Facultad de Informática. Postgrado

Universidad Nacional de La Plata

Tesis presentada para obtener el grado de
Doctor en Ciencias Informáticas

Análisis de datos educativos aplicado en el estudio de la incidencia de factores socioeconómicos en el rendimiento escolar

Autor

JORGE IVÁN PINCAY PONCE

Director

ING. ARMANDO EDUARDO DE GIUSTI

Junio de 2023

Resumen

La investigación que corresponde con esta tesis se desarrolló en el campo de la Minería de Datos Educativos, en un sentido más amplio, en la Ciencia Informática aplicada en la Educación. El documento articula el análisis de datos con el problema multifactorial del rendimiento académico en las escuelas.

Así, el objetivo general es el análisis de la incidencia de los factores socioeconómicos en el aprovechamiento académico a nivel escolar, con la finalidad de contribuir a su entendimiento y mejora, mediante la aplicación de modelos de análisis de datos predictivos o supervisados y descriptivos o no supervisados. También se ha incluido un análisis confirmatorio que tiene relaciones entre sus elementos, a priori sustentados en las exploraciones estadísticas de los datos que anteceden al desarrollo de los modelos supervisados y no supervisados y también a dichos modelos.

Los datos objeto de estudio corresponden a dos escuelas de Ecuador, dado que la cantidad de datos entre una y otra difería considerablemente no se presenta un análisis comparativo, sino uno con base en la información consolidada que totaliza 6808 instancias o registros de calificaciones y 88 columnas que lo describen. El análisis gira en torno a cada registro de calificaciones y no de cada alumno, porque en el sistema escolar ecuatoriano las bajas calificaciones en una materia, simplificadas como rendimiento académico, pueden llegar a determinar la reprobación del año básico cursado por el alumno.

El proceso de análisis ejecutado es iterativo, permite ir hacia adelante y hacia atrás entre las fases que lo componen, siempre que resulte necesario tener mejores resultados. Se basa en el ciclo de vida conocido como CRISP-DM, siglas del Proceso Estándar Intersectorial para Minería de Datos. Además, se adicionó algunas prácticas sugeridas en el Proceso Estándar Intersectorial para el Desarrollo de Aplicaciones de Aprendizaje Automático con Metodología de Garantía de Calidad o CRISP-ML (Q), cómo, por ejemplo, cumplir con requisitos que promuevan la calidad de datos, robustez del modelo y evaluación de riesgos, para así aminorar problemas de sesgo, sobreajuste y falta de reproducibilidad de los modelos hacia nuevas escuelas y regiones.

Se utilizó el modelado predictivo para ayudar a las instituciones educativas con la identificación temprana de los estudiantes con dificultades para sostener su

rendimiento académico escolar. Se desarrolló modelos predictivos que utilizan datos de calificaciones, factores socioeconómicos y de comportamiento de los estudiantes, mismos que se han recopilado de sistemas provistos por el Estado y del departamento de orientación estudiantil de las escuelas ecuatorianas. Con ello se buscó clasificar con precisión si un estudiante está en riesgo de reprobar un curso o experimentar problemas en cierta materia del curso. La identificación de patrones de estudiantes en riesgo es de ayuda a los docentes y más actores educativos en la toma de medidas proactivas que favorezcan la participación efectiva en las aulas de clases y en que se aminore las eventuales brechas educativas relacionadas con el rendimiento académico.

Se recurrió a 13 modelos supervisados, 5 no supervisados y un análisis confirmatorio. La relación entre los resultados obtenidos a partir de ellos guarda consistencia. Los datos fueron estudiados desde cinco ejes (1) Modelos no supervisados, (2) Modelos de clasificación considerando notas intermedias, (3) Modelos de clasificación sin considerar notas intermedias, (4) Modelos de regresión sin considerar notas intermedias y (5) Modelos de clasificación con datos reducidos en su dimensionalidad, balanceados y sin considerar notas intermedias.

Cuando no se incluyó a las notas intermedias fue porque era de esperar que el promedio final se vea muy influenciado por las calificaciones progresivas de los alumnos, por tanto, la no inclusión de dichas calificaciones ilustra de mejor manera la incidencia de los factores socioeconómicos sobre el rendimiento académico. Existen calificaciones que en el sistema escolar ecuatoriano se registran, pero no condicionan la aprobación del año básico por parte del alumno, estas son el comportamiento de cada alumno y la calificación de su participación en los denominados proyectos escolares, que tienen como finalidad evaluar a las habilidades sociales de los alumnos. Con la reducción de la dimensionalidad se favoreció los tiempos de entrenamiento de los modelos supervisados a la par de prevenir la indisponibilidad de ciertos datos para los análisis posteriores.

La información resultante de los modelos se combinó con el aporte de la revisión sistemática de la literatura. De modo general, los métodos de ensamblado reportaron los mejores valores en las diversas métricas, entonces, los resultados de las clasificaciones y regresiones logradas son confiables y no casuales, reflejan los patrones en los datos, porque en tales métodos de ensamblado se empleó 50 estimadores basados en árboles de decisión. Como referencia a una métrica, la Exactitud de la clasificación siempre superó el 90% y las regresiones tuvieron una

efectividad de hasta el 85% porque las predicciones de promedios en los mejores casos pueden efectuase con un error de hasta 1.5 puntos sobre 10 posibles.

En esta investigación doctoral, se ha combinado la objetividad de las métricas en las tareas de clasificación y regresión, con la subjetiva pero importante interpretabilidad de los resultados, apoyados en estudios referidos a técnicas de puntuación de características y su respectiva ilustración visual, con ello se ha pretendido que los modelos resulten interpretables por los usuarios posibles al tiempo de fortalecer su confianza en las decisiones de los modelos de las instituciones escolares.

Parte de los resultados obtenidos muestran que los alumnos que no alcanzan los aprendizajes requeridos, es decir, que obtienen las calificaciones más bajas posibles, tienen como tendencia a un padre en estado civil de unión libre, un bajo número de hermanos, suelen presentar alguna discapacidad, su comportamiento en principio es A o el más alto, pero tiende a bajar conforme avanza el periodo lectivo, en sus proyectos escolares tienen una muy buena calificación B pero que no es la mejor A, su padre suele tener una ocupación laboral informal (por ejemplo, guardia de seguridad), el ingreso familiar suele ser bajo y también suelen vivir en familias reconstruidas.

A futuro, estudios como el presente pueden ser fortalecidos con la incorporación de más escuelas de distintas regiones para obtener un abordaje más significativo por disponer de más datos y así producir resultados más fiables y extrapolables.

Agradecimientos

A Dios,

por ser siempre la fuente de mi paz y esperanza.

Al Ing. Armando De Giusti,

por haber aceptado ser mi tutor y haberme compartido sus conocimientos,

por cada detalle que me mostró y por cada momento que me orientó en el largo camino recorrido.

A la Dirección de Postgrado de la Facultad de Informática y a toda la UNLP,

por haberme permitido ser parte de tan prestigiosa institución.

A la Universidad Laica Eloy Alfaro de Manabí,

en especial a las autoridades y leales amigos de las carreras de Tecnologías de la Información,

por aportar de varias maneras en este camino de formación doctoral.

A las unidades educativas,

que facilitaron los datos para desarrollar esta investigación y siempre estuvieron prestos a los diversos requerimientos de esta.

A los amigos ecuatorianos con los que inicié estos estudios, por cada momento de formación y camaradería compartido.

Dedicatoria

En especial a mis padres,

que de muchos modos comprenden el esfuerzo de esta travesía, me expresan su apoyo, su amor y eso ha significado para mi mucha fortaleza, sacrificio y compromiso.

A mi familia en general,

que fue un grato complemento con sus muchos mensajes y frases de aliento.

Tabla de contenidos

Índice Tablas

Índice de Figuras

Índice de Gráficos

Aspectos de redacción

Se ha empleado como separador decimal al punto, dado que en el documento se recurre a diversas capturas de pantalla de las herramientas de software utilizadas y estas utilizan el punto como separador decimal.

Se ha incorporado referencias cruzadas en el documento con el fin de facilitar su navegación y legibilidad entre diversas secciones.

Algunos gráficos generados con Orange son de considerable resolución, aunque se han realizado e insertado capturas de las partes más pertinentes en este documento, en algunos casos se ha preferido almacenarla para su disponibilidad desde una carpeta disponible en https://tinyurl.com/5n7sh6ta.

Capítulo 1:

1. Introducción

La educación es un elemento estratégico para el desarrollo y bienestar de cualquier sociedad, por tanto, si las personas no tienen acceso a la educación sus logros educativos serán limitados y esto impacta en toda la sociedad en la que viven. Por ello, las posibles soluciones a algunos de los problemas multifactoriales actuales que se presentan a los estudiantes, como el bajo rendimiento en las materias que se estudian en la escuela, la reprobación de los años básicos cursados o incluso el abandono escolar, requieren del aporte de todos los actores que intervienen en el proceso educativo.

En 2020, en el reporte de las Naciones Unidas acerca de la Educación durante la pandemia del COVID-19 y sus impactos disruptivos en el mundo, se invocó a la aceleración de los cambios positivos en la educación que pueden resultar de la experiencia de la pandemia y que lleven al incremento del aprendizaje y la disminución de la deserción estudiantil, de modo especial por parte de los alumnos de sectores sociales más desfavorecidos (De Giusti, 2020).

En 2022, de nuevo la UNESCO (2022), informó que a nivel mundial cada vez hay más niños que están más alejados de la educación, que corren el riesgo de reprobar años, no terminar sus estudios y no aprender en la escuela, pese a que la educación es un derecho humano de incidencia en el crecimiento económico. Los niños que no completan su educación primaria pueden afectar sus perspectivas y satisfacción en lo laboral, así como en sus comportamientos y salud infantil, que a su vez impacta su posterior rol en la sociedad y la formación de sus familias.

En ese sentido, la principal motivación de esta tesis es el análisis de datos educativos aplicado en el estudio de la incidencia de factores socioeconómicos en el rendimiento escolar a partir de datos de dos escuelas ecuatorianas, para ello se ha empleado metodológicamente los principios, técnicas y algoritmos de aprendizaje automático apropiados a las características de los datos. Con estas acciones se ha

conseguido conocimiento válido para el actual y nuevos conjuntos de datos, comprensible por parte de escuelas y útil para la futura toma de decisiones o acciones. Estas consecuciones se respaldan con las métricas pertinentes y favorables que reportan los modelos de aprendizaje automático.

Los datos objeto de estudio se corresponden con una muestra transversal de calificaciones y factores socioeconómicos de dos escuelas ecuatorianas.

1.1. Motivación

1.1.1. El problema del rendimiento escolar

Para muchos niños, dejar la escuela es el paso final en un largo proceso de desconexión gradual y reducción de la participación en el currículo escolar formal, así como en la vida social de la escuela. El detonante de la deserción puede asociarse a la reprobación o no aprobación de una o más materias a causa de un insuficiente rendimiento cuantitativo o cualitativo.

En la actualidad las escuelas están interesadas en que sus estudiantes aprueben y avancen en su formación y no la abandonen. Toman como referencia una serie de políticas y programas para evitar el fracaso escolar, sin embargo, este es un problema multifactorial. Entre aquellos factores están los socioeconómicos.

Las instituciones relacionadas con la educación, así como otros sectores, empiezan a concebir el análisis de la creciente cantidad de almacenes de datos, como una forma posible de garantizar una mayor calidad, eficiencia e inclusión educativa, sin embargo, en muchas de ellas se recopilan cantidades considerables de datos porque las personas los han estado almacenando durante muchos años, en lugar de por unas razones intencionales de analítica de relaciones entre los puntajes y factores que los afectan de modo clave (Q.-K. Fu & Hwang, 2018; McNeish & Wolf, 2020; Schildkamp, 2019; Soncin & Cannistrà, 2022).

Según la UNESCO (2017), los datos y la información pueden sustentar una mejor responsabilidad externa e interna de las escuelas, a lo que se adiciona la concepción que tiene la Organización para la Cooperación y el Desarrollo Económico, (OCDE, 2016), de que, con oportunidades educativas significativas, los estudiantes que se sepa son menos favorecidos tienen probabilidades de mejorar, permanecer y aprovechar al máximo su educación escolar.

Entonces, con base en la disponibilidad de datos que contienen las condiciones

socioeconómicas y el nivel de aprendizaje alcanzado por los estudiantes de dos escuelas ecuatorianas, la pregunta que direcciona esta investigación es la siguiente:

¿Cómo inciden los factores socioeconómicos en el aprovechamiento escolar?

1.1.2. Análisis de datos educativos

La masificación de los datos está configurando diferentes sectores debido al creciente número de sistemas automatizados, que almacenan datos de diferentes fuentes y el sector de la educación es uno de los más involucrados, dado el potencial subyacente en los datos para apoyar la enseñanza y el aprendizaje efectivo que mejoren las formas en que las generaciones futuras construirán la realidad con y a través de los datos (Soncin & Cannistrà, 2022).

En tal sentido Schildkamp (2019), considera que las escuelas han comenzado a manejar el análisis como forma posible de garantizar una mayor calidad que propicie la eficiencia y la inclusión, como objetivos cruciales de cualquier institución educativa, sin embargo, sucede que en muchas escuelas se recopilan cantidades considerables de datos durante muchos años, pero, sin unas razones de analítica de datos predefinidas.

La revisión periódica de los hábitos de estudio, la investigación de los temas abordados en cada asignatura y el afán de observación para poder analizar las condiciones en las que se desarrollan las actividades de los estudiantes, permiten mejorar la eficiencia del aprendizaje de los alumnos y en consecuencia su rendimiento, entendido como mejores calificaciones. En la actualidad es importante extraer información significativa de una gran cantidad de datos y en ese sentido según Sun (2022), la minería de datos con enfoque educativo como disciplina emergente, se ocupa de desarrollar métodos para explorar los datos a gran escala que provienen de entornos educativos y así comprender mejor a la enseñanza, a la gestión de los maestros y en especial a los estudiantes.

De acuerdo con Tarik y colaboradores (2021), un número importante de estudiantes experimenta diversos tipos de dificultades en su vida escolar y debido a estas, adoptan comportamientos antisociales o en lo posterior se sitúan en la franja más baja del mercado laboral, lo que se suma a sus dificultades de adaptación con la sociedad.

Ahora, si bien, se debe reconocer la proliferación de muchos sistemas de gestión de la información educativa, la utilización de los datos allí almacenados se limita a

consultas o cálculos de promedios, de rangos u otros, pero no se suelen buscar relaciones entre los puntajes o, determinar factores que afectan de modo clave el rendimiento de los estudiantes o cuales serían las posibles tendencias de desarrollo futuro a partir del rendimiento y del comportamiento actual (Q.-K. Fu & Hwang, 2018; McNeish & Wolf, 2020).

A inicios de la década de 1990 un estudio de Entwisle y Alexander (1992), señaló que, por ejemplo, el aprovechamiento en matemáticas tiene relación con las diferencias en la situación económica familiar, seguidas de la segregación escolar, aunque también puso en el escenario que las configuraciones familiares biparentales o padre-presente versus monoparental o padre-ausente, son probablemente insignificantes como causa cuando el estatus económico sí es favorable.

Una revisión sistemática de la literatura efectuada en 2018 amplió la lista de estos factores de incidencia con habilidades psicomotoras, el rendimiento antes y en el transcurso del curso, la participación de los estudiantes, la demografía de los estudiantes, el género y la autorregulación. Sin embargo, tal cual denota la misma revisión, las tasas de deserción se vieron influenciadas principalmente por la motivación de los estudiantes, los hábitos, los problemas sociales y financieros (Hellas et al., 2018).

A decir de varios autores, que han evaluado las asociaciones entre el momento y la duración del nivel socioeconómico bajo durante la infancia y el rendimiento académico en la escuela, la educación tiene implicaciones de por vida para el bienestar (Liu et al., 2020; van Zwieten et al., 2021).

En épocas recientes, investigadores como Asif y colaboradores (2017) y Yağcı (2022) se han enfocado en el momento oportuno de emplear las predicciones del rendimiento académico de los estudiantes, siendo sus opciones al final de un programa de estudio de cuatro años o examinar el desarrollo de los estudiantes y combinarlos con resultados predictivos, entre sus encuentros claves estos autores han descubierto que es importante que los educadores se centren en un pequeño número de materias que indiquen un rendimiento particularmente bueno o malo para ofrecer advertencias oportunas, apoyar a los estudiantes de bajo rendimiento y ofrecer consejos y oportunidades a los estudiantes de alto rendimiento.

Tal cual se ha indicado, varios de los trabajos mencionados, que datan de fechas recientes, no se han centrado específicamente en el nivel escolar, ni menos en lo que sería un análisis de incidencias o afectaciones de los factores socioeconómicos en

determinados momentos de un curso escolar o más en concreto de una materia que forme parte del curso, nivel o grado escolar.

En lo que respecta al sistema escolar ecuatoriano, con el afán de comprenderlo, este consta de seis años de la llamada educación básica, los grados 2, 3 y 4 corresponden a la Educación Básica Elemental (EBE) y los grados 5, 6 y 7 corresponden a la Educación Básica Media (EBM) (Anchundia-Delgado et al., 2022). Los seis grados de la educación básica también se relacionan con los llamados procesos del Sistema de Educación Intercultural Bilingüe (SEIB), ideado para fortalecer la calidad de la educación con pertinencia cultural y lingüística a fin de desarrollar las habilidades y destrezas cognitivas, psicomotrices y afectivas de los estudiantes de nacionalidades y pueblos en las instituciones educativas interculturales bilingües. La referida relación se muestra en la siguiente figura:

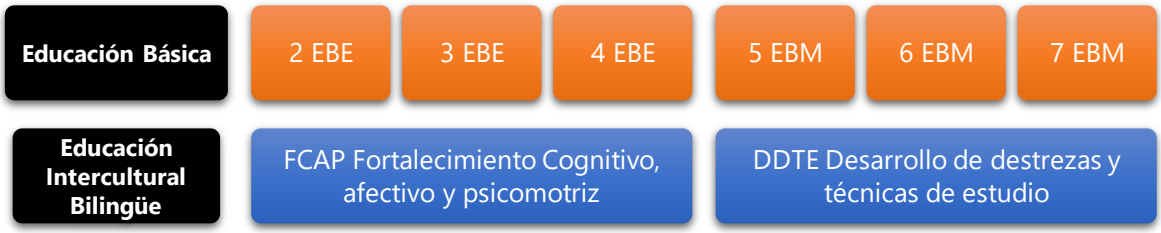


Figura 1: Relación entre el Sistema Nacional de Educación y el del Sistema de Educación Intercultural Bilingüe
Fuente: Ministerio de Educación de Ecuador (2016)

Es de indicar que los datos objeto de análisis en esta tesis, una vez preparados contienen las condiciones socioeconómicas y el nivel de aprendizaje alcanzado por los estudiantes de dos escuelas ecuatorianas, con base en la escala de calificaciones propuesta por el Ministerio de Educación de Ecuador (2013), misma que se muestra en la **Tabla 1**:

Tabla 1: Escala de calificaciones propuesta por el Ministerio de Educación de Ecuador

Nº	Siglas	Significado	Rango numérico
1	**DAR**	Domina los aprendizajes requeridos	desde 9.00 hasta 10.00
2	**AAR**	Alcanza los aprendizajes requeridos	desde 7.00 a 8.99.
3	**PAAR**	Próximo a alcanzar los aprendizajes requeridos	desde 4.001 a 6.99.
4	**NAAR**	No alcanza los aprendizajes requeridos	menos o igual que 4.

Fuente: Ministerio de Educación (2013)

De modo general, el conjunto de datos que se emplea en esta tesis guarda calificaciones del promedio anual de los alumnos en cada una de siete materias. El promedio anual deriva de 2 promedios quimestrales y dicho promedio deriva de 3

promedios parciales por quimestre. En los aspectos socioeconómicos se almacena información sobre: Tipo de familia, ingreso mensual familiar, parentesco del representante, sector de dirección domiciliaria, cantidad de personas que viven en casa, escolaridad alcanzada por los padres y los servicios básicos de los que se dispone en el hogar.

Sobre la disponibilidad de los servicios básicos que incluyen computador, internet y dispositivos móviles relacionados, a decir de la UNESCO (2021a), la reciente pandemia del COVID-19 reveló la fragilidad y la falta de preparación de los sistemas educativos, pues casi un tercio de los estudiantes del mundo, a octubre de 2020, no tenían acceso a la educación a distancia, en parte debido a la falta de los dispositivos necesarios para que se puedan conectar desde el hogar o simplemente a la falta de conectividad, así como de sus habilidades digitales para acceder a contenidos pedagógicos que requieran del uso de la tecnología.

En este sentido, durante y después de la pandemia, el desafío es potenciar normas sociales, la seguridad, la privacidad en línea, las habilidades y demás pautas para conectar a niños y jóvenes con soluciones digitales que puedan brindar experiencias personalizadas y aprendizajes relevantes (De Giusti, 2020; UNESCO, 2021a).

Con el análisis de datos y tal como se indica en la sección de Objetivos de este documento, se aspira a comprender la influencia de estos factores en el aprovechamiento escolar, lo que está en concordancia con lo concluido por la Organización para la Cooperación y el Desarrollo Económico, de que, con oportunidades educativas significativas, los estudiantes menos favorecidos tienen más probabilidades de permanecer en el sistema escolar y aprovechar al máximo su educación (OCDE, 2016), entre tanto que la UNESCO (2017) destaca que los datos y la información pueden sustentar una mejor responsabilidad externa e interna de las escuelas.

Varios de los trabajos mencionados en esta motivación, que datan de fechas recientes, no se han centrado específicamente en las escuelas, ni menos en lo que sería un análisis de incidencias o afectaciones de los factores socioeconómicos en determinados momentos de un curso, nivel o grado escolar, tal cual es el aporte de la presente propuesta de tesis, además de determinar las características de parámetros e hiperparámetros adecuados a los modelos en un escenario como el indicado.

1.1.3. Aprendizaje Automático, Minería de Datos

El Aprendizaje Automático es una rama de la Inteligencia Artificial que estudia sistemas capaces de aprender a realizar tareas a partir de datos. Es de naturaleza inductiva, a diferencia de la inteligencia artificial clásica y comprende técnicas y métodos para realizar clasificación, optimización y predicción, en dominios en donde por lo general los problemas no pueden programarse de forma explícita o no existen soluciones analíticas aplicables (F. M. Quiroga, 2020).

En años recientes, se han empleado diversos modelos que implican algoritmos de aprendizaje automático en el campo educativo, más en concreto, algoritmos de minería de datos cada vez más precisos desde la mirada informática y estadística (Hellas et al., 2018; Krumm et al., 2018; Li & Zhai, 2018).

El referido campo de investigación relacionado con la aplicación de la minería de datos, recientemente se lo denomina minería de datos educativos, este es un campo que busca desarrollar y mejorar los métodos para explorar estos datos, que a menudo tienen múltiples niveles de jerarquía significativa, con el fin de descubrir nuevos conocimientos sobre cómo las personas aprenden en el contexto de ciertos entornos (Fernandes et al., 2019; Pincay-Ponce et al., 2022). Además, de que con su aplicabilidad se ha contribuido a las teorías del aprendizaje y se ha fortalecido la analítica del aprendizaje comparando y contrastando datos (Baker, 2010; Siemens & Baker, 2012).

En algunos estudios efectuados en el campo, según Han y colaboradores (2011), cuando se ha dispuesto de una variedad de datos y el enfoque sea analizar las causas del éxito o el fracaso, se pueden emplear métodos estadísticos como la regresión logística y las series temporales; cuando el enfoque sea el pronóstico se pueden emplear redes neuronales, máquinas vectoriales de soporte, bosques aleatorios y los mismos árboles de decisión.

En esta tesis, se prueban diversos algoritmos en la construcción de modelos de análisis de datos, tanto con sus configuraciones de parámetros e hiperparámetros y la valoración de sus métricas de evaluación, así como un apropiado proceso previo de preparación de los datos.

1.2. Objetivos

El objetivo general en esta tesis es el análisis de la incidencia de los factores socioeconómicos en el aprovechamiento académico a nivel escolar y de ese modo

contribuir a su entendimiento y mejora, mediante la aplicación de modelos de análisis de datos.

Para ello se establecieron los siguientes objetivos específicos:

1. Reconocer las aplicaciones de análisis de datos en los problemas del contexto educativo escolar.
2. Preparar los datos acordes con la dimensionalidad a un número efectivo de características bajo consideración.
3. Estudiar comparativamente algoritmos de minería de datos en función del proceso de minería de datos que resulte idóneo.
4. Establecer parámetros e hiperparámetros apropiados a los datos y los modelos.
5. Interpretar los resultados del conocimiento descubierto y su eficiencia según métricas pertinentes a los modelos.

Debido a la existencia de múltiples algoritmos de aprendizaje automático, tanto supervisados, no supervisados y de refuerzo para lograr construir modelos en diversos contextos, el alcance de esta tesis se limita a su empleo en el análisis de datos educativos y socioeconómicos en la etapa de formación escolar y no propone nuevos algoritmos de aprendizaje automático.

1.3. Alcance

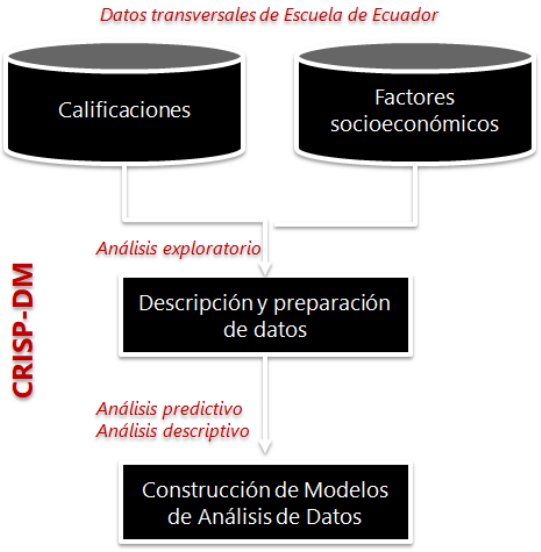


Figura 2: Alcance de la tesis

Esta investigación transversal que parte de la consolidación de una muestra de datos de alumnos de primaria de dos escuelas ecuatorianas, usa los promedios finales de cada alumno, sus componentes individuales y un conjunto de datos de factores socioeconómicos correspondientes con cada alumno.

Luego, se emplea análisis de datos, para comprender las influencias de los factores socioeconómicos en su aprovechamiento escolar y contribuir al mejor abordaje de este problema multifactorial.

Los tipos de análisis de datos empleados son exploratorio, predictivo y descriptivo. Esta selección se ilustra en la **Figura 2**. El desarrollo del componente práctico de la tesis sigue las etapas sugeridas en el Proceso Estándar Intersectorial para Minería de Datos CRISP-DM. Si bien el rendimiento académico es un objeto de estudio en todos los niveles del sistema educativo, en esta investigación se lo aborda a nivel escolar.

1.4. Metodología

1.4.1. Tipo de investigación

La investigación seguida en esta tesis es mixta, es decir, aplicada con detalles descriptivos. Aplica algoritmos y técnicas de análisis de datos conocidos siguiendo un ciclo de vida también conocido como lo es CRISP-DM, el Proceso Estándar Intersectorial para Minería de Datos. Luego se construyen modelos para el análisis de la incidencia de los factores socioeconómicos en el aprovechamiento académico a nivel escolar y de ese modo se contribuye a su entendimiento y mejora.

1.4.2. Ciclo de vida de los modelos de Aprendizaje Automático

Se utilizó el Proceso Estándar Intersectorial para Minería de Datos, conocido también como Metodología CRISP-DM por sus siglas en inglés y surgido en 1996. CRISP-DM proporciona una descripción normalizada del ciclo de vida de minería de datos, tal cual se observa en la **Figura 3** y es adaptable al análisis de datos en general. Además, sus fases iterativas y exploratorias permiten ir hacia adelante y hacia atrás siempre que resulte necesario tener mejores resultados en los negocios o la investigación (Azevedo & Santos, 2008; Wirth & Hipp, 2000).

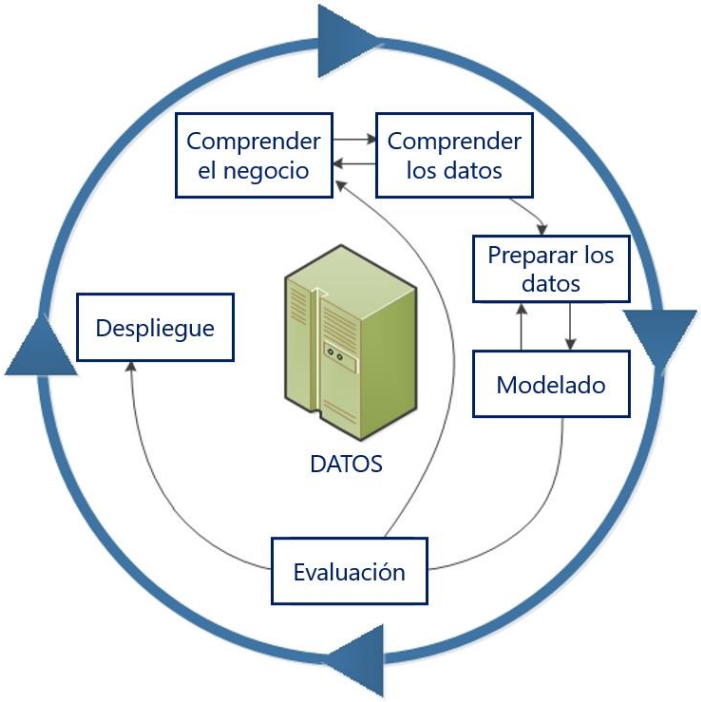


Figura 3: Proceso Estándar Intersectorial para Minería de Datos, CRISP-DM.

Tabla 2: Actividades generales en cada fase de CRISP-DM.

Fase	Actividades generales
[1] Comprender el negocio	En lo concerniente a esta investigación, se documenta el aprovechamiento escolar y lo concerniente de su contexto para con los objetivos del proyecto, luego se define el problema de minería de datos y se construye un cronograma preliminar junto con el reconocimiento de responsabilidades, posibles riesgos tecnológicos que retrasen el cumplimiento de los objetivos.
[2] Comprensión de datos	En lo concerniente a esta investigación, se recopilan los datos de las calificaciones y los factores socioeconómicos de los alumnos, se familiariza con ellos con acciones como identificar múltiples problemas, descubrir conocimientos previos sobre los datos o encontrar subconjuntos de interés para formular hipótesis sobre lo oculto.
[3] Preparación de datos	Se crea un conjunto de datos final a partir de los datos iniciales sin procesar. Las tareas incluyen la selección de tablas, conjuntos de datos y atributos y la transformación y limpieza de datos para su posterior modelado.
[4] Modelado	Se seleccionan y aplican métodos de modelado relacionados con el problema, tantos como sea posible, al tiempo de ajustar sus parámetros a valores óptimos, pues existen varios métodos para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos para el formato de datos, que suelen requerir de nuevos pasos de preparación de datos en cualquier proyecto.
[5] Evaluación	Se valoran los modelos creados que parecen tener suficiente calidad desde el punto de vista del análisis de datos o en función de las métricas. Se revisan los pasos de

	creación de los modelos resultantes y se contrastan con los objetivos comerciales para definir si los problemas se abordaron adecuadamente. Al final de esta etapa, se decide sobre la implementación de los resultados del proceso de análisis de datos.
[7] Despliegue	Se organizan y presentan los conocimientos adquiridos para que las organizaciones puedan utilizarlos, mediante informes, procesos de análisis de datos cíclico o se automatiza lo pertinente en una organización.

Fuente: Elaboración propia a partir de IBM (2021), Azebedo y Santos (2008).

Cada una de las fases resumidas en la **Tabla 3** se detallan y contextualizan en el capítulo 3, correspondiente al Desarrollo del Análisis de Datos de esta investigación.

En cada fase de CRISP-DM se incorporaron algunas prácticas sugeridas en el Proceso Estándar Intersectorial para el Desarrollo de Aplicaciones de Aprendizaje Automático con Metodología de Garantía de Calidad o CRISP-ML (Q) por sus siglas en inglés. Los investigadores que lo propusieron en 2021 sugieren que a cada fase se le definan requisitos de calidad de datos, robustez del modelo y evaluación de riesgos. Para así aminorar problemas de sesgo, sobreajuste y falta de reproducibilidad de los modelos (Studer et al., 2021).

Si bien CRISP-ML (Q) tienes sus propias fases: (1) Comprensión del negocio y los datos, (2) Ingeniería de datos, (3) Ingeniería de modelos de aprendizaje, (4) Garantía de calidad para aplicaciones de aprendizaje automático, (5) Despliegue y (6) Monitoreo y mantenimiento, en esta tesis se siguen las fases de CRISP-DM con los agregados de calidad sugeridos en CRISP-ML (Q).

1.4.3. Conjuntos de datos

Los datos se han recopilado de dos fuentes: (1) Las calificaciones se descargan en formato CSV desde un sistema proporcionado por el Estado y (2) Las fichas socioeconómicas se obtienen del Departamento de Consejería Estudiantil (DECE) en formatos DOCX. Respecto de las calificaciones, cada alumno tiene varias instancias o registros asociados con él, que se corresponden con las calificaciones de cada materia que estudia y cuyos principales componentes se observan en la **Figura 4**.

El DECE es la instancia responsable de la atención integral de los estudiantes, brinda apoyo y acompañamiento psicológico, psicoeducativo, emocional y social, en concordancia con el marco legal vigente en Ecuador. Los datos que de modo principal guarda el DECE acerca de los alumnos son los mostrados en la **Figura 5**.

En principio para el conjunto de datos se dispuso de datos de 557 estudiantes y un total de 6808 registros o instancias de calificaciones y de factores socioeconómicos

asociados. Estos datos aún requerían del proceso de preparación para su empleo con los modelos de análisis de datos.

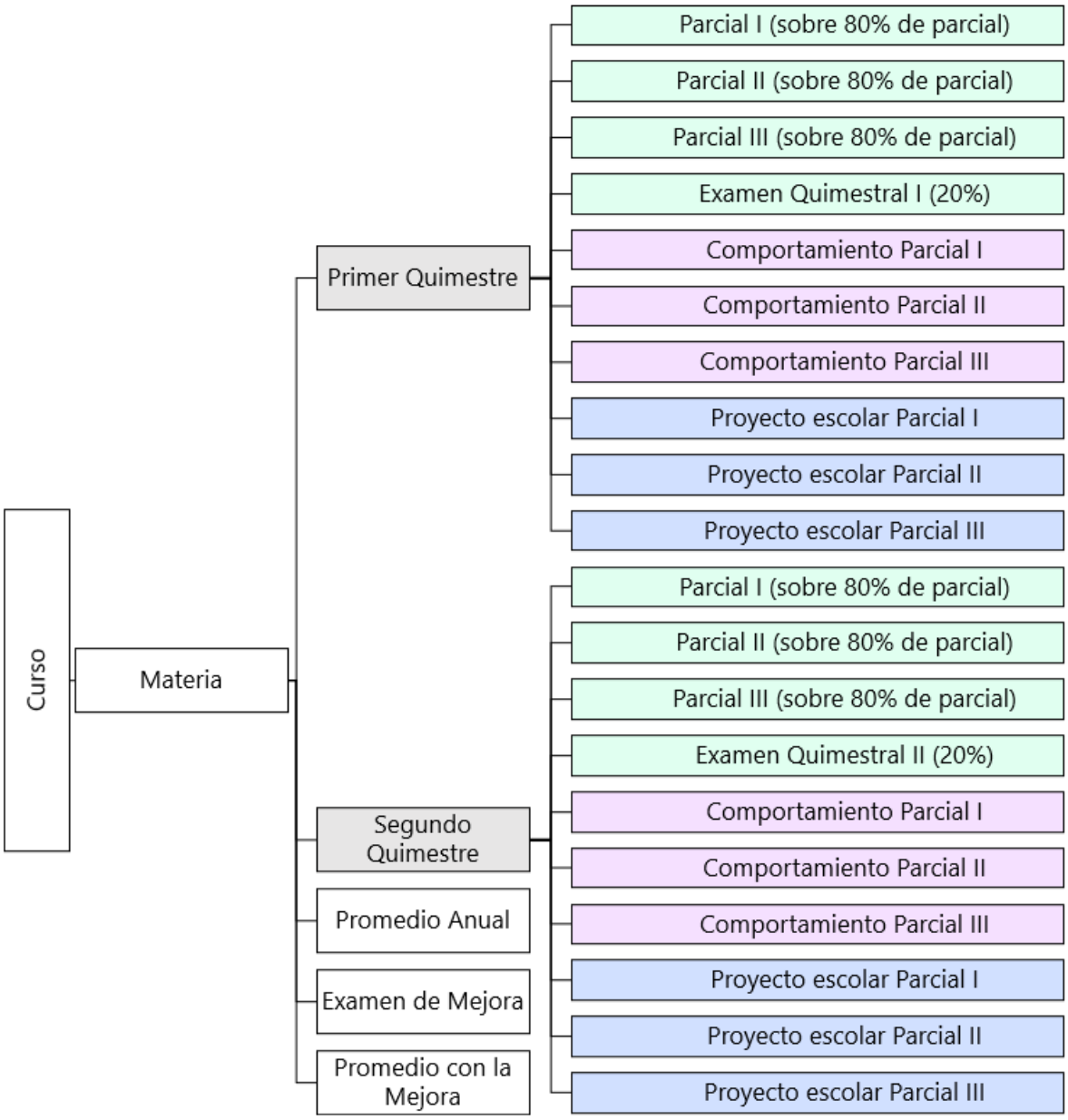


Figura 4: Componentes de las calificaciones de los alumnos.

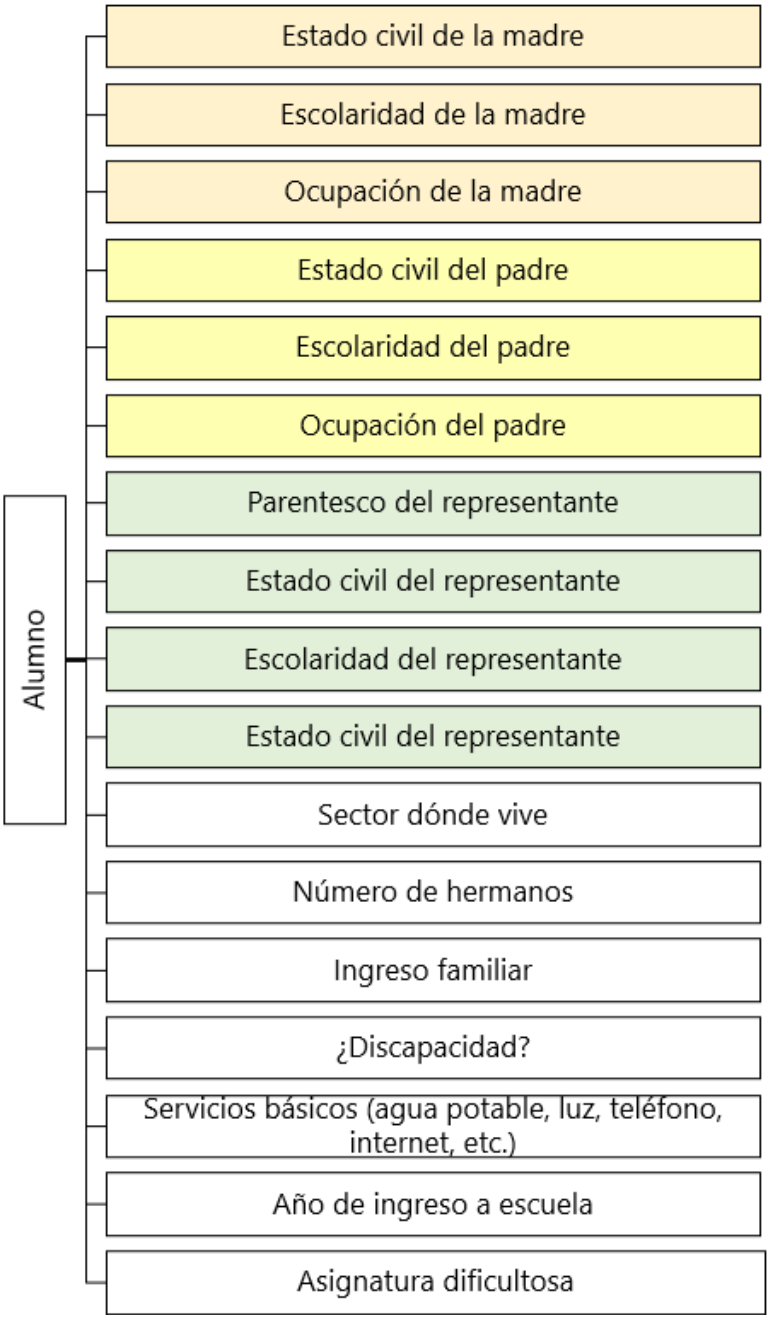


Figura 5: Datos que el DECE guarda de cada estudiante.

1.4.4. Niveles de análisis de datos

José Supo (2017) plantea los siguientes métodos de análisis de datos: (1) De nivel exploratorio, (2) De nivel descriptivo, (3) De nivel relacional, (4) De nivel explicativo, (5) De nivel predictivo y (6) De nivel aplicativo. En esta tesis doctoral se emplean los análisis exploratorio, descriptivo y predictivo, tal cual se detalla en la **Tabla 3.**

Tabla 3: Niveles de análisis de datos empleados en la tesis doctoral

Análisis de datos	Usos frecuentes
Análisis exploratorio de datos	Empleado para resumir características principales del conjunto de datos, utilizando gráficos estadísticos y otros métodos de visualización de datos. El análisis exploratorio de datos ha sido promovido por John Tukey desde 1970 para alentar a la exploración de los datos y formular hipótesis que podrían conducir a una nueva recopilación de datos (Baillie et al., 2022). Como ejemplos de este tipo de análisis, en el capítulo 3 se realiza exploraciones del rendimiento académico con base en las enfermedades auto reportadas, dificultades auto reportadas en las asignaturas, ingresos familiares, estructura familiar, familias reconstruidas, entre otros análisis.
Análisis descriptivo de datos	Empleado para extraer datos y especificar datos actuales sobre eventos pasados. Olson y Lauhoff (2019) consideran que el análisis de la cesta de compra, asociaciones, agrupaciones, visualizaciones y el análisis de enlaces, son los ejemplos más recurridos en cuanto a uso de este análisis. Como ejemplos de este tipo de análisis, en el capítulo 3 se realiza perfiles de datos de alumnos en riesgo de bajar su rendimiento, de alumnos por calificaciones cualitativas obtenidas, entre otros análisis.
Análisis predictivo de datos	Empleado para predecir valores de nuevas instancias o ejemplos. Esos valores pueden ser numéricos (regresión) o categóricos (clasificación). Varias técnicas clásicas de aprendizaje automático se pueden usar tanto para la regresión como para la clasificación. La calidad de los modelos predictivos se evalúa midiendo su poder predictivo, es decir, qué tan bien predicen los valores de las nuevas instancias de prueba que no se usan para entrenar el modelo, para conjuntos de datos pequeños, generalmente se emplean la llamada validación cruzada (Vodencarevic & Fett, 2015). Como ejemplos de este tipo de análisis, en el capítulo 3 se realizan clasificaciones y regresiones con base en las características socioeconómicas y los diversos momentos del periodo lectivo, como lo son los tres promedios parciales de cada uno de los dos quimestres del periodo lectivo escolar.

1.4.5. Consideraciones éticas

Como parte del desarrollo de la ética en la investigación, se proporcionó una declaración escrita en lenguaje sencillo a los propietarios de los datos para obtener su consentimiento por escrito de usar los datos para el desarrollo de la tesis doctoral. Como parte del acuerdo, el tesista firmó un acuerdo de confidencialidad, en el que se aclaró que solo se podían usar los datos anonimizados para el propósito de su investigación y las publicaciones académicas de los resultados de la investigación. Ver **Anexo 15.**

1.5. Contribuciones

En esta sección, se presenta de forma resumida las contribuciones principales de la tesis:

1. Se trata de hacer una contribución al área de análisis de datos para el mejoramiento educativo, teniendo en cuenta información disponible en Ecuador, en el área donde se desempeña el Tesista. Los resultados pueden ser significativos a nivel local pero también se pueden proyectar en forma general, dado el impacto del aspecto socioeconómico en el rendimiento académico.
2. Se analiza datos educativos tanto de rendimiento académico como socioeconómicos en una etapa temprana de formación como lo es la escolar. Además de que se revisa no sólo las notas al final de los periodos lectivos, sino también los avances progresivos, capaz de que, a los actores implicados, especialmente representantes de los alumnos, se les genere recomendaciones tempranas y que estos tomen acciones sustentadas en datos, que disipen posibles afectaciones o brechas en el aprovechamiento escolar por causas socioeconómicas.
3. En términos de experimentos se exploran y configuran los datos con criterios de ingeniería de datos y con base en varias hipótesis y sus respectivas pruebas en varios modelos de clasificación y predicción, hasta evidenciar métricas favorables que sugieran que los resultados son susceptibles de convertirse en recomendaciones y por ende en una apropiación social de la tecnología para lograr beneficios trascendentes en la vida escolar.

1.6. Publicaciones

Las siguientes publicaciones están relacionadas directamente con las contribuciones que se presentan en esta tesis:

- Pincay Ponce, Jorge Iván y colaboradores <<Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar>>" Revista Ibérica de Sistemas e Tecnologias de Informação. ISSN 1696-9895. Nº E49. (2022): 654-667. Ver **Anexo 1**.
- Pincay Ponce, Jorge Iván y colaboradores <<Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar. Revisión sistemática>>" Revista Ibérica de Sistemas e Tecnologias de Informação. ISSN 1696-9895. Nº E53. (2023): 654-667. Ver **Anexo 2**.

- Pincay-Ponce, Jorge Iván y colaboradores <<CatBoost: Aprendizaje automático en conjunto para la analítica de los factores socioeconómicos que inciden en el rendimiento escolar>>. Enviado a la Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología (TE&ET).

Las siguientes publicaciones constituyen trabajos publicados antes y durante la tesis:

- Pincay Ponce, Jorge Iván y colaboradores <<Análisis de implementaciones de sistemas tutores inteligentes y afectivos. Revisión sistemática>>. REFCalE: Revista Electrónica Formación y Calidad Educativa. ISSN 1390-9010. Vol. 7. Nº 2 (2019): 218-234. Ver **Anexo 3**.
- Anchundia-Delgado, Isabel Marina y colaboradores <<La autoestima en los adolescentes que cursan el bachillerato. Realidad y expectativas>>. REFCALE: Revista Electrónica Formación y Calidad Educativa. ISSN 1390-9010. Vol. 10 Nº 3 (2022), 104-118. Ver **Anexo 4**.

Los siguientes son trabajos publicados y filiados con la UNLP que se generaron durante el curso del doctorado:

- <<Accesibilidad web: Retos de las Universidades Ecuatorianas>> Presentado como ponencia en el V Congreso Científico Internacional Investigación para la Innovación en las Ciencias. Guayaquil, Ecuador, 2018. Ver **Anexo 6**.
- <<Clasificación de pacientes según su posibilidad de adquirir diabetes mellitus empleando algoritmos de machine learning>> Presentado como ponencia en el IV Congreso Internacional: Tecnologías de la Información y Computación. Calceta, Ecuador, 2020. https://tinyurl.com/33hmenmt. Ver **Anexo 7**.
- <<Legibilidad y Accesibilidad en los Sitios Web de las Universidades de la Provincia de Manabí-Ecuador>> Publicado en la Revista Electrónica Formación y Calidad Educativa (REFCALE). Manta, Ecuador, 2020. https://tinyurl.com/5n97ujwc. Ver **Anexo 8**.
- <<Usabilidad en sitios web oficiales de las universidades del Ecuador>> Publicado en la Revista Ibérica de Sistemas e Tecnologias de Informação (RISTI). Portugal, 2020. https://tinyurl.com/2p97fu8x. Ver **Anexo 9**.
- <<Algunas experiencias de investigación basada en ciencia ciudadana para el beneficio de África>> Publicado en la Revista Electrónica Formación y Calidad Educativa (REFCALE). Manta, Ecuador, 2020. https://tinyurl.com/bddrcsr4. Ver **Anexo 10**.

- <<La usabilidad y la escala diferencial de emociones en aplicaciones para Android. Un estudio de caso>> Publicado en la Revista Científica Multidisciplinaria Mikarimin. Santo Domingo, Ecuador, 2021. https://tinyurl.com/yn4nvk2x. Ver **Anexo 11**.
- <<La usabilidad de los sitios web oficiales de destinos turísticos de países miembros de la OMT>> Publicado en la Revista Electrónica Formación y Calidad Educativa (REFCALE). Manta, Ecuador, 2022. https://tinyurl.com/3kpdava3. Ver **Anexo 12**.
- Pincay Ponce, Jorge Iván y colaboradores <<Modelo Matemático de predicción de Graduados>> Aceptado para publicarse en la Revista Ibérica de Sistemas e Tecnologias de Informação. Ver **Anexo 13**.
- Macías Espinales, Adriana Virginia y colaboradores <<Innovación en la enseñanza - aprendizaje en universidades sudamericanas mediante gestión del conocimiento>> Aceptado para publicarse en la Revista Ibérica de Sistemas e Tecnologias de Informação. Ver **Anexo 14**.
- Pincay Ponce, Jorge Iván y colaboradores. <<Evaluate software designs without jumping down the cool path of creative thinking. Evaluation of designs with User tasks.>> Capítulo de libro enviado para publicarse en la Editorial Universitaria Uleam.

1.7. Organización de la tesis

El Capítulo 1, que es la presente sección introductoria, muestra aspectos relacionados con la motivación de esta tesis, sus objetivos, la metodología de investigación seguida, así como una mención de las publicaciones relacionadas tanto antes como durante el desarrollo de la investigación.

El Capítulo 2 presenta el marco teórico de la tesis, en el cual se describe en más detalle el funcionamiento de los algoritmos de aprendizaje automático, en especial, los de minería de datos. Además, se analiza la bibliografía existente en la actualidad acerca de la minería de datos, tanto a nivel de modelos descriptivos como predictivos, en particular aplicados al análisis de datos socioeconómicos y su incidencia en el rendimiento académico escolar.

El Capítulo 3, contempla el desarrollo de la propuesta de la tesis y en consecuencia las principales contribuciones de esta. El capítulo se estructura con base en las fases del Proceso Estándar de la Industria para Minería de Datos CRISP-DM.

El Capítulo 4, se construye sobre la base de los resultados de los modelos predictivos

y descriptivos construidos metodológicamente en el capítulo 3, contempla la interpretación de los resultados y su transformación en conocimiento útil para las escuelas y actores interesados. La interpretabilidad es el grado en que un humano puede "predecir" consistentemente el resultado del modelo, es decir, es el grado en que un humano puede entender la causa de una decisión (Miller, 2018).

En el Capítulo 5, se presenta las conclusiones relacionadas con los objetivos de esta tesis y se plantean posibles trabajos futuros.

Capítulo 2:

2. Marco teórico

En esta sección se proporcionan detalles de la revisión sistemática de la literatura que complementa esta tesis y de la que el tesista es el autor principal: Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar. Revisión sistemática (Pincay-Ponce et al., 2023). La revisión incluyó 19 estudios primarios publicados entre 2012 y 2022 y se rigió por las siguientes preguntas:

(1) ¿Cuáles son las técnicas de análisis de datos que se han utilizado en este contexto?,
(2) ¿Cuáles son los factores socioeconómicos de mayor incidencia sobre el rendimiento escolar? y
(3) ¿Cuál es el tamaño del conjunto de datos analizados?

La metodología seguida en la revisión fue la propuesta por Kitchenham y Charters (2007), cuyos cuatro pasos se muestran en la **Figura 6**, junto con los números asociados a cada ítem, por ejemplo, para el paso 1 se emplearon 3 preguntas de investigación, en el paso 3 se obtuvieron 19 documentos entre los distribuidos en las bases de datos IEEE Explore y Scopus.

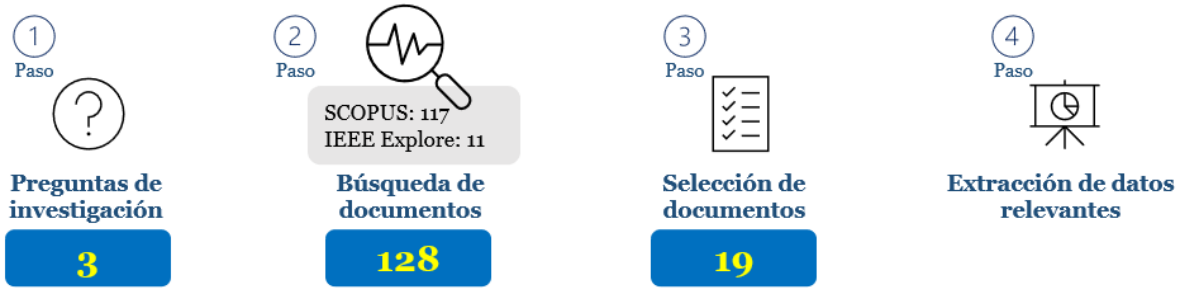


Figura 6: Protocolo de la revisión sistemática de la literatura

En la **Tabla 4** se muestran los 19 artículos seleccionados durante el paso 3, además del tipo de estudio transversal [T] o longitudinal [L]. Los 2 artículos codificados como

A4 y A5 se encontraron en la base de datos IEEE Explore y los 17 restantes en Scopus.

Tabla 4: Documentos seleccionados para la revisión sistemática

Cód.	Tipo de estudio y título	Año	País
A1	[T] Academic performance of Peruvian elementary school children: The case of schools in Lima at the 6th grade	2012	Perú
A2	[T] Parental involvement in homework: Relations with parent and student achievement-related motivational beliefs and achievement	2014	Grecia
A3	[T] Effortful control and early academic achievement of Chinese American children in immigrant families	2015	EEUU
A4	[T] A hybrid method based on MLFS approach to analyze students' academic achievement	2016	Taiwan
A5	[L] Prediction models of learning strategies and learning achievement for lifelong learning	2017	Tainlandia
A6	[L] Home learning environment and development of child competencies from kindergarten until the end of elementary school		Alemania
A7	[L] Childhood Social Skills as Predictors of Middle School Academic Adjustment		EEUU
A8	[T] An Appraisal Model Based on a Synthetic Feature Selection Approach for Students' Academic Achievement		Taiwan
A9	[L] Effortful control is associated with children's school functioning via learning-related behaviors	2018	España
A10	[L] Effects of early childhood education attendance on achievement, social skills, behaviour, and stress		Brasil
A11	[L] Home Visiting Among Inner-City Families: Links to Early Academic Achievement		EEUU
A12	[L] Skin color and academic achievement in young, Latino children: Impacts across gender and ethnic group.	2019	EEUU
A13	[L] The Many (Subtle) Ways Parents Game the System: Mixed-method Evidence on the Transition into Secondary-school Tracks in Germany		Alemania
A14	[L] Longitudinal Associations Linking Elementary and Middle School Contexts with Student Aggression in Early Adolescence	2020	EEUU
A15	[L] Impacts of School Racial Composition on the Mathematics and Reading Achievement Gap in Post Unitary Charlotte-Mecklenburg Schools		EEUU
A16	[T] Parental Self-Efficacy in Helping Children Succeed in School Favors Math Achievement	2021	Canadá
A17	[L] Predicting Students' Mathematics Achievement Through Elementary and Middle School.		EEUU
A18	[T] The Pathway to Enrolling in a High-Performance High School: Understanding Barriers to Access	2022	EEUU
A19	[T] Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar		Ecuador

Fuente: Elaboración a partir de Pincay-Ponce y colaboradores (2023)

2.1. Minería de datos educativos

En 2010, surgió una de las primeras definiciones de minería de datos educativos, Baker (2010) la describió como una disciplina emergente, preocupada por el desarrollo de métodos para explorar tipos de datos que provienen de entornos educativos y utilizar esos métodos para comprender mejor a los estudiantes y a los entornos en los que aprenden.

Romero y Ventura (2013) vincularon la definición con la de Minería de Datos al describirla como la aplicación de técnicas de minería de datos a conjuntos de datos que provienen de entornos educativos para abordar sus cuestiones inherentes, lo que según Peña-Ayala (2014), no se debe confundir con Analítica del Aprendizaje, descrita por Chatti y colaboradores (2013), como la medición, recopilación, análisis e informe de datos sobre el aprendizaje de alumnos y el contexto en que ocurre.

Peña-Ayala (2014) considera a la minería de datos educativos como la combinación de tres áreas principales, tal cual se muestra en la **Figura 7**: Ciencias de la computación, Educación y Estadística, que a su vez forman otras tres subáreas: Analítica de aprendizaje, e Learning y Minería de datos. En consecuencia, la minería de datos utiliza métodos y técnicas de estadística, aprendizaje automático, minería de datos, con: recuperación de información, sistemas de recomendación, psicopedagogía, psicología cognitiva y psicometría.

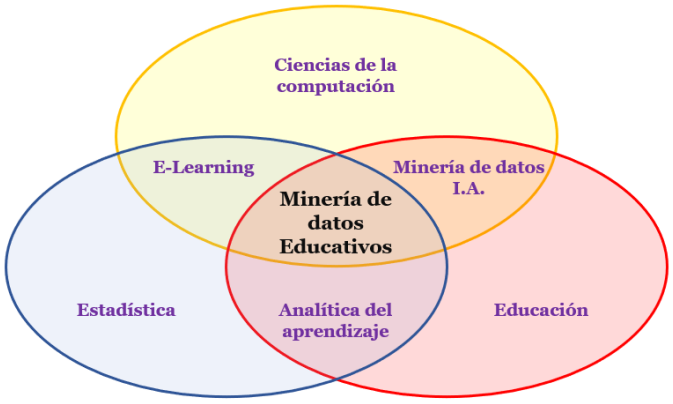


Figura 7: Áreas y subáreas relacionadas con la Minería de Datos Educativos.
Fuente: Peña-Ayala (2014)

En años recientes, Soncin y Cannistrà (2022), valoran este tipo de exploración como clave en el dominio de los datos utilizados, su transformación y apoyo a la toma de decisiones en la práctica educativa y en la investigación.

2.1.1. Campos de aplicación

Es evidente la importancia que actualmente tiene la minería de datos educativos, sin embargo, muchos de los problemas que enfrentan las escuelas no tienen soluciones obvias o por lo menos no son situaciones donde resulte efectivo actuar sin datos. Schildkamp (2019) ejemplifica con una escuela que invierte en materiales curriculares para tratar de mejorar el rendimiento de sus estudiantes en un área temática determinada, no obstante, si la causa o las causas del bajo rendimiento de los estudiantes se encuentran en una falta de apoyo específico, el problema seguirá sin resolverse o incluso puede exacerbarse, pese a la inversión ejemplificada.

Para esclarecer tal situación, en la **Tabla 5** se muestra una clasificación de los campos de aplicación de la Minería de Datos Educativos propuesta por Peña-Ayala (2014), en función de los beneficiarios.

Tabla 5: Campos de aplicación de la minería de datos educativos

Beneficiarios	Ejemplo de aplicación
Estudiantes	• Reflexionar sobre la situación de los alumnos • Proporcionar retroalimentación o recomendaciones adaptativas • Mejorar el rendimiento
Docentes	• Comprender los procesos de aprendizaje de los alumnos • Reflexionar sobre sus propios métodos de enseñanza • Comprender aspectos sociales, cognitivos y conductuales de los alumnos • Mejorar el desempeño docente
Investigadores	• Desarrollar y comparar técnicas de minería de datos para poder recomendar la más útil para cada tarea • Evaluar
Directivos	• Evaluar la mejor manera de organizar los recursos institucionales, sean humanos o materiales y su oferta educativa

Fuente: Peña-Ayala (2014)

2.1.2. Objetivos generales de la minería de datos educativos

Cuando una aplicación de minería de datos educativos está relacionada con más de un beneficiario de los mostrados en la **Tabla 5**, puede resultar conveniente clasificar su alcance en función de la misma aplicabilidad, tal cual se muestra en la **Tabla 6.** A final de cuentas, según Peña-Ayala, se persigue avanzar en el conocimiento científico sobre el aprendizaje y los alumnos, mediante la construcción, el descubrimiento o la mejora de modelos del alumno, del dominio y del apoyo pedagógico

Tabla 6: Aplicaciones de la minería de datos educativos

Objetivos	Ejemplo de aplicación
Modelado de	• Caracterizar estados de los estudiantes, tales como conocimientos,

estudiantes	habilidades, motivación, satisfacción, metacognición, actitudes, experiencias y progreso en su aprendizaje • Caracterizar problemas que impactan negativamente en sus resultados de aprendizaje • Mejorar un modelo de estudiante a partir de la información
Modelado del comportamiento del estudiante	• Detectar comportamientos tales como: dormirse, desmotivarse, adivinar en las evaluaciones, preguntar, solicitar ayuda, predisposición, entre otros. • Adaptar sistemas o aplicaciones de acuerdo con el comportamiento de los alumnos.
Predicción del desempeño y resultados del aprendizaje	• Predecir las calificaciones finales de un estudiante • Predecir la retención en un curso o programa de grado • Predecir la capacidad futura de aprender según los datos de las actividades del curso
Recomendaciones	• Recomendar a los alumnos qué contenidos son los más apropiados para ellos en determinado momento • Agrupar a los alumnos según su perfil y con fines de adaptación y personalización
Directivos	• Analizar las actividades de los estudiantes y el uso de la información en los cursos
Mantenimiento y mejora de los cursos	• Mejorar los cursos a nivel de contenidos, actividades, enlaces, etc. • Relacionar información, en particular sobre el uso y aprendizaje de los estudiantes • Estudiar los efectos de los diferentes tipos de apoyo pedagógico que puede proporcionar un software de aprendizaje
Análisis de estructura de dominio	• Mejorar los modelos de dominio que caracterizan el contenido a aprender y a las secuencias de instrucción, utilizando la capacidad de predecir el desempeño del estudiante como una medida de calidad de un modelo de estructura de dominio

Fuente: Elaboración a partir de Peña-Ayala (2014) y Umer (2020)

2.1.3. Tipos de datos usados con frecuencia en el contexto escolar

De acuerdo con la revisión sistemática de la literatura que complementa esta tesis, que incluyó 19 estudios primarios publicados entre 2012 y 2022 y en aras de la simplicidad, los datos de los estudiantes de escuela a los que se recurren en dichos estudios se agrupan según se muestra en la **Tabla 7**. Los siguientes cinco estudios utilizan además las calificaciones de los alumnos en diversas materias:

- Academic performance of Peruvian Elementary school children: The case of schools in Lima at the 6th grade (Manrique Millones et al., 2011).
- Prediction models of learning strategies and learning achievement for lifelong learning (Bussaman et al., 2017).
- The Many (Subtle) Ways Parents Game the System: Mixed-method Evidence on the Transition into Secondary-school Tracks in Germany (Dumont et al., 2019).
- Longitudinal Associations Linking Elementary and Middle School Contexts with

Student Aggression in Early Adolescence (Sanders et al., 2020).
- The Pathway to Enrolling in a High-Performance High School: Understanding Barriers to Access (Sartain & Barrow, 2022).

Tabla 7: Tipos de datos más recurridos para analizar el rendimiento escolar

Tipo de datos	Estudios primarios
Datos preescolares (jardín)	• (Niklas & Schneider, 2017), • (Nievar et al., 2018), • (Sartain & Barrow, 2022), • (Han & Neuharth-Pritchett, 2021), • (Correia-Zanini et al., 2018)
Participación de los padres en las actividades académicas de los niños	• (Gonida & Cortina, 2014), • (Niklas & Schneider, 2017), • (Y. Liu & Leighton, 2021)
Autocontrol y habilidades sociales de los niños	• (Chen et al., 2015), • (Sánchez-Pérez et al., 2018)
Visitas domiciliarias a niños por parte de personal de la escuela	• (Nievar et al., 2018)
Factores socioeconómicos	Etnia: • (Kim & Calzada, 2019), • (Wang et al., 2020)
	Escolaridad de padres, ingreso familiar, tipología de familias, disponibilidad de servicios básicos en el hogar: • (W.-X. Liu & Cheng, 2016), • (Cheng & Liu, 2017), • (Pincay-Ponce et al., 2022).

Fuente: Elaboración a partir de Pincay-Ponce y colaboradores (2023)

Además de la agrupación precedente, en diversos estudios primarios se argumentó que la diferencia basada en el género, niño o niña influye en el desempeño de los estudiantes (Correia-Zanini et al., 2018; Manrique Millones et al., 2011; Niklas & Schneider, 2017; Wang et al., 2020).

2.2. Delimitación del término rendimiento académico

El término rendimiento surgió en la sociedad industrial, luego se derivó a otros ámbitos de la ciencia y de la técnica, hasta llegar al ámbito escolar y ser con frecuencia identificado con aprendizaje, que, al mismo tiempo, se lo entiende como un indicador para medir la productividad de un sistema, que involucra a su vez alumnos, profesores y procesos de evaluación que buscan una educación de calidad (Grasso, 2020; Solano Luengo, 2015).

La definición más común de rendimiento académico y que se sigue en esta tesis, se

asocia a las calificaciones cuantitativas obtenidas en el ámbito académico, como el indicador operativo más frecuente del nivel promedio de educación adquirido en determinada área (Grasso, 2020; Hernández, 1994; Tejedor, 1998). Además, según Tejedor (2003), en este contexto, el éxito implica terminar la carrera en los tiempos estipulados y por el tanto abarcar un lapso mayor al previsto representa una forma de fracaso.

Ahora bien, el bajo rendimiento, de acuerdo con De la A Muñoz (2018), es el resultado de varios factores de riesgo, surge de la combinación y acumulación de muchas barreras y desventajas que afectan a los estudiantes, en ocasiones a través de su entorno escolar y familiar, en el transcurso de sus vidas.

2.3. Factores de riesgo del rendimiento

El rendimiento escolar es un problema que preocupa a nivel mundial, lo que es demostrado por múltiples estudios efectuados durante años en la búsqueda por entender por qué se quedan atrás ciertos estudiantes y cómo se les puede ayudar.

Uno de los informes más referidos en Iberoamérica, es el producido en el marco de la Investigación Iberoamericana sobre Eficacia Escolar (2007), que señala a modo global que el país de residencia del alumno marca diferencias en el logro cognitivo, al menos en el área de matemáticas, dado que el 15% de la varianza en el logro en Iberoamérica se explica por el país. Luego, entre un 14% y 18% es la escuela la que marca influencia especialmente en áreas de matemática y de lengua y para finalizar, la influencia del aula de clases sobre el rendimiento académico global bordea el 10%. En 2016, la OCDE (2016), determinó como tres factores de riesgo del rendimiento académico al estudiante, la escuela y el sistema educativo (ver **Tabla 8**).

Tabla 8: Factores de riesgo en el rendimiento académico según la OCDE (2016)

Grupo	Factor
Sistema educativo	• Políticas de estratificación
	• Asignación de recursos
Escuela	• Organización
	• Ambiente de estudio
Estudiante	• Contexto familiar y nivel socioeconómico y demográfico
	• Actitudes y comportamiento para el progreso educativo

Fuente: OCDE (2016)

Como complemento a lo señalado, el informe del Programa para la Evaluación Internacional de Alumnos (PISA), del que Ecuador forma parte desde el 2014, señala

que la probabilidad de un bajo rendimiento es resultado de una combinación y acumulación de varios factores como una situación socioeconómica desfavorecida, el origen inmigrante, o hablar una lengua distinta en casa respecto de la empleada en el centro escolar, vivir en una familia monoparental, o asistir a escuelas en el medio rural (Programme for International Student Assessment, PISA 2015, 2018).

Según un informe del Banco Mundial (2022), en Ecuador, desde dónde surge este estudio, en 2020 el estado presupuestó para educación el 4.1% de su PIB, ocupando el puesto 9 entre los 13 países sudamericanos. Aunque el año de reporte del valor difiere entre los países tal cual se muestra en el **Gráfico 1**.

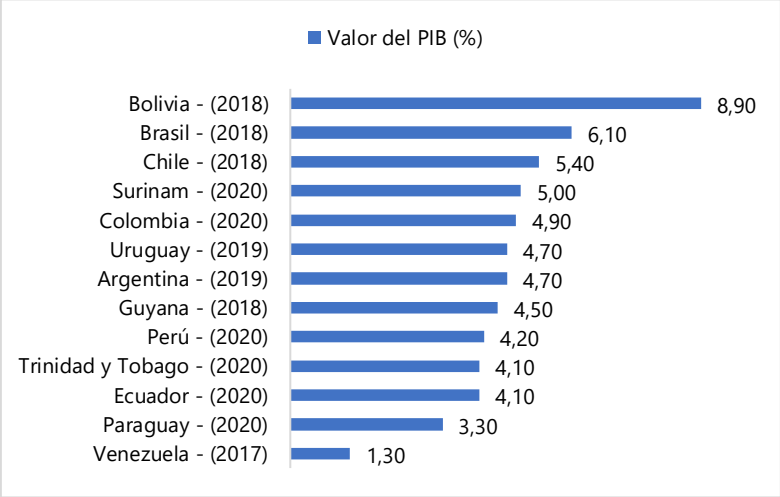


Gráfico 1: Porcentaje del PIB destinado a Educación en países de América del Sur
Fuente: Banco Mundial (2022)

Si bien, en el informe del Banco Mundial, países del denominado primer mundo como Estados Unidos, Australia, Nueva Zelanda y Japón, reportaron en 2018 4.9, 5.1, 6 y 3.1 como presupuesto porcentual de su PIB para la educación, es de considerar que sus presupuestos generales son mayores. Para complementar la lista de países de primer mundo, el valor facilitado por Canadá al banco mundial es de 3.1% de su PIB correspondiente a 2011. Finalmente, el porcentaje de PIB promedio dedicado a Educación en América Latina y el Caribe, excluido países de altos ingresos, es del 4.6% según el Banco Mundial (2022).

En el caso de Ecuador, el presupuesto se destina entre otras cosas a sueldos de docentes, gastos de infraestructura, servicios básicos, servicios complementarios de colación escolar, uniformes y textos, con el afán de reducir la brecha de ausentismo o deserción escolar y lograr fomentar la calidad académica (De La A Muñoz, 2018).

Esto se puede asociar con que el financiamiento del estado tiene como meta la finalización de los estudios de los educandos, incluyendo niños de escuelas, por lo tanto, los estudios incompletos o prolongados más de lo previsto son preocupaciones importantes que pueden llegar a afectar incluso a la economía en

un largo plazo. Por lo tanto, las instituciones educativas necesitan desarrollar estrategias para mejorar la calidad de su educación y asegurar mayores tasas de matriculación y retención.

Tal como se estableció en los objetivos de la tesis, se buscó contribuir al entendimiento y mejora de la progresión del estudiante de escuela en el transcurso del estudio de las materias correspondientes al año de educación básica: 2do, 3ero, 4to, 5to, 6to o 7mo. Se pone especial atención a la incidencia de los factores socioeconómicos en el posible bajo rendimiento escolar y la consecuente reprobación del año básico cursado o peor aún, el abandono temporal o incluso la deserción del sistema educativo.

Es necesario acotar que, a nivel de Latinoamérica y El Caribe, el informe de los resultados de logros de aprendizaje y factores asociados del Estudio Regional Comparativo y Explicativo (ERCE), que abordó el periodo 2013 – 2019 y que contó con el respaldo de la UNESCO, reportó que, en promedio en los 16 países de la región, el 40% de los estudiantes de 3º grado y el 60% de 6º grado de primaria, no alcanzan el nivel mínimo de competencias fundamentales en Lectura y Matemática. Por tanto, la región enfrenta a una crisis educativa en el progreso en el logro de contenidos básicos de aprendizaje, lo que supone un reto para la concreción del derecho a una educación de calidad (UNESCO, 2021).

De acuerdo con la revisión sistemática de la literatura que complementa esta tesis, en la **Tabla 9**, se muestran los factores socioeconómicos que cada estudio primario reportó como más influyente sobre el aprovechamiento escolar, acorde con el contexto y alcance de cada estudio.

Tabla 9: Factores socioeconómicos de más afectación al rendimiento escolar

Estudio	Principal factor socioeconómico
1 - 2012 Academic performance of Peruvian elementary school children: The case of schools in Lima at the 6th grade	Género Edad de los alumnos
2 - 2014 Parental involvement in homework: Relations with parent and student achievement-related motivational beliefs and achievement	Participación de los padres
3 - 2015 Effortful control and early academic achievement of Chinese American children in immigrant families	Cultura de los padres Padres autoritarios

Estudio	Principal factor socioeconómico
4 - 2016 A hybrid method based on MLFS approach to analyze students' academic achievement	Edad de los padres Economía
6 - 2017 Home learning environment and development of child competencies from kindergarten until the end of elementary school	Género Entorno de aprendizaje hogareño
7 - 2017 Childhood Social Skills as Predictors of Middle School Academic Adjustment	Entorno de aprendizaje hogareño Habilidades sociales Economía
8 - 2017 An Appraisal Model Based on a Synthetic Feature Selection Approach for Students' Academic Achievement	Escolaridad de los padres
9 - 2018 Effortful control is associated with children's school functioning via learning-related behaviors	Habilidades sociales Economía
10 - 2018 Effects of early childhood education attendance on achievement, social skills, behavior, and stress	Género Economía
11 -2018 Home Visiting Among Inner-City Families: Links to Early Academic Achievement	Participación de los padres
12 - 2019 Skin color and academic achievement in young, Latino children: Impacts across gender and ethnic group.	Raza
13 - 2019 The Many (Subtle) Ways Parents Game the System: Mixed-method Evidence on the Transition into Secondary-school Tracks in Germany	Economía
14 - 2020 Longitudinal Associations Linking Elementary and Middle School Contexts with Student Aggression in Early Adolescence	Escolaridad de los padres
15 - 2020 Impacts of School Racial Composition on the Mathematics and Reading Achievement Gap in Post Unitary Charlotte-Mecklenburg Schools	Género Raza
16 - 2021 Parental Self-Efficacy in Helping Children Succeed in School Favors Math Achievement	Participación de los padres
17 - 2021 Predicting Students' Mathematics Achievement Through	Economía

Estudio	Principal factor socioeconómico
Elementary and Middle School: The Contribution of State-Funded Prekindergarten Program Participation	
18 - 2022 The Pathway to Enrolling in a High-Performance High School: Understanding Barriers to Access	Economía Distancia casa - escuela
19 - 2022 Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar	Economía Escolaridad de los padres

Fuente: Elaboración a partir de Pincay-Ponce y colaboradores (2023)

Nótese que en 9 de los 19 trabajos se menciona a los padres de los niños de modo directo o indirecto, por lo que cabe resaltar lo descrito por el Ministerio de Educación de Ecuador (2016), que conmina a los docentes escolares a resaltar a los padres la importancia de una buena relación con sus hijos, traducida en conversaciones, diálogos y orientaciones que:

- Inculquen en sus hijos una actitud positiva hacia la realización de las tareas escolares.
- Escuchen siempre a los hijos para conocer los problemas o éxitos que les quieran compartir.
- Dediquen tiempo cada día para el trabajo escolar en casa.
- Expresen a sus hijos cariño, afecto tanto verbal como físico.
- Valoren el esfuerzo y la superación de dificultades y limitaciones en su trabajo.
- Asistan a los llamados de los docentes a la institución educativa.

2.4. Abandono y deserción escolar

En las definiciones precedentes se puso especial atención a la incidencia de los factores socioeconómicos en el posible bajo rendimiento escolar y la posible reprobación del año básico cursado o peor aún, el abandono temporal o incluso la deserción del sistema educativo.

A efectos conceptuales, se considera que el abandono escolar está dado por la diferencia de la cantidad de estudiantes entre uno y otro período académico, se calcula como la resta entre la matrícula total de un año básico, menos los promovidos al siguiente año básico. Sin embargo, existen situaciones no numéricas que en el transcurso de los estudios conducen al abandono temporal y el consecuente retraso académico, incluso tal escenario es precursor de la deserción

escolar (Russo, 2019).

La referida deserción escolar, se define como abandonar el sistema educativo formal antes de obtener un título apropiado en algún nivel educativo. Debido a su importante impacto en el desarrollo cultural y profesional de la sociedad, existen multiples iniciativas para combatirlo, pero se lo reconoce como problema común tanto en el mundo industrializado como en los llamados países del Tercer Mundo. La diferencia es que en el primer caso sucede con frecuencia en la educación universitaria, mientras que en el segundo caso se da en escuelas, colegios y universidades (Editorial Etecé, 2021).

A nivel mundial, la UNESCO, en su informe de deserción escolar expresa que hay niños que cada vez están más alejados de la educación, corren el riesgo de reprobar años, no terminar sus estudios y no aprender en la escuela. Aunque la educación es un derecho humano de impacto positivo en el crecimiento económico, para los niños que no completan su educación primaria, el precio puede ser muy alto, afectando sus perspectivas y satisfacción en lo laboral, así como en sus comportamientos y salud infantil, que a su vez impacta su posterior rol en la sociedad y la formación de sus familias (UNESCO, 2022).

2.5. Aprendizaje automático

El aprendizaje automático es un tema central en la Inteligencia Artificial, es de naturaleza inductiva y comprende técnicas y métodos para realizar clasificación, optimización y predicción. Se caracteriza por ser uno o más programas no codificados explícitamente para resolver un problema, sino por basarse en algoritmos de aprendizaje sobre los registros de bases de datos, a los que también se les suele denominar ejemplos o instancias y así, generar un modelo de un problema (F. M. Quiroga, 2020; El-Naqa & Murphy, 2015).

Con las técnicas de aprendizaje automático en la minería de datos educativos, se desarrollan modelos para descubrir patrones ocultos significativos y explorar información útil de entornos educativos. Algunos datos recurridos en este tipo de análisis suelen ser los demográficos, antecedentes académicos y características de comportamiento.

Se clasifican con base en las salidas que generan, en algoritmos de aprendizaje supervisado o predictivos y algoritmos de aprendizaje no supervisado o descriptivos, tal cual se detalla en la **Tabla 10**. Muchas veces también se amplía esta clasificación

a modelos semi supervisados y de refuerzo, aunque este documento se centra en los dos primeros.

Tabla 10: Algoritmos de aprendizaje automático categorizados según las tareas que realizan

Aprendizaje	Tareas	Tipo de algoritmos
Supervisado	Clasificación	Máquinas de soporte vectorial
		Análisis discriminante
		Método de Bayes
		Vecino más cercano
		Árboles de decisión
	Regresión	Regresión lineal (Generalized Linear Model, GLM)
		Regresión vectorial de soporte (Support Vector Regression, SVR)
		Regresión de procesos gaussianos (Gaussian process regression, GPR)
		Aprendizaje en conjunto (Ensemble methods)
		Árboles de decisión
		Redes neuronales
No supervisado	Agrupamiento	K-Medias
		K-Medoids
		Fuzzy C-Means (algoritmo difuso)
		Jerárquico
		Gaussian Mixture Models
		Redes neuronales
		Modelos ocultos de Markov
	Asociación	A priori
		Árbol de patrones frecuentes, FP Growth

Fuente: Elaboración propia a partir de Duc y colaboradores (2020)

Cada uno de los algoritmos mencionados se desagrega en otros más, por ejemplo, los árboles de decisión encierran varios algoritmos como C4.5, C5, ID3, Chi-square automatic interaction detection (Chaid), Classification And Regression Tree (CART), Multivariate adaptive regression splines (MARS), entre otros. En el desarrollo de esta investigación se empleó los árboles de decisión implementados en el software Orange que son C 4.5 y CART que se implementa en los modelos de aprendizaje ensamblados o de en conjuntos.

Cada algoritmo, dentro de la construcción de los modelos, tienen de posible la configuración de varios hiperparámetros y parámetros. A continuación, se ofrece una definición general de estos términos, pero se las concreta en la documentación de los modelos del Capítulo 3, sección **3.4. Fase de Modelado**.

En las secciones siguientes de este capítulo se documentan los siguientes aspectos de cada modelo empleado:

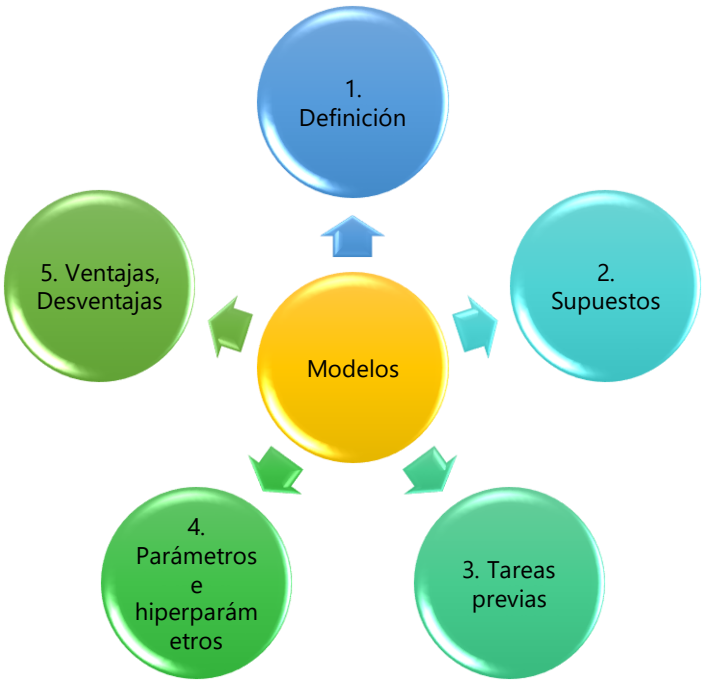


Figura 8: Generalidades documentadas en la investigación respecto de cada modelo

2.5.1. Parámetros e hiperparámetros generales

Los hiperparámetros son ajustes, el prefijo hiper sugiere que son ajustes de 'nivel superior' que controlan el proceso de aprendizaje y determinan los parámetros del modelo que resultan de él. Estos ajustes que utilizará el algoritmo de aprendizaje se definen antes de que comience el entrenamiento del modelo y no pueden ser cambiados durante el mismo, esto significa, que no es posible conocer qué valores de los hiperparámetros que se utilizaron para entrenar un modelo a partir del modelo en sí, solo conocemos los parámetros del modelo que se aprendieron (Nyuytiymbiy, 2022).

De acuerdo con Nyuytiymbiy (2022), algunos ejemplos de hiperparámetros son: (1) Relación del tamaño del conjunto de datos para entrenamiento y prueba, (2) Tasa de aprendizaje en algoritmos de optimización, (3) Elección del algoritmo de optimización, (4) Elección de la función de activación en una capa de red neuronal, (5) Número de capas ocultas de una red neuronal, (6) Número de clústeres en una tarea de agrupación, entre otros. Los hiperparámetros son muy críticos en la construcción de modelos robustos y precisos, porque posibilitan el equilibrio entre

el sesgo y la varianza y por lo tanto, evitan que el modelo se sobreajuste o se ajuste menos.

Los parámetros, por otro lado, son internos al modelo y se aprenden a partir de los datos durante el entrenamiento al tiempo que el algoritmo utilizado intenta aprender el mapeo entre las características de entrada y las etiquetas objetivos. El entrenamiento de modelos normalmente comienza con hiperparámetros que se inicializan en algunos valores aleatorios o establecidos en ceros. A medida que avanza el entrenamiento - aprendizaje, los valores iniciales se actualizan utilizando un algoritmo de optimización. El algoritmo de aprendizaje actualiza continuamente los valores de los parámetros a medida que avanza el aprendizaje, pero los hiperparámetros permanecerán sin cambios.

Kuhn y Johnson (2013) señalan como ejemplos de parámetros a: (1) Los coeficientes o pesos de los modelos de regresión lineal y logística, (2) Pesos y sesgos de una red neuronal y (3) Los centroides del clúster en la agrupación en clústeres.

En el desarrollo de esta investigación, en el Capítulo 3, se documentan los parámetros e hiperparámetros de acuerdo con cada modelo implementado en el software Orange y en Script de Python implementados como parte del modelo en Orange.

2.5.2. Modelos supervisados

En el aprendizaje automático y en la Inteligencia Artificial en general, los modelos predictivos se corresponden con el aprendizaje supervisado. Estos modelos utilizan un subconjunto de datos de entrenamiento que incluye entradas y salidas correctas para aprender y luego producir el resultado deseado con los consecuentes patrones e hipótesis generales. El restante subconjunto de datos, usualmente más pequeño, se usa de prueba y mide la precisión del modelo a través de una función de pérdida, ajustándose hasta que el error se ha minimizado lo suficiente (Duc et al., 2020; Nelli, 2018).

Entre las tareas predictivas se encuentran la clasificación y la regresión, además, también se pueden efectuar tareas de inferencia para comprender cómo se ve afectada una variable de respuesta cuando cambian las predictoras (Duc et al., 2020):

- En las tareas de clasificación se utilizan algoritmos para asignar con precisión los datos de prueba en categorías específicas. Se reconoce entidades específicas dentro del conjunto de datos y se intenta sacar algunas

conclusiones sobre cómo deben etiquetarse o definirse esas entidades. Muchas veces, una instancia puede pertenecer a varias categorías, en tal caso se necesitará determinar cuáles son esas categorías y cuánta confianza tiene el algoritmo en sus predicciones (A. Singh et al., 2016; Duc et al., 2020).

- La regresión consiste en aprender una función real que asigna a cada instancia un valor continuo o real, a diferencia de la clasificación que emplea valores categóricos. El formato de las regresiones, como problema, a menudo sigue un formato lineal (Russo, 2019).

Tal cual se mostró en la **Tabla 10**, para las tareas de clasificación los tipos de algoritmos más referidos en la literatura son: Máquinas de soporte vectorial, Análisis discriminante, Método de Bayes, Vecino más cercano y Árboles de decisión.

2.5.2.1. Máquinas de soporte vectorial (SVM)

Las SVM se pueden utilizar para problemas de clasificación como clasificación de vectores de soporte (SVC) y regresión de vectores de soporte (SVR). Se utilizan para conjuntos de datos pequeños y etiquetados para encontrar un hiperplano óptimo que separe mejor a las características en diferentes clases o valores de clase (Nelli, 2018; Sarkar et al., 2018).

No hay supuestos de tipo estadístico que validar para SVM, sin embargo, Yerutu (2020) menciona algunos aspectos que SVM asume: (1) El margen debe ser lo más grande posible, ver **Figura 9**, (2) Los vectores de soporte son los puntos de datos más útiles porque son los que tienen más probabilidades de clasificarse incorrectamente, ver **Figura 9**. (3) SVM asume que los datos son independientes y se distribuyen de manera idéntica.

En cuanto a las tareas comunes de preparación de los datos para este modelo, en Orange (2015ac): (1) Se elimina instancias con etiquetas de clase desconocidas. (2) Se convierte el conjunto de datos multidimensional en forma binaria utilizando el método de conversión uno contra el resto o uno contra uno. (3) Se recodifica las características categóricas como numéricas, usualmente con una codificación en caliente. (4) Se elimina columnas con valores vacíos. (5) Se imputan los valores faltantes con los valores medios.

Referente de los parámetros de SVM, estos pueden tener alguna variación dependiendo del software que se utilice, pero en general existen dos métodos de SVM, SVC y ν-SVC, estos dan puntuaciones por clase para cada instancia o una puntuación única por instancia en el caso de clases binarias. SVM y ν-SVM se basan

en una minimización diferente de la función de error. A continuación, sus parámetros con base en Orange, que a su vez hace uso del paquete multiclase LIBSVM (2022).

- **SVM. –** Implica ajustar:
  (1) **Coste:** Plazo de penalización por pérdida y se aplica a tareas de clasificación y regresión,
  (2) **épsilon ε:** Un parámetro para el modelo épsilon-SVR, se aplica a las tareas de regresión y define la distancia desde los valores verdaderos dentro de los cuales no se asocia ninguna penalización con los valores previstos.
  ε es un parámetro de regularización, igual que Coste, con valores más grandes permite errores residuales más grandes, lo que reduce el número de vectores de soporte y por ende el tiempo de ejecución. Además, el modelo se vuelve más suave o útil, por ejemplo, para manejar datos ruidosos y evitar el sobreajuste (Bartz et al., 2023; Orange, 2015ac).
- **ν-SVM. -** Implica ajustar: (1) **Costo** o plazo de penalización por pérdida y se aplica solo a tareas de regresión, (2) **ν,** es un parámetro del modelo ν-SVR, se aplica a tareas de clasificación y regresión. Es un límite superior en la fracción de errores de entrenamiento y un límite inferior de la fracción de vectores de soporte.
- **Tolerancia numérica:** El valor sugerido en la literatura es entre $10^{-5}\ a\ 10^{-1}$ (van Rijn & Hutter, 2018). En un inicio si $Tolerancia_i = 0$ los puntos pueden considerarse correctamente clasificados y si $Tolerancia_i > 0$ los puntos son clasificados incorrectamente. Entonces la $Tolerancia$ es un término de error asociado con X_i (variable). El error promedio se puede dar como $\frac{1}{n}\sum_{i=1}^{n} Tolerancia_i$
- **Kernel. -** Es una función que transforma el espacio de características en un nuevo espacio de características para ajustarse al hiperplano desde el margen máximo, para obtener el modelo con núcleos lineales, polinomios, sigmoides y Radial Basis Function (RBF, que es el valor predeterminado). Involucra las siguientes constantes: (1) **g** o gamma en la función kernel, cuyo valor recomendado es 1/k, donde k es el número de atributos, (2) **c** para la constante c_0 en la función kernel, cuyo valor predeterminado 0 y (3) **d** que es el grado de la función polinómica del núcleo, pero que no aplica a núcleos lineales y RBF.

C es un parámetro de regularización que controla la complejidad del modelo, que en este escenario no significa la cantidad de coeficientes como lo es en los modelos lineales, o las divisiones como lo es en los árboles de decisión, sino el potencial para generar funciones más activas o resistentes. C influye en el número de vectores de soporte, comprendiendo que muchos vectores de soporte pueden crear funciones

con muchos picos y conducir a un sobreajuste (Bartz et al., 2023).

Sobre el Kernel, a menudo puede entenderse como una medida que describe qué tan similares son dos instancias entre sí, según los valores de sus características (Bartz et al., 2023).

Sobre el valor de Gamma, como parámetro de RBF en el Kernel, la literatura sugiere un gamma: $\gamma \epsilon (0, \propto)$, aunque se reporta como razonable el uso de una escala logarítmica, por ejemplo 2^{-10} ... 2^{10} para cubrir un amplio espectro de valores muy pequeños y grandes (Bartz et al., 2023).

En la **Figura 9** se ilustra como los puntos más cercanos al hiperplano se denominan puntos vectoriales de soporte, las distancias de los vectores desde el hiperplano o líneas punteada central se denominan márgenes. Los puntos rojos y verdes representan la clasificación final (Nelli, 2018).

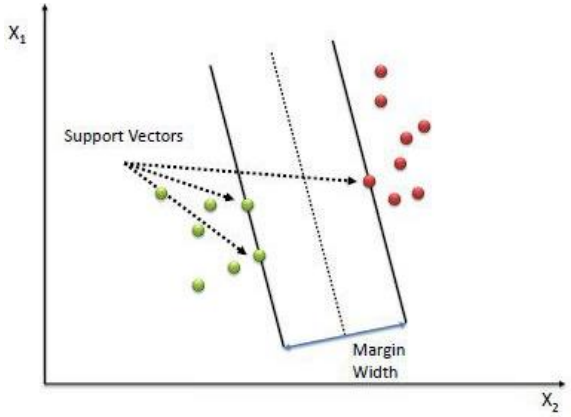


Figura 9: Representación gráfica de SVM

En cuanto a las principales ventajas de SVM se tiene que: (1) Es eficaz cuando las clases son separables, (2) El hiperplano se ve afectado solo por los vectores de soporte, por lo que los valores atípicos tienen menos impacto y (3) Es adecuado para la clasificación binaria (Brownlee, 2016; Yeturu, 2020). Los valores atípicos son aquellos cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos (El-Naqa & Murphy, 2015).

En cuanto a desventajas: (1) Requiere gran tiempo de procesamiento cuando se tienen más de 10 mil instancias, (2) No funciona bien para clases superpuestas, (3) Requiere de cuidado en sus hiperparámetros para lograr un rendimiento de generalización suficiente, (4) Puede ser complicado seleccionar la función del kernel apropiada y (5) Clasifica colocando puntos de datos, por encima y por debajo del hiperplano, pero no aclara las probabilidades de clasificación (Brownlee, 2016; Yeturu, 2020).

2.5.2.2. Análisis discriminante lineal

El análisis discriminante lineal, o LDA por sus siglas en inglés, se utiliza como

herramienta para la clasificación y regresión, la reducción de dimensiones y la visualización de datos. Referente a la clasificación, dos o más grupos son conocidos a priori, por lo que se lo considera un clasificador Ad Hoc y las nuevas observaciones se clasifican en uno de ellos en función de sus características, haciendo uso del teorema de Bayes. La puntuación de características se determina por la Exactitud de clasificación o el MSE de K-vecinos más cercanos en los datos bidimensionales proyectados para el caso de la regresión (Basantia et al., 2019).

Además de que LDA demande de características independientes continuas y una etiqueta de clase como categórica, Boedeker y Kearns (2019) resaltan a los siguientes supuestos o asunciones de LDA:

- **Normalidad multivariante** de las características independientes para cada nivel de la característica de agrupación.
- **Homogeneidad de varianza, covarianza u homocedasticidad**, las varianzas entre las características de grupo son las mismas en todos los niveles de predictores.
- **Multicolinealidad,** el poder predictivo puede disminuir con una mayor correlación entre las características predictoras.
- **Independencia,** se supone que las instancias son muestreadas al azar y se supone que la puntuación de una instancia es independiente de las puntuaciones en esa variable en todas las demás instancias.

Con respecto de los supuestos listados, en Orange, no se ejecuta tareas adicionales de preparación de los datos. Referente de los parámetros, estos pueden tener alguna variación dependiendo del software que se utilice. Para este algoritmo Orange utiliza la librería Scikit-Learn (Orange, 2015m; Pedregosa et al., 2011).

Antes de referir a las desventajas, en la **Figura 10** se ilustra un modelado de pacientes enfermos y no. En (A) se identifica los límites entre las dos clases teniendo en cuenta la distancia entre las medias de las clases, los puntos negros y la varianza dentro de cada clase. En (B) se crea una función discriminante (DF) representada por la línea ahora reducidos a 1 dimensión. En (C) no se haría otra función discriminante porque no separaría bien las diferentes clases entre sí. (Basantia et al., 2019).

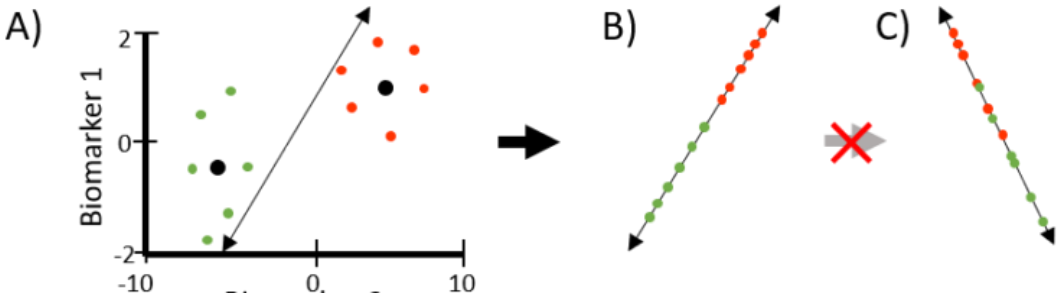


Figura 10: Representación gráfica de LDA

En cuanto a las principales ventajas de LDA se tiene que: (1) Si las clases están bien separadas, los parámetros estimados por LDA son estables y por ejemplo, en la regresión logística no lo serían. (2) Otra vez, respecto de la regresión logística, cuando el número de observaciones es bajo y la distribución de los predictores es aproximadamente normal en cada una de las clases, LDA será muy estable. (3) Produce resultados de clasificación robustos, decentes e interpretables.

En el aprendizaje automático, se considera que los usuarios de modelos de aprendizaje automático confían mejor en las predicciones (o regresiones) si el modelo también puede explicar por qué hizo la predicción (Miller, 2018).

En cuanto a desventajas la literatura sugiere que sus supuestos son difíciles de cumplir (Boedeker & Kearns, 2019).

2.5.2.3. Método de Bayes

Bayes Naïve, Ingenuo de Bayes en español, se utiliza como herramienta de clasificación y se basa en los teoremas de Bayes. La denominación de ingenuo es porque el algoritmo utiliza técnicas bayesianas, pero sin tener en cuenta las dependencias que puedan existir entre las características, asume que son independientes entre sí. Se define por:

$$P(A|B) = \frac{P(R|A)P(A)}{P(R)}$$

Dónde:

- P (A). – Probabilidad de A
- P (R |A). – Probabilidad de que se dé R dado A
- P(R). – Probabilidad de R

- P(A|R). – Probabilidad posterior de que se dé A dado R

Respecto de los supuestos de Bayes Naïve, Yerutu (2020) menciona que: (1) Todas las columnas son independientes entre sí y solo dependen de la etiqueta y (2) Todas las filas son independientes entre sí… Su mayor suposición es la de independencia condicional.

En cuanto a las tareas comunes de preparación de los datos para este modelo: (1) Se elimina columnas con datos vacíos, (2) Se discretiza valores numéricos a 4 contenedores con igual frecuencia. (Orange, 2015).

En Orange 3.34 no se requieren de ajustes para parámetros de Bayes Naïve (Orange, 2015p).

Yeturu (2020) compiló algunas ventajas y desventajas de Bayes Naïve, en cuanto a las principales ventajas de Bayes Naïve se tiene que: (1) El algoritmo es muy rápido para procesar características discretas. (2) Puede manejar un gran número de características sin mermar de modo considerable su rendimiento. (3) Se comporta bien incluso ante la presencia de características irrelevantes.

En cuanto a las principales desventajas de Bayes Naïve se tiene que: (1) Funciona lento para características continuas. (2) Cuando el conjunto de datos de prueba tiene una característica que no ha sido observada en el conjunto de entrenamiento, el modelo le asigna una probabilidad de cero, para evitar aquello, en Orange se implementa como técnica de suavizado a la estimación de Laplace (Kikuchi et al., 2015). (3) Es probable que la presunción de independencia no refleje cómo son los datos en el mundo real.

2.5.2.4. Vecino más cercano, KNN

El algoritmo k-vecinos más cercanos (KNN) es un método de clasificación de datos para estimar la probabilidad de que un punto de datos se convierta en miembro de un grupo u otro en función del grupo al que pertenecen los puntos de datos más cercanos. También se lo emplea en la regresión. En el caso de la clasificación, los pasos globales por seguir son:

Cargar los datos
Elija el valor K
Para cada punto de datos de los datos:
- Buscar la distancia, por ejemplo, euclidiana, a todas las muestras de datos de entrenamiento

- Almacenar las distancias en una lista y ordenarla
- Elija las primeras entradas K de la lista ordenada
- Etiquete el punto de prueba en función de la mayoría de las clases presentes en los puntos seleccionados

Fin

En el caso de la regresión KNN, en lugar de asignar la clase con los votos más altos, el promedio de los valores de los vecinos se calcula y se asigna al punto de datos desconocido.

Si se quiere clasificar un objeto, con k= 5 vecinos, con 3 vecinos de la clase A y 2 de la clase B, por mayoría el objeto se clasifica como clase A. Para evitar empates, se prefiere que el número de vecinos k seleccionado sea un número impar, tal cual se muestra en la **Figura 11** (Bartz et al., 2023).

Como kNN es un algoritmo no paramétrico, no tiene supuestos para la distribución de datos subyacente. También es un algoritmo considerado perezoso.

En cuanto a las tareas comunes de preparación de los datos, en Orange (2015l) kNN requiere de al menos lo siguiente: (1) Quitar las instancias con etiquetas de clase desconocidas, (2) Recodificar las características categóricas como numéricas, usualmente con una codificación en caliente, (3) Eliminar columnas con datos faltantes, (4) Imputar los valores faltantes con los valores medios y (5) Normalizar los datos centrándolos en la media y escalando a una desviación estándar de 1.

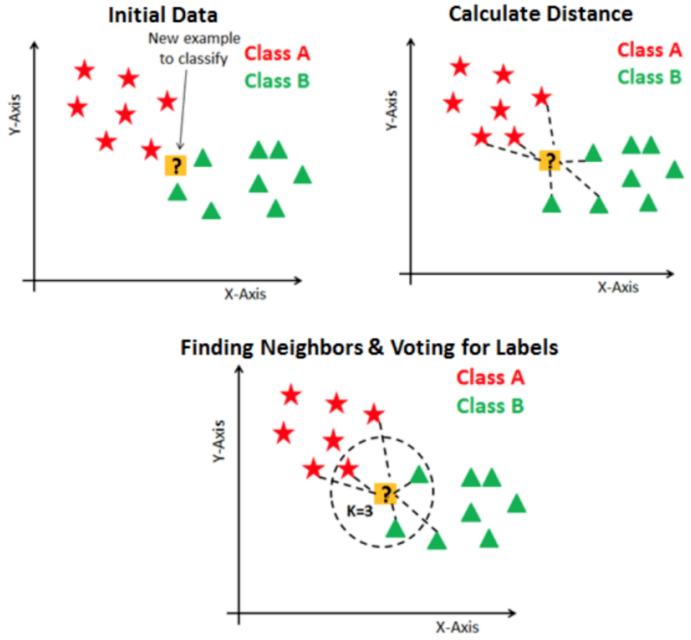


Figura 11: Representación gráfica de Vecinos más cercanos

Orange (2015l) requiere de ajustar ciertos parámetros para kNN. La métrica que puede ser (1) Euclidiana, es la más usada y supone una línea recta como distancia entre dos puntos. (2) Manhattan, suma de las diferencias absolutas de todos los atributos, (3) Máxima, la mayor de las diferencias absolutas entre atributos o (4) Mahalanobis, distancia entre punto y distribución. Además, se debe de especificar los pesos que pueden ser: (1) Uniforme, todos los puntos de cada vecindario se ponderan por igual y (2) Distancia, los vecinos más cercanos de un punto de consulta tienen una mayor influencia que los vecinos más lejanos.

Abu y Maghari (2017) compilaron y documentaron algunas ventajas de kNN: (1) Fácil de entender y simple de implementar, (2) Utilizable para problemas de clasificación y regresión, (3) Ideal para datos no lineales porque no hay suposiciones sobre los datos subyacentes y (4) Puede operar casos multiclase.

También documentaron como desventajas que: (1) Requiere mucho almacenamiento de memoria, (2) Puede ser complejo determinar K, (3) La predicción es lenta si el número de instancias es alto, (4) Es sensible a características irrelevantes y (5) Es sensible a valores atípicos.

Respecto de la estimación de K, si fuese igual a 1, entonces usaremos solo el vecino más cercano para determinar la clase de un punto de datos. Si fuese a 10, entonces usaremos los diez vecinos más cercanos y así sucesivamente.

2.5.2.5. Árboles de decisión

Los árboles de decisión se utilizan como herramienta para la clasificación y regresión. Estos toman como entrada una tabla X, de características numéricas o categóricas y la dividen recursiva y sucesivamente en subtablas, mejorando la llamada puntuación de pureza, que es un mecanismo basado en la proporción de clases individuales respecto de la etiqueta de clase numérica, cuanto mayor es la proporción de una de las clases, más pura es la colección. Una característica numérica puede dividir la tabla en dos partes en torno a su valor medio y una categórica puede dividir la tabla en tantas partes como sus posibles valores individuales. (Yeturu, 2020).

El desafío principal en la implementación del árbol de decisión es seleccionar las características como nodo raíz en cada partición, para lo cual se han desarrollado diversas soluciones como lo son la entropía, ganancia de información, índice de Gini

y relación o ratio de ganancia (Witten & Witten, 2017). En esta investigación, la tasa de Ganancia es utilizada por el algoritmo C4.5 que se implementa en Orange, el índice Gini es utilizado por el algoritmo CART que a su vez está inmerso en los métodos de aprendizaje en conjuntos de Orange (Demšar et al., 2013).

Respecto de los supuestos, los árboles de decisión no hacen suposiciones distributivas, independientes o de varianza constante respecto de los datos de entrenamiento. Sin embargo, Kotu y Deshpande (2019) mencionan supuestos no estadísticos cómo que: (1) Al principio, todo el conjunto de entrenamiento se considera como la raíz. (2) Se prefiere que los valores de las características sean categóricos, si los valores son continuos, entonces se discretizan antes de construir el modelo. (3) Los registros se distribuyen de forma recursiva sobre la base de valores de las características. (4) La selección de características se realiza con algunos de los métodos mencionados en el párrafo precedente.

El software Orange no realiza preprocesamiento de los datos para este modelo (Orange, 2015ae). Pero si requiere de ajustar ciertos hiperparámetros para el árbol C4.5: (1) Seleccionar o no la inducción de un árbol binario. (2) Establecer o no un número mínimo de instancias en las hojas del árbol. (3) Impedir o no la división de los nodos con menos del número dado de instancias. (4) Limitar la profundidad máxima del árbol a N niveles. (5) Detener la división de los nodos después de alcanzar un umbral de mayoría N% especificado. (Orange, 2015ae).

Shobha y Rangaswamy (2018) compilaron y documentaron algunas ventajas de los árboles de decisión: (1) Fáciles de entender, interpretar y visualizar. (2) Pueden manejar datos categóricos y numéricos. (3) Resistentes a valores atípicos, por tanto, requieren de poco preprocesamiento de datos. (4) Aportan con una selección implícita de características. (5) Su rendimiento no se afecta por la relación no lineal de los parámetros. Estos autores documentaron como desventajas al hecho de que son propenso al sobreajuste y además pueden resultar sesgados en caso de clases desbalanceadas.

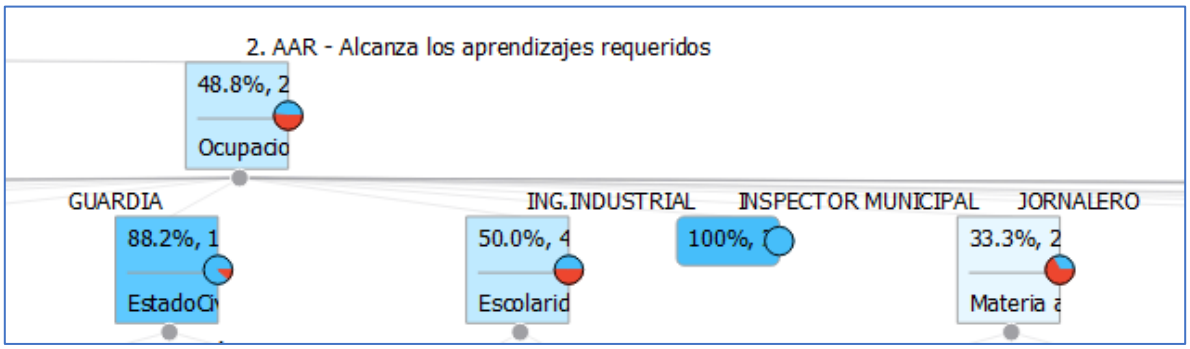


Figura 12: Captura parcial de un árbol de decisión en Orange que expresa relación entre las ocupaciones del padre del alumno y su alcance de promedios del tipo <Alcanza los aprendizajes requeridos> (con poco más de lo justo)

Como se mostró en la **Tabla 10,** para las tareas de regresión los tipos de algoritmos más referidos en la literatura son: Regresión lineal o Generalized Linear Model (GLM), Regresión vectorial de soporte (explicados también como clasificadores), Regresión de procesos gaussianos, aprendizaje en conjunto, árboles de decisión (explicados también como clasificadores) y Redes neuronales.

2.5.2.6. Regresión lineal

El objetivo principal de la regresión es la construcción de un modelo eficiente para predecir variables dependientes a partir de variables o características independientes. Un problema de regresión es cuando la variable de salida es real o un valor continuo, es decir, salario, peso, área o un promedio de calificaciones. Se parte del supuesto de que la variable independiente tiene una relación lineal con la variable dependiente, por lo que es común representarla con una línea recta que exprese el mejor ajuste a un problema dado. La línea se puede representar por la ecuación:

$$Y = b_0 + b_1 x + e$$

Dónde:

- Y: Es la variable dependiente que se va a predecir.
- b_0: Es la intersección, una línea que toca el eje Y.
- b_1: Es la pendiente de la recta,
- x: Representa las variables independientes que determinan la predicción de Y.
- e: Es el error en la predicción resultante.

Luego, una función de costo debe proporcionar los mejores valores posibles para b_0 y b_1, de tal modo que se consiga la línea más adecuada para los puntos de datos (ver **Figura 13**). Para aquello este problema se convierte en un problema de minimización del error entre el valor real y el valor predicho. Para ello, con la siguiente ecuación se eleva al cuadrado la diferencia de error y se suma el error sobre todos los puntos de datos. Luego, se lo divide entre el número total de puntos de datos n.

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

El valor producido por la ecuación es el denominado Error Cuadrático Medio documentado en la sección **2.5.2.11.8. Error cuadrático medio, MSE**.

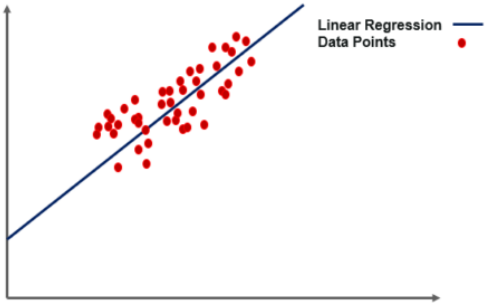


Figura 13: Representación gráfica de Regresión Lineal

Respecto de los supuestos de la regresión lineal, Nørskov (2021) resalta los siguientes: (1) Linealidad, la relación entre X y la media de Y es lineal. (2) Homocedasticidad, la varianza del residual es la misma para cualquier valor de X. (3) Independencia, las observaciones son independientes entre sí y (4) Normalidad, para cualquier valor fijo de X, Y se distribuye normalmente.

En cuanto a las tareas comunes de preparación de los datos, en Orange (2015n) la regresión lineal requiere de al menos lo siguiente: (1) Quitar las instancias con etiquetas de clase desconocidas, (2) Recodificar las características categóricas como numéricas, usualmente con una codificación en caliente, (3) Eliminar columnas con datos faltantes y (4) Imputar los valores faltantes con los valores medios.

Referente de los parámetros de la regresión lineal, en Orange (2015n) se debe especificar por lo menos: (1) Intercepción de ajuste, que por defecto es 0, (2) El modelo para entrenar que puede ser sin regularización, Regularización Ridge (de

penalización con norma L2), Regularización Lasso (de penalización con norma L1) o Elastic Net que combina linealmente las penalizaciones L1 y L2 de Lasso y Ridge.

Respecto de L1, también se conoce como desviaciones mínimas absolutas, errores mínimos absolutos. Básicamente se trata de minimizar la suma de las diferencias **absolutas** $\boldsymbol{S}$ entre el valor objetivo y_i y los valores estimados $f(x_i)$: $S = \sum_{i=1}^{n} |y_i - f(x_i)|$.

Respecto de L2 también se conoce como error de mínimos cuadrados. Básicamente se trata de minimizar la suma del cuadrado de las diferencias $\boldsymbol{S}$ entre el valor objetivo y_i y los valores estimados $f(x_i)$: $S = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2$

El principal beneficio de la regularización es mitigar el sobreajuste controlando el proceso de aprendizaje de un modelo agregando otro término a la función de pérdida (costo) que se busca minimizar. L1 se define como $\alpha \sum(valores\ al\ cuadrado\ de\ los\ coeficientes)$, L2 se define como $\alpha \sum(valores\ absolutos\ de\ los\ coeficientes)$. Alfa es un hiperparámetro que controla la fuerza de regularización, debe ser un float positivo. El valor predeterminado es 1. Los valores más grandes de alfa implican una regularización más fuerte, es decir menos sobreajuste. Los valores más pequeños implican una regularización débil, es decir, sobreajuste (Liu et al., 2018).

En cuanto a las principales ventajas de la regresión lineal se tienen: (1) Funciona muy bien para datos separables linealmente. (2) Se considera fácil de implementar, interpretar y eficiente de entrenar. (3) Gestiona bien el sobreajuste utilizando técnicas de reducción dimensional, regularización y validación cruzada. (4) Se considera que la extrapolación más allá de un conjunto de datos específico es otra gran ventaja de este modelo. Por otra parte, se le considera como desventajas que: (1) Si bien el modelo obliga a la predicción a ser una combinación lineal de características, esto es tanto su mayor fortaleza como su mayor limitación. (2) Es propenso a la multicolinealidad. (3) Es sensible a los valores atípicos. (Hoffmann & Shafer, 2015).

Respecto de la multicolinealidad, es habitual que las características predictoras de la regresión estén correlacionadas, de cierto modo se espera que esta correlación sea fuerte, pero no perfecta o de valor 1. Cuando existe una fuerte relación lineal, pero no perfecta, que se produce sólo entre dos variables explicativas, decimos que se trata de un caso de colinealidad. Sería multicolinealidad cuando la relación lineal

fuerte se produce entre más de dos variables independientes (Hoffmann & Shafer, 2015).

2.5.2.7. Regresión logística

La Regresión Logística, muchas veces referida como modelo Logit, es un modelo estadístico de la probabilidad de que ocurra un evento haciendo que las probabilidades logarítmicas para el evento sean una combinación lineal de una o más variables independientes. Los resultados de salida son probabilidades de variables dependientes categóricas, mismas que pueden ser: (1) Binarias del tipo Pasa o Reprueba, (2) Múltiple o Multinomial, como gatos, perros u ovejas y (3) Ordinal, como baja, media o alta. Un paso intermedio e importante, es estimar los parámetros de un modelo logístico, los cuales son los coeficientes en la combinación lineal (Körner & Waaijer, 2020). En esta investigación se emplea y se documenta la regresión logística multinomial.

La ecuación de partida de los modelos de regresión logística es:

$$P(Y = 1 \mid X) = \frac{exp(b_0 + \sum_{i=1}^{n} b_i x_i)}{1 + exp(b_0 + \sum_{i=1}^{n} b_i x_i)}$$

Dónde:

- $P(Y = 1 \mid X)$ es la probabilidad de que Y tome el valor de 1 o presencia de la característica estudiada.
- X es el conjunto de n covariables $x_1, \dots x_2$ que forma parte del modelo
- b_0 es la constante del modelo o término dependiente
- b_i los coeficientes de las covariables

En cuanto a las tareas comunes de preparación de los datos, en Orange (2015o) la regresión logística requiere de al menos lo siguiente: (1) Quitar las instancias con etiquetas de clase desconocidas, (2) Eliminar columnas con datos faltantes, (3) Recodificar las características categóricas como numéricas, usualmente con una codificación en caliente y (4) Imputar los valores faltantes con los valores medios.

Además de las tareas mencionadas, para la aplicación de la regresión logística se deben cumplir los siguientes supuestos: **(1) Linealidad:** El supuesto de linealidad en regresión logística es que existe una relación lineal entre cada variable predictora continua y el logaritmo de la variable de respuesta. **(2) Independencia de los errores:** Los distintos casos de los datos no deben estar relacionados, por ejemplo,

no podemos medir a la misma gente en diferentes puntos del tiempo. **(3) Multicolinealidad:** Es de considerar que las variables predictoras no deben estar altamente correlacionadas, porque su efecto sería el incremento exagerado de los errores estándar y en ocasiones, del valor estimado para los coeficientes de regresión, lo que hace las estimaciones poco creíbles (Stoltzfus, 2011).

Referente de los parámetros de la regresión logística, en Orange (2015o) se debe especificar por lo menos: (1) El tipo de regularización L1 o Lasso, L2 o Ridge, (2) La fuerza de costo (el valor predeterminado es C = 1).

Una de las principales diferencias entre Lasso y Ridge es que, en la regresión de Ridge, a medida que aumenta la penalización, todos los parámetros se reducen sin dejar de ser cero, mientras que, en Lasso, aumentar la penalización hará que más y más parámetros se lleven a cero. Esta es una ventaja de Lasso sobre la regresión de Ridge, porque los parámetros de conducción a cero anulan la selección de las características de la regresión. Por lo tanto, Lasso selecciona automáticamente las características más relevantes y descarta las demás, mientras que Ridge no descarta por completo característica alguna. Además, La regularización Lasso es más robusta que Ridge porque toma los valores absolutos de los pesos, por lo que el costo solo aumenta linealmente, en tanto que, Ridge toma el cuadrado de los pesos, por lo que el costo de los valores atípicos presentes en los datos aumenta exponencialmente.

En cuanto a las principales ventajas de la regresión logística se tienen: (1) Es simple de comprender y explicar, (2) Es efectiva en la selección de características con el uso de la regularización L1 o Lasso y L2 o de Ridge, (3) Es un modelo rápido de entrenar, (4) No necesita escalar las características de entrada y (5) Tiende a funcionar de manera más eficiente cuando se omiten los atributos que no están relacionados con la variable de salida. Tiene como desventajas que: (1) Requiere de adaptaciones para problemas no lineales, (2) Puede ser sensible a los valores atípicos, (3) Sólo se puede utilizar para predecir un resultado categórico y (4) Tiende a sobre ajustarse. (Körner & Waaijer, 2020; Stoltzfus, 2011).

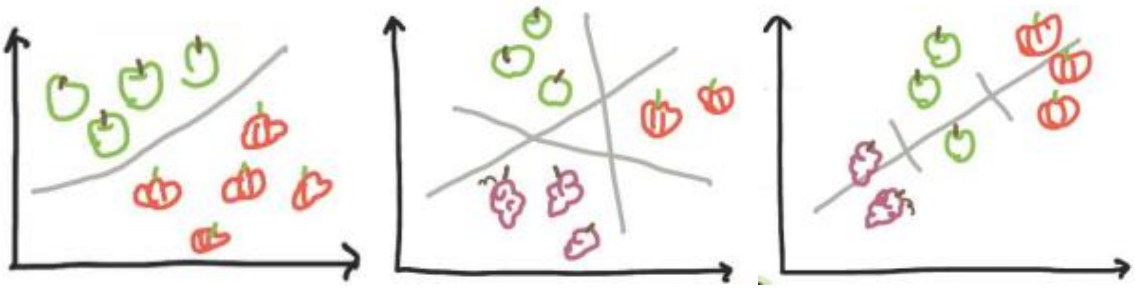


Figura 14: Representación gráfica de regresión logística binaria, multinomial y ordinal.

2.5.2.8. Aprendizaje en conjunto

El aprendizaje en conjuntos, o también referido como algoritmos ensamblados, es un método en el que se utilizan varios algoritmos al mismo tiempo, para lograr un rendimiento predictivo más alto que si se utilizara un algoritmo individual por sí mismo, aunque aumente la complejidad computacional en comparación con los clasificadores individuales. Estos algoritmos construyen muchos árboles en el proceso y realizan la predicción final con base en todos los árboles (Brownlee, 2021a). Los hay de dos tipos, a continuación, se mencionan los aplicados en esta tesis:

- ***Boosting***. – En esta investigación se emplea los algoritmos ADA Boost, Gradient Boost, XG Boost, XG Boost Random Forest y Cat Boost.
- ***Bagging***. – En esta investigación se emplea los Bosques Aleatorios, referidos comúnmente en inglés como Random Forest.

En el ***Boosting***, traducido como impulso o refuerzo, los modelos aprenden secuencialmente con las primeras instancias que ajustan modelos simples a los datos y luego analizan los datos en busca de errores. En los árboles consecutivos el objetivo es mejorar la precisión del árbol anterior, porque cuando una instancia se clasificó erróneamente ante una hipótesis, su peso aumenta para que la siguiente hipótesis tenga más probabilidad de clasificación correcta. En los algoritmos de Boosting, los modelos simples son utilizados secuencialmente, es decir, cada modelo simple va delante o detrás de otro modelo simple.

El ***Bagging***, traducido generalmente como embolsado, es una agregación de Bootstrap que ayuda a mejorar el rendimiento y la precisión de los algoritmos de regresión y clasificación al tratar las compensaciones sesgo-varianza y reducir la varianza de modelos de predicción. Además, de evitar el sobreajuste, porque si una instancia se clasificó incorrectamente, se intenta compensar el peso de esta instancia

para construir modelos predictivos sólidos. Esta naturaleza recursiva de seleccionar las muestras al azar con reemplazo mejora la precisión de un modelo inestable y mejora su generalización a nuevos datos. Los métodos de Bagging son métodos donde los algoritmos simples son usados en paralelo, para aprovecharse de la independencia que hay entre los algoritmos simples, pues el error se puede reducir bastante al promediar las salidas de los modelos simples.

De modo general, los algoritmos que se basan en **Boosting** tienen como ventajas que: (1) Reducen el sobreajuste del modelo, (2) Manejan muy bien los datos de mayor dimensionalidad o número de características y (3) Mantienen la precisión a pesar de la posible ausencia de datos. Tienen como desventaja que, dado que la predicción final se basa en las predicciones medias de los árboles que fungen de subconjuntos, no proporcionan valores precisos para el modelo de clasificación y regresión.

De modo general, los algoritmos que se basan en **Bagging** tienen como ventajas que permiten que muchos árboles débiles combinen esfuerzos para superar a un árbol fuerte. También ayuda en la reducción de la varianza, eliminando así el sobreajuste de modelos en el procedimiento. Como desventajas tienen que: (1) Introducen una pérdida de interpretabilidad de los modelos, (2) A expensas de la precisión que logran son costosos desde el punto de vista computacional y (3) Ignoran el valor con el resultado más alto y el más bajo de cada árbol intermedio y en su lugar proporcionan un resultado promedio de los modelos.

En cuanto a las tareas comunes de preparación de los datos ADA Boost, XG Boost, Gradient Boost y Cat Boost requieren de las siguientes:

- Quitar las instancias con etiquetas de clase desconocidas.
- Eliminar columnas vacías.
- Imputar los valores faltantes con los valores medios
- Con excepción de Cat Boost, en los demás algoritmos es necesario convertir o preprocesar las características categóricas como números, usualmente con una codificación en caliente.

Como estos modelos se basan en árboles, no tienen supuestos estadísticos por validar, pues los modelos basados en árboles son robustos a los valores atípicos en el sentido de que no se requiere que las variables dependientes cumplan con los supuestos de normalidad.

A continuación, se revisan de modo general algunos de los algoritmos de aprendizaje en conjunto:

2.5.2.8.1. ADA Boost

El meta algoritmo Adaptive Boosting, comúnmente referido ADA Boost se clasifica dentro de los modelos de Boosting y fue el primero surgido en esta categoría en 1995 (Freund & Schapire, 1995). Su concepto de adaptativo es con base en la ponderación con pesos a cada estimador (árbol) a partir de la tasa de error de clasificación de las instancias de cada árbol del modelo, al que se denomina aprendiz débil. Una vez actualizados los pesos de las muestras, se normalizan dividiéndolos por la suma total de los pesos, de esta forma, el siguiente aprendiz se ve obligado a concentrarse en aquellas muestras que hayan resultado con pesos mayores. Si bien ADA Boost se emplea en la clasificación binaria, puede generalizarse a múltiples clases (Hastie et al., 2009; Sumathi, 2021).

En cuanto a las principales ventajas de ADA Boost se tienen: (1) Puede utilizarse en conjuntos de datos desbalanceados, (2) Requiere de pocos ajustes de hiperparámetros y parámetros y (3) Es más resistente al sobreajuste o entrenamiento con datos anómalos que muchos otros algoritmos de aprendizaje automático, excepto cuando las distribuciones condicionales de clase tienen una superposición significativa. En cuanto a las principales desventajas: (1) A menudo es sensible a datos ruidosos y valores atípicos y (2) Es más lento que algoritmos como XGBoosting (Zhou et al., 2022).

En cuanto a los principales hiperparámetros por configurar desde Orange (2015q), ADA Boost requiere de especificar al menos lo siguiente:

- **Número de estimadores**. - cuyo valor por defecto es de 50 árboles.
- **Tasa de aprendizaje**. - peso aplicado a cada clasificador en cada iteración de impulso. A más alto aumenta la contribución de cada clasificador.
- **Semilla**. - fija para permitir la reproducción de los resultados.
- **Método de impulso**. - Para la clasificación se tiene como opción SAMME (Stagewise Additive Modeling), que actualiza los pesos del estimador base con los resultados de la clasificación o SAMME. R, que actualiza el peso del estimador base con estimaciones de probabilidad. Para la regresión se tiene como opción de función de pérdida a la Lineal, Cuadrada y Exponencial.

2.5.2.8.2. Gradient Boosting

El algoritmo Gradient Boosting se clasifica dentro de los modelos de Boosting, surgió en 1999 (Friedman, 2001). El aumento de gradiente es una técnica de aprendizaje automático utilizada en tareas de regresión y clasificación. Proporciona un modelo de predicción en forma de un conjunto de árboles de predicción CART débiles construidos de modo secuencial. El algoritmo resultante se suele denominar árboles potenciados por gradiente.

En cuanto a las principales ventajas de **Gradient Boost** se tienen: (1) Puede aproximar la mayoría de las funciones no lineales, (2) Gestiona automáticamente los valores faltantes y (3) No requiere de transformar ninguna variable. En cuanto a las principales desventajas de Gradient Boost se tienen: (1) Puede sobre ajustarse si se ejecuta durante demasiadas iteraciones, (2) Es sensible a valores atípicos y (3) No funciona de la mejor manera sin parámetros (Smirani et al., 2022).

En cuanto a los principales hiperparámetros, Gradient Boosting, XG Boost, XG Boost Random Forest y Cat Boost, comparten los mismos, salvo cuando se indique lo contrario (Orange, 2015i).

2.5.2.8.3. XG Boost

El algoritmo Extreme Gradient Boosting, comúnmente referido XG Boost se clasifica dentro de los modelos de Boosting, surgió en 2014. Se puede usar para modelado predictivo de clasificación o regresión. Los conjuntos se construyen a partir de árboles de decisión que se agregan uno a la vez al conjunto y se ajustan para corregir los errores de predicción cometidos por los anteriores utilizando funciones de pérdida diferenciable y arbitraria y el algoritmo de optimización de descenso de gradiente, lo que le da su nombre de aumento de gradiente, porque el gradiente de pérdida se minimiza a medida que el modelo se ajusta, al igual que una red neuronal (Chen & Guestrin, 2016; Sumathi, 2021).

En cuanto a las principales ventajas de **XGBoosting** se tienen: (1) Es útil para grandes conjuntos de datos, (2) Permite regularización para evitar sobreajustes y (3) Tiene un manejo eficiente de datos faltantes. En cuanto a desventajas, si bien el algoritmo implica el aumento de gradiente, esto se convierte en problema porque vuelve lento el entrenamiento del modelo, más aún en grandes conjuntos de datos (Yu & Liu, 2022).

2.5.2.8.4. XG Boost Random Forest

El algoritmo XG Boost Random Forest es una variante de XG Boost que emplea Random Forest para obtener menores tiempos de entrenamiento. En este sentido, el número de árboles del bosque se constituye en un hiperparámetro clave, pues normalmente, el número de árboles aumenta hasta que el rendimiento del modelo se estabiliza, sin conducir a un sobreajuste dada la naturaleza estocástica del algoritmo (Brownlee, 2021a).

2.5.2.8.5. CatBoost

El algoritmo CatBoost se clasifica dentro de los modelos de Boosting, es de código abierto, surgió en 2017. Proporciona un aumento de gradiente que intenta resolver las características categóricas utilizando una alternativa impulsada e imparcial por permutaciones para construir las divisiones de árbol y elegir los valores hojas a partir de conjuntos de datos diferentes en cada una de las iteraciones. Con esto se mejora a los árboles potenciados por gradiente que tienden a sobre ajustarse en un conjunto de datos pequeño, aunque una vez construidos todos los árboles, los valores hojas del modelo final se calculan mediante el procedimiento estándar del aumento de gradiente (Prokhorenkova et al., 2019).

En cuanto a ventajas de **CatBoost**, la literatura indica que: (1) Es más rápido que muchos otros algoritmos de aprendizaje automático, (2) Soporta tipos de datos categóricos, (3) El muestreo de varianza mínima hace que el número de ejemplos necesarios para cada iteración disminuya, (4) La estructura equilibrada de cada árbol está optimizada para ser más rápidos en GPU y CPU. Respecto de desventajas, no hay mucha documentación, sólo lo referente a la molestia de posible dificultad para ajustar los parámetros de optimización con características categóricas (Ibragimov & Gusev, 2019).

En cuanto a los principales parámetros por configurar en Orange (2015i), Gradient Boosting, XG Boost, XG Boost Random Forest y Cat Boost, comparten los siguientes:

- **Número de árboles.** - indica cuántos árboles potenciados se incluirán. Un gran número suele dar como resultado un mejor rendimiento.
- **Tasa de aprendizaje.** -Establece la tasa de aprendizaje de cada impulso o iteración. La tasa de aprendizaje reduce la contribución de cada árbol.
- **Entrenamiento replicable.** - Corrige la semilla aleatoria, lo que permite la replicabilidad de los resultados.

- **Regularización**. – Especifica el plazo de regularización L2. Disponible solo para los métodos XGBoost y CatBoost.
- **Control de crecimiento.** – (1) Profundidad máxima de cada árbol, (2) Subconjunto más pequeño que se puede dividir, (3) Fracción de instancias de entrenamiento (XG Boost), (4) Porcentaje de características que se usarán al construir cada árbol (XG Boost y Cat Boost), (5) Porcentaje de características que se usarán para cada nivel (XG Boost) y (6) Fracción de características para cada división (XG Boost).

2.5.2.8.6. Random Forests

El algoritmo de bosque aleatorio, comúnmente referido en inglés como Random Forests se clasifica dentro de los modelos de Bagging, su creación fue liderada por Leo Breiman en 2001. Este algoritmo construye un bosque de árboles de decisión, dónde cada árbol genera una predicción a partir de un conjunto de características. Una vez que se han generado todas las predicciones, se toma un voto mayoritario y la clase más predicha forma la predicción final. El nombre bosque aleatorio proviene de la combinación de la aleatoriedad que se utiliza para elegir el subconjunto de datos para los árboles de decisión CART, Classification and Regression Tree (Breiman, 2001; Smith, 2017).

Random Forest no asume supuestos de distribución formales, no son paramétricos, por lo tanto, pueden manejar datos sesgados y multimodales, así como datos categóricos que son ordinales o no ordinales.

En cuanto a las principales ventajas de Random Forest se tienen: (1) Pueden trabajar con cómputo paralelo, (2) Gestionan bien los valores faltantes, (3) No requiere de transformar ninguna variable y (4) No es necesario modificar los hiperparámetros. En cuanto a las principales desventajas de Random Forest se tienen que: (1) Pueden resultar difíciles de interpretar, (2) Tiene desventajas para la regresión al estimar valores en los extremos de la distribución de los valores de respuesta y (3) Son sesgados hacia clases más frecuentes en problemas multiclase (H. Fu & Qi, 2022).

En Orange (2015u), se requiere de especificar al menos lo siguiente para Random Forest:

- **Número de árboles** que se incluirán en el bosque.
- **Número de atributos considerados en cada división,** cuyo valor por defecto es igual a la raíz cuadrada del número de atributos en los datos.

- **Entrenamiento replicable. -** arregla la semilla para la generación de árboles, lo que permite la replicabilidad de los resultados.
- **Pesos de clase para conjuntos de datos desequilibrados.**
- **Control de crecimiento:** (1) Limite o poda de la profundidad de árboles individuales y (2) Especificación del subconjunto más pequeño en que se puede dividir cada árbol.

Para finalizar esta sección de **Boosting** y **Bagging**, la elección de uno u otro depende de ciertas situaciones: Bagging es poco útil en caso de sesgo o falta de ajuste en los datos, en los mismos casos Boosting genera un modelo unificado con menores errores porque optimiza las ventajas y reduce las deficiencias de los árboles en cada iteración. Cuando el análisis de datos presenta como desafío el sobreajuste, el método de Bagging funciona mejor que la técnica de Boosting, pues este viene con el sobreajuste en sí mismo.

2.5.2.9. Redes neuronales

Existen varios tipos de redes neuronales y clasificaciones más detalladas según varias fuentes bibliográficas. Entre la clasificación más referida está el Perceptron Multicapa, Redes Neuronales Convolucionales, Redes Neuronales Recurrentes y Redes Neuronales de Base Radial (Tian et al., 2021).

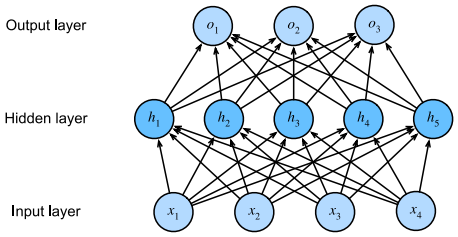


Figura 15: MLP con cinco nodos ocultos en una capa oculta. **Fuente:** Scikit learn (2022b)

En el software Orange (2015k) se implementa el algoritmo de perceptrón multicapa (MLP) con retropropagación, haciendo uso de la librería Sklearn, que puede aprender modelos no lineales y lineales, además, es útil en tareas de clasificación y regresión.

MLP es un algoritmo que aprende una función $f(.): R^m \rightarrow R^0$ desde un conjunto de datos de entrenamiento, donde m es el número de dimensiones para la entrada y o es el número de dimensiones para la salida. Dado un conjunto de características $X = x_1, x_2, \ldots, x_m$ y un objetivo y, una red puede aprender un aproximador de funciones no lineal para tareas de clasificación o regresión. Entre la capa de entrada y la de salida, puede haber una o más capas no lineales llamadas ocultas. La **Figura 15** muestra un MLP con una capa oculta (Scikit Learn, 2022b).

La capa de entrada consiste en un conjunto de neuronas que representan a las características de entrada. Cada neurona en la capa oculta transforma los valores de la capa de entrada con una suma lineal ponderada, seguida por una función de activación no lineal como la hiperbólica, sigmoide logística o de unidad lineal rectificada. La capa de salida recibe los valores de la última capa oculta y los transforma en valores de salida $\{x_i|x_1, x_2, \dots, x_m\}\, w_1x_1 + w_2x_2 + \cdots + w_mx_m\; g(.): R \rightarrow R$

En cuanto a las tareas comunes de preparación de los datos, en Orange (2015q) se requiere de lo siguiente para el Perceptron Multicapa: (1) Quitar las instancias con etiquetas de clase desconocidas, (2) Recodificar las características categóricas como numéricas, usualmente con una codificación en caliente, (3) Eliminar columnas con datos faltantes, (4) Imputar los valores faltantes con los valores medios y (5) Normalizar los datos centrándolos en la media y escalando a la desviación estándar de 1.

Referente de los hiperparámetros del Perceptron Multicapa, en Orange (2015q) se debe de especificar lo listado en las siguientes líneas, mismo que se ha ampliado en definiciones de (Zhang et al., 2022):

- **Neuronas por capa oculta:** definida como el elemento i-ésimo, representa el número de neuronas en la i-ésima capa oculta. Por ejemplo, una red neuronal con 3 capas se puede definir como 2, 3, 2.
- **Función de activación para la capa oculta,** estas deciden si una neurona debe activarse o no calculando la suma ponderada y agregando más sesgo con ella, las opciones disponibles en Orange son:
    - La función Identity, es considerada la más simple de todas, aplica la operación de identidad en sus datos y los datos de salida son proporcionales a los datos de entrada. Tiene el problema de que su derivada es una constante, su gradiente también será una constante y el descenso estará en un gradiente constante. Se formula como $f(x) = x$.
    - La función sigmoide logística, que transforma sus entradas, para las cuales los valores se encuentran en el dominio $\mathbb{R}$, a las salidas que se encuentran en el intervalo (0, 1). Por esa razón, el sigmoide a menudo se llama una función de squashing (aplastamiento), porque "aplasta" cualquier entrada

en el rango (-inf, inf) a algún valor en el Rango (0, 1). Se formula como: $sigmoid(x) = \frac{1}{1+\exp(-x)}$.

- Función hiperbólica o TANho, "aplasta" sus entradas, transformándolas en elementos en el intervalo entre -1 y 1. Se formula como: $\tanh(x) = \frac{1-\exp(-2x)}{1+\exp(-2x)}$.
- Función de unidad lineal rectificada o ReLu, es la función de activación empleada en esta investigación, porque es la opción más popular, simple de implementar y tiene buen desempeño en una variedad de tareas predictivas. Proporciona una transformación no lineal simple, dado un elemento x, la función se define como el máximo de ese elemento y 0. Se formula como $RelU(z) = \max(x, 0)$.

- **Solver** para la optimización del peso, cuyas opciones son:
  - L-BFGS-B, es un optimizador en la familia de métodos cuasi-Newton, que se aproxima a la matriz hessiana que representa la derivada parcial de segundo orden de una función. Además, se aproxima a la inversa de la matriz hessiana para realizar actualizaciones de parámetros (Pedregosa et al., 2011; Scikit Learn, 2022b).
  - Descenso de gradiente estocástico, SGD, que actualiza los parámetros utilizando el gradiente de función de pérdida con respecto a un parámetro que necesita adaptación, es decir, $w \leftarrow w - n\left(\alpha \frac{\vartheta R(w)}{\vartheta w} + \frac{\vartheta LOSS}{\vartheta w}\right)$, donde n es la tasa de aprendizaje que controla el tamaño del paso en la búsqueda en el espacio de parámetros. $LOSS$ es la función de pérdida utilizada para la red.
  - Adam, es un optimizador estocástico basado en gradiente, pero puede ajustar automáticamente valores para actualizar los parámetros en función de las estimaciones adaptables de momentos de orden inferior, que a su vez son valores que resumen o sintetizan propiedades de la variable aleatoria.
- **Alfa,** que es un parámetro de penalización o regularización L2 que ayuda a evitar el sobreajuste penalizando pesos con grandes magnitudes como se explicó para las regresiones lineal y logística en las secciones precedentes.
- **Número** máximo de iteraciones, comúnmente llamado épocas.

En cuanto a las principales ventajas del Perceptron Multicapa se tiene: (1) Capacidad para aprender modelos no lineales y (2) Capacidad para aprender modelos en

tiempo real o en línea. Se le considera como desventajas que: (1) Con capas ocultas tiene una función de pérdida no convexa donde existe más de un mínimo local. Por lo tanto, con diferente peso aleatorio las inicializaciones pueden conducir a una precisión de validación diferente, (2) Requiere ajustar una serie de hiperparámetros, como el número de neuronas, capas e iteraciones ocultas y (3) Es sensible al escalado de características. (Pedregosa et al., 2011; Scikit Learn, 2022b).

MLP admite la clasificación multiclase que es la necesaria para esta investigación, aplicando Softmax como función de salida, misma que convierte un vector de K números reales en una distribución de probabilidad de K posibles resultados. Softmax se usa a menudo como la última función de activación de una red neuronal para normalizar la salida de una red a una distribución de probabilidad sobre las clases de salida predichas, basada en el axioma de elección de Luce.

Para la regresión con MLP se utiliza la retropropagación sin función de activación en la capa de salida, por lo tanto, utiliza el Error Cuadrático Medio (MSE) como función de pérdida y la salida es un conjunto de valores continuos (Pedregosa et al., 2011; Scikit Learn, 2022b).

2.5.2.10. Descenso de gradiente estocástico, SGD

En Orange (2015ab), SGD utiliza un descenso de gradiente estocástico que minimiza una función de pérdida elegida con una función lineal. El algoritmo se aproxima a un gradiente real considerando una muestra a la vez, simultáneamente, actualiza el modelo en función del gradiente de la función de pérdida. Para la regresión, devuelve predictores como minimizadores de la suma, es decir, estimadores M y es especialmente útil para conjuntos de datos dispersos y de gran escala, de hasta de 10^5 instancias de entrenamientos y más de 10^5 características en el caso de Scikit Learn (2022b).

SGD es una técnica de optimización del entrenamiento de un modelo y no corresponden a una familia específica de modelos de aprendizaje automático (Scikit Learn, 2022b). SGD puede considerarse como una aproximación estocástica de la optimización del descenso de gradiente, porque reemplaza el gradiente real calculado a partir de todo el conjunto de datos, por una estimación de este calculada a partir de un subconjunto de datos seleccionado aleatoriamente.

En cuanto a las tareas comunes de preparación de los datos, en Orange (2015ab) se requiere de lo siguiente para SGD: (1) Quitar las instancias con etiquetas de clase

desconocidas, (2) Recodificar las características categóricas como numéricas, usualmente con una codificación en caliente, (3) Eliminar columnas con datos faltantes, (4) Imputar los valores faltantes con los valores medios y (5) Normalizar los datos centrándolos en la media y escalando a la desviación estándar de 1.

Referente de los hiperparámetros de SGD, en Orange (2015s) se debe especificar al menos los siguientes, tal y como también se ilustra en la **Figura 16**:

- **Función de pérdida para la clasificación:**
  - Hinge, linear SVM.
  - Logistic Regression SGD.
  - Modified Huber, pérdida suave que brinda tolerancia a los valores atípicos, así como a las estimaciones de probabilidad.
  - Squared Hinge, Hinge cuadráticamente penalizado.
  - Perceptron, pérdida lineal utilizada por el algoritmo perceptrón.
  - Squared Loss, ajustado a mínimos cuadrados ordinarios.
  - Huber, cambia a pérdida lineal más allá de ε.
  - Epsilon insensitive, ignora errores dentro de ε.
  - Squared epsilon insensitive, la pérdida se eleva al cuadrado más allá de la región ε.
- **Función de pérdida de regresión:**
  - Pérdida al cuadrado, ajustada a mínimos cuadrados ordinarios.
  - Huber, cambia a pérdida lineal más allá de ε.
  - Epsilon insensible, ignora errores dentro de ε, lineal más allá.
  - Insensible a épsilon al cuadrado, la pérdida se eleva al cuadrado más allá de la región ε.
- **Normas de regularización para evitar el sobreajuste.** La intensidad de la regularización define cuánta regularización se aplicará, cuanto menos se regularice, más se permite que el modelo se ajuste a los datos. La red elástica será la relación entre la pérdida L1 y L2, si se establece en 0, entonces la pérdida es L2, si se establece en 1, entonces es L1. L1 y L2 se explicaron en las secciones precedentes de las regresiones lineales y logísticas. Las opciones de regularización son:
  - Ninguno.
  - Lasso, L1, conduce a soluciones escasas.
  - Ridge, L2, regularizador estándar.

- Red elástica, combina linealmente L1 y L2.

- **Hiperparámetros de aprendizaje.**
  - Tasa de aprendizaje:
    - Constante: la tasa de aprendizaje se mantiene igual en todas las épocas o pases.
    - Óptimo: una heurística propuesta por Leon Bottou
    - Escalado inverso: la tasa de ganancia está inversamente relacionada con el número de iteraciones
  - Tasa de aprendizaje inicial.
  - Exponente de escala inversa: decaimiento de la tasa de aprendizaje.
  - Número de iteraciones: el número de pasadas a través de los datos de entrenamiento.
  - Mezclar o no los datos después de cada iteración o pasada.
  - Establecer a la semilla fija para la mezcla aleatoria o no. Sí, permitirá replicar los resultados.

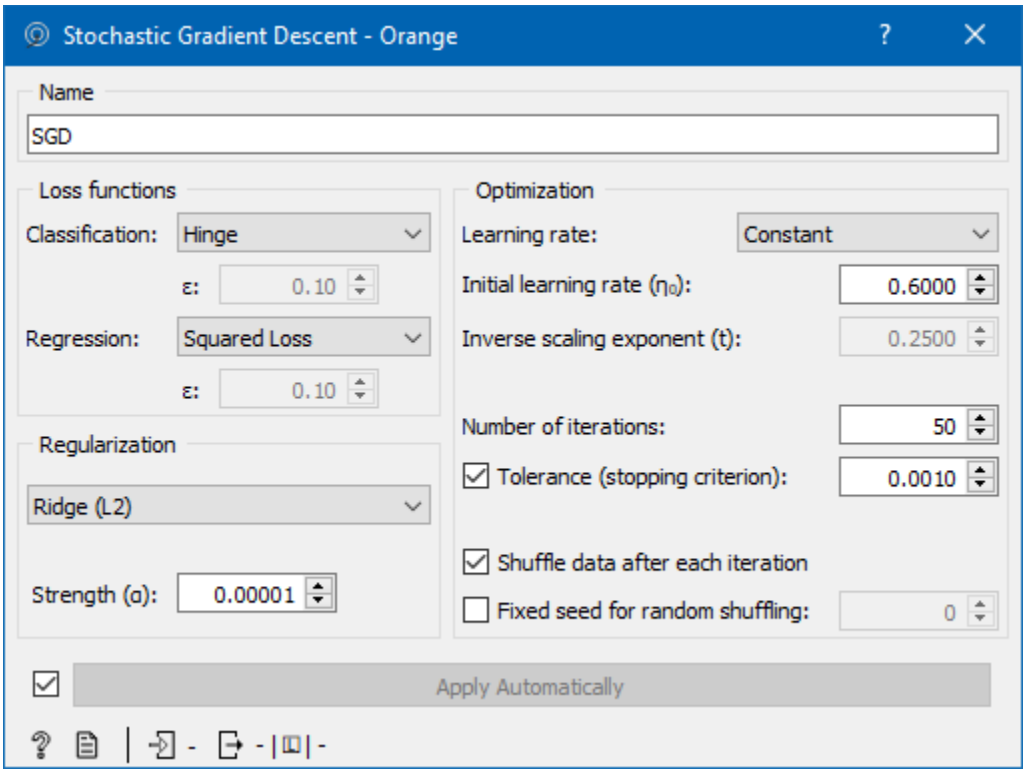


Figura 16: Interfaz de configuración de hiperparámetros de SGD en Orange

En cuanto a las principales ventajas de SGD se tienen su eficacia y personalización en la implementación. Como desventajas se cuentan (1) Requiere una serie de hiperparámetros como la regularización y el número de iteraciones y (2) Es sensible al escalado de características (Scikit Learn, 2022b).

Por mencionar un detalle de la eficiencia, esta es básicamente lineal con el número de instancias de entrenamiento. Si X es una matriz de tamaño (n, p) el entrenamiento tiene un coste de $O\ (kn\bar{p})$, donde k es el número de iteraciones (épocas) y $\bar{p}$ es el número medio de atributos distintos de cero por muestra.

En esta sección referida a SGD por fines optimización de espacio no se han documentado las funciones de pérdida, pero se lo hace en específico para los parámetros empleados en la Fase de Modelado del capítulo de Desarrollo de este documento.

2.5.2.11. Métricas de evaluación de modelos supervisados

En las siguientes secciones, especialmente para las métricas de los modelos de clasificación, se hará uso de algunos términos que se aclaran a continuación (Witten & Witten, 2017):

- Verdadero Positivo TP. - Son las instancias que el modelo predice correctamente como positivas.
- Falso positivo FP. - Son las instancias que el modelo predice como positivas, cuando en realidad son negativas.
- Verdaderos negativos TN. - Son las instancias que el modelo predice correctamente como negativas.
- Falsos negativos FN. - Son las instancias que en realidad son positivas, pero se predicen como negativas.

2.5.2.11.1. Matriz de confusión

En los modelos de clasificación, la matriz de confusión presenta una visión gráfica de los errores cometidos por modelos de clasificación por medio de una tabla de NxN. N representa el número de clases, que en la **Figura 17** son cuatro. En la diagonal principal muestra los valores clasificados correctamente conocidos también como Verdaderos Positivos, TP, los valores fuera de la diagonal que está resaltada en tonalidad verde en la figura, son las clasificaciones incorrectas (Gironés et al., 2017; Orange, 2015b; Pincay Ponce et al., 2020).

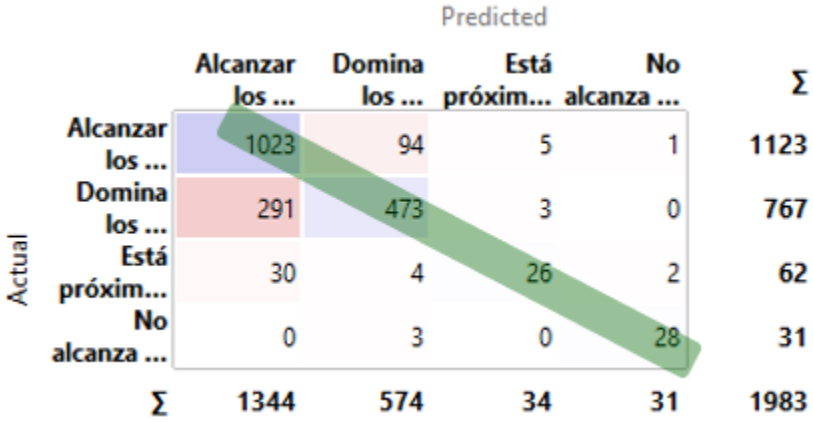


Figura 17: Matriz de confusión 4x4 generada en Orange

2.5.2.11.2. Precisión

En los modelos de clasificación, es una medida de corrección que indica cuántas predicciones son realmente positivas de todo el total positivo predicho. La precisión se define como la relación entre el número total de clases positivas correctamente clasificadas dividido por el número total de clases positivas previstas. La precisión es una métrica que debe ser alta, idealmente 1, es útil en los casos en que los falsos positivos son una preocupación mayor que los falsos negativos (Mukhopadhyay, 2018).

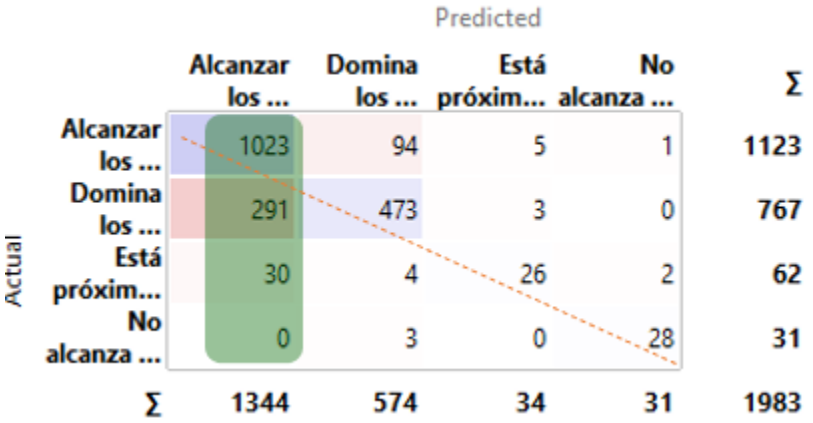


Figura 18: Matriz de precisión con datos resaltados para determinar la Precisión de la clase <Alcanza los aprendizajes requeridos> de los alumnos estudiados.

En función de lo indicado en el párrafo anterior, la precisión de la clase <Alcanza los aprendizajes requeridos> se calcula como:

$$Precisión = \frac{TP}{TP + FP}$$

$$= \frac{Predicciones\ actualmente\ positivas}{Total\ de\ predicciones\ positivas} = \frac{1023}{1023 + 291 + 30 + 0} = \frac{1023}{1344} = 0.761$$

El total de la precisión del clasificador se calcula promediando la precisión de todas las clases: (0.761+0.824+0.765+0.903) / 4=0.88

2.5.2.11.3. Exactitud (Accuracy)

En los modelos de clasificación, la exactitud, a veces referida como CA, de classification accuracy, mide la frecuencia con la que el clasificador hace la predicción correcta, es la relación entre el número de predicciones correctas y el número total de predicciones. Esta métrica no es adecuada para clases desequilibradas, porque en tal caso, cuando el clasificador predice que cada punto pertenece a la etiqueta de clase mayoritaria, la exactitud será alta, pero, el modelo no es preciso (Gironés et al., 2017; Mukhopadhyay, 2018).

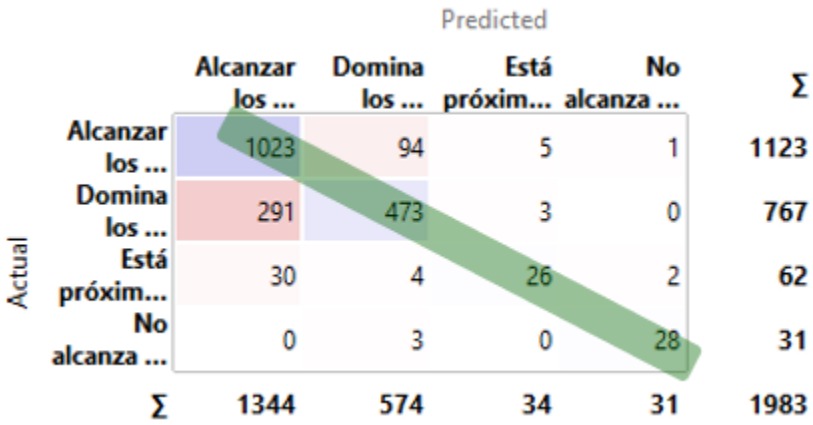


Figura 19: Matriz de precisión con datos resaltados para determinar la Exactitud del clasificador de los alumnos estudiados.

Esta métrica indica cuántas predicciones son realmente positivas de todo el total positivo predicho. Se calcula como la suma de todos los valores verdaderos dividida por los valores totales:

$$Exactitud = \frac{TP_1 + TP_2 + \cdots + TP_n}{TP + FP}$$

$$= \frac{Suma\ de\ Verdaderos\ Positivos}{Suma\ de\ todos\ los\ valores\ de\ las\ predicciones}$$

$$= \frac{1023 + 473 + 26 + 28}{1023 + 94 + 5 + 1 + 291 + 473 + 3 + 0 + 30 + 4 + 26 + 2 + 0 + 3 + 0 + 28} = \frac{1550}{1983} = 0.781$$

Si se desea conocer la exactitud de cada clase, existen el método 1 vs 1 que produce una matriz de confusión, con formato 2x2, para cada par de clases existentes. Por otra parte, la aproximación 1 vs Todo, que produce una matriz de confusión 2x2 para cada clase (Gironés et al., 2017).

2.5.2.11.4. Recuerdo (Recall)

En los modelos de clasificación, es una medida de observaciones reales que se predicen correctamente, también se conoce como sensibilidad, es útil cuando se desea capturar tantos aspectos positivos como sea posible, porque que el falso negativo resulte mayor que el falso positivo. El recuerdo debe ser alto, idealmente 1, en todo caso, la métrica es importante en situaciones médicas o académicas como compete a esta investigación, donde no importa si se levanta una falsa alarma, pero los casos positivos reales no deben pasar desapercibidos (Gironés et al., 2017; Mukhopadhyay, 2018).

		Predicted				
		Alcanzar los ...	Domina los ...	Está próxim...	No alcanza ...	Σ
Actual	Alcanzar los ...	1023	94	5	1	1123
	Domina los ...	291	473	3	0	767
	Está próxim...	30	4	26	2	62
	No alcanza ...	0	3	0	28	31
	Σ	1344	574	34	31	1983

Figura 20: Matriz de precisión con datos resaltados para determinar el Recuerdo de la clase <Domina los aprendizajes requeridos> de los alumnos estudiados.

El recuerdo se define como la relación entre el número total de clases positivas correctamente clasificadas, 473 en la matriz de ejemplo, dividido por el número total de clases positivas, 767 en la matriz de ejemplo.

$$Recuerdo = \frac{TP}{TP + FN}$$

$$= \frac{Predicciones\ actualmente\ positivas}{Total\ de\ predicciones\ de\ la\ clase}$$

$$= \frac{473}{291 + 473 + 3 + 0} = \frac{473}{767} = 0.617$$

El total del Recuerdo del clasificador se calcula promediando los valores de Recuerdo de todas clases: (0.911+0.617+0.419+0.903) / 4=0.713

2.5.2.11.5. F1 Score

En los modelos de clasificación, es una media armónica ponderada de la precisión y la recuperación, reporta un valor entre 0 y el ideal que es 1. Emplea la media armónica porque no es sensible a valores grandes como si lo son los promedios simples. Si la precisión es baja, F1 también, si el recuerdo es bajo nuevamente F1 también. En la práctica, cuando la precisión de un modelo es mayor, el recuerdo disminuye y viceversa. La puntuación F1 captura ambas tendencias en un solo valor.

		Predicted				
		Alcanzar los ...	Domina los ...	Está próxim...	No alcanza ...	Σ
Actual	Alcanzar los ...	1023	94	5	1	1123
	Domina los ...	291	473	3	0	767
	Está próxim...	30	4	26	2	62
	No alcanza ...	0	3	0	28	31
	Σ	1344	574	34	31	1983

Figura 21: Matriz de precisión con datos resaltados para determinar F1 en la clase <Domina los aprendizajes requeridos> de los alumnos estudiados.

A continuación, se calcula F1, concebido como la media armónica de 2 valores, Recuerdo y Precisión:

$$F1\ Score = 2 * \left(\frac{Recuerdo\ x\ Precisión}{Recuerdo + Precisión}\right)$$

$$= 2 * \frac{0.617x\ 0.824}{0.617 +\ 0.824} = 2 * \left(\frac{0.508}{1.441}\right) = 0.705$$

2.5.2.11.6. Especificidad

En los modelos de clasificación, es una métrica que indica cuántas de las instancias negativas reales se predijeron correctamente. Es el negativo verdadero dividido por el número total de valores negativos reales. La especificidad es importante cuando

se considera que los verdaderos negativos son más importantes que los falsos negativos, por ejemplo, una prueba de dopaje en la que no se quiere que ningún atleta libre de drogas sea clasificado y prohibido erróneamente. Como otro ejemplo, en el área de la salud, la especificidad es la capacidad de poder identificar los casos de pacientes sanos entre todos los sanos. Se define como:

$$Especificidad = \frac{TN}{TN + FP}$$

2.5.2.11.7. Curva ROC

En los modelos de clasificación, la curva ROC, acrónimo de Receiver Operating Characteristic, mide el rendimiento respecto de dos métricas que se obtienen de la matriz de confusión: Tasa de Verdaderos Positivos (TPR) y Tasa de Falsos Positivos (FPR). La TPR es lo mismo que la métrica de Recuerdo, es decir TPR= TP / (TP+FN). A su vez, FPR es la relación entre las predicciones de falsos positivos y el número total de muestras negativas reales. Es lo mismo que 1 — Especificidad: FPR= FP / (FP + FN). La curva ROC se traza en base a TPR y FPR (Brownlee, 2021; Herrera et al., 2016).

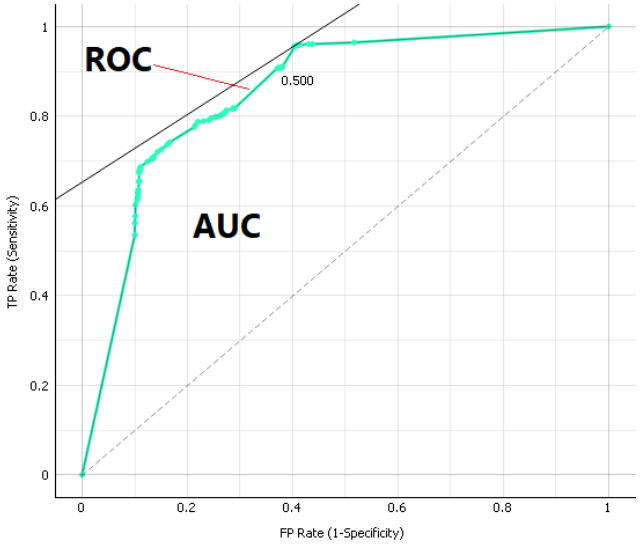


Figura 22: Curva ROC para la clasificación de la clase <Alcanza los promedios requeridos> = 1 en un 57%.

En la **Figura 22**, de forma predeterminada, el umbral de clasificación es 0.5. Cualquier probabilidad superior a 0.5 se clasificará como clase 1 = <Alcanza los promedios requeridos> y por debajo de 0.5 se clasificará como clase 0, es decir, otra clase (Ver **Tabla 1**). En la Figura, cuanto más cerca siga la curva celeste del borde izquierdo y

luego del borde superior del espacio ROC, más preciso será el clasificador. La probabilidad obtenida para la clase del ejemplo fue de 57%, que significa que tal clase se clasificará de modo correcto con esa posibilidad.

Ahora que se han visto, algunas de las principales métricas para los modelos de clasificación, se puntualiza en que la Exactitud se utiliza cuando los verdaderos positivos y los verdaderos negativos son más importantes y los datos están equilibrados. La Precisión es ideal cuando el falso positivo sea más relevante al estudio. El Recuerdo es ideal cuando el falso negativo sea más relevante al estudio. F1 es ideal cuando los falsos negativos y falsos positivos sea relevantes al estudio, además, de que es ideal para datos desequilibrados.

De acuerdo con Brownlee (2016), son varias las métricas para evaluar el rendimiento de un modelo de regresión, dónde se predice un valor numérico. En general los modelos cometen errores, disminuyendo el rendimiento, así que cuanto mayor sea la diferencia entre el resultado real 'y' y el resultado predicho 'ŷ' el modelo es impreciso, en tanto que, cuanto más cercanos sean los valores el modelo es mejor. Es importante determinar una puntuación de error para resumir la habilidad predictiva de los modelos.

2.5.2.11.8. Error cuadrático medio, MSE

En los modelos de regresión, mide el promedio de los cuadrados de los errores o desviaciones, es decir, la diferencia entre el estimador y lo que se estima. Siempre es positivo. Utilizando esta métrica se penaliza los residuos grandes generados por valores atípicos. En este sentido, si el modelo se aproxima correctamente a gran parte de las instancias del conjunto de datos, pero comete importantes errores en unos pocos, la penalización será alta (Brownlee, 2016). Se define como:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_j\right)^2$$

Siendo N el número de instancias, 'y' el resultado real y 'ŷ' el resultado predicho.

Un MSE perfecto es 0, lo que significa que todas las predicciones coincidieron exactamente con los valores esperados, aunque tal caso quizá signifique que el problema de modelado predictivo es trivial, por lo que Brownlee (2016), sugiere establecer un MSE de referencia mediante un modelo predictivo ingenuo y luego, el modelo que logre un mejor MSE que aquel, es mejor.

2.5.2.11.9. Raíz del error cuadrático medio, RMSE

En los modelos de regresión, es la raíz cuadrada de la media aritmética de los cuadrados de un conjunto de números, una medida de imperfección del ajuste del estimador a los datos.

$$RMSE = \sqrt{MSE}$$

Un RMSE perfecto es 0, aunque tal caso quizá signifique que el problema de modelado predictivo es trivial, por lo que Brownlee (2016), sugiere establecer un RMSE de referencia mediante un modelo predictivo ingenuo y luego, el modelo que logre un mejor RMSE que aquel, es mejor.

2.5.2.11.10. Error absoluto medio, MAE

En los modelos de regresión, se utiliza para medir qué tan cerca están los pronósticos o predicciones de los resultados eventuales.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_i - \hat{y}_j|$$

Siendo N el número de instancias, 'y' el resultado real y 'ŷ' el resultado predicho. Como MAE toma el valor absoluto, siempre es positivo y todos los errores se ponderarán en la misma escala lineal. Por lo tanto, a diferencia del MSE, no da demasiado peso a los valores atípicos y la función de pérdida proporciona una medida genérica y uniforme de qué tan bien está funcionando un modelo. Como desventaja, en los errores mayores provenientes de los valores atípicos, se terminan ponderando igual que los errores más bajos (Brownlee, 2016; Gironés et al., 2017).

Al igual que MSE y RMSE, Brownlee (2016) sugiere establecer un MAE de referencia mediante un modelo predictivo ingenuo y luego, el modelo que logre un mejor MAE que aquel, es mejor.

2.5.2.11.11. R cuadrado, R^2

En los modelos de regresión, se interpreta como la proporción de la varianza en la variable dependiente que es predecible a partir de la variable independiente. Si el R^2 de un modelo es 0.80, entonces aproximadamente el 80% de la variación observada puede explicarse por las características del modelo (Brownlee, 2016).

Se calcula como: R^2= Variación explicada / Variación total

Dónde:

- **La variación explicada**. - Se determina como $\hat{y} - \bar{y}$ siendo $\bar{y}$ la media de los valores de y. Es la variación en 'y' que se explica por un modelo de regresión.
- **Variación inexplicable**. – Se determina como: y – ŷ. Es la variación en 'y' que no es capturada por un modelo de regresión. También se conoce como el residuo de un modelo de regresión.
- **Variación total**. - Es la suma de la variación inexplicada y la variación explicada, que también es $y - \bar{y}$, siendo $\bar{y}$ la media de los valores de y.

Por tanto:

$$R^2 = \frac{\hat{y} - \bar{y}}{y - \bar{y}}$$

R^2 está estrechamente relacionada con la MSE, pero tiene la ventaja de estar libre de escala, es decir, no importa si los valores de salida son muy grandes o pequeños, el R^2 siempre estará entre -∞ y 1.

2.5.2.11.12. N Error cuadrático medio de la raíz, NRMSE

En los modelos de regresión, es el RMSE normalizado por el valor medio de los valores reales $\bar{y}$, se expresa como un porcentaje, donde los valores más bajos indican menos varianza residual. En muchos casos, especialmente para muestras más pequeñas, es probable que el rango de muestra se vea afectado por el tamaño de la muestra, lo que dificultaría las comparaciones, aunque de modo general, la normalización del RMSE facilita la comparación entre conjuntos de datos o modelos con diferentes escalas, pese a que no existe un medio consistente de normalización en la bibliografía (Brownlee, 2016; Karlsson, 2021).

$$NRMSE = \frac{RMSE}{\bar{y}}$$

En el desarrollo de esta tesis, el análisis predictivo permite encontrar un modelo relacional, basado en minería de datos, que explica las variaciones en las calificaciones obtenidas por los alumnos de distintos años básicos e incluso de otras segmentaciones, esto brinda sustento para ejecutar intervenciones que impulsen la mejora del rendimiento en las materias curriculares analizadas.

De acuerdo con la revisión sistemática de la literatura que complementa esta tesis, a nivel del empleo de modelos predictivos para el estudio de incidencias sobre el

rendimiento académico escolar, se ha frecuentado a los clasificadores basados en árboles, este tipo de análisis da como resultado una forma de árbol de sencilla interpretación, alta precisión, poder predictivo y de uso regular en la selección de características para análisis de datos tanto continuos como discretos. Algunos de los métodos utilizados en la literatura son el Algoritmo C4.5, Bosques aleatorios, Reglas de Clasificación y diversas variantes de Redes Neuronales (Pincay-Ponce et al., 2023).

También se ha frecuentado a los clasificadores basados en funciones, tales como máquinas de vectores de soporte, perceptron multicapa, redes de base radial, regresión lineal jerárquica, regresión lineal simple, regresión logística multinomial y regresión logística simple. Además, algunos investigadores emplearon método de selección de características fuertemente relacionadas con la clase por predecir, o emplearon para tal finalidad, entre ellos al análisis de componentes principales y las correlaciones bivariadas (Pincay-Ponce et al., 2023).

2.5.3. Modelos no supervisados

En el aprendizaje automático y en la Inteligencia Artificial en general, los modelos descriptivos que se corresponden con el aprendizaje no supervisado. Se utilizan cuando el problema requiere una cantidad masiva de datos sin etiquetar, por ejemplo, en las redes sociales, dónde al no haber etiquetas, no es posible hacer predicciones, pues no hay respuestas asociadas a cada observación. Estos modelos requieren de algoritmos que ejecutan principalmente tareas de agrupación de instancias y de reducción la dimensionalidad con base en patrones encontrados mediante procesos iterativos ejecutados sin intervención humana, tal cual ilustra la **Figura 23** (Nelli, 2018).

El agrupamiento es la tarea de dividir la población o los puntos de datos en varios grupos de observaciones similares según un criterio prefijado. Cada grupo de datos será mutuamente excluyente de tal manera que cada miembro de un grupo esté lo más cercano posible a otro elemento y los grupos diferentes estén lo más lejos posible entre sí. La agrupación, sus propiedades y la cuantificación de la calidad de los grupos difiere de un algoritmo a otro (Russo, 2019; Yeturu, 2020).

La reducción de dimensionalidad es la tarea de reducir un conjunto de datos de alta dimensión a uno con menos dimensiones de modo que cada una de las dimensiones más bajas transmita mucha más información, dado que procesar datos de gran tamaño es complejo por requerir de más potencia de procesamiento y espacio (Nelli, 2018).

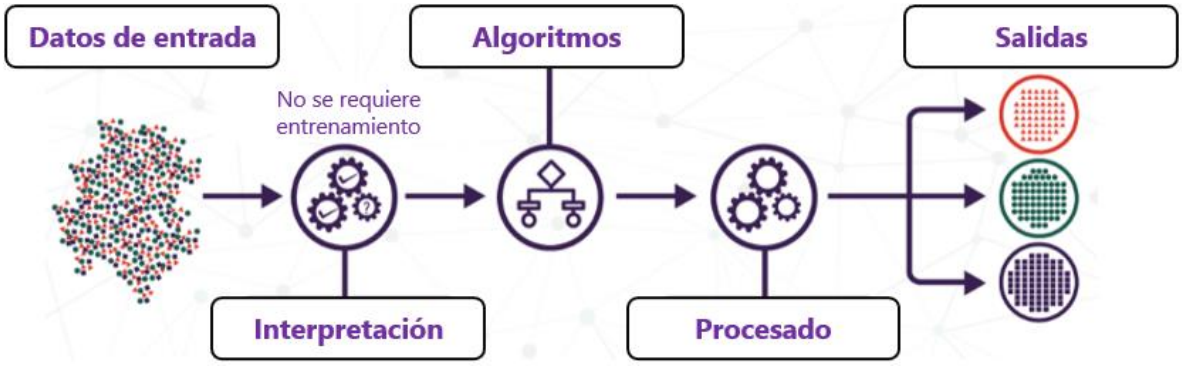


Figura 23: Esquema general del funcionamiento de los modelos no supervisados

A continuación, se hace un repaso de los modelos no supervisados que se emplean en esta investigación:

2.5.3.1. Patrones frecuentes, FP-Growth

FP-Growth es un algoritmo para extraer conjuntos de elementos que aparecen con frecuencia en una base de datos transaccional, tal frecuencia es un umbral de soporte especificado por el usuario. Existen varios algoritmos para este problema, como Apriori (Agrawal & Srikant, 1994), Eclat (Zaki, 2000) y FP-Growth (Han et al., 2004). Surgió como una alternativa respecto de A priori, con quien se lo suele comparar en la literatura, a diferencia de este, FP-Growth utiliza una estructura de datos llamada árbol FP para agregar ítems a los patrones, en lugar de ampliarlos con un ítem candidato a la vez (generación de candidatos), con esto se consigue un algoritmo particularmente atractivo para grandes conjuntos de datos (Shabtay et al., 2021).

En Orange (2015h), la implementación de FP-Gwowth permite parametrizar o filtrar los itemsets frecuentes según el soporte porcentual de instancias en cada grupo respecto del total de instancias, establecer un máximo N de itemsets frecuentes ordenados por soporte, seleccionar de conjuntos con un mínimo o máximo de instancias, así como el filtrado por expresiones regulares.

2.5.3.2. K-Means

K-Means determina puntos de datos similares en grupos minimizando la distancia media entre puntos geométricos. Para ello, particiona iterativamente los conjuntos de datos en K subgrupos no superpuestos, en los que cada punto de datos pertenece al clúster con el centroide o centro de clúster medio más cercano. Existen variaciones de este método no supervisado, como lo son K-Medoids o Fuzzy K-Means, así como

formulaciones como un algoritmo de descenso de gradiente que converge mucho más rápido que el enfoque iterativo mencionado por lo que resulta idóneo para, por ejemplo, el aprendizaje en línea, pues la actualización del centroide corresponde a la actualización del peso (Yeturu, 2020).

En Orange (2015k), K-Means que también es conocido como Lloyd, calcula y muestra las k agrupaciones y sus puntuaciones acordes con la métrica de la Silueta, dónde cuanto mayor sea la puntuación de Silueta, mejor será la agrupación. Existen métricas adicionales no cubiertas en este documento.

En cuanto a las tareas comunes de preparación de los datos, en Orange (2015k) se requiere de lo siguiente para K-Means: (1) Recodificar las características categóricas como numéricas, usualmente con una codificación en caliente y (2) Imputar los valores faltantes con los valores medios.

Referente de los hiperparámetros de K-Means, en Orange (2015k) se debe especificar al menos:

- **El número de clústeres**, que por defecto es 3 pero ocasionalmente se debe incrementar para generar más centroides y mejorar la separación entre grupos.
- **Desde X a Y,** es un hiperparámetro de Orange que permite visualizar puntajes de agrupación para el rango de clúster seleccionado usando el puntaje de Silhouette, lo que hace es contrastar la distancia promedio a los elementos en el mismo clúster con la distancia promedio a los elementos en otros clústeres.
- **Preprocesar o no,** si se activa las columnas se normalizan, es decir, la media es centrada en 0 y la desviación estándar es escalada en 1.
- **Método de inicialización**, es la forma en que el algoritmo comienza a agruparse. Las opciones son:
    - k-Means ++, el primer centroide se selecciona al azar, los siguientes se eligen de los puntos restantes con probabilidad proporcional a la distancia al cuadrado desde el centro más cercano. El algoritmo implementado es voraz k-Means ++, que se diferencia de otro importante como Vanilla k-Means++ en que realiza varios ensayos en cada paso de muestreo y elige el mejor centroide entre ellos (Scikit Learn, 2022a).
    - Inicialización aleatoria, los grupos se asignan aleatoriamente al principio y luego se actualizan con más iteraciones.

- **Repeticiones**, indica cuántas veces se ejecuta el algoritmo desde posiciones iniciales aleatorias; se usará el resultado con la suma de cuadrados más baja dentro del grupo. Número de veces que se ejecuta el algoritmo k-Means con semillas de centroide diferentes. Se recomiendan varias ejecuciones para problemas dispersos de alta dimensión. (Scikit Learn, 2022a).
- **Iteraciones máximas**, es el número máximo de iteraciones dentro de cada ejecución del algoritmo

En cuanto a las principales ventajas de K-Means se tiene: (1) Relativamente fácil de implementar, (2) Escala a grandes conjuntos de datos, (3) Garantiza la convergencia, (4) Puede iniciar en caliente las posiciones de los centroides y (5) Se adapta con facilidad a nuevas instancias. Se le considera como desventajas que: (1) Requiere de elegir K manualmente, de preferencia yendo con un K de menos a más, (2) Es propenso a problemas de formación de grupos de diferentes tamaños y densidades, por lo que se aconseja escalar los datos, (3) Los centroides pueden ser arrastrados por valores atípicos, o los valores atípicos pueden obtener su propio grupo en lugar de ignorarse, por lo que es necesario eliminar o recortar los valores atípicos antes de agrupar y (4) A medida que aumenta el número de dimensiones, una medida de similitud basada en la distancia converge a un valor constante entre las instancias dadas, por lo que se recomienda reducir la dimensionalidad de los datos. (Google Developers, 2022a).

2.5.3.3. Clúster jerárquico

La agrupación jerárquica es un método de aprendizaje no supervisado para agrupar puntos de datos en una jerarquía, dónde el punto de enlace es un conjunto de clústeres distintos entre sí y los objetos dentro de cada clúster son muy similares entre sí. Gráficamente los grupos son dendrogramas visualizados a modo de árbol de puntos basados en métricas de similitud o diferencia (Brownlee, 2016; Yeturu, 2020).

El dendrograma es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de instancias en cada paso y sus niveles de similitud, mismos que medidos en el eje vertical, en tanto que las diferentes observaciones se especifican en el eje horizontal, con relación a la figura de la derecha (Minitab, 2023).

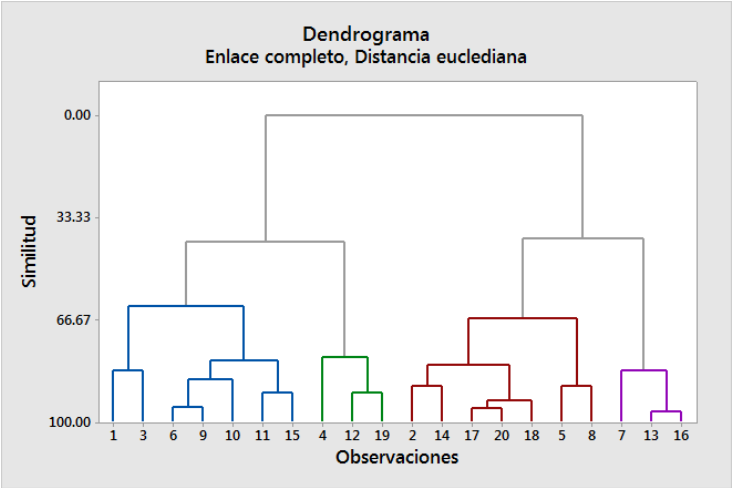


Figura 24: Representación gráfica de un dendrograma

Las puntuaciones de distancia se definen entre cada par de puntos. En Orange se dispone del widget Distances (2015e), desde dónde la métrica de distancia es altamente personalizable y captura cualquier noción de disimilitud, entre las opciones están la medida Euclidiana, Manhattan, Hamming, Mahalanobis, Bhattacharyya, Jaccard o el puntaje negativo de similitud. Una tarea previa del cálculo de las distancias es la normalización en columnas, cuyos valores son centrados en cero y escalados. Los valores faltantes se imputan con el valor promedio de la fila o la columna. En el caso de datos categóricos, la distancia es 0 si los dos valores son iguales ('verde' y 'verde') y 1 si no lo son ('verde' y 'azul').

Si bien, cada medida tiene sus ventajas según el contexto de uso, para esta investigación dónde se tiene una considerable cantidad de atributos, es decir, una alta dimensionalidad, se empleó la medida de Manhattan, que en términos de Aggarwal, Hinneburg y Keim (2001), es ideal para aplicaciones de alta dimensión y se define como la sumatoria del valor absoluto de las diferencias de las distancias entre los puntos en un plano cartesiano: $\sum_{j=1}^{p}|y_{1j} - y_{2j}|$, dónde p es el número de características, y_{1j} es el valor del descriptor j en la entidad 1 y y_{2j} es el valor del descriptor j en la entidad 2.

Una vez que se definen las distancias entre cada par de puntos, se procede con la identificación iterativa de subgrupos mediante el widget Hierarchical Clustering (2015j). Luego, generalmente se filtran las filas y se exploran los clústeres generados tal cual se muestra en la **Figura 25**, que corresponde al flujo parcial construido en Orange.

El procedimiento de identificación puede ser de arriba hacia abajo (divisorio) o de abajo hacia arriba (aglomerativo). En el proceso de arriba hacia abajo, al principio, se supone que todos los puntos están en un solo grupo. A medida que itera el algoritmo, el grupo único se subdivide. Sin embargo, el agrupamiento divisivo tiene que examinar un número exponencial de subconjuntos para determinar dónde dividir. Este problema se mitiga con un enfoque ascendente. En esta formulación, al principio, todos los puntos se asignan a grupos individuales. Luego, los grupos se fusionan para dar como resultado el siguiente nivel al definir la noción de distancias de grupo a grupo (Yeturu, 2020).

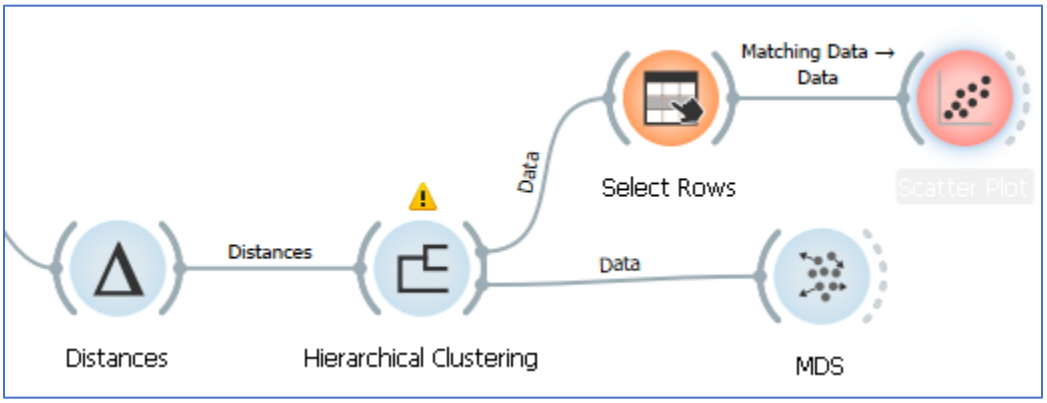


Figura 25: Vista parcial de un flujo para construcción de agrupamiento jerárquico en Orange

En cuanto a las tareas comunes de preparación de los datos para el Clúster jerárquico, en Orange (2015j) se requiere del mismo cumplimiento para el cálculo de las distancias explicados en párrafos precedentes. Respecto de los parámetros, la agrupación jerárquica requiere de una medida de disimilitud y un método de vinculación. Algunas herramientas ofrecen parámetros como la visualización de N clústeres, un porcentaje de los N clústeres, entre otros. Respecto de los métodos de vinculación, estos definen la forma y fórmula de medir las distancias entre clústeres, en durante los pasos de actualización de la matriz de distancias. En Orange se dispone de Single, Average, Weighted, Complete y Ward.

- Single calcula la distancia entre los elementos más cercanos de los dos clústeres. Se formula como $D_{uk} = \min(D_{ik}, D_{jk})$.
- Average calcula la distancia promedio entre los elementos de los dos grupos.
- Weighted utiliza el método ponderado de grupos de pares utilizando promedios aritméticos, WPGMA por sus siglas en inglés. Se formula como $D_{uk} = \frac{1}{2}(D_{ik} + D_{jk})$. Es el método empleado en esta investigación.
- Complete calcula la distancia entre los elementos más distantes de los grupos.

Este método tiende a encontrar grupos compactos de diámetros aproximadamente iguales. Se formula como $D_{uk} = \max(D_{ik}, D_{jk})$.

- Ward calcula el aumento de la suma de errores de cuadrados, es decir, minimiza la varianza total dentro del grupo.

Una apreciable ventaja del agrupamiento jerárquico es la posibilidad de cortar a un N número de niveles de clústeres. Este tipo de agrupamiento se adapta bien a datos jerárquicos, como las taxonomías. (Google Developers, 2022b). Una forma gráfica y análoga a los dendrogramas para representar al agrupamiento jerárquico y no jerárquico es la siguiente:

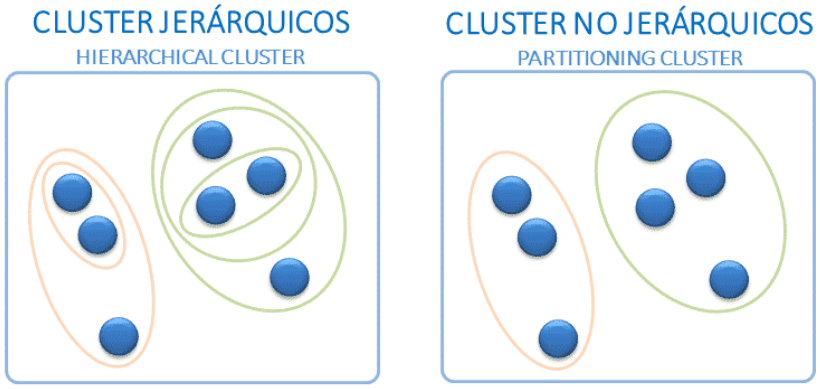


Figura 26: Representación gráfica de las diferencias entre agrupamiento jerárquico y no jerárquico.

2.5.3.4. Reglas de asociación

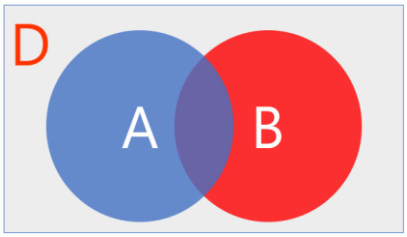
Las reglas de asociación se utilizan para explorar conjuntos de datos transaccionales e identificar patrones e interacciones entre características, para expresarlas e interpretarlas en reglas de la forma $if\ (antecedente)\ then\ (consecuente)$, dónde tanto el antecedente como el consecuente se expresan como $Atributo1 = Valor1\ y\ Atributo2 = Valor2\ y\ \ldots AtributoN = ValorN$. Ejemplo: $if\ (Color = 'Claro')\ then\ (Diámetro = 'Chico')$. En Orange, las reglas de asociación están implementadas con el algoritmo de patrones frecuentes FP-Growth y con optimización de bucketing. (Abdulhamit, 2020; Orange, 2016; F. Quiroga, 2020).

Bucketing es una técnica de optimización que consiste en agrupar los datos de una tabla con base en el valor de una o varias columnas, de forma que las instancias con igual valor caigan en el mismo bucket, aunque un bucket puede tener valores

distintos de la columna sobre la que se hace bucketing (Pincay Ponce et al., 2020). FP-Growth es explicado en la siguiente sección.

De acuerdo con F. Quiroga (2020), el objetivo con las reglas es obtenerlas con alto valor en algunas de sus cuatro principales métricas: Cobertura, Confianza, Soporte e Interés. Si se denota a |A| como el antecedente de la regla y a |B| como el consecuente, con el apoyo del siguiente Diagrama de Venn se comprenderá cada métrica. Esta explicación se amplía en la **Tabla 11**.

|A| = Número de instancias que cumplen A
|B| = Número de instancias que cumplen B
|D| = Número de instancias en total
|A y B| = Número de instancias que cumplen A y B



El soporte es el número de casos del conjunto de datos que contienen la combinación X de elementos. Se formula como: $Soporte(X) = \frac{|X|}{D}$. Entonces, el soporte representa el número de veces que los antecedentes y consecuentes aparecen juntos en los datos, por tanto:

$$Soporte\ (A \rightarrow B) = \frac{|A\ y\ B|}{|D|} = Soporte\ (A\ y\ B).$$

La confianza también denominada probabilidad, representa la fracción de casos del conjunto de datos que contienen A y que también contienen B. Se formula como:

$$Confianza\ (A \rightarrow B) = \frac{|A|}{|B|} = \frac{Soporte\ (A)}{Soporte\ (B)}$$

La cobertura o soporte de antecedentes, es el número de instancias que la regla predice correctamente, para expresarla en porcentaje se formula como:

$$Cobertura\ (A \rightarrow B) = \frac{|A|}{|D|} = Soporte\ (A)$$

El Interés (lift) expresa la proporción del soporte de la regla observado en |D| respecto del soporte teórico de ese conjunto dado el supuesto de independencia. Un valor de Interés = 1 indica que ese conjunto aparece una cantidad de veces acorde a lo esperado bajo condiciones de independencia. Por tanto:

$$Interés\ (A \rightarrow B) = \frac{Confianza\ (A \rightarrow B)}{Soporte\ (B)} = \frac{Soporte\ (A\ y\ B)}{Soporte\ (A) * Soporte\ (B)}$$

Tabla 11: Ilustración de las principales métricas de las reglas de asociación

Regla	Soporte	Confianza	Interés
$A \rightarrow D$	$\frac{2}{5}$	$\frac{2}{3}$	$\frac{10}{9}$
$C \rightarrow A$	$\frac{2}{5}$	$\frac{2}{4}$	$\frac{5}{6}$
$A \rightarrow C$	$\frac{2}{5}$	$\frac{2}{3}$	$\frac{5}{6}$
$(B\ y\ C) \rightarrow D$	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{5}{9}$

2.5.3.5. Análisis de componentes principales

El análisis de componentes principales, PCA por sus siglas en inglés, es una técnica para la reducción de la dimensionalidad. La idea es calcular un grupo de vectores a partir de los datos de entrada que sean representativos de todos los datos. Cualquier entrada nueva se puede representar ahora, simplemente como productos punto con respecto a los vectores representativos. El efecto es que la dimensionalidad de entrada se reduce drásticamente.

Una de las aplicaciones de PCA es la reducción de dimensionalidad (variables), perdiendo la menor cantidad de información (varianza) posible: cuando se cuenta con un gran número de variables cuantitativas posiblemente correlacionadas y que indiquen posibles redundancias, PCA permite reducirlas a un número menor de variables transformadas llamadas componentes principales que expliquen gran parte de la variabilidad en los datos. Cada dimensión o componente principal generada por PCA será una combinación lineal de las variables originales y serán además independientes o no correlacionadas entre sí. Las componentes principales generadas pueden utilizarse a su vez en métodos de aprendizaje supervisado.

Respecto de las tareas comunes de preparación de los datos para PCA considérese que, al trabajar con varianzas, el método PCA es altamente sensible a los valores atípicos, por lo que es recomendable determinar si los hay. En Orange Data Mining se ejecuta las siguientes tareas de preparación de los datos para este modelo no supervisado: (1) Codificar las características categóricas como numéricas, de modo predeterminado con la codificación en caliente. (2) Imputar los valores faltantes con los valores medios y (3) Normalizar variables. (Orange, 2015s).

Respecto de tales pasos, Orange permite seleccionar la normalización o no, pero como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1, aquellas variables cuya escala sea mayor dominarán al resto.

El análisis descriptivo permite hacer un primer acercamiento a la realidad objeto de estudio en virtud de la muestra transversal de datos académicos y socioeconómicos de los alumnos escolares, por ejemplo, la distribución por cursos, materias, géneros, tipos de familia, educación parental, entre otros.

De acuerdo con la revisión sistemática de la literatura que complementa esta tesis, a nivel del empleo de modelos descriptivos para el estudio de incidencias sobre el rendimiento académico escolar, se ha frecuentado a los clasificadores basados en reglas, por ser un tipo de algoritmo de aprendizaje automático que representa el conocimiento en una forma de reglas que favorecen la interpretabilidad de los resultados de la clasificación y que además son adecuados para el análisis de datos numéricos y categóricos. Algunos de los métodos no supervisados reportados como más utilizados en la literatura son A priori (Reglas de asociación) y el Análisis de Componentes Principales (Pincay-Ponce et al., 2023).

Capítulo 3:

3. Desarrollo

En el presente capítulo se estudió las incidencias de los factores socioeconómicos en el rendimiento escolar, con base en el Proceso Estándar Intersectorial para Minería de Datos, conocido como Metodología CRISP-DM, mismo que se estructura de seis fases iterativas y exploratorias:

1. Comprensión del aprovechamiento escolar
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue

Además, se incluyó tareas relacionadas con la mejora de la calidad de los modelos de aprendizaje automático, sugeridas en el Proceso Estándar Intersectorial para el desarrollo de aplicaciones de Aprendizaje Automático con Metodología de Garantía de Calidad o CRISP-ML (Q) propuesto recientemente por Studer y colaboradores (2021).

El análisis exploratorio de los datos obtenidos contemplado entre los objetivos de la tesis se desarrolló en el software Orange con el eventual apoyo en Scripts de Python. Orange es una plataforma abierta para el análisis, visualización y predicción con datos, que además cuenta con configuraciones bajo pago para la realización de tareas relacionadas al despliegue de modelos de aprendizaje automático (Demšar et al., 2013). En esta investigación no se ofrece una guía de uso del software Orange, pero se sugiere remitirse a la documentación oficial de cada Widget del software disponible desde este enlace https://orangedatamining.com/widget-catalog/.

3.1. Fase 1. Comprensión del aprovechamiento escolar

En esta Fase 1, que en CRISP-DM se denomina Comprensión del Negocio, se abordaron cuestiones inherentes a las escuelas y más en específico al aprovechamiento escolar. Se clarificó los objetivos de las escuelas, los objetivos del análisis de datos y se bosquejó la situación actual en cuanto a tratar problemáticas similares con soluciones de aprendizaje automático. Al final de la fase se presentó un plan de proyecto utilizando la información que se contiene en esta documentación.

3.1.1. Sobre las escuelas y el rendimiento académico

Según la denominada Ley Orgánica de Educación Intercultural de Ecuador, el Sistema Nacional de Educación tiene tres niveles: Inicial, Básica y Bachillerato y cada nivel cuenta con subniveles. Las escuelas pertenecen al nivel básico, dividido en Segundo, Tercero y Cuarto año básico que corresponden a la Educación Básica Elemental (EBE). Quinto, sexto y séptimo que corresponden a la Educación Básica Media (EBM) (Anchundia-Delgado et al., 2022).

Las escuelas del país almacenan regularmente datos en medios electrónicos como hojas de cálculo o bases de datos para eventuales procesamientos, así como en medios parcialmente manuscritos. Para el desarrollo de esta tesis se dispuso de datos de una muestra transversal con respecto del periodo lectivo del año 2019, tanto de calificaciones de los alumnos y datos de sus factores socioeconómicos.

Continuando con el escenario ecuatoriano, existen dos cronogramas, el de la Región Interandina o Sierra, que inicia en septiembre y termina en junio del siguiente año. El de las regiones Costa, Oriente y Galápagos, que inician en mayo y culminan en febrero del año siguiente. En ambos casos la duración de dichos periodos es de 200 días hábiles.

Cada periodo lectivo o año escolar, se divide en dos periodos de cinco meses o dos quimestres, que incluyen feriados y días de holgura por varias razones, cada quimestre tiene una duración ideal de 100 días hábiles. Cada Quimestre a su vez se divide en tres parciales de una duración idealmente similar pero que puede variar según diversos factores.

Cada Quimestre tiene un Examen que aporta al 20% de la calificación anual, el restante 80% es el promedio de calificaciones de los tres parciales que conforman cada quimestre. El promedio de los dos Quimestres determina la aprobación de los

alumnos, es el llamado promedio anual. Todas las calificaciones se corresponden a la escala mostrada en la **Tabla 1**.

En cada parcial se registra una valoración cualitativa del comportamiento del alumno, tal cual se observa en la **Tabla 12**. La calificación es determinada en las juntas de docentes con base en informes de un docente que hace las funciones de tutor. Además, se registra una calificación del comportamiento por cada quimestre, pero la misma no corresponde con el promedio de la de los parciales, sino a la del tercer parcial del quimestre actual, porque, según el Ministerio de Educación (2016), un estudiante que pudo haber demostrado un comportamiento poco satisfactorio en los primeros parciales, puede cambiar conforme transcurre su estancia de estudios. En todo caso, esta calificación cualitativa no condiciona la aprobación del estudiante.

Tabla 12: Escala de evaluación cualitativa del comportamiento estudiantil

Valoración		Descripción
Muy satisfactorio	A	Lidera el cumplimiento de los compromisos establecidos para la sana convivencia social.
Satisfactorio	B	Cumple con los compromisos establecidos para la sana convivencia social.
Poco satisfactorio	C	Falla ocasionalmente en el cumplimiento de los compromisos establecidos para la sana convivencia social.
Mejorable	D	Falla reiteradamente en el cumplimiento de los compromisos establecidos para la sana convivencia social.
Insatisfactorio	E	No cumple con los compromisos establecidos para la sana convivencia social.

Fuente: Ministerio de Educación de Ecuador (2016)

También, en cada parcial se desarrolla un denominado Proyecto Escolar, definido como un espacio de aprendizaje interactivo que busca desarrollar tanto las habilidades cognitivas, como las socioemocionales, es decir, contribuir al desarrollo integral del estudiante tal cual lo establece la Constitución de Ecuador (Ministerio de Educación del Ecuador, 2016) y se muestra en la **Tabla 13**.

Tabla 13: Escala de evaluación cualitativa de los proyectos escolares

Valoración		Descripción
Excelente	EX	Demuestra destacado desempeño en cada fase del desarrollo del proyecto escolar lo que constituye un excelente aporte a su formación integral.
Muy buena	MB	Demuestra fiabilidad en el desempeño para cada fase del desarrollo del proyecto escolar lo que constituye un aporte a su formación integral.
Buena	B	Demuestra un desempeño aceptable, en cada fase del desarrollo del proyecto escolar lo que contribuye parcialmente a su formación integral.
Regular	R	Demuestra dificultad en atender cada fase del desarrollo del proyecto escolar lo que contribuye escasamente a su formación integral.

Fuente: Ministerio de Educación de Ecuador (2016)

En principio, los datos correspondientes a las calificaciones estaban disponibles en hojas de cálculo que el personal designado por las escuelas descarga desde un sistema informático proporcionado por el Gobierno del País, estos datos habían sido subidos previamente a dicho sistema por personal de las propias escuelas. Por otra parte, los datos de condiciones socioeconómicas estaban al custodio del Departamento de Consejería Estudiantil (DECE), que es la instancia responsable de la atención integral de los estudiantes, en la mayoría de los casos se llenan de modo directo con un procesador de textos y en otros casos el DECE imprime las fichas y el psicólogo responsable las llena manualmente en una entrevista con el representante del alumno y ocasionalmente el alumno. Luego de preparar los datos, se los consolidó en un archivo CSV de 6808 filas y 88 columnas.

Toda vez que se ha comprendido lo elemental del contexto escolar conforme con el problema a abordar, se establecen los objetivos escolares.

3.1.2. Sobre los objetivos escolares

Las escuelas no habían realizado analítica de datos alguna, por lo que los datos descritos fungían de herramientas para que mediante su uso mejoren sus procesos relacionados con la calidad de la enseñanza y el aprendizaje. A partir de la detección de la incidencia de los factores socioeconómicos en el rendimiento escolar, en resumen, se buscó:

- Comprender los datos mediante un análisis exploratorio que genere evidencia cuantificable de la incidencia de los factores socioeconómicos sobre el rendimiento escolar.
- Buscar mediante análisis descriptivo, asociaciones o patrones de datos de los estudiantes, que describan el comportamiento de las instancias o ejemplos del conjunto de datos sin tener hipótesis predefinidas.
- Buscar mediante análisis predictivo, respuestas a la posibilidad de fracaso o aproximación al fracaso escolar a partir de datos de calificaciones y factores socioeconómicos, tanto al final del curso como durante el trayecto del curso y a nivel de cada materia.
- Analizar la incidencia de los factores socioeconómicos en el aprovechamiento académico a nivel escolar y de ese modo contribuir a su entendimiento y mejora, mediante la aplicación de modelos de análisis de datos.

Una vez establecidos los objetivos escolares, se procedió a describir el abordaje de la problemática en la actualidad, conforme con la literatura existente.

3.1.3. Sobre la situación actual

La situación actual respecto de las principales técnicas o algoritmos de análisis de datos aplicados para el estudio de factores socioeconómicos y su incidencia en la educación escolar fue determinada mediante una revisión sistemática de la literatura ejecutada en las bases de datos Scopus y Web of Science, que se corresponde con el periodo 2012 – 2022 (Pincay-Ponce et al., 2023). La revisión, tal cual muestra la **Tabla 14** proporcionó información de las diferentes técnicas de minería de datos y de estadística aplicadas con frecuencia en este contexto.

Tabla 14: Técnicas y métodos de análisis de datos empleados para detectar incidencias de factores socioeconómicos en el aprovechamiento escolar

Estudio	Técnica de análisis de datos	Principal factor socioeconómico
1 - 2012 Academic performance of Peruvian elementary school children: The case of schools in Lima at the 6th grade	Análisis de componentes principales, PCA	Género Edad de los alumnos
	Análisis de potencia	
	Correlaciones bivariadas	
	Regresión lineal jerárquica	
2 - 2014 Parental involvement in homework: Relations with parent and student achievement-related motivational beliefs and achievement	Correlaciones bivariadas	Participación de los padres
	Prueba de Chi Cuadrado	
3 - 2015 Effortful control and early academic achievement of Chinese American children in immigrant families	Correlaciones bivariadas	Cultura de los padres Padres autoritarios
	Modelo de ecuaciones estructurales	
4 - 2016 A hybrid method based on MLFS approach to analyze students' academic achievement	Bosques aleatorios	Edad de los padres Economía
	Correlaciones bivariadas	
	Máquinas de vectores de soporte, SVM	
	Perceptron multicapa, ANN	
	Redes de base radial, ANN	
5 - 2017 Prediction models of learning strategies and learning achievement for lifelong learning	Árboles de decisión (C4.5)	Entorno de aprendizaje Políticas de gobierno Materias
	Correlaciones bivariadas	
	Reglas de asociación (Apriori)	
6 - 2017 Home learning environment and development of child competencies from kindergarten until the end of elementary school	Correlaciones bivariadas	Género Entorno de aprendizaje hogareño
	Modelo de ecuaciones estructurales	
	Regresión logística multinomial	

Estudio	Técnica de análisis de datos	Principal factor socioeconómico
7 - 2017 Childhood Social Skills as Predictors of Middle School Academic Adjustment	Correlaciones bivariadas	Entorno de aprendizaje hogareño Habilidades sociales Economía
	Regresión lineal jerárquica	
8 - 2017 An Appraisal Model Based on a Synthetic Feature Selection Approach for Students' Academic Achievement	Análisis discriminante	Escolaridad de los padres
	Bosques aleatorios	
	Correlaciones bivariadas	
	Máquinas de vectores de soporte, SVM	
	Perceptron multicapa, ANN	
	Red de correlación en cascada, ANN	
	Redes de base radial, ANN	
9 - 2018 Effortful control is associated with children's school functioning via learning-related behaviors	Correlaciones bivariadas	Habilidades sociales Economía
	Modelo de ecuaciones estructurales	
10 - 2018 Effects of early childhood education attendance on achievement, social skills, behaviour, and stress	Correlaciones bivariadas	Género Economía
	Prueba de Chi Cuadrado	
	Regresión lineal jerárquica	
11 -2018 Home Visiting Among Inner-City Families: Links to Early Academic Achievement	Correlaciones bivariadas	Participación de los padres
	Modelo de ecuaciones estructurales	
12 - 2019 Skin color and academic achievement in young, Latino children: Impacts across gender and ethnic group.	Modelo de ecuaciones estructurales	Raza
	Regresión lineal jerárquica	
13 - 2019 The Many (Subtle) Ways Parents Game the System: Mixed-method Evidence on the Transition into Secondary-school Tracks in Germany	Correlaciones bivariadas	Economía
	Regresión lineal simple	
	Regresión logística multinomial	
14 - 2020 Longitudinal Associations Linking Elementary and Middle School Contexts with Student Aggression in Early Adolescence	Regresión lineal jerárquica	Escolaridad de los padres
15 - 2020 Impacts of School Racial Composition on the Mathematics	Análisis de componentes principales, PCA	Género Raza
	Análisis de la varianza, ANOVA	

Estudio	Técnica de análisis de datos	Principal factor socioeconómico
and Reading Achievement Gap in Post Unitary Charlotte-Mecklenburg Schools	Correlaciones bivariadas	
16 - 2021 [T] Parental Self-Efficacy in Helping Children Succeed in School Favors Math Achievement	Correlaciones bivariadas Modelo de rutas	Participación de los padres
17 - 2021 Predicting Students' Mathematics Achievement Through Elementary and Middle School: The Contribution of State-Funded Prekindergarten Program Participation	Correlaciones bivariadas Prueba de Chi Cuadrado	Economía
18 - 2022 The Pathway to Enrolling in a High-Performance High School: Understanding Barriers to Access	Correlaciones bivariadas Regresión logística simple	Economía Distancia casa - escuela
19 - 2022 Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar	Árboles de decisión (C4.5) Naïve Bayes Reglas de asociación (Apriori) Reglas de clasificación (CN2)	Economía Escolaridad de los padres

Fuente: Elaboración propia a partir de Pincay-Ponce y colaboradores (2023)

Entre enero de 2012 y septiembre de 2022, el rendimiento académico escolar contemplado como un objeto de estudio multifactorial, ha sido analizado con modelos surgidos en las matemáticas, estadística y en años recientes con modelos de aprendizaje automático. Se ha perseguido identificar dependencias, independencias y poder de predicción de las características o datos disponibles de los alumnos. En este sentido las técnicas más recurridas en este tiempo fueron las ecuaciones estructurales, los análisis correlacionales, los análisis factoriales como en el caso del Análisis de Componentes Principales (PCA) y varios algoritmos relacionados con el aprendizaje automático, más en concreto la minería de datos, tales como modelos de regresión, redes neuronales, bosques aleatorios, clasificadores como árboles de decisión, reglas de asociación y Naïve Bayes (Pincay-Ponce et al., 2023).

Entre los aspectos socioeconómicos más reportados en los estudios listados en la **Tabla 14**, figuran la economía, género y habilidades sociales de los niños y otros más dependientes de los padres como su escolaridad, ambiente en el hogar, cultura

y autoritarismo, sin embargo, en la misma revisión, que se desarrolló como complemento de esta tesis, se encontró como limitante que apenas 4 de 15 estudios primarios reportaron más de mil sujetos en su población estudiada, en algunos casos combinando padres e hijos (Pincay-Ponce et al., 2023).

Como se ha indicado en secciones precedentes, para fines de esta tesis se contó con un conjunto de datos de una muestra transversal correspondiente al periodo lectivo 2019, tanto de calificaciones como de factores socioeconómicos.

Conocida la situación actual y la forma de abordar la problemática conforme con la literatura, se procedió a establecer los objetivos del análisis de datos.

3.1.4. Sobre los objetivos de análisis de datos

A partir de los objetivos escolares se definieron los siguientes objetivos de análisis de datos:

- Realizar un análisis exploratorio que ayude en la identificación de dependencias, independencias y poder de predicción de las características o datos de calificaciones y socioeconómicos disponibles de los alumnos.
- Creación de modelos utilizando datos de calificaciones y socioeconómicos de utilidad en la predicción de las posibilidades fracaso, abandono o bajo rendimiento académico. En este caso se valoran la exactitud y precisión alcanzada por los modelos, así como las especificaciones de sus parámetros e hiperparámetros.
- Creación de modelos que encuentren asociaciones o patrones en datos de calificaciones y socioeconómicos disponibles de los estudiantes sin tener hipótesis predefinidas.
- Asignación de un rango a cada estudiante basado en las posibilidades de fracaso, abandono o bajo rendimiento académico.

Luego de definir estos objetivos, se procedió a planificar el modelado de los datos.

3.1.5. Sobre planificación del modelado de datos

En función de los tres tipos de análisis recurridos en la tesis, en la **Tabla 15** se muestran las herramientas de software libre empleadas. En lo concerniente a la

preparación de los datos se empleó: Microsoft SQL Server 2019 Developer, Microsoft Excel de Office 365 y de nuevo Orange Data Mining 3.34.

Tabla 15: Tipos de análisis de datos y herramientas de software empleadas en la tesis

Tipo de Análisis	Orange Data Mining 3.34 Python Script	SPSS Amos 24.0.0.
Exploratorio	√	√
Predictivo	√	
Descriptivo	√	

La **Tabla 16** muestra la planificación básica para la construcción de modelos de análisis de datos cuya finalidad es informar a todos los usuarios relacionados con los objetivos, recursos, riesgos del proyecto, así como la estimación de la duración de las fases de análisis de datos.

Tabla 16: Planificación básica de la construcción de modelos de análisis de datos

Fase	Semanas	Recursos	Riesgos
Comprensión del negocio	2	Todos los analistas (Escuela, Doctorando)	Cambio económico
Comprensión de los datos	4	Todos los analistas (Escuela, Doctorando)	Problemas de datos, Problemas tecnológicos
Preparación de los datos	10	Doctorando Tiempo de análisis de base de datos	Problemas de datos, Problemas tecnológicos
Modelado	6	Doctorando Tiempo de análisis de base de datos	Problemas de tecnología, incapacidad para encontrar un modelo adecuado
Evaluación	2	Todos los analistas (Escuela, Doctorando)	Cambio económico, incapacidad para implementar resultados
Despliegue	1	Doctorando Tiempo de análisis de base de datos	Cambio económico, incapacidad para implementar resultados

3.2. Fase 2. Comprensión de los datos

Para el cumplimiento de la segunda fase fue necesario realizar tres actividades. La primera consistió en realizar una ficha técnica de las características (Ver **Tabla 17**), de cada una se detalló un identificador, nombre de la variable, descripción y tipo =Texto, 123 = Número. La segunda actividad, fue acceder a los datos y explorarlos con la ayuda de tablas, gráficos y varias técnicas estadísticas que ayuden a determinar su calidad. Además, de ofrecer diversas agrupaciones y resúmenes procurando no sacrificar información importante. La tercera actividad fue la identificación de relaciones, tendencias, anomalías y poder de predicción de los datos.

3.2.1. Recopilación inicial de datos

Las Unidades Educativas emplean diversos orígenes de datos de utilidad para el desarrollo de un proyecto de análisis de datos como el que contempla esta tesis:

- **Sistema de calificaciones.** – Es proporcionado por el estado para el registro de notas finales y datos de promoción o no de los estudiantes. De este sistema web se han descargado las calificaciones empleadas en el presente trabajo.
- **Fichas Socioeconómicas.** – Son elaboradas por el Departamento de Consejería Estudiantil (DECE), en la mayoría de los casos se llenan de modo directo con un procesador de textos y en otros casos el DECE imprime las fichas y el psicólogo responsable las llena manualmente en una entrevista con el representante del alumno y ocasionalmente el alumno. Este proceso se hace al iniciar el año lectivo. De las fichas se han obtenido datos de los factores socioeconómicos de los alumnos.

Una vez obtenidos los datos precedentes se siguió los siguientes pasos:

1. Se elaboró un libro de Microsoft Excel con datos de los factores socioeconómicos.
2. Se elaboró un libro de Microsoft Excel con las calificaciones de los alumnos.
3. Se creó una Base de Datos en Microsoft SQL Server 2019 con dos tablas que se corresponden con los dos libros mencionados, se las relacionó mediante la cédula de identidad de los alumnos, misma que en el paso 5, se anonimizó para cumplir con el acuerdo de confidencialidad apropiado a estos casos.
4. En Microsoft SQL Server 2019, se construyó una vista que relacione y combine ambos tipos de datos disponibles ahora en tablas, en formato tabular. Esta vista resultante se estructuró de 6808 filas y 61 columnas y se guardó en un archivo de formato CSV. Es de indicar que de cada alumno se registran siete filas de calificaciones que se corresponden con la cantidad de materias que ellos estudian en cada año básico. Luego de completar las tareas de preparación de datos, el conjunto de datos se consolidó en 6808 filas y 88 columnas.
5. Se importó el archivo CSV creado desde Orange Data Mining 3.34, para iniciar la preparación de los datos, la evaluación de su calidad y la concerniente construcción de modelos.
6. Luego de la recopilación inicial de datos se realizó la descripción del conjunto de datos.

3.2.2. Descripción del conjunto de datos

En la siguiente tabla se describe el conjunto de datos inicial compuesto de 6808 filas y 61 columnas, en términos de su denominación, breve descripción y tipo.

Tabla 17: Ficha de las características predictoras y etiquetas de clase con las que se inicia el análisis

N°	Atributo	Descripción	N°	Atributo	Descripción
1	Curso	123 Nivel del estudiante	35	PQP1	123 Nota de 1er quimestre parcial 1
2	ApellidosNombres	Apellidos y Nombres	36	PQP2	123 Nota de 1er quimestre parcial 2
3	EstadoCivilMadre	Estado civil de la Madre	37	PQP3	123 Nota de 1er quimestre parcial 3
4	EscolaridadMadre	Nivel de estudio de la Madre	38	PROM1	123 Promedio de tres parciales, 1er quimestre sobre 10
5	Ocup. Madre	Profesión u Ocupación de la Madre	39	1PRO80	123 Promedio 1er quimestre sobre 80%
6	EstadoCivilPadre	Estado civil del Padre	40	1EXA20	123 Examen 1er quimestre sobre 20%
7	EscolaridadPadre	Escolaridad máxima del Padre	41	EXA1	123 Examen 1er quimestre sobre 10
8	Ocup. Padre	Profesión u Ocupación del Padre	42	QUI1	123 Nota 1er quimestre
9	ParentescoRepresentante	Parentesco del Representante	43	SQP1	123 Nota de 2do quimestre parcial 1
10	EscolaridadRepresentante	Escolaridad máxima del Representante	44	SQP2	123 Nota de 2do quimestre parcial 2
11	Ocup. Representante	Profesión u Ocupación de la Madre	45	SQP3	123 Nota de 2do quimestre parcial 3
12	Dirección	Lugar donde vive	46	PROM2	123 Promedio de tres parciales, 2do quimestre sobre 10
13	NumeroHermanos	123 Número de hermanos	47	2PRO80	123 Promedio 2do quimestre sobre 80%
14	EstructuraFamiliar	Estructura familiar	48	2EXA20	123 Examen 2do quimestre sobre 20%
15	IngresoMensual	Ingreso mensual del núcleo familiar	49	EXA2	123 Examen 2do quimestre sobre 10
16	LuzElectrica	Posee luz eléctrica (Si/No)	50	QUI2	123 Nota 2do Quimestre
17	AguaPotable	Posee agua potable (Si/No)	51	PromAnual	**123 Promedio Anual (etiqueta de clase)**
18	Telefono	Posee teléfono fijo (Si/No)	52	ComportamientoPQP1	Comportamiento 1er quimestre parcial 1

N°	Atributo	Descripción	N°	Atributo	Descripción
19	Alcantarillado	Posee Alcantarillado (Si/No)	53	ComportamientoPQP2	Comportamiento 1er quimestre parcial 2
20	Internet	Posee Internet (Si/No)	54	ComportamientoPQP3	Comportamiento 1er quimestre parcial 3
21	TVCable	Posee TV Cable (Si/No)	55	ComportamientoPQ	Comportamiento 1er quimestre
22	Celular	Posee celular (Si/No)	56	ComportamientoSQP1	Comportamiento 2do quimestre parcial 1
23	Computador	Posee computador (Si/No)	57	ComportamientoSQP2	Comportamiento 2do quimestre parcial 2
24	Discapacidad	Discapacidad del estudiante	58	ComportamientoSQP3	Comportamiento 2do quimestre parcial 3
25	FechaIngreso	Año de ingreso a escuela	59	ProyEscPQP1	Comportamiento 2do quimestre
26	MateriaDificultosa	Materia dificultosa del estudiante	60	ComportamientoSQ	Proyecto escolar 1er quimestre parcial 1
27	ProcedeDeOtraInstitucion	Procedente de otra institución (Si/No)	61	ProyEscPQP2	Proyecto escolar 1er quimestre parcial 2
28	Repetidor	Repetidor de año básico (Si/No)	62	ProyEscPQP3	Proyecto escolar 1er quimestre parcial 3
29	Enfermedad	Enfermedad del estudiante	63	ProyEscPQ	Proyecto escolar 1er quimestre
30	Grado	Grado estudiantil	64	ProyEscSQP1	Proyecto escolar 2do quimestre parcial 1
31	Paralelo	Paralelo del estudiante	65	ProyEscSQP2	Proyecto escolar 2do quimestre parcial 2
32	Jornada	Jornada de estudios	66	ProyEscSQP3	Proyecto escolar 2do quimestre parcial 3
33	Nombres	Apellidos y nombres	67	ProyEscSQ	Proyecto escolar 2do quimestre
34	Materia	Materia			

En relación con el análisis de la incidencia de los factores socioeconómicos son de especial relevancia los siguientes datos: (1) Estado civil del padre, (2) Estado civil de la madre, (3) Estado civil del representante, (4) Escolaridad del padre, (5) Escolaridad de la madre, (6) Escolaridad del representante, (7) Ocupación del padre, (8) Ocupación de la madre, (9) Ocupación del representante, (10) Parentesco del representante, (11) Número de hermanos, (12) Estructura familiar, (13) Ingreso mensual en el hogar, (14) servicio de luz eléctrica regularizada, (15) Servicio de agua potable regularizado, (16) Servicio de alcantarillado regularizado, (17) Servicio de internet regularizado, (18) Servicio de televisión por cable regularizado, (19) Teléfono celular, (20) Computador, (21) Discapacidad, (22) Año de la fecha de ingreso, (23) Materia dificultosa, (24) Procedencia desde otra institución, (25) Repetidor de año

básico y (26) Enfermedad auto reportada.

En relación con el análisis de la incidencia de los factores socioeconómicos, en principio no parecen relevantes los siguientes datos y ciertas calificaciones:

1. Curso, porque concatena el año básico con el paralelo del alumno, ejemplo 2C. En lugar de este dato se utiliza sólo el año básico.
2. Disponibilidad de teléfono convencional.
3. Grado porque es equivalente al año básico.
4. Paralelo porque para tal efecto la institución solo busca que se equilibren por género al número de alumnos por aula.
5. Jornada porque todos estudian en las mañanas.
6. Nombres porque se deben anonimizar.
7. Calificación de exámenes quimestrales sobre 2 puntos, porque también se dispone de la calificación sobre 10 puntos, tal cual se registran las demás calificaciones.

Si bien, en principio se dispone de una buena cantidad de atributos, para la presente investigación se construyeron nuevos atributos y esta actividad se documentó en la Fase 3 de este capítulo, dónde también se muestra el perfil de los datos una vez preparados.

3.2.3. Exploración de datos

La exploración implica buscar patrones, conexiones y relaciones entre los datos y expresarlas en gráficas o estadística de ayuda para comprender el tipo de información que se ha recopilado, en combinación con la información adquirida durante la definición de los objetivos del negocio y de minería. En los últimos años, la visualización de datos se ha convertido en una disciplina en sí misma, también se ha acompañado del desarrollo de numerosas herramientas de software que le soportan (Nelli, 2018).

Para fines de obtener visualizaciones más comprensivas en las siguientes tablas y gráficos, se han expresado en texto el año básico cursado por los alumnos, se han abreviado los nombres de las materias y se ha especificado el género de cada alumno (hombre, mujer) mismo que se ha deducido del nombre de cada uno. De modo progresivo e iterativo junto con la Fase 3 de Preparación de Datos, se generó nuevas características que amplíen las posibilidades de exploración de los datos. En los casos pertinentes se ofrece una justificación bibliográfica.

2.3.3.1. Con base en la cantidad de alumnos

La **Tabla 18**, que muestra la distribución de cantidad de calificaciones de los alumnos por cada año básico y cada materia, es importante porque el análisis de datos de esta investigación se centra en las calificaciones de cada materia, debido a que, en el sistema educativo escolar ecuatoriano, si un alumno reprueba una materia entonces reprueba el año básico en curso.

Tabla 18: Distribución de cantidad de calificaciones de los alumnos por año básico y materia.

Año básico	Arte	CC NN	EE SS	Ed. F	Inglés	Lenguaje	Matemática	Total
Segundo	204	204	204	204	204	204	204	**1428**
Tercero	184	184	184	184	184	184	184	**1288**
Cuarto	80	151	151	151	151	151	151	**986**
Quinto	153	152	152	152	152	153	153	**1067**
Sexto	181	182	182	181	181	181	181	**1269**
Séptimo	110	110	110	110	110	110	110	**770**
Totales	**912**	**983**	**983**	**982**	**982**	**983**	**983**	**6808**

Arte, cultura: Educación cultural y artística
CC NN: Ciencias naturales
EE SS: Estudios sociales
Ed. F: Educación física
Lenguaje: Lengua y literatura

La **Tabla 19** y la **Tabla 20**, en cambio muestran datos a nivel de la cantidad de alumnos analizados, en la 20 se puede observar que al menos 159 alumnos tienen retrasos en la completitud de sus estudios en curso.

Tabla 19: Distribución de cantidad de alumnos por año básico y género.

Año básico	Hombre	Mujer	Total
2. AB	58	61	**119**
3. AB	52	47	**99**
4. AB	38	43	**81**
5. AB	43	35	**78**
6. AB	49	43	**92**
7. AB	67	42	**109**
Totales	**307**	**271**	**578**

Tabla 20: Distribución de cantidad de alumnos por años de retraso en sus estudios.

Años de retraso	Hombre	Mujer	Total
3	46	50	**96**
4	20	25	**45**
5	7	6	**13**
6	2	2	**4**
7	0	1	**1**
Totales	**75**	**84**	**159**

Como complemento a la **Tabla 20**, en la **Tabla 21** se evidencia que al menos 67 alumnos ingresaron antes de 2014, por tanto, al 2019 han tenido retrasos para culminar estudios con normalidad, es decir, en el periodo de 6 años.

Tabla 21: Distribución de cantidad de alumnos según años de llegada

Año	Hombre	Mujer	Total	Frecuencia Acumulativa
2008	0	1	1	**1**
2011	1	4	5	**6**
2012	16	10	26	**32**
2013	20	15	35	**67**
2014	43	49	92	
2015	48	41	89	
2016	43	53	96	
2017	25	14	39	
2018	36	39	75	
2019	75	45	120	

Extendiendo la información de las dos tablas precedentes, en la **Tabla 22** se muestra la situación de fin de curso más actual de los 159 alumnos denotados en la **Tabla 20,** cómo era de esperarse su situación respecto del promedio anual obtenido mejoró en el año básico cursado más recientemente.

Tabla 22: Distribución de cantidad de alumnos con retrasos en complementar sus estudios, ordenados por género y nivel de logro anual obtenido

Tipo de promedio obtenido	Hombre	Mujer	Total	%
1. DAR - Domina los aprendizajes requeridos	37	45	82	51.6%
2. AAR - Alcanza los aprendizajes requeridos	37	38	75	47.2%
3. PAAR - Próximo a alcanzar los aprendizajes requeridos	1	1	2	1.2%
Total	**75**	**84**	**159**	**100%**

9 hasta 10: 1. DAR **7 hasta 08.99:** 2. AAR **4.01 hasta 6.99:** 3. PAAR

2.3.3.2. Con base en los registros de notas de cada materia

En esta sección se ofrecen una serie de segmentaciones como lo son las de materias con calificaciones en riesgo, materias con dificultad auto reportada. Además, de análisis a partir del año básico en cuestión, tipos de familia, niveles de ingreso, género y otras que se muestran en las siguientes páginas.

Tabla 23: Distribución de dificultades auto reportadas en las asignaturas, según año básico, género y materia

Resumen gráfico	Año básico	Género	Materia	Cantidad
	2. Segundo	Hombre	EE SS	2

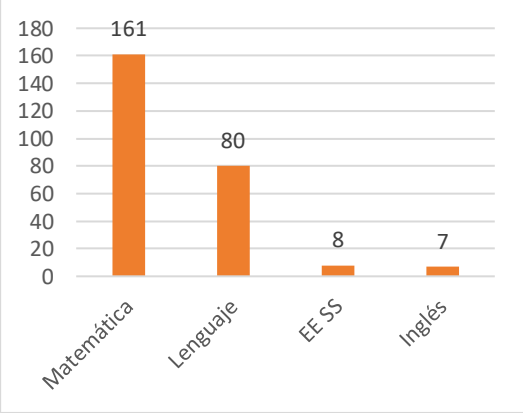


Gráfico 2: Dificultades auto reportadas en total

		Lenguaje	14
		Matemática	4
	Mujer	Lenguaje	14
		Matemática	8
3. Tercero	Hombre	Lenguaje	12
		Matemática	8
	Mujer	Lenguaje	6
		Matemática	12
4. Cuarto	Hombre	Inglés	2
		Lenguaje	6
		Matemática	10
	Mujer	EE SS	2
		Lenguaje	2
		Matemática	20
5. Quinto	Hombre	Lenguaje	8
		Matemática	18
	Mujer	Inglés	2
		Lenguaje	6
		Matemática	16
6. Sexto	Hombre	EE SS	2
		Inglés	2
		Lenguaje	10
		Matemática	30
	Mujer	EE SS	2
		Inglés	1
		Matemática	35
7. Séptimo	Mujer	Lenguaje	2
		Total	**256**

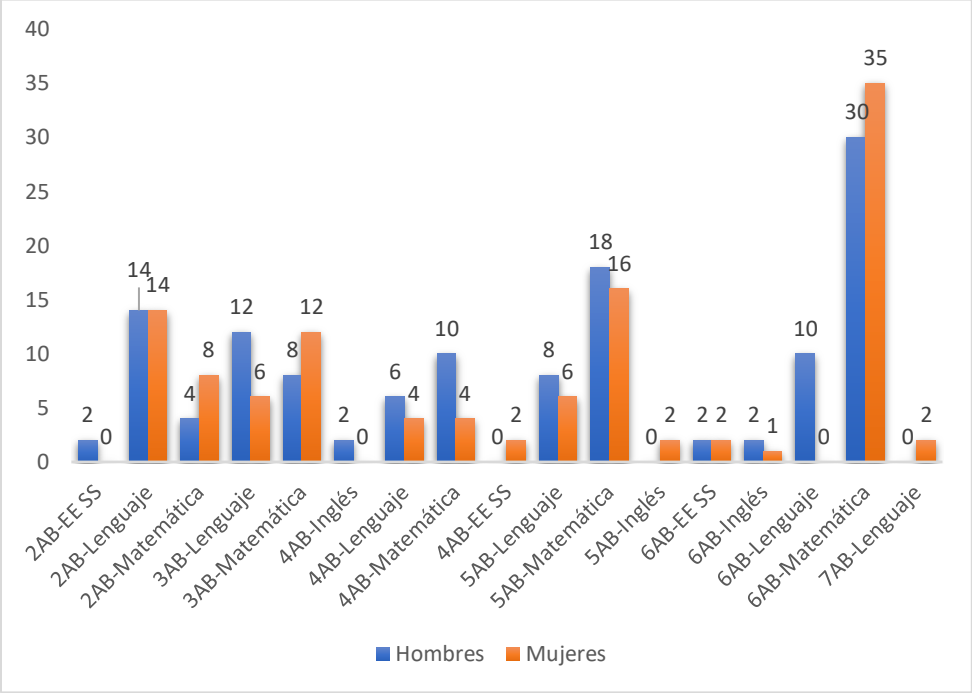


Gráfico 3: Dificultades auto reportadas por años básicos, materias y género

Tabla 24: Distribución de dificultades auto reportadas en las asignaturas, según género, año básico, materia e ingresos familiares

Sueldos Básicos	Año Básico (Género)	Materia	Auto reportes	Porcentaje
Menos de 1 SBU	2AB (H)	Lenguaje	2	**4.20**
	3AB (H)	Matemática	2	
	4AB (H)	Matemática	2	
	5AB (M)	Matemática	2	
	6AB (M)	Matemática	6	
1.0 SBU	2AB (H)	EE SS	2	**54.65**
	2AB (H)	Lenguaje	6	
	2AB (M)	Lenguaje	4	
	2AB (H)	Matemática	4	
	2AB (M)	Matemática	4	
	3AB (H)	Arte, Cultura	1	
	3AB (M)	Arte, Cultura	1	
	3AB (H)	CC NN	1	
	3AB (M)	CC NN	1	
	3AB (H)	EE SS	1	
	3AB (M)	EE SS	1	
	3AB (H)	Ed. F	1	

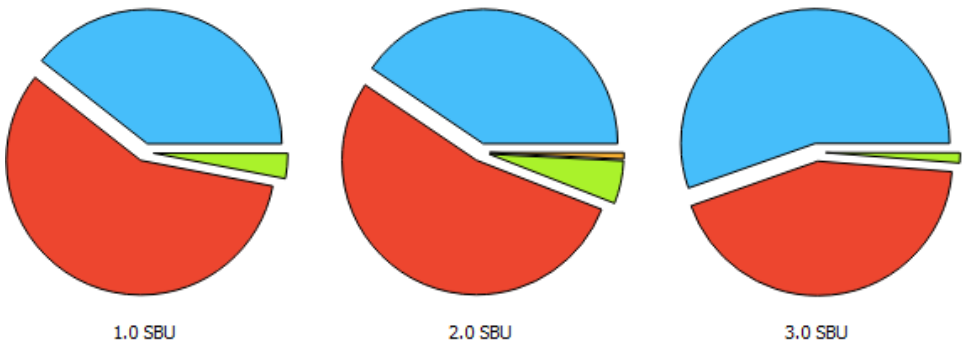
Sueldos Básicos	Año Básico (Género)	Materia	Auto reportes	Porcentaje
	3AB (M)	Ed. F	1	
	3AB (H)	Inglés	1	
	3AB (M)	Inglés	1	
	3AB (H)	Lenguaje	7	
	3AB (M)	Lenguaje	5	
	3AB (H)	Matemática	5	
	3AB (M)	Matemática	7	
	4AB (H)	Arte, Cultura	2	
	4AB (H)	CC NN	2	
	4AB (H)	EE SS	2	
	4AB (H)	Ed. F	2	
	4AB (H)	Inglés	4	
	4AB (H)	Lenguaje	8	
	4AB (M)	Lenguaje	2	
	4AB (H)	Matemática	10	
	4AB (M)	Matemática	18	
	5AB (M)	Arte, Cultura	1	
	5AB (M)	CC NN	1	
	5AB (M)	EE SS	1	
	5AB (M)	Ed. F	1	
	5AB (M)	Inglés	3	
	5AB (H)	Lenguaje	6	
	5AB (M)	Lenguaje	3	
	5AB (H)	Matemática	14	
	5AB (M)	Matemática	9	
	6AB (H)	Arte, Cultura	1	
	6AB (H)	CC NN	1	
	6AB (H)	EE SS	1	
	6AB (M)	EE SS	2	
	6AB (H)	Ed. F	1	
	6AB (H)	Inglés	3	
	6AB (M)	Inglés	1	
	6AB (H)	Lenguaje	5	
	6AB (H)	Matemática	5	
	6AB (M)	Matemática	19	
	2AB (H)	Lenguaje	4	
2.0 SBU	2AB (M)	Lenguaje	8	**31.53**
	2AB (M)	Matemática	2	

Sueldos Básicos	Año Básico (Género)	Materia	Auto reportes	Porcentaje
	3AB (H)	Arte, Cultura	1	
	3AB (M)	Arte, Cultura	2	
	3AB (H)	CC NN	1	
	3AB (M)	CC NN	2	
	3AB (H)	EE SS	1	
	3AB (M)	EE SS	2	
	3AB (H)	Ed. F	1	
	3AB (M)	Ed. F	2	
	3AB (H)	Inglés	1	
	3AB (M)	Inglés	2	
	3AB (H)	Lenguaje	1	
	3AB (M)	Lenguaje	4	
	3AB (H)	Matemática	3	
	3AB (M)	Matemática	6	
	4AB (M)	Arte, Cultura	1	
	4AB (M)	CC NN	1	
	4AB (M)	EE SS	3	
	4AB (M)	Ed. F	1	
	4AB (M)	Inglés	1	
	4AB (M)	Lenguaje	1	
	4AB (M)	Matemática	3	
	5AB (H)	Lenguaje	2	
	5AB (M)	Lenguaje	4	
	5AB (H)	Matemática	4	
	5AB (M)	Matemática	4	
	6AB (H)	Arte, Cultura	1	
	6AB (H)	CC NN	1	
	6AB (H)	EE SS	1	
	6AB (H)	Ed. F	1	
	6AB (H)	Inglés	1	
	6AB (H)	Lenguaje	5	
	6AB (H)	Matemática	17	
	6AB (M)	Matemática	8	
	7AB (M)	Lenguaje	2	
3.0 SBU	3AB (H)	Lenguaje	4	**3.60**
	5AB (M)	Matemática	2	
	6AB (H)	EE SS	2	
	6AB (H)	Lenguaje	2	

Sueldos Básicos	Año Básico (Género)	Materia	Auto reportes	Porcentaje
	6AB (H)	Matemática	2	
4.0 SBU	2AB (M)	Matemática	2	**3.00**
	6AB (H)	Matemática	6	
	6AB (M)	Matemática	2	
5.0 SBU	2AB (H)	Lenguaje	2	**1.80**
	2AB (M)	Lenguaje	2	
	3AB (M)	Matemática	2	
8.0 SBU	3AB (H)	Lenguaje	2	**1.20**
	6AB (H)	Matemática	2	
		Totales	**333**	**100.00**

En la **Tabla 24** se observa que el 85% de auto reportes de dificultad estan asociados con alumnos cuyas familias reportan 0, 1 o 2 sueldos básicos como ingresos. Para contrastar con la dificultad auto reportada, en el **Gráfico 4** generado en Orange Data Mining se resume la distribución de promedios obtenidos por los alumnos según sueldos básicos unificados reportados como ingresos familiares. Se puede notar que existe similitud en las proporciones de casos para ambos análisis. Los subsiguientes gráficos de pastel también se centran en los promedios anuales obtenidos, observados desde diferentes perspectivas.

En el **Gráfico 5** se ilustra que las niñas obtuvieron calificaciones de la clase <4 No alcanza los aprendizajes requeridos> lo que representa un alto riesgo para su rendimiento académico, en tanto que la proporción de calificaciones de la clase <3 Próximo a alcanzar los aprendizajes requeridos> que representa un riesgo moderado fue similar entre niños y niñas.



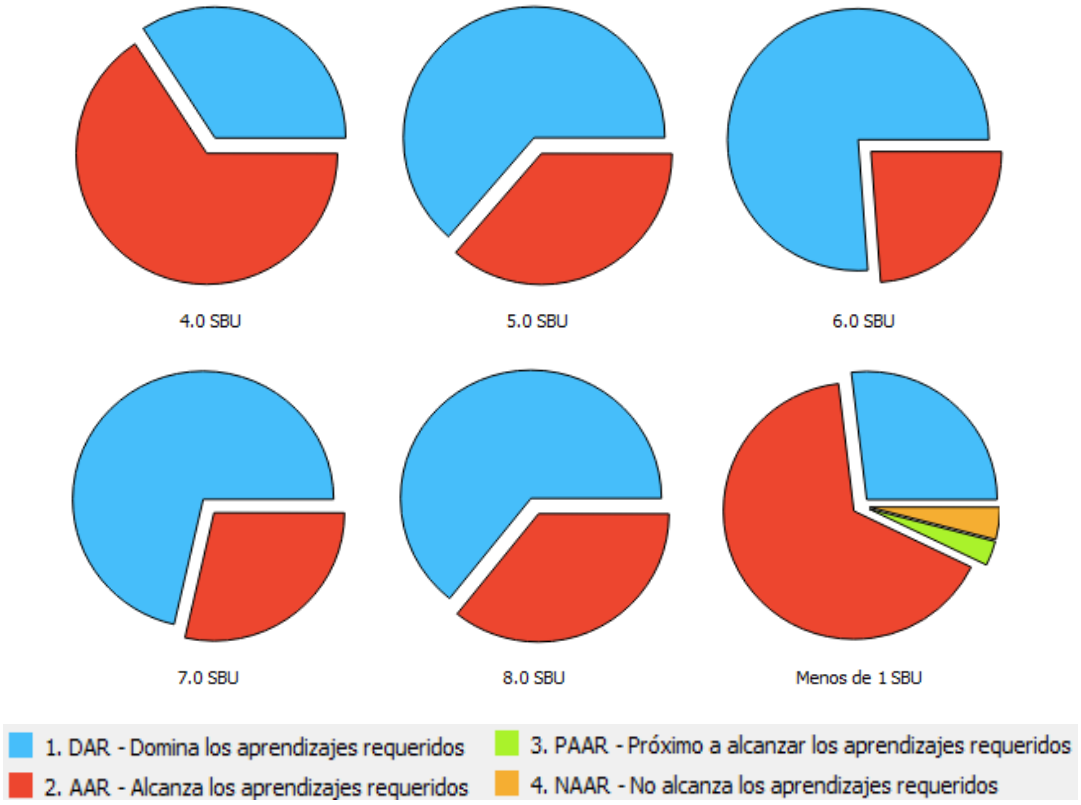


Gráfico 4: Distribución de promedios obtenidos por los alumnos según sueldos básicos unificados reportados como ingresos familiares

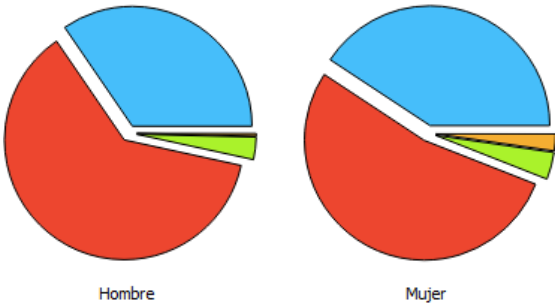


Gráfico 5: Distribución de promedios obtenidos por los alumnos según su género

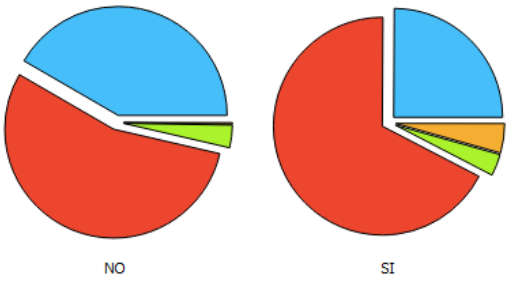
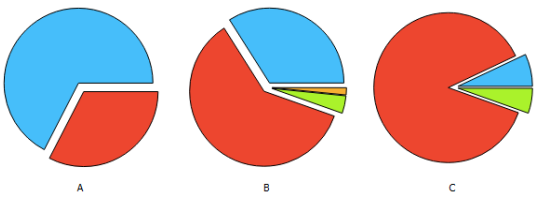


Gráfico 6: Distribución de promedios obtenidos por los alumnos según discapacidad autoreportada por sus representantes



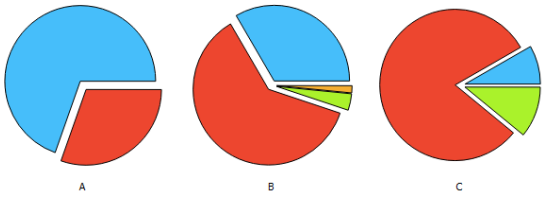


Gráfico 7: Distribución de promedios obtenidos por los alumnos según su comportamiento en el Parcial I, Quimestre 1 (A, B, C)

Gráfico 8: Distribución de promedios obtenidos por los alumnos según su comportamiento en el Parcial II, Quimestre 1 (A, B, C)

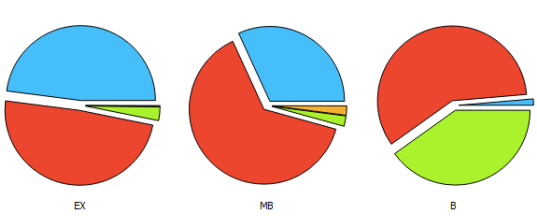


Gráfico 9: Distribución de promedios obtenidos por los alumnos según su calificación en proyectos escolares en el Parcial I, Quimestre 1 (EX, MB, B)

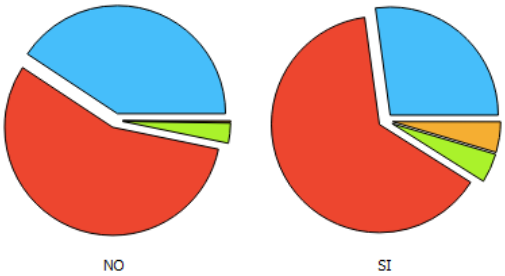


Gráfico 10: Distribución de promedios obtenidos por los alumnos según si su familia es reconstruida

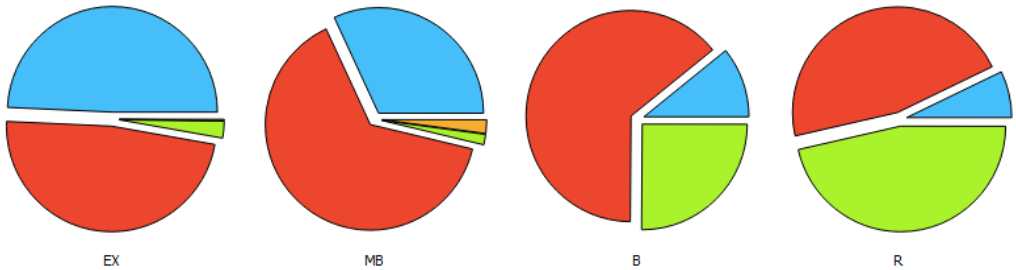


Gráfico 11: Distribución de promedios obtenidos por los alumnos según su calificación en proyectos escolares en el Parcial II, Quimestre 1 (EX, MB, B, R)

En el **Gráfico 6** se ilustra que los alumnos de los que se auto reportó alguna discapacidad obtuvieron una proporción de calificaciones de la clase <4 No alcanza los aprendizajes requeridos> en tanto que de los alumnos que se indicó ausencia de discapacidad no se registran calificaciones de esta clase que representa un alto riesgo para el rendimiento académico. La proporción de calificaciones de la clase <3 Próximo a alcanzar los aprendizajes requeridos> que representa un riesgo moderado es similar entre ambos casos. En el **Gráfico 10** se observa una proporción que guarda similitud en el caso de alumnos que viven en familias reconstruidas y los que no, además, la información de este gráfico se amplía en la **Tabla 25:** Distribución de promedios obtenidos por alumnos con familias reconstruidas, agrupados por materias.

En el **Gráfico 7** se ilustra las proporciones de calificaciones de las clases <3 Próximo a alcanzar los aprendizajes requeridos> y <4 No alcanza los aprendizajes

requeridos> en los alumnos que registran comportamientos B y C en el primer parcial del Quimestre 1. Se reitera que los comportamientos en las escuelas ecuatorianas no se promedian a final de los quimestres, sino que en cada quimestre se registra la del parcial 3 o más actual. Es de esperar que las calificaciones del Parcial 1 y 2 del Quimestre 1, reflejen de modo más espontáneo a los comportamientos de los alumnos. En el **Gráfico 8** se observa que en el Parcial 2 del Quimestre 1 la situación empeoró levemente en aquellos alumnos con comportamientos de calificación C.

En el **Gráfico 9** y en el **Gráfico 11** se ilustra las proporciones de las calificaciones de los proyectos escolares en el Parcial 1 y 2 del Quimestre 1, si se considera que los proyectos escolares son espacios de aprendizaje interactivos que buscan desarrollar tanto las habilidades cognitivas, como las socioemocionales de los alumnos, entonces reviste de importancia valorar la relación entre calificaciones menores en los proyectos escolares con los promedios anuales.

En el **Gráfico 12** se ilustra las proporciones de las calificaciones correspondientes a cada clase para cada una de las materias, en lo que se puede ver leves diferencias con la proporción de dificultades auto reportadas, por ejemplo, existen alumnos con calificaciones de riesgo en Ciencias Naturales en mayor medida que las dificultades reportadas para dicha materia. Comparar con la **Tabla 22:** Distribución de cantidad de alumnos con retrasos en complementar sus estudios, ordenados por género y nivel de logro anual obtenido.

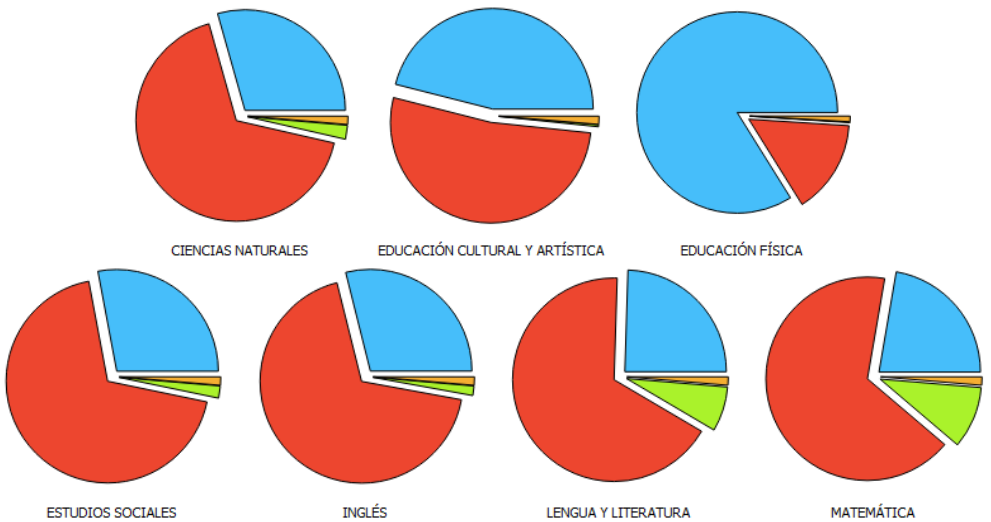


Gráfico 12: Distribución de promedios obtenidos por los alumnos según cada asignatura

Para complementar al **Gráfico 12** en el Diagrama de Venn del **Gráfico 13:** Proporción de coincidencias entre dificultades auto reportadas en las materias y la obtención de calificaciones PARA y NAAR que indican riesgos reales, se observa que 14 de 333 auto reportes de dificultad concuerdan con la obtención de calificaciones que suponen riesgos como lo son las de las clases <Próximo a alcanzar los aprendizajes requeridos> y <No alcanza los aprendizajes requeridos>. En sentido contrario 14 de 144 registros de calificaciones que representan riesgos coincidieron con el auto reporte respectivo y fue en las materias de Lenguaje y Matemática. Las barras del **Gráfico 14** ilustran acerca de los registros de calificaciones parciales equivalentes a riesgos pero que no fueron auto reportadas como dificultosas, a pesar de que la mayoría de las calificaciones se corresponden con la clase <No alcanza los aprendizajes requeridos>.

Gráfico 13: Proporción de coincidencias entre dificultades auto reportadas en las materias y la obtención de calificaciones PARA y NAAR que indican riesgos reales

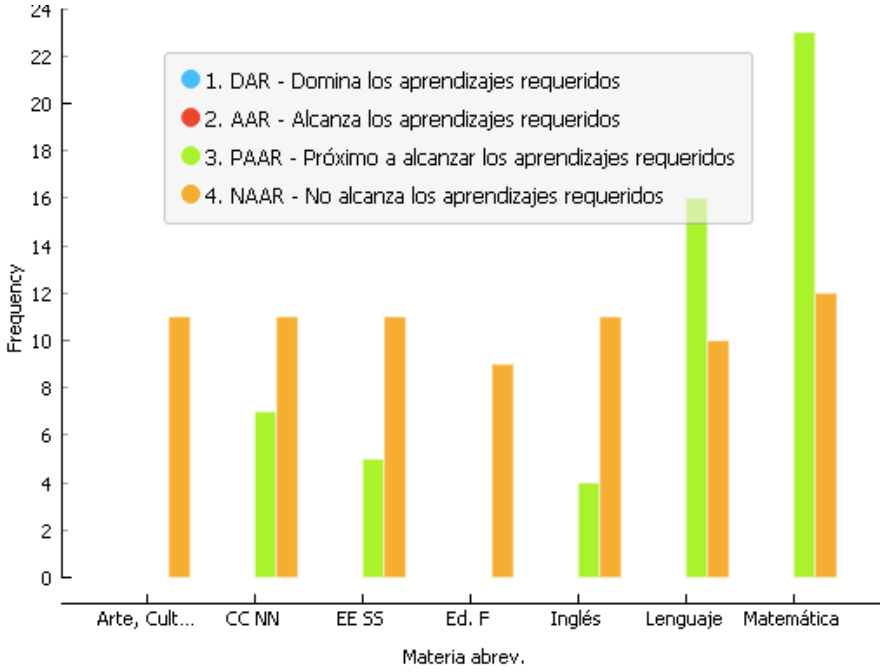


Gráfico 14: Calificaciones con riesgos que no fueron auto reportadas como dificultosas

Tabla 25: Distribución de promedios obtenidos por alumnos con familias reconstruidas, agrupados por materias

Promedio	Materia	Cantidad	Porcentaje
1. DAR - Domina los aprendizajes requeridos	Arte, Cultura	84	**27.79**
	CC NN	39	
	EE SS	36	
	Ed. F	187	
	Inglés	32	
	Lenguaje	37	
	Matemática	35	
2. AAR - Alcanza los aprendizajes requeridos	Arte, Cultura	122	**63.37**
	CC NN	179	
	EE SS	179	
	Ed. F	32	
	Inglés	183	
	Lenguaje	168	
	Matemática	163	
3. PAAR - Próximo a alcanzar los aprendizajes requeridos	Arte, Cultura	3	**4.26**
	CC NN	4	
	EE SS	5	
	Ed. F	1	

Promedio	Materia	Cantidad	Porcentaje
	Inglés	3	
	Lenguaje	24	
	Matemática	29	
	Arte, Cultura	10	
	CC NN	11	
	EE SS	11	
4. NAAR - No alcanza los aprendizajes requeridos	Ed. F	9	**4.57**
	Inglés	11	
	Lenguaje	11	
	Matemática	11	
		1619	**100.00**

El dato de las familias reconstruidas hace referencia a los matrimonios que acaban en divorcio, según Quirantes y colaboradores (2016), con anterioridad los matrimonios tenían mayor tiempo de duración, pero esta dinámica familiar ha cambiado y hasta 2016 entre el 40-50% de matrimonios terminaban en separación, en consecuencia, según los mismos autores, 4 o 5 de cada 10 niños pasarán su infancia con un solo progenitor. Para el caso de esta investigación se observa que cerca del 10% de 1610 registros de promedios de alumnos de familias reconstruidas, obtuvieron promedios con los que <No alcanzan los aprendizajes requeridos> o están <Próximos a alcanzar los aprendizajes requeridos>. El dato se complementa con que un 63% <Alcanzan los aprendizajes requeridos> pero sin llegar al nivel más alto de calificaciones.

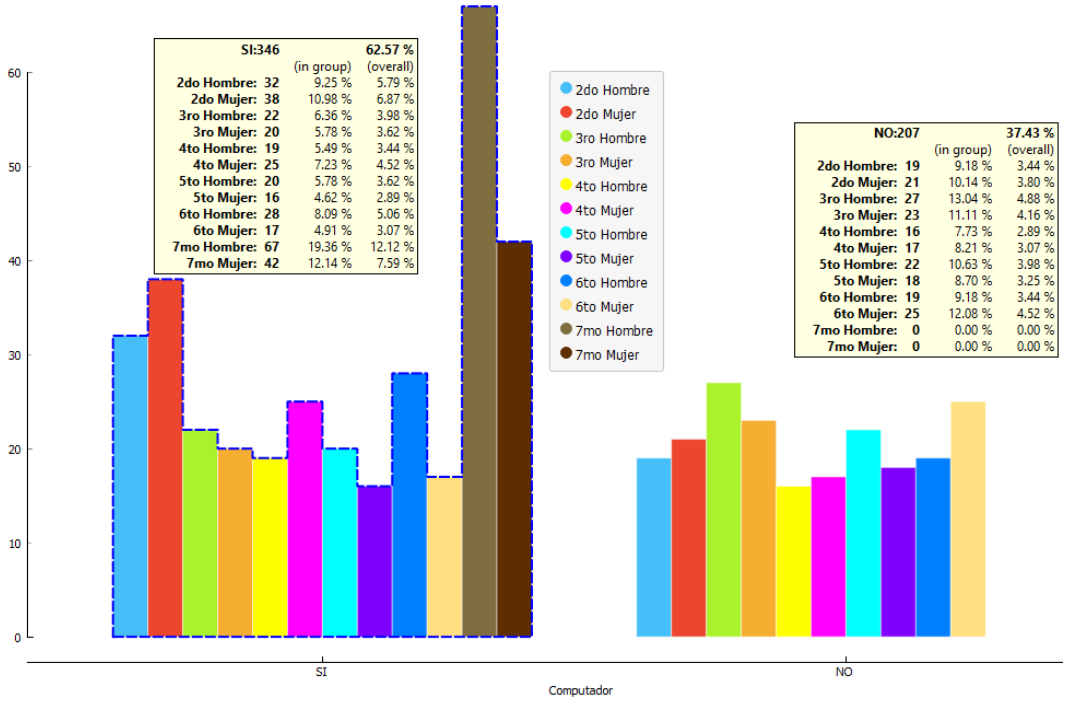


Gráfico 15: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de computador en casa, según año básico cursado por el alumno y género. Cada par de barras juntas representan un año básico

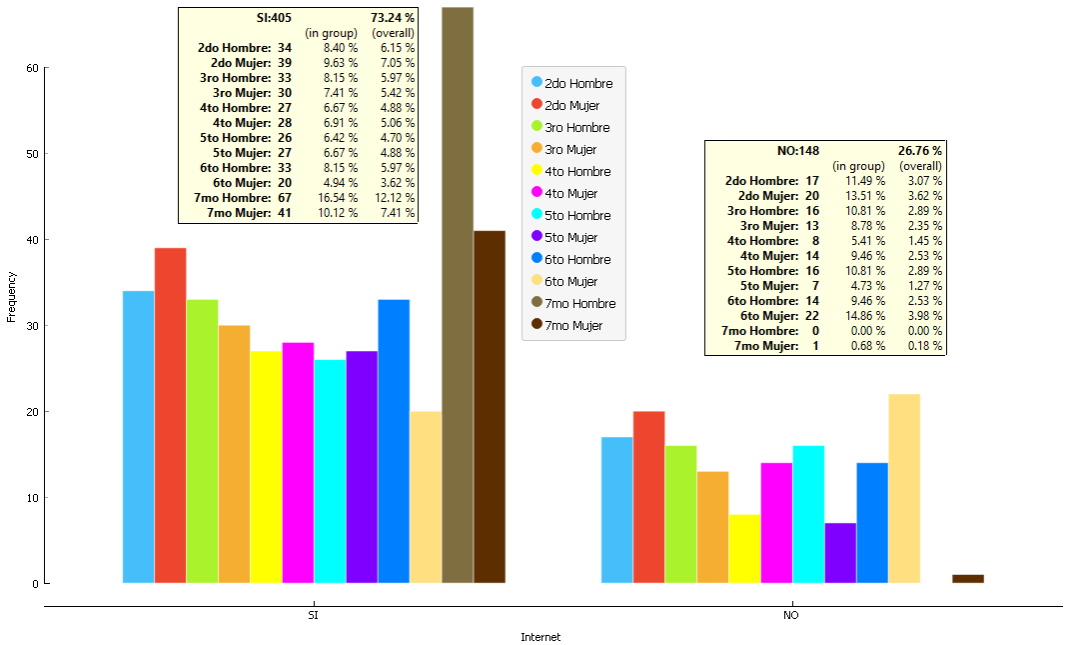


Gráfico 16: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de servicio de internet en casa, según año básico cursado por el alumno y género. Cada par de barras juntas representan un año básico

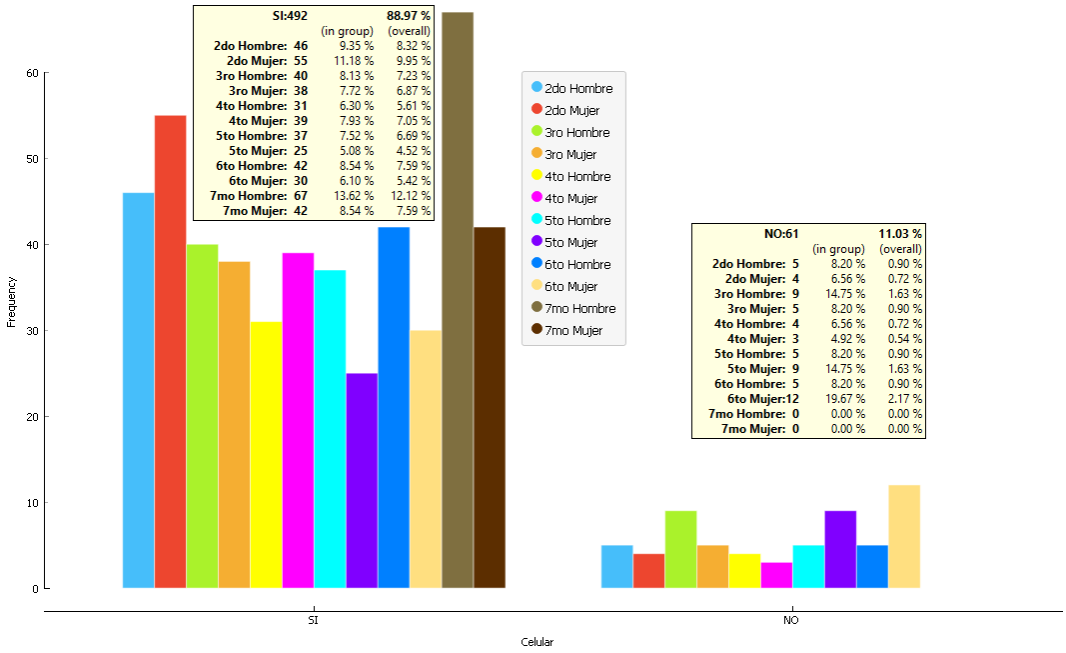


Gráfico 17: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de teléfono celular en casa para uso del alumno, según año básico cursado por el alumno y género. Cada para de barras juntas representan un año básico o curso

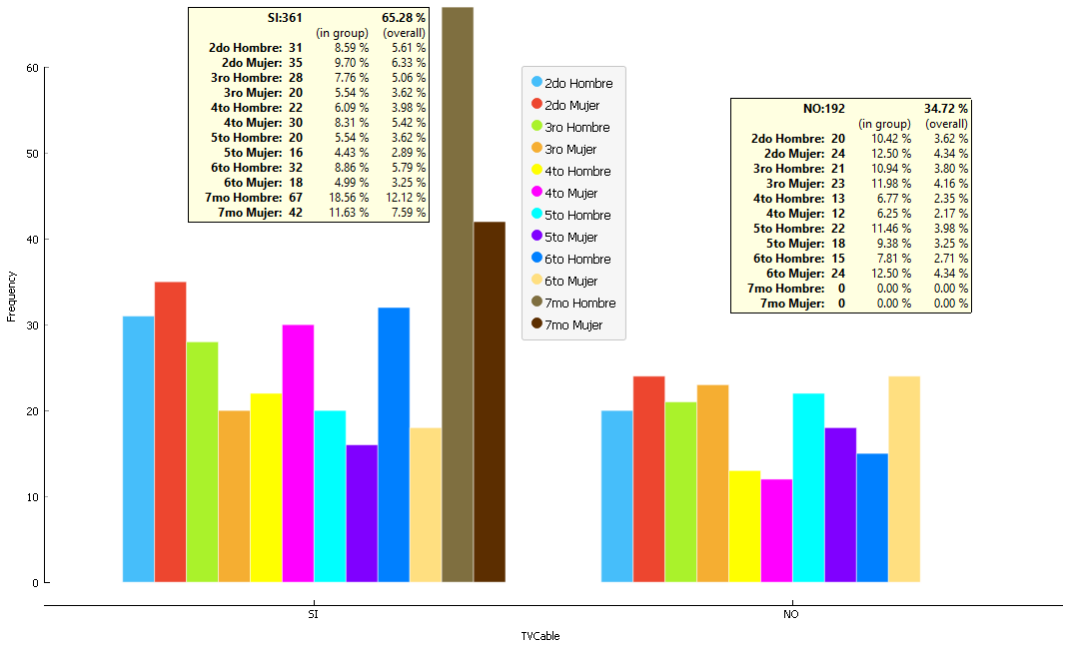


Gráfico 18: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de TV Cable regularizado en casa, según año básico cursado por el alumno y género. Cada par de barras juntas representan un año básico

El **Gráfico 18** refleja que cerca del 35% de estudiantes no dispone de servicio de TV por cable, es que si bien los estudiantes que tengan acceso a internet en sus hogares, tienen la posibilidad de ampliar sus estudios por la vía virtual, según la UNESCO (2020), especialmente en el reciente contexto del COVID-19, para la parte de la población estudiantil que no cuenta con conectividad en el hogar, las alternativas pueden ser la televisión e incluso la radio, por lo que este dato cobra relevancia.

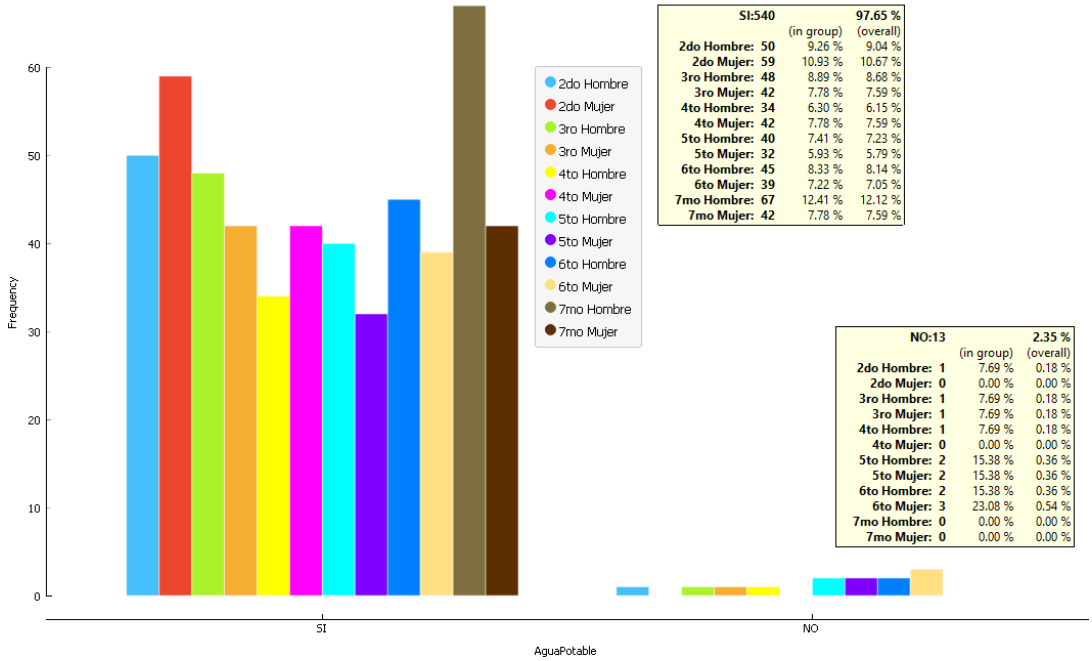


Gráfico 19: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de servicio regularizado de agua potable en casa, según año básico cursado por el alumno y género. Cada par de barras juntas representan un año básico

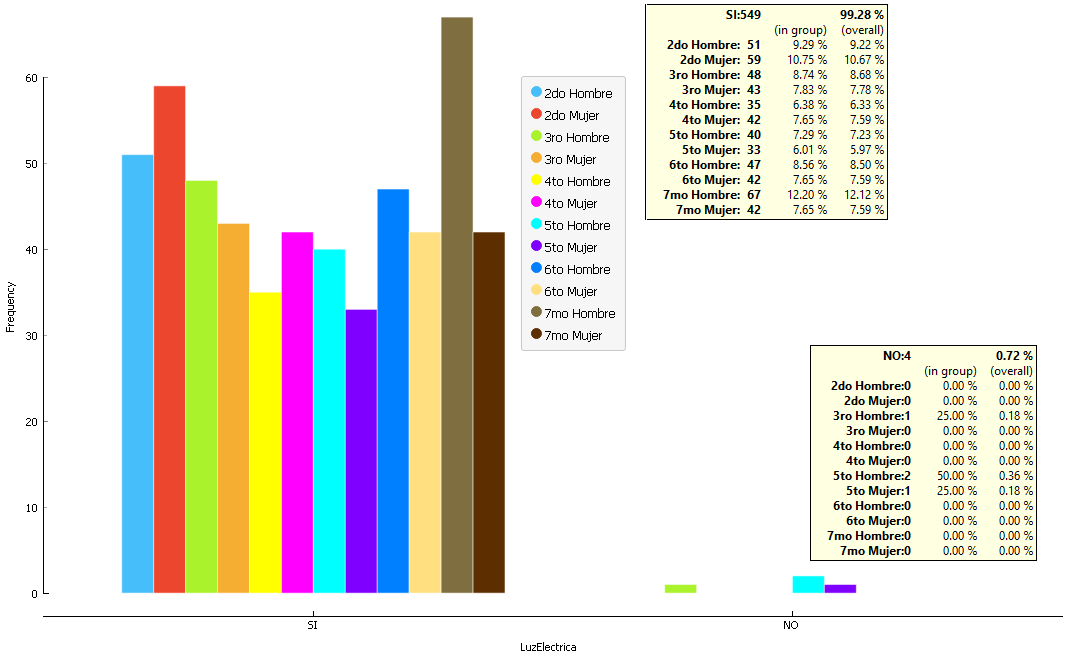


Gráfico 20: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de servicio regularizado de energía eléctrica en casa, según año básico cursado por el alumno y género. Cada par de barras juntas representan un año básico

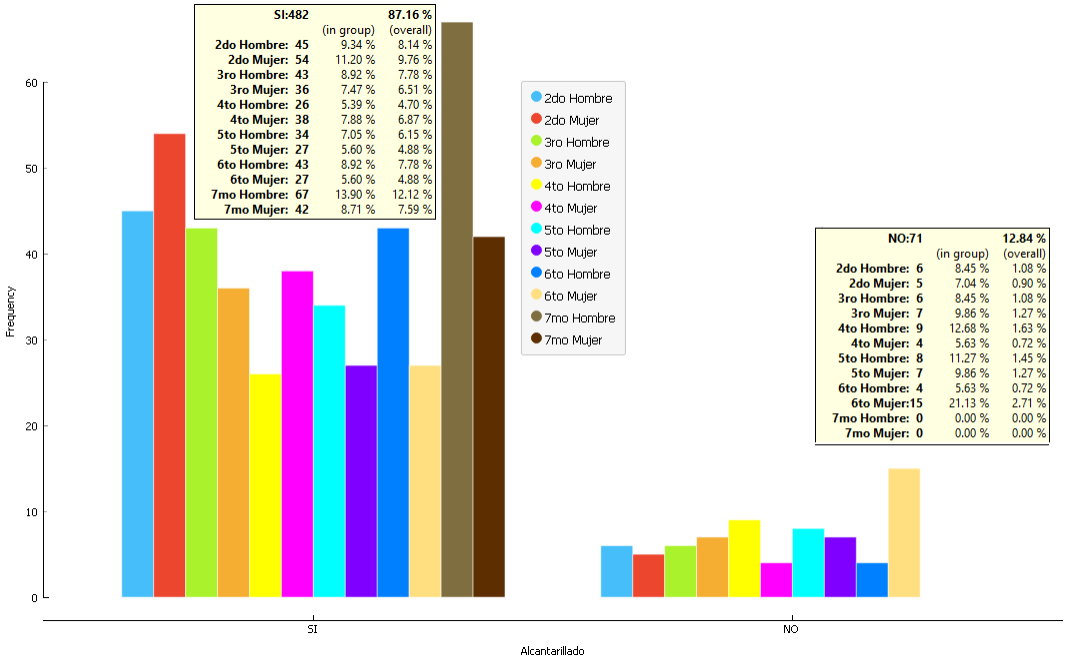


Gráfico 21: Frecuencias y porcentajes de disponibilidad (izq) e indisponibilidad (der) de servicio regularizado de alcantarillado para la casa, según año básico cursado por el alumno y género. Cada par de barras juntas representan un año básico

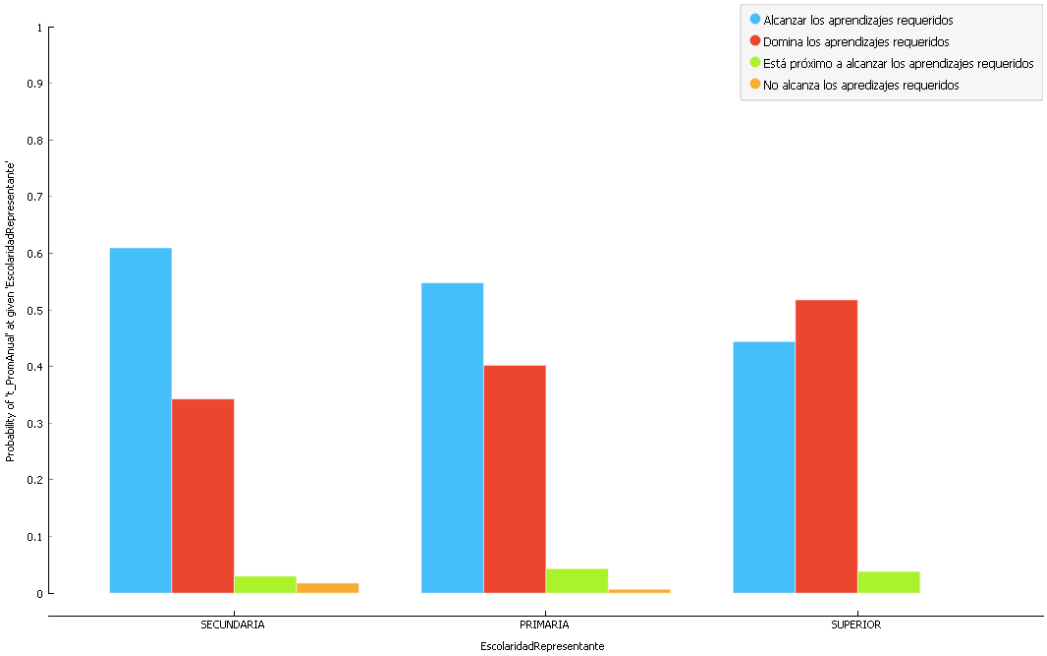


Gráfico 22: Proporción de promedios según la escolaridad del representante del alumno

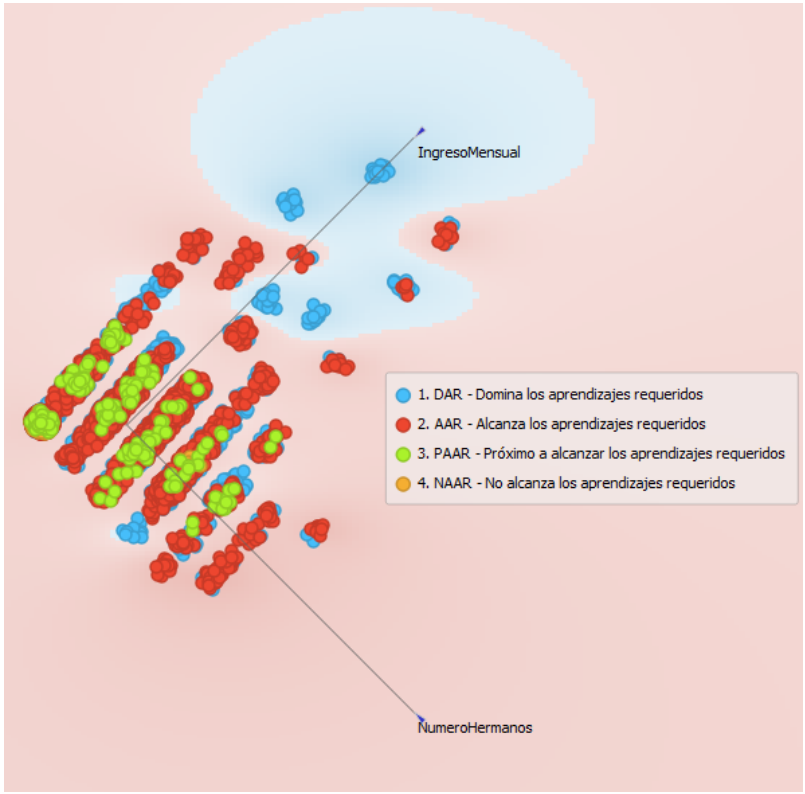


Gráfico 23: Proyección lineal que ilustra como un mayor ingreso mensual clasifica mejor a alumnos de promedios anuales DAR y AAR. Un mayor número de hermanos clasifica mejor a los alumnos con promedios AAR, que son los que aprueban con poco más que lo justo

2.3.3.3. Correlaciones

Previo a la construcción de una matriz correlacional se evaluó la normalidad de las características (variables) numéricas del conjunto de datos. En la **Tabla 26** se observa que la mediana representa con mayor precisión a la mayoría de las características numéricas en el contexto del problema estudiado, en especial a las que no siguen una distribución normal. La distribución Normal se determinó con base en la observación de la Campana de Gauss generada en Orange Data Mining para cada característica.

Tabla *26*: Media, mediana, desviación estándar y distribución normal de las características numéricas

Nº	Característica (Variable)	Media	Mediana	Desviación estándar	Mín.	Máx.	Normal
1	Año Básico	4.260	4	0.398	2	7	SI
2	Número de hermanos	1.437	1	0.899	0	6	SI
3	Ingreso mensual familiar	452.194	395	0.955	0	3100	SI
4	Año de llegada	2015.660	2016	0.001	2008	2019	SI
5	Años de retraso en estudios	1.163	0	1.479	0	7	NO
6	PQ Parcial 1	8.433	8.6	0.161	0	10	SI
7	PQ Parcial 2	8.455	8.6	0.162	0	10	SI
8	PQ Parcial 3	8.505	8.8	0.168	0	10	SI
9	Promedio Parciales Quimestre 1	8.463	8.66	0.155	0	10	SI
10	Examen Quimestre 1	8.219	8.5	0.217	0	10	SI
11	**Quimestre 1**	**8.413**	**8.6**	**0.159**	**0**	**10**	**SI**
12	SQ Parcial 1	8.452	8.6	0.172	0	10	SI
13	SQ Parcial 2	8.444	8.6	0.174	0	10	SI
14	SQ Parcial 3	8.510	8.6	0.174	0	10	SI
15	Examen Quimestre 2	8.527	9	0.205	0	10	SI
16	Promedio Parciales Quimestre 2	8.466	8.56	0.168	0	10	SI
17	**Quimestre 2**	**8.477**	**8.6**	**0.169**	**0**	**10**	**SI**
18	**Promedio Anual**	**8.441**	**8.58**	**0.161**	**0**	**10**	**SI**
19	Distancia Km, casa - escuela	1.028	0.4	1.514	0.07	7.7	NO
20	Riesgo 1er Quim. Parcial 1	0.066	0	4.354	0	2	NO
21	Riesgo 1er Quim. Parcial 2	0.057	0	4.805	0	2	NO
22	Riesgo 1er Quim. Parcial 3	0.068	0	4.390	0	2	NO
23	Riesgo 2do Quim. Parcial 1	0.064	0	4.567	0	2	NO
24	Riesgo 2do Quim. Parcial 2	0.068	0	4.436	0	2	NO
25	Riesgo 2do Quim. Parcial 3	0.064	0	4.636	0	2	NO

Con base en las ilustraciones previas de esta sección de exploración de datos, se tiene que los datos disponibles de los estudiantes no siempre siguen una relación lineal, por lo que en primera instancia se descarta el empleo del análisis correlacional (paramétrico) de Pearson, dado que para su aplicación se recomienda cumplir con este supuesto (Schober et al., 2018).

Entonces, se optó por la correlación de Spearman que evalúa qué tan bien se puede describir la relación entre dos variables utilizando una función monótona, es decir, sea la relación entre las características lineal o no. La función monótona puede ser que: (1) Aumenta monótonamente, cuando la variable x aumenta y la variable y nunca disminuye, (2) Disminuye monótonamente, cuando la variable x aumenta, pero la variable y nunca, (3) No monótona, cuando la variable x aumenta y la variable y a veces aumenta o a veces disminuye (de Winter et al., 2016).

El coeficiente de correlación de Spearman se define como el coeficiente de correlación de Pearson entre las variables de rango. Para una muestra de tamaño n, las n puntuaciones brutas X_i, Y_i se convierten rangos $R(X_i)$, $R(Y_i)$ y r_s se calcula como:

$$r_s = \rho R(X), R(Y) = \frac{cov(R(X), R(Y))}{\sigma R(X)\ \sigma R(Y)}$$

Dónde:

- ρ: Denota el coeficiente de correlación de Pearson habitual, pero aplicado a las variables de rango.
- $cov(R(X), R(Y))$: es la covarianza de las variables de rango.
- $\sigma R(R)\ \sigma R(Y)$: son las desviaciones estándar de las variables de rango.

Solo si todos los n rangos son enteros distintos, se puede calcular usando la fórmula:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Dónde:

- $d_i = R(X_i) - R(Y_i)$: es la diferencia entre los dos rangos de cada observación.
- n: es el número de observaciones.

Las correlaciones bivariadas de Spearman de cada característica con respecto del promedio anual que se obtuvieron en Orange Datamining, mediante el Widget Correlations (2015c) son las siguientes:

Tabla 27: Correlaciones entre características

Nº	Característica 1	Característica 2	Correlación	
1	PromAnual	QUI2	0.975	
2	PROM2	PromAnual	0.969	
3	PromAnual	QUI1	0.967	
4	PROM1	PromAnual	0.952	
5	PromAnual	SQP3	0.94	
6	PromAnual	SQP2	0.936	Muy fuerte
7	PromAnual	SQP1	0.935	
8	PQP3	PromAnual	0.924	
9	PQP2	PromAnual	0.912	
10	PQP1	PromAnual	0.873	
11	EXA2	PromAnual	0.839	
12	EXA1	PromAnual	0.82	
13	PromAnual	aniosRetraso	0.186	
14	IngresoMensual	PromAnual	0.162	
15	AnioIngreso	PromAnual	0.15	
16	NumeroHermanos	PromAnual	0.113	Muy débil
17	PromAnual	distanciaKM	0.062	
18	PromAnual	anioBasico	-0.073	
19	AnioLlegada	PromAnual	-0.097	
20	PromAnual	RiesgoPQP1	-0.657	
21	PromAnual	RiesgoPQP2	-0.683	
22	PromAnual	RiesgoSQP3	-0.694	Fuerte
23	PromAnual	RiesgoSQP2	-0.694	
24	PromAnual	RiesgoSQP1	-0.699	
25	PromAnual	RiesgoPQP3	-0.705	

Interpretación	Desde	Hasta
Muy débil	0.000	0.199
Débil	0.200	0.399
Moderada	0.400	0.599
Fuerte	0.600	0.799
Muy fuerte	0.800	1.000

Fuente: (Ridwan et al., 2018)

Como interpretación, en la Fila 1 se tiene que los alumnos con promedio anual más alto tuvieron a su vez una nota más alta en el segundo quimestre, por sobre la nota del primer quimestre (Fila 3), la nota del parcial 3 del segundo quimestre (Fila 5) o la nota del parcial 2 del segundo quimestre (Fila 6). Las correlaciones de los promedios anuales con respecto de los años de retraso en completar los estudios, el ingreso mensual familiar, el año de ingreso a la institución, número de hermanos o la distancia en Km desde la casa a la escuela, fueron muy débiles.

Estas correlaciones de cierto modo eran esperadas dado que la mayoría de los factores socioeconómicos se expresan como valores discretos de tipo nominal o texto y ciertamente el promedio anual se deriva de las distintas calificaciones parciales. Lo significativo de esto, es que en la mayoría de los casos sucede que los promedios anuales más altos se suscitan porque hubo un segundo quimestre con mejores calificaciones respecto del primero. Desde la Fila 20 a la 25, se observa que el promedio anual fue más alto, cuando el valor del riesgo fue menor en los parciales. Respecto de la valoración de riesgo, tal cual se muestra en la **Tabla 31**, este se estimó según el siguiente criterio: (0) para Calificaciones mayores o iguales que 7, (1) para calificaciones entre 4.00 y 6.99, (2) para calificaciones menores que 4.

2.3.3.4. Ganancia de Información e Información Mutua

El concepto de ganancia de información es clave para la selección de características durante la construcción de un árbol de decisión, que a su vez es un modelo muy recurrido en la Fase 4 de CRISP-DM en este documento. La Ganancia de la información evalúa la calidad de una variable con base en la entropía de la variable objetivo que este caso será el promedio anual y la de las restantes 51 características listadas en la **Tabla 28**. La entropía de la característica objetivo, promedio anual, se define como:

$$Entropía\ (promedioAnual) = \sum_{i=1}^{n} -p_i * log_2(p_i)$$

Dónde p_i es la proporción de ejemplos de cada clase de promedio, es decir cuántos registros de promedios corresponden con <Domina los aprendizajes requeridos>, <Alcanza los aprendizajes requeridos>, está <Próximo a alcanzar los aprendizajes requeridos> y <No alcanza los aprendizajes requeridos> de entre los 6808 registros. La entropía de cada una de las 51 características categóricas (A) restantes se define, tomando como ejemplo la característica materia, como:

$$Entropía\ (promedioAnual, A) = \sum_{v \in V_A} \frac{|E_v|}{|E|} Entropía(E_v)$$

Dónde:

- V_A = Conjunto de valores únicos y distintos de A, en este caso: Arte y Cultura, Estudios Sociales, Ciencias Naturales, Matemática, Inglés, Educación Física y Lenguaje.

- E_v= Es el conjunto de ejemplos o instancias dónde A = v. Por ejemplo, Si A = Materia y v = Matemática, entonces E_v sería el conjunto de instancias de calificaciones de Matemática. Este se repite para cada una de las 7 materias ($v \in V_A$).

Para el caso de los valores numéricos se realizan unos pasos previos: (1) se ordenan sus valores de menor a mayor, (2) para cada par de valores consecutivos (Vi, Vi+1) se calcula la media M = (Vi + Vi+1) / 2, (3) se discretiza el atributo en dos intervalos: (-∞, M] (M, +∞) y (4) se calcula la entropía con las fórmulas precedentes. Una vez que se conocen los valores de entropía de cada característica y la del atributo objetivo, la ganancia de información de cada atributo se calcula como:

Ganancia (promedioAnual, A) =

Entropía (promedioAnual) – Entropía (promedioAnual, A)

En otros términos, Maximizar Ganancia (promedioAnual, A) = Minimizar Entropía (promedioAnual, A). Ahora bien, la ganancia de información, info gain, favorece a los atributos con muchos valores, pero la información mutua o gain ratio compensa el hecho de que un atributo pueda tener muchos valores dividiendo la ganancia de información por la medida denominada información de la división (Mohammad, 2018). La información de la división se calcula de acuerdo con la siguiente fórmula:

$$InfoDivision = (promedioAnual) = -\sum_{v \in V_A} P_i log_2 \left(\frac{E_v}{E}\right)$$

Dónde los términos de la fórmula son los mismos que en las fórmulas precedentes, luego la información mutua o gain ratio de cada atributo A se calcula como:

$$Gain\ Ratio = \frac{Ganancia\ (A)}{InfoDivision\ (A)}$$

Un valor más alto de información mutua indica que la distribución de dos variables es similar, por ejemplo, en la **Tabla 28**, desde las filas 42 a 46 se observa que el número de hermanos, la escolaridad de la madre, sexo, estado civil de la madre y escolaridad del representante, comparten el valor de 0.010 de información mutua o gain ratio. Entonces estas características en particular no son determinantes la una respecto de la otra en lo que se refiere a la obtención del promedio anual que funge como objetivo en los datos de la tabla.

La Ganancia de Información (info gain) y la Información mutua (gain ratio) de cada característica que actúa como variable independiente con respecto del promedio anual que actúa como variable dependiente, que se obtuvieron en Orange Datamining mediante el Widget Rank son las mostradas en la **Tabla 28**. En la tabla se han ordenado los valores de mayor a menor para distinguir el poder de predicción de cada característica con respecto de la variable dependiente, tanto de la Ganancia y la Información mutua. La columna Nº muestra la cantidad de valores distintos que hay en cada característica de tipo texto.

La columna gain ratio que representa la información mutua con un valor mínimo de 0 y máximo de 1, señala que los comportamientos, los proyectos escolares, que entre otros aspectos evalúan la integración con otros compañeros, la disponibilidad regulada de energía eléctrica, la materia, la dificultad auto reportada, haber repetido o no un año básico, la ocupación del representante, el año básico en curso, reportar discapacidad, SEIB que finalmente determina si está en los primeros años básicos o los últimos, si vive en una familia reconstruida, entre otros, figuran como los atributos de más poder de predicción respecto del posible promedio anual.

Tabla 28: Ganancia e información mutua de cada característica con respecto del promedio anual

	Característica	Nº	Info. gain	Característica	Nº	Gain ratio
1	Materia abrev.	7	0.146	ComportamientoSQP3	3	0.083
2	Ocup. Padre	60	0.106	ComportamientoSQ	3	0.083
3	Ocup. Representante	46	0.089	ComportamientoSQP2	3	0.079
4	ProyEscSQP2	4	0.082	ComportamientoPQP3	3	0.078
5	t_anioBásico	6	0.081	ComportamientoPQ	3	0.075
6	ProyEscSQP3	4	0.076	ComportamientoSQP1	3	0.071
7	ComportamientoPQP3	3	0.073	ProyEscSQP2	4	0.065
8	anioBasico		0.073	ProyEscSQP3	4	0.063
9	ProyEscSQ	4	0.070	ComportamientoPQP2	3	0.062
10	ProyEscSQP1	4	0.069	ProyEscPQP2	3	0.059
11	ProyEscPQP2	3	0.069	ProyEscSQP1	4	0.058
12	ProyEscPQP3	3	0.069	ProyEscSQ	4	0.056
13	ComportamientoSQ	3	0.067	ProyEscPQP3	3	0.055
14	ComportamientoSQP3	3	0.066	LuzElectrica	2	0.053
15	ComportamientoSQP2	3	0.065	Materia abrev.	7	0.052
16	ComportamientoPQ	3	0.064	ProyEscPQP1	4	0.051
17	Ocup. Madre	42	0.064	ProyEscPQ	3	0.050
18	ProyEscPQP1	4	0.060	DificultadAutoreportada	2	0.045
19	ComportamientoSQP1	3	0.057	ReprobadoRepetido	2	0.045

	Característica	Nº	Info. gain	Característica	Nº	Gain ratio
20	ComportamientoPQP2	3	0.056	Ocup. Representante	46	0.039
21	ProyEscPQ	3	0.055	anioBasico		0.038
22	SBU	9	0.039	Discapacidad	2	0.037
23	SEIB	2	0.036	SEIB	2	0.036
24	IngresoMensual		0.035	t_anioBásico	6	0.032
25	Discapacidad	2	0.030	t_familiaReconstruida	2	0.031
26	aniosRetraso		0.026	Ocup. Padre	60	0.030
27	t_familiaReconstruida	2	0.025	Ocup. Madre	42	0.028
28	Enfermedad	20	0.024	Enfermedad	20	0.022
29	AnioLlegada		0.023	SBU	9	0.020
30	NumeroHermanos		0.020	aniosRetraso		0.020
31	EstadoCivilPadre	7	0.018	AguaPotable	2	0.019
32	EstadoCivilMadre	7	0.017	IngresoMensual		0.018
33	TVCable	2	0.015	ComportamientoPQP1	3	0.016
34	ComportamientoPQP1	3	0.014	TVCable	2	0.015
35	Computador	2	0.013	Computador	2	0.014
36	EscolaridadMadre	3	0.013	ParentescoRepresentante	6	0.013
37	distanciaKM		0.013	Alcantarillado	2	0.012
38	EscolaridadPadre	3	0.013	AnioLlegada		0.012
39	DificultadAutoreportada	2	0.013	EscolaridadPadre	3	0.011
40	EscolaridadRepresentante	3	0.012	EstructuraFamiliar	3	0.011
41	Sexo	2	0.010	EstadoCivilPadre	7	0.011
42	ParentescoRepresentante	6	0.009	NumeroHermanos		0.010
43	EstructuraFamiliar	3	0.008	EscolaridadMadre	3	0.010
44	Alcantarillado	2	0.007	Sexo	2	0.010
45	Telefono	2	0.006	EstadoCivilMadre	7	0.010
46	LuzElectrica	2	0.004	EscolaridadRepresentante	3	0.010
47	Internet	2	0.004	distanciaKM		0.007
48	AguaPotable	2	0.004	ProcedeDeOtraInstitucion	2	0.007
49	ProcedeDeOtraInstitucion	2	0.003	Telefono	2	0.006
50	Celular	2	0.003	Celular	2	0.005
51	ReprobadoRepetido	2	0.002	Internet	2	0.005

En la tabla anterior se observa que las columnas de ganancia de la información difieren en algunos casos respecto del orden en las columnas referidas a gain ratio, esto es porque la ganancia de información suele sesgarse hacia las características con más valores distintos, como lo son la ocupación del padre y del representante en las filas 2 y 3.

Para finalizar esta sección, se debe indicar que existen diferentes métodos de selección de atributos utilizados por los algoritmos conforme se va generando un árbol, para la selección de aquel atributo que mejor distribuye los ejemplos de acuerdo con su clasificación objetivo, por ejemplo, la ganancia de Información es usada por el Algoritmo ID3, la tasa de Ganancia es utilizada por el algoritmo C4.5 que se implementa en Orange, el índice Gini es utilizado por el algoritmo CART que se implementa en los métodos de aprendizaje en conjuntos o ensamblados de Orange (Demšar et al., 2013).

2.3.3.5. Análisis confirmatorio

Si bien el análisis confirmatorio es una técnica estadística que permite explorar con mayor precisión a factores subyacentes, constructos o variables latentes de las variables observadas o medidas en la investigación. En esta tesis no se ha construido un documento de recolección de datos, sino que se emplea datos colectados mediante fichas que las escuelas aplican con base en directrices del Ministerio de Educación de Ecuador.

El análisis confirmatorio se ha desarrollado en el software SPSS Amos 24.0.0. Se ha empleado el método de la Máxima Verosimilitud (Robitzsch, 2022). El conjunto de datos se ha agrupado en cuatro factores (que en ocasiones son referidas con la denominación de dimensiones) representadas con elipses en la **Figura 27**. Se ha prescindido de los datos que puedan indicar multicolinealidad o que los modelos de minería de datos documentados en las Fases 4 y 5 de este capítulo hayan reportado como poco incidentes en las clasificaciones o regresiones efectuadas. Estos son los cuatro factores o variables no observadas y sus respectivas variables observadas:

- *Calificaciones*. – Incluye como variables observadas a las notas de cada parcial pero no a las quimestrales para evitar multicolinealidad dado que estas derivan de las parciales. También incluye al año básico y la materia porque en la exploración de datos se ha observado que en Lenguaje y Matemática es dónde se concentran los promedios menores.
- *Factores socioeconómicos (FSE)*. – Incluye como variables observadas a los datos de la escolaridad, estado civil y ocupación de padres, madres y representantes de los alumnos. Además, género, número de hermanos, estructura familiar, distancia casa-escuela, ingresos familiares (SBU), disponibilidad o no de servicios básicos, padecimiento o no de enfermedades, padecimiento o no de discapacidad y vivir o no en familia reconstruida.
- *Sociales*. – Incluye como variables observadas a las calificaciones de los parciales

del comportamiento y los proyectos escolares, que miden las habilidades sociales de los alumnos. No se incluye a las notas quimestrales para evitar multicolinealidad dado que estas derivan de las parciales. Estos datos no se incluyeron con el resto de las calificaciones porque si bien se registran en los sistemas, no condicionan la aprobación del año académico del alumno.

- *Historial*. – Incluye como variables observadas a los años de retraso que va teniendo el alumno en sus estudios, procedencia o no desde otra escuela, reprobado o no de algún año básico y los nombres de las materias que auto reporta como dificultosas.

En la **Figura 27**, las flechas que salen de la elipse que representa al factor de habilidades sociales indican que también se desea observar tal factor mediante las variables: Representante (papá, mamá o de otro parentesco), número de hermanos, ingreso familiar, estructura familiar y por la procedencia o no de una familia reconstruida. Se hace aquello con base en la información obtenida con la exploración de datos documentada en las secciones que preceden, en otros términos, esta es la información que se requiere para el análisis confirmatorio como teoría articulada de base para su elaboración y contrastación empírica.

En la **Figura 27**, en las variables observadas que se conectan desde la elipse que representa al factor Historial Académico, las líneas con flechas en ambos extremos de las variables observadas Retraso y Repetidor, indican que estas se pueden afectar mutuamente, pues tener años de retraso en los estudios puede ser resultado de un abandono temporal de los estudios y no siempre por haber reprobado un año básico. Según Elastika y Dewanto (2021), cuando una variable observada se ve afectada por otras variables del sistema se la clasifica como variable endógena observada y cuando no es afectada se la clasifica como variable exógena observada.

En el factor de Calificaciones, las calificaciones parciales y la materia son variables endógenas observadas, porque la exploración de datos evidencia diferencias en las calificaciones de Matemática, Lenguaje y Educación Física.

En el análisis graficado también se plantean factores correlacionados mediante flechas en ambos extremos. Se ha planteado que el historial académico y las habilidades sociales se afectan mutuamente, que las habilidades sociales y los factores socioeconómicos se afectan mutuamente, que el historial académico y las calificaciones se afectan mutuamente y que los factores socioeconómicos afectan a las calificaciones. Por ejemplo, con la exploración de datos se ha observado que las bajas calificaciones pueden ocasionar la lógica reprobación del alumno, pero un

alumno que reprueba usualmente mejora sus calificaciones en el periodo siguiente (ver **Tabla 22**).

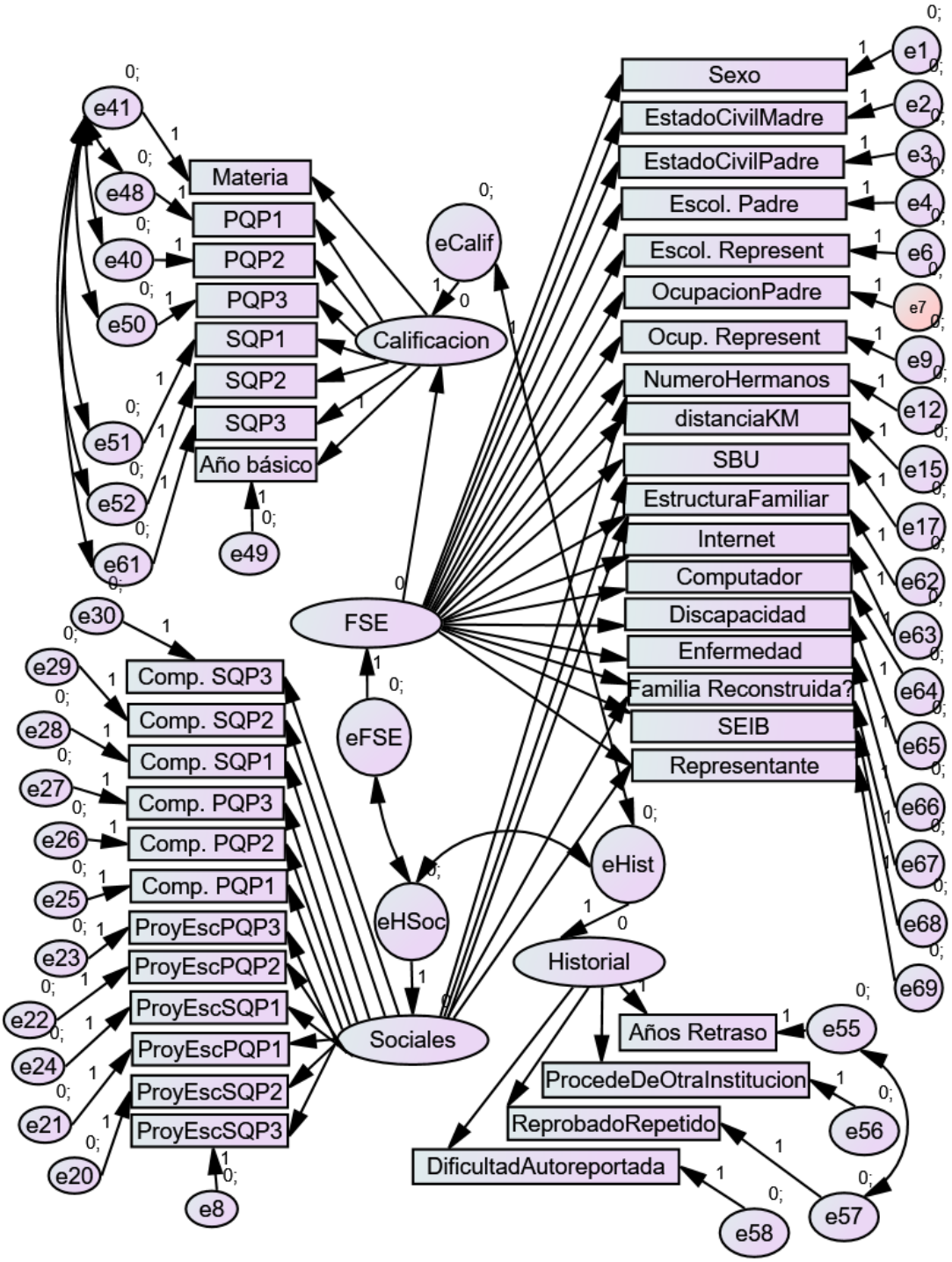


Figura 27: Imagen ilustrativa del análisis confirmatorio

En AMOS se proporcionan diversas medidas entre 0 y 1 para valorar el análisis, los valores cercanos a 1 indican un buen ajuste. Estas medidas se informan para cada modelo especificado por el usuario y para dos modelos adicionales denominados "saturado" y de "independencia". En el modelo saturado, no se imponen restricciones a los momentos de población, es un modelo vacío que se ajusta perfectamente a cualquier conjunto de datos. El modelo independentista va al extremo opuesto porque supone que las variables observadas no están correlacionadas entre sí, es un modelo severamente limitado que cabría esperar que no se ajuste bien a ningún conjunto de datos interesante. Los modelos saturados y de independencia pueden verse como dos extremos entre los que se encuentran nuestros modelos de análisis (*Appendix C: Measures of Fit*, 2016).

En la siguiente tabla se presentan los valores de la Función de Discrepancia en su Mínimo (CMIN): NPAR, CMIN, DF, P y CMIN/DF. Además, se muestran los valores de las denominadas Comparaciones con un Modelo de Referencia: NFI, RFI, IFI, TLI y CFI. Estos últimos valores se calculan con base en los valores de CMIN y los resultados de la siguiente tabla confirman lo planteado en la **Figura 27**.

Tabla 29: Valores medidos para CMIN y Baseline Comparisons

Model - CMIN	**NPAR**	**CMIN**	**DF**	**P**	**CMIN/DF**
Default model	143	3709.873	25	0.000	148.395
Saturated model	54	0.000	0		
Independence model	18	56548.427	36	0.000	1570.790
Model - Baseline	**NFI Delta1**	**RFI rho1**	**IFI Delta2**	**TLI rho2**	**CFI**
Default model	0.934	0.906	0.935	0.906	0.935
Saturated model	1.000		1.000		1.000
Independence model	0.000	0.000	0.000	0.000	0.000

De acuerdo con Uedufy (2023) y la documentación del software AMOS 24.0.0 estos valores significan:

- NFI = Índice de ajuste normado también conocido como Delta 1. Consiste en una escala de valores entre el modelo de independencia (terriblemente ajustado, que en la tabla tiene 0) y el modelo saturado (perfectamente ajustado, que en la tabla tiene 1). Un NFI de 1 muestra un ajuste perfecto, mientras que los modelos valorados < 0.9 se pueden mejorar sustancialmente. A NFI se lo denota como $NFI = \Delta_1 = 1 - \frac{\hat{C}}{\hat{C}_b} = 1 - \frac{\hat{F}}{\hat{F}_b}$. Dónde $\hat{C} = n\hat{F}$ es la discrepancia mínima del modelo que se está evaluando y $\hat{C}_b = n\hat{F}_b$ es la discrepancia mínima del modelo de

referencia. En esta investigación $NFI = 1 - \frac{3709.873}{56548.427} = 0.934$ que confirma lo graficado en la **Figura 27**.

- RFI = Índice de ajuste relativo, se deriva del NFI. Si su valor es cercano a 1 se indica un ajuste muy bueno mientras que 1 indica un ajuste perfecto. A RFI se lo denota como $RFI = \rho_1 = 1 - \frac{\frac{\hat{C}}{d}}{\frac{\hat{C}_b}{d_b}} = 1 - \frac{\hat{F}/d}{\hat{F}_b/d_b}$. Dónde al igual que en NFI, IFI, TLI y CFI, $\hat{C}$ y d son la discrepancia y los grados de libertad del modelo que se está evaluando y $\hat{C}_b$ y d_b son la discrepancia y los grados de libertad del modelo de referencia. En esta investigación $RFI = 1 - \frac{\left(\frac{3709.873}{25}\right)}{\left(\frac{56548.427}{36}\right)} = 0.906$ que confirma lo graficado en la **Figura 27**.
- IFI = Índice de ajuste incremental. Si su valor es cercano a 1 indica un ajuste muy bueno, mientras que 1 indica un ajuste perfecto. A NFI se lo denota como $IFI = \Delta_2 = \frac{\hat{C}_b - \hat{C}}{\hat{C}_b - d_b}$. En esta investigación $IFI = \frac{56548.427 - 3709.873}{56548.427 - 36} = 0.935$ que confirma lo graficado en la **Figura 27**.
- TLI = El coeficiente de Tucker-Lewis, varía entre 0 y 1 pero no se limita a dicho rango. Si su valor es cercano a 1 se indica un ajuste muy bueno, mientras que 1 representa un ajuste perfecto. A TLI se lo denota como $TLI = \rho_2 = \frac{\frac{\hat{C}_b}{d_b} - \frac{\hat{C}}{d}}{\frac{\hat{C}_b}{d_b} - 1}$. El 0.906 obtenido confirma lo graficado en la **Figura 27**.
- CFI = El índice de ajuste comparativo tiene un valor truncado entre 0 y 1. Si su valor es cercano a 1 se indica un ajuste muy bueno. Esto se debe a que los valores mayores que 1 se informan como 1, mientras que los valores menores que 0 se informan como 0. A CFI se lo denota como $CFI = 1 - \frac{\max(\hat{C} - d, 0)}{\max(\hat{C}_b - d_b)} = 1 - \frac{\max(3709.873 - 25,\ 0)}{\max(56548.427 - 36,\ 0)} = 1 - \frac{3648.873}{56512.427} = 0.935$. Cómo un valor CFI de ≥ 0.95 se considera de un ajuste excelente para el modelo, el valor obtenido confirma lo graficado en la **Figura 27** de modo muy aceptable.

AMOS ofrece varios estimadores que ayudan a comprender a detalle resultados como los indicados en las líneas anteriores. Uno de ellos es el Efecto Total Estandarizado. En la celda resaltada de amarillo de la siguiente tabla, el Efecto Total Estandarizado (directo e indirecto) del factor Calificación en SQP3 es 0.941. Es decir, debido a los efectos directos (sin mediación) e indirectos (mediados) de la

Calificación en SQP3, cuando la Calificación aumenta en 1 su desviación estándar, SQP3 aumenta 0.941 de desviación estándar. Así en los demás casos.

Tabla 30: Estimadores basados en el efecto total estandarizado

Variables observadas	Factores (Dimensiones)			
	FSE	Historial	Calificación	Sociales
Calificación	0.162	0	0	0
Parentesco de representante	0.105	0	0	0.095
Procede de otra institución	0	0.071	0	0
Años de retraso	0	0.707	0	0
SQP3	0.152	0	**0.941**	0
Reprobado o Repetido	0	0.071	0	0
Dificultad Auto reportada	0	0.071	0	0
Estructura Familiar	0.105	0	0	0.095
Internet	0.106	0	0	0
Computador	0.106	0	0	0
Discapacidad	0.106	0	0	0
Enfermedad	0.106	0	0	0
Familia Reconstruida	0.105	0	0	0.095
SEIB	0.106	0	0	0
Año básico	0.122	0	0.754	0
Materia	0.018	0	0.114	0
ComportamientoPQP1	0	0	0	0.096
ComportamientoSQP3	0	0	0	0.096
ComportamientoSQP2	0	0	0	0.096
ComportamientoSQP1	0	0	0	0.096
ComportamientoPQP3	0	0	0	0.096
ComportamientoPQP2	0	0	0	0.096
SQP2	0.153	0	0.946	0
SQP1	0.151	0	0.937	0
PQP3	0.145	0	0.899	0
PQP2	0.146	0	0.903	0
PQP1	0.138	0	0.856	0
ProyEscSQP3	0	0	0	0.096
ProyEscSQP1	0	0	0	0.096
ProyEscPQP3	0	0	0	0.096
ProyEscPQP2	0	0	0	0.096
ProyEscPQP1	0	0	0	0.096
ProyEscSQP2	0	0	0	0.096
SBU	0.105	0	0	0.095
Distancia casa-escuela KM	0.299	0	0	0
Número de hermanos	0.65	0	0	0.349
Ocupación de representante	0.106	0	0	0
Ocupación del padre	0.106	0	0	0
Escolaridad del representante	0.106	0	0	0
Escolaridad del padre	0.106	0	0	0
Estado civil del padre	0.106	0	0	0

Variables observadas	Factores (Dimensiones)			
	FSE	Historial	Calificación	Sociales
Estado civil de la madre	0.106	0	0	0
Sexo	0.728	0	0	0

3.3. Fase 3. Preparación de los datos

En esta fase se obtienen los datos aún faltantes, se limpian, normalizan y transforman en un conjunto de datos tabular y optimizado, adecuado para los modelos empleados en la Fase 4. Es de indicar que todas las actividades concernientes a la preparación de datos se han realizado en el software Orange Datamining 3.34, aunque en principio los datos socioeconómicos fueron llenados en un libro de Microsoft Excel, copiándolos desde fichas elaboradas en Microsoft Word o digitándolos porque se encontraban en fichas impresas.

A nivel de filas las tareas efectuadas se resumen en remover duplicados, remover valores atípicos, consolidar las filas de los datos de las dos escuelas estudiadas y sobre muestrear las clases minoritarias. A nivel de columnas las tareas efectuadas se resumen en selección de características, reducción de la dimensionalidad, convertir a numéricos los datos discretos mediante la técnica de One Hot Encoding y Label Encoding cuando los algoritmos lo requieran y la creación de 27 nuevas columnas. A nivel de valores las tareas efectuadas se resumen en imputaciones y cálculos de valores faltantes, cambio de valores y escalado de datos. Este proceso se resume en la **Figura 30:** Vista parcial de la preparación de datos en Orange Data Mining 3.34.

Sobre las técnicas de codificación de variables categóricas, las principales técnicas son: (1) Encontrar y reemplazar, donde cada aparición coincidente de un carácter o caracteres se reemplaza unos nuevos, (2) Codificación de etiquetas o Label Encoding, dónde a cada etiqueta se le asigna un número entero único según el orden alfabético y (3) Codificación en caliente o One Hot Encoding, que es dónde se convierte cada variable categórica en una columna y se le asigna un valor de 1 o 0, algunos autores denominan a One Hot Encoding como binarización (Hwang, 2019).

3.3.1. Selección de los datos

De modo general, se utilizó el 100% de las instancias o ejemplos disponibles en las dos escuelas. También se empleó a todas las columnas con excepción de: Curso, Disponibilidad de teléfono convencional, Grado, Paralelo, Jornada, Nombres (porque se deben anonimizar) y Calificación de exámenes quimestrales sobre 2 puntos, tal cual se justificó en la sección **3.2.2. Descripción del conjunto de datos** Las características se seleccionan para cada algoritmo con base en su función dentro del

modelo. No se consideró otros criterios de inclusión o exclusión de datos para esta selección.

3.3.2. Limpieza de los datos

Respecto de los datos faltantes, que en ningún caso superó el 5% de instancias, se empleó diversos modos de imputación, que es definida por Brownlee (2016) como la sustitución de datos faltantes o inconsistentes por valores estimados, destinados a crear un conjunto de datos sin valores faltantes. Se procedió de la siguiente manera:

- En casos de datos faltantes del *año básico* cursado por el alumno se los calculó en función del año de llegada del alumno a la escuela, en otros casos se buscó en documentos que listen a los alumnos por cursos.
- En casos de datos faltantes del *año de llegada a las escuelas* se los calculó en función del año básico que cursa el alumno, este dato es necesario para estimar casos de alumnos que han tardado más de lo previsto en culminar sus estudios.
- En casos de datos faltantes de la *escolaridad del padre y de la madre* se los imputó con un modelo de regresión, específicamente C4.5.
- En casos de datos faltantes de la *escolaridad del representante,* cuando fue posible se los igualó con el del padre o madre, determinado por el parentesco del representante con el alumno. En los otros casos se los imputó con C4.5.
- En casos de datos faltantes del *estado civil del padre y de la madre* se los imputó con el modelo C4.5.
- En casos de datos faltantes de *estado civil del representante,* cuando fue posible se los igualó con el del padre o madre, determinado por el parentesco del representante con el alumno. En los otros casos se los imputó con C4.5.
- En casos de datos faltantes de *ocupación del padre y de la madre* se los imputó con un modelo de regresión, específicamente C4.5.
- En casos de datos faltantes de *ocupación del representante,* cuando fue posible se los igualó con el del padre o madre, determinado por el parentesco del representante con el alumno. En los otros casos se los imputó con C4.5.
- En casos de datos faltantes de *estructura familiar,* se los completó con la Moda, que en este caso fue la Nuclear o compuesta por padre y madre.
- En casos de datos faltantes de servicios regularizados de *agua, luz, alcantarillado, teléfono celular, internet, televisión por cableoperadora y computador,* se los

imputó con la Moda, que en todos los casos fue de que SI disponían de tales servicios.

- En casos de datos faltantes de poseer *discapacidad,* se los completó con la Moda, que en este caso fue de que NO poseían discapacidad.
- En casos de datos faltantes de poseer *enfermedad,* se los completó con la Moda, que en este caso fue de NINGUNA enfermedad.
- En casos de datos faltantes de *dificultad auto reportada* en las materias, se los completó con la Moda, que en este caso fue de NO tener dificultad. Un alumno puede tener cero o más dificultades auto reportadas.
- En casos de datos faltantes de calificaciones de *Comportamiento y de Proyectos escolares,* se los imputó con un modelo de regresión, específicamente C4.5. Estas calificaciones son cualitativas.
- En casos de datos faltantes de calificaciones de los *Parciales,* se los imputó con un modelo de regresión, específicamente C4.5. Estas calificaciones son cuantitativas. Luego, se calculó los promedios Quimestrales que se derivan de los parciales.

En lo que respecta a errores de datos, estos se dieron en los nombres de algunas enfermedades, denominaciones de ocupaciones de padres, madres y representantes. Estas correcciones se efectuaron con el Editor de Dominios de Orange (Widget Edit Domain), que permite editar el dominio de un conjunto de datos, renombrar características, combinar valores de las características categóricas, agregar un valor categórico y asignar etiquetas (Demšar et al., 2013; Orange, 2015f). Con este mismo widget se corrigió tipos de datos que Orange interpretó como nominales o textos, siendo continuos o numéricos. Las imputaciones mencionadas en la lista previa se efectuaron con el widget Impute, tal cual se muestra de modo parcial en la **Figura 28**. Además, fue necesario complementar valores faltantes con código de Python.

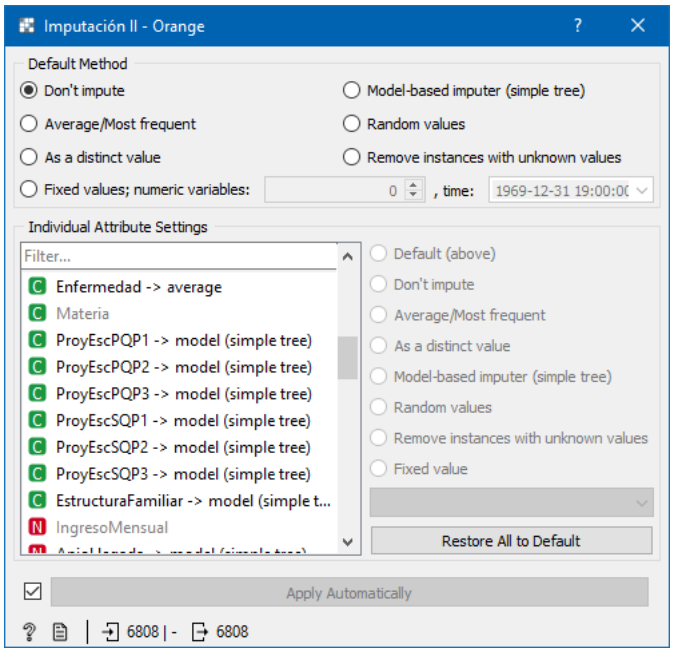


Figura 28: Vista parcial de las imputaciones ejecutadas con el Widget Impute de Orange

3.3.3. Construcción de nuevos datos

Como actividad de transformación de datos se agregaron nuevas columnas a partir de las existentes. Estas se muestran en la **Tabla 31**, dónde además se especifica un identificador, nombre de la variable, descripción (💬 =Texto, 123 = Número). La información se complementa con la **Figura 29**. La **Figura 30** presenta una vista parcial del flujo de trabajo construido en Orange Datamining para las detalladas actividades de preparación de datos.

Tabla 31: Ficha de las nuevas características predictoras y de respuestas agregadas para el análisis

Nº	Nueva columna	Tipo y descripción	
1	Género	💬 Llenado manualmente en función de los nombres de cada alumno	
2	SEIB	💬 Correspondencia del Sistema de Educación Intercultural Bilingüe (SEIB) con los grados de Educación Básica.	
		Grados de educación básica 2, 3 y 4:	*FCAP Fortalecimiento Cognitivo, afectivo y psicomotriz*
		Grados de educación básica 5, 6 y 7:	*DDTE Desarrollo de destrezas y técnicas de estudio*
3	t_anioBasico	💬 *Año básico cursado por el alumno en formato de texto, ejemplo, 2 = Segundo*	
4	RiesgoPQP1	123 Calificación de riesgo en función de la nota del Parcial 1 del Quimestre 1	

Nº	Nueva columna	Tipo y descripción	
		Nota de primer parcial del Quimestre 1 mayor o igual a 7	*0*
		Nota de primer parcial del Quimestre 1 de entre 4.00 y 6.99	*1*
		Nota de primer parcial del Quimestre 1 menor que 4	*2*
5	RiesgoPQP2	123 Igual criterio que la nueva columna anterior, pero con el parcial 2	
6	RiesgoPQP3	123 Igual criterio que la nueva columna anterior, pero con el parcial 3	
7	RiesgoSQP1	123 Calificación de riesgo en función de la nota del Parcial 1 del Quimestre 2	
		Nota de primer parcial del Quimestre 2 mayor o igual a 7	*0*
		Nota de primer parcial del Quimestre 2 de entre 4.00 y 6.99	*1*
		Nota de primer parcial del Quimestre 2 menor que 4	*2*
8	RiesgoSQP2	123 Igual criterio que la nueva columna anterior, pero con el parcial 2	
9	RiesgoSQP3	123 Igual criterio que la nueva columna anterior, pero con el parcial 3	
10	t_RiesgoPQP1	Calificación cualitativa de riesgo en función de la nueva columna RiesgoPQP1, correspondiente al primer parcial, quimestre 1.	
		RiesgoPQP1 igual 0	*Sin riesgo*
		RiesgoPQP1 igual 1	*Moderado*
		RiesgoPQP1 igual 2	*Alto*
11	t_RiesgoPQP2	Igual criterio que la nueva columna anterior, pero con RiesgoPQP2, correspondiente al segundo parcial, quimestre 1.	
12	t_RiesgoPQP2	Igual criterio que la nueva columna anterior, pero con RiesgoPQP3, correspondiente al tercer parcial, quimestre 1.	
13	t_RiesgoSQP1	Calificación cualitativa de riesgo en función de la nueva columna RiesgoSQP1, correspondiente al primer parcial, quimestre 2.	
14	t_RiesgoSQP2	Igual criterio que la nueva columna anterior, pero con RiesgoSQP2, correspondiente al segundo parcial, quimestre 2.	
15	t_RiesgoSQP3	Igual criterio que la nueva columna anterior, pero con RiesgoSQP3, correspondiente al tercer parcial, quimestre 2.	
16	DistanciaKM	123 Distancia calculada con Google Maps en Km entre la escuela y la zona central de residencia auto reportada por el alumno. En todos los casos se siguió las calles y avenidas principales.	
17	t_familiaReconstruida	Indicador de SI o NO el alumno vive con padrastro o madrastra.	
		Estado civil diferente entre padre y madre	*SI*
		Estado civil igual entre padre y madre	*NO*
18	SBU	123 Cantidad de sueldos básicos unificados (SBU) que representa el ingreso familiar, se obtiene dividiendo el ingreso familiar entre 394 que es la cantidad en dólares equivalente al SBU de 2019.	
19	t_QUI1	Calificación cualitativa correspondiente al primer quimestre según **Tabla 1**	
20	t_PQP1	Calificación cualitativa correspondiente al primer parcial, del primer quimestre	

Nº	Nueva columna	Tipo y descripción
21	t_PQP2	Calificación cualitativa correspondiente al segundo parcial, del primer quimestre
22	t_PQP3	Calificación cualitativa correspondiente al tercer parcial, del primer quimestre
23	t_QUI2	Calificación cualitativa correspondiente al primer quimestre según **Tabla 1**
24	t_SQP1	Calificación cualitativa correspondiente al segundo parcial, del primer quimestre
25	t_SQP2	Calificación cualitativa correspondiente al segundo parcial, del primer quimestre
26	t_SQP3	Calificación cualitativa correspondiente al segundo parcial, del primer quimestre

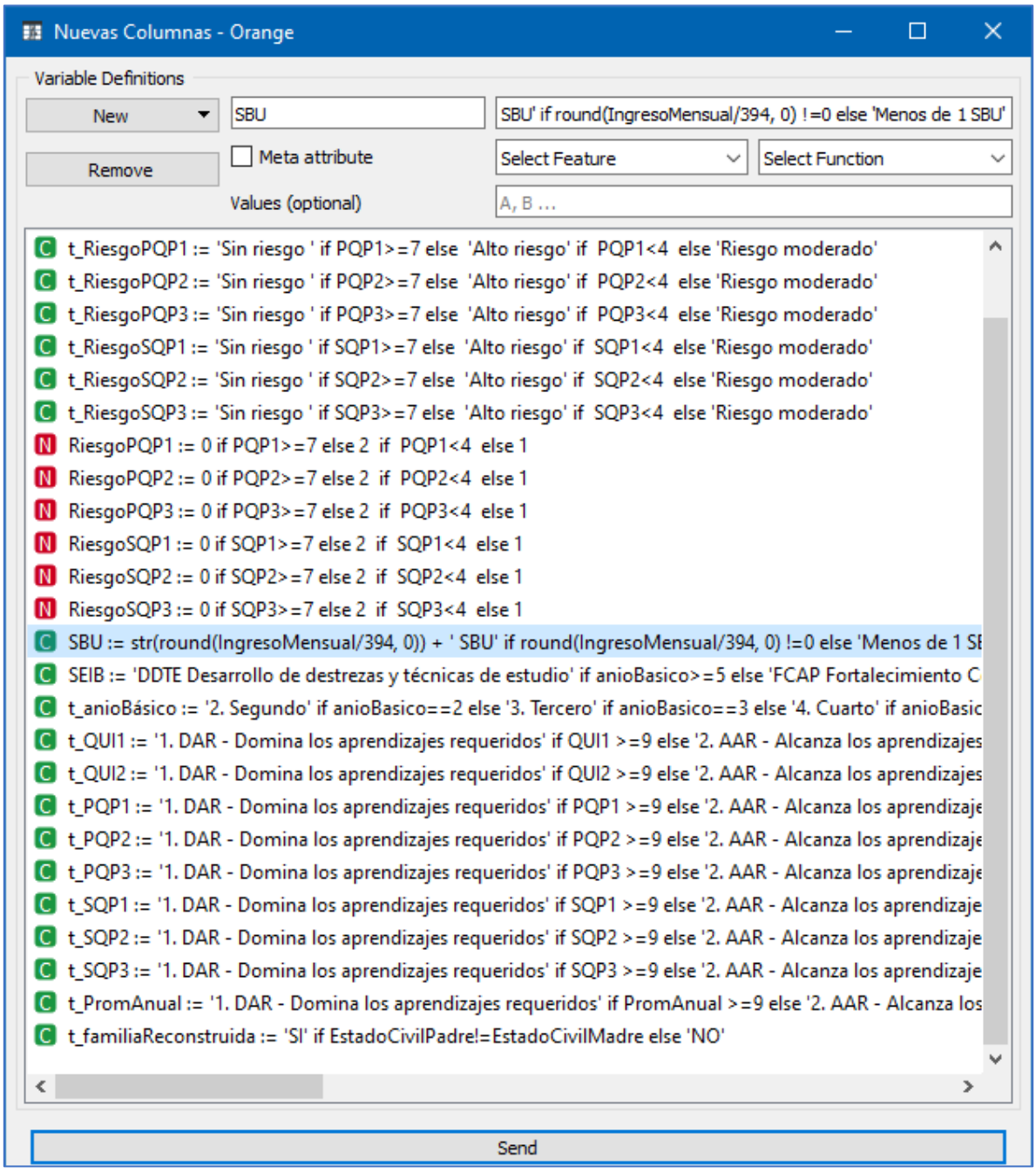


Figura 29: Vista parcial de la creación de nuevas columnas desde el Widget Feature Constructor de Orange

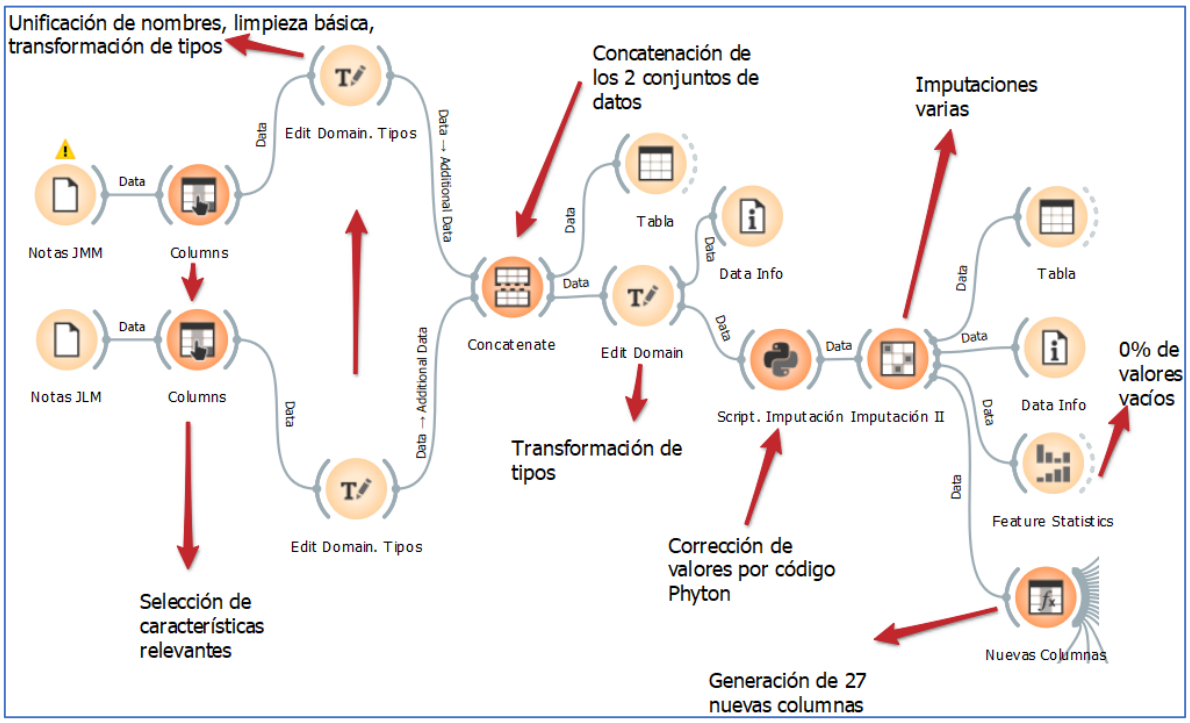


Figura 30: Vista parcial de la preparación de datos en Orange Data Mining 3.34

3.3.4. Aumento de datos

La preparación de los datos para esta investigación implicó resolver problemas que ocurren cuando se usa un conjunto de datos desequilibrado, en este caso, respecto de los alumnos en riesgo, pues los alumnos que <No alcanzan los aprendizajes requeridos> y los <Próximos a alcanzar los aprendizajes requeridos> eran minorías en los promedios anuales y en los parciales respecto de aquellos que sí <Alcanzan los aprendizajes requeridos> y <Dominan los aprendizajes requeridos>.

La literatura existente sugiere varias técnicas para sobrellevar tales casos, en aras de no perder instancias o filas para el análisis como sucede con el submuestreo de clases mayoritarias (Grina et al., 2022), se optó por el empleo de SMOTE para el sobre muestreo de las indicadas clases minoritarias.

SMOTE significa Synthetic Minority Oversampling Technique o Técnica de sobre muestreo de minorías sintéticas, es una técnica de aprendizaje automático que realiza el aumento de datos mediante la creación de puntos de datos sintéticos ligeramente diferentes de los puntos de datos originales, es decir, sin generar duplicados, con ello, se evita que el modelo casi nunca prediga clases minoritarias (Chawla et al., 2002; Fernández et al., 2018).

Con SMOTE se consiguió reducir los falsos negativos, a costa de aumentar los falsos positivos, en consecuencia, se mejoró los valores en la métrica del Recuerdo (Recall) aunque se aminoró en algo la Precisión, tal cual se explica en la Fase 4 de este capítulo. En otros términos, se agregó más predicciones de la clase minoritaria, algunas de ellas correctas favoreciendo el Recuerdo y otras erróneas que decrementan la Precisión. En la **Tabla 32** se muestra el bloque de código de Python insertado en un Script, junto con la distribución de instancias respecto de la clase <Promedio Anual>.

Para el uso de SMOTE se debe importar al entorno de Orange la librería imbalanced-learn, para ello se ejecuta la instrucción <pip install -U imbalanced-learn>. La librería es de código abierto, con licencia MIT, se basa en scikit-learn. En la línea 9 se establece la estrategia de muestreo como auto, lo que significa que al tratarse de un sobre muestreo u over sampling (línea 10), las clases minoritarias se igualan en número con la mayoritaria que tal cual se muestra en el **Gráfico 24** es <Alcanza los aprendizajes requeridos>. En el **Gráfico 25** se muestran los datos balanceados. La documentación oficial de la librería se encuentra en el sitio web oficial de Imbalanced-Learn (Lemaître et al., 2017).

Tabla 32: Código Python en un script de Orange y la distribución de clases antes y después del sobre muestreo sintético

```
def python_script(in_data):
```

```
 1 import Orange
 2 import sklearn
 3 import numpy as np
 4 from imblearn.over_sampling import SMOTE
 5 from imblearn.under_sampling import RandomUnderSampler
 6 from Orange.data import Domain, Table
 7 from collections import Counter
 8 #Transformación del conjunto de datos de la salida
 9 oversample = SMOTE(sampling_strategy='auto')
10 X, Y = oversample.fit_resample(in_data.X, in_data.Y)
11 #Creación de una nueva tabla para la salida
12 #a partir de los datos de entrada (in_data)
13 domain = Domain(in_data.domain.attributes, in_data.domain.class_vars)
14 out_data = Table.from_numpy(domain, X, Y)
15 print ('Entradas %s' %Counter(in_data.Y))
16 print ('Salida   %s' %Counter(out_data.Y))
```

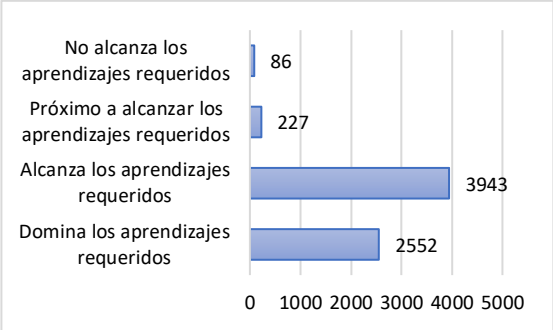


Gráfico 24: Datos desbalanceados respecto de la clase

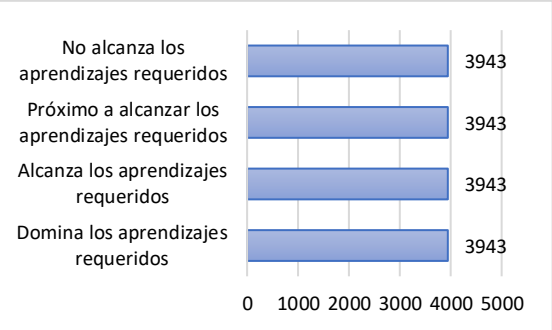


Gráfico 25: Datos balanceados respecto de la clase

En el **Gráfico 24** se observa que las dos clases minoritarias tienen 86 y 227 instancias respectivamente. Para fines de obtener un balanceo de datos personalizado se construyó otro conjunto de datos sobre muestreados con base en cinco pasos:

1) Se tomó las instancias de la clase mayoritaria <Alcanza los aprendizajes requeridos> con el widget Select Rows (Orange, 2015y).
2) De las filas seleccionadas en el paso 1, se seleccionó el 78% de instancias de modo aleatorio, esto corresponde a 3076. Para ello se empleó el widget Data Sampler (Orange, 2015d).
3) Aquel 78% de instancias referidas en el paso 2, se concatenó con las instancias de las restantes del paso 1, mediante el widget Concatenate (Orange, 2015a).
4) Se especificó como columna objetivo tipo categórica al Promedio Anual, mediante el widget Select Column (Orange, 2015x).
5) Se sobre muestreó las clases minoritarias con un script de Python (Orange, 2015t). En lugar de sobre muestrear las clases al número de instancias de la clase mayoritaria, en las líneas 9 y 10 de la **Figura 31** se especifica las instancias por sobre muestrear en cada clase. En la misma figura se resalta en el rectángulo amarillo la cantidad de instancias entrantes y salientes con la aplicación de SMOTE.

La **Figura 32** ofrece una vista parcial del flujo de trabajo (workflow) construido en Orange para la realización de los cinco pasos detallados previamente. Nótese que se ha mantenido el nombre de los Widget con la finalidad de ser ilustrativos respecto de cada paso seguido. Posterior al sobre muestreo de clases minoritarias se estableció como objetivo (target) el atributo **cuantitativo** del promedio anual para efectuar las regresiones documentadas en la Fase 4 de este capítulo, también se efectuaron clasificaciones con el sobre muestreo personalizado.

Editor

```
def python_script(in_data):
```

```
 1 import Orange
 2 import sklearn
 3 import numpy as np
 4 from imblearn.over_sampling import SMOTE
 5 from imblearn.under_sampling import RandomUnderSampler
 6 from Orange.data import Domain, Table
 7 from collections import Counter
 8 #Smote personalizado
 9 strategy = {1:3076, 0:2552, 2:1500, 3:1000}
10 oversample = SMOTE(sampling_strategy=strategy)
11 X, Y = oversample.fit_resample(in_data.X, in_data.Y)
12 #Creación de una nueva tabla para la salida
13 #a partir de los datos de entrada (in_data)
14 domain = Domain(in_data.domain.attributes, in_data.domain.class_vars)
15 out_data = Table.from_numpy(domain, X, Y)
16 print ('Entradas %s' %Counter(in_data.Y))
17 print ('Salida   %s' %Counter(out data.Y))
```

```
    return out_data, out_learner, out_classifier, out_object
```

Console

```
Running script:
Entradas Counter({1.0: 3076, 0.0: 2552, 2.0: 227, 3.0: 86})
Salida   Counter({1.0: 3076, 0.0: 2552, 2.0: 1500, 3.0: 1000})
```

Figura 31: Sobre muestreo ponderado

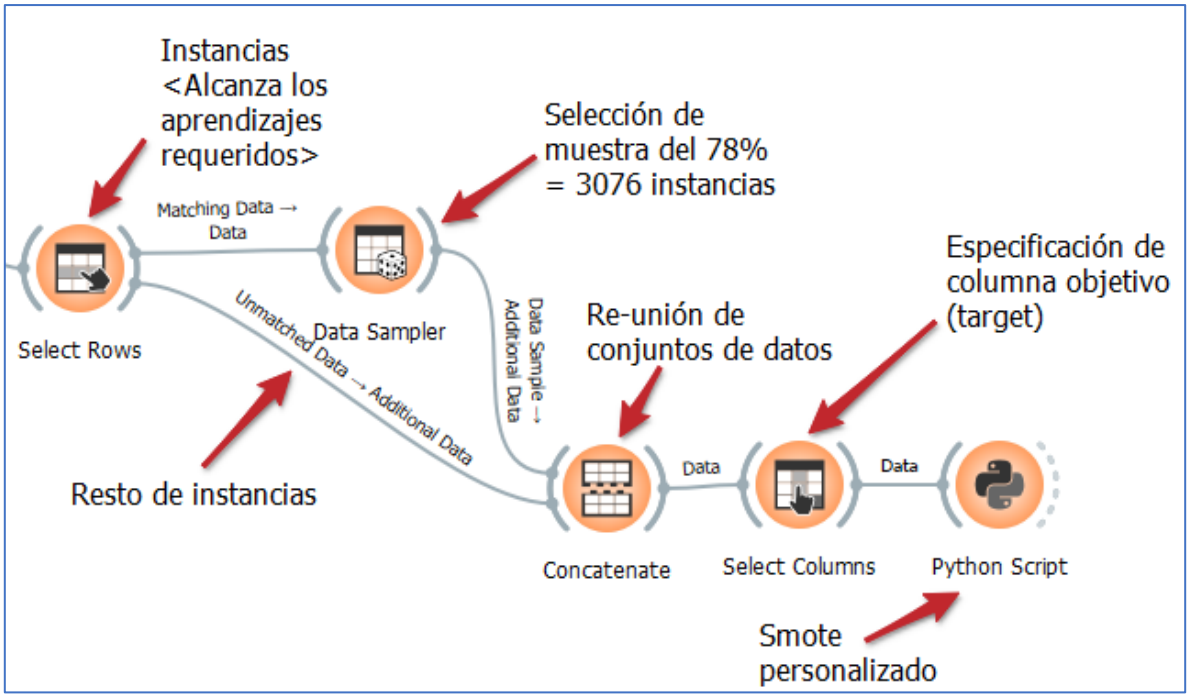


Figura 32: Vista parcial del modelo en Orange para el sobremuestreo ponderado de clases minoritarias

3.3.5. Reducción de la dimensionalidad

Cómo se explicó en la sección **2.5.3.5. Análisis de componentes principales**, PCA es un algoritmo que reduce la dimensión de un conjunto de datos a la vez que conserva la mayor variabilidad posible, para ello identifica las direcciones posibles llamadas componentes principales, dentro de las que la variación de los datos es la máxima. Con ello se facilita la exploración visual de datos de altas dimensiones y se aminora el tiempo de procesado de los modelos de aprendizaje automático que se apliquen a los datos. En suma, se redujo la dimensionalidad de los datos a 15 componentes y se obtuvo un 30% de varianza explicada, como se explicará en la Fase 6 de evaluación de modelos.

3.3.6. Formato de datos

Como paso final antes de la generación del modelo, se debe de comprobar si algunos algoritmos requieren de aplicar un formato concreto o la clasificación de los datos. Por ejemplo, un algoritmo de secuencia requiere que los datos estén clasificados de forma previa antes de ejecutar el modelo, o que sea recomendable asegurar la aleatorización de las filas previo de una validación cruzada. Cada uno de estos particulares requerimientos se especifican antes de la aplicación de cada modelo en la Fase 4.

De modo general, habiendo consolidado los datos de las dos escuelas, igualando en tipos y nombres a cada característica y ejecutado las tareas de preparación de datos descritas a este momento, el flujo parcial de la preparación de datos construido hasta esta parte se corresponde con la **Figura 30:** Vista parcial de la preparación de datos en Orange Data Mining 3.34.

3.4. Fase 4. Modelado

En esta Fase 4 se construyó modelos supervisados y no supervisados con la intención de cumplir los objetivos de la investigación, se empleó 13 algoritmos a los datos preparados en la Fase 3. Además, se realizó varias iteraciones con los concernientes ajustes de hiperparámetros y parámetros hasta obtener los resultados y tiempos documentados en la Fase 6, lo cual ayudó a determinar cuándo dejar de hacer ajustes y cambiar a un nuevo modelo. En promedio se realizó cinco sesiones de ajustes de parámetros.

Además, se buscó cumplir los supuestos estadísticos y no estadísticos de cada modelo, como el balanceo de instancias con respecto de las clases, el escalado de

datos, problemas generales de calidad de los datos y otros aspectos documentados para cada modelo en el capítulo **2. Marco teórico**. Respecto de los métodos no supervisados se empleó las métricas disponibles desde Orange.

3.4.1. Generalidades

A continuación, se mencionan algunas generalidades a los modelos en esta fase:

1. El análisis se ha dividido en dos grupos, los modelos analizando todos los datos posibles y los modelos analizando los datos sin incluir las calificaciones, porque es de esperar que el promedio final se correlacione con las calificaciones progresivas que lo constituyen. Esto último, porque puede ocasionar multicolinealidad, que es la relación de dependencia lineal fuerte entre más de dos características en la regresión múltiple, de modo especial, en el paso 6 de esta lista.
2. Los modelos no supervisados, también llamados descriptivos, se emplean sobre el conjunto de datos original compuesto por 6808 filas y 88 columnas. Aunque por su naturaleza se realiza una preparación adicional de datos para suministrarles la mayor cantidad de atributos categóricos posibles que ellos requieren como entradas.
3. Los modelos supervisados, también llamados predictivos, se emplean sobre el conjunto de datos balanceado con respecto de las clases mayoritarias, es decir, sobre muestreando las clases minoritarias. Este conjunto de datos se compone de 15772 filas y 88 columnas, tal cual se explicó en la sección **3.3.4. Aumento de datos**.
4. En los modelos supervisados se emplea el valor cualitativo y el cuantitativo del promedio anual como objetivo para las tareas de clasificación y regresión respectivamente.
5. Los modelos supervisados se emplean con todas las características disponibles y también reducidos en dimensiones mediante una aplicación previa del Análisis de Componentes Principales, PCA. Tras varias pruebas para obtener altos valores de clasificación, el valor seleccionado para retener la mayor variación de ellos en vectores llamados componentes principales fue de 15. Con 15 componentes principales se alcanzó una varianza explicada del 30%.
6. En todas las aplicaciones de los modelos supervisados, el método de muestreo utilizado para evaluar el rendimiento de los modelos es el aleatorio estratificado. En este tipo de muestreo se ha dividido al conjunto de datos en 10 subgrupos o estratos, en función del promedio anual. Luego se selecciona el 80% de los datos de cada subgrupo para el entrenamiento y la media de las métricas será el

resultado global para cada modelo supervisado.

7. Se incorporó a la validación cruzada a los modelos supervisados de regresión, configurada con 10 pliegues o folds y en cada pliegue se usó el 80% de datos para el entrenamiento y el 20% para la prueba.

Se prescindió de la validación con un conjunto de datos de prueba porque los datos para el entrenamiento son sobre muestreados y los de prueba no, entonces era poco probable que se tuviese una proporción representativa de instancias con clases minoritarias al momento de probar y las que hubiese tenderían a reportar altos valores de Exactitud (Classification Accuracy, CA) poco realistas dado el desequilibrio de las clases, por lo que Mukhopadhyay (2018) señala que la métrica no se adecúa para clases desequilibradas. Los 7 pasos precedentes se ilustran en la siguiente figura:

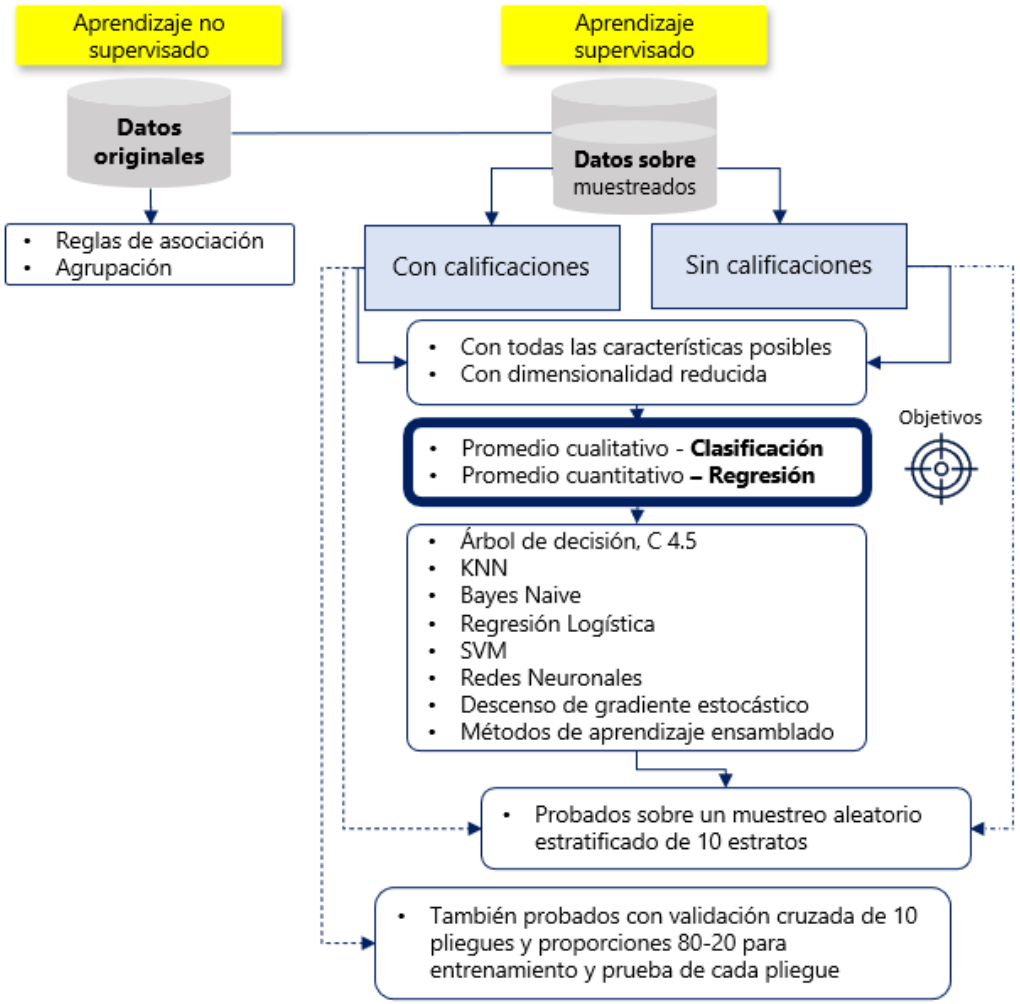


Figura 33: Resumen de la Fase de modelado

3.4.2. Parámetros e hiperparámetros

La mayoría de las técnicas de modelado tienen diferentes hiperparámetros o configuraciones que se pueden ajustar para controlar el proceso de entrenamiento con datos, por ejemplo, de los árboles de decisión como C4.5 se puede controlar su profundidad, divisiones y otros ajustes. Por lo común este proceso requiere de refinamientos una y otra vez hasta obtener los mejores resultados respecto de los objetivos.

3.4.3. Aprendizaje no supervisado

Una de las formas más comunes de aprendizaje no supervisado es explorar regularidades, en este caso relacionadas con los registros de calificaciones y factores socioeconómicos de los estudiantes de escuela.

	Itemsets	Support	%
1	ReprobadoRepetido=NO	6775	99.52
	LuzElectrica=SI	6738	98.97
	AguaPotable=SI	6607	97.05
2	t_RiesgoPQP2=Sin riesgo	6492	95.36
3	DificultadAutoreportada=NO	6475	95.11
4	t_RiesgoSQP3=Sin riesgo	6468	95.01
5	t_RiesgoSQP1=Sin riesgo	6459	94.87
6	t_RiesgoSQP2=Sin riesgo	6439	94.58
7	t_RiesgoPQP3=Sin riesgo	6433	94.49
8	t_RiesgoPQP1=Sin riesgo	6430	94.45
9	ParentescoRepresentante=MAMÁ	6078	89.28
10	ProcedeDeOtraInstitucion=NO	6038	88.69
11	Celular=SI	5925	87.03
	Alcantarillado=SI	5723	84.06

Figura 34: Patrones frecuentes y sus soportes detectados en el conjunto de datos

Con la aplicación de FP-Growth documentado en la sección **2.5.3.1. Patrones frecuentes, FP-Growth**, se detectó 11 grupos de interés, como el 2 de la **Figura 34** que se corresponde con un 95.36% de registros de calificaciones sin riesgo en el parcial 2 del primer quimestre. En la **Figura 35** se ajustaron los parámetros de detección a un 20% de soporte y un mínimo de 15 ítems por grupo detectado.

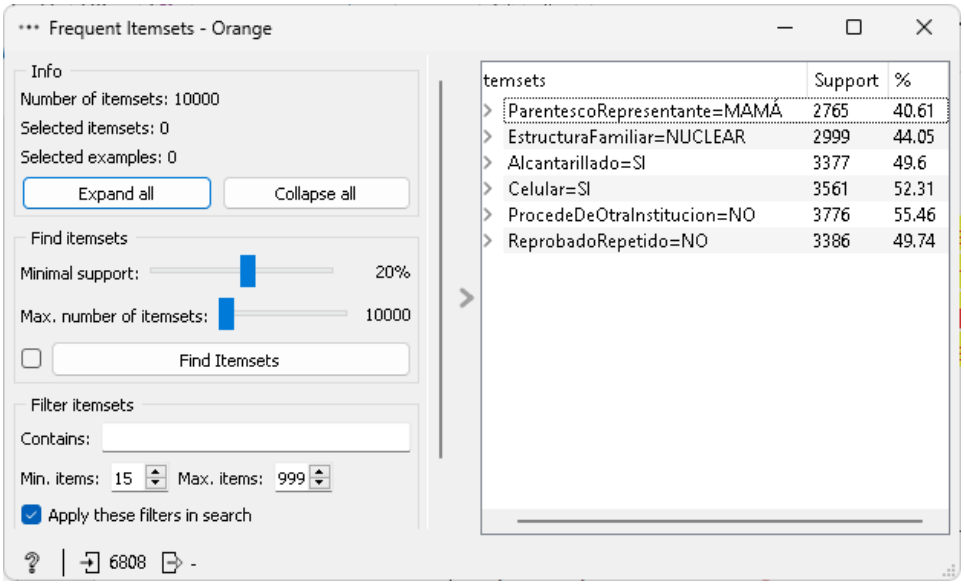


Figura 35: Patrones frecuentes filtrados por soporte e ítems (instancias) por grupo

Cada grupo detectado con FP-Growth se puede desagregar para explorar que lo constituye, como se muestra en la **Figura 36** que explora el patrón de instancias referida a alumnos que no auto reportan dificultad en asignaturas, viven en familia nuclear y les representa su mamá. Estos alumnos generalmente tienen un patrón de conducta B, que no es el más alto, pero es muy aceptable.

Itemsets	Support	%
ParentescoRepresentante=MAMÁ	2765	40.61
EstructuraFamiliar=NUCLEAR	2765	40.61
Alcantarillado=SI	2765	40.61
Celular=SI	2925	42.96
ProcedeDeOtraInstitucion=NO	3215	47.22
ReprobadoRepetido=NO	3379	49.63
DificultadAutoreportada=NO	3252	47.77
ComportamientoPQP1=B	3252	47.77
ComportamientoPQ=B	3252	47.77
ComportamientoSQP1=B	3252	47.77
ComportamientoSQP2=B	3252	47.77
ComportamientoSQP3=B	3252	47.77
ComportamientoSQ=B	3252	47.77
t_RiesgoPQP1=Sin riesgo	3252	47.77
t_RiesgoPQP2=Sin riesgo	3252	47.77
t_RiesgoPQP3=Sin riesgo	3252	47.77
t_RiesgoSQP1=Sin riesgo	3252	47.77
t_RiesgoSQP2=Sin riesgo	3252	47.77
Enfermedad=NINGUNA	2703	39.7

Figura 36: Exploración de un patrón frecuente

Otro de los modelos de aprendizaje no supervisado fue el de la sección **2.5.3.2. K-Means**, K-Means es un modelo muy popular para encontrar grupos interesantes en los datos, aporta diversas métricas para comprender que tan bien se han constituido dichos grupos, aunque este documento se centra en la Puntuación de la Silueta, dónde cuanto mayor sea la puntuación de Silueta, mejor será la agrupación.

La **Figura 37** es una vista parcial del widget Silhouette Plot (2015aa) de 40 instancias con puntuación negativa. Dada la alta resolución del gráfico de silueta no es posible mostrarlo en este documento, pero está disponible desde https://tinyurl.com/5n7sh6ta. Cuando se observó a los grupos por el promedio anual obtenido, resultó que los grupos con puntuación de silueta más altas corresponden a alumnos que <Dominan los aprendizajes requeridos> y <No alcanzan los aprendizajes requeridos>, es decir, ambos extremos. Es evidente que también reportan instancias con puntuación negativa, pero en menor medida respecto de los otros dos tipos de promedios.

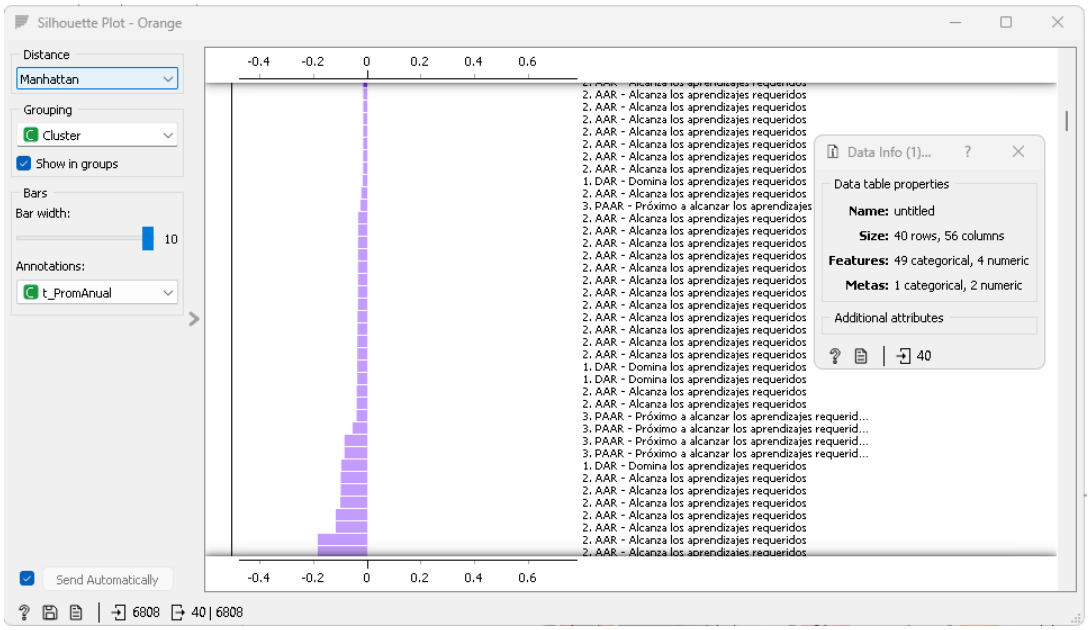


Figura 37: Exploración de un grupo de 40 instancias que K-Means puntúa con valores negativos

Para los resultados anteriores se configuró los siguientes parámetros:

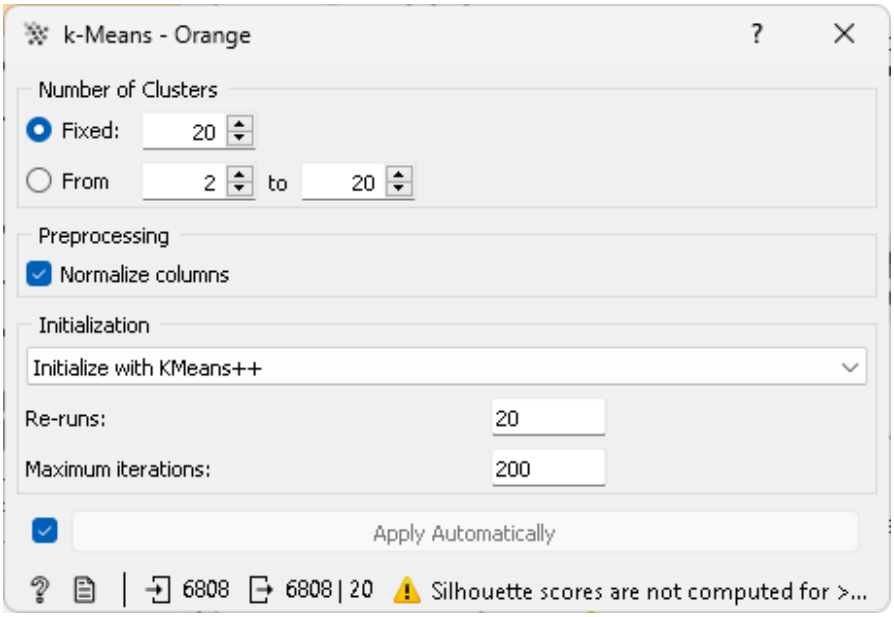


Figura 38: Configuración de parámetros para K-Means

Otro de los modelos de aprendizaje no supervisado que se empleó, fue el de la sección **2.5.3.3. Clúster jerárquico**.

Para la aplicación del clúster jerárquico, se calculó la distancia entre instancias, con la medida de Manhattan, ver **Figura 39**. También se normalizó los datos.

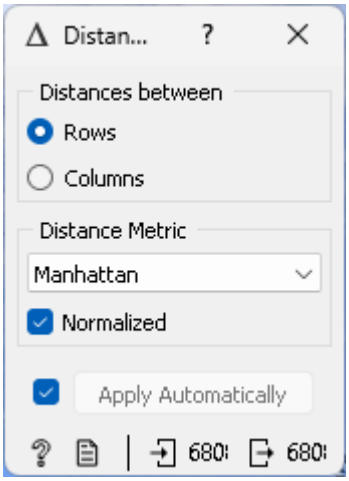


Figura 39: Configuración para el cálculo de Distancias

Al disponer de datos de tipo continuo y categóricos, la elección de medidas de distancias se redujo a la Euclidiana, Manhattan o Hamming, pues medidas como la disimilitud del Coseno, Spearman, Absoluta de Spearman, Pearson, Absoluta de Pearson, Bhattacharyya o Jaccard ignoran a las características categóricas, en tanto que, Mahalanobis se limita al manejo de 1000 instancias (Orange, 2015e).

En términos de Aggarwal, Hinneburg y Keim (2001), la medida seleccionada, la de Manhattan, es ideal para aplicaciones de alta dimensión y se define como la sumatoria del valor absoluto de las diferencias de las distancias entre los puntos en un plano cartesiano: $\sum_{j=1}^{p}|y_{1j} - y_{2j}|$, dónde p es el número de características, y_{1j} es el valor del descriptor j en la entidad 1 y y_{2j} es el valor del descriptor j en la entidad 2. La **Figura 40** corresponde a la vista parcial del flujo para la construcción de un

agrupamiento jerárquico en Orange, en ella se usan en orden los widgets Distances (2015e), Hierarchical Clustering (2015j), Select Rows (2015z) para la selección del grupo a explorar con Scatter Plot (2015w).

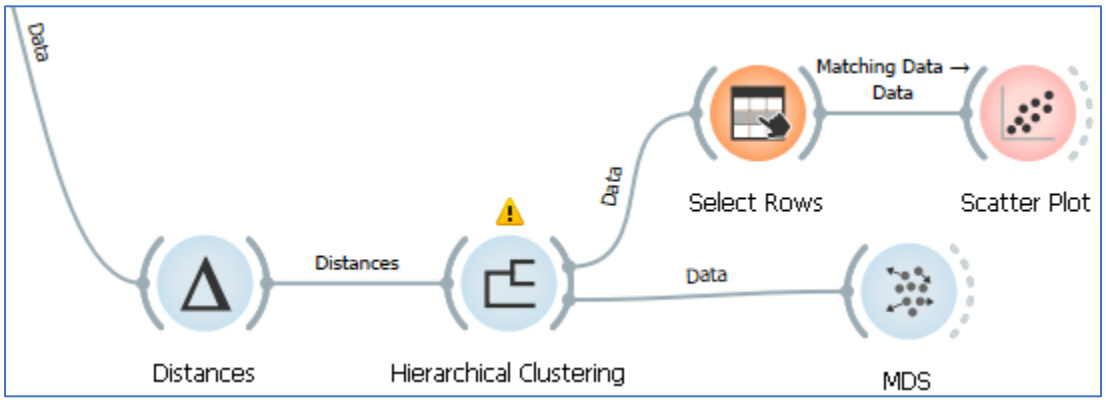


Figura 40: Vista parcial del flujo para la construcción de agrupamiento jerárquico en Orange

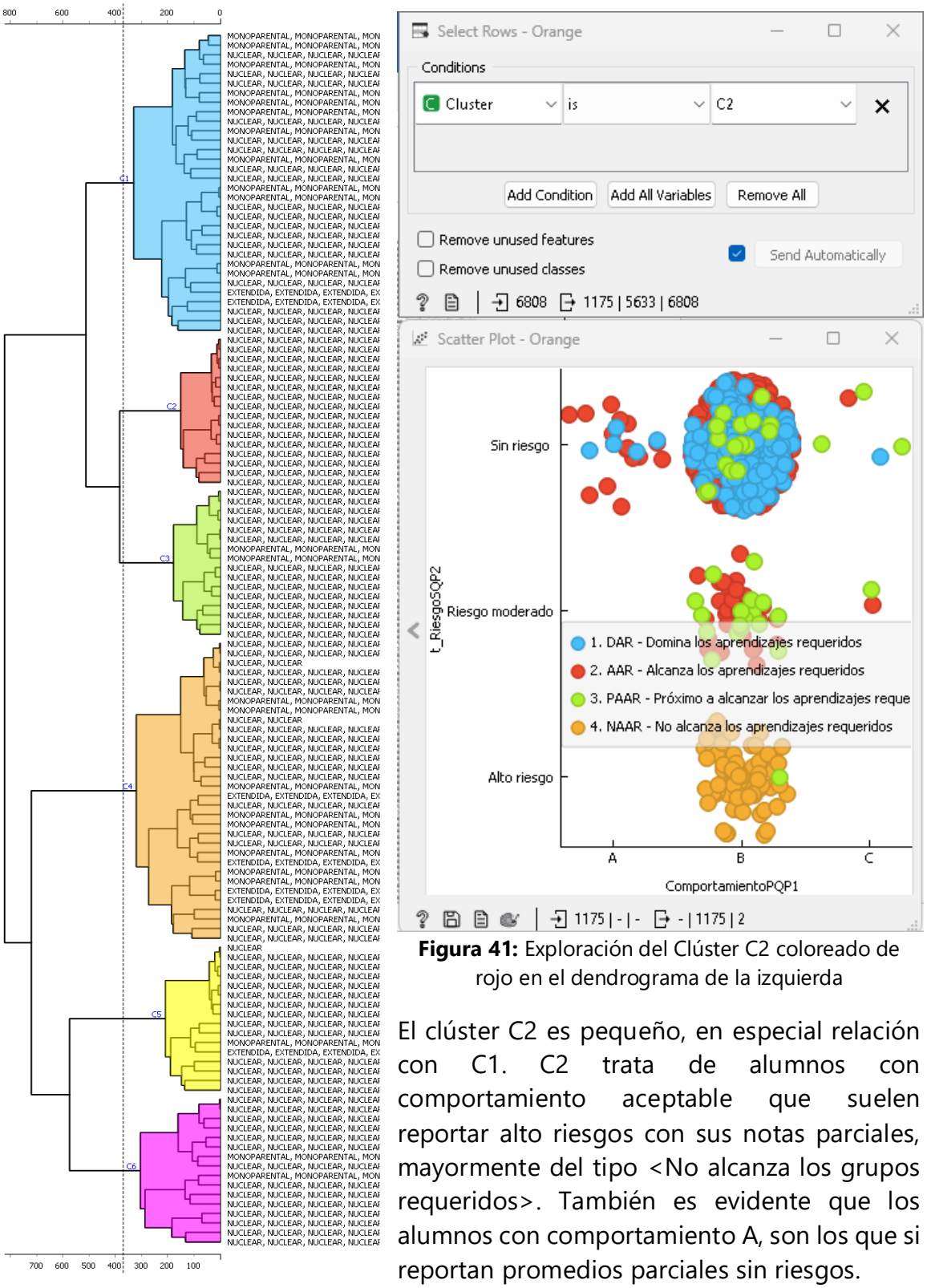


Figura 41: Exploración del Clúster C2 coloreado de rojo en el dendrograma de la izquierda

El clúster C2 es pequeño, en especial relación con C1. C2 trata de alumnos con comportamiento aceptable que suelen reportar alto riesgos con sus notas parciales, mayormente del tipo <No alcanza los grupos requeridos>. También es evidente que los alumnos con comportamiento A, son los que si reportan promedios parciales sin riesgos.

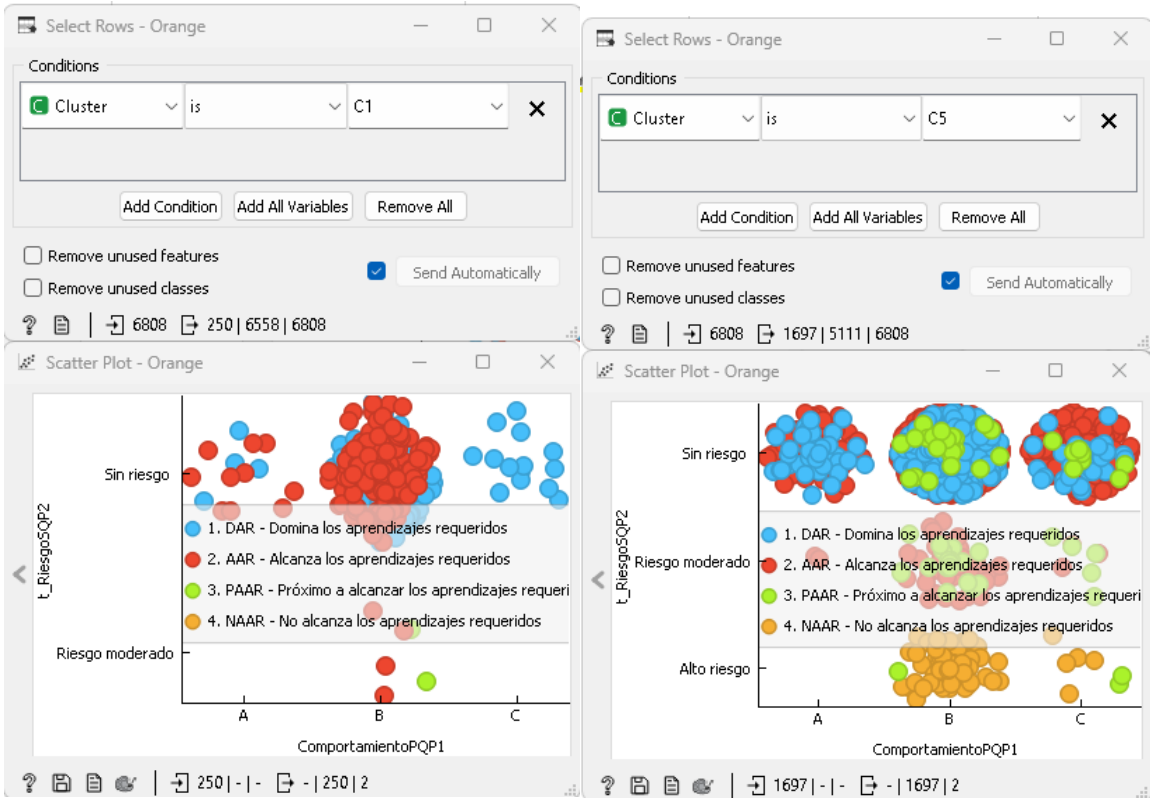


Figura 42: Exploración de los Clúster C1 y C5, con Select Rows mostrando el tamaño de cada clúster

De la gráfica que precede, se destaca la singularidad de C5, es un clúster grande en términos de instancias, tiene alumnos de alto riesgo con comportamientos B y C, pero también tiene muchos alumnos con los dos tipos de promedios más altos y con comportamientos C. En la configuración se utilizó el método de vinculación ponderado y una profundidad máxima de 7 niveles para el dendrograma, aunque la visualización se limitó a 5 niveles. C1 es un grupo pequeño en términos de número de instancias, no incluye alumnos en alto riesgo, pero si alumnos sin riesgo y con comportamiento C.

Además de FP-Growth, K-Means y el agrupamiento jerárquico, se empleó las reglas de asociación, cuyos fundamentos y métricas se documentaron en **2.5.3.4. Reglas de asociación**. Si bien, las métricas más recurridas para apreciar el aporte de las reglas, suelen ser el soporte y la confianza, en esta sección se hará énfasis en el interés. Esto porque cuando el soporte del consecuente B de una regla sea alto, la $Confianza\ (A \rightarrow B)$ siendo A el antecedente de la regla, no resultará un buen indicador.

Además de lo indicado respecto del soporte y confianza, otro aspecto por considerar es la gran cantidad de reglas que se pueden generar, las que se aproximan a $\frac{M \times 2 \times (M-1) \times 2}{2}$, siendo M la cantidad de atributos y 2 la cantidad (sólo ejemplo) de valores distintos por atributos. Es imperioso aplicar diversos filtros para obtener reglas representativas. Ver **Figura 43**.

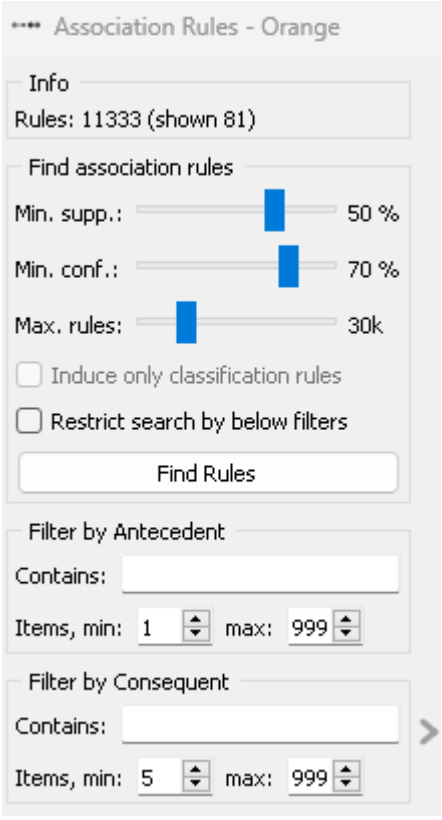


Figura 43: Parámetros para las reglas de asociación

Es posible filtrar por el Interés a reglas desde una hoja de cálculo, pero de momento no desde Orange. El Interés compara la frecuencia observada de una regla con la frecuencia esperada y determina si esta sucede por azar. Dada una regla $A \rightarrow B$, el Interés se obtiene acorde con la ecuación $\frac{\text{Soporte}(\text{Union}(A,\ B))}{\text{Soporte}\ (A) \times \text{Soporte}\ (B)}$. Si el resultado es menor que 1 la regla es de poco interés, si es igual que 1 es porque A y B son independientes y si es mayor que 1 es porque la regla es de alto interés y no es producto del azar.

Por ejemplo, en la **Tabla 33** la regla 1 tiene un soporte de 51.3% y una confianza de 83.3%, su interés es de 1.61, lo que denota que no es una regla producto del azar. La regla 1 dice que cuando la escolaridad alcanzada por la madre es la secundaria, entonces se reporta un 83.3% de probabilidad (confianza) de que ella sea la representante del alumno que se ha mantenido estudiando en la misma escuela, que no auto reporta dificultad en materias y no ha repetido algún año básico.

La regla 2 se refiere a representantes de ocupación como amas de casa que viven en una familia nuclear. La regla reporta un 84.6% de posibilidad (confianza) de que se trate de la madre de algún alumno que no reporta mayores problemas. La búsqueda de reglas que aporten utilidad a las escuelas puede suponer un buen tiempo de búsqueda manual, pero se considera que un importante filtro es obtener el valor de la métrica del Interés de cada regla.

Tabla 33: Reglas de asociación con sus respectivas métricas

	Antecedente	Consecuente	Soporte	Conf.	Interés
1	EscolaridadMadre= SECUNDARIA	ParentescoRepresentante=MAMÁ, EscolaridadRepresentante=SECUNDARIA, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.513	0.833	1.611
2	OcupacionRepresentante=AMA DE CASA, EstructuraFamiliar=NUCLEAR	OcupacionMadre=AMA DE CASA, ParentescoRepresentante=MAMÁ, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.506	0.846	1.576
3	EscolaridadRepresentante=SECUNDARIA	EscolaridadMadre=SECUNDARIA, ParentescoRepresentante=MAMÁ, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.513	0.813	1.575
4	OcupacionMadre=AMA DE CASA, EstructuraFamiliar=NUCLEAR	ParentescoRepresentante=MAMÁ, OcupacionRepresentante=AMA DE CASA, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.506	0.87	1.562
5	OcupacionRepresentante=AMA DE CASA, ProcedeDeOtraInstitucion=NO	OcupacionMadre=AMA DE CASA, ParentescoRepresentante=MAMÁ, EstructuraFamiliar=NUCLEAR, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.506	0.829	1.552
6	OcupacionMadre=AMA DE CASA, ParentescoRepresentante=MAMÁ	OcupacionRepresentante=AMA DE CASA, EstructuraFamiliar=NUCLEAR, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.506	0.834	1.548
7	OcupacionMadre=AMA DE CASA, ProcedeDeOtraInstitucion=NO	ParentescoRepresentante=MAMÁ, OcupacionRepresentante=AMA DE CASA, EstructuraFamiliar=NUCLEAR, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.506	0.841	1.535
8	ParentescoRepresentante=MAMÁ, OcupacionRepresentante=AMA DE CASA	OcupacionMadre=AMA DE CASA, EstructuraFamiliar=NUCLEAR, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO, DificultadAutoreportada=NO	0.506	0.804	1.519
9	OcupacionRepresentante=AMA DE CASA, DificultadAutoreportada=NO	OcupacionMadre=AMA DE CASA, ParentescoRepresentante=MAMÁ, EstructuraFamiliar=NUCLEAR, ProcedeDeOtraInstitucion=NO, ReprobadoRepetido=NO	0.506	0.786	1.519

3.4.4. Aprendizaje supervisado

3.4.4.1. Support Vector Machine

Como se documentó en **2.5.2.1. Máquinas de soporte vectorial (SVM)**, las SVM se pueden utilizar para tareas de clasificación como clasificación de vectores de soporte (SVC) y de regresión de vectores de soporte (SVR). Para ambos casos se aplicó el siguiente ajuste de parámetros:

- **SVM, Coste:** 1, aplica para clasificación y regresión (Orange, 2015ac). Un rango de valores sugeridos en la literatura es 2^{-10} … 2^{10} (Probst et al., 2018).
- **SVM, épsilon ε:** 0.10, aplica para clasificación y regresión. Un rango de valores sugeridos en la literatura es 0 a 10 (Probst et al., 2018).
- **Kernel:** Aunque SVM es un modelo lineal, es posible usar kernels para modelar datos linealmente no separables, se ha optado por la Función de Base Radial, RBF, con base en los resultados obtenidos en la revisión sistemática de la literatura que es parte de esta tesis y que además es el valor por defecto para SVM en Orange (Orange, 2015ac; Pincay-Ponce et al., 2023).
- **Gamma:** 1/k, siendo K el número de atributos. Un rango de valores sugeridos en la literatura es 2^{-10} … 2^{10} (Bartz et al., 2023).
- **Tolerancia numérica:** 0.0010 o 10^{-3}. El valor sugerido en la literatura es entre 10^{-5} a 10^{-1} (van Rijn & Hutter, 2018).
- **Límite de iteraciones permitidas:** 50, con ellos se guarda consonancia con el número de estimadores empleados en los métodos de ensamblado.

3.4.4.2. Análisis discriminante lineal, LDA

Como se indicó en la **sección 2.5.2.2** del marco teórico, LDA tiene usos en la clasificación, regresión y visualización de datos, que es como se empleó en la sección **3.2.3. Exploración de datos**. En el **Gráfico 26** los ejes de la proyección ingreso mensual familiar y número de hermanos se contrastan respecto del tipo de promedio anual. Los ingresos más altos coinciden con tipos de promedios más altos DAR y AAR, en tanto que, un mayor número de hermanos tiene más relación con AAR que de cierto modo significa aprobar, pero no con el tipo de promedios más altos que es DAR.

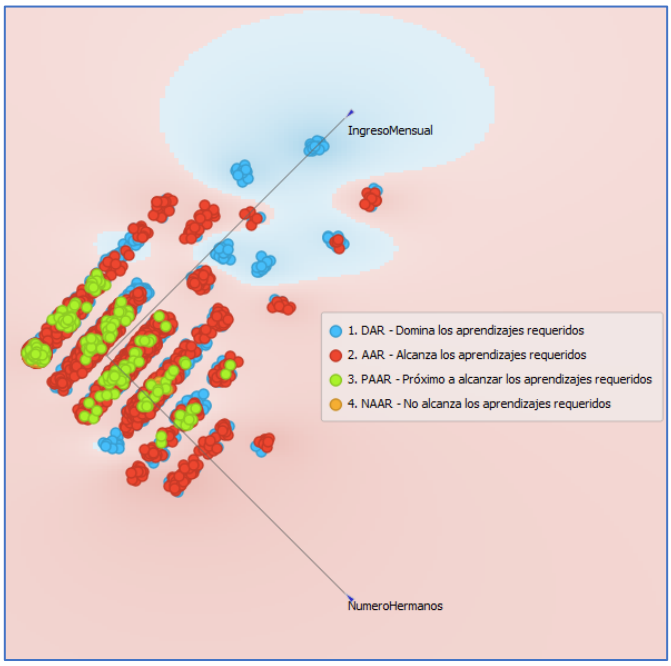


Gráfico 26: Proyección lineal que ilustra como un mayor ingreso mensual clasifica mejor a alumnos de promedios anuales DAR y AAR. Un mayor número de hermanos clasifica mejor a los alumnos con promedios AAR

El software Orange (2015m) incorpora la sugerencia automática de características que para el caso de LDA se denominan ejes. LDA selecciona las características con base en la puntuación de la exactitud para las tareas de clasificación o el MSE que reporte kNN en el caso de las tareas de regresión.

2.4.4.3. Método de Bayes

Respecto de Bayes Naïve, se ha cumplido con los supuestos no estadísticos y la preparación de los datos enfocada en resolver problemas de datos faltantes y discretizar valores numéricos cuando fuere posible, pues Orange ignora las columnas de tipo continuo (Orange, 2015). Bayes Naïve se utiliza para las tareas de clasificación tal cual se indicó en **2.5.2.3. Método de Bayes**.

3.4.4.4. KNN

El modelo KNN determina para cada x los k vecinos con la menor distancia a x, siendo posibles de calcular con distancia Euclidiana, Manhattan, Chebyshev y Mahalanobis (Ver sección **2.5.2.4. Vecino más cercano, KNN**). Para la regresión se utiliza la media

de los vecinos y para la clasificación, la predicción del modelo es la clase más frecuente observada en el vecindario. Tanto para la clasificación como regresión se empleó:

- Número de vecinos (k): 5, porque la literatura sugiere un valor impar (Bartz et al., 2023). Se realizó pruebas con valores impares mayores.
- Distancia (metric): Euclidiana.
- Peso: Uniforme.

3.4.4.5. Árbol de decisión, C4.5

En Orange (2015ae) se implementa como árbol de decisión al algoritmo C4.5, mismo que divide los datos en nodos mediante la ganancia de información para tareas de clasificación y con el MSE para las tareas de regresión. Para ambas tareas se empleó:

- **Forzar un árbol binario:** No, porque se generaría un árbol con mucha profundidad y un alto costo de procesamiento.
- **Mínimo número de instancias en las hojas:** 10, porque un análisis previo determinó gran diferencia respecto del balanceo de clases, aunque luego las clases minoritarias fueron balanceadas con relación a las mayoritarias.
- **Mínimo número de instancias en los nodos:** 10, por la misma razón precedente.
- **Profundidad del árbol:** 20 niveles, porque aun cuando se redujo la dimensionalidad con PCA se pueden obtener árboles muy profundos y ralentizar el entrenamiento del modelo.
- **Detener cuando la mayoría alcance:** 95%, que significa que se deje de dividir los nodos después de alcanzar un umbral de mayoría específico con los datos reducidos

2.4.4.6. Regresión lineal

En la sección **2.5.2.6. Regresión lineal** se ha documentado lo esencial respecto de lo requerido para su aplicación en las tareas de regresión de esta investigación. Los hiperparámetros se ajustaron como:

- **Intercepción de ajuste:** calculada por Orange mediante la ecuación $y = b_0 + b_1 x$ dónde b_0 es la intersección en Y y b_1 es la pendiente.
- Variante con **Regularización Ridge**: que emplea penalización con norma L2 por tanto es computacionalmente más eficiente que L1.

- Variante con **Regularización Lasso:** que emplea penalización con norma L1, menos eficiente que L2, pero aporta selección con selección de características porque los coeficientes que equivalen a 0 reflejan características de las que puede prescindirse.
- **Aplha:** 1, tanto para la Regularización Ridge como para la Regularización Lasso.

2.4.4.7. Regresión Logística

En la sección **2.5.2.7. Regresión logística** se ha documentado lo esencial respecto de lo requerido para su aplicación en las tareas de clasificación de esta investigación. Los hiperparámetros se ajustaron como:

- Variante con **Regularización Ridge**: que emplea penalización con norma L2.
- Variante con **Regularización Lasso**: que emplea penalización con norma L1.
- La **fuerza de costo:** C = 1.
- Balance class Distribution: No, porque los datos han sido previamente balanceados con SMOTE, tal cómo se indicó en **3.3.4. Aumento de datos**.

2.4.4.8. Métodos de aprendizaje en conjunto o ensamblados

La **Tabla 34** muestra la configuración empleada para cada hiperparámetro tanto de los modelos ensamblados por Boosting, como lo son ADA Boost, Gradient Boost, XGBoost, XGBoost Random Forest y Cat Boost y el ensamblado por Bagging como es el caso de Random Forest.

Tabla 34: Configuración empleada para los algoritmos de aprendizaje en conjunto

	ADA Boost	Gradient Boost	XGBoost.	XGBoost. Random Forest	Cat Boost	Random Forest
Numbers of estimators	50	50	50	50	50	50
Learning rate	0.5	0.5	0.5	0.5	0.5	
Fixed seed	5 (Default)					
Algorithm classification	SAMME.R					
Regression loss function	Exponential					
Limit Depth		20	20	20	10	20
Do not Split subsets <		5				5

	ADA Boost	Gradient Boost	XGBoost.	XGBoost. Random Forest	Cat Boost	Random Forest
Fraction of training instances		1	1	1		
Fraction of features for each tree			1	1	1	
Fraction of features for each level			1	1		
Fraction of features for each split			1	1		
Regularization			Lambda 1	Lambda 1	Lambda 1	
Replicable training		SI	SI	SI	SI	
Balance class Distribution						NO
# Attributes considered at each split						All (Default)
	4	**6**	**9**	**9**	**6**	**5**

Las configuraciones de la tabla precedente aplican para:

- Modelos de clasificación considerando notas intermedias
- Modelos de clasificación sin considerar notas intermedias
- Modelos de regresión sin considerar notas intermedias
- Modelos de clasificación con datos reducidos en dimensionalidad con PCA, Smote ponderado y sin considerar notas intermedias

2.4.4.9. Redes neuronales

En la sección **2.5.2.9. Redes neuronales** se ha documentado lo esencial respecto de lo requerido para su aplicación en las tareas de clasificación y regresión de esta investigación. Los hiperparámetros se ajustaron como:

- **Neuronas por capa oculta:** 3 capas ocultas con 3 neuronas cada una.
- **La función de activación ReLu** es la empleada en esta investigación, porque es la opción más popular, simple de implementar y tiene buen desempeño en una variedad de tareas predictivas (Zhang et al., 2022).

- **Solver**: Adam, porque es un optimizador estocástico basado en gradiente, pero puede ajustar automáticamente valores para actualizar los parámetros en función de las estimaciones adaptables de momentos de orden inferior, que a su vez son valores que resumen o sintetizan propiedades de la variable aleatoria.
- **Alfa:** L2 con un valor de 10^{-4} que ayuda a evitar el sobreajuste penalizando pesos con grandes magnitudes.
- **Número máximo de iteraciones**: 200.

2.4.4.10. Descenso del gradiente estocástico, SGD

Con base en lo documentado en la sección **2.5.2.10. Descenso de gradiente estocástico, SGD**, se han establecido los parámetros de esta técnica de aproximación estocástica para la optimización del entrenamiento de modelos, tal cual se muestra en la **Figura 44**.

Los resultados de SGD en las tareas de regresión de esta investigación fueron desfavorables en cuanto a las métricas MSE, RMSE, MAE y R^2. De modo especial en RMSE, métrica de la cual se espera un error pequeño porque es asociable con los puntos de los promedios de los alumnos, por ejemplo, si el RMSE es de 0.8 se interpreta como que a lo mucho el modelo tiene un error de 0.8/10 puntos, sin embargo, los valores obtenidos fueron siempre mayores que 10, por ende, imprácticos. Entonces, SGD se aplicó sólo para los siguientes casos de clasificación:

- Modelos de clasificación considerando notas intermedias.
- Modelos de clasificación sin considerar notas intermedias.
- Modelos de clasificación con datos reducidos en dimensionalidad con PCA, Smote ponderado y sin considerar notas intermedias.

La función pérdida para las tareas de clasificación con SGD que se finalmente se consideró fue Modified Huber (ver **Figura 44**), como el nombre lo sugiere, se trata de una modificación de la función de pérdida de Huber que combina lo mejor de MSE y MAE pero se orienta a tareas de regresión, la modificación está diseñada para las tareas de clasificación y a la mayor tolerancia a los valores atípicos respecto de las funciones de pérdida como Huber, Error medio cuadrado (MSE), Error absoluto medio (MAE), Pérdida de probabilidad logarítmica o Pérdida de Hinge o Bisagra (Guo et al., 2021).

La función pérdida para las tareas de clasificación especificada como Pérdida al Cuadrado o Squared Loss será ignorada porque se empleará SGD sólo para tareas de clasificación.

La regularización empleada fue la de Lasso, por las ventajas explicadas en las secciones de Regresión Logística y Lineal en el marco teórico, además de que, en caso de colinealidad, que una variable X1 sea combinación lineal de otra X2, Lasso tiende a reducir algunos de estos coeficientes de entidad a cero.

Respecto de la tasa de aprendizaje n, Learning Rate, puede ser constante o decaer gradualmente, lo que se estima que es óptimo y viene dado por $n^{(t)} = \frac{1}{\infty(t_0+t)}$, dónde donde t es el paso y existen un $\text{Número de instancias} * \textit{Número de iteraciones}$ de pasos. t_0 se determina con base a una heurística propuesta por Léon Bottou (2018) de modo que las actualizaciones iniciales esperadas sean comparables con el tamaño esperado de los pesos.

Con la tasa de aprendizaje óptima, el algoritmo alcanza el mínimo en un corto período de tiempo con un número considerablemente menor de épocas. Este detalle es importante porque si es demasiado grande, el algoritmo puede pasar por alto el mínimo local y rebasarlo. Si es demasiado pequeño, podría aumentar el tiempo total de cálculo en gran medida.

El criterio se evalúa una vez por época, es decir, por el conjunto de pasos que el algoritmo SGD o incluso una red neuronal, ha completado durante el entrenamiento o después de haber llegado al número máximo de iteraciones que se ha fijado en 200. La mejora se evalúa con tolerancia, que es un valor del tamaño del paso cercano a 0, es un criterio de parada.

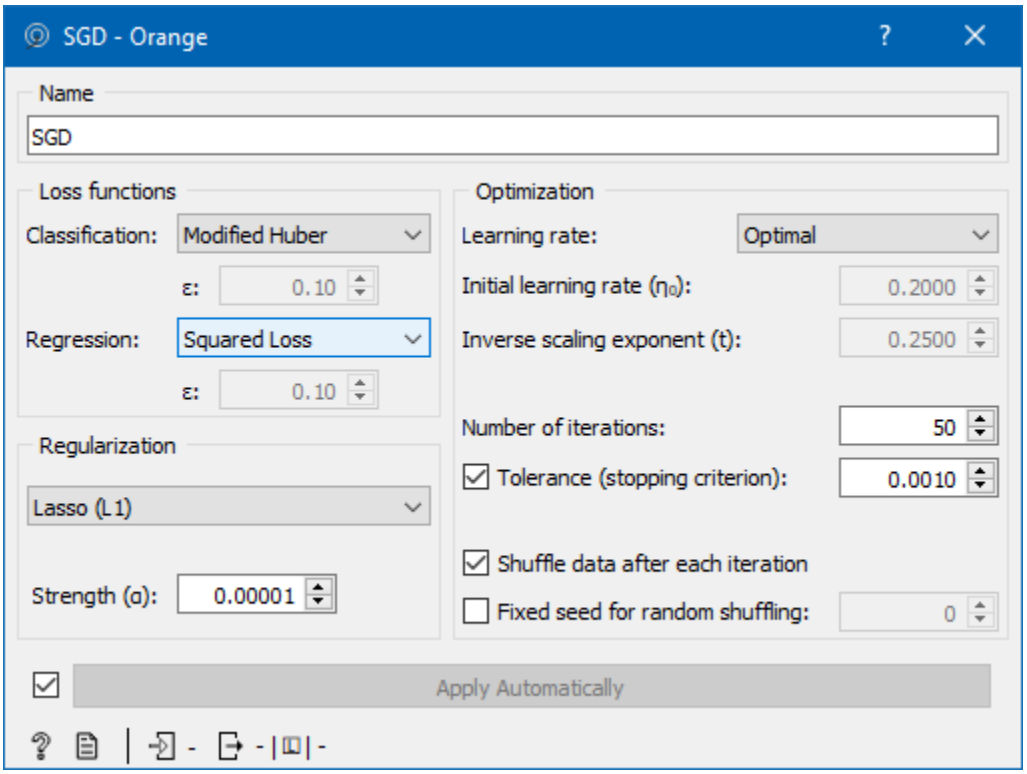


Figura 44: Establecimiento de hiperparámetros para SGD.

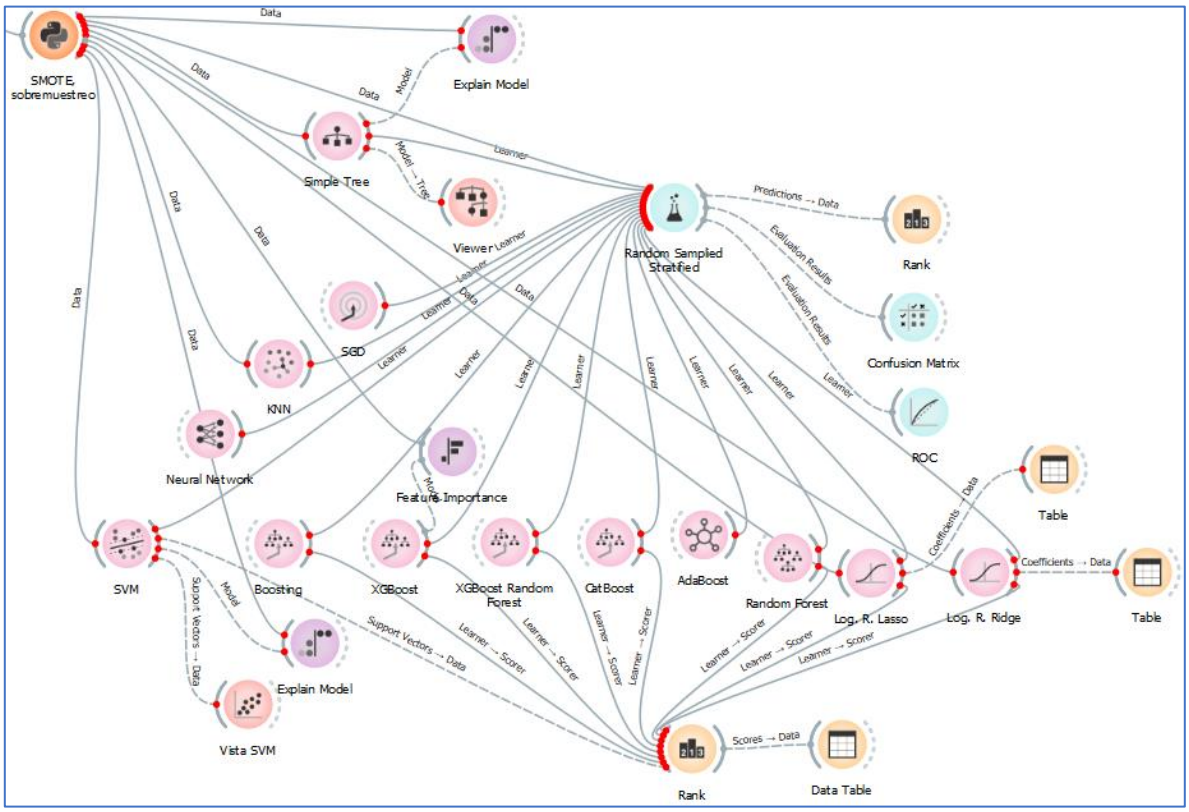


Figura 45: Vista parcial del modelo para las tareas de clasificación, tanto incluyendo las calificaciones como no incluyéndolas. La diferencia entre un tipo y otro se establece en la selección de columnas con el widget Select Column (2015x)

Para fines de clasificar a las características, atributos o variables disponibles, se considerará como variables dependientes o en términos de minería de datos, variables de respuestas o columnas objetivo al Promedio anual, tanto cuantitativo para los problemas de regresión y cualitativo para los problemas de clasificación.

3.5. Fase 5. Evaluación

En esta Fase 5 se comparan y validan los resultados de las métricas producidos por los algoritmos modelados, de cierto modo se revisa el proceso de minería de datos que llevó hasta este punto y se revisan los pasos por seguir, por ejemplo, repetir pasos de fases anteriores o abrir nuevas líneas de investigación.

Los procesos de entrenamiento y prueba de todos los modelos empleados en la investigación se ejecutaron en un computador Intel Core i3 de quinta generación de 2 GHz de velocidad y 4 núcleos, disco SDD de 456 GB y 16 GB de memoria RAM.

En esta sección no se hará referencia a la evaluación de los modelos no supervisados o descriptivos, porque lo mínimo necesario para ellos fue explicado en la Fase 4.

3.5.1. Modelos de clasificación considerando notas intermedias

En las siguientes tres tablas se muestran los resultados para las métricas AUC, CA, F1, Precision, Recall y Specificity considerando como características para el análisis a las calificaciones intermedias obtenidas por los alumnos. La inclusión de dichas calificaciones ilustra sobre la importancia de los momentos para el rendimiento académico, pues es conocido que las calificaciones se suscitan en diversos momentos del año básico de estudio. El muestreo configurado en el widget Test and Score (2015ad), es aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos que se fijó.

En las tablas concernientes en esta Fase de evaluación se han coloreado en tonos más verdes a los valores más altos o mejores y en tonos rojos a los más bajos o peores. Se invirtió el sentido de la coloración en el caso de los tiempos de entrenamiento y prueba cuando se lo requiera, así los tiempos más altos se pintaron en rojo. Como ejemplo, en la **Tabla 35** el tiempo de entrenamiento de la Regresión Logística con la Regularización Ridge (cresta) tomó aproximadamente 41 de los 57 minutos ocupados por todos los modelos, con la Regularización de Lasso en cambio el tiempo ocupado fue de poco más de 3 minutos (195.95 segundos). Luego, las matrices de confusión que se muestran en la **Tabla 38** detallan la Exactitud alcanzada para cada clase por cada modelo (Orange, 2015b).

Tabla 35: Métricas porcentuales resultantes para la clasificación de todas las clases, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
AdaBoost	164.79	9.50	1.000	0.995	0.995	0.995	0.995	0.998
Grad. XGBoost	224.62	2.66	1.000	0.994	0.994	0.994	0.994	0.998
Grad. Boosting	485.37	4.31	1.000	0.994	0.994	0.994	0.994	0.998
Neural Network	178.57	5.08	1.000	0.993	0.993	0.993	0.993	0.998
Random Forest	19.42	3.64	1.000	0.993	0.993	0.993	0.993	0.998
Grad. CatBoost	731.58	8.49	1.000	0.992	0.992	0.992	0.992	0.997
Grad. XGBoost Random Forest	265.75	2.19	1.000	0.987	0.987	0.987	0.987	0.996
Log. Regression, Ridge	1025.97	2.94	1.000	0.986	0.986	0.986	0.986	0.995
Log. Regression, Lasso	195.95	2.74	0.999	0.984	0.984	0.984	0.984	0.995
SGD	20.57	4.54	0.987	0.980	0.980	0.980	0.980	0.993
kNN	9.97	181.16	0.996	0.969	0.969	0.969	0.969	0.990

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
Tree Balanced	10.60	0.00	0.985	0.963	0.963	0.964	0.963	0.988
SVM	104.25	18.09	0.986	0.903	0.903	0.906	0.903	0.968
TOTAL (segundos)	**3437.40**	**245.32**						
TOTAL (minutos)	**57.29**	**4.09**						

Tabla 36: Métricas porcentuales resultantes para la clasificación de la clase <No alcanza los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
AdaBoost	164.79	9.50	1.000	1.000	1.000	1.000	1.000	1.000
Grad. Boosting	485.37	4.31	1.000	1.000	1.000	1.000	1.000	1.000
Grad. CatBoost	731.58	8.49	1.000	1.000	1.000	0.999	1.000	1.000
Grad. XGBoost	224.62	2.66	1.000	1.000	1.000	1.000	1.000	1.000
Grad. XGBoost Random Forest	265.75	2.19	1.000	1.000	1.000	1.000	1.000	1.000
kNN	9.97	181.16	1.000	1.000	1.000	1.000	1.000	1.000
Log. Regression, Lasso	195.95	2.74	1.000	1.000	0.999	0.999	1.000	1.000
Log. Regression, Ridge	1025.97	2.94	1.000	1.000	1.000	1.000	1.000	1.000
Neural Network	178.57	5.08	1.000	1.000	1.000	1.000	1.000	1.000
Random Forest	19.42	3.64	1.000	1.000	1.000	1.000	1.000	1.000
SGD	20.57	4.54	0.997	0.996	0.992	0.984	1.000	0.994
SVM	104.25	18.09	1.000	1.000	0.999	0.998	1.000	0.999
Tree Balanced	10.60	0.00	0.999	0.999	0.997	0.995	1.000	0.998
TOTAL (segundos)	**3437.40**	**245.32**						
TOTAL (minutos)	**57.29**	**4.09**						

Tabla 37: Métricas porcentuales resultantes para la clasificación de la clase <Próximo a alcanzar los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
Random Forest	19.42	3.64	1.000	0.999	0.998	0.998	0.999	0.999
AdaBoost	164.79	9.50	1.000	0.999	0.998	0.998	0.998	0.999
Grad. XGBoost	224.62	2.66	1.000	0.999	0.998	0.997	0.999	0.999
Neural Network	178.57	5.08	1.000	0.999	0.998	0.997	0.999	0.999
Grad. Boosting	485.37	4.31	1.000	0.999	0.998	0.997	0.998	0.999
Grad. CatBoost	731.58	8.49	1.000	0.999	0.997	0.996	0.999	0.999
Log. Regression, Ridge	1025.97	2.94	1.000	0.998	0.996	0.994	0.998	0.998
Log. Regression, Lasso	195.95	2.74	1.000	0.997	0.995	0.993	0.997	0.998
Grad. XGBoost Random Forest	265.75	2.19	1.000	0.997	0.994	0.990	0.998	0.997
SGD	20.57	4.54	0.995	0.996	0.993	0.992	0.993	0.997

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
kNN	9.97	181.16	0.999	0.995	0.991	0.984	0.998	0.995
Tree Balanced	10.60	0.00	0.965	0.979	0.958	0.977	0.939	0.993
SVM	104.25	18.09	0.998	0.975	0.948	0.977	0.919	0.993
TOTAL (segundos)	**3437.40**	**245.32**						
TOTAL (minutos)	**57.29**	**4.09**						

A continuación, se hacen algunas apreciaciones en especial relación con las clases que representan eventuales clasificaciones de nuevos alumnos en riesgo, es decir, aquellos que <No alcanzan los aprendizajes requeridos> o <Están próximos a alcanzar los aprendizajes requeridos>:

- Con excepción del Descenso de gradiente estocástico (SGD) y C4.5 (Tree), todos los modelos alcanzan una exactitud (CA) y Recall del 100% para la clasificación de alumnos que <No alcanzan los aprendizajes requeridos> tal cual muestra la **Tabla 36**. Esto tiene como explicación el haber contemplado a las calificaciones las cuales determinan el promedio final del alumno y luego de ellas recién figuran las incidencias de los factores socioeconómicos.
- Con excepción de C4.5 (Tree) y SVM, todos los modelos alcanzan una Exactitud (CA) y Recall superior o igual al 99% para la clasificación de alumnos que están <Próximos a alcanzar los aprendizajes requeridos> tal cual muestra la **Tabla 37**.
- Con excepción de kNN, C4.5 (Tree) y SVM, todos los modelos alcanzan una Exactitud (CA) y Recall superior o igual al 98% para la clasificación promedio de todas las clases, tal cual muestra la **Tabla 35**.
- Los tiempos de prueba de todos los modelos son relativamente breves, siendo el mayor kNN con algo más de tres minutos.
- En tales circunstancias la elección de un modelo puede depender de los marcados tiempos de entrenamiento, que no favorecen a la Regresión Logística con regularización de Ridge o a los métodos de ensamblado CatBoost, Boosting y XG Boosting.

Tabla *38*: Matrices de confusión de los modelos para todas las clases, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Matriz de confusión

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	96.5 %	2.6 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	3.5 %	92.1 %	2.3 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	5.3 %	97.7 %	0.7 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	99.3 %	7888
Σ	7956	8072	7576	7946	31550

Matriz de confusión

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	98.8 %	0.9 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	1.2 %	99.1 %	0.2 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.8 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7919	7846	7895	7890	31550

Tree (C4.5)

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	98.4 %	1.3 %	0.1 %	0.0 %	7890
2. AAR - Alcanza...	1.6 %	98.6 %	0.4 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.5 %	0.1 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	99.9 %	7888
Σ	7905	7838	7916	7891	31550

Random Forest

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	98.7 %	0.9 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	1.3 %	99.0 %	0.4 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.6 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7918	7835	7907	7890	31550

Gradiente CatBoost

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	97.5 %	2.9 %	0.1 %	0.0 %	7890
2. AAR - Alcanza...	2.5 %	96.9 %	0.8 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.2 %	0.1 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	99.9 %	7888
Σ	7854	7866	7935	7895	31550

Gradiente Boosting

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	97.8 %	2.6 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	2.2 %	97.3 %	0.6 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.4 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7861	7872	7926	7891	31550

Regresión Logística, Lasso

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	97.9 %	1.8 %	0.2 %	0.0 %	7890
2. AAR - Alcanza...	2.1 %	98.1 %	1.1 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	98.8 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7899	7788	7974	7889	31550

Regresión Logística, Ridge

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	98.9 %	0.9 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	1.1 %	99.0 %	0.3 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.7 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7909	7852	7899	7890	31550

XGBoost Random Forests

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	99.0 %	0.8 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	1.0 %	99.0 %	0.2 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.8 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7906	7863	7891	7890	31550

XGBoost

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	97.3 %	2.4 %	0.1 %	0.5 %	7890
2. AAR - Alcanza...	2.7 %	96.7 %	0.6 %	0.8 %	7885
3. PAAR - Próxim...	0.0 %	0.9 %	99.2 %	0.2 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.1 %	98.6 %	7888
Σ	7875	7823	7859	7993	31550

Matriz de confusión — Ada Boost

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	82.8 %	10.2 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	15.2 %	85.8 %	1.8 %	0.0 %	7885
3. PAAR - Próxim...	2.0 %	4.0 %	98.2 %	0.3 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	99.7 %	7888
Σ	8603	7504	7532	7911	31550

Matriz de confusión — SGD

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	94.6 %	6.0 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	5.4 %	93.9 %	1.5 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	98.4 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7843	7811	8005	7891	31550

SVM

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	98.8 %	0.9 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	1.2 %	99.0 %	0.4 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	0.1 %	99.6 %	0.0 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	100.0 %	7888
Σ	7913	7842	7904	7891	31550

KNN

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	94.6 %	3.4 %	0.0 %	0.0 %	7890
2. AAR - Alcanza...	5.4 %	79.5 %	0.7 %	0.0 %	7885
3. PAAR - Próxim...	0.0 %	17.1 %	99.3 %	0.5 %	7887
4. NAAR - No ...	0.0 %	0.0 %	0.0 %	99.5 %	7888
Σ	8006	9315	6300	7929	31550

Neural Network — Naïve Bayes

A continuación, se hacen algunas apreciaciones respecto de las matrices de confusión y los resultados de Exactitud que muestran:

- Naïve Bayes es un buen clasificador para todas las clases, pero lo es en menor medida para aquellos alumnos que <Alcanzan los aprendizajes requeridos>, allí obtiene una Exactitud del 79.5%, la diferencia respecto del 100% la distribuye con un 17.1% de alumnos clasificados incorrectamente como <Próximos a alcanzar los aprendizajes requeridos> y un 3.4 % clasificados incorrectamente como que <Dominan los aprendizajes requeridos>.
- SVM es un buen clasificador para los alumnos que se encuentran en riesgo, es decir, los <Próximos a alcanzar los aprendizajes requeridos> y los que <No alcanzan los aprendizajes requeridos, allí obtiene unas exactitudes superiores al 98%. Los errores de clasificación se concentran en las clases como <Próximos a alcanzar los aprendizajes requeridos> y <Domina los aprendizajes requeridos>, si bien estas clases no representan un riesgo de problemas de aprobación del año básico que curse el alumno. En todo caso, las clasificaciones erróneas ocurren con la clase más cercana.

En la **Tabla 39** se han ordenado los valores de la tasa de ganancia o información mutua de mayor a menor, para distinguir el poder de predicción de cada

característica con respecto de la variable dependiente <Promedio Anual>. La columna Nº muestra la cantidad de valores distintos que hay en cada característica de tipo texto. A diferencia de la **Tabla 28,** en esta tabla se han considerado y se muestran 62 características, además, el Gain Ratio reflejado es la media de las reportada por los modelos empleados y que soportan dicha métrica: Gradiente Boosting, XG Boosting, XG Boosting Random Forest, CatBoost, Random Forest, Regresión Logística con regularización Lasso y Regresión Logística con regularización Ridge.

Tabla 39: Ganancia e información mutua de cada característica con respecto del promedio anual

Nº	Característica	#	Gain ratio	Nº	Característica	#	Gain ratio
1	t_RiesgoPQP3	3	0.779	**32**	LuzElectrica	2	0.175
2	t_RiesgoSQP1	3	0.755	**33**	Computador	2	0.173
3	t_RiesgoSQP3	3	0.720	**34**	ProyEscSQP1	4	0.171
4	t_RiesgoSQP2	3	0.712	**35**	ComportamientoPQP2	3	0.159
5	t_RiesgoPQP2	3	0.707	**36**	ReprobadoRepetido	2	0.155
6	t_QUI2	4	0.703	**37**	ProyEscPQP3	3	0.152
7	t_QUI1	4	0.662	**38**	SBU	9	0.140
8	t_RiesgoPQP1	3	0.621	**39**	EscolaridadPadre	3	0.139
9	t_SQP1	4	0.609	**40**	EstadoCivilMadre	7	0.132
10	t_PQP3	4	0.598	**41**	AnioLlegada		0.129
11	t_SQP3	4	0.590	**42**	NumeroHermanos		0.127
12	t_SQP2	4	0.583	**43**	Ocup. Representante	46	0.121
13	t_PQP2	4	0.560	**44**	aniosRetraso		0.121
14	t_PQP1	4	0.459	**45**	Alcantarillado	2	0.108
15	Discapacidad	2	0.239	**46**	Ocup. Padre	60	0.106
16	t_familiaReconstruida	2	0.219	**47**	Ocup. Madre	42	0.104
17	ProyEscSQP2	4	0.218	**48**	Enfermedad	20	0.099
18	ComportamientoPQP3	3	0.206	**49**	ProcedeDeOtraInstitucion	2	0.095
19	ComportamientoPQ	3	0.198	**50**	distanciaKM		0.094
20	ComportamientoSQP3	3	0.198	**51**	ParentescoRepresentante	6	0.090
21	anioBasico		0.197	**52**	EscolaridadMadre	3	0.085
22	ComportamientoSQP2	3	0.197	**53**	EstructuraFamiliar	3	0.080
23	ComportamientoSQ	3	0.196	**54**	EstadoCivilPadre	7	0.077
24	ProyEscPQP2	3	0.192	**55**	Sexo	2	0.075
25	ProyEscSQ	4	0.192	**56**	DificultadAutoreportada	2	0.068

Nº	Característica	#	Gain ratio	Nº	Característica	#	Gain ratio
26	t_anioBásico	6	0.188	**57**	ComportamientoPQP1	3	0.066
27	ComportamientoSQP1	3	0.187	**58**	EscolaridadRepresentante	3	0.056
28	ProyEscSQP3	4	0.186	**59**	AguaPotable	2	0.055
29	TVCable	2	0.186	**60**	Celular	2	0.052
30	ProyEscPQP1	4	0.185	**61**	Materia abrev.	7	0.049
31	ProyEscPQ	3	0.177	**62**	Internet	2	0.048

Como lectura general de la tabla que precede, se observa que la calificación del Segundo Quimestre incide más que la del Primer Quimestre sobre el promedio anual, como se sabe que estos promedios derivan de los promedios parciales, entre los números 1 al 8 que representan el orden de las características se muestra el poder de predicción de las notas parciales por medio de los riesgos que pueden ser de tipo alto, moderado o sin riesgos, como se explicó en la **Tabla 31:** Ficha de las nuevas características predictoras y de respuestas agregadas para el análisis.

En el lugar 15 y 16 aparecen factores socioeconómicos como la discapacidad o no del alumno, vivir en una familia reconstruida o no. Sus habilidades sociales aparecen de modo indirecto en el orden 17 a través de las calificaciones de los proyectos escolares, así como su comportamiento que se evidencia en los órdenes siguientes.

Si bien, la tabla anterior ilustra de la importancia o poder de predicción de las características, para conocer los valores inmersos en estos resultados es necesario revisar a cada modelo. Por ejemplo, al final de la **Tabla 38** se observa que el modelo Naïve Bayes alcanza un 79.5% de Exactitud en la clasificación de los alumnos con promedios de la clase <Próximo a alcanzar los aprendizajes requeridos>, conocido aquello, se puede emplear un nomograma que nos permita explorar a detalle dicho valor.

Un nomograma en Orange Data Mining permite la representación visual interactiva de clasificadores como Naive Bayes y la Regresión Logística (Orange, 2015r). Ofrece una visión de la estructura de los datos de entrenamiento y los efectos de todas o de las n mejores características en las probabilidades de ocurrencia de una clase, para el ejemplo de la **Figura 46** se ha tomado la clase <Próximo a alcanzar los aprendizajes requeridos>. La probabilidad para la clase objetivo se calcula mediante el principio 1 contra todos, porque la clase tiene cuatro posibles tipos de promedios. En tal caso se deben normalizar las probabilidades.

Para hacer una predicción, la contribución de cada atributo se puntúa, luego, tales puntuaciones individuales se suman para determinar la probabilidad. En el nomograma de la figura dada la personalización de valores de cada característica se ha obtenido una suma de -30 puntos y una probabilidad del 4% de que el alumno esté <Próximo a alcanzar los aprendizajes requeridos> si: (a) su comportamiento en el segundo quimestre parcial 2 es B. (b) su calificación en el proyecto escolar del primer quimestre parcial 1 es MB. (c) la ocupación de la madre es Docente… entre otras posibilidades que el nomograma permite explorar de manera interactiva.

Respecto de la Regresión Logística con la Regularización de Lasso, en la misma **Tabla 38**, se observa que alcanza un 98% de Exactitud (Classification Accuracy, CA) sobre el promedio de todas las clases. El 98% es favorable si se considera que las clases están balanceadas con SMOTE, sin embargo, resulta abstracto si no se conoce a detalle los valores de las multiples características que lo ocasionan. En ese escenario, resulta útil conocer los coeficientes de Logit del modelo que nos reporta Orange, además de recordar que cuanto más extremo sea el coeficiente en lo positivo o negativo, mayor será la dependencia marginal en la variable objetivo.

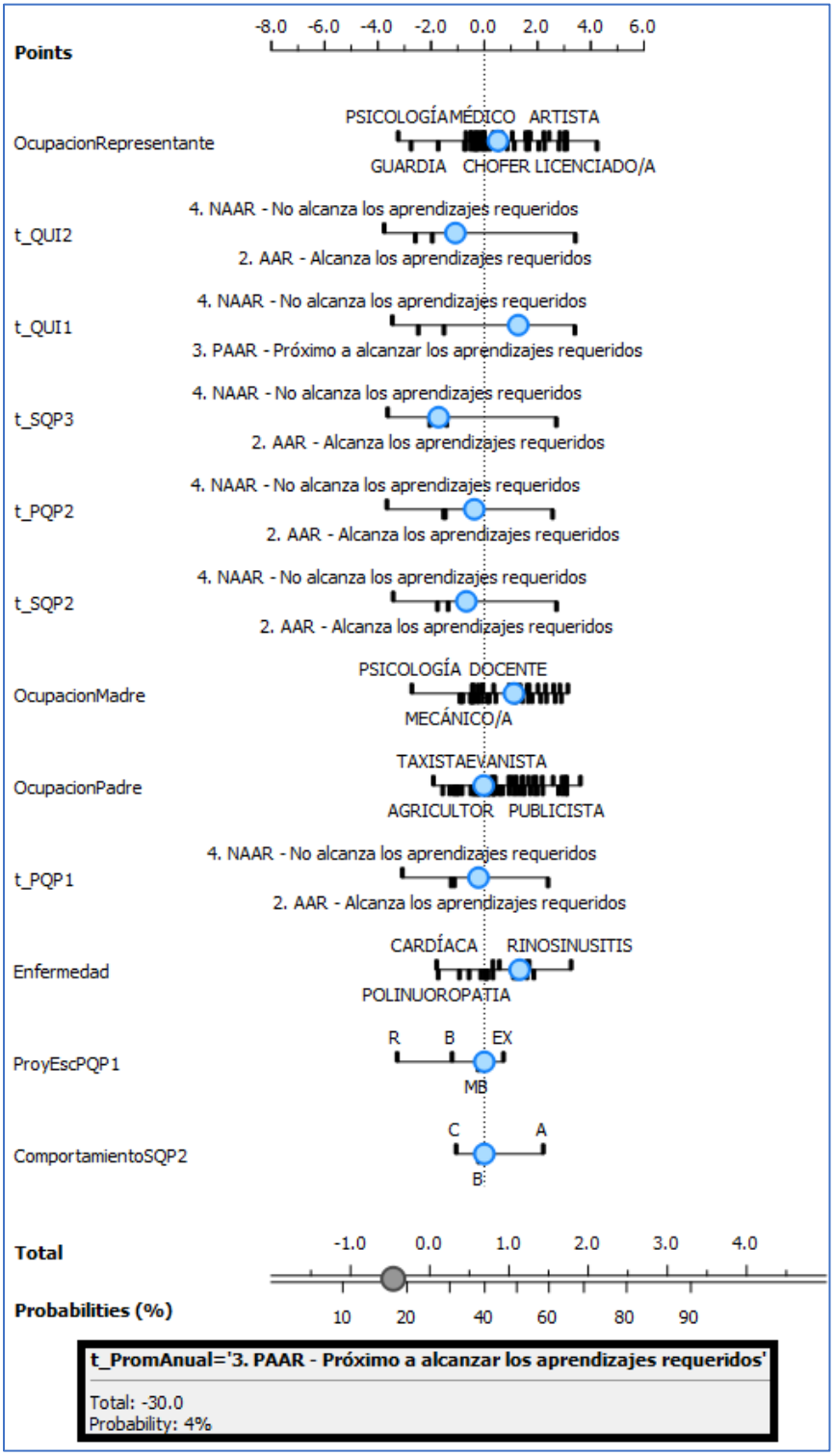


Figura 46: Nonograma de ejemplo para evaluar la probabilidad de que un alumno alcance un promedio de la clase <Próximo a alcanzar los aprendizajes requeridos>

En la **Tabla 40**, **Tabla 41** y la **Tabla 42** se lista las características ordenadas y coloreadas en tonos de verdes a rojos según su mayor a menor importancia, además de los valores que tomaron las características en tales escenarios. Por optimización de espacios se eliminó de las tablas a aquellas filas con coeficientes cercanos a 0 e iguales a 0 porque tales valores resultan poco influyentes y en ese sentido, esa es una ventaja que ofrece la regularización de Lasso y por lo que la literatura señala es preferida por sobre la regularización de Ridge que, aunque se acerca a cero, no llega a cero. Lasso, por tanto, refleja una selección de características y de sus valores.

Tabla 40: Coeficientes Logit de la Regresión Logística con la Regularización de Lasso, sobre la clase <1. Domina los aprendizajes requeridos>

	Características	Coef.		Características	Coef.
1	t_QUI1=1. DAR - Domina los aprendizajes requeridos	3.432	44	Ocup. Padre=MARINERO	-0.363
2	Ocup. Representante=PESCADOR	2.327	45	anioBasico	-0.381
3	t_RiesgoSQP1=Sin riesgo	2.078	46	t_anioBásico=6. Sexto	-0.392
4	Ocup. Padre=GUARDIA	2.060	47	Materia abrev.=Lenguaje	-0.421
5	t_QUI2=1. DAR - Domina los aprendizajes requeridos	2.059	48	t_PQP1=2. AAR - Alcanza los aprendizajes requeridos	-0.428
6	ComportamientoSQP2=A	1.464	49	TVCable=SI	-0.438
7	ComportamientoPQP1=C	1.151	50	Ocup. Representante=ESTUDIANTE	-0.465
8	ProyEscSQP2=MB	1.131	51	EscolaridadMadre=PRIMARIA	-0.471
9	EstadoCivilPadre=DIVORCIADO/A	1.036	52	t_familiaReconstruida=NO	-0.473
10	t_RiesgoPQP3=Sin riesgo	1.010	53	Ocup. Madre=ESTUDIANTE	-0.483
11	t_SQP3=1. DAR - Domina los aprendizajes requeridos	0.976	54	ParentescoRepresentante=PAPÁ	-0.513
12	t_SQP1=1. DAR - Domina los aprendizajes requeridos	0.858	55	EscolaridadPadre=SUPERIOR	-0.523
13	Enfermedad=RINOSINUSITIS	0.843	56	t_PQP2=2. AAR - Alcanza los aprendizajes requeridos	-0.552
14	SBU=3.0 SBU	0.741	57	ProyEscPQ=MB	-0.559
15	t_RiesgoPQP2=Sin riesgo	0.728	58	ComportamientoPQP1=B	-0.564
16	t_RiesgoSQP2=Sin riesgo	0.724	59	t_PQP2=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	-0.616
17	Ocup. Representante=IMPULSADOR/A	0.695	60	t_SQP3=4. NAAR - No alcanza los aprendizajes requeridos	-0.647
18	Ocup. Representante=LICENCIADO/A	0.626	61	t_QUI2=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	-0.668
19	EstadoCivilMadre=UNIÓN LIBRE	0.616	62	t_RiesgoPQP1=Riesgo moderado	-0.831
20	ComportamientoPQ=A	0.601	63	AguaPotable=NO	-0.873
21	Ocup. Padre=EMPLEADO PRIVADO	0.600	64	SBU=5.0 SBU	-0.896
22	Enfermedad=CONVULSIONES	0.597	65	Enfermedad=BRONCONEUMONÍA	-0.915

	Características	Coef.		Características	Coef.
23	ProyEscSQ=EX	0.586	66	ComportamientoSQP1=A	-0.953
24	t_PQP2=1. DAR - Domina los aprendizajes requeridos	0.566	67	Ocup. Padre=RECAUDADOR	-0.995
25	t_RiesgoPQP1=Sin riesgo	0.515	68	t_SQP2=2. AAR - Alcanza los aprendizajes requeridos	-1.014
26	t_PQP1=1. DAR - Domina los aprendizajes requeridos	0.503	69	t_PQP3=2. AAR - Alcanza los aprendizajes requeridos	-1.041
27	Ocup. Madre=COMERCIANTE	0.498	70	Materia abrev.=Matemática	-1.073
28	Materia abrev.=Ed. F	0.489	71	t_SQP1=2. AAR - Alcanza los aprendizajes requeridos	-1.076
29	Ocup. Padre=EMPLEADO/A	0.425	72	Materia abrev.=Inglés	-1.078
30	Ocup. Madre=LICENCIADO/A	0.424	73	DificultadAutoreportada=SI	-1.103
31	Ocup. Padre=OBRERO/A	0.416	74	Discapacidad=SI	-1.130
32	Celular=NO	0.409	75	SBU=1.0 SBU	-1.140
33	Ocup. Padre=ABOGADO/A	0.403	76	EstructuraFamiliar=MONOPARENTAL	-1.272
34	SBU=2.0 SBU	0.400	77	Ocup. Representante=ASESORA DE VENTAS	-1.273
35	t_SQP2=1. DAR - Domina los aprendizajes requeridos	0.349	78	Materia abrev.=Arte, Cultura	-1.340
36	Alcantarillado=SI	0.327	79	Ocup. Padre=MECÁNICO/A	-1.382
37	Ocup. Representante=OBRERO/A	0.314	80	t_QUI1=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	-1.526
38	Ocup. Representante=ARTISTA	0.311	81	ComportamientoSQP2=C	-1.831
39	t_anioBásico=3. Tercero	0.305	82	t_RiesgoSQP3=Alto riesgo	-2.133
40	Ocup. Padre=ARTISTA	0.305	83	t_anioBásico=2. Segundo	-2.415
41	Ocup. Representante=PSICOLOGÍA	-0.306	84	t_QUI1=2. AAR - Alcanza los aprendizajes requeridos	-2.579
42	ProyEscSQP2=B	-0.329	85	Ocup. Madre=ASESORA DE VENTAS	-3.246
43	EstadoCivilPadre=UNIÓN LIBRE	-0.330	86	Enfermedad=ANEMIA	-3.674
			87	t_QUI2=2. AAR - Alcanza los aprendizajes requeridos	-3.949

Como lectura general de la **Tabla 40** se observa que en los casos de alumnos que <Dominan los aprendizajes requeridos> ha influido que sus calificaciones del primer quimestre sean las más altas (Nº 1), que su representante sea de profesión pescador (Nº 2), guardia de seguridad (Nº 4) o impulsador de ventas (Nº 17), que en general las calificaciones parciales no hayan sido de riesgos (Nº 3, 10, 15, 16…), la materia de Educación Física (Nº 28) figura como la más asociada a este tipo de promedios, que en el comportamiento registre altas valoraciones (Nº 6) aunque haya empezado con comportamientos mejorables en el primer parcial del primer quimestre (Nº 7), que en los proyectos escolares registren altas valoraciones (Nº 8, 23), que el ingreso familiar ronde los tres salarios básicos unificados, entre otros. Es de indicar que todos

los datos combinados en la **Tabla 40** revisten de interés, porque la regularización de Lasso aplicada al conjunto de datos implica de cierto modo una selección de mejores características y de valores tal cual se mostró en la tabla.

Los coeficientes de la **Tabla 40**, llamados de Logit, aún son ilustrativos, pues la probabilidad para 4 clases (K = 4) como es el caso de los cuatro tipos de posibles promedios de los estudiantes se define por las siguientes fórmulas:

$$p_1 = \frac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}+e^{z_4}} \qquad p_2 = \frac{e^{z_2}}{e^{z_1}+e^{z_2}+e^{z_3}+e^{z_4}}$$

$$p_3 = \frac{e^{z_3}}{e^{z_1}+e^{z_2}+e^{z_3}+e^{z_4}} \qquad p_4 = \frac{e^{z_4}}{e^{z_1}+e^{z_2}+e^{z_3}+e^{z_4}}$$

Luego, todas las probabilidades deben sumar 1 y la probabilidad más alta determina la clase, es decir:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Con respecto de z_1, z_2, z_3 y z_4, se definen como:

$$z_i = \alpha_{i,1}x_1 + \alpha_{i,2}x_2 + \cdots + \alpha_{i,n}x_n + \beta_i$$

z_i es el resultado de la combinación lineal de los valores de las características o variables independientes x multiplicada por los parámetros alfa $\propto$, sumados al sesgo beta β. Esto se repite para cada clase posible.

Tabla 41: Coeficientes Logit de la Regresión Logística con la Regularización de Lasso, sobre la clase <3. Próximo a alcanzar los aprendizajes requeridos>

	Características	Coef.		Características	Coef.
1	t_QUI2=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	6.327	34	ComportamientoPQP3=A	-1.013
2	EscolaridadPadre=SECUNDARIA	3.042	35	Materia abrev.=CC NN	-1.022
3	t_QUI1=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	1.881	36	Enfermedad=NINGUNA	-1.027
4	EscolaridadMadre=PRIMARIA	1.707	37	t_anioBásico=4. Cuarto	-1.041
5	ProyEscPQ=EX	1.503	38	EscolaridadMadre=SECUNDARIA	-1.042
6	EscolaridadPadre=SUPERIOR	1.411	39	ComportamientoPQ=A	-1.043
7	Ocup. Representante=GUARDIA	1.291	40	Ocup. Padre=COMERCIANTE	-1.077

	Características	Coef.		Características	Coef.
8	t_SQP3=2. AAR - Alcanza los aprendizajes requeridos	1.278	41	Sexo=Hombre	-1.126
9	t_anioBásico=2. Segundo	1.160	42	t_PQP3=2. AAR - Alcanza los aprendizajes requeridos	-1.170
10	t_PQP2=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	1.129	43	ProyEscSQP1=B	-1.175
11	Ocup. Padre=PESCADOR	1.049	44	Enfermedad=ASMÁTICO/A	-1.188
12	NumeroHermanos	0.730	45	EstructuraFamiliar=EXTENDIDA	-1.341
13	Enfermedad=ALERGIA	0.652	46	t_PQP3=1. DAR - Domina los aprendizajes requeridos	-1.379
14	t_RiesgoPQP2=Riesgo moderado	0.408	47	t_PQP3=4. NAAR - No alcanza los aprendizajes requeridos	-1.435
15	Materia abrev.=Matemática	0.345	48	ProyEscSQP3=MB	-1.538
16	SBU=2.0 SBU	0.244	49	ProyEscPQP3=EX	-1.607
17	t_SQP1=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	-0.380	50	t_RiesgoSQP3=Sin riesgo	-1.736
18	Ocup. Padre=CHOFER	-0.413	51	Ocup. Padre=ALBAÑIL	-1.744
19	ComportamientoPQP2=B	-0.468	52	t_SQP3=1. DAR - Domina los aprendizajes requeridos	-1.767
20	t_RiesgoSQP2=Riesgo moderado	-0.552	53	t_RiesgoSQP1=Sin riesgo	-1.849
21	t_RiesgoSQP2=Alto riesgo	-0.577	54	t_RiesgoPQP3=Alto riesgo	-1.980
22	EscolaridadPadre=PRIMARIA	-0.587	55	t_QUI2=2. AAR - Alcanza los aprendizajes requeridos	-2.045
23	Materia abrev.=EE SS	-0.716	56	t_PQP1=2. AAR - Alcanza los aprendizajes requeridos	-2.076
24	ProyEscSQP2=MB	-0.754	57	t_SQP1=1. DAR - Domina los aprendizajes requeridos	-2.143
25	ProyEscPQ=MB	-0.780	58	t_familiaReconstruida=NO	-2.229
26	ComportamientoPQP1=B	-0.782	59	ComportamientoPQP1=A	-3.036
27	ComportamientoPQP2=C	-0.803	60	t_RiesgoSQP1=Alto riesgo	-3.541
28	ProyEscSQP1=EX	-0.803	61	t_QUI2=1. DAR - Domina los aprendizajes requeridos	-3.708
29	t_PQP2=2. AAR - Alcanza los aprendizajes requeridos	-0.823	62	Materia abrev.=Inglés	-4.246
30	t_anioBásico=6. Sexto	-0.895	63	t_QUI1=2. AAR - Alcanza los aprendizajes requeridos	-4.895
31	t_SQP2=3. PAAR - Próximo a alcanzar los aprendizajes requeridos	-0.931	64	t_PQP2=1. DAR - Domina los aprendizajes requeridos	-4.927
32	Sexo=Mujer	-0.934	65	t_SQP1=4. NAAR - No alcanza los aprendizajes requeridos	-5.095
33	t_SQP2=1. DAR - Domina los aprendizajes requeridos	-0.963			

En la **Tabla 41** se observa que en los casos de alumnos que están <Próximos a alcanzar los aprendizajes requeridos> ha influido que sus calificaciones del segundo

y primer quimestre se correspondan con dicha clase (Nº 1, 3, 10, 17…), que la escolaridad alcanzada por el padre sea secundaria (Nº 2) y en una ligera menor medida que sea escolaridad superior (Nº 6), que la escolaridad de la madre sea primaria (Nº 4), que la ocupación del padre sea de pescador (Nº 11), que el número de hermanos sea más alto (Nº 12), que la materia ahora sea Matemática (Nº 15), que el ingreso familiar ronde los dos salarios básicos unificados (Nº 16), que en el comportamiento registre valoraciones muy buenas aunque no excelentes (Nº 24, 25), entre otros.

Tabla 42: Coeficientes Logit de la Regresión Logística con la Regularización de Lasso, sobre la clase <4. No alcanza los aprendizajes requeridos>

	Características	Coef.		Características	Coef.
1	t_RiesgoSQP1=Alto riesgo	3.965	7	t_QUI2=4. NAAR - No alcanza los aprendizajes requeridos	1.315
2	t_SQP1=4. NAAR - No alcanza los aprendizajes requeridos	3.918	8	t_RiesgoSQP2=Alto riesgo	1.024
3	t_PQP3=4. NAAR - No alcanza los aprendizajes requeridos	2.996	9	t_SQP3=4. NAAR - No alcanza los aprendizajes requeridos	0.816
4	t_SQP2=4. NAAR - No alcanza los aprendizajes requeridos	1.662	10	EscolaridadPadre=SECUNDARIA	-0.812
5	t_RiesgoPQP3=Alto riesgo	1.411	11	t_RiesgoPQP3=Sin riesgo	-2.561
6	t_PQP1=2. AAR - Alcanza los aprendizajes requeridos	1.407	12	t_RiesgoPQP2=Sin riesgo	-2.698

En la tabla anterior se observa que en los casos de alumnos que <No alcanzan los aprendizajes requeridos> ha influido por sobre otros factores las calificaciones de modo principal y como factor socioeconómico se observa a la escolaridad del padre, que se corresponde con la secundaria (Nº 10).

3.5.2. Modelos de clasificación sin considerar notas intermedias

En las siguientes tablas se muestran los resultados para las métricas AUC, CA, F1, Precision, Recall y Specificity sin considerar como características para el análisis a las calificaciones intermedias obtenidas por los alumnos. Es de esperar que el promedio final se vea muy influenciado por las calificaciones progresivas de los alumnos, por tanto, la no inclusión de dichas calificaciones ilustra de mejor manera la incidencia de los factores socioeconómicos sobre el rendimiento académico, aun así, se incluyó la calificación del comportamiento de cada alumno y la calificación de su participación en los proyectos escolares porque en ellos se considera en buena medida a las habilidades sociales de los alumnos. El muestreo utilizado es aleatorio

estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos que se fijó.

En las tablas se ha coloreado en tonos más verdes a los valores más altos o mejores y en tonos rojos a los más bajos o peores, se invirtió el sentido de la coloración en el caso de los tiempos de entrenamiento y prueba, así los tiempos más altos se pintaron en rojo.

Tabla 43: Métricas porcentuales resultantes para la clasificación de todas las clases, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
XGBoost	588.28	10.65	0.998	0.971	0.971	0.971	0.971	0.990
Neural Network	1372.11	8.27	0.997	0.967	0.967	0.967	0.967	0.989
Boosting	1102.79	8.31	0.993	0.966	0.966	0.966	0.966	0.989
AdaBoost	248.25	21.86	0.995	0.959	0.959	0.959	0.959	0.986
Random Forest	44.30	10.20	0.995	0.954	0.954	0.955	0.954	0.985
CatBoost	653.92	7.55	0.995	0.949	0.949	0.949	0.949	0.983
XGBoost Random Forest	650.90	9.51	0.991	0.939	0.938	0.938	0.939	0.980
Log. R. Ridge	1852.60	7.76	0.985	0.911	0.911	0.911	0.911	0.970
Log. R. Lasso	1219.58	4.90	0.983	0.909	0.909	0.909	0.909	0.970
kNN	18.46	172.10	0.979	0.898	0.897	0.896	0.898	0.966
SGD	35.23	11.57	0.910	0.866	0.865	0.864	0.866	0.955
C4.5	60.49	0.03	0.929	0.830	0.833	0.847	0.830	0.943
SVM	163.94	31.58	0.793	0.544	0.540	0.550	0.544	0.848
TOTAL (segundos)	**8010.84**	**304.29**						
TOTAL (minutos)	**133.51**	**5.07**						

A continuación, se hacen algunas apreciaciones en especial relación con la tabla precedente referida a la clasificación promedio de todas las clases sin considerar las calificaciones progresivas de los alumnos:

- Los valores de Exactitud de la clasificación (CA) son menores que cuando se considera a las calificaciones que claramente son más determinantes respecto del promedio final, sin embargo, todos los modelos obtienen entre un 83% y 97% de CA, con excepción de SVM que tal cual se indicó en **2.5.2.1. Máquinas de soporte vectorial (SVM)**, funciona mejor para clases dicotómicas, pese a que se puede generalizar a multiclases como es el caso de esta investigación. Incluso

Recall, entendida como la capacidad de generalización a nuevos datos, es poco favorable para SVM.

- Continuando con SVM, la especificidad que reporta es alta, lo cual es importante porque se trata de los casos negativos que el algoritmo ha clasificado correctamente, es decir, expresa cuan bien puede el modelo detectar esa clase.
- A partir de la **Tabla 43**, se denota que los métodos de aprendizaje en conjunto o ensamblado son los mejores para la clasificación general si se incluyen los factores socioeconómicos y no las calificaciones.
- El balanceo de instancias respecto de la clase mayoritaria favorece la generalización de cada modelo para con nuevos datos, lo cual se denota con la métrica de Recall.
- Un factor que puede determinar la elección de un modelo u otro, de nuevo es el tiempo que en ellos se emplea para el entrenamiento y prueba. Los métodos de regresión y más con la regularización de Ridge ocupan de un mayor tiempo de entrenamiento, en ese sentido también es alta la cantidad de tiempo de las Redes Neuronales y Boosting, aunque en el caso de estos dos últimos, adicional al proceso propio de probar los modelos con muestras aleatorias de 10 estratos, emplean sus propios procesos iterativos dada la naturaleza de sus algoritmos.
- Se resalta que los datos para el entrenamiento y prueba estaban balanceados, por lo que la importancia de la métrica F1 que es de gran utilidad cuando la distribución de las clases es desigual, queda algo relegada en esta ocasión.
- De modo general, los tiempos de entrenamiento y prueba rondan en su incremento a un 100% respecto de los casos dónde no se incluyó calificaciones.

Tabla 44: Métricas porcentuales resultantes para la clasificación de la clase <No alcanza los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
Random Forest	44.30	10.20	1.000	0.997	0.993	0.989	0.998	0.996
Boosting	1102.79	8.31	0.999	0.997	0.993	0.988	0.998	0.996
XGBoost	588.28	10.65	1.000	0.997	0.993	0.988	0.998	0.996
XGBoost Random Forest	650.90	9.51	0.999	0.997	0.993	0.988	0.998	0.996
CatBoost	653.92	7.55	1.000	0.996	0.993	0.988	0.998	0.996
Log. R. Lasso	1219.58	4.90	1.000	0.996	0.993	0.987	0.998	0.996
kNN	18.46	172.10	0.999	0.996	0.992	0.988	0.997	0.996
Log. R. Ridge	1852.60	7.76	1.000	0.996	0.992	0.987	0.998	0.995
Neural Network	1372.11	8.27	1.000	0.996	0.992	0.990	0.995	0.997

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
AdaBoost	248.25	21.86	0.999	0.993	0.986	0.976	0.997	0.992
SGD	35.23	11.57	0.987	0.986	0.972	0.954	0.990	0.984
C4.5	60.49	0.03	0.971	0.976	0.951	0.984	0.920	0.995
SVM	163.94	31.58	0.958	0.907	0.814	0.811	0.817	0.937
TOTAL (segundos)	**8010.84**	**304.29**						
TOTAL (minutos)	**133.51**	**5.07**						

En relación con la tabla que precede, referida a la clasificación de alumnos que <No alcanzan los aprendizajes requeridos>, solo SVM resulta algo relegado con un 90% de Exactitud (CA) y las redes neuronales presentan la mejor combinación de CA y Recall, sin embargo, el tiempo de entrenamiento empleado no les favorece.

Tabla 45: Métricas porcentuales resultantes para la clasificación de la clase <Próximo a alcanzar los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
XGBoost	588.28	10.65	1.000	0.993	0.986	0.993	0.979	0.998
Boosting	1102.79	8.31	0.999	0.992	0.985	0.992	0.978	0.997
AdaBoost	248.25	21.86	0.999	0.992	0.984	0.995	0.974	0.998
Random Forest	44.30	10.20	0.999	0.991	0.983	0.986	0.979	0.995
Neural Network	1372.11	8.27	0.999	0.991	0.981	0.983	0.979	0.994
CatBoost	653.92	7.55	0.999	0.988	0.976	0.975	0.976	0.992
XGBoost Random Forest	650.90	9.51	0.998	0.985	0.971	0.965	0.976	0.988
Log. R. Ridge	1852.60	7.76	0.997	0.976	0.953	0.950	0.956	0.983
Log. R. Lasso	1219.58	4.90	0.996	0.975	0.951	0.948	0.953	0.983
kNN	18.46	172.10	0.995	0.975	0.951	0.925	0.979	0.973
SGD	35.23	11.57	0.950	0.962	0.924	0.922	0.927	0.974
C4.5	60.49	0.03	0.890	0.922	0.826	0.935	0.739	0.983
SVM	163.94	31.58	0.747	0.760	0.439	0.528	0.376	0.888
TOTAL (segundos)	**8010.84**	**304.29**						
TOTAL (minutos)	**133.51**	**5.07**						

En relación con la tabla anterior, referida a la clasificación de alumnos que están <Próximos a alcanzar los aprendizajes requeridos>, C4.5 y SVM resultan algo relegados respecto de su CA, incluso Precision, F1 y Recall les son menos favorables. Las redes neuronales presentan la mejor combinación de CA y Recall, sin embargo, el tiempo de entrenamiento empleado no les favorece. De nuevo, los métodos de

aprendizaje en conjunto o ensamblado son los mejores para la clasificación en general si se incluyen los factores socioeconómicos y no las calificaciones.

Tabla 46: Métricas porcentuales resultantes para la clasificación de la clase <Alcanza los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
XGBoost	588.28	10.65	0.996	0.975	0.950	0.940	0.961	0.979
Neural Network	1372.11	8.27	0.994	0.972	0.943	0.935	0.952	0.978
Boosting	1102.79	8.31	0.988	0.970	0.940	0.934	0.947	0.978
AdaBoost	248.25	21.86	0.992	0.963	0.927	0.926	0.928	0.975
kNN	16.82	192.23	0.981	0.960	0.919	0.922	0.916	0.974
Random Forest	44.30	10.20	0.989	0.959	0.919	0.901	0.938	0.966
CatBoost	653.92	7.55	0.989	0.953	0.907	0.901	0.914	0.967
XGBoost Random Forest	650.90	9.51	0.982	0.945	0.890	0.895	0.885	0.965
Log. R. Ridge	1852.60	7.76	0.968	0.917	0.839	0.820	0.857	0.938
Log. R. Lasso	1219.58	4.90	0.964	0.916	0.834	0.825	0.842	0.941
SGD	35.23	11.57	0.840	0.880	0.760	0.762	0.759	0.921
C4.5	60.49	0.03	0.915	0.858	0.746	0.676	0.831	0.867
SVM	163.94	31.58	0.695	0.718	0.402	0.428	0.379	0.831
TOTAL (segundos)	8009.20	324.42						
TOTAL (minutos)	133.49	5.41						

La tabla anterior se refiere a la clasificación de alumnos que <Alcanzan los aprendizajes requeridos>, la clase mayoritaria previo del balanceo de datos con SMOTE, los valores de las métricas se encuentran en el rango de 91% a 97% para los modelos, con excepción del Descenso de Gradientes Estocástico (SGD), C4.5 y SVM. En cada caso F1, Precision y Recall resultaron menores.

Tabla 47: Métricas porcentuales resultantes para la clasificación de la clase <Domina los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
Random Forest	44.30	10.20	1.000	0.997	0.993	0.989	0.998	0.996
Boosting	1102.79	8.31	0.999	0.997	0.993	0.988	0.998	0.996
XGBoost	588.28	10.65	1.000	0.997	0.993	0.988	0.998	0.996
XGBoost Random Forest	650.90	9.51	0.999	0.997	0.993	0.988	0.998	0.996
CatBoost	653.92	7.55	1.000	0.996	0.993	0.988	0.998	0.996
Log. R. Lasso	1219.58	4.90	1.000	0.996	0.993	0.987	0.998	0.996
kNN	18.46	172.10	0.999	0.996	0.992	0.988	0.997	0.996

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
Log. R. Ridge	1852.60	7.76	1.000	0.996	0.992	0.987	0.998	0.995
Neural Network	1372.11	8.27	1.000	0.996	0.992	0.990	0.995	0.997
AdaBoost	248.25	21.86	0.999	0.993	0.986	0.976	0.997	0.992
SGD	35.23	11.57	0.987	0.986	0.972	0.954	0.990	0.984
C4.5	60.49	0.03	0.971	0.976	0.951	0.984	0.920	0.995
SVM	163.94	31.58	0.958	0.907	0.814	0.811	0.817	0.937
TOTAL (segundos)	**8010.84**	**304.29**						
TOTAL (minutos)	**133.51**	**5.07**						

En relación con la tabla precedente referida a la clasificación de alumnos que <Dominan los aprendizajes requeridos>, los valores de las métricas se encuentran en el rango de 90% a 99% para los modelos. En cada caso F1, Precision y Recall también resultaron menores, especialmente en el caso de SVM.

La elección de un modelo para clasificar esta clase dependerá de los tiempos que demandan su entrenamiento y prueba. De nuevo, los métodos de aprendizaje en conjunto o ensamblado son los mejores para la clasificación en general si se incluyen los factores socioeconómicos y no las calificaciones.

La **Figura 47** corresponde al widget Explain Model (Orange, 2021), que enumera las características que más contribuyen al promedio anual del tipo <No alcanza los aprendizajes requeridos> NAAR, clasificadas por C4.5, Las figuras siguientes hacen lo propio con las clases <Domina los aprendizajes requeridos> DAR, <Próximo a alcanzar los aprendizajes requeridos> PAAR y <Alcanza los aprendizajes requeridos> AAR. Como ejemplo, la **Figura 47** muestra los valores SHAP (2018) que tienen un alto impacto para alumnos que <No alcanzan los aprendizajes requeridos> hacia la derecha y los de menos hacia la izquierda. El color del punto representa el valor del atributo, rojo para los valores más altos y azul para los más bajos. Como en este caso las características son categóricas los valores numéricos asociados corresponden con una codificación One Hot ejecutada por el Explain Model, por lo que la contribución más importante del widget es hacia la selección de características.

Para el caso de C4.5 Orange emplea Tree SHAP, que es un algoritmo para calcular valores exactos SHapley Additive exPlanation, SHAP, para modelos basados en árboles de decisión. SHAP es un enfoque de la teoría de juegos para explicar el resultado de cualquier modelo de aprendizaje automático. SHAP se basa en los

valores de Shapley, dónde una predicción se puede explicar asumiendo que cada valor de característica de la instancia es un jugador en un juego donde la predicción es el pago (Van den Broeck et al., 2022).

Los valores de Shapley son métodos de la teoría de juegos de coalición, expresan cómo distribuir de manera justa el pago entre las características. El objetivo de SHAP es explicar la predicción para cualquier instancia x_i como una suma de contribuciones de sus valores de características individuales (Lundberg, 2018; Van den Broeck et al., 2022).

$$\emptyset_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(|M| - |S| - 1)!}{M!} [fx(S \cup \{i\}) - f_x(S)]$$

Dónde $|M|$ es el número total de características. S representa cualquier subconjunto de características que no incluya la i-ésima característica y $|S|$ es el tamaño de ese subconjunto. $f_x(S)$ representa la función de predicción para el modelo para el subconjunto S.

La fórmula es una suma de todos los posibles subconjuntos (S) de valores de entidad, excluyendo el i-ésimo valor de entidad. $|S|!$ representa el número de permutaciones de valores de entidad que aparecen antes del i-ésimo valor de entidad. Del mismo modo, $(|M| - |S| - 1)!$ representa el número de permutaciones de valores de entidad que aparecen después del i-ésimo valor de entidad. El término de diferencia en la ecuación anterior es la contribución marginal de agregar el valor de la característica i-ésima a S. Tenga en cuenta también que la ecuación anterior requiere que calculemos la predicción del modelo para cualquier subconjunto de características (Lundberg, 2018).

Los valores SHAP son las soluciones a la ecuación anterior bajo los supuestos: $f_x(S) = \in [f(x|x_s)]$, es decir, la predicción para cualquier subconjunto S de valores de características es el valor esperado de la predicción para $f_{(x)}$ dado el subconjunto x_x. El cálculo exacto de los valores SHAP es computacionalmente desafiante. (Lundberg & Lee, 2017; Miller, 2018)

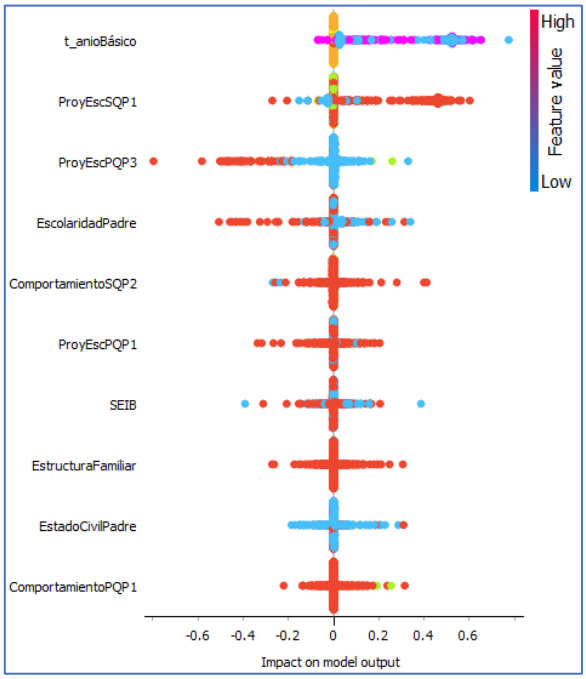


Figura 47: Explain Model C4.5, clase NAAR

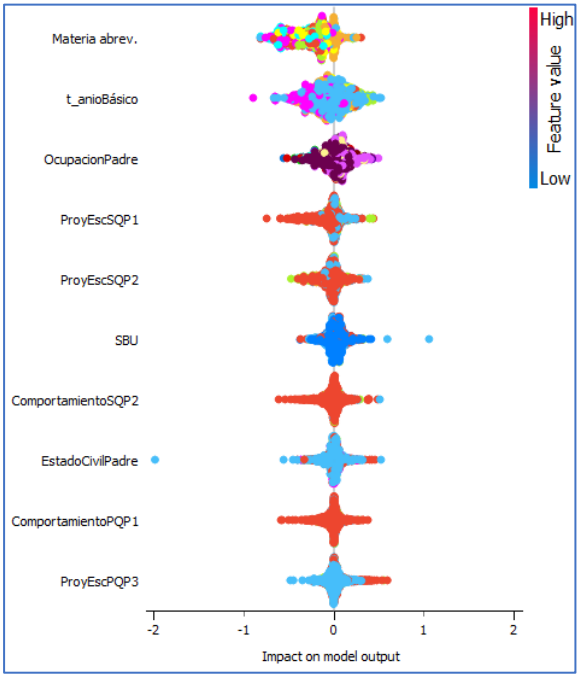


Figura 48: Explain Model C4.5, clase DAR

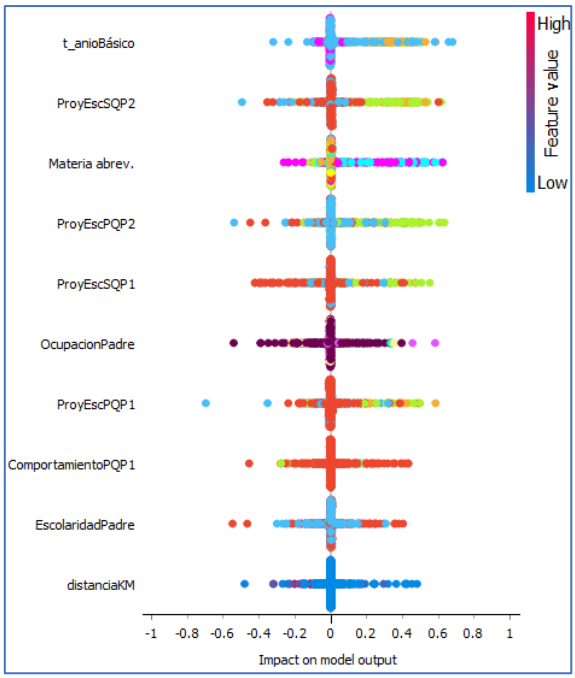


Figura 49: Explain Model C4.5, clase PAAR

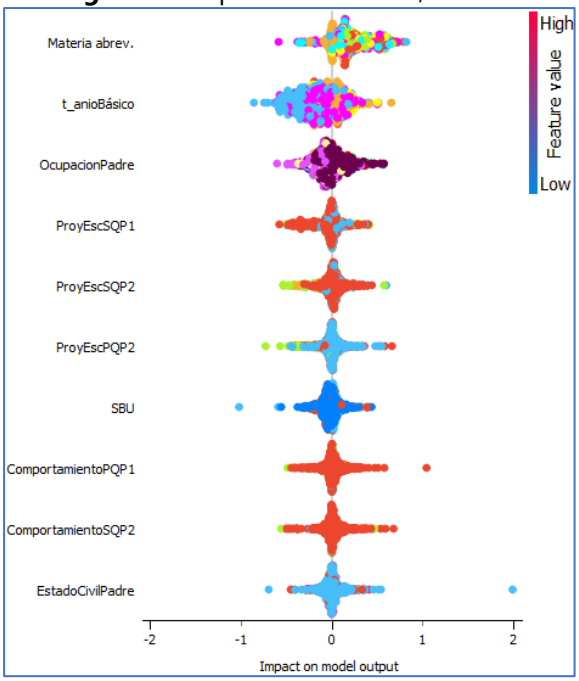


Figura 50: Explain Model C4.5, clase AAR

Si bien SVM tiene la Exactitud global más baja entre los modelos de esta sección con un 54% (Ver **Tabla 43:** Métricas porcentuales resultantes para la clasificación de todas las clases, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos), tiene resultados de

valores de métricas altos para las clases <No alcanza los aprendizajes requeridos> y <Domina los aprendizajes requeridos, con el 90% de CA (ver **Tabla 44** y **Tabla 47**). Por tal razón, en la **Tabla 48** se indican los mejores Valores de Shapley de SVM para estas clases.

Tabla 48: Mejores valores de Shapley para las clases con mejor CA acorde con SVM sobre un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Nº	Domina los aprendizajes requeridos	Nº	No alcanza los aprendizajes requeridos
1	Ocupación de la madre: *Comerciante*	1	Estado civil de padre: *Unión libre*
2	Año básico: *Séptimo*	2	Número de hermanos: *bajo, hacia 0*
3	Comportamiento 1er Quim, P3: *A*	3	Discapacidad: *Si*
4	Estado civil de madre: *Casada*	4	Comportamiento 1er Quim, P3: *A*
5	Ocupación del representante: *Comerciante*	5	Proy. Escolar, 2do Quim, P3: *B*
6	Enfermedad (del alumno): *Alergia*	6	Comportamiento 2do Quim, P3: B
7	Disponibilidad de Internet: *Si*	7	Proy. Escolar, 1er Quim, P3: *B*
8	Escolaridad de representante: *Superior*	8	Ocupación del padre: *Guardia*
9	Procede de otra institución: *Si*	9	Enfermedad: *Asmático*
10	Ocupación del representante: *Administrador*	10	Año básico: *Segundo*

3.5.3. Modelos de regresión sin considerar notas intermedias

En esta sección se presenta y evalúan los resultados de los siguientes modelos para la regresión: Regresión Lineal con regularización de Lasso, Regresión Lineal con regularización de Ridge, CatBoost, Random Forest, kNN, C4.5, Neural Network, Extreme Gradient Boosting (XGBoost), Gradient Boosting, AdaBoost, XGBoost Random Forest y SVM. Los modelos se validaron con la validación cruzada con 10 pliegues o folds. Los 10 pliegues son tomados por sugerencia de la literatura (Nelli, 2018; Witten & Witten, 2017).

Una buena práctica cuando se aplica Validación Cruzada es aleatorizar el 100% de instancias sobre muestreadas respecto de las clases, para asegurar que los datos no estén ordenados de ninguna manera, pues en Orange basta con hacer clic sobre el encabezado de una columna en una tabla y ordenar los datos según la misma. El widget Randomize permite tal acción desde Orange (2015v).

Previo al modelado, en lo concerniente al preprocesamiento y preparación de los datos se ha aleatorizado todas las filas que habían sido muestreadas con SMOTE. En la preparación de datos se depuró a los datos de posibles valores anómalos, lo que beneficia de modo especial a las tareas de regresión porque MSE, RMSE, MAE y R^2 son sensibles a los valores anómalos. Entonces, resultó la siguiente vista parcial respecto de todo el modelo.

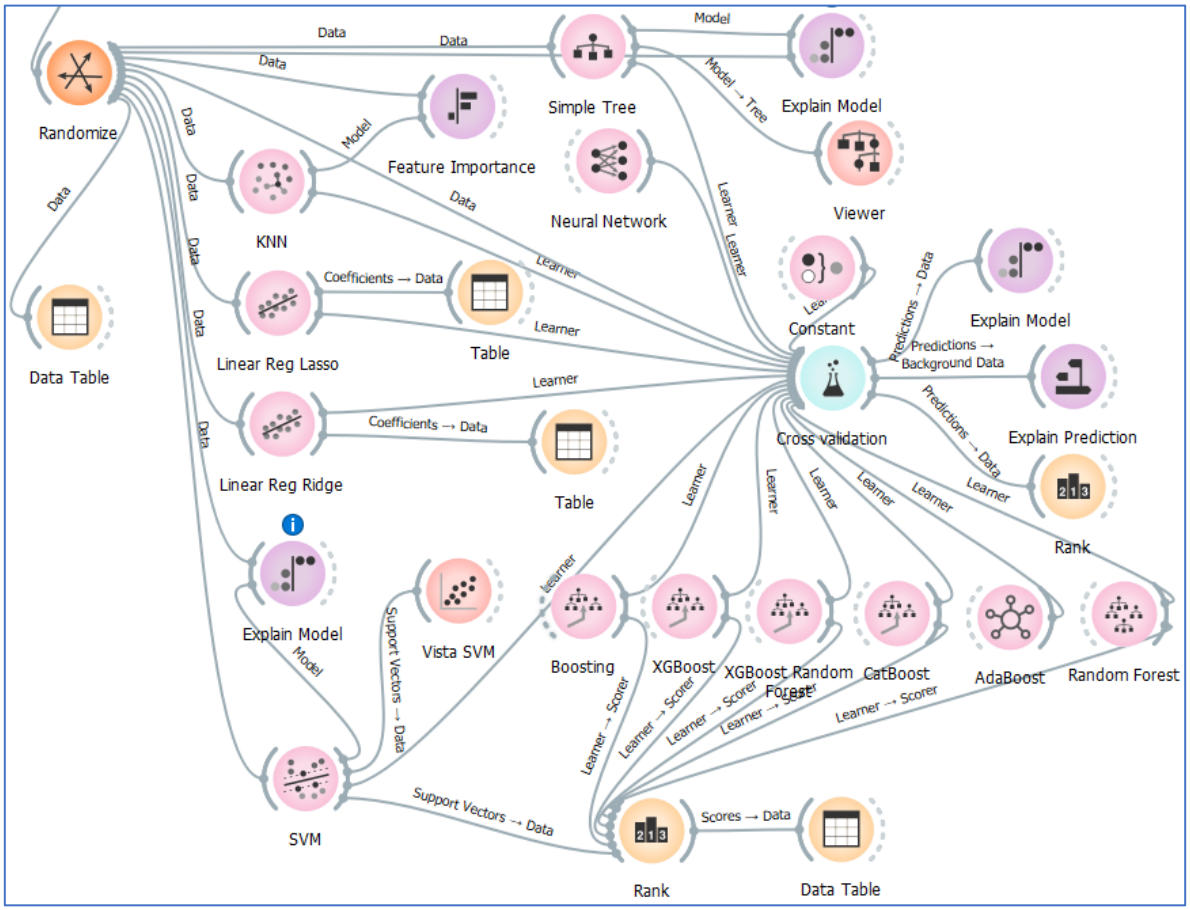


Figura 51: Vista parcial de los modelos para las tareas de regresión sin considerar las calificaciones progresivas de los alumnos

Tabla 49: Métricas resultantes para la regresión y prueba de los modelos con validación cruzada de 10 pliegues

Model	Train time	Test time	MSE	RMSE	MAE	R^2
Linear Reg Lasso	10.36	3.39	1.853	1.361	0.932	0.000
Linear Reg Ridge	9.00	2.86	1.922	1.386	0.954	-0.037
CatBoost	94.79	0.82	1.943	1.394	0.958	-0.049
Random Forest	126.97	11.33	2.103	1.450	0.988	-0.135
kNN	14.23	4.72	2.228	1.493	1.038	-0.203
C4.5	249.43	0.02	2.556	1.599	1.102	-0.380
Neural Network	2241.50	10.98	2.620	1.619	1.122	-0.414
XGBoost	87.28	8.22	3.264	1.807	1.217	-0.762
Boosting	199.45	8.21	3.300	1.817	1.224	-0.781
AdaBoost	253.62	14.52	3.417	1.849	1.229	-0.845
XGBoost Random Forest	9.16	2.67	11.947	3.457	3.308	-5.450

Model	Train time	Test time	MSE	RMSE	MAE	R^2
SVM	12.97	4.69	13.722	3.704	3.564	-6.408
TOTAL (segundos)	**3308.74**	**72.42**				
TOTAL (minutos)	**55.15**	**1.21**				

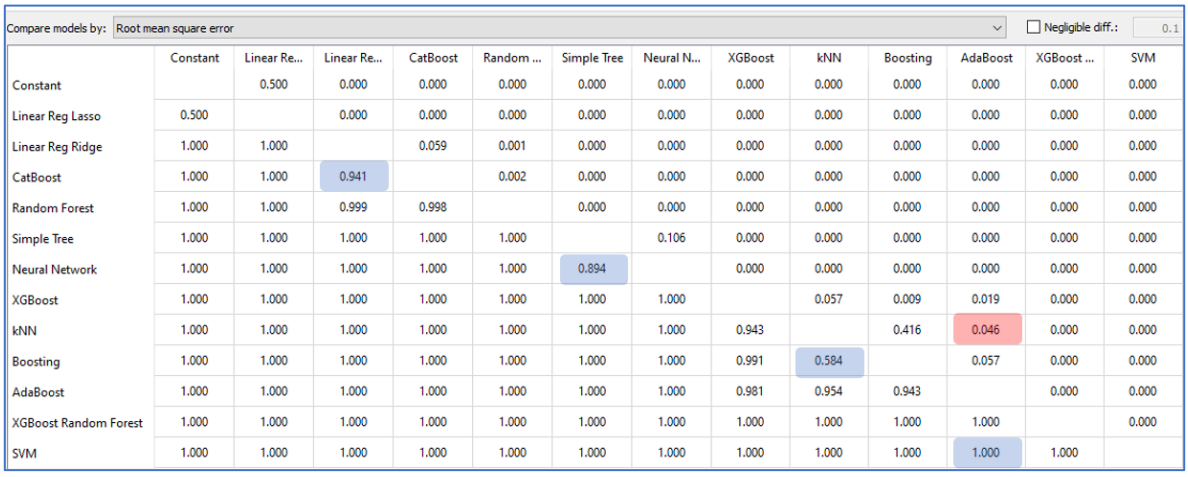
Un beneficio de usar RMSE es que la métrica que produce es en términos de la unidad que se predice, en el caso de esta investigación se predicen promedios de 1 a 10, por lo que RMSE reporta el error en términos de promedios de los alumnos, eso deja comprender que el rendimiento de cada modelo, de mejor a peor y con un error máximo de hasta 1.49 puntos es: Regresión Lineal con regularización de Lasso, Regresión Lineal con regularización de Ridge, CatBoost, Random Forest y kNN. Por otra parte. MSE es igual a RMSE^2.

Con base en el párrafo precedente, a partir de la **Tabla 49** se denota que C4.5, Neural Network, XGBoost, Boosting, AdaBoost, XGBoost Random Forest y SVM, calculan la relación estimada entre el promedio anual y las características, sin incluir las calificaciones progresivas mediante RMSE con errores desde 1.5 puntos hasta 3.70 respectivamente en el proceso.

Otra métrica que resulta de interés es R^2, que como se documentó en el marco teórico, siempre estará entre -∞ y 1. Los mejores valores son los cercanos a 1 y los peores los negativos… que tienden a -∞.

MAE, que se documentó en **2.5.2.11.10. Error absoluto medio, MAE** coincide con las restantes métricas respecto de cuáles son los mejores modelos para la regresión del caso que compete a la investigación.

La comparativa de la **Figura 52**, se lee por pares de modelos utilizando la puntuación seleccionada, que en este caso es RMSE. En el primer rectángulo azul translúcido desde arriba hacia abajo se indica que CatBoost tiene un 94.1% de posibilidad de mayor RMSE que la regresión lineal con regularización de Ridge, el segundo rectángulo translúcido indica de Neural Network tiene un 89.4% de posibilidad de un mayor RMSE que C4.5 (Tree)… Si se desea leer la tabla de otro modo, el rectángulo translúcido rojo señala que AdaBoost tiene un 4.6% de posibilidad de ser menor que kNN. Los números pequeños, incluso 0, muestran la probabilidad de que la diferencia sea insignificante, como para el caso lo es entre Tree y kNN.


Compare models by: Root mean square error — Negligible diff.: 0.1

	Constant	Linear Re...	Linear Re...	CatBoost	Random ...	Simple Tree	Neural N...	XGBoost	kNN	Boosting	AdaBoost	XGBoost ...	SVM
Constant		0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Linear Reg Lasso	0.500		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Linear Reg Ridge	1.000	1.000		0.059	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CatBoost	1.000	1.000	0.941		0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random Forest	1.000	1.000	0.999	0.998		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Simple Tree	1.000	1.000	1.000	1.000	1.000		0.106	0.000	0.000	0.000	0.000	0.000	0.000
Neural Network	1.000	1.000	1.000	1.000	1.000	0.894		0.000	0.000	0.000	0.000	0.000	0.000
XGBoost	1.000	1.000	1.000	1.000	1.000	1.000	1.000		0.057	0.009	0.019	0.000	0.000
kNN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.943		0.416	0.046	0.000	0.000
Boosting	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.584		0.057	0.000	0.000
AdaBoost	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.981	0.954	0.943		0.000	0.000
XGBoost Random Forest	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000		0.000
SVM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

Figura 52: Comparativa de probabilidad de puntuación mayor en RMSE entre el algoritmo de las filas vs el algoritmo de las columnas

En la **Tabla 50** se ha empleado el algoritmo RefiefF, que fue creado en 1997 y se lo emplea para encontrar los pesos predictores de las características respecto del promedio anual continuo, porque los algoritmos de esta sección se enmarcan en tareas de regresión. También se lo puede emplear para tareas de clasificación respecto de clases categóricas como es el caso de la **Tabla 57**. RefiefF penaliza a las características que dan diferentes valores a vecinos de la misma clase y recompensa a las características que dan diferentes valores a vecinos de diferentes clases. Es decir, estima la calidad de las características sobre la base de qué tan bien pueden distinguir entre instancias que están cerca unas de otras, se apoya en kNN (Kononenko et al., 1997; Orange, 2015r).

Tabla 50: ReliefF aplicado en tareas de regresión

	Característica	Relief		Característica	ReliefF
1	Materia abrev.	0.566704	22	distanciaKM	0.076214
2	Ocup. Padre	0.305417	23	ProyEscPQP1	0.07111
3	Ocup. Madre	0.255175	24	AnioLlegada	0.069459
4	Sexo	0.218114	25	ProyEscPQP2	0.064354
5	EscolaridadPadre	0.181124	26	ProyEscSQP3	0.061742
6	EstadoCivilMadre	0.179729	27	ProyEscSQP2	0.058585
7	Ocup. Representante	0.155038	28	EscolaridadRepresentante	0.055973
8	EstadoCivilPadre	0.153863	29	EscolaridadMadre	0.055186
9	t_anioBásico	0.145793	30	aniosRetraso	0.055008
10	ParentescoRepresentante	0.136397	31	Celular	0.053984
11	EstructuraFamiliar	0.134918	32	ComportamientoPQP1	0.052086

	Característica	Relief		Característica	ReliefF
12	Enfermedad	0.129034	33	ProyEscSQP1	0.046688
13	SBU	0.118293	34	t_familiaReconstruida	0.040032
14	TVCable	0.113539	35	Alcantarillado	0.039257
15	ProyEscPQP3	0.110443	36	ComportamientoPQP3	0.032979
16	Computador	0.107532	37	ComportamientoSQP2	0.023156
17	ProcedeDeOtraInstitucion	0.105678	38	ComportamientoPQP2	0.02236
18	DificultadAutoreportada	0.09902	39	AguaPotable	0.02122
19	NumeroHermanos	0.087954	40	ComportamientoSQP1	0.019936
20	Internet	0.087202	41	Discapacidad	0.016024
21	SEIB	0.078711	42	ComportamientoSQP3	0.003222

En el siguiente enlace se puede apreciar la gráfica del árbol de decisión C4.5 para modelos de regresión con balanceo ponderado https://tinyurl.com/5n7sh6ta. No fue posible insertarla en este documento dada su alta resolución.

La importancia de las características se refiere a una clase de técnicas para asignar puntuaciones e importancia relativa a las características de entrada en un modelo supervisado, tanto para las tareas de clasificación como de regresión. Estas técnicas proporcionan una mejor comprensión de los datos y del modelo e incluso ayudan a reducir el número de características de entrada, eso sí, demandan un alto costo a nivel de cálculo computacional. La **Figura 53** se refiere a la técnica de importancia de la característica de permutación (Orange, 2015g).

En síntesis, con estas técnicas se consigue explicar algunas características de los datos objeto de estudio, e ir más allá de los valores de las métricas para evitar modelos abstractos, sino que resulten interpretables desde la perspectiva de las características del conjunto de datos, es decir, de cómo los predictores influyen en las decisiones de los modelos. Las características con puntuaciones cercanas a 0 son menos relevantes. Los pasos generales del cálculo de la importancia se resumen en:

Calcule la puntuación de referencia del modelo sobre los datos, por ejemplo, RMSE
Para cada característica D_j:
Por cada K repetición:
Mezcla aleatoriamente los datos de la columna $D_j \breve{D}_{k,j}$
Calcule la puntuación del modelo, ejemplo, RMSE: $S_{k,j} \breve{D}_{k,j}$
Calcular la importancia de la característica con base en: $i_j = S - \frac{1}{K}\sum_{K=1}^{K} S_{k,j}$

Es de señalar que los coeficientes que reportan la regresión lineal y la logística también se pueden usar como puntuación de importancia de las características.

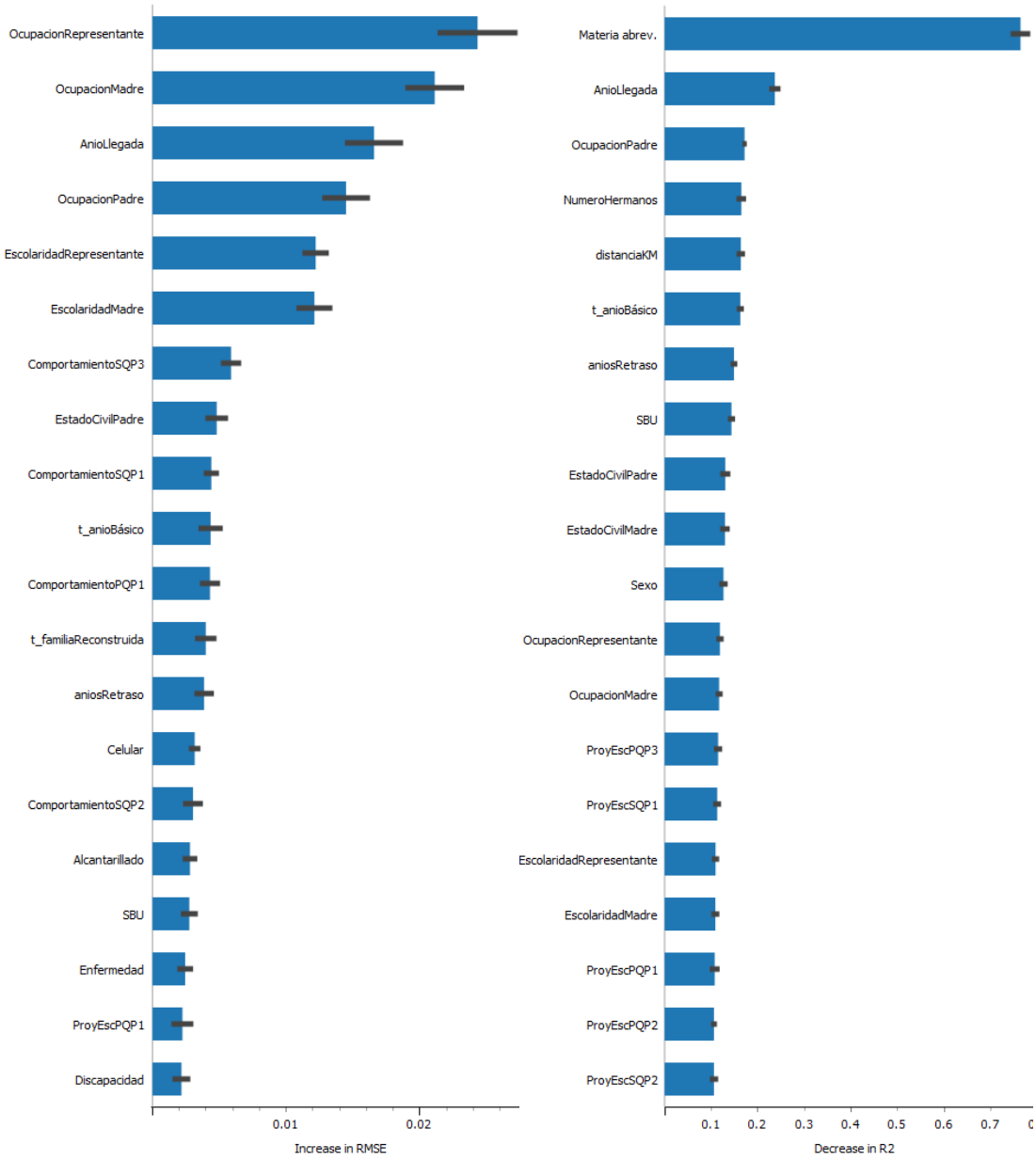


Figura 53: Feature importance con base en RMSE de la Regresión Logística (Izq) con Regularización de Ridge y Feature importance con base en R^2 de kNN (Der). Ambos ejecutados con 10 permutaciones.

Los algoritmos de árbol de decisión como C4.5 o CART ofrecen puntuaciones de importancia basadas en la reducción del criterio utilizado para seleccionar puntos de división, como Gini o la entropía. Ese mismo enfoque se puede utilizar para conjuntos

de árboles de decisión, como el bosque aleatorio y los algoritmos de aumento de gradiente estocástico.

3.5.4. Modelos de clasificación con PCA, Smote ponderado y sin considerar notas intermedias

Con diferencia de las secciones precedentes, en este grupo de algoritmos que forman parte del modelo general, el sobre muestreo de las clases minoritarias es ponderado y se ha seleccionado sólo una muestra del 70% de la clase mayoritaria <Alcanza los aprendizajes requeridos> con el afán de evitar posibles sobre ajustes pese a que los datos estan equilibrados respecto de la clase categórica Promedio Anual. Los pasos ilustrados en la **Figura 54**, se detallaron junto con el bloque de código respectivo en la sección **3.3.4. Aumento de datos**.

Además, en este grupo de algoritmos se redujo la dimensionalidad de los datos expresando en 15 componentes a las 43 características seleccionadas, para lo cual se empleó PCA y se obtuvo un 30% de varianza explicada, previo de estandarizar los datos. Los componentes principales calculados sobre características estandarizadas son auto vectores que se toman de una matriz de correlaciones. La elección se realiza de manera que la primera componente principal sea la que mayor varianza recoja; la segunda debe recoger la máxima variabilidad no recogida por la primera y así sucesivamente, eligiendo un número que recoja un porcentaje suficiente de varianza total, en este caso 30%.

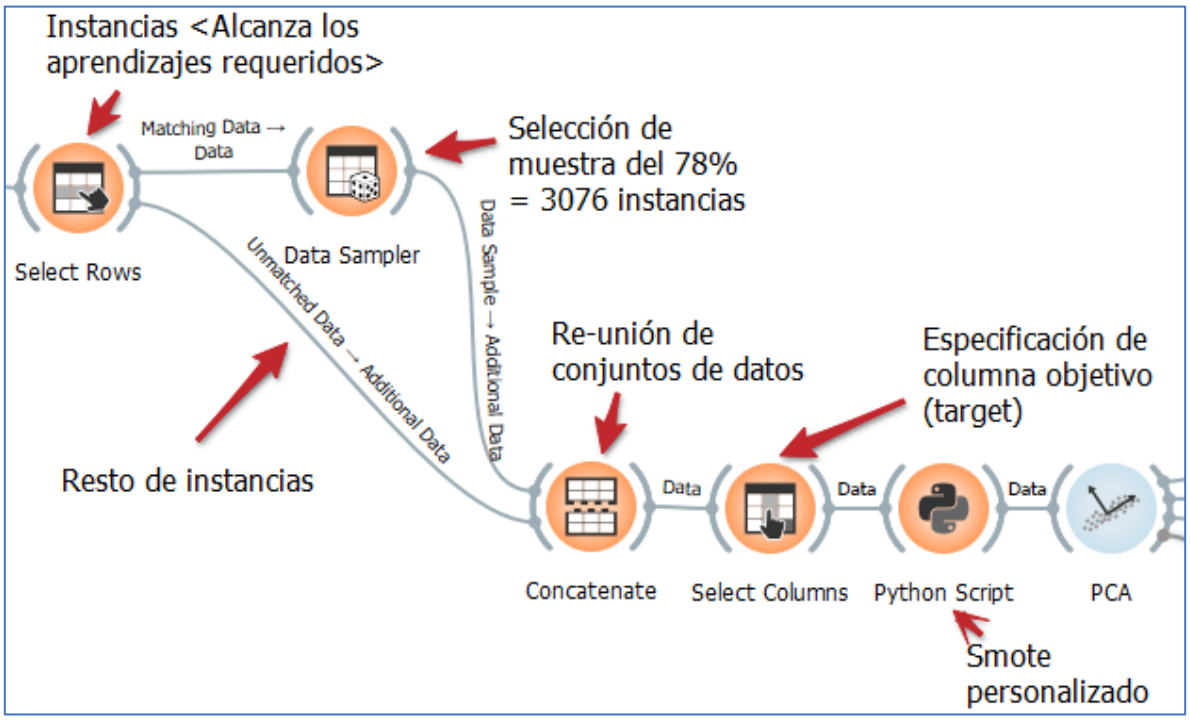


Figura 54: Pasos generales del balanceo ponderado de clases

Respecto de PCA, se tuvo como objetivo identificar las combinaciones lineales que mejor representan a las características $x_1, x_2, \dots, x_p$ con $z_1, z_2, \dots, z_m$ y que resulte $m < p$ combinaciones lineales de las p (43) características originales, es decir:

$$z_m = \sum_{j=1}^{p} \Phi_{jm} x_j$$

Donde $\Phi_{1m}, \Phi_{2m}, \dots, \Phi_{pm}$ son las cargas de los componentes principales, por ejemplo, Φ_{11} corresponde a la primera carga de la primera componente principal. Las cargas o loadings representan el peso de cada característica en cada componente. Cada vector de *cargas* $[\Phi_{1m}, \Phi_{2m}, \dots, \Phi_{pm}]$, de longitud igual a p, define además la dirección en el espacio sobre el cual la varianza de los datos es mayor. Luego, la combinación lineal se normaliza para no inflar la varianza, para lo cual la suma de cuadrados de las cargas se iguala a 1.

$$\sum_{j=1}^{p} \Phi_{j1}^2 = 1$$

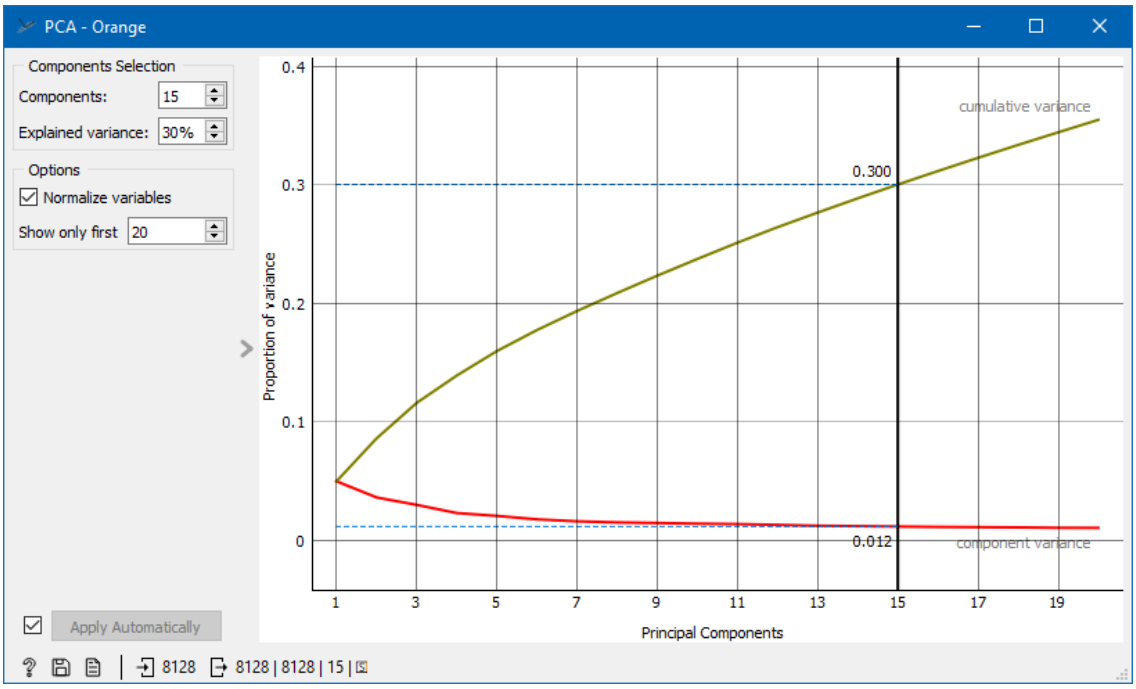


Figura 55: Reducción de 43 características a 15 componentes principales con el widget PCA. La varianza explicada alcanza el 30%

La primera componente principal (z_1) es aquella cuya dirección contiene la mayor variabilidad y por ende información de los datos. Es un vector que define la línea lo más próxima posible a los datos y que minimiza la suma de las distancias perpendiculares entre cada dato y la línea representada por la componente, la proximidad corresponde al promedio de la distancia euclídea al cuadrado:

$$z_i = \Phi_{11}x_{i1} + \Phi_{21}x_{i2} + \cdots + \Phi_{p1}x_{ip}$$

La segunda componente principal (z_2) será una combinación lineal de las variables, que recoja la segunda dirección con mayor varianza de los datos, pero que no esté correlacionada con z_1, es decir, que la dirección de z_2 es perpendicular respecto de z_1. Para comprender la varianza de los m (15) componentes, es de indicar que la varianza total presente en los datos se define como:

$$\sum_{j=1}^{p} Var(x_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

Luego, la varianza explicada por la m-ésima componente principal se corresponde con:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} \Phi_{jm} x_{ij}\right)^2$$

Y la proporción de varianza explicada de la m-ésima componente principal es:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p} \Phi_{jm} x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}$$

El resultado será un número positivo y la suma de todas las varianzas explicadas de los m componentes principales será 1. Con PCA interesa conocer la proporción de varianza explicada por cada uno de los componentes principales, es decir, cuanta información presente en los datos se pierde por la proyección de las observaciones sobre los primeros 15 componentes principales. Como se explicó anteriormente, cada eigenvalor se corresponde con la varianza del componente Z_i definido por el eigenvector $\overrightarrow{v_i}$

$$Var(Z_i) = \lambda_i$$

Por lo que la proporción de varianza total que explica la componente Z_i será:

$$\frac{\lambda_i}{\sum_{i=1}^{p}} = \frac{\lambda_i}{\sum_{i=1}^{p} Var(x_i)}$$

Donde la suma de las varianzas de los componentes principales y las variables originales son iguales. Multiplicando esta proporción por 100 se obtiene el porcentaje y sumando todos los autovalores λ_i se obtiene la varianza total de todos los componentes, que en el caso de esta investigación es 30%.

$$\sum_{i=1}^{p} Var(Z_i) = \sum_{i=1}^{p} \lambda_i$$

En las siguientes tablas se muestran los resultados para las métricas AUC, CA, F1, Precision, Recall y Specificity sin considerar como características para el análisis a las calificaciones intermedias obtenidas por los alumnos y previendo que no la no inclusión de dichas calificaciones ilustra de mejor manera la incidencia de los factores socioeconómicos sobre el rendimiento académico. El muestreo utilizado es aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos que se fijó.

Tabla 51: Métricas porcentuales resultantes para la clasificación de todas las clases, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
XGBoost	61.351	0.441	0.980	0.930	0.930	0.930	0.930	0.964
AdaBoost	412.412	1.943	0.977	0.929	0.929	0.929	0.929	0.963
Boosting	379.941	1.161	0.978	0.929	0.929	0.929	0.929	0.963
Random Forest	12.512	0.864	0.976	0.914	0.913	0.914	0.914	0.954
XGBoost Random Forest	76.215	0.594	0.964	0.879	0.878	0.878	0.879	0.938
C4.5	12.508	0.014	0.941	0.877	0.877	0.876	0.877	0.940
CatBoost	225.595	0.559	0.959	0.864	0.863	0.864	0.864	0.929
kNN	0.739	1.819	0.949	0.850	0.851	0.852	0.850	0.922
Neural Network	158.065	0.155	0.921	0.789	0.789	0.794	0.789	0.891
Log. R. Lasso	37.743	0.061	0.836	0.650	0.650	0.655	0.650	0.824
Log. R. Ridge	21.885	0.079	0.842	0.643	0.643	0.657	0.643	0.814
SGD	0.872	0.095	0.670	0.539	0.544	0.551	0.539	0.776
SVM	18.571	4.566	0.648	0.412	0.402	0.431	0.412	0.753
TOTAL (segundos)	**1418.41**	**12.35**						
TOTAL (minutos)	**23.64**	**0.21**						

Como antecedente de la explicación de la tabla precedente y las tres similares que siguen, es de recordar que F1 es ideal cuando los falsos negativos y falsos positivos sean relevantes al estudio, además, de que también lo es para datos desequilibrados, como lo están ligeramente después del sobre muestreo ponderado. A continuación, se hacen algunas apreciaciones en especial relación con la tabla precedente referida a la clasificación promedio de todas las clases.

- El haber reducido la dimensionalidad de los datos y ponderado el sobre muestreo aminoró el tiempo de entrenamiento a apenas 23 minutos y el de prueba a 0.2 minutos.
- Como las instancias se desbalancearon al respecto de las clases, los modelos que dependen con más énfasis del balanceo se afectaron en los valores de sus métricas, en especial la Regresión Logística con Regularización de Lasso, la Regresión Logística con Regularización de Ridge, el Descenso de Gradientes Estocástico (SGD) y SVM. Aunque, la métrica de especificidad que indica cuántas de las instancias negativas reales se predijeron de forma correcta alcanza valores muy favorables en ellos.

- Los métodos de ensamblado XGBoost, AdaBoost, Boosting, Random Forest, XGBoost Random Forest y CatBoost de nuevo se presentan como las mejores opciones para los casos de clasificación promedio de todas las clases. Si bien, AdaBoost y Boosting obtienen los tiempos más altos respecto del entrenamiento, entre todos los algoritmos considerados, estos son apenas de entre 6 a 8 minutos.
- La diferencia entre C4.5 y CatBoost es de milésimas en casi todas las métricas en favor de C4.5. CatBoost resultó como el peor de los métodos de ensamblado.

Tabla 52: Métricas porcentuales resultantes para la clasificación de la clase <No alcanza los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
AdaBoost	412.412	1.943	0.993	0.997	0.990	0.987	0.993	0.998
Boosting	379.941	1.161	0.998	0.997	0.989	0.986	0.993	0.998
Random Forest	12.512	0.864	0.997	0.997	0.989	0.985	0.993	0.998
XGBoost	61.351	0.441	1.000	0.997	0.989	0.985	0.993	0.998
kNN	0.739	1.819	0.996	0.997	0.989	0.987	0.991	0.998
CatBoost	225.595	0.559	1.000	0.997	0.989	0.985	0.993	0.998
XGBoost Random Forest	76.215	0.594	0.997	0.997	0.988	0.984	0.992	0.998
Neural Network	158.065	0.155	0.999	0.996	0.981	0.996	0.968	0.999
C4.5	12.508	0.014	0.995	0.994	0.977	0.963	0.991	0.995
Log. R. Lasso	37.743	0.061	0.998	0.993	0.972	0.976	0.968	0.997
Log. R. Ridge	21.885	0.079	0.997	0.990	0.960	0.959	0.962	0.994
SGD	0.872	0.095	0.895	0.969	0.865	0.948	0.796	0.994
SVM	18.571	4.566	0.953	0.951	0.783	0.869	0.712	0.985
TOTAL (segundos)	**1418.41**	**12.35**						
TOTAL (minutos)	**23.64**	**0.21**						

En relación con la tabla precedente referida a la clasificación de alumnos que <No alcanzan los aprendizajes requeridos>, todos los algoritmos alcanzan valores altos, con excepción del Descenso de Gradientes Estocástico (SGD) y SVM que en Recall, entendida como la capacidad de generalización del algoritmo a otros datos, tiene valores menores.

Tabla 53: Métricas resultantes para la clasificación de la clase <Próximo a alcanzar los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
AdaBoost	412.412	1.943	0.998	0.991	0.975	0.983	0.968	0.996
Boosting	379.941	1.161	0.998	0.990	0.974	0.980	0.967	0.996
XGBoost	61.351	0.441	0.998	0.990	0.974	0.979	0.969	0.995
Random Forest	12.512	0.864	0.998	0.989	0.969	0.974	0.965	0.994
CatBoost	225.595	0.559	0.996	0.979	0.943	0.933	0.953	0.985
XGBoost Random Forest	76.215	0.594	0.990	0.977	0.937	0.935	0.939	0.985
kNN	0.739	1.819	0.991	0.974	0.929	0.951	0.908	0.989
C4.5	12.508	0.014	0.980	0.974	0.929	0.930	0.927	0.984
Neural Network	158.065	0.155	0.981	0.946	0.853	0.855	0.851	0.967
Log. R. Ridge	21.885	0.079	0.905	0.880	0.617	0.753	0.523	0.961
Log. R. Lasso	37.743	0.061	0.901	0.872	0.621	0.687	0.567	0.941
SGD	0.872	0.095	0.708	0.839	0.535	0.574	0.500	0.916
SVM	18.571	4.566	0.586	0.713	0.277	0.259	0.297	0.808
TOTAL (segundos)	**1418.41**	**12.35**						
TOTAL (minutos)	**23.64**	**0.21**						

En relación con la tabla precedente referida a la clasificación de alumnos que están <Próximos a alcanzar los aprendizajes requeridos>, todos los algoritmos de ensamblado alcanzan los mejores valores en todas las métricas.

Tabla 54: Métricas porcentuales resultantes para la clasificación de la clase <Alcanza los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
XGBoost	61.351	0.441	0.970	0.933	0.912	0.908	0.915	0.943
Boosting	379.941	1.161	0.967	0.932	0.911	0.904	0.918	0.940
AdaBoost	412.412	1.943	0.968	0.932	0.910	0.903	0.918	0.940
Random Forest	12.512	0.864	0.963	0.916	0.892	0.870	0.915	0.917
XGBoost Random Forest	76.215	0.594	0.944	0.885	0.850	0.838	0.861	0.899
C4.5	12.508	0.014	0.907	0.884	0.845	0.853	0.836	0.913
CatBoost	225.595	0.559	0.935	0.868	0.832	0.805	0.861	0.873
kNN	0.739	1.819	0.922	0.856	0.814	0.796	0.832	0.870
Neural Network	158.065	0.155	0.873	0.797	0.749	0.705	0.799	0.796
Log. R. Lasso	37.743	0.061	0.742	0.683	0.609	0.570	0.654	0.700
Log. R. Ridge	21.885	0.079	0.754	0.665	0.603	0.546	0.674	0.659
SGD	0.872	0.095	0.588	0.609	0.494	0.484	0.504	0.673
SVM	18.571	4.566	0.547	0.587	0.289	0.413	0.222	0.809
TOTAL (segundos)	**1418.41**	**12.35**						

TOTAL (minutos)	**23.64**	**0.21**

En relación con la tabla precedente referida a la clasificación de alumnos que <Alcanzan los aprendizajes requeridos>, todos los algoritmos de ensamblado alcanzan los mejores valores en todas las métricas, siendo mínimas las diferencias entre CatBoost y C4.5 que no es un algoritmo de ensamblado. Esta clase es la mayoritaria.

Tabla 55: Métricas resultantes para la clasificación de la clase <Domina los aprendizajes requeridos>, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Model	Train time	Test time	AUC	CA	F1	Precision	Recall	Specificity
XGBoost	61.351	0.441	0.969	0.940	0.904	0.907	0.900	0.958
Boosting	379.941	1.161	0.966	0.938	0.901	0.908	0.895	0.958
AdaBoost	412.412	1.943	0.966	0.938	0.901	0.907	0.895	0.958
Random Forest	12.512	0.864	0.963	0.925	0.877	0.904	0.850	0.959
C4.5	12.508	0.014	0.922	0.902	0.845	0.839	0.852	0.925
XGBoost Random Forest	76.215	0.594	0.949	0.899	0.835	0.852	0.819	0.935
CatBoost	225.595	0.559	0.940	0.883	0.804	0.849	0.763	0.938
kNN	0.739	1.819	0.925	0.873	0.795	0.808	0.783	0.915
Neural Network	158.065	0.155	0.895	0.840	0.725	0.788	0.672	0.917
Log. R. Lasso	37.743	0.061	0.814	0.752	0.590	0.613	0.569	0.836
Log. R. Ridge	21.885	0.079	0.818	0.751	0.582	0.615	0.552	0.842
SGD	0.872	0.095	0.618	0.661	0.483	0.463	0.504	0.732
SVM	18.571	4.566	0.661	0.572	0.464	0.383	0.590	0.563
TOTAL (segundos)	**1418.41**	**12.35**						
TOTAL (minutos)	**23.64**	**0.21**						

En relación con la tabla precedente referida a la clasificación de alumnos que <Dominan los aprendizajes requeridos>, los algoritmos de ensamblado XGBoost, Boosting, AdaBoost y Random Forest, alcanzan los mejores valores en todas las métricas, aunque entre los valores para sus distintas métricas se observan diferencias ocasionadas por el desbalanceo de las clases.

Las calificaciones de Recall, valorado como la capacidad de generalización para con nuevos datos, si bien son algo menores guardan proporción con la Exactitud (CA). Por otro lado, los siguientes fueron los algoritmos que obtuvieron menores valoraciones: CatBoost, kNN, Neural Network, Regresión Logística con

Regularización de Lasso, la Regresión Logística con Regularización de Ridge, el Descenso de Gradientes Estocástico (SGD) y SVM.

En la siguiente tabla se muestran las matrices de confusión de los algoritmos XGBoost, Boosting, AdaBoost, Random Forest, C4.5, XGBoost Random Forest, CatBoost, kNN, Neural Network y regresión logística con regularización de Lasso.

Tabla *56*: Matrices de confusión de los modelos para todas las clases, empleando un muestreo aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos

Matriz de confusión

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	90.7 %	8.0 %	0.4 %	0.1 %	5110
2. AAR - Alcanza...	9.2 %	90.8 %	1.8 %	0.0 %	6150
3. PAAR - Próxim...	0.1 %	1.0 %	97.9 %	1.4 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.0 %	98.5 %	2000
Σ	5071	6203	2970	2016	16260

XGBoost

Matriz de confusión

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	90.8 %	8.3 %	0.5 %	0.1 %	5110
2. AAR - Alcanza...	9.2 %	90.4 %	1.5 %	0.0 %	6150
3. PAAR - Próxim...	0.1 %	1.1 %	98.0 %	1.3 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.0 %	98.6 %	2000
Σ	5039	6246	2960	2015	16260

Boosting

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	90.7 %	8.4 %	0.3 %	0.1 %	5110
2. AAR - Alcanza...	9.3 %	90.3 %	1.3 %	0.0 %	6150
3. PAAR - Próxim...	0.0 %	1.1 %	98.3 %	1.2 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.0 %	98.7 %	2000
Σ	5044	6250	2954	2012	16260

AdaBoost

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	90.4 %	11.6 %	0.5 %	0.1 %	5110
2. AAR - Alcanza...	9.5 %	87.0 %	2.1 %	0.0 %	6150
3. PAAR - Próxim...	0.0 %	1.2 %	97.4 %	1.4 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.0 %	98.5 %	2000
Σ	4806	6467	2971	2016	16260

Random Forest

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	83.9 %	11.8 %	1.1 %	0.8 %	5110
2. AAR - Alcanza...	15.5 %	85.3 %	5.8 %	1.5 %	6150
3. PAAR - Próxim...	0.6 %	2.7 %	93.0 %	1.4 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.1 %	96.3 %	2000
Σ	5187	6025	2990	2058	16260

Tree, C4.5

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	85.2 %	13.9 %	1.5 %	0.1 %	5110
2. AAR - Alcanza...	14.3 %	83.8 %	5.0 %	0.0 %	6150
3. PAAR - Próxim...	0.5 %	2.0 %	93.5 %	1.5 %	3000
4. NAAR - No ...	0.0 %	0.3 %	0.0 %	98.4 %	2000
Σ	4911	6320	3013	2016	16260

XGBoost Random Forest

Matriz de confusión

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	84.9 %	17.7 %	1.4 %	0.1 %	5110
2. AAR - Alcanza...	15.1 %	80.5 %	5.3 %	0.0 %	6150
3. PAAR - Próxim...	0.1 %	1.7 %	93.3 %	1.4 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.0 %	98.5 %	2000
Σ	4596	6584	3065	2015	16260

CatBoost

Matriz de confusión

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	89.5 %	7.9 %	0.4 %	0.1 %	5110
2. AAR - Alcanza...	10.5 %	90.7 %	1.9 %	0.0 %	6150
3. PAAR - Próxim...	0.1 %	1.2 %	97.4 %	1.3 %	3000
4. NAAR - No ...	0.0 %	0.2 %	0.2 %	98.6 %	2000
Σ	5152	6123	2975	2010	16260

kNN

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	78.8 %	22.6 %	3.2 %	0.1 %	5110
2. AAR - Alcanza...	20.7 %	70.5 %	11.3 %	0.0 %	6150
3. PAAR - Próxim...	0.5 %	6.0 %	85.5 %	0.3 %	3000
4. NAAR - No ...	0.0 %	0.9 %	0.0 %	99.6 %	2000
Σ	4361	6970	2986	1943	16260

Neural Network

	1. DAR - Domin...	2. AAR - Alcanza...	3. PAAR - Próxim...	4. NAAR - No ...	Σ
1. DAR - Domin...	61.3 %	26.8 %	11.0 %	1.9 %	5110
2. AAR - Alcanza...	34.1 %	57.0 %	20.3 %	0.3 %	6150
3. PAAR - Próxim...	4.5 %	15.3 %	68.7 %	0.3 %	3000
4. NAAR - No ...	0.0 %	0.9 %	0.0 %	97.6 %	2000
Σ	4740	7059	2478	1983	16260

Log. R. Lasso

Si bien algunos resultados varían entre los modelos dada su naturaleza estocástica, todos son buenos clasificadores entre los alumnos que <No alcanzan los aprendizajes requeridos> o están <Próximos a alcanzar los aprendizajes requeridos>. Sin embargo, las tasas de Exactitud son menores para clasificar a los alumnos que <Alcanzan los aprendizajes requeridos> o <Dominan los aprendizajes requeridos>. Esto da cuenta de que el factor socioeconómico si se relaciona con calificaciones más bajas, pero no marcan diferencia sustancial entre los alumnos clasificados en los niveles más altos posibles. Además:

- XGBoost, Boosting, AdaBoost, Random Forest y KNN obtuvieron una Exactitud, CA, superior al 97.4% para clasificar a alumnos que <No alcanzan los aprendizajes requeridos> y que están <Próximos a los aprendizajes requeridos>. El porcentaje de clasificaciones incorrectas de los clasificadores se corresponde casi en total con la clase de calificación inmediata superior o inmediata inferior. Sólo C4.5 clasificó a un 0.8% de alumnos que <No alcanzan los aprendizajes requeridos> como que <Dominan los aprendizajes requeridos> y en los casos de modelos restantes el porcentaje fue mucho menor.
- Para poner en contexto a la **Tabla 56**, la regresión logística con regularización de Lasso tiene una Exactitud del 68% en la clasificación de alumnos que están <Próximos a alcanzar los aprendizajes requeridos>, pero 2 de cada diez se

pueden clasificar de forma incorrecta como que <Alcanzan los aprendizajes requeridos> y otro puede ser clasificado incorrectamente como que <Domina los aprendizajes requeridos>.

- Siguiendo con los alumnos de las clases que <Alcanzan los aprendizajes requeridos> y <Dominan los aprendizajes requeridos>, Random Forest, C4.5, XGBoost Random Forest, CatBoost, Neural Network y la Regresión Logística con Regularización de Lasso tienen los porcentajes más altos de las clasificaciones incorrectas entre una clase y otra, aunque se resalta que los alumnos con estas calificaciones tienen poco riesgo de un bajo rendimiento.
- En general los métodos de aprendizaje automático en conjunto son los mejores clasificadores para casos de alumnos con los rendimientos más bajos.

En la **Tabla 57**, se colorean y muestran los resultados de Gain Ratio (ver sección **2.3.3.4**) y el Factor de Alivio o ReliefF, con base en la semejanza en el orden e importancia de la información reducida a 15 componentes. Gain Ratio mide la información o reducción de la incertidumbre (entropía) del promedio anual. La reducción de la dimensionalidad aminoró los tiempos de entrenamiento y prueba de los modelos, sin embargo, resulta abstracto conocer a detalle los valores de las multiples características expresadas mediante sus componentes principales.

Si bien la Regresión Logística con Regularización de Lasso, tiene para sus métricas a los valores menos favorables respecto de los demás modelos, combinando los widgets Rank (2015r) y Data Table (2015e) de Orange se exploran los coeficientes de Logit del modelo y por ende la incidencia de cada valor entre las características más influyentes.

Componente	Gain ratio	ReliefF
PC1	0.121	0.047
PC2	0.060	0.055
PC3	0.033	0.038
PC4	0.132	0.049
PC5	0.126	0.031
PC6	0.061	0.020
PC7	0.061	0.019
PC8	0.067	0.010
PC9	0.065	0.032
PC10	0.075	0.014
PC11	0.019	0.019
PC12	0.061	0.019
PC13	0.022	0.029
PC14	0.059	0.021
PC15	0.042	0.028

Tabla *57*: Gain Ratio y ReliefF de los datos con dimensionalidad reducida

En la **Tabla 58** se muestran los 10 mayores y 10 menores coeficientes Logit para las clases <Próximo a alcanzar los aprendizajes requeridos> y <No alcanza los aprendizajes requeridos> abreviados como PARA y NAAR. En la **Tabla 58** se muestran los 12 mayores y 12 menores coeficientes para las clases <Domina los

aprendizajes requeridos> y <Alcanza los aprendizajes requeridos> abreviados como DAR y AAR.

Tabla *58*: Coeficientes Logit de la Regresión Logística con la Regularización de Lasso, sobre las clases <Próximo a alcanzar los aprendizajes requeridos> y <No alcanza los aprendizajes requeridos>

	Características	PAAR		Características	NAAR
1	Enfermedad=HIPOTIROIDISMO	5.43	1	ProyEscPQP2=EX	3.26
2	ProyEscPQP1=R	3.51	2	ProyEscSQP1=MB	3.19
3	ProyEscSQP3=R	3.50	3	ComportamientoPQP2=B	2.99
4	Ocup. Padre=AMA DE CASA	3.25	4	ProyEscPQP3=EX	2.31
5	Ocup. Representante=ESTUDIANTE	2.84	5	ProyEscSQP3=MB	2.30
6	EstadoCivilPadre=CASADO/A	2.59	6	EscolaridadPadre=PRIMARIA	1.75
7	Ocup. Padre=PERIODISTA	1.87	7	ComportamientoPQP3=B	1.57
8	EscolaridadPadre=SECUNDARIA	1.65	8	EstadoCivilMadre=UNIÓN LIBRE	1.24
9	Ocup. Padre=SOLDADOR/A	1.56	9	ComportamientoSQP1=B	1.18
10	Ocup. Madre=ING. COMERCIAL	1.50	10	ProyEscPQP1=MB	1.11
11	EstadoCivilMadre=CASADO/A	-3.12	11	SBU=Menos de 1 SBU	1.10
12	ComportamientoSQP2=A	-3.33	12	Materia abrev.=Inglés	-2.67
13	ComportamientoPQP2=C	-3.41	13	Materia abrev.=EE SS	-2.92
14	Ocup. Padre=MARINERO	-3.58	14	Materia abrev.=Arte, Cultura	-2.94
15	Materia abrev.=CC NN	-3.73	15	Materia abrev.=Lenguaje	-2.94
16	t_anioBásico=4. Cuarto	-4.54	16	Materia abrev.=Matemática	-3.23
17	Materia abrev.=EE SS	-4.73	17	ComportamientoPQP1=B	-3.32
18	Materia abrev.=Inglés	-5.29	18	Materia abrev.=Ed. F	-4.02
19	Materia abrev.=Arte, Cultura	-8.39	19	EscolaridadPadre=SECUNDARIA	-4.84
20	Materia abrev.=Ed. F	-9.95	20	ProyEscPQP2=MB	-5.80

De la **Tabla 58** se extrae como información que los alumnos <Próximos a alcanzar los aprendizajes requeridos> evidencian enfermedad especialmente Hipotiroidismo, según sus calificaciones Regulares (R) en sus proyectos escolares no han desarrollados sus habilidades sociales, sus madres son amas de casa por sobre otras ocupaciones, sus padres son estudiantes por sobre ocupaciones como periodista o soldador. Respecto de los coeficientes negativos de la tabla, por ejemplo, que Ciencias Naturales incide menos que las materias mostradas en las filas 17 a 20. Por consecuencia las materias que inciden más para llegar a esta situación serían Matemática y Lenguaje.

De la **Tabla 58** se extrae como información que los alumnos que <No alcanzan los aprendizajes requeridos> si bien tienen buenas habilidades sociales según sus calificaciones muy buenas y excelentes en sus proyectos escolares, su comportamiento no es el más alto, sus papás alcanzan la escolaridad primaria, sus mamás son de estado civil unión libre y el ingreso familiar más común es de menos de 1 salario básico unificado. Respecto de los coeficientes negativos de la tabla, por ejemplo, se deduce que inciden las materias, pero menos que las características con coeficientes positivos, especialmente incide menos la Educación Física.

Tabla 59: Coeficientes Logit de la Regresión Logística con la Regularización de Lasso, sobre las clases <Domina los aprendizajes requeridos> y <Alcanza los aprendizajes requeridos>

Nº	Características	DAR	Nº	Características	AAR
1	Materia abrev.=Ed. F	5.87	1	ProyEscPQP2=EX	3.26
2	Ocup. Padre=TÉCNICO-MECÁNICO/A	4.98	2	ProyEscSQP1=MB	3.19
3	Ocup. Madre=CONTADOR/A	4.13	3	ComportamientoPQP2=B	2.99
4	Ocup. Padre=JUBILADO	4.10	4	ProyEscPQP3=EX	2.31
5	Ocup. Padre=INSPECTOR MUNICIPAL	3.54	5	ProyEscSQP3=MB	2.30
6	Ocup. Padre=PINTOR/A DE AUTOMÓVILES	3.29	6	EscolaridadPadre=PRIMARIA	1.75
7	Materia abrev.=Arte, Cultura	3.11	7	ComportamientoPQP3=B	1.57
8	Ocup. Madre=COCINERO/A	2.99	8	EstadoCivilMadre=UNIÓN LIBRE	1.24
9	Ocup. Padre=TECNÓLOGO-ELÉCTRICO	2.93	9	ComportamientoSQP1=B	1.18
10	Enfermedad=BRONQUITIS	2.78	10	ProyEscPQP1=MB	1.11
11	t_anioBásico=4. Cuarto	2.65	11	SBU=Menos de 1 SBU	1.10
12	Ocup. Representante=LICENCIADO/A	2.64	12	EscolaridadRepresentante=SUPERIOR	-1.61
13	Ocup. Representante=SECRETARIA	-1.82	13	Ocup. Padre=CARPINTERO/A	-1.70
14	Enfermedad=INFECCIONES VIRALES	-1.82	14	Ocup. Padre=INSPECTOR MUNICIPAL	-1.75
15	Ocup. Madre=FARMACEUTICO/A	-1.83	15	Ocup. Padre=JUBILADO	-1.84
16	ProyEscSQP3=R	-1.87	16	Ocup. Representante=LICENCIADO/A	-2.06
17	Ocup. Representante=ABOGADO/A	-2.17	17	Ocup. Representante=ARTISTA	-2.08
18	Ocup. Padre=PINTOR	-2.30	18	Ocup. Padre=TÉCNICO-MECÁNICO/A	-2.57
19	Ocup. Madre=COMERCIO EXTERIOR	-2.31	19	Ocup. Madre=CONTADOR/A	-2.64
20	ProyEscPQP3=B	-2.34	20	ReprobadoRepetido=NO	-2.80
21	Enfermedad=CONVULSIONES	-2.52	21	Ocup. Representante=ATENCIÓN AL CLIENTE	-2.81
22	ComportamientoSQP2=C	-2.67	22	Ocup. Padre=TECNÓLOGO-ELÉCTRICO	-3.35
23	Ocup. Representante=MODELO	-3.07	23	Enfermedad=BRONQUITIS	-4.51

Nº	Características	DAR	Nº	Características	AAR
24	Ocup. Representante=GUARDIA	-3.33	24		
25	ReprobadoRepetido=SI	-3.84			

De la **Tabla 59** se extrae como información que los alumnos que <Dominan los aprendizajes requeridos> lo hacen mayormente en Educación Física y Arte (Nº 1, 7), las ocupaciones del padre y de la madre suelen ser diversas (Nº 3, 4, 5, 6, 8, 9, 12). Respecto de los coeficientes negativos de la tabla, mayormente no son reprobados o repetidores (Nº 25), no tienen comportamiento de calificación C, por consecuencia tienen A o B que son los más altos.

De la **Tabla 59** se extrae como información que los alumnos que <Alcanzan los aprendizajes requeridos> tienen buenas habilidades sociales según los proyectos escolares (Nº 1, 2, 4, 5, 10), tienen muy buen comportamiento, pero no el más alto que es A (Nº 3, 7, 9) lo hacen mayormente en Educación Física y Arte (Nº 1, 7), las ocupaciones del padre y de la madre suelen ser diversas (Nº 3, 4, 5, 6, 8, 9, 12) y el ingreso familiar ronda 1 sueldo básico unificado. Respecto de los coeficientes negativos de la tabla, mayormente se refieren a las ocupaciones de los padres y en este grupo suelen estar los alumnos que han sido reprobados o repetidores y que ahora aprueban con poco más de lo justo.
En el siguiente enlace se puede apreciar la gráfica del árbol de decisión C4.5 para datos reducidos con PCA https://tinyurl.com/5n7sh6ta. No se insertó en este documento dado el tamaño de la imagen.

3.6. Fase 6. Despliegue

Esta es la fase final del proceso de análisis, tiene como objetivo presentar los resultados, es decir, las conclusiones del análisis traducidas en un beneficio para el cliente que lo ha encargado (Nelli, 2018). En entornos técnicos o científicos, se traduce en soluciones de diseño o publicaciones científicas, las cuales se listan en la sección **1.6. Publicaciones**.

En las fases precedentes se presentó modelos agrupados en cinco ejes:

- Modelos no supervisados
- Modelos de clasificación considerando notas intermedias
- Modelos de clasificación sin considerar notas intermedias

- Modelos de regresión sin considerar notas intermedias
- Modelos de clasificación con datos reducidos en dimensionalidad con PCA, sobre muestreados con Smote (ponderado) y sin considerar notas intermedias

Además, la Fase 2 aportó visualizaciones que ayudaron a comprender e interpretar los resultados de los modelos en las Fases 4 y 5.

En una eventual implementación, con independencia del software por emplear, aunque mencionando que en caso de Orange Data Mining existe plena compatibilidad con Python, lo cual facilita la labor de la construcción de un sistema web, se sugiere tomar al menos las siguientes acciones que aseguren un despliegue correcto.

- Integrar en o a los sistemas transaccionales la información socioeconómica, incluso de modo histórico por cada año básico que los alumnos cursen. Esto posibilita usar modelos como series temporales y aprovechar sus ventajas para estudios de investigación de diseño longitudinal.
- Entre las ventajas que otorga llevar el registro histórico está la facilidad para medir el impacto de los factores socioeconómicos y otros datos de los que se disponga longitudinalmente entre diversas cohortes en las instituciones educativas.
- Ejecutar de forma periódica los modelos de aprendizaje automático para leer e incorporar nuevos registros desde los sistemas transaccionales a los formatos tabulares de la minería de datos educativos y el aprendizaje automático en general.
- Cumplimentar la preparación de los datos desde las herramientas de modelado, prevenir los problemas más comunes para cada columna, respecto del tratamiento de valores nulos, erróneos, atípicos u otros. De este modo se consigue atender anomalías no presentadas en los datos objeto de la presente investigación.
- Cumplir los supuestos estadísticos y no estadísticos de cada algoritmo, entrenar a los modelos y buscar las mejores métricas, en especial para las clases de interés.
- Ajustar con periodicidad los hiperparámetros de los modelos en aras de lograr resultados significativos en menos tiempo, costos, esfuerzo y recursos.

- Posibilitar la incorporación de visualizaciones interactivas con base en la navegación del usuario en un sistema web apropiado y las tareas de exploración de datos que se muestran en este documento, sección **3.2.3. Exploración de datos**, junto con la información de patrones que muestran los modelos no supervisados. Esto es de ayuda para brindar respuestas sustentadas y en tiempo real ante inquietudes de docentes, alumnos u otros actores educativos. Además, del cumplimiento de posibles nuevas normativas respecto del uso transparente de datos y de las decisiones concernientes a ellos.
- Posibilitar la incorporación de visualizaciones de los valores de Shapley, por su contribución marginal promedio de un valor de característica en todas los modelos dónde se incorpore, pues aquello aumenta la interpretabilidad de los modelos y genera confianza en los usuarios de estos.
- La información de los ítems anteriores es susceptible de semaforizar o en función de los riesgos académicos o del tipo de promedio por alcanzar.

Una de las opciones sugeridas por la literatura es que los resultados del modelo se envíen al almacén de datos, lo cual es de mencionar, aunque en el caso de las escuelas públicas de Ecuador estas decisiones pasan por entes gubernamentales. Lo cierto, es que, cumplidos los ítems previos, pueden suscitarse algunas cuestiones:

- Como en el principio del año básico no se tiene las calificaciones, para un eventual diseño se puede optar por una interfaz que presente al interesado, por ejemplo, el psicólogo, una predicción a modo de advertencia con los (a) Modelos de clasificación sin considerar notas intermedias, (b) Modelos de regresión sin considerar notas intermedias ó (c) Modelos de clasificación con datos reducidos en dimensionalidad con PCA, Smote ponderado y sin considerar notas intermedias. Dentro de este grupo de modelos, las opciones recomendadas son los métodos de ensamblado (1) AdaBoost, (2) Gradiente XGBoost, (3) Gradiente Boosting, (4) Random Forest, (5) CatBoost y (6) XGBoost Random Forest, además de (7) Neural Network. Cómo en ese momento aún no se dispone de las calificaciones de los comportamientos o de los proyectos escolares, estos se deben de imputar, de sugerencia con kNN, con base en la información almacenada en los modelos. En la sección **3.3.2.** referida a la limpieza de los datos, se sugirió opciones de imputación para valores faltantes, con base en los datos disponibles para esta investigación.

- Aunque la sugerencia de emplear modelos es hacia los de ensamblado, dado los resultados reportados en la Fase 5, es de esperar que estos varíen conforme se carguen nuevos datos, por lo que una aplicación deberá de leer las métricas con los entrenamientos más actuales y a partir de allí seleccionar con cual modelo predecir el resultado para los datos que se estén introduciendo.
- Como el enfoque actual es medir la incidencia de los factores socioeconómicos en el rendimiento escolar, la lectura de métricas para tareas de clasificación que se ha indicado en el ítem precedente, se debe de enfocar en las clases que representan riesgos que actualmente son <No alcanza los aprendizajes requeridos> y <Próximo a alcanzar los aprendizajes requeridos>

La **Figura 56** ilustra sobre una posible interfaz del sistema, en este caso una exploración individual de cada alumno con visualizaciones basadas en el entrenamiento de los modelos:

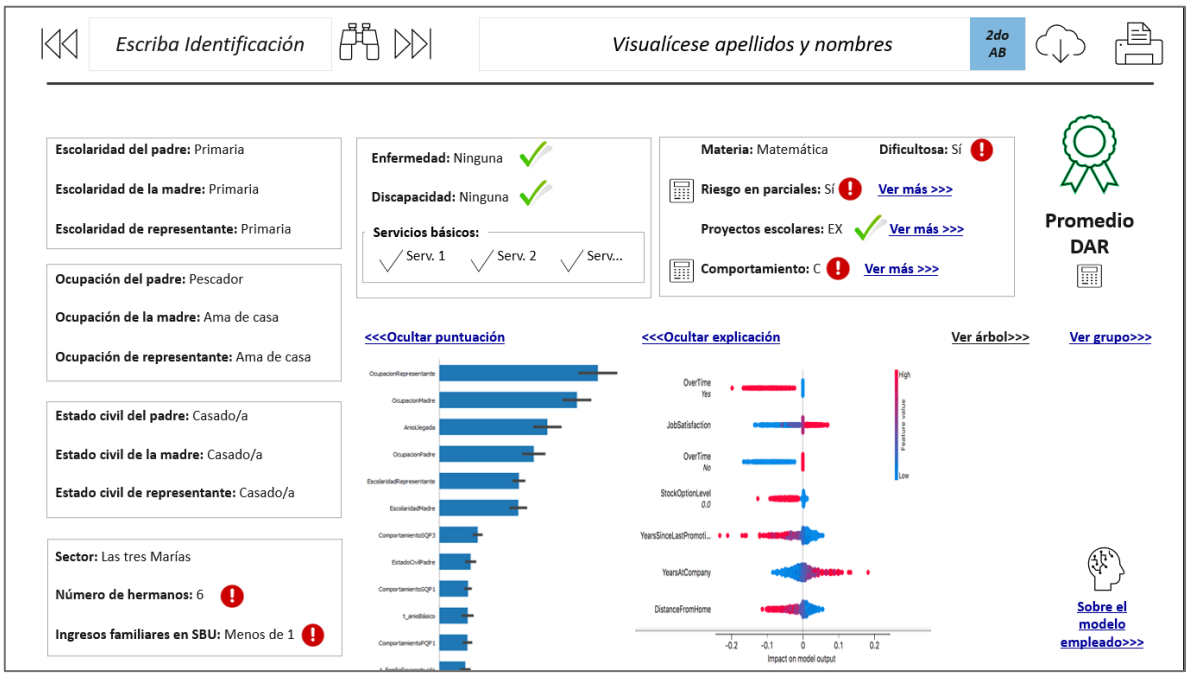


Figura 56: Imagen ilustrativa de la interfaz de la aplicación de aprendizaje automático

La interfaz anterior lee los valores de las características a partir de una identificación, que puede ser la cédula, DNI, Pasaporte u otro dato disponible en un sistema o aplicación transaccional. A partir de allí muestra los apellidos, nombres y año básico cursado por el alumno. Además de los restantes datos agrupados por afinidad entre ellos.

Se ha agregado una calculadora a la izquierda de la etiqueta de cada característica con la finalidad de orientar al usuario de sí ese es un dato almacenado en el sistema transaccional o es un dato imputado y mostrado en ese momento. Cuando compete se agrega un ícono a la derecha de los valores para ilustrar sobre lo ideal o riesgoso del valor de la característica. Los enlaces con leyenda "Ver más >>>" permiten obtener detalles, por ejemplo, las calificaciones de comportamiento en cada parcial, las calificaciones de proyectos escolares en cada parcial.

El ícono de medalla adoptaría colores con base en una semaforización de riesgos a causa del promedio, mismo que puede ser producto de la predicción lo cual se indica con un ícono de varita mágica o es un dato almacenado en el sistema transaccional.

En la parte inferior de la pantalla están los componentes de interpretabilidad que ayudan a transparentar ante el usuario la decisión del modelo. El primer gráfico oculta o muestra la puntuación e importancia de cada característica en la decisión del modelo. El segundo gráfico oculta o muestra las características y valores que influyen en la decisión con base en los valores de Shapley. Un tercer gráfico mostraría la estructura del árbol de decisión en caso de que la regresión o clasificación se efectúe con C4.5. El quinto gráfico ilustraría mediante un diagrama de dispersión sobre a qué clúster de alumnos pertenecería el alumno en cuestión. Un sexto gráfico sería una ilustración pre almacenada que muestra que algoritmo se ha empleado en la clasificación o regresión.

Para el caso de tareas de regresión se construiría una interfaz similar que mostraría el posible puntaje del alumno con base en los valores de las características disponibles, junto con los gráficos que apoyan la interpretabilidad del modelo y las métricas más orientadoras, que serían RMSE o R^2.

Capítulo 4:

4. Resultados

Los resultados de la investigación se han presentado de forma progresiva en cada fase de CRISP-DM documentada en el capítulo 3 y se han complementado con reflexiones documentadas en el capítulo de Conclusiones, Reconocimiento de Limitaciones y Planteo de Trabajos Futuros. Con base en lo expresado, a continuación, se sintetizan algunos aspectos:

- En las Fases de modelado, evaluación y despliegue de CRISP-DM, se presentó diversos algoritmos agrupados en cinco ejes: (1) Modelos no supervisados, (2) Modelos de clasificación considerando notas intermedias, (3) Modelos de clasificación sin considerar notas intermedias, (4) Modelos de regresión sin considerar notas intermedias y (5) Modelos de clasificación con datos reducidos en dimensionalidad mediante PCA, sobre muestreados con Smote y sin considerar notas intermedias.
- Con esos ejes se abordó a las tareas de regresión y clasificación, a la posibilidad de incluir o no las calificaciones, a la reducción de la dimensionalidad para favorecer los tiempos de entrenamiento y prever la indisponibilidad de ciertos datos. Además, se abordó posibles problemas de predicción derivados del desbalanceo de instancias respecto de las clases.
- La información resultante de los modelos se combinó con el aporte de la revisión sistemática de la literatura. De modo general, los métodos de ensamblado reportaron los mejores valores en las diversas métricas, por lo tanto, técnicamente, es conveniente interpretar que las clasificaciones y predicciones logradas son confiables y no casuales, es decir, que en realidad reflejan los patrones en los datos, porque en tales métodos se empleó 50 estimadores o árboles de decisión CART.
- Los resultados reportados por los métodos de ensamblado son en general favorables y estables entre los cinco ejes de análisis mencionados, por

consecuencia se puede indicar que los mismos favorecen la concepción de una ocurrencia de predicciones confiables la mayor parte del tiempo, tanto para las tareas de clasificación como de regresión.

- Los resultados han combinado la objetividad de las métricas en las tareas de clasificación y regresión, con la subjetiva pero importante interpretabilidad de los resultados, apoyados en estudios referidos a técnicas de puntuación de características por permutación, por algoritmos de relieve (ReliefF) y por puntuación con base en los valores de Shapley, que a su vez tienen soporte en la teoría de juegos, dónde una predicción se puede explicar asumiendo que cada valor de las características de la instancia es un jugador en un juego donde la predicción es el pago.
- En este documento se ha insistido en la importancia de la interpretabilidad de los resultados de los modelos, porque, aunque los usuarios de este sistema no necesariamente van a tener formación en ciencia de datos, los modelos deben ser interpretables por los usuarios posibles al tiempo de fortalecer su confianza en las decisiones de los modelos de las instituciones escolares.

En el siguiente capítulo se desagregan los ítems precedentes y se los relaciona con los objetivos de la investigación doctoral.

Capítulo 5:

5. Conclusiones, limitaciones y trabajos futuros

Al término de la presente tesis doctoral y con base en la experimentación realizada mediante los análisis descriptivos y predictivos, además de su comparativa favorable con el análisis confirmatorio, abordando tanto la parte conceptual, como empírica relacionada con cada algoritmo que formó parte del flujo o modelo en Orange, así como antecedentes relevantes reportados en la literatura, fue posible llegar a las conclusiones que se plantean en las siguientes secciones.

5.1. Respecto del objetivo de reconocer las aplicaciones de análisis de datos en los problemas del contexto educativo escolar

Aplicados los modelos de análisis de datos, junto con la revisión sistemática de la literatura, es posible describir el problema del rendimiento escolar como multifactorial, también, que los estudiantes que tienen diversos tipos de dificultades en esta etapa suelen tener problemas futuros de adaptación con la sociedad. Pero no todo es negativo, porque se han observado estudios primarios con soluciones a este problema desde múltiples aristas.

En lo referente a los alumnos, que en esta etapa educativa viven su niñez, se les puede orientar a la autorreflexión sobre su situación académica y gestionarles retroalimentación que les ayude en la mejora de su rendimiento, esto en concordancia con su temprana edad, que en los primeros años básicos se relaciona con un desarrollo afectivo, cognitivo y psicomotriz, en tanto que en los últimos años escolares se corresponden con el desarrollo de sus destrezas y técnicas de estudio.

En lo referente a los docentes, ellos pueden comprender de modo colectivo los procesos de aprendizaje de los alumnos e incluso reflexionar sobre sus propios métodos de enseñanza, sin descuidar aspectos relacionados con los niños, tales como sus habilidades sociales que pueden incidir sobre sus conductas y rendimiento, además, de la estructura familiar, factores socioeconómicos y el ambiente de

estudios en el hogar. En este sentido, ha llamado la atención la tendencia en los niños a obtener mejores calificaciones en los parciales del segundo quimestre, por sobre los del primer quimestre, algo que invita a observarse por parte de los profesionales competentes o por mediación de los padres.

En esta tesis, además de haber presentado los resultados del análisis en las Fases de Modelado y Evaluación de la metodología CRISP-DM (Capítulo 3), se ha hecho énfasis, en la Fase de Despliegue, de mejorar las formas de captura y preparación de los datos para los modelos de un modo recurrente, porque las revisiones y aumentaciones periódicas de mejoras en el entrenamiento repercuten en sus predicciones más efectivas. Es necesario capturar la mayor cantidad de datos posibles incluso para que las imputaciones de datos faltantes sean más fieles con la realidad de cada registro, que representa la situación del alumno en cada materia cursada en la escuela. Esto lleva a comprender mejor cada situación desde la perspectiva de distintos datos sociales, académicos e incluso de situaciones fortuitas como lo es la pandemia del Covid-19 y sus secuelas.

Estas aplicaciones, debe contar con el apoyo de directivos institucionales e incluso estatales, además, es apropiado de que estén mediados por personal de formación en análisis de datos e incluso en Storytelling, pero sin dejar de lado la construcción de modelos interpretables, como se ha insistido en la Fase de Evaluación del capítulo 3, referente al desarrollo de los modelos. Es deseable que cada vez los modelos de aprendizaje automático presenten interfaces amigables que cuenten historias para "conectar" con sus usuarios académicos.

5.2. Respecto del objetivo de preparar los datos de acuerdo con la dimensionalidad a un número efectivo de características

Aunque en esta tesis se presentó la opción de reducir la dimensionalidad de los datos mediante el análisis de componentes principales (PCA), expresándolos en 15 componentes y con un 30% de varianza explicada, siendo en principio 88 características, durante el capítulo 2, referido al marco teórico, además de explicar ventajas, desventajas y tareas comunes de preparación de datos para cada algoritmo en el modelo, también se detalló ciertas características que cada algoritmo incorpora y que pueden ser empleadas para la selección de características de un modo en primera instancia manual pero configurable mediante programación.

De cierto modo esta es una selección de características más controlada y válida como opción, como aporte a una mayor interpretabilidad de los modelos. Tal es el caso de

C4.5 o las regresiones lineales y logísticas con regularización de Lasso.

En ambos casos, la reducción de la dimensionalidad aporta con tiempos más cortos para el entrenamiento de los modelos. Además, de que puede evitar la necesidad de imputar ciertos datos durante la predicción de un caso concreto de un nuevo alumno, porque a lo mejor un dato faltante no forma parte de los datos reducidos en dimensionalidad.

La naturaleza multifactorial del problema del rendimiento escolar posibilita diversas perspectivas de análisis, como, por ejemplo, observar el tipo de estructura familiar del alumno, el estado civil del padre y de la madre, la escolaridad de los padres, el ingreso familiar vs la ocupación de los padres, número de hermanos en casa, disponibilidad de servicios básicos, posibles enfermedades, dificultad en aprender materias específicas, la distancia casa – escuela, entre otros.

Cuando estos multiples factores son objetos de análisis corresponde intercambiar los roles entre características (datos) predictoras y de respuestas en varias ocasiones, e incluso generar nuevas características a partir de las existentes. Si bien el conocimiento del contexto escolar es importante, también lo es una adecuada selección de características y reducción de la dimensionalidad de estas mediante técnicas provenientes del álgebra o la programación recursiva.

Con esto se obtienen visualizaciones de datos que proyecten de forma simplificada a un conjunto de características, que en el presente caso se partió de 88, ante una clasificación o regresión, con el fin de adaptarse mejor a un modelo predictivo en términos de rendimiento computacional y entendimiento, como los que conllevan árboles decisión, máquinas de soporte vectorial o bosques aleatorios. Todo ello sin descuidar métricas como la precisión y tasa de recuerdo de los modelos predictivos.

Además de la dimensionalidad y de más tareas comunes de preparación de los datos, resulta útil muestrear adecuadamente los casos para el entrenamiento de los modelos. Por ejemplo, para estudiar a las familias monoparentales y el rendimiento académico, comportamiento, materia dificultosa u otro aspecto relacionado, comprendiendo que los modelos de predicción y clasificación funcionan mejor con clases aproximadamente balanceadas, se sobre muestreó la minoría de familias monoparentales para aumentar el número de casos y mejorar la tasa de precisión y recuerdo de los modelos de predicción de problemas por esta causa.

5.3. Respecto del objetivo de estudiar comparativamente la idoneidad de los algoritmos de minería de datos.

Se ha mencionado que existe aplicación de la minería de datos para múltiples contextos y que el educativo es sólo uno de ellos, por lo que es justo expresar la dificultad de generalizar efectivamente aún a través de muchos análisis de datos diferentes, porque cada uno de los cuales tiene aspectos únicos importantes. Entonces para considerar una posible idoneidad se partió de la base reportada por la revisión sistemática de la literatura que forma parte de esta tesis. La revisión sugiere técnicas como Redes Neuronales (en especial el Perceptron Multicapa y las Redes de Base Radial), Máquinas de vectores de soporte (SVM), Árboles de decisión, Reglas de Asociación, Regresión Lineal (lineal, multinomial, jerárquica), Regresión Logística y Naïve Bayes.

Otro de los aspectos en los que se enfatizó fue en la posibilidad de interpretabilidad de cada algoritmo, haciendo posible el entendimiento y por ende generar acciones de mejora del rendimiento académico, a nivel de las tareas de clasificación y regresión, así como la posibilidad de transparentar las decisiones del modelo en general. Los árboles de decisión, regresiones lineales y logísticas, reglas de asociación o el análisis discriminante lineal (LDA) reportan por su naturaleza una mayor interpretabilidad. Para los otros casos se asume un costo computacional alto a nivel de entrenamiento, pero se consigue puntuaciones de características y valores que aumentan la interpretabilidad.

Con base en los ítems precedentes, se buscó potenciar algunos aspectos, por ejemplo, el uso de métodos de aprendizaje en conjunto o ensamblados, que a su vez se apoyan en otros muy recurridos como lo son los árboles de decisión CART, es decir, se basan en la potenciación del gradiente y es de esperar que al tratarse de métodos relativamente nuevos e iterativos aporten ventajas respecto de sus antecesores. Aún dentro de ellos, por poner un ejemplo, CatBoost logra mejores resultados cuando se emplea con características de entrada mayormente de tipo categórico.

Con base en lo expuesto, se indica que, para las tareas de clasificación, los métodos de aprendizaje en conjunto resultaron como los más eficaces con leves diferencias de valores entre las métricas de cada uno. Para las tareas de regresión si hubo diferencias, en favor de la regresión lineal con Regularización de Lasso y de Ridge, aunque a nivel de RMSE la diferencia con CatBoost fue de apenas centésimas. Se

reconoce que una simulación paramétrica más profunda puede ayudar a mejorar los resultados de algoritmos como el Perceptron Multicapa, cuyo tiempo de ejecución resultó siempre alto en los entrenamientos efectuados.

Finalmente, la idoneidad de cada algoritmo está ligada al cumplimiento de los supuestos estadísticos y no estadísticos de cada uno de ellos. Además, pese a adoptar una metodología como CRISP-DM, que sigue un proceso iterativo y que contempla vueltas hacia atrás para llevar de mejor manera el modelado, es cada algoritmo el que posiblemente obligue a hacer revisiones particulares conforme a sus características.

5.4. Respecto del objetivo de establecer parámetros e hiperparámetros que pueden ser apropiados a los datos y los modelos

Resulta primordial que previo del establecer hiperparámetros y parámetros de los modelos, se realice una minuciosa preparación de los datos, pensada en cada columna y con independencia de los casos de requerimientos de imputación presentados en la Fase de Preparación de los Datos del Capítulo 3 de esta tesis. Por ejemplo, la imputación de una calificación faltante en una materia X podría realizarse en función del género del alumno en el caso de que se haya determinado que las calificaciones en dicha materia suelen diferir en función del género, tal como se documentó en la revisión sistemática de la literatura incorporada parcialmente en el Capítulo 2. La imputación podría efectuarse con KNN que ahora valoraría a instancias vecinas más pertinentes. Además, de cumplir con los supuestos estadísticos y no estadísticos de cada modelo que se utilice. Con ello se consiguen predicciones con mejores tasas, pero por sobre todo más significativas, aunque el tiempo del proceso de entrenamiento se vea levemente incrementado.

Partiendo de la premisa de que los hiperparámetros con alta capacidad de ajuste son de mayor importancia para el entrenamiento de los modelos, sino se logra valores mínimos esperados en las métricas respectivas con la parametrización predeterminada, recién se debe continuar con la optimización de los hiperparámetros del modelo, debido a la combinación y complejidad que se puede suscitar, por ejemplo, en la Fase de Evaluación del Capítulo 3, no se documentó al clasificador Naïve Bayes y al Stochastic Gradient Descent (SGD) en las tareas de regresión porque pese a efectuar diversas combinaciones de hiperparámetros tal como sugiere la literatura y la misma naturaleza de los datos, los valores obtenidos

con la métrica del Error cuadrático medio de la raíz (RMSE), cuyo error es asociable con los puntos en el promedio, es decir, si el error es 0.80, entonces el error de SGD está por el orden de 0.80 puntos sobre 10 posibles. En todo caso el error siempre fue mayor a 10, por ende, inapropiado.

En esta tesis no se realizó una simulación paramétrica, pero se reconoce lo recomendable que es optar por ajustes óptimos y automatizados con librerías apropiadas y que conduzcan a un mejor rendimiento con una validación cruzada o el muestreo por estratos, que son los dos mecanismos que se han empleado. Además, con la optimización se consiguen modelos que reducen la función de pérdida predefinida.

No se debe descartar los tiempos de entrenamiento, porque también repercuten en las tarifas de un eventual contrato por demanda de servicios de Cloud Computing para proyectos de aprendizaje automático, por ejemplo, la Regresión Logística con Regularización de Ridge ocupó cerca del 30% del tiempo de entrenamiento entre 13 modelos para las tareas de clasificación considerando notas intermedias. Este modelo, junto con la Regresión Logística con Regularización de Lasso, Neural Network (Perceptron Multicapa) y Boosting, ocuparon un 80% del tiempo de entrenamiento en las tareas de clasificación sin considerar notas intermedias.

Por contraparte, las Neural Network (Perceptron Multicapa) reportaron menores tiempos de entrenamiento en las tareas de clasificación con PCA, Smote ponderado y sin considerar notas intermedias, aunque la Exactitud, Precisión, Recuerdo, Especificidad y F1 fueron siempre menores que los métodos de ensamblado, especialmente XGBoost, AdaBoost, Boosting y Random Forest (en ese orden). Esto significa que el tiempo de entrenamiento es una importante variable por considerar, más cuando el conjunto de datos tiende siempre a crecer.

En lo que respecta a las tareas de regresión sin considerar notas intermedias, es dónde destacan, como era de esperarse, la Regresión Lineal con Regularización de Lasso y la Regresión Lineal con Regularización de Ridge. Además, se les sumaron CatBoost, Random Forest y kNN (en ese orden) con los mejores valores en sus métricas, siendo las más interpretables RMSE y R^2. Los tiempos de entrenamiento de estos modelos también estuvieron entre los mejores. En el caso del RMSE este fue desde 1.361, 1.386, 1.394, 1.450 y 1.493 puntos respectivamente, si lo desea, esto se puede interpretar como una efectividad de hasta un 85% aproximadamente.

5.5. Respecto del objetivo de interpretar los resultados del conocimiento descubierto y su eficiencia según métricas pertinentes a los modelos

En la Fase 2 de CRISP-DM, referida a la comprensión de los datos, se realizó la exploración de datos y producto de aquello se obtuvieron algunos patrones y relaciones interesantes, previo de la construcción de los modelos:

- Un 11% de alumnos, es decir, 1 de cada 10, tiene al menos un año de atraso en sus estudios.
- En el año más actual de estudios, el 47% de alumnos que han tenido atrasos en sus estudios alcanzan los promedios requeridos y el 51% los dominan. Es decir, obtienen los dos tipos de promedios más altos posibles.
- Las asignaturas en las que más auto reportes de dificultad se dan, son Lenguaje y Matemática. La muestra estudiada no reflejó distingo de esta situación en función del género del alumno.
- Las asignaturas en las que efectivamente más promedios bajos se registran son Lenguaje y Matemática, pero existe apenas un 4% de concordancia entre la dificultad auto reportada en las materias y la dificultad evidenciada con los promedios obtenidos (Ver **Gráfico 14**). Es decir, el dato de la dificultad auto reportada requiere quizá de pruebas diagnósticas más allá de sólo lo indicado en una entrevista psicólogo – padre y alumno.
- La escolaridad del representante del alumno no parece incidir marcadamente en el tipo de promedio que obtiene el alumno, aunque los alumnos que obtuvieron promedios del tipo <Domina los aprendizajes requeridos>, es decir, el más alto posible, estan asociados con padres de escolaridad superior, en un 15% más que los de primaria y un 20% más que los de escolaridad secundaria.
- El 90% de alumnos que auto reportaron al menos una materia como dificultosa, también reportaron por medio de sus padres, un ingreso familiar que ronda entre 3 y menos de 1 salario básico unificado.
- Los promedios del tipo <Próximo a alcanzar los aprendizajes requeridos> y <No alcanza los aprendizajes requeridos>, es decir los menores posibles, se suscitaron en alumnos provenientes de familias que reportaron un ingreso familiar entre 3 y menos de 1 salario básico unificado.
- Los promedios del tipo <Próximo a alcanzar los aprendizajes requeridos> y <No alcanza los aprendizajes requeridos>, es decir los menores posibles, se suscitaron

principalmente en alumnos de los que se reportó provenir de familias reconstruidas.

- Los promedios del tipo <Próximo a alcanzar los aprendizajes requeridos>, se suscitaron principalmente en alumnos que obtuvieron calificaciones Buenas y Regulares en el desarrollo de proyectos escolares, que son insumos de medición de sus habilidades sociales.
- Del grupo de alumnos del que se reportó alguna discapacidad, se reportó calificaciones del tipo <No alcanzan los aprendizajes requeridos>, es decir, la más baja posible.
- Los promedios del tipo <Próximo a alcanzar los aprendizajes requeridos> se suscitan principalmente en los alumnos que registraron al menos un parcial con comportamiento de tipo C, es decir, el más bajo del que se tiene registros en la muestra estudiada.

Respecto de los modelos, como se dispuso de varios algoritmos de aprendizaje automático, resultó importante considerar y presentar una métrica de error adecuada, por ejemplo, en las tareas de regresión efectuadas en esta investigación, el resultado de los algoritmos es un valor numérico, dentro de un conjunto infinito de promedios anuales de los estudiantes como posibles resultados, yendo desde 0 a 10 puntos. Si bien el error cuadrático medio (MSE) acercándose a cero, el error absoluto medio (MAE) acercándose a cero, R^2 acercándose a 1, resultan de orientación sobre que algoritmo seleccionar para predecir valores con la regresión, es el error cuadrático medio de la raíz, RMSE, la métrica más interpretable porque tiene la propiedad útil de estar en las mismas unidades que el promedio anual. Si el valor de RMSE es 1.1 significaba que el valor real se acercaba al predicho hasta en 1.1 puntos. En tal sentido los mejores algoritmos para las tareas de regresión resultaron ser la Regresión lineal con regularización de Lasso, Regresión lineal con regularización de Ridge, CatBoost, Random Forest y kNN, porque con ellos los valores reales se acercan a los predichos en un rango desde 1.36 hasta 1.49 puntos según lo mostrado en la **Tabla 49**.

En continuación con las tareas de regresión y privilegiando la interpretabilidad, se ha visto que el algoritmo (métrica) ReliefF, es eficiente y consciente de la información contextual y puede estimar correctamente la calidad de las características en problemas como el analizado, así, las 13 características de mejor calidad para obtener los valores referidos en el párrafo precedente son la materia, ocupación del padre, ocupación de la madre, género, escolaridad del padre, estado civil de la

madre, ocupación del representante, estado civil del padre, año básico en curso, parentesco del representante, estructura familiar y cantidad de sueldos básicos expresados como el ingreso familiar. Se ha referido a ReliefF como "consciente" de la información contextual porque este algoritmo estima la calidad de las características sobre la base de qué tan bien pueden distinguir entre instancias vecinas. Además, desde el punto de vista de la eficiencia, ReliefF es conveniente porque computacionalmente se ejecuta en tiempo polinómico de bajo orden. Se conoce que una de sus desventajas es que no discrimina entre características redundantes, pero este problema se trató al seleccionar características no redundantes, además de eliminar instancias con valores atípicos que afectan a RMSE con más énfasis a MSE y en menor medida a MAE.

Como también se ha documentado, la Regresión Lineal con Regularización de Ridge y KNN resultaron entre los mejores modelos para las tareas de regresión, por tal razón luego del entrenamiento de los diversos algoritmos se analizó la importancia de las características con base en el RMSE y R^2. En la **Figura 53** se expuso que después del entrenamiento estos dos algoritmos muestran como características más importantes a la ocupación del representante, materia, ocupación de la madre, año de llegada (por consecuencia año básico), ocupación del padre, número de hermanos, escolaridad del representante, distancia casa-escuela, años de retraso, estado civil del padre e ingresos familiares, como las características de mayor impacto en la decisión. A diferencia de ReliefF el cálculo de importancia que se ejecutó con 10 permutaciones se efectuó luego del entrenamiento de los modelos, ReliefF se ejecutó antes. Otra diferencia que penaliza al cálculo de la importancia de las características es el tiempo de procesamiento que demanda, en función de las posibles permutaciones que se escojan.

En lo que respecta a las tareas de clasificación, en general los métodos de aprendizaje en conjunto resultaron como los más eficaces con leves diferencias de valores entre las métricas de cada uno. Cuando se redujo la dimensionalidad de los datos, como era de esperarse los tiempos de entrenamiento y de prueba global se aminoraron hasta en un 75% que cuando se empleó a todas las características. En esas condiciones, además de que las instancias estaban ligeramente desbalanceadas respecto de la clase, XGBoost, AdaBoost, Boosting y Random Forest obtuvieron mejores valores en la Exactitud de la clasificación (CA), Precisión, Recuerdo, F1 y Especificidad. La lista de modelos con mejores métricas difiere entre las tareas de regresión y de clasificación.

Continuando con las tareas de clasificación, cuando no se redujo la dimensionalidad

de los datos, la Exactitud de los seis mejores clasificadores estuvo entre el 93% al 97% en el siguiente orden: XGBoost, Neural Network, Boosting, AdaBoost, Random Forest y CatBoost. Es decir, con diferencia de cuando se redujo los datos sólo se incluyó ahora el Perceptron Multicapa (Neural Network).

Con diferencia de las tareas de Regresión, para las tareas de clasificación se dispuso del cálculo de valores SHAP, que se basan en los valores de Shapley, dónde una predicción se puede explicar asumiendo que cada valor de característica de la instancia es un jugador en un juego donde la predicción es el pago. Así, para los alumnos que no alcanzan los aprendizajes requeridos la lista de atributo y valor se calculó y ordenó como: Estado civil de padre = Unión libre, Número de hermanos = hacia 0, Discapacidad = Si, Comportamiento del primer quimestre P3 = A, Proyecto escolar del segundo quimestre P3 = B, Comportamiento del segundo quimestre P3 = B, Proyecto escolar del primer quimestre P3 = B, Ocupación del padre = Guardia… Es decir, los valores de Shapley contribuyen a una mejor explicación sobre las decisiones tomadas por los modelos de aprendizaje automático, en especial, en las tareas de clasificación.

Con la información el párrafo anterior se responde a la pregunta de investigación de ¿Cómo inciden los factores socioeconómicos en el aprovechamiento escolar? Y con párrafos precedentes se responde a en qué medida (métrica).

5.6. Limitaciones y trabajos futuros

El presente estudio se limitó a dos escuelas, aunque no se realizó comparaciones entre ambas dado que una de ellas facilitó una menor cantidad de información respecto de la otra y por tanto no cabía la comparación sino la unificación de los datos. A futuro, este aspecto puede ser fortalecido al incorporar más escuelas, de distintas regiones para obtener un abordaje más significativo por disponer de más datos y así producir resultados más fiables y extrapolables.

En consonancia con la limitación expuesta, es recomendable que el Estado, el sistema escolar y las familias se empoderen de modo conjunto sobre la importancia del ambiente del hogar, como auxiliar en la disipación de la afectación del rendimiento de los alumnos. De ese modo se pueden generar acciones respaldadas con datos que ayuden a los niños a que su estancia académica escolar sea más agradable, en cuanto a calidad de vida, aprendizajes de calidad y el desarrollo exitoso de la labor docente.

Esta investigación inició con una revisión sistemática de la literatura a partir de las

bases de datos bibliográficas IEEE Explore y Scopus, que eran a las cuales se tenía acceso. Se buscó estudios primarios del tipo artículos y conferencias publicados en inglés o español durante los últimos 10 años. Por lo tanto, a futuro puede ampliarse el campo de revisión a más bases de datos e idiomas, emplear técnicas de inteligencia artificial para la revisión y así refrescar la cantidad y la calidad disponible de literatura.

Continuando con la revisión sistemática, para conseguir una estructura organizada de las revisiones bibliográficas sobre minería de datos educacional se recomienda la actualización periódica de matrices resumen que contengan las investigaciones pertinentes, desde cada factor socioeconómico posible, de la interpretabilidad que ofrezcan los modelos inmersos en los estudios y de situaciones anómalas que puedan fungir de alerta temprana acerca de este problema multifactorial. La intención es disponer de estados del arte concretos que den sustento a nuevos elementos y situaciones de análisis.

En la actualidad muchos de los modelos de aprendizaje automático son percibidos por los usuarios como cajas negras, sin embargo, en esta investigación se intentó presentar resultados interpretables, porque la interpretabilidad es importante porque no todos los usuarios de este tipo de sistemas son de formación en estadística o de ciencia de datos. Entonces resulta imperioso generar soluciones de aprendizaje automático interactivas e interpretativas que con base en normativas emergentes como es el caso del Reglamento General de Protección de Datos (GDPR) de la Unión Europea, transparenten a los usuarios la razón de sus decisiones.

Si bien los datos analizados, se han obtenido en formato tabular siguiendo un detallado proceso de extracción y preparación, se puede ampliar el estudio a otros datos como los de tipo psicológico, estilos de aprendizaje estudiantil, autoeficacia (de la que se ha tomado en consideración la dificultad auto reportada en las asignaturas), metas de logro, motivación, intereses u otros. Esta información puede seguirse mediante encuestas periódicas y el apoyo por software para el registro de su variabilidad en el tiempo. Esto supone más esfuerzo, costos computacionales y económicos, pero aportará más elementos al análisis. Incluso cuando se trate de información que contenga como respuestas a textos cortos, se puede plantear el uso de minería de textos. En el caso de los psicólogos de los departamentos de orientación estudiantiles de Ecuador, usualmente recurren a textos manuscritos por los alumnos, esta es una información que puede ser enriquecida con anotaciones del profesional para tener más insumos de valoración como lo serían las habilidades motoras finas, lenguaje, memoria, concentración, etc. y contrastarlos con los

rendimientos académicos.

La revisión sistemática de la literatura reportó la existencia parcial de estudios similares al presente, pero no con el abordaje desde diversos modelos supervisados y no supervisados. Por tanto, es de esperar soluciones futuras basadas en Inteligencia Artificial para acelerar la ayuda o solución temprana a problemas de rendimiento académico de los niños, pues la literatura evidencia que los posibles problemas en esta primera etapa de la formación educativa pueden impactar diversas aristas de la vida personal y académica a futuro.

Con base en los presentes y futuros resultados de la incidencia de los factores socioeconómicos en el rendimiento escolar, en un sentido más amplio respecto de la minería de datos educativos, la Inteligencia Artificial, puede ayudar a identificar los posibles vacíos en los recursos didácticos de los maestros y sugerir ajustes cuando sea necesario determinar dónde existen estancamientos académicos por parte de los niños, ayudarlos a mejorar y en última instancia a sobresalir.

Desde la perspectiva de la ciencia de datos se abordó técnicas y tecnologías de demanda en la actualidad, aunque se debe reconocer que en el mundo se generan datos e información por y para la sociedad, en distintos escenarios demandantes de eficiencia, productividad e incluso para forjar un estilo de vida ergonómico apoyado cada vez más con tecnologías e información. Y se espera que la ciencia de datos sea tan popular y de tendencia como otras tecnologías relevantes utilizadas en la actualidad.

6. Referencias

A. Singh, N. Thakur, & A. Sharma. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310-1315.

Abdulhamit, S. (2020). *Practical Machine Learning for Data Analysis Using Python*. Elsevier. https://doi.org/10.1016/C2019-0-03019-1

Abu Amra, I. A., & Maghari, A. Y. A. (2017). Students performance prediction using KNN and Naïve Bayesian. *2017 8th International Conference on Information Technology (ICIT)*, 909-913. https://doi.org/10.1109/ICITECH.2017.8079967

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the surprising behavior of distance metrics in high dimensional space*. 420-434.

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. *1215*, 487-499.

AMOS (24.0.0.). (2016). IBM.

Anchundia-Delgado, I. M., Pincay-Ponce, J. I., & Delgado-Muentes, W. R. (2022). La autoestima en los adolescentes que cursan el bachillerato. Realidad y expectativas. *REFCALE: Revista electrónica formación y calidad educativa*, *10*(3), 1-10.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

Azevedo, A., & Santos, M. (2008). *KDD, semma and CRISP-DM: A parallel overview* (p. 185).

Baillie, M., le Cessie, S., Schmidt, C. O., Lusa, L., Huebner, M., & for the Topic Group "Initial Data Analysis" of the STRATOS Initiative. (2022). Ten simple rules for initial data analysis. *PLOS Computational Biology*, *18*(2), e1009819. https://doi.org/10.1371/journal.pcbi.1009819

Baker, R. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118.

Banco Mundial. (2022). *Gasto público en educación, total (% del PIB)—Latin America & Caribbean*.

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., & Mersmann, O. (Eds.). (2023). *Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide*. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5170-1

Basantia, N. C., Nollet, L. M. L., & Kamruzzaman, M. (Eds.). (2019). *Hyperspectral imaging analysis and applications for food quality*. CRC Press, Taylor & Francis Group.

Boedeker, P., & Kearns, N. T. (2019). Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer. *Advances in Methods and Practices in Psychological Science*, *2*(3), 250-263. https://doi.org/10.1177/2515245919849378

Bottou, L. (2018). *Stochastic Gradient Descent (v.2)*. Leon.Bottou.Org. https://leon.bottou.org/projects/sgd

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2016). *Machine Learning Mastery With Python* (1.4). Machine Learning Mastery.

Brownlee, J. (2021a). *Ensemble Learning Algorithms With Python* (1.11). Machine Learning Mastery.

Brownlee, J. (2021b). *Imbalanced Classification with Python* (1.3). Machine Learning Mastery.

Bussaman, S., Nuankaew, W., Nuankaew, P., Rachata, N., Phanniphong, K., & Jedeejit, P. (2017). Prediction models of learning strategies and learning achievement for lifelong learning. *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 192-197. https://doi.org/10.1109/TALE.2017.8252332

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2013). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5-6), 318-331.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chih-Chung, C., & Chih-Jen, L. (2022). *LIBSVM -- A Library for Support Vector Machines*. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Correia-Zanini, M. R. G., Marturano, E. M., & Fontaine, A. M. G. V. (2018). Effects of early childhood education attendance on achievement, social skills, behaviour, and stress. *Estudos de Psicologia (Campinas)*, *35*(3), 287-297. https://doi.org/10.1590/1982-02752018000300007

De Giusti, A. (2020). Book Review: Policy Brief: Education during COVID-19 and beyond. *Revista Iberoamericana de Tecnología En Educación y Educación En Tecnología*, *26*, 110-111.

De La A Muñoz, G. F. (2018). *Análisis del rendimiento académico en los/as estudiantes de octavo año de educación básica de la Unidad Educativa Fiscal "31 de Octubre" del cantón Samborondón, provincia del Guayas, periodo lectivo 2016-2017*.

de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*(3), 273-290. https://doi.org/10.1037/met0000079

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, *14*, 2349-2353.

Duc, T. L., Leiva, R. G., Casari, P., & Östberg, P.-O. (2020). Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. *ACM Computing Surveys*, *52*(5), 1-39. https://doi.org/10.1145/3341145

Dumont, H., Klinge, D., & Maaz, K. (2019). The Many (Subtle) Ways Parents Game the System: Mixed-method Evidence on the Transition into Secondary-school Tracks in Germany. *Sociology of Education*, *92*(2), 199-228. https://doi.org/10.1177/0038040719838223

Elastika, R. W., & Dewanto, S. P. (2021). Analysis of Factors Affecting Students' Mathematics Learning Difficulties Using SEM as Information for Teaching Improvement. *International Journal of Instruction*, *14*(4), 281-300.

El-Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? En *Machine Learning in Radiation Oncology* (pp. 3-11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Entwisle, D. R., & Alexander, K. L. (1992). Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School. *American Sociological Review*, *57*(1), 72. https://doi.org/10.2307/2096145

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905.

Freund, Y., & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. En P. Vitányi (Ed.), *Computational Learning Theory* (Vol. 904, pp. 23-37). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-59119-2_166

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189-1232.

Fu, H., & Qi, K. (2022). Evaluation Model of Teachers' Teaching Ability Based on Improved Random Forest with Grey Relation Projection. *Scientific Programming*, *2022*.

Fu, Q.-K., & Hwang, G.-J. (2018). Trends in mobile technology-supported collaborative learning: A systematic review of journal publications from 2007 to 2016. *Computers & Education*, *119*, 129-143.

Gironés, J., Casas, J., & Minguillón, Julià. (2017). *Minería de datos*. Editorial UOC.

Google Developers. (2022a). *Machine Learning*. k-Means Advantages and Disadvantages. https://tinyurl.com/4jrz6c8f

Google Developers. (2022b). *Machine Learning*. Clustering Algorithms. https://tinyurl.com/4uydaw4k

Grasso, P. (2020). Rendimiento académico: Un recorrido conceptual que aproxima a una definición unificada para el ámbito superior. *Revista de educación*, *11*(20), 87-102.

Grina, F., Elouedi, Z., & Lefevre, E. (2022). Learning from Imbalanced Data Using an Evidential Undersampling-Based Ensemble. En F. Dupin de Saint-Cyr, M. Öztürk-Escoffier, & N. Potyka (Eds.), *Scalable Uncertainty Management* (Vol. 13562, pp. 235-248). Springer International Publishing. https://doi.org/10.1007/978-3-031-18843-5_16

Guo, Z., Min, A., Yang, B., Chen, J., & Li, H. (2021). A Modified Huber Nonnegative Matrix Factorization Algorithm for Hyperspectral Unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 5559-5571. https://doi.org/10.1109/JSTARS.2021.3081984

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, *8*(1), 53-87.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, *2*(3), 349-360. https://doi.org/10.4310/SII.2009.v2.n3.a8

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, 175-199.

Hernández, M. J. (1994). Competencia social: Intervención preventiva en la escuela. *Infancia y Sociedad: Revista de estudios*, *24*, 21-48.

Herrera, F., Charte, F., Rivera, A. J., & del Jesus, M. J. (2016). *Multilabel Classification*. Springer International Publishing. https://doi.org/10.1007/978-3-319-41111-8

Hoffmann, J. P., & Shafer, K. (2015). *Linear regression analysis: Assumptions and applications*. NASW Press, National Association of Social Workers.

Hwang, Y. H. (2019). *Hands-on data science for marketing: Improve your marketing strategies with machine learning using Python and R*. Packt Publishing.

IBM. (2021, agosto 17). *IBM Documentation*. https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Ibragimov, B., & Gusev, G. (2019). *Minimal Variance Sampling in Stochastic Gradient Boosting* (arXiv:1910.13204). arXiv. http://arxiv.org/abs/1910.13204

Karlsson, N. (2021). *Comparison of linear regression and neural networks for stock price prediction*.

Kikuchi, M., Yoshida, M., Okabe, M., & Umemura, K. (2015). Confidence interval of probability estimator of Laplace smoothing. *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1-6. https://doi.org/10.1109/ICAICTA.2015.7335387

Körner, C., & Waaijer, K. (2020). *Mastering Azure machine learning: Perform large-scale end-to-end advanced machine learning in the cloud with Microsoft Azure Machine Learning*. Packt Publishing.

Kotu, V., & Deshpande, B. (2019). Classification. En *Data Science* (pp. 65-163). Elsevier. https://doi.org/10.1016/B978-0-12-814761-0.00004-6

Krumm, A., Means, B., & Bienkowski, M. (2018). *Learning analytics goes to school: A collaborative approach to improving education*. Routledge.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1-5.

Li, Y., & Zhai, X. (2018). Review and Prospect of Modern Education using Big Data. *2017 International Conference on Identification, Information and Knowledge on The Internet of Things*, *129*, 341-347. https://doi.org/10.1016/j.procs.2018.03.085

Liu, J., Liang, G., Siegmund, K. D., & Lewinger, J. P. (2018). Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*, *19*(1), 369. https://doi.org/10.1186/s12859-018-2401-1

Liu, J., Peng, P., & Luo, L. (2020). The Relation Between Family Socioeconomic Status and Academic Achievement in China: A Meta-analysis. *Educational Psychology Review*, *32*(1), 49-76. https://doi.org/10.1007/s10648-019-09494-0

Lundberg, S. (2018). *SHAP*. API Reference. https://tinyurl.com/yhcdt2w8

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://doi.org/10.48550/ARXIV.1705.07874

Manrique Millones, D. L., Van Leeuwen, K., & Ghesquière, P. (2011). Academic performance of Peruvian elementary school children: The case of schools in Lima at the 6th grade. *Interdisciplinaria*, *28*(2), 323-343.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods*, *52*(6), 2287-2305.

Miller, T. (2018). *Explanation in Artificial Intelligence: Insights from the Social Sciences* (arXiv:1706.07269). arXiv. http://arxiv.org/abs/1706.07269

Ministerio de Educación. (2016). *Instructivo para la aplicación de la evaluación estudiantil (actualizado a julio 2016)*. Ministerio de Educación. https://tinyurl.com/ycc6tdvz

Ministerio de Educación del Ecuador. (2013). *Instructivo para la aplicación de la evaluación estudiantil*. Ministerio de Educación del Ecuador.

Ministerio de Educación del Ecuador. (2016). *Proyectos escolares. Instructivo*. Ministerio de Educación del Ecuador.

Minitab. (2023). *Dendrograma*. Soporte de Minitab® 21. https://tinyurl.com/3kscrfru

Mohammad, A. H. (2018). Comparing two feature selections methods (information gain and gain ratio) on three different classification algorithms using Arabic Dataset. *Journal of Theoretical & Applied Information Technology*, *96*(6).

Mukhopadhyay, S. (2018). *Advanced Data Analytics Using Python*. Apress. https://doi.org/10.1007/978-1-4842-3450-1

Nelli, F. (2018). *Python Data Analytics: With Pandas, NumPy, and Matplotlib*. Apress. https://doi.org/10.1007/978-1-4842-3913-1

Niklas, F., & Schneider, W. (2017). Home learning environment and development of child competencies from kindergarten until the end of elementary school. *Contemporary Educational Psychology*, *49*, 263-274. https://doi.org/10.1016/j.cedpsych.2017.03.006

Nørskov, A. K., Lange, T., Nielsen, E. E., Gluud, C., Winkel, P., Beyersmann, J., de Uña-Álvarez, J., Torri, V., Billot, L., & Putter, H. (2021). Assessment of assumptions of statistical analysis methods in randomised clinical trials: The what and how. *BMJ evidence-based medicine*, *26*(3), 121-126.

Nyuytiymbiy, K. (2022, marzo 28). *Parameters and Hyperparameters in Machine Learning and Deep Learning*. Medium. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

OCDE. (2016). *Estudiantes de bajo rendimiento: Por qué se quedan atrás y cómo ayudarles a tener éxito*.

Olson, D. L., & Lauhoff, G. (2019). Descriptive Data Mining. En D. L. Olson & G. Lauhoff, *Descriptive Data Mining* (pp. 129-130). Springer Singapore. https://doi.org/10.1007/978-981-13-7181-3_8

Orange. (2015a). *Concatenate*. Orange Visual Programming. https://tinyurl.com/yc5yff86

Orange. (2015b). *Confusion matrix*. Orange Visual Programming. https://tinyurl.com/mr2f29cy

Orange. (2015c). *Correlations*. Orange Visual Programming. https://tinyurl.com/y7cvjz24

Orange. (2015d). *Data sampler*. Orange Visual Programming. https://tinyurl.com/5n8vxrjy

Orange. (2015e). *Distances*. Orange Visual Programming. https://tinyurl.com/mw4sv7y6

Orange. (2015f). *Edit domain*. Orange Visual Programming. https://tinyurl.com/2vj5e79z

Orange. (2015g). *Feature Importance*. Orange 3 - Explain. https://tinyurl.com/2jt3bz84

Orange. (2015h). *Frequent Itemsets*. Orange Visual Programming. https://tinyurl.com/mryprcs9

Orange. (2015i). *Gradient Boosting*. Orange Visual Programming. https://tinyurl.com/yn3at3sz

Orange. (2015j). *Hierarchical Clustering*. Orange Visual Programming. https://tinyurl.com/482rbd5m

Orange. (2015k). *K Means*. Orange Visual Programming. https://tinyurl.com/4v4bbhc8

Orange. (2015l). *KNN*. Orange Visual Programming. https://tinyurl.com/emj7yd7b

Orange. (2015m). *Linear projection*. Orange Visual Programming. https://tinyurl.com/3apw5m5f

Orange. (2015n). *Linear regression*. Orange Visual Programming. https://tinyurl.com/4yz4duvb

Orange. (2015o). *Logistic regression*. Orange Visual Programming. https://tinyurl.com/38eevv74

Orange. (2015p). *Naïve Bayes*. Orange Visual Programming. https://tinyurl.com/4fxrzze4

Orange. (2015q). *Neural Network*. Orange Visual Programming. https://tinyurl.com/nu8xm87t

Orange. (2015r). *Nomogram*. Orange Visual Programming. https://tinyurl.com/ye27pfyt

Orange. (2015s). *PCA*. Orange Visual Programming. https://tinyurl.com/y884za4m

Orange. (2015t). *Python script*. Orange Visual Programming. https://tinyurl.com/y5w8uypy

Orange. (2015u). *Random forest*. Orange Visual Programming. https://tinyurl.com/4566kvre

Orange. (2015v). *Randomize*. Orange Visual Programming. https://tinyurl.com/mtpjtp38

Orange. (2015w). *Scatter Plot*. Orange Visual Programming. https://tinyurl.com/ye762ynp

Orange. (2015x). *Select columns*. Orange Visual Programming. https://tinyurl.com/4sshywht

Orange. (2015y). *Select rows*. Orange Visual Programming. https://tinyurl.com/5n8vxrjy

Orange. (2015z). *Select rows*. Orange Visual Programming. https://tinyurl.com/5n8vxrjy

Orange. (2015aa). *Silhouette Plot*. Orange Visual Programming. https://tinyurl.com/3rkmjyy4

Orange. (2015ab). *Stochastic Gradient Descent*. Orange Visual Programming. https://tinyurl.com/4za4yckv

Orange. (2015ac). *SVM*. Orange Visual Programming. https://tinyurl.com/mpr64ja7

Orange. (2015ad). *Test and scores*. Orange Visual Programming. https://tinyurl.com/yrwsmu35

Orange. (2015ae). *Tree*. Orange Visual Programming. https://tinyurl.com/8xfratcm

Orange. (2016). *Association Rules*. Orange Visual Programming. https://tinyurl.com/39j884au

Orange. (2021). *Explain model*. Explaining Predictive Models. https://tinyurl.com/yhcdt2w8

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Peña-Ayala, A. (Ed.). (2014). *Educational Data Mining* (Vol. 524). Springer International Publishing. https://doi.org/10.1007/978-3-319-02738-8

Pincay Ponce, J., Sánchez-Andrade, D., Caicedo-Ávila, I., & Macías-Valencia, D. (2020, noviembre 27). *Clasificación de pacientes según su posibilidad de adquirir Diabetes Mellitus empleando algoritmos de Machine Learning*. IV Congreso Internacional Tecnologías de la Información y Computación (CITIC 2020), Calceta, Ecuador. https://tinyurl.com/yve333v7

Pincay-Ponce, J. I., Herrera-Tapia, J. S., Terranova-Ruiz, J., Cruz-Felipe, M., Sendón-Varela, J. C., & Fernández-Capestany, L. (2022). Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *E49*, 654-667.

Pincay-Ponce, J. I., Herrera-Tapia, J. S., Terranova-Ruiz, J., Cruz-Felipe, M., Sendón-Varela, J. C., & Fernández-Capestany, L. (2023). Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar. Revisión sistemática. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *E52*, 654-667.

Probst, P., Bischl, B., & Boulesteix, A.-L. (2018). *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*. https://doi.org/10.48550/ARXIV.1802.09596

Programme for International Student Assessment, PISA 2015. (2018). The Organisation for Economic Co-operation and Development. https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features* (arXiv:1706.09516). arXiv. http://arxiv.org/abs/1706.09516

Quirantes, M. I. R., Carrillo, M. I. S., & Gómez, I. G. (2016). Trastornos Biopsicosociales en Hijos de Padres Divorciados y Prevención desde el Ámbito Sanitario. *Perspectivas y Análisis de la Salud*, 91.

Quiroga, F. (2020). *Reglas de Asociación: Métricas*.

Quiroga, F. M. (2020). *Medidas de Invarianza y Equivarianza a Transformaciones en Redes Neuronales Convolucionales. Aplicaciones al reconocimiento de formas de mano*. Universidad Nacional de La Plata.

Ridwan, F., Subagio, B., & Rahman, H. (2018). Porous concrete basic property criteria as rigid pavement base layer in indonesia. *MATEC Web of Conferences*, *147*, 02008. https://doi.org/10.1051/matecconf/201814702008

Robitzsch, A. (2022). Comparing the robustness of the structural after measurement (SAM) approach to structural equation modeling (SEM) against local model misspecifications with alternative estimation approaches. *Stats*, *5*(3), 631-672.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Russo, C. C. (2019). Minería de datos aplicada a estrategias para minimizar la deserción universitaria en carreras de Informática de la UNNOBA. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, *24*, 94-95.

Sanders, M. T., Bierman, K. L., & Heinrichs, B. S. (2020). Longitudinal Associations Linking Elementary and Middle School Contexts with Student Aggression in Early Adolescence. *Journal of Abnormal Child Psychology*, *48*(12), 1569-1580. https://doi.org/10.1007/s10802-020-00697-6

Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical Machine Learning with Python*. Apress. https://doi.org/10.1007/978-1-4842-3207-1

Sartain, L., & Barrow, L. (2022). The Pathway to Enrolling in a High-Performance High School: Understanding Barriers to Access. *Education Finance and Policy*, *17*(3), 379-407. https://doi.org/10.1162/edfp_a_00349

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257-273. https://doi.org/10.1080/00131881.2019.1625716

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768.

Scikit Learn. (2022a). *KMeans*. Scikit Learn. https://tinyurl.com/342jzavb

Scikit Learn. (2022b). *Stochastic Gradient Descent*. Scikit Learn. https://tinyurl.com/yyhzdwhh

Shabtay, L., Fournier-Viger, P., Yaari, R., & Dattner, I. (2021). A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. *Information Sciences*, *553*, 353-375.

Shobha, G., & Rangaswamy, S. (2018). Machine Learning. En *Handbook of Statistics* (Vol. 38, pp. 197-228). Elsevier. https://doi.org/10.1016/bs.host.2018.07.004

Siemens, G., & Baker, R. S. d. (2012). *Learning analytics and educational data mining: Towards communication and collaboration*. 252-254.

Smirani, L. K., Yamani, H. A., Menzli, L. J., & Boulahia, J. A. (2022). Using ensemble learning algorithms to predict Student failure and enabling customized educational paths. *Scientific Programming*, *2022*.

Smith, C. (2017). *Decision trees and random forests: A visual introduction for beginners*. Blue Windmill Media.

Solano Luengo, L. O. (2015). *Rendimiento académico de los estudiantes de secundaria obligatoria y su relación con las aptitudes mentales y las actitudes ante el estudio*.

Soncin, M., & Cannistrà, M. (2022). Data analytics in education: Are schools on the long and winding road? *Qualitative Research in Accounting & Management*, *19*(3), 286-304. https://doi.org/10.1108/QRAM-04-2021-0058

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, *18*(10), 1099-1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392-413. https://doi.org/10.3390/make3020020

Sumathi, S. (2021). *Advanced decision sciences based on deep learning and ensemble learning algorithms: A practical approach using Python*. Nova Science Publishers.

Sun, G. (2022). Application of GA-BP Neural Network in Online Education Quality Evaluation in Colleges and Universities. *Mobile Information Systems*, *2022*.

Supo, J. (2017). *Portafolio de Aprendizaje Para la Docencia en Investigación Científica* (1.ª ed.).

Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science*, *184*, 835-840. https://doi.org/10.1016/j.procs.2021.03.104

Tejedor, F. J. T. (1998). *Los alumnos de la Universidad de Salamanca. Características y rendimiento académico* (Vol. 34). Universidad de Salamanca.

Tejedor, F. J. T. (2003). Poder explicativo de algunos determinantes del rendimiento en los estudios universitarios. *Revista española de pedagogía*, 5-32.

Torrecilla, F. J. M., & Bernal, E. C. (2007). *Investigación iberoamericana sobre eficacia escolar*. Convenio Andrés Bello.

Uedufy. (2023). *How to interpret model fit results in AMOS*. Uedufy. https://tinyurl.com/4szy8m6w

Umer Baloch, R. (2020). *Prediction of students' performance through data mining: A thesis presented in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science, Massey University, Auckland, New Zealand*.

UNESCO. (2020). *Aprendiendo en casa: Educación a distancia para todos*. https://www.unesco.org/es/articles/aprendiendo-en-casa-educacion-distancia-para-todos

UNESCO. (2021a). *Reforzar el aprendizaje y las capacidades digitales en los países más poblados del mundo para estimular la recuperación de la educación*. https://www.unesco.org/es/articles/reforzar-el-aprendizaje-y-las-capacidades-digitales-en-los-paises-mas-poblados-del-mundo-para

UNESCO. (2021b). *Resultados de logros de aprendizaje y factores asociados del Estudio Regional Comparativo y Explicativo (ERCE 2019)*. https://www.unesco.org/es/articles/resultados-de-logros-de-aprendizaje-y-factores-asociados-del-estudio-regional-comparativo-y##

UNESCO. (2022). *Leave no child behind. Global report on boys' disengagement from education*. https://unesdoc.unesco.org/ark:/48223/pf0000381106/PDF/381106eng.pdf.multi

UNESCO, G. (2017). Accountability in education: Meeting our commitments. *Global education monitoring report Available at: https://unesdoc. unesco. org/ark:/48223/pf0000259338 (accessed 20 October 2020)*.

Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the Tractability of SHAP Explanations. *Journal of Artificial Intelligence Research*, *74*, 851-886. https://doi.org/10.1613/jair.1.13283

van Rijn, J. N., & Hutter, F. (2018). Hyperparameter Importance Across Datasets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2367-2376. https://doi.org/10.1145/3219819.3220058

van Zwieten, A., Teixeira-Pinto, A., Lah, S., Nassar, N., Craig, J. C., & Wong, G. (2021). Socioeconomic Status During Childhood and Academic Achievement in Secondary School. *Academic Pediatrics*, *21*(5), 838-848. https://doi.org/10.1016/j.acap.2020.10.013

Vodencarevic, A., & Fett, T. (2015). Data analytics for manufacturing systems. *2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA)*, 1-4. https://doi.org/10.1109/ETFA.2015.7301541

Wang, C., Fan, X., & Pugalee, D. K. (2020). Impacts of School Racial Composition on the Mathematics and Reading Achievement Gap in Post Unitary Charlotte-Mecklenburg Schools. *Education and Urban Society*, *52*(7), 1112-1132. https://doi.org/10.1177/0013124519894970

Witten, I. H., & Witten, I. H. (Eds.). (2017). *Data mining: Practical machine learning tools and techniques* (Fourth Edition). Elsevier.

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yeturu, K. (2020). Machine learning algorithms, applications, and practices in data science. En *Handbook of Statistics* (Vol. 43, pp. 81-206). Elsevier. https://doi.org/10.1016/bs.host.2020.01.002

Yu, F., & Liu, X. (2022). Research on Student Performance Prediction Based on Stacking Fusion Model. *Electronics*, *11*(19), 3166.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, *12*(3), 372-390.

Zhang, A., Lipton, Z., Li, M., & Smola, A. (2022). *Dive into deep learning*. https://tinyurl.com/55jn6wzs

Zhou, Y., Lu, Z., & Cheng, K. (2022). Adaboost-based ensemble of polynomial chaos expansion with adaptive sampling. *Computer Methods in Applied Mechanics and Engineering*, *388*, 114238.

Anexo 1: Artículo "Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar"

risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 13/12/2021
Aceitação/Acceptance: 07/02/2022

Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar

Jorge Pincay-Ponce[1, 2], Jorge Herrera-Tapia[1], Jackeline Terranova-Ruiz[1], Marely Cruz-Felipe[2], Juan Sendón-Varela[1], Lytyet Fernández-Capestany[1]

jorge.pincay@uleam.edu.ec; jorge.herrera@uleam.edu.ec; jackeline.terranova@uleam.edu.ec; juan.sendon@uleam.edu.ec; marely.cruz@utm.edu.ec; lytyet.fernandez@uleam.edu.ec

[1] Universidad Laica Eloy Alfaro de Manabí, Manta, 130802, Manabí, Ecuador.

[2] Universidad Nacional de la Plata, 1900, La Plata, Argentina.

[2] Universidad Técnica de Manabí, Portoviejo, 130105, Manabí, Ecuador.

Pages: 654-667

Resumen: La minería de datos educativos es un compendio de métodos eficaces para detectar patrones de situaciones que afectan diversos aspectos de la escolaridad, sin embargo, no es del todo claro en qué momentos de un curso académico los factores socioeconómicos pueden tener más incidencia en el aprovechamiento. En este estudio se ha tomado una muestra transversal aleatoria de datos de calificaciones parciales progresivas y de factores socioeconómicos de alumnos de una escuela ecuatoriana, para estimar las influencias de estos factores en su aprovechamiento mediante algoritmos de clasificación. Con el empleo de Reglas de Asociación y Árboles de Decisión, se precisó en un 90% que los factores socioeconómicos influyen el aprovechamiento de los estudiantes de modo especial en el primero de los dos quimestres que componen un periodo académico, y, en el segundo el mayor dominio de aprendizajes reflejado en notas se relaciona más con las calificaciones de los componentes parciales. Este modelo de análisis puede ser aplicado a todos los niveles de educación.

Palabras-clave: Mining Student; CN2; Árboles de decisión; Clasificación; Naive Bayes.

Educational data mining: Incidence of socioeconomic factors on school achievement

Abstract: Educational data mining is a compendium of effective methods to detect patterns of situations that affect various aspects of schooling; however, it is not entirely clear at what times of an academic year socioeconomic factors may have more impact on achievement. In this study, a random cross-sectional sample of data from progressive partial grades and socioeconomic factors of students from an Ecuadorian school has been taken, to estimate the influences of these factors on their achievement through classification algorithms. With the use of Association Rules and Decision Trees, it was specified in 90% those socioeconomic factors influence

Anexo 2: Artículo "Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar. Revisión sistemática"

ICITS 23

February 8-10, Cusco, Peru

ACCEPTANCE LETTER

Dear Jorge Iván Pincay-Ponce
Universidad Laica Eloy Alfaro de Manabí
Ecuador

On behalf of the ICITS'23 - The 2023 International Conference on Information Technology & Systems, I am pleased to inform you that your submission "***Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar. Revisión sistemática.***" has been accepted as a Full Paper for publication and oral presentation in this conference.

So, you are cordially invited to participate and present the paper in the ICITS'23 (http://www.icits.me/) to be held in Cusco, Peru, between the 8th and the 10th of February of 2023, an international scientific event sponsored and organized by Universidad Nacional de San Antonio Abad del Cusco, IEEE SMC and ITMA.

We sincerely hope that you will join us in making ICITS'23 a success. We look forward to seeing you next February.

Sincerely,

Álvaro Manuel Reis da Rocha

ICITS'23, General Chair

Anexo 3: Artículo "Análisis de implementaciones de sistemas tutores inteligentes y afectivos. Revisión sistemática"

ANÁLISIS DE IMPLEMENTACIONES DE SISTEMAS TUTORES INTELIGENTES Y AFECTIVOS. REVISIÓN SISTEMÁTICA

TUTORES INTELIGENTES AFECTIVOS

AUTORES: Jorge Pincay Ponce[1]

Pablo Pintado[2]

Julio Biset[3]

DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 2019-06-28

Fecha de aceptación: 2019-07-26

RESUMEN

Uno de los aspectos más importantes para el desarrollo de sistemas de tutoría inteligentes, además de su conocida computarización para dar tutoría a los estudiantes como lo haría un experto en enseñanza individualizada, es considerar el factor emocional en los estudiantes. Aunque la literatura muestra algunos avances a nivel de los modelos teóricos, el número de implementaciones de tales procesos es escaso, por lo que el objetivo de este trabajo fue identificar investigaciones que hayan implicado la implementación de sistemas tutores inteligentes afectivos en ámbitos educativos, mediante la detección del compromiso emocional del alumno mientras permanece en un entorno virtual de aprendizaje. La revisión sistemática de la literatura permitió sintetizar los estudios disponibles y proporcionar un marco para la realización de nuevas investigaciones, en tal sentido es de destacar el progreso de las implementaciones hasta obtener sistemas tutores cada vez más automatizados en lo que a acciones tutoriales se refiere, empleando modernas técnicas de análisis de datos e incluso la revisión de factores fisiológicos.

PALABRAS CLAVE: tutores inteligentes afectivos; computación afectiva; sistemas empáticos.

[1] Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. Doctorando del Programa Doctorado en Informática de la Universidad Nacional del La Plata. Email: jorge.pincay@uleam.edu.ec, jorge.pincayp@info.unlp.edu.ar. Código ORCID: http://orcid.org/0000-0003-4711-8850.

[1] Docente en la Escuela de Ingeniería de Sistemas, Facultad de Administración de Empresas, Universidad del Azuay. Cuenca, Ecuador. Doctorando del Programa Doctorado en Informática de la Universidad Nacional del La Plata. Email: pablopintado@hotmail.com, pablo.pintadoz@info.unlp.edu.ar.

[1] Docente en la Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN). Doctorando del Programa Doctorado en Informática de la Universidad Nacional del La Plata. Tandil, Argentina. Email: juliotandil88@gmail.com, julio.biset@info.unlp.edu.ar. Código ORCID: http://orcid.org/0000-0002-5438-0411

Anexo 4: Artículo "La autoestima en los adolescentes que cursan el bachillerato. Realidad y expectativas"

Revista Electrónica Formación y Calidad Educativa (RefCalE). ISSN 1390-9010
AUTOESTIMA COMO UN SENTIMIENTO HACIA UNO MISMO

La autoestima en los adolescentes que cursan el bachillerato. Realidad y expectativas.

AUTOESTIMA COMO UN SENTIMIENTO HACIA UNO MISMO

AUTORES: Isabel Marina Anchundia Delgado [1]
Jorge Iván Pincay Ponce [2]
Wilian Richart Delgado Muentes [3]

DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 03/11/2022
Fecha de aceptación: 27/12/2022

RESUMEN

El presente trabajo tiene como objetivo ofrecer un análisis de la autoestima en una muestra transversal de alumnos del bachillerato de una Unidad Educativa, de la Ciudad de Manta – Ecuador, a partir de la aplicación por medios digitales, de un cuestionario estandarizado con base en la Escala de Autoestima de Rosenberg de 1965, misma que aporta una fiabilidad del 92%. Además, relaciona la realidad de dichos resultados con las expectativas que se tienen de alumnos en ese momento de su educación. Los resultados obtenidos muestran patrones emocionales estables en los alumnos y de cierto modo positivos, lo que favorece el camino para que las instituciones educativas preparen a los estudiantes para la vida laboral y en sociedad, así como para continuar con sus estudios universitarios.

PALABRAS CLAVES/PALAVRAS-CHAVE: autoestima; educación; Rosenberg; sociedad

1 Docente en la Unidad Educativa Galileo Galilei - Ecuador. Licenciada en Administración Educativa, Máster en Gerencia Educativa por la Universidad Estatal del Sur de Manabí, Doctora en Educación por la Universidad Católica Andrés Bello. ORCID: 0000-0001-6729-8671. Manta, Manabí, Ecuador. Email: marina.anchundia@educacion.gob.ec. Móvil: +593988538880.
2 Docente en la Carrera de Tecnologías de la Información de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero en Sistemas. Máster Universitario en Ingeniería de Software por la Universidad de Alcalá. Doctor en Informática por la Universidad Nacional de La Plata. ORCID: 0000-0003-4711-8850. Manta, Manabí, Ecuador. Email: jorge.pincay@uleam.edu.ec, jorge.pincayp@info.unlp.edu.ar. Móvil: +593992621369.
3 Docente en la Carrera de Tecnologías de la Información de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero en Sistemas Computacionales. Máster en Informática de Gestión y Nuevas Tecnologías por Universidad Santa María. ORCID: 0000-0002-5136-0677. Manta, Manabí, Ecuador. Email: wilian.delgado@uleam.edu.ec. Móvil: +593980778865.

Anexo 6: Artículo "Accesibilidad web: Retos de las Universidades Ecuatorianas"

V CONGRESO CIENTÍFICO INTERNACIONAL

Accesibilidad web: Retos de las Universidades Ecuatorianas

Ing. Jorge Iván Pincay Ponce, MSc.

Profesor Ocasional de la Facultad de Ciencias Informáticas de la Universidad Laica ELOY ALFARO de Manabí. Doctorando del Doctorado en Ciencias Informáticas, de la Universidad Nacional de la Plata de Argentina. Código postal: 130802, Manta, Ecuador. Correo electrónico: jorge.pincay@uleam.edu.ec

Lic. Kléver Alfredo Delgado Reyes, Mg.

Profesor Titular Principal de la Facultad de Ciencias Informáticas de la Universidad Laica ELOY ALFARO de Manabí. Magíster en Educación, Mención en Inclusión Educativa y Atención a la Diversidad, de la Universidad Laica VICENTE ROCAFUERTE de Guayaquil. Código postal: 130802, Manta, Ecuador. Correo electrónico: klever.delgado@uleam.edu.ec

Resumen

El propósito de este estudio fue determinar la situación actual de la accesibilidad de los sitios webs de las universidades ecuatorianas y los retos de estas, ante la publicación de recientes pautas para la accesibilidad del contenido en la web por parte del Consorcio Mundial de Internet (W3C), el análisis efectuado abordó todas las universidades; como un referente importante para la percepción de la accesibilidad web, considerada factor importante de inclusión en la educación superior. El procedimiento básico seguido en la presente investigación fue el siguiente: (1) Reconocer la situación actual de la accesibilidad de los sitios webs universitarios, con base en información bibliográfica publicada en los últimos dos años. (2) Elaborar una lista de chequeo incluyendo a los nuevos criterios de conformidad para la accesibilidad del contenido en la Web. (3) Evaluar el cumplimiento de cada criterio de conformidad en cada universidad, y (4) Analizar y presentar los resultados. Si bien los resultados evidencian la incorporación de mejoras en el diseño de los sitios webs universitarios, aún reflejan la necesidad de un mayor esfuerzo en su diseño, para hacer de estas herramientas oficiales de comunicación de las instituciones de educación superior, medios que contribuyan a disminuir barreras que afectan en general a todas las personas en situaciones de limitación del contexto o de los dispositivos desde donde se esté navegando, independientemente de sus capacidades.

Palabras clave: Accesibilidad Web; WCAG 2.0; WCAG 2.1; Instituciones de Educación Superior; NTE INEN ISO/IEC 40500; WAI.

Abstract

The purpose of this study was to determine the current situation of the accessibility of the

Anexo 7: Artículo "Clasificación de pacientes según su posibilidad de adquirir diabetes mellitus empleando algoritmos de machine learning"

CLASIFICACIÓN DE PACIENTES SEGÚN SU POSIBILIDAD DE ADQUIRIR DIABETES MELLITUS EMPLEANDO ALGORITMOS DE MACHINE LEARNING

Jorge Iván Pincay-Ponce
Universidad Laica Eloy Alfaro de Manabí, EC130850, Manta, Ecuador.
Universidad Nacional de la Plata, 1900, La Plata, Argentina.
Email: jorge.pincay@uleam.edu.ec

Diana Alexandra Sánchez-Andrade
Universidad Internacional de La Rioja (UNIR), La Rioja, España.
Universidad de Guayaquil, EC090101 - EC090158, Guayaquil, Ecuador.
Email: diana.sancheza@ug.edu.ec

Ingrid Vanessa Caicedo-Ávila
Universidad Laica Eloy Alfaro de Manabí, EC130850, Manta, Ecuador.
Email: vanec027a@gmail.com

David Gabriel Macías-Valencia
Universidad Laica Eloy Alfaro de Manabí, EC130850, Manta, Ecuador.
Email: david.macias@uleam.edu.ec

Anexo 8: Artículo "Legibilidad y Accesibilidad en los Sitios Web de las Universidades de la Provincia de Manabí-Ecuador"

Revista Electrónica Formación y Calidad Educativa (REFCalE) ISSN 1390-9010
ACCESIBILIDAD Y LEGIBILIDAD EN LOS SITIOS WEB

LEGIBILIDAD Y ACCESIBILIDAD EN LOS SITIOS WEB DE LAS UNIVERSIDADES DE LA PROVINCIA DE MANABÍ-ECUADOR

ACCESIBILIDAD Y LEGIBILIDAD EN LOS SITIOS WEB

AUTORES: Jorge Iván Pincay Ponce [1]
José Jacinto Reyes Cárdenas [2]
Pedro Emilio Delgado Franco [3]
Oscar Armando González López [4]
DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 2020-03-05
Fecha de aceptación: 2020-04-10

RESUMEN.

La accesibilidad web en el Ecuador ha tomado gran importancia, tanto así que el país dispone de la Norma Técnica Ecuatoriana NTE INEN ISO/IEC 40500 cuyo objetivo es garantizar que los sitios web ecuatorianos que presten un servicio público sean accesibles de acuerdo con la Norma Internacional ISO/IEC 40500:2012 Information Technology - W3C Web Content Accessibility Guidelines (WCAG) 2.0. Por lo tanto, para las universidades nacionales el cumplimiento de la norma INEN es de carácter obligatorio. El objetivo de este estudio es evaluar la accesibilidad web de cinco páginas representativas de cada sitio de las cinco universidades que tienen matriz sede en la provincia de Manabí, e identificar mediante herramientas de valoración automática los errores más comunes en estos medios de comunicación. Como resultado se presentan recomendaciones con respecto a los errores que se han identificado en el diseño de los sitios web y que generan barreras para un gran número de usuarios

[1] Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. Doctorando del Programa Doctorado en Informática de la Universidad Nacional de La Plata. E-mail: jorge.pincay@uleam.edu.ec, jorge.pincayp@info.unlp.edu.ar. Móvil: +593992621369
[2] Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí, Calle 12 y vía San Mateo, Km 1.5. Manta, Manabí, Ecuador. Código Postal: EC130802. Email: jose.reyes@uleam.edu.ec. Móvil: +593996018350
[3] Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí, Calle 12 y vía San Mateo, Km 1.5. Manta, Manabí, Ecuador. Código Postal: EC130802. Email: pedro.delgado@uleam.edu.ec. Móvil: +593969654508
[4]Docente ocasional en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí, Calle 12 y vía San Mateo, Km 1.5. Manta, Manabí, Ecuador. Código Postal: EC130802. Email: oscar.gonzalez@uleam.edu.ec. Móvil: +593985723603

Anexo 9: Artículo "Usabilidad en sitios web oficiales de las universidades del Ecuador"

risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 26/12/2019
Aceitação/Acceptance: 04/02/2020

Usabilidad en sitios web oficiales de las universidades del Ecuador

Jorge Pincay Ponce[1, 2], Vanessa Caicedo Ávila[1], Jorge Herrera-Tapia[1], Wilian Delgado Muentes[1], Pedro Delgado Franco[1].

jorge.pincay@uleam.edu.ec, vanec027a@gmail.com, jorge.herrera@uleam.edu.ec, wilian.delgado@uleam.edu.ec, pedro.delgado@uleam.edu.ec

[1] Universidad Laica Eloy Alfaro de Manabi, EC130850, Manta, Ecuador.

[2] Universidad Nacional de la Plata, 1900, La Plata, Argentina.

Pages: 106–119

Resumen: Los sitios web son un componente clave en las organizaciones del competitivo y globalizado mundo actual, paralelamente su usabilidad se reviste de importancia en términos de satisfacer las necesidades y expectativas de sus usuarios. En este escenario se encuentran las universidades, de cuyos sitios web se espera que cubran diversas necesidades de información para los estudiantes, profesores, personal y exalumnos, que se forman o formaron en ellas. En esta investigación se presenta una evaluación heurística multicriterio basado en la norma ISO 9241-151, donde se valora la ergonomía de la interacción hombre-sistema, de las interfaces de usuario web en función de los criterios de diseño, presentación, búsqueda, diseño de contenido y navegación; que se aplicaron en la construcción de estos sitios de las 59 universidades de Ecuador. Finalmente, se discuten las recomendaciones para el diseño de estos sitios y se reflexiona sobre los problemas detectados en las pruebas de usabilidad.

Palabras-clave: ISO 9241-151, HCI, evaluación heurística, sitios web.

Usability in official websites of the universities of Ecuador.

Abstract: Websites are a key component in the organizations of today's competitive and globalized world, at the same time their usability is important in terms of meeting the needs and expectations of their users. In this scenario are the universities, whose websites are expected to cover various information needs for students, teachers, staff and alumni, who are trained or trained in them. This research presents a multicriteria heuristic evaluation based on the ISO 9241-151 standard, where the ergonomics of the man-system interaction, of the web user interfaces according to the criteria of design, presentation, search, design of content and navigation; that were applied in the construction of these sites of the 59 universities in Ecuador. Finally, the recommendations for the design of these sites are discussed and the problems detected in the usability tests are discussed.

Keywords: ISO 9241-151, HCI, heuristic evaluation, websites.

Anexo 10: Artículo "Algunas experiencias de investigación basada en ciencia ciudadana para el beneficio de África"

Revista Electrónica Formación y Calidad Educativa (REFCalE) ISSN 1390-9010
LA CIENCIA CIUDADANA EN ÁFRICA. PERSPECTIVAS.

ALGUNAS EXPERIENCIAS DE INVESTIGACIÓN BASADA EN CIENCIA CIUDADANA PARA EL BENEFICIO DE ÁFRICA

LA CIENCIA CIUDADANA EN ÁFRICA. PERSPECTIVAS.

AUTORES: Jorge Iván Pincay Ponce[1]
Wilian Richart Delgado Muentes[2]
Leo Antonio Cedeño Cabezas[3]
Pedro Emilio Delgado Franco[4]

DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 10-08-2020

Fecha de aceptación: 26-11-2020

RESUMEN.

África es un continente culturalmente caracterizado como el de mayor pobreza, enfermedades o guerras civiles en el mundo; cuando entre otros aspectos positivos se trata de un continente con una creciente capacidad científica. El objetivo del presente es ofrecer una perspectiva del potencial científico de este continente, especialmente a nivel de proyectos que contemplan ciencia ciudadana, y cuyos resultados se reflejen en artículos publicados recientemente. La metodología aplicada ha sido la compilación y análisis de diez artículos revisados por pares, nueve de ellos se publicaron entre los años 2016 y 2018 y uno se publicó en 2014. Como resultado se presentan los detalles de cómo se recolectaron y analizaron los datos, los beneficiarios, la cantidad de participantes y su nivel de actividad en los proyectos. Como resultados se evidencia que la ciencia ciudadana se ha desarrollado a nivel del alcance y la eficiencia de la recopilación de datos para estudios en ecología, conservación de la biodiversidad, labores de rescate ante desastres o preservación del conocimiento indígena, entre otros. Como conclusiones se

[1] Docente titular en la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero en Sistemas. Máster Universitario en Ingeniería de Software, por la Universidad de Alcalá. Doctorando del Programa Doctorado en Informática de la Universidad Nacional de La Plata. ORCID: 0000-0003-4711-8850. Manta, Manabí, Ecuador. Email: jorge.pincay@uleam.edu.ec, jorge.pincayp@info.unlp.edu.ar. Móvil: +593992621369.
[2] Docente titular en la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero en Sistemas Computacionales. Máster en Informática de Gestión y Nuevas Tecnologías, por la Universidad Santa María de Chile. ORCID: 0000-0002-5136-0677. Manta, Manabí, Ecuador. Email: wilian.delgado@uleam.edu.ec. Móvil: +593980778865
[3] Docente titular en la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero Civil. Máster en Educación Matemática Universitaria, por la Universidad de Holguín de Cuba. Manta, Manabí, Ecuador. Email: leo.cedeno@uleam.edu.ec. Móvil: +593984282932
[4] (+) Docente titular en la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero Eléctrico. Máster en Administración de Empresas Mención en Gestión de Recursos Humanos, por la ULEAM. Manta, Manabí, Ecuador.

Anexo 11: Artículo "La usabilidad y la escala diferencial de emociones en aplicaciones para Android. Un estudio de caso"

La usabilidad y la escala diferencial de emociones en aplicaciones para Android. Un estudio de caso

AUTORES: Jorge Iván Pincay Ponce[1]
Jorge Sergio Herrera Tapia[2]
Wilian Richart Delgado Muentes[3]

DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 23-02-2021

Fecha de aceptación: 12-04-2021

RESUMEN

El objetivo principal de esta investigación fue evaluar los resultados de las escalas de emociones reportadas por las aplicaciones Emotion Detector y Image & Emotion Recognizer, seleccionadas de entre varias existentes para dispositivos Android, así como su usabilidad mediante la reconocida Escala de Usabilidad de un Sistema (SUS). La emoción es una parte integral de la existencia humana, que desempeña un papel importante en la vida cotidiana, por lo que el campo ha sido investigado significativamente en los últimos años, surgiendo diversas escalas discretas en variados contextos. Los resultados obtenidos en esta comparativa muestran cercanías en las escalas diferenciales de emociones reportadas, así como de la usabilidad que tienen las aplicaciones seleccionadas.

PALABRAS CLAVE: computación afectiva; SUS; Escala de usabilidad del sistema; escala diferencial de emociones.

Usability and the differential scale of emotions in Android applications

ABSTRACT

The main objective of this research was to evaluate the results of the emotional scales reported by the Emotion Detector and Image & Emotion Recognizer applications, selected among several existing for Android devices, as well as their usability through the recognized System Usability Scale (SUS). Emotion is an integral part of human existence, which plays an important role in everyday life, so that the field has been researched significantly in recent years, arising discrete scales in various contexts. The results obtained in this comparison show closeness in the differential scales of reported emotions, as well as the usability of the selected applications.

KEYWORDS: Affective computing; SUS; System usability scale; differential scale of emotions.

[1] Ingeniero en Sistemas. Alumno del programa de Doctorado en Informática de la Universidad Nacional de la Plata. La Plata, Buenos Aires, Argentina. Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. E-mail: jorge.pincayp@info.unlp.edu.ar.
[2] Ingeniero en Sistemas. Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. E-mail: jorge.herrera@uleam.edu.ec
[3] Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. E-mail: wilian.delgado@uleam.edu.ec

Anexo 12: Artículo "La usabilidad de los sitios web oficiales de destinos turísticos de países miembros de la OMT"

Revista Electrónica Formación y Calidad Educativa (REFCalE) ISSN 1390-9010
UNA VALORACIÓN HEURÍSTICA DE LA USABILIDAD

LA USABILIDAD DE LOS SITIOS WEB OFICIALES DE DESTINOS TURÍSTICOS DE PAÍSES MIEMBROS DE LA OMT.

UNA VALORACIÓN HEURÍSTICA DE LA USABILIDAD

AUTORES: Karla Guadalupe Palma Laaz [1]
Jorge Iván Pincay Ponce [2]
David Gabriel Macías Valencia [3]
Jorge Sergio Herrera Tapia [4]
DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 13-03-2022
Fecha de aceptación: 10-06-2022

RESUMEN

El turismo es una actividad económica importante en muchos países, en los cuales tanto entidades oficiales como no oficiales, compiten por atraer a los turistas, mediante canales de comunicación y promoción, entre los que se incluyen a los sitios web, sin embargo, no existe un método integral para determinar si estos medios se diseñan cumpliendo con criterios de usabilidad web que generen una percepción positiva por parte de sus usuarios, así como una interacción efectiva y eficiente en ellos. El objeto de estudio de esta investigación fueron los sitios web oficiales de información turística de cada país afiliado a la Organización Mundial del Turismo (OMT). El objetivo de la investigación fue presentar una evaluación heurística multicriterio de la ergonomía de la interacción hombre-sistema basada en la norma ISO 9241-151, la Norma permite evaluar los sitios web con base en los criterios de navegación, diseño general, diseño de contenidos, búsqueda, presentación y navegación; que se aplicaron en su construcción. Los resultados se obtuvieron según cada criterio con su grupo de indicadores, tanto a nivel de

1 Ingeniera en Sistemas por la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. Email: karlapalmam13@gmail.com. Móvil: +593988264512.
2 Docente titular en la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero en Sistemas. Máster Universitario en Ingeniería de Software, por la Universidad de Alcalá. Doctorando del Programa Doctorado en Informática de la Universidad Nacional de La Plata. ORCID: 0000-0003-4711-8850. Manta, Manabí, Ecuador. Email: jorge.pincay@uleam.edu.ec, jorge.pincayp@info.unlp.edu.ar. Móvil: +593992621369.
3 Docente titular en la Facultad de Contabilidad y Auditoría de la Universidad Laica Eloy Alfaro de Manabí. Tecnólogo en Computación Administrativa en Sistemas. Especialista en Diseño Curricular por Competencias por la Universidad del Mar, Chile. ORCID: 0000-0003-1945-753X. Manta, Manabí, Ecuador. Email: david.macias@uleam.edu.ec. Móvil: +593985803987.
4 Docente titular en la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Ingeniero en Sistemas. Doctor en Informática por la Universidad Politécnica de Valencia, España. ORCID: 0000-0002-8673-0236. Manta, Manabí, Ecuador. Email: jorge.herrera@uleam.edu.ec. Móvil: +59399951006.

Anexo 13: Artículo "Modelo Matemático de predicción de Graduados"

ICITS 23
February 8-10,
Cusco, Peru

ACCEPTANCE LETTER

Dear PINCAY PONCE JORGE IVÁN
Facultad de Ciencias Informáticas
Ecuador

On behalf of the ICITS'23 - The 2023 International Conference on Information Technology & Systems, I am pleased to inform you that your submission "Modelo Matemático de predicción de Graduados", en la International Conference on Information Technology & Systems (ICITS)-2023" has been accepted as a Full paper for publication in RISTI and oral presentation in this conference.

So, you are cordially invited to participate and present the paper in the ICITS'23 (http://www.icits.me/) to be held in Cusco, Peru, between the 8th and the 10th of February of 2023, an international scientific event sponsored and organized by Universidad Nacional de San Antonio Abad del Cusco, IEEE SMC and ITMA.

We sincerely hope that you will join us in making ICITS'23 a success. We look forward to seeing you next February.

Sincerely,

Alvaro Manuel Reis da Rocha

ICITS'23, General Chair

Anexo 14: Artículo "Innovación en la enseñanza - aprendizaje en universidades sudamericanas mediante gestión del conocimiento"

ICITS 23

February 8-10, Cusco, Peru

ACCEPTANCE LETTER

Dear Jorge Iván Pincay-Ponce
Universidad Laica Eloy Alfaro de Manabí
Ecuador

On behalf of the ICITS'23 - The 2023 International Conference on Information Technology & Systems, I am pleased to inform you that your submission "***Innovación en la enseñanza - aprendizaje en universidades sudamericanas mediante gestión del conocimiento***" has been accepted as a Full Paper for publication and oral presentation in this conference.

So, you are cordially invited to participate and present the paper in the ICITS'23 (http://www.icits.me/) to be held in Cusco, Peru, between the 8th and the 10th of February of 2023, an international scientific event sponsored and organized by Universidad Nacional de San Antonio Abad del Cusco, IEEE SMC and ITMA.

We sincerely hope that you will join us in making ICITS'23 a success. We look forward to seeing you next February.

Sincerely,

Álvaro Manuel Reis da Rocha

ICITS'23, General Chair

Anexo 15: Ejemplo de acuerdo de confidencialidad de los datos de una de las escuelas participantes

UNIDAD EDUCATIVA PARTICULAR "JUAN LEÓN MERA"
Código AMIE 13H04278
Jaramijó - Manabí - Ecuador

Acuerdo de confidencialidad y no divulgación de la información sensible para la Unidad Educativa "Juan León Mera"

Intervienen en el presente **Acuerdo de Confidencialidad**, por una parte, Ing. Juan Alberto Figueroa Suárez, con cédula de ciudadanía Nro. 130516265-1, Rector de la Unidad Educativa Juan León Mera en calidad de **Proveedor de Información** Académica y Socioeconómica de los alumnos del ciclo escolar correspondiente al periodo lectivo 2019 - 2020; y Jorge Iván Pincay Ponce, con cédula de ciudadanía Nro. 131091554-9 en calidad de **Receptor de la Información.**

En este acuerdo, ambas partes reconocen recíprocamente su capacidad para obligarse al trato anonimizado de la información, con la finalidad de su uso exclusivo en la investigación de grado doctoral "Análisis de datos educativos aplicado en el estudio de la incidencia de factores socioeconómicos en el rendimiento escolar", desarrollada por el docente investigador Sr. Pincay Ponce.

Por lo que suscriben el presente Acuerdo de Confidencialidad de Información.

Jaramijó, Ecuador.

Enero de 2023

Ing. Juan Alberto Figueroa Suárez
CI. 130516265-1
Rector de Unidad Educativa "Juan León Mera"

Jorge Iván Pincay Ponce
CI. 131091554-9
Docente Investigador

UNIDAD EDUCATIVA PARTICULAR "JUAN LEÓN MERA" Resolución No. 018 Jaramijó - Manabí - Ecuador

DIRECCIÓN: Calle Haraway entre Aníbal San Andrés e Isaac Mendoza **E-MAIL:** uejuanleonmerajaramijo@hotmail.com **TELÉFONO:** 0988424074