Pontificia Universidad Católica del Ecuador

Facultad de Ingeniería

Maestría en Sistemas de Información mención Ciencia de Datos

Proyecto de titulación

Rendimiento académico de los estudiantes de educación secundaria en el Ecuador período 2021-2022

Autora: Fanny Narcisa Cabrera Barbecho

Director: Ing. Rafael Melgarejo, PhD.

Cuenca, mayo 2022

Tabla de contenido

Índice de Tablas

Índice de Gráficos

Índice de Anexos

Dedicatoria

A mis hijos Abel y Alexander fuente de mi disciplina y trabajo, motores de mi vida.

Agradecimientos

A Dios por sus infinitas bendiciones, a mis padres Carlos y Narcisa, mis hermanos Carlos y Liliana por su apoyo incondicional en esta etapa. A mi tutor Rafael por su importante soporte y dirección en la elaboración de este proyecto.

Resumen

El acceso a la educación superior en Ecuador enfrenta obstáculos significativos debido a diversos factores, como la calidad de la educación secundaria. El rendimiento académico de los estudiantes de bachillerato desempeña un papel crucial en la probabilidad de acceder a una institución de educación superior. Para ser admitidos, los estudiantes deben cumplir ciertos requisitos mínimos, incluyendo el puntaje del examen Ser Estudiante y el promedio académico logrado en la secundaria. Este estudio examina el rendimiento académico de los estudiantes de bachillerato en Ecuador durante el período 2021-2022, utilizando la metodología CRISP DM y técnicas de aprendizaje automático basadas en los datos del Instituto Nacional de Evaluación Educativa. La investigación emplea una variedad de métodos supervisados para predecir el rendimiento académico, incluyendo Regresión lineal múltiple, K vecinos cercanos, Árbol de decisión, Random forest, Elastic NET y Gradient Boosting. Además, se aborda la clasificación del nivel de rendimiento como elemental o insuficiente mediante el uso de modelos como Logit, K vecinos más cercanos, Árbol de decisión, Random forest, Naive bayes, Gradient boosting y Multilayer perceptron. En última instancia, se implementan enfoques no supervisados, tales como K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) y Agglomerative clustering, con el objetivo de identificar grupos distintivos de estudiantes. Los resultados revelan que el rendimiento académico se ve afectado por factores como el nivel socioeconómico, la región y el área de ubicación de la institución educativa. Asimismo, aspectos institucionales, como el tipo de institución, y características educativas del estudiante, como la cantidad de horas y días dedicados a las tareas y la pérdida de años escolares, influyen en el rendimiento del bachiller. Se determinó que la satisfacción del estudiante con su entorno familiar también incide en su desempeño académico. A través del análisis, se identificaron tres grupos de estudiantes para los cuales se pueden proponer estrategias de mejora específicas. Estos hallazgos brindan información valiosa para apoyar la toma de decisiones informadas en relación con la educación secundaria en Ecuador.

Palabras clave: rendimiento académico, aprendizaje automático, predicción, clasificación, cluster

Abstract

Access to higher education in Ecuador faces significant obstacles due to various factors, such as the quality of secondary education. The academic performance of high school students plays a crucial role in the likelihood of accessing a higher education institution. To be admitted, students must meet certain minimum requirements, including the Ser Estudiante exam score and the academic average achieved in high school. This study examines the academic performance of high school students in Ecuador during the 2021-2022 period, using the CRISP DM methodology and machine learning techniques based on data from the National Institute of Educational Evaluation. The research employs a variety of supervised methods to predict academic performance, including Multiple Linear Regression, K nearest neighbors, Decision Tree, Random Forest, Elastic NET, and Gradient Boosting. In addition, the classification of the performance level as elementary or insufficient is addressed using models such as Logit, K nearest neighbors, Decision Tree, Random Forest, Naive Bayes, Gradient Boosting, and Multilayer Perceptron. Ultimately, unsupervised approaches, such as K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Agglomerative Clustering, are implemented to identify distinctive groups of students. The results reveal that academic performance is affected by factors such as socioeconomic level, region, and area of location of the educational institution. Likewise, institutional aspects, such as the type of institution, and student educational characteristics, such as the number of hours and days devoted to tasks and the loss of school years, influence the performance of the high school student. It was determined that the student's satisfaction with their family environment also affects their academic performance. Through the analysis, three groups of students were identified for which specific improvement strategies can be proposed. These findings provide valuable information to support informed decision-making regarding secondary education in Ecuador.

Keywords: academic performance, machine learning, prediction, classification, cluster

1. Introducción

La educación es un factor clave que influye en la capacidad de las personas para encontrar empleo y determina sus salarios (Lucero, 2019). Según la teoría del capital humano, la educación contribuye al desarrollo de habilidades, lo que a su vez mejora la productividad y la competitividad de los individuos (Quintero Montaño, 2020). Como resultado, aquellos que invierten en educación suelen obtener mayores beneficios económicos en comparación con

quienes no lo hacen (Martínez, 2019). Esta teoría sostiene que cuanto mayor es el nivel educativo alcanzado, mayor es el retorno económico que las personas pueden obtener. En este contexto, estudios como el de Ñiquen (2019) demuestran que quienes cuentan con educación superior tienen una mayor probabilidad de encontrar empleos de calidad y disfrutan de un mayor retorno en términos de salario en comparación con aquellos que poseen un nivel educativo inferior. Los hallazgos de Austria-Carlos et al. (2018) respaldan esta idea, ya que muestran que aquellos con educación universitaria tienden a tener salarios más altos que quienes no cuentan con dicho nivel de instrucción.

En el Ecuador, nueve de cada diez estudiantes aspiran a ingresar a una institución de educación superior, sin embargo, solo cinco de cada diez postulantes logran el acceso (SENESCYT, 2021). Diversos factores influyen en la capacidad de acceder a la educación superior, incluyendo aspectos motivacionales, socioeconómicos y el rendimiento académico durante la educación secundaria (Han et al., 2022; Silva et al., 2020). Considerando este último factor, El bachillerato, como etapa previa a la educación superior, desempeña un papel crucial en este proceso. Un buen desempeño en la educación secundaria no solo puede facilitar el ingreso a la universidad, sino también reducir la probabilidad de abandono durante los primeros ciclos de la carrera universitaria (Chigbu & Nekhwevha, 2021; Vieira et al., 2018).

Entre 2013 y 2020, en Ecuador se llevó a cabo la prueba Ser Bachiller con el objetivo de evaluar el rendimiento de los estudiantes de bachillerato y emplearla como parte de la calificación para el ingreso a la universidad (INEVAL, 2020). Durante el año lectivo 2019-2020, el desempeño promedio de los bachilleres ecuatorianos en el examen preuniversitario "Ser Bachiller" fue elemental, alcanzando una calificación media de 7.62/10 (INEVAL, 2022). Este resultado sugiere que los estudiantes poseen únicamente conocimientos básicos, lo cual dificulta su acceso a instituciones de educación superior (SENESCYT, 2021).

Aunque la prueba "Ser Bachiller" fue derogada en 2020, el Instituto Nacional de Evaluación Educativa (INEVAL) introdujo el instrumento de evaluación "Ser Estudiante" para diagnosticar el estado de la educación en estudiantes de nivel de bachillerato (INEVAL, 2021). En el período 2021-2022, los estudiantes evaluados obtuvieron un promedio global de 683/1000 puntos en el

instrumento "Ser Estudiante" (INEVAL, 2022). Dicha nota es considerada como un rendimiento insuficiente, ya que está por debajo de los 700 puntos.

En este contexto, es crucial evaluar y analizar el desempeño de los estudiantes de bachillerato, ya que constituye uno de los factores clave para el éxito en el acceso a instituciones de educación superior (IES) (Villarruel-Meythaler et al., 2020). Esto permitirá identificar las características que influyen en su rendimiento y, en última instancia, en su futuro académico y laboral. Un bajo rendimiento académico disminuye las probabilidades de ingresar a la universidad, lo que a su vez se relaciona con menores niveles de empleabilidad, trabajos con salarios bajos o condiciones laborales inadecuadas (Oreopoulos & Salvanes, 2011). A largo plazo, estas circunstancias dificultan la mejora de las condiciones de vida de los individuos (ver Gráfico 1).

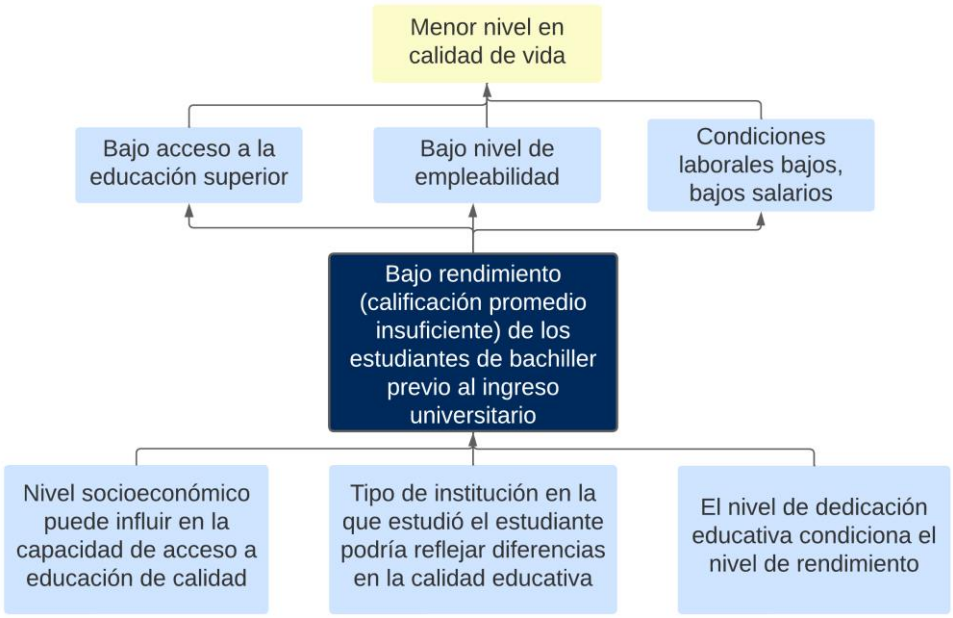


Gráfico 1. Problematización
Elaboración: propia basado en Lucero (2019); Ñiquen (2019)

El estudio del rendimiento académico previo al acceso a la educación superior en Ecuador se ha centrado principalmente en investigaciones de carácter descriptivo. Comprender y diagnosticar el estado actual, así como identificar los factores que influyen en el rendimiento preuniversitario, resulta esencial para proporcionar información necesaria en la formulación de programas y proyectos destinados a abordar los problemas de acceso a la educación superior (Villarruel-

Meythaler et al., 2020). Es importante destacar que el acceso a la educación superior no es el único desafío; también existe el problema de la sostenibilidad en los estudios universitarios durante los primeros ciclos de la carrera profesional (Tinto, 2012). Por lo tanto, reforzar el proceso educativo en el nivel de bachillerato se convierte en un punto crítico o cuello de botella que requiere atención y soluciones efectivas (SENESCYT, 2022).

En este contexto, la investigación tiene como objetivo predecir y clasificar el rendimiento de los estudiantes de bachillerato utilizando técnicas de aprendizaje automático. Para ello, se realiza un diagnóstico descriptivo del nivel de rendimiento académico en función de las características socioeconómicas y educativas. En segundo lugar, se predice y clasifica el rendimiento a través de múltiples técnicas de aprendizaje supervisado. En tercer lugar, se identifican los factores que influyen en el desempeño académico. Por último, se establecen grupos o clústeres de estudiantes por medio de técnicas de aprendizaje no supervisado, de forma que permita identificar patrones y características comunes dentro de cada grupo, facilitando la comprensión de las similitudes y diferencias en el rendimiento académico, lo que puede proporcionar información valiosa para diseñar estrategias educativas más eficaces y adaptadas a las necesidades específicas de cada grupo de estudiantes. De esta forma, las preguntas que esta investigación busca responder son:

1. ¿Cuáles son las características socioeconómicas, geográficas, educativas y de percepción que influyen significativamente en el desempeño académico de los estudiantes de bachillerato?
2. ¿Cómo pueden las técnicas de aprendizaje automático no supervisado, como el análisis de clústeres, revelar grupos de estudiantes con características y necesidades similares?

Este estudio se centra en Ecuador durante el período 2021-2022, enfocándose específicamente en los estudiantes que cursaron el bachillerato general unificado en el país durante ese tiempo. La información utilizada proviene del Instituto Nacional de Evaluación Educativa (INEVAL), a través de la base de datos Ser Estudiante, que recopila datos sobre el rendimiento académico, aspectos socioeconómicos, factores educativos y otros temas relevantes.

La metodología empleada sigue el marco del proceso estándar CRISP-DM (Cross-Industry Standard Process for Data Mining), que proporciona una estructura organizada y rigurosa para abordar proyectos de análisis de datos. Con respecto a las técnicas de análisis, se aplica enfoques

de aprendizaje automático supervisado y no supervisado. Inicialmente, se lleva a cabo un análisis descriptivo y exploratorio de las variables para comprender las tendencias y relaciones en los datos. Luego, se implementa modelos de aprendizaje supervisado para predecir y clasificar el rendimiento académico de los estudiantes. Además, se explorarán variables importantes y su relación con el rendimiento. Finalmente, se utiliza el análisis de clústeres para identificar grupos de estudiantes con características y necesidades similares, lo que permite obtener una visión más detallada y personalizada del rendimiento académico en función de los factores identificados. Esta información es invaluable para diseñar estrategias educativas y políticas públicas que aborden de manera efectiva las necesidades específicas de cada grupo de estudiantes.

La relevancia de esta investigación radica en que el análisis del rendimiento académico proporciona una base sólida y técnica para los responsables de políticas educativas, ya que permite identificar las características y factores que influyen en el rendimiento estudiantil. Esto facilita la toma de decisiones basada en datos para focalizar esfuerzos encaminados a mejorar el proceso de enseñanza-aprendizaje, como por ejemplo, mediante programas de acceso a la educación superior. Además, este estudio permite identificar grupos de estudiantes en situación de vulnerabilidad o riesgo académico, lo que contribuye a proporcionar asistencia adecuada para prevenir la deserción estudiantil a largo plazo en carreras universitarias (Ahmad et al., 2015).

La aplicación de métodos de aprendizaje automático enriquece la investigación empírica en este ámbito, ofreciendo herramientas técnicas para un análisis más profundo de la problemática. Instituciones educativas y responsables políticos pueden utilizar estos hallazgos para desarrollar intervenciones y estrategias personalizadas que mejoren el acceso y el éxito en la educación superior. Asimismo, estos resultados pueden ser útiles para identificar áreas en las que se requieren ajustes o innovaciones. También pueden servir de base para investigaciones futuras, que profundicen en los mecanismos subyacentes que influyen en el rendimiento académico y en el desarrollo de soluciones más efectivas y sostenibles para mejorar la calidad y la equidad en la educación.

La estructura de esta investigación se presenta en varias secciones. La segunda sección contiene una revisión de la literatura, donde se recopilan estudios que brindan evidencia empírica acerca de los factores que inciden en el rendimiento académico, así como los enfoques metodológicos

utilizados en dichos estudios. En la tercera sección, se describen los métodos, datos, variables y técnicas empleadas en el desarrollo de la presente investigación. Esto incluye una explicación detallada del proceso de recolección y tratamiento de los datos, así como de la aplicación de las técnicas de aprendizaje automático en el análisis.

La cuarta sección presenta los principales hallazgos obtenidos a partir del análisis de datos, con un enfoque en la identificación de los factores que influyen en el rendimiento académico y en la predicción y clasificación de los estudiantes según su desempeño. Además, se discuten los resultados en el contexto de los estudios previos y se proponen posibles explicaciones para las tendencias observadas. Por último, en la quinta sección, se resumen las conclusiones más relevantes de la investigación y se abordan las implicaciones prácticas y políticas de los hallazgos. También se sugieren áreas de investigación futura, con el objetivo de continuar profundizando en el conocimiento de los factores que determinan el rendimiento académico y en el desarrollo de estrategias efectivas para mejorar la calidad y la equidad en la educación.

2. Revisión de la literatura y marco teórico

2.1. Educación secundaria y su relación con la educación superior

La educación secundaria desempeña un papel fundamental en el proceso de educación superior, ya que estas instituciones preparan a los estudiantes para la etapa postsecundaria al proporcionarles una base sólida en su desarrollo académico y social, así como al promover las habilidades necesarias para alcanzar el éxito académico (Kuh et al., 2006). En cuanto a la preparación académica, la educación secundaria es responsable de establecer fundamentos sólidos en materias como matemáticas, inglés, entre otras, a la vez que fomenta el pensamiento crítico, las habilidades de aprendizaje, las técnicas de estudio y el desarrollo de la capacidad para resolver problemas (Conley, 2007; Venezia & Jaeger, 2013; M. Wang & Degol, 2014).

Además, durante la educación secundaria se refuerzan aspectos como la autodisciplina, motivación y autoestima (Schunk & Greene, 2017). Estas habilidades y actitudes son fundamentales no solo para afrontar desafíos académicos en la educación superior, sino también para el desarrollo integral del estudiante, preparándolos para enfrentar con éxito las demandas futuras en la vida profesional y personal (Zins et al., 2007). Por ello, es esencial que las instituciones de educación secundaria apliquen enfoques pedagógicos efectivos y ofrezcan apoyo

adecuado a sus alumnos, asegurando una transición exitosa a la educación superior y, en última instancia, mejorando las oportunidades y bienestar de los jóvenes (Farrington et al., 2012).

Múltiples instituciones de educación superior requieren ciertos puntajes o calificaciones mínimas durante la secundaria para admitir a los estudiantes. En este contexto, el rendimiento académico en la secundaria se considera un factor importante para la admisión y uno de los indicadores del potencial de un estudiante para tener éxito en la educación superior (Farrington et al., 2012).

En la actualidad, la educación superior enfrenta desafíos no solo en términos de acceso equitativo, sino también en relación con el alto nivel de deserción estudiantil (Pérez et al., 2020). La transición exitosa hacia la educación superior depende de varios factores, entre ellos, las competencias desarrolladas durante la educación secundaria, ya que una transición inadecuada puede ocasionar un bajo rendimiento académico y, en algunos casos, la deserción temprana de la carrera (M. Wang & Degol, 2014). En el Ecuador, 5 de cada 10 postulantes a la educación superior logran ingresar a una universidad, pero además, existe un 28% de estudiantes que abandona la carrera en el primer año de estudios, lo que demuestra que no solo el acceso sino también la sostenibilidad o permanencia en una carrera universitaria son desafíos importantes (SENESCYT, 2021). En este contexto, es fundamental promover la colaboración entre las instituciones de educación secundaria y superior para facilitar una transición exitosa y abordar las brechas en la preparación de los estudiantes, ofreciendo orientación y recursos adicionales según se requiera (Kirst & Venezia, 2004).

2.2. Rendimiento académico

El rendimiento académico se define como el desempeño que refleja los conocimientos, habilidades y valores adquiridos por los estudiantes al finalizar un curso o al culminar sus estudios (Namoun & Alshanqiti, 2020). De este modo, el rendimiento o desempeño trasciende las calificaciones o puntajes y se manifiesta en la habilidad para alcanzar objetivos postsecundarios, como obtener empleo o ingresar a una carrera profesional (Pascarella & Terenzini, 2005). York et al. (2015) sostienen que el éxito académico es un concepto amplio que abarca no solo la adquisición de conocimientos, habilidades y competencias, sino también la satisfacción y la capacidad de permanecer en la educación superior. En consecuencia, se considera que el éxito académico es un constructo complejo de medir (Kuh et al., 2006).

Para evaluar el rendimiento académico, se han propuesto diversos enfoques, como el uso de calificaciones finales, promedio de calificaciones, perspectivas laborales futuras, logro de resultados de aprendizaje y continuidad en estudios de nivel superior (Namoun & Alshanqiti, 2020). A menudo, se recurre a las calificaciones promedio del curso o programa a pesar de sus limitaciones, ya que se consideran una aproximación útil para medir el logro de aprendizaje y la adquisición de habilidades y competencias, además, estas calificaciones suelen estar fácilmente disponibles para su análisis (Astin, 1997).

En este contexto, predecir el rendimiento académico puede ofrecer múltiples beneficios, como mejorar la planificación y estrategias de gestión en la educación secundaria. Basándose en los resultados, es posible identificar medidas de apoyo para estudiantes con menor rendimiento y buscar acciones para reducir las tasas de deserción estudiantil (Brahim, 2022). Adewale et al. (2018) sostienen que este tipo de investigaciones permite focalizar los recursos de aprendizaje acorde a las capacidades de los diferentes grupos de estudiantes según su desempeño académico. De esta manera, la educación resulta más efectiva y permite identificar de forma temprana a aquellos grupos con mayores dificultades en el desempeño.

2.3. Determinantes del rendimiento académico

Numerosas investigaciones, tanto a nivel nacional como internacional, examinan los factores socioeconómicos, académicos, motivacionales e institucionales que influyen en el rendimiento académico de los estudiantes de bachillerato. York et al. (2015) sostienen que los resultados en un curso o programa educativo dependen de los insumos, características demográficas y socioeconómicas, así como de los aspectos educativos recibidos por los estudiantes.

A nivel nacional, en Ecuador, Villarruel-Meythaler et al. (2020) emplearon una regresión logística para evidenciar que aspectos de preparación privada y factores internos relacionados con la motivación influyen en la obtención de mejores calificaciones en los estudiantes de educación media en el examen Ser Bachiller durante el período 2016-2017. Un estudio más reciente, utilizando regresión lineal múltiple, encontró que el índice de estratificación socioeconómica y la etnia del estudiante también afectan el rendimiento académico en el período 2018-2019 (Sandoval Vega, 2022).

Dentro de las características demográficas, la etnia y el género, suelen ser empleadas en modelos de predicción del rendimiento académico (Rastrollo-Guerrero et al., 2020). Sanchez (2011) encontró que existen brechas entre diferentes etnias lo que se ve reflejado en el rendimiento académico. El factor geográfico también puede influir en el rendimiento, los estudiantes que residen en áreas geográficas remotas pueden enfrentarse a instituciones con niveles más bajos de calidad educativa, lo que afecta su desempeño (Gershenson, 2013).

Entre los factores socioeconómicos, un ingreso familiar bajo puede incidir en la limitada accesibilidad a herramientas tecnológicas necesarias para el estudio en la actualidad, lo cual perjudica el rendimiento (Park et al., 2011). Los grupos de estudiantes socialmente excluidos, ya sea por etnia o recursos económicos, enfrentan mayores desafíos para mantener su desempeño académico, puesto que pueden disponer de menos recursos o incluso experimentar dificultades de acceso a la educación (Reardon & Owens, 2014). Otro factor considerado como influyente en el desempeño es el trabajo, de acuerdo a Carrillo & Ríos (2013) un estudiante que trabaja puede disponer de menos tiempo para las actividades de educación afectando negativamente el rendimiento.

Los factores como programas de formación orientados a la preparación previa al examen también ejercen una influencia positiva en el rendimiento, según el estudio de Theobald (2021). Dubuc et al. (2020) encontraron que los hábitos de estudio, el uso de redes sociales y los estilos de vida están correlacionados con el rendimiento académico, además de observar diferencias entre hombres y mujeres. En relación con las herramientas académicas, Rodríguez Rosero et al. (2021) descubrieron que el acceso a tecnologías de aprendizaje aumenta la probabilidad de obtener un mejor rendimiento.

Diversas estrategias de aprendizaje pueden contribuir a mejorar el rendimiento académico. Según la evidencia hallada por Bressane et al. (2022), la revisión de lecciones y la lectura bibliográfica tienen un impacto más significativo en el rendimiento que la asistencia a clase y el control de emociones. La gestión de las instituciones también es analizada, ya que diferentes estructuras administrativas pueden generar distintos resultados en el rendimiento, según Masud et al. (2019). A nivel institucional, además del tipo de gestión, el acceso limitado a recursos humanos calificados puede afectar negativamente el desempeño al proporcionar una educación de baja

calidad (Hanushek & Rivkin, 2012). Otro factor que puede influir negativamente en el rendimiento, es la pérdida de años escolares, Rumberger (1995) sostiene que un estudiante se puede ver afectado en el rendimiento al sufrir desmotivación por la pérdida escolar, así como por la falta de estrategias para la reincorporación.

Con respecto a los determinantes motivacionales, Pintrich (2003) sostiene que estos factores desempeñan un papel significativo en la explicación del rendimiento académico, ya que reflejan aspectos relacionados con el bienestar y las percepciones del entorno académico, así como el logro de los objetivos del estudiante. Adicionalmente, aspectos como la autoestima y los estilos de crianza han sido analizados en menor medida por estudios como el de Masud et al. (2019), quienes encontraron evidencia de su influencia en el rendimiento académico.

Por otro lado, Namoun & Alshanqiti (2020) afirman que la participación familiar en el proceso educativo también influyen en el rendimiento académico. Así, Wang & Eccles (2012) indican que un entorno familiar que provea de apoyo emocional al estudiante puede mejorar el rendimiento académico.

2.4. Métodos de aprendizaje automático utilizados para predecir el rendimiento

El método más utilizado como procedimiento para la aplicación de técnicas de aprendizaje automático se basa en el modelo CRISP-DM (Kabathova & Drlik, 2021; Namoun & Alshanqiti, 2020). Dentro de las técnicas de aprendizaje automático, de acuerdo a estudios de revisión de literatura, los métodos supervisados son los aplicados en los múltiples estudios para la predicción del rendimiento (Namoun & Alshanqiti, 2020), como por ejemplo: Regresión lineal múltiple, Regresión logística, Deep learning (Li & Liu, 2021), Árboles de decisión, Grandient Boosted Trees (Nagy & Molontay, 2018), Redes neuronales (Adewale et al., 2018), Support vector machine (Brahim, 2022) y otros (Rastrollo-Guerrero et al., 2020).

Según el estudio de revisión de literatura de Namoun & Alshanqiti (2020) las técnicas Random Forest híbrido, Feedforward Neural Network y Naïve Bayes, son los que mejor resultado presentan en la evaluación del modelo de rendimiento. Mientras que, Bressane et al. (2022) hallaron que el sistema de inferencia difuso tiene mayor capacidad de predicción del desempeño académico, y las técnicas de peor rendimiento fueron la regresión lineal y la regresión logística de efectos mixtos.

A priori no es posible seleccionar una u otra técnica de aprendizaje, pues cada uno presenta ventajas y desventajas y funcionan mejor de acuerdo a los datos y el contexto en el que se encuentre realizando el estudio. En este sentido métricas de desempeño de los modelos debe ser utilizado para la evaluación del nivel de predicción. Xiao et al. (2022) menciona que algunas métricas comúnmente utilizadas son: precisión, ROC AUC, exactitud y sensibilidad para modelos de variable dependiente cualitativa y la Raíz del error cuadrático medio (RMSE por sus siglas en inglés), R cuadrado y Error absoluto medio (MAE por sus siglas en inglés) para variable dependiente cuantitativa. Se debe enfatizar que el tipo de métrica utilizado depende del tipo de variable a analizar, ya que no todas las métricas pueden ser calculadas para todos los tipos.

3. Metodología

La investigación analiza el rendimiento académico de los estudiantes de bachillerato del período lectivo 2021-2022 en el Ecuador, para ello, se aplica un estudio de tipo descriptivo, correlacional y aplicado. En primer lugar, es descriptivo porque busca identificar y caracterizar perfiles de estudiantes con mayor rendimiento. En segundo lugar, la investigación es de tipo correlacional porque tiene como objetivo determinar los factores que explican el rendimiento académico de los estudiantes de bachillerado en Ecuador. De forma que, se establezcan las relaciones entre el desempeño y los distintos factores a través de correlaciones. Además, es una investigación aplicada porque se aporta conocimiento de un fenómeno en específico (rendimiento académico) a través de los resultados empíricos o prácticos que se generan de la aplicación de técnicas y métodos para resolver el problema planteado, con base al marco teórico desarrollado.

3.1. Flujo del proceso aplicado para el análisis del rendimiento académico

Para responder las preguntas de investigación el estudio aplica la metodología CRISP-DM y los pasos de desarrollo de un proyecto de inteligencia artificial de acuerdo al diagrama de flujo descrito en el Gráfico 2. En la primera etapa, se analiza la problemática referente a la importancia del rendimiento académico de los bachilleres presentado en la sección uno y dos del estudio.

En esta etapa y con base al objetivo planteado en la presente investigación se recopila los datos cuya unidad de análisis lo conforman los estudiantes que cursaron el tercer grado de bachillerato general unificado en Ecuador en el período 2021-2022. Para ello, se fusiona la base de datos “micro” y “factores asociados estudiantes” publicado por INEVAL. La primera contiene

información socioeconómica general y notas de los estudiantes en diferentes ramas como la matemática , lengua y literatura, ciencias naturales y ciencias sociales. La segunda base contiene variables que miden características educacionales y de percepción del estudiante.

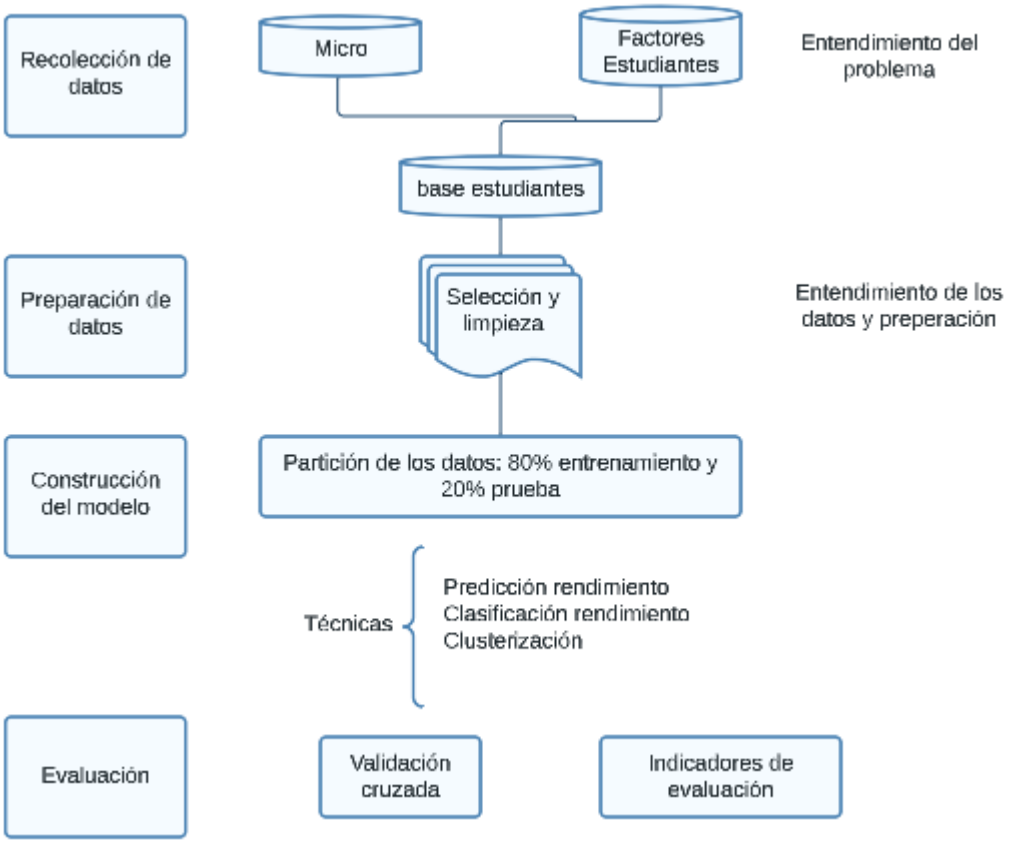


Gráfico 2. Diagrama de flujo del proceso para el análisis del rendimiento académico

3.2. Preparación de los datos

La base fusionada “estudiantes”, de acuerdo a lo descrito en la sección 3.1, tiene una estructura de corte transversal con 245 variables y 18697 registros, de los cuales un 33.06% corresponde a estudiantes de tercero de bachillerato y el restante pertenece a estudiantes de séptimo y décimo de básica. La investigación se enfoca en los estudiantes de tercero de bachillerato, por lo que, en esta segunda etapa, se filtra la base y se trabaja inicialmente con 6181 registros.

Con respecto a las variables, el rendimiento académico de los estudiantes es la característica de interés del estudio. La base contiene ocho materias evaluadas con un puntaje de 0 a 1000: matemática, lengua y literatura, física, química, biología, historia, filosofía y educación para la ciudadanía. La variable rendimiento académico denominado “nota” fue obtenido a través del

promedio de las calificaciones de las ocho materias. Además, los estudiantes fueron categorizados por tipo de rendimiento: elemental (mayor o igual a 700 y menores a 800), y con rendimiento insuficiente (menores a 700).

Las variables explicativas fueron seleccionadas de acuerdo a la revisión de literatura descrita en la sección 2.3 y según la disponibilidad de datos (ver Tabla 1). Un total de 18 variables explicativas de tipo cuantitativas y cualitativas fueron elegidas, mismas que contienen información de factores socioeconómicos, de educación y de percepción de estado del estudiante.

Tabla 1. Variables de rendimiento, socioeconómicas, y educación

Variable	**Descripción**	**Dominio**
Rendimiento escolar		
Nota	nota promedio de ocho materias evaluadas	0-1000
Elemental	Estudiantes que obtuvieron una nota elemental	1 si obtuvo una nota elemental, 0 si la nota fue insuficiente
Características socioeconómicas		
Hombre	sexo del estudiante	1 hombre, 0 mujer
Edad	años del estudiante	16-22
Mestizo	Estudiante que se autoidentifica como mestizo	1 mestizo, 0 otros
Sierra	Estudiantes que viven en la región sierra	1 sierra, 0 otras regiones
Urbano	Estudiantes que viven en la región urbano	1 urbano, 0 rural
Trabajo	Estudiantes que trabajan	1 si, 0 no
computador	Número de días a la semana que utiliza el computador en el hogar	0-7
Quintil	Quintil socioeconómico del estudiante	1-5
Características de educación		
Días	Número de días a la semana que repasa la materia o hace tareas escolares en la casa	0-7
Horas	Estudiantes que repasan las materias o hacen las tareas en la casa más de dos horas al día	1 si, 0 no
privado	Estudiantes que provienen de institutos privados	1 si, 0 no
fiscomisional	Estudiantes que provienen de institutos fiscomisionales	1 si, 0 no
Faltas	Estudiantes que faltan a clases una o más veces	1 si, 0 no

perdido	Estudiantes que han perdido algún año escolar desde educación básica hasta bachillerato	1 si, 0 no
Carrera	Estudiantes que indican que desean continuar estudiando una carrera	1 si, 0 no
Percepción de estado del estudiante		
satisfacción	Satisfacción con las personas con las que vive el estudiante	1 nada satisfecho al 5 totalmente satisfecho
contento	Estudiante contento con su vida	1 totalmente en desacuerdo al 5 totalmente de acuerdo
Estado	Estudiantes que indican que se sienten contentos en la escuela	1 si, 0 no

Elaboración: propia
Fuente: INEVAL (2022)

Después de seleccionar las características de la base fusionada “estudiantes” que son analizadas en el presente estudio conforme a la Tabla 1, se llevó a cabo una depuración y limpieza que incluyó los siguientes pasos: (1) eliminar los registros de los estudiantes que no fueron evaluados, (2) descartar los registros de los estudiantes que carecen de información sobre su rendimiento académico, y (3) eliminar los datos extremos en la variable de rendimiento estudiantil (calificación = 400 puntos). Tras este proceso, la base de datos depurada cuenta con 5069 registros. Finalmente, se llevó a cabo la imputación de datos en variables como el quintil del nivel de ingresos y la etnia, utilizando la moda como método de imputación, tomando en consideración la ubicación geográfica a nivel provincial.

3.3. Construcción del modelo

Antes de desarrollar el modelo, se llevó a cabo la partición del conjunto de datos en un 80% para entrenamiento y un 20% para prueba. En la tercera etapa, con el objetivo de predecir y clasificar el rendimiento estudiantil, se aplicaron varios métodos supervisados, mientras que para identificar clústeres se utilizaron técnicas no supervisadas. El Gráfico 3 resume las técnicas empleadas en cada caso.

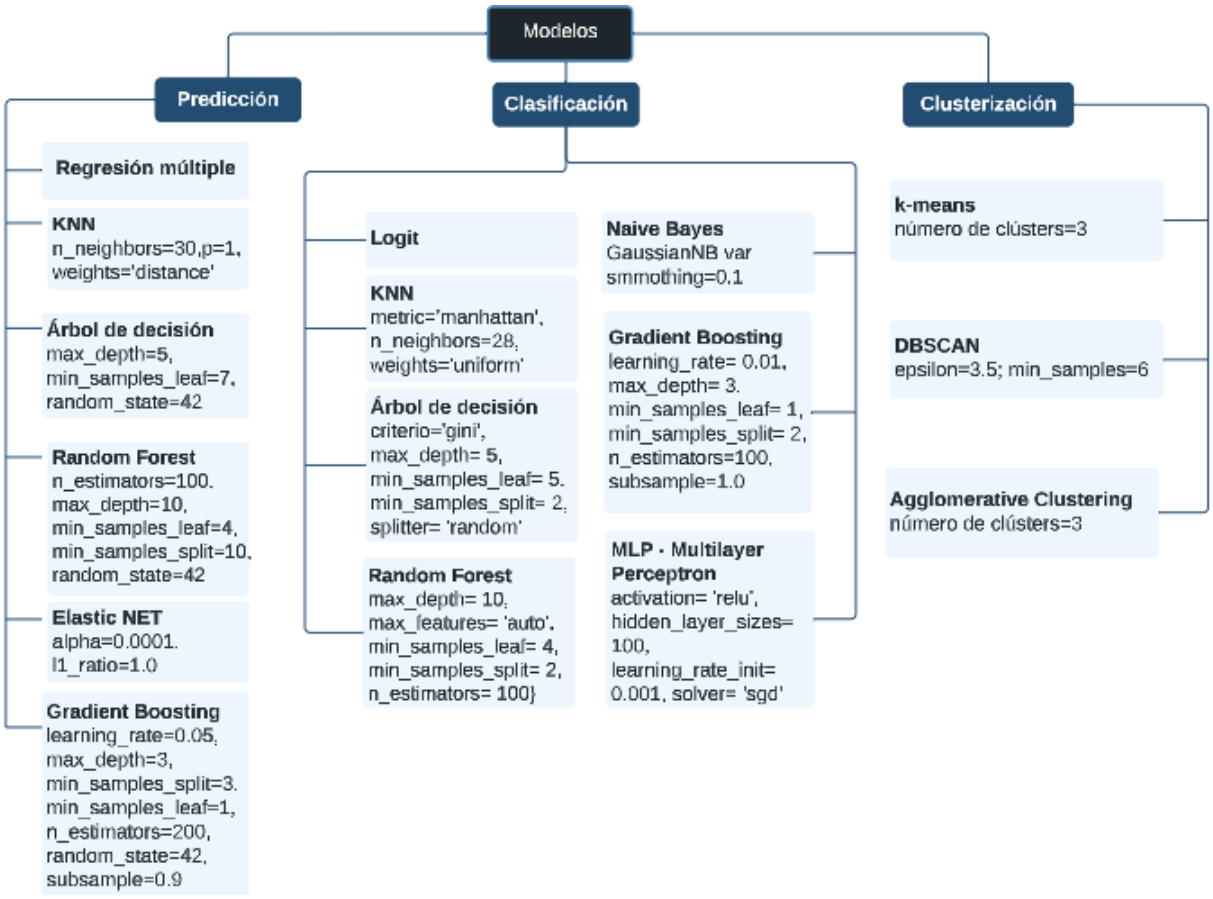


Gráfico 3. Técnicas de aprendizaje automático aplicados en el análisis del rendimiento académico

Elaboración: propia

Predicción del rendimiento y clasificación

Para predecir el rendimiento, la variable dependiente "nota" es de tipo cuantitativa, y en el caso de la clasificación del rendimiento la variable "elemental" es de tipo cualitativa dicotómica. Para ambos casos se aplica las técnicas: K vecinos más cercanos (KNN), Árbol de decisión, Random Forest y Gradient Boosting, pues son modelos supervisados que se puede aplicar a ambos casos. Para el caso de la predicción del rendimiento también se aplica la regresión lineal múltiple y Elastic Net. Mientras que, en el caso de la clasificación del rendimiento se utilizaron adicionalmente el modelo logit, Naive Bayes y Multilayer perceptrón (MLP). Se aplicaron varias técnicas debido a que no es posible seleccionar a priori una técnica que permita obtener mejores resultados en términos de pronóstico, pues cada una presenta sus ventajas y desventajas. La Tabla 2 muestra un resumen de las ventajas y desventajas de cada método de acuerdo a la literatura.

Tabla 2. Ventajas y desventajas de los modelos implementados

Método	Ventajas	Desventajas
K-vecinos más cercanos (KNN)	- Fácil de implementar - Robusto frente al ruido - Efectivo con grandes conjuntos de entrenamiento - No paramétrico	- Requiere determinar el valor de k - No funciona bien con un gran número de características - Sensible a la escala de las variables
Árbol de decisión	- Simple de entender - Factible para resolver problemas de decisiones - Maneja datos categóricos y numéricos - Menos sensible a las escalas de las variables	- Puede presentar sobreajuste con datos con ruido - Inestable a cambios en los datos
Random Forest	- Reduce el sobreajuste en comparación al árbol de decisión - Maneja bases con alta dimensionalidad	- Menos interpretable que el árbol de decisión - Requiere más tiempo para entrenar y predecir - Computacionalmente costoso
Gradient Boosting	- Maneja datos de alta dimensionalidad - Robusto a los valores atípicos	- Menos interpretable - Costos computacionales más grandes - Sensible a la configuración de hiperparámetros
Regresión lineal múltiple	- Fácil de entender e interpretar - Adecuado cuando existe una relación lineal entre las variables	- No funciona correctamente si no hay relación lineal - Supuestos como homocedasticidad, no autocorrelación y no multicolinealidad deben analizarse
Elastic Net	- Combina las fortalezas de Lasso y Ridge - Maneja la multicolinealidad	- Menos interpretable - Requiere seleccionar coeficientes de regularización y penalización - Costos computacionales más grandes
Modelo Logit	- Fácil de implementar e interpretar - Adecuado cuando existe una relación lineal entre las características y la transformación log-odds de la variable objetivo	- No funciona en casos donde existe alta dimensionalidad - Sensible a la multicolinealidad
Naive Bayes	- Fácil de implementar e interpretar - Funciona bien con alta dimensionalidad o un gran número de características - Rápido en tiempo de entrenamiento y predicción	- Asume independencia entre las características - Puede no funcionar bien cuando las características están altamente correlacionadas - La estimación de las probabilidades puede ser incorrecta si la suposición de independencia no se cumple
Multilayer Perceptron (MLP)	- Puede aprender relaciones no lineales - Funciona bien con conjuntos de datos de alta dimensionalidad o un gran número de características	- Menos interpretable - Requiere selección y ajuste de hiperparámetros - Tiempo de entrenamiento más largo - Puede ser propenso al sobreajuste

	- Capaz de aprender y aproximar funciones complejas

Elaboración: propia con base a Chollet & Allaire (2018); Géron (2022); James et al. (2013); Kelleher et al. (2015); Raschka & Mirjalili (2019).

Cada modelo cuenta con sus propios hiperparámetros, y los resultados pueden variar al ajustar estos valores. Por esta razón, se utilizó la librería GridSearchCV en Python para buscar los hiperparámetros óptimos. El Gráfico 3 muestra los hiperparámetros óptimos obtenidos. Para los modelos empleados en la predicción, la búsqueda de hiperparámetros se centró en minimizar el error cuadrático medio (MSE, por sus siglas en inglés). En cambio, para el caso de la clasificación, se buscaron hiperparámetros óptimos que maximizaran la precisión de la predicción.

Clusterización

Con el fin de identificar los perfiles relacionados con el rendimiento académico, se utiliza el modelo no supervisado de clusterización, mismo que permite identificar los segmentos de estudiantes que presentan características similares (Wang, 2022). Se aplican las técnicas k-means, DBSCAN y Agglomerative Clustering.

K-Means es una técnica de clustering que busca agrupar datos en k grupos con centroides representativos (Jain, 2010; MacQueen, 1967). Una de las principales ventajas de K-Means es su simplicidad y rapidez para procesar grandes conjuntos de datos (Jain, 2010). Sin embargo, una K-Means requiere la especificación previa del número de grupos (k), lo que puede ser difícil de determinar de antemano (MacQueen, 1967).

DBSCAN es una técnica de clustering que agrupa puntos en función de su proximidad y densidad (Ester et al., 1996; Schubert et al., 2017). Una de las ventajas de DBSCAN es que puede identificar grupos de forma automática sin necesidad de especificar un número de grupos previamente, es eficaz para identificar grupos de diferentes formas y tamaños y puede manejar ruido en los datos (Ester et al., 1996). Sin embargo, una desventaja de DBSCAN es que su desempeño puede ser sensible a la elección de parámetros, como el radio y la densidad mínima, y puede ser menos eficiente para datos de alta dimensionalidad (Schubert et al., 2017).

Agglomerative Clustering es una técnica de clustering que agrupa datos de forma jerárquica, fusionando grupos similares (Murtagh & Legendre, 2014). Una ventaja de Agglomerative Clustering es que puede crear una estructura jerárquica de grupos, lo que permite una mayor comprensión de la relación entre los grupos, es eficaz para manejar diferentes tipos de similitud y distancia (Maimon & Rokach, 2005). Sin embargo, una desventaja de Agglomerative Clustering es que puede ser computacionalmente costoso para grandes conjuntos de datos y puede ser sensible a la elección de parámetros (Tan et al., 2018).

Cabe destacar que, al igual que para la predicción y clasificación, cada uno de estos métodos cuenta con diferentes hiperparámetros que deben ser seleccionados de manera adecuada para obtener los mejores resultados. Por lo que, los hiperparámetros fueron obtenidos maximizando el coeficiente de silueta (ver Gráfico 3) que es uno de los indicadores utilizados para medir la calidad de los clústeres (Kaufman & Rousseeuw, 2009; Rousseeuw, 1987).

3.4. Evaluación del modelo

Una vez que los modelos son entrenados, es necesario evaluar su rendimiento para determinar cuál de ellos es el mejor para predecir y clasificar el rendimiento académico. Para ello en la cuarta etapa, se utilizan diferentes indicadores de evaluación, dependiendo del tipo de problema (regresión o clasificación). El Gráfico 4 resume los indicadores utilizados para cada caso. Es importante tener en cuenta que no hay un único indicador que sea el mejor para evaluar el rendimiento de un modelo, ya que cada indicador proporciona una información diferente sobre el modelo. Por lo tanto, es recomendable evaluar varios indicadores y seleccionar el modelo que tenga el mejor desempeño en términos generales y que sea el más adecuado para el problema que se está resolviendo.

El modelo que presenta el menor error de predicción medido a través del Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y Error Porcentual Absoluto Medio (MAPE) y un mayor coeficiente de determinación (R2) se considera como el mejor modelo para predecir el rendimiento académico.

El MAE mide la diferencia promedio absoluta entre los valores reales y los valores predichos por el modelo. El MSE mide el promedio de los errores al cuadrado entre los valores reales y los valores predichos por el modelo (Hanke & Wichern, 2009). La raíz cuadrada del MSE es el

RMSE. El MAPE mide el promedio del porcentaje del error absoluto entre los valores reales y los valores predichos por el modelo (Greene, 2003). Cuanto menor sea el valor del MAE, MSE, RMSE y MAPE, mejor será el modelo para predecir el rendimiento académico (Li & Liu, 2021). El coeficiente de determinación (R2) es una medida de ajuste del modelo que mide la proporción de la variabilidad de la variable objetivo que es explicada por el modelo. R2 toma valores entre 0 y 1, y cuanto mayor sea el valor de R2, mayor será la proporción de la variabilidad explicada por el modelo (Wooldridge, 2010).

En el caso de la clasificación, todos los indicadores presentados en el Gráfico 4 toman valores entre 0 y 1. Cuanto más altos sean estos valores, mejor será la capacidad de clasificación del modelo. La precisión indica el porcentaje de predicciones correctas para una clase específica realizadas por el modelo, mientras que la exactitud mide la proporción de veces que el modelo predice correctamente todas las clases (Sokolova & Lapalme, 2009). La sensibilidad, también conocida como recall, mide el porcentaje en el que el modelo predice correctamente una clase con respecto a las predicciones correctas de la clase analizada y las predicciones incorrectas del resto de clases (Jawad et al., 2022). El área bajo la curva ROC (AUC) es una medida de calidad de la clasificación que evalúa la capacidad del modelo para diferenciar entre las clases (M. S. Ahmad et al., 2021). Un valor de 1 indica que el modelo clasifica perfectamente. Sin embargo, a menudo un alto indicador de exactitud (accuracy) puede estar acompañado de bajos niveles de precisión, sensibilidad (recall) u otros indicadores.

Por lo tanto, para el caso de estudio se selecciona el modelo que tenga un mayor equilibrio entre recall y precisión, ya que esto asegurará un buen desempeño en la identificación correcta de las clases, así como en la minimización de las predicciones incorrectas. Al buscar este equilibrio, se obtendrá un modelo más robusto y confiable para la clasificación de las clases analizadas.

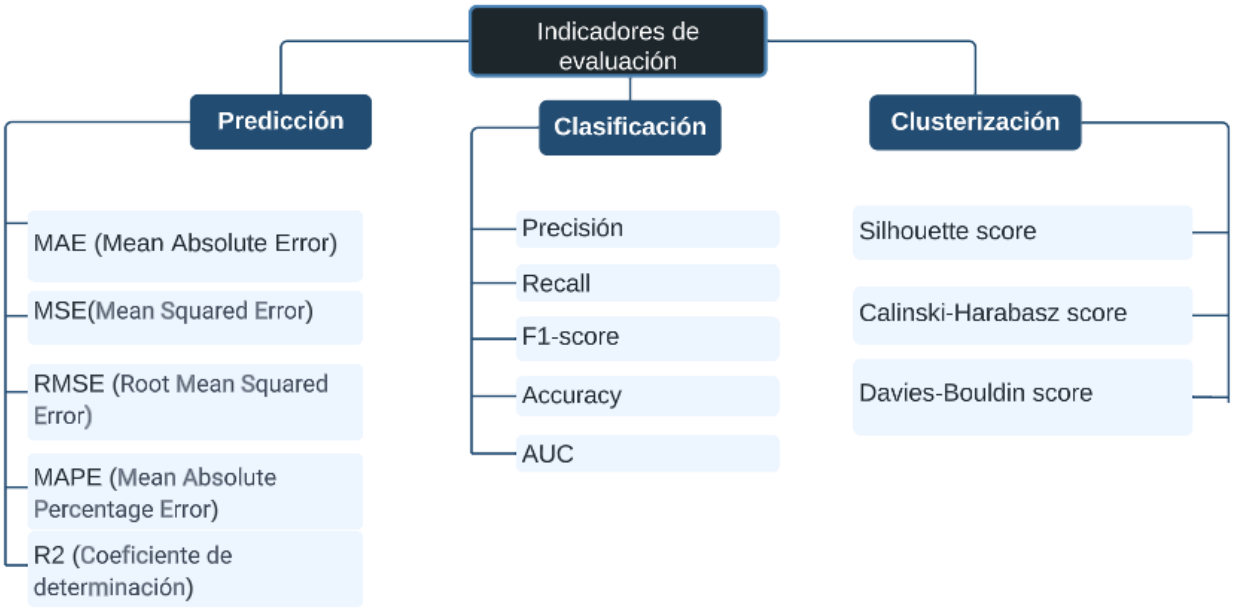


Gráfico 4. Indicadores de evaluación del modelo
Elaboración: propia

Para la clusterización, el modelo utilizado es aquel que presenta mejores puntajes del coeficiente de silueta, Calinski-Harabasz y Davies-Bouldin. El coeficiente de silueta toma valores entre -1 y 1; en general, un coeficiente de silueta más alto es preferible, ya que indica una mejor calidad de clustering (Rousseeuw, 1987). Un valor por encima de 0.5 se considera aceptable y sugiere que los clústeres están razonablemente separados y tienen una buena cohesión interna (Kaufman & Rousseeuw, 2009). Un valor por encima de 0.7 indica un clustering de alta calidad.

El índice de Calinski-Harabasz mide la relación entre la dispersión dentro del clúster y la dispersión entre clústeres, por lo que a mayor valor del índice, se espera una mejor calidad de los grupos generados (Caliński & Harabasz, 1974). Por otro lado, el índice de Davies-Bouldin evalúa la similitud entre los clústeres, teniendo en cuenta tanto la separación entre ellos como la cohesión interna, de forma que, mientras más bajo sea este coeficiente, se espera que los grupos generados presenten mejor calidad (Davies & Bouldin, 1979).

Por último, con el fin de evitar un posible sobreajuste o sesgo en las evaluaciones de los modelos, se aplica la validación cruzada basada en k folds igual a 5, como sugieren James et al. (2013). Esto implica particionar los datos en cinco partes iguales, utilizar uno de ellos como prueba y los cuatro restantes como entrenamiento (80% - 20% entrenamiento y prueba respectivamente), así

sucesivamente para cada partición. La partición se basa en la aleatoriedad de la selección, de esta manera, todos los datos son parte de la prueba y del entrenamiento. Los valores finales de los indicadores de evaluación se calculan por medio del promedio de los resultados de todas las particiones k (Kohavi, 1995). Como resultado, la precisión de las estimaciones mejora y se reduce el sesgo en la evaluación del rendimiento (Varoquaux et al., 2017).

4. Resultados y discusión

Esta sección se enfoca en analizar y presentar los resultados principales obtenidos a partir de la investigación realizada. A través de estos hallazgos, se identifica los factores clave que determinan el rendimiento académico en esta población estudiantil y se explora grupos de estudiantes con características y necesidades similares.

4.1. Análisis descriptivo

A través del análisis descriptivo se proporciona una visión general del rendimiento académico y las características socioeconómicas, de educación y de percepción del estado emocional de los estudiantes. La Tabla 3 revela que el rendimiento académico de los estudiantes tiene una calificación promedio de 689.78/1000, con aproximadamente un 30.20% de estudiantes que obtienen una calificación elemental (700 a menos de 800 puntos). En términos de características socioeconómicas, el 48.53% de los estudiantes son hombres y la edad promedio es de 17.73 años. La mayoría de los estudiantes (84.55%) se autoidentifican como mestizos y el 52.46% estudia en la región de la sierra. Alrededor del 57.57% de los estudiantes estudian en instituciones ubicadas en áreas urbanas y aproximadamente el 32.98% trabaja. En promedio, los estudiantes utilizan una computadora en casa 3.35 días a la semana y están distribuidos en diferentes quintiles socioeconómicos.

Con respecto a las características educativas, los estudiantes dedican un promedio de 4.40 días a la semana para repasar materia o hacer tareas en casa. Aproximadamente el 39.46% estudia más de dos horas al día. Alrededor del 28.84% de los estudiantes asisten a instituciones privadas, mientras que el 25.82% están matriculados en instituciones fiscomisionales. Un total del 42.95% de los estudiantes ha faltado a clases una o más veces y el 13.28% ha repetido un año escolar o más en algún momento (educación básica a secundaria). Es notable que el 90.85% de los estudiantes expresan el deseo de continuar su educación en una carrera.

En cuanto a la percepción del bienestar de los estudiantes, la satisfacción con las personas con las que viven es de 3.97 en promedio en una escala del 1 (totalmente de acuerdo) al 5 (totalmente en desacuerdo). En general, los estudiantes se sienten contentos con sus vidas, con un promedio de 3.69 en una escala del 1 (totalmente en desacuerdo) al 5 (totalmente de acuerdo). Además, el 32.73% de los estudiantes informa sentirse feliz en el colegio. El Anexo 1 muestra las correlaciones lineales de Pearson entre las variables utilizadas.

Tabla 3. Estadísticos descriptivos de las variables

Variable	Promedio	Desv. Std	Min	Max
Rendimiento escolar				
Nota	689.7830	28.8676	505.25	793
Elemental	0.3020	0.4592	0	1
Características socioeconómicas				
Hombre	0.4853	0.4998	0	1
Edad	17.7325	0.8076	16	22
Mestizo	0.8455	0.3614	0	1
Sierra	0.5246	0.4994	0	1
Urbano	0.5757	0.4943	0	1
Trabaja	0.3298	0.4702	0	1
Computador	3.3474	2.5110	0	7
Quintil	2.9817	1.4689	1	5
Características de educación				
Días	4.3985	1.7146	0	7
Horas	0.3946	0.4888	0	1
Privado	0.2884	0.4531	0	1
Fiscomisional	0.2582	0.4377	0	1
Faltas	0.4295	0.4950	0	1
Perdido	0.1328	0.3394	0	1
Carrera	0.9085	0.2884	0	1
Percepción de estado del estudiante				
Satisfacción	3.9671	1.0057	1	5
Contento	3.6861	1.0139	1	5
Estado	0.3273	0.4693	0	1

Elaboración: propia
Fuente: INEVAL (2022)

Desde un análisis exploratorio, se observa que la nota o rendimiento promedio no difiere por sexo hombre o mujer (ver Gráfico 5). Si se analiza la etnia, aquellos que se autoidentifican como mestizos (promedio de 691.25) tienen aproximadamente 10 puntos más que aquellos con otras etnias (691.75) como afroecuatorianos, montubios, indígenas y otros. También se observa diferencias cuando se analiza la localización. Aquellos que estudian en la región sierra (694.6 en promedio) tienen aproximadamente 10 puntos más en comparación con quienes estudian en instituciones ubicados en las regiones costa, oriente y región insular (684.47 en promedio).

Con respecto al área urbano-rural donde se encuentra la institución educativa, se observa que no existe diferencias significativas respecto a las medidas de localización como la mediana y los cuartiles, pero existe una pequeña diferencia promedio a favor de quienes estudian en el área urbana (691.27) con relación al área rural (687.76).

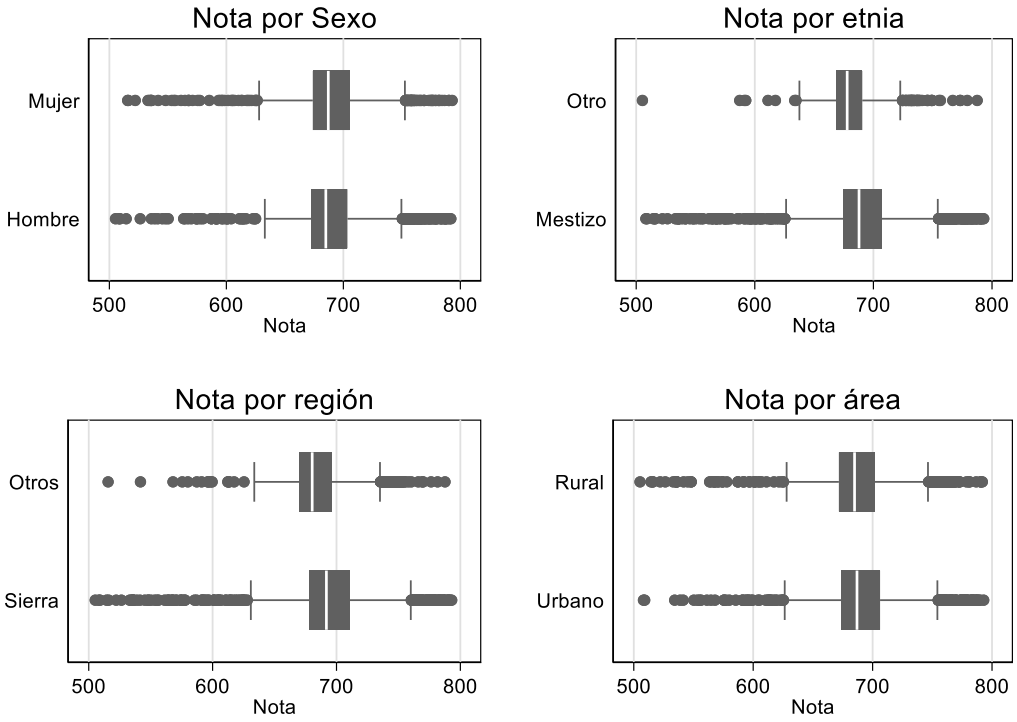


Gráfico 5. Rendimiento académico según características demográficas y de localización
Fuente: INEVAL (2022)

El Gráfico 6 indica que el rendimiento se ve afectado si el estudiante trabaja, pues este grupo tiene en promedio aproximadamente 8 puntos menos (694.61 en promedio) que aquellos que no trabajan (692.33 en promedio). También se observa diferencias respecto al número de días de uso del computador. Así, quienes tienen un rendimiento elemental utilizan el computador en el hogar 4.38 días promedio, en comparación con quienes tienen un rendimiento insuficiente (2.9 días en promedio). Al observar el quintil socioeconómico, se aprecia que existen diferencias de rendimiento a favor de mayores niveles socioeconómicos, así los estudiantes del primer quintil presenta un promedio de rendimiento de 679.84, que es aproximadamente 24 puntos menos que el rendimiento promedio de estudiantes que se encuentran en el quintil cinco (703.86 en promedio). Son los últimos quintiles quienes disponen de mejores condiciones socioeconómicas

que el resto, pudiendo esto influir en su preparación para el examen. Pero no se encontró diferencias promedio en la edad entre los estudiantes que tienen un rendimiento elemental o insuficiente.

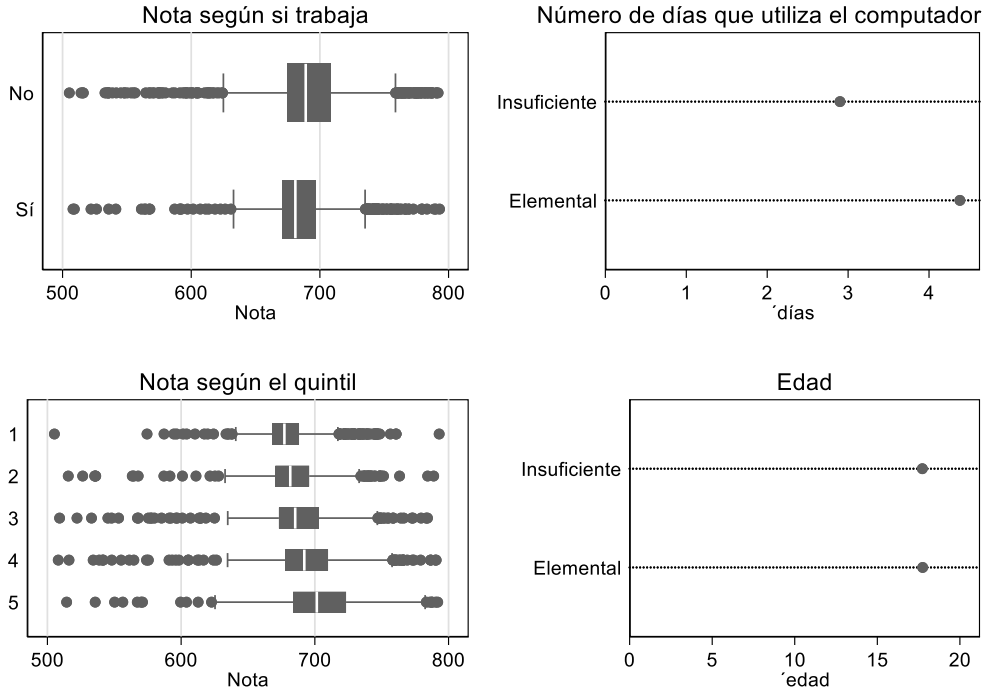


Gráfico 6. Rendimiento académico según características socioeconómicas
Fuente: INEVAL (2022)

Con respecto a las características educacionales (ver Gráfico 7), el número de horas que repasan para realizar tareas o estudiar y el tipo de gestión financiera de la institución parece influir en el tipo de rendimiento. Así, la proporción de estudiantes que tienen una calificación elemental (700 a 800 puntos) es mayor en aquellos grupos que estudian más de dos horas (34.2%) y estudian en instituciones privadas (53.8%) en comparación con quienes estudian dos horas o menos (27.6%) y estudian en otras instituciones (20.6%).

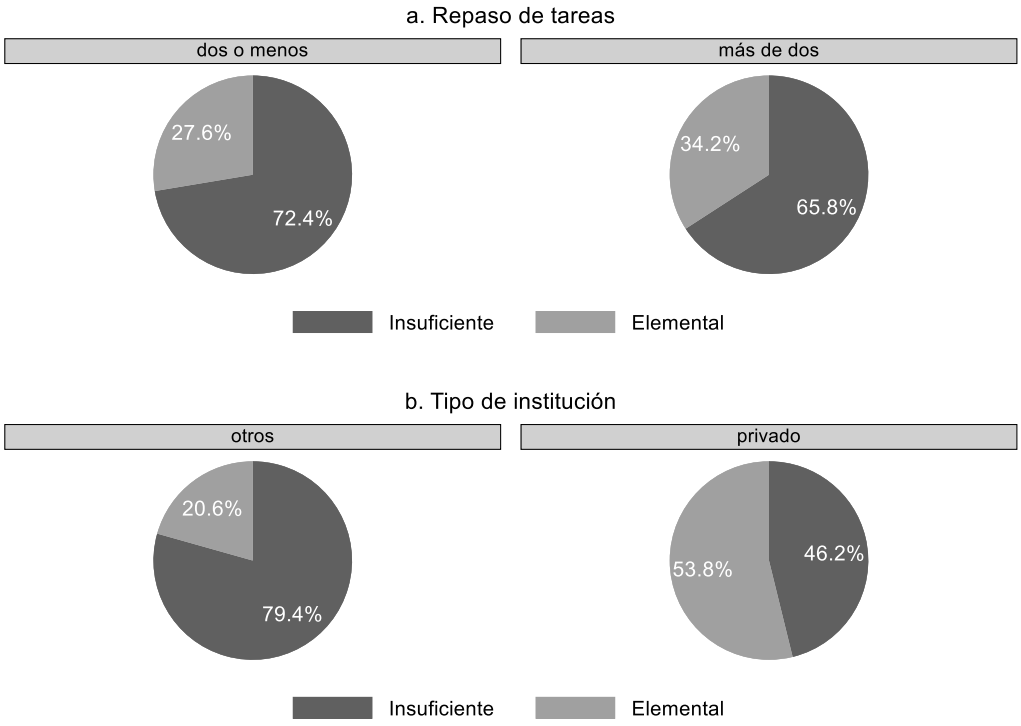


Gráfico 7. Tipo de rendimiento académico según el número de horas que repasan tareas y el tipo de institución
Fuente: INEVAL (2022)

De acuerdo al Gráfico 8, se observa que un 32.1% de estudiantes que faltan una o más veces durante el ciclo lectivo, tienen un rendimiento elemental, frente al 28.8% de quienes no faltan. Mientras que solo un 16% de quienes han pedido un año lectivo tienen una calificación elemental frente al 32.4% del grupo que no han perdido ningún año lectivo. Por lo tanto, las faltas escolares no necesariamente disminuyen el rendimiento académico, pero si la pérdida de año escolar.

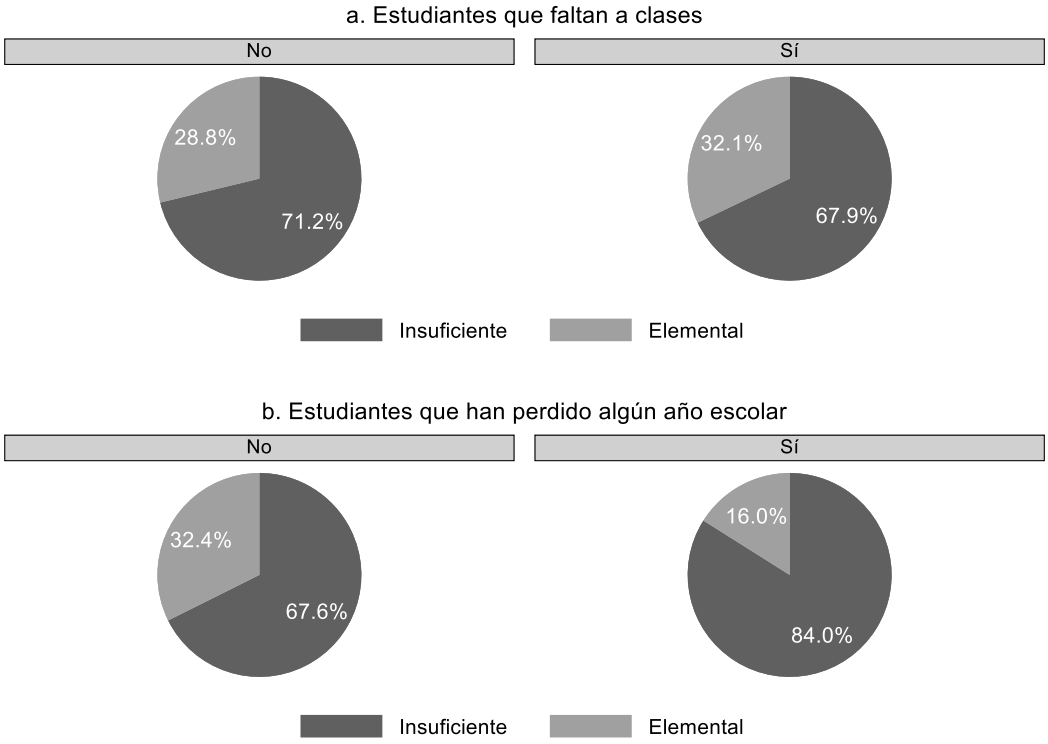


Gráfico 8. Rendimiento académico por a. faltas; b. pérdida de año escolar
Fuente: INEVAL (2022)

Con respecto a la percepción de estado emocional del estudiante, se observa que la proporción de estudiantes con rendimiento elemental incrementa a medida que el estudiante indica que se encuentra muy satisfecho con las personas con las que vive (ver Gráfico 9). Sin embargo, los estudiantes que indican que no se encuentran contentos con su vida o con la escuela presentan mayor puntaje que aquellos que indican que si están contentos. Por último, un 31.6% de los estudiantes que indican que si desean continuar con sus estudios luego de obtener su título de bachiller tienen una calificación elemental, en comparación con quienes no desean continuar estudiante en cuyo caso solo un 16.8% dispone de una nota elemental.

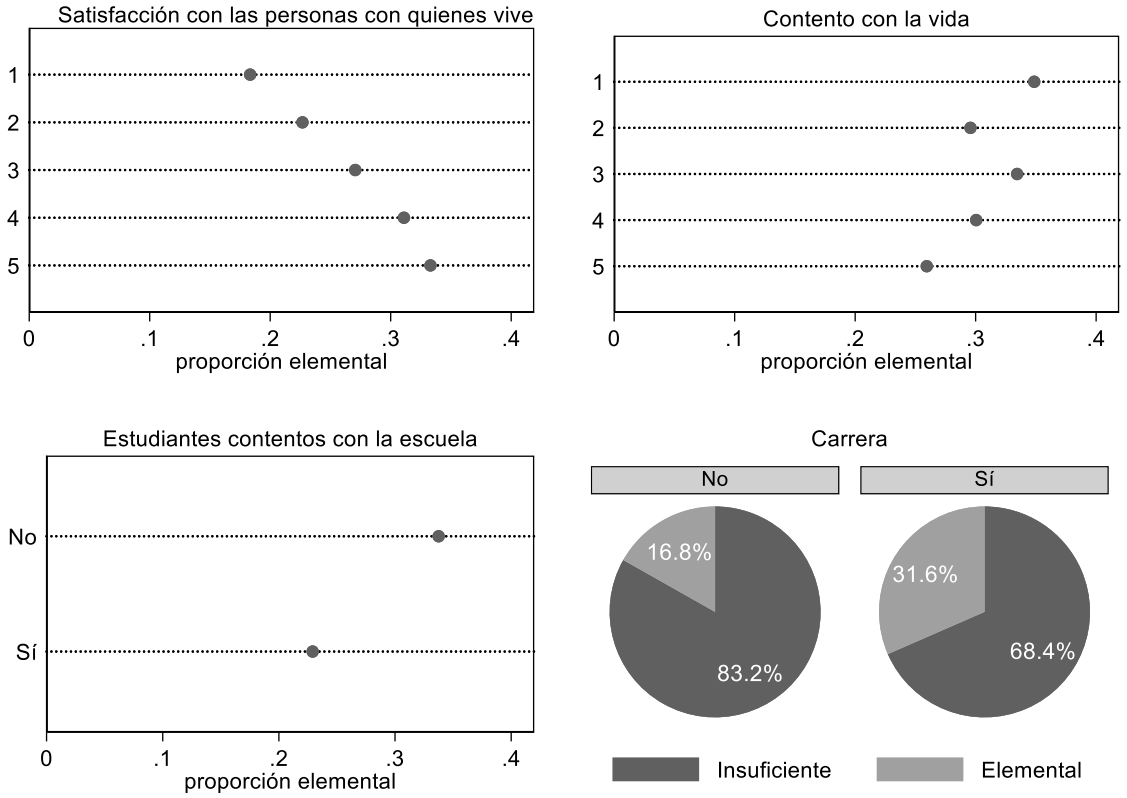


Gráfico 9. Tipo de rendimiento por percepción de estado emocional y continuidad de estudios
Fuente: INEVAL (2022)

En resumen, el análisis descriptivo revela una serie de hallazgos importantes en relación al rendimiento académico y las características socioeconómicas, educativas y emocionales de los estudiantes. Algunos de los principales resultados incluyen la influencia del estatus socioeconómico, el tipo de institución educativa, el tiempo dedicado al estudio y el trabajo en el rendimiento académico. Además, se observa que factores como la satisfacción con las personas con las que viven y el deseo de continuar con estudios superiores también pueden estar relacionados con el rendimiento académico. Sin embargo, algunas variables como el sexo, la edad y la ubicación de la institución educativa (área urbana - rural) no parecen tener diferencias significativas en el rendimiento. Estos hallazgos proporcionan una base para comprender a nivel descriptivo las diferencias en rendimiento académico de los estudiantes.

A través del análisis exploratorio, se determina que el rendimiento académico promedio de los estudiantes de bachillerato es insuficiente (689.78 puntos), lo que podría afectar sus posibilidades de ingresar a estudios superiores, dependiendo de la carrera y universidad seleccionada. Por

ejemplo, en general las instituciones públicas admiten a los estudiantes con puntajes más altos, mientras que las instituciones privadas ofrecen mayor accesibilidad, aunque requieren inversiones significativas.

Dado que el 90.85% de los estudiantes tiene la intención de continuar sus estudios, resulta esencial identificar los factores determinantes del rendimiento académico. Este conocimiento puede servir como base para que los responsables de las políticas educativas se enfoquen en aquellos aspectos que permitan ampliar las oportunidades de acceso a la educación superior.

4.2. Evaluación, predicción y clasificación del nivel de rendimiento

Evaluación y predicción del rendimiento académico

Con el objetivo de predecir el rendimiento académico, se emplearon los métodos detallados en la sección 3.3. y se evaluaron los modelos utilizando los indicadores propuestos en la sección 3.4. Tras llevar a cabo la validación cruzada, los resultados mostrados en la Tabla 4 revelan que el modelo Gradient Boosting supera a los demás en cuanto a predicción, lo cual se evidencia por sus menores valores de error en MAE, MSE, RMSE y su mayor coeficiente de determinación R2. Además, Gradient Boosting registra el menor error absoluto porcentual medio (MAPE), lo que indica que este modelo se adapta de forma más efectiva a los datos en comparación con los otros modelos evaluados. Resultados que se encuentran en línea con el estudio de Wang et al. (2022) quienes muestran que modelos derivados del método Gradient Boosting resultaron más efectivos para la predicción del rendimiento académico.

Por su parte, el modelo Random Forest ocupa el segundo lugar en términos de rendimiento, mientras que el modelo KNN muestra un desempeño inferior en todos los indicadores en comparación con los demás modelos, seguido por el Árbol de decisión. Estos hallazgos sugieren que los modelos KNN y Árbol de decisión no resultan los más apropiados para este conjunto de datos en particular.

Tabla 4. Promedio de los indicadores de evaluación según la técnica utilizada obtenidas luego de la validación cruzada

Indicador	Regresión múltiple	KNN	Árbol de decisión	Random Forest	Elastic NET	Gradient Boosting
MAE	17.47	18.15	17.85	17.48	17.47	17.21
MSE	662.65	684.65	667.73	644.53	662.65	640.8
RMSE	25.72	26.14	25.83	25.37	25.72	25.29
MAPE	2.56%	2.66%	2.61%	2.56%	2.56%	2.52%
R2	0.2041	0.1773	0.1968	0.2252	0.2041	0.2308

Elaboración: propia
Fuente: INEVAL (2022)

Según el valor de importancia relativa de cada variable, calculado a través del modelo Gradient Boosting (ver Gráfico 10), se destaca que el tipo de institución privada es el factor más relevante para el modelo, lo que implica que esta variable ejerce el mayor impacto en las predicciones. De acuerdo con las estimaciones del modelo, los estudiantes de instituciones privadas tienen un rendimiento promedio de 704.3 puntos, aproximadamente 20 puntos más que aquellos que estudian en otras instituciones (con un promedio de 683.88 puntos) (ver Gráfico 11).

Una posible explicación para que el tipo de institución privada sea el factor más determinante en el modelo de Gradient Boosting, según Benalcázar (2017), en el contexto ecuatoriano, podría estar relacionada con el nivel socioeconómico de los estudiantes. Esta interpretación es coherente con los resultados del presente estudio, en el cual la segunda y tercera características de mayor relevancia son el quintil socioeconómico del estudiante y la cantidad de días que utiliza el computador. El quintil socioeconómico puede influir en el acceso a recursos educativos adicionales como clases particulares, materiales didácticos y entornos favorables para el estudio, hallazgos que concuerdan con la investigación de Sirin (2005). Además, la frecuencia con la que un estudiante utiliza el computador puede ser un indicativo de su capacidad para acceder a recursos de aprendizaje en línea y desarrollar habilidades digitales, aspectos que cobran cada vez más importancia en la educación contemporánea.

Otras características, como sierra, urbano, fiscomisional, días, perdido y edad, presentan cierta relevancia en el modelo, aunque su importancia es menor en comparación con las tres primeras variables mencionadas. Las características restantes, como satisfacción, contento, estado, mestizo, horas, carrera, hombre, trabaja y faltas, poseen una importancia relativamente baja en el modelo, lo que significa que estas variables tienen un impacto menor en las predicciones en

comparación con las características más importantes. Es fundamental resaltar que la importancia de las características no implica necesariamente una relación causal. Únicamente indica cuánto contribuye cada característica a las predicciones del modelo en términos de la estructura interna de los árboles de decisión de Gradient Boosting.

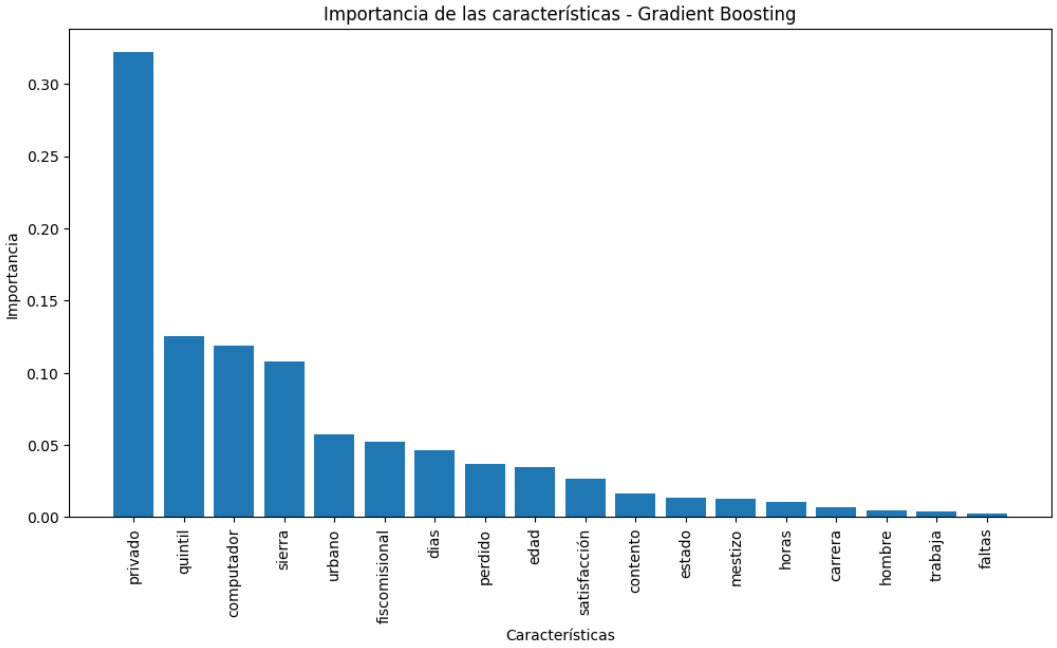


Gráfico 10. Importancia relativa de las variables de acuerdo al método Grandient Boosting
Fuente: INEVAL (2022)

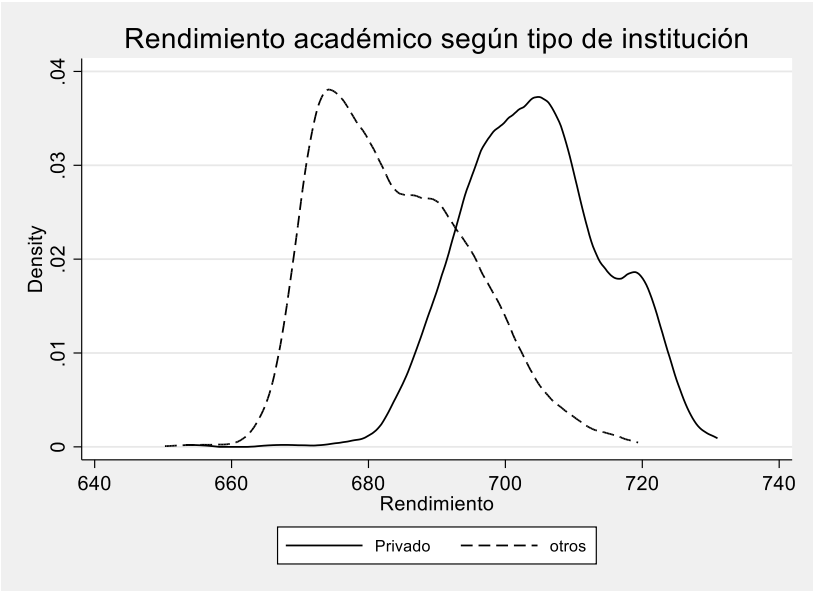


Gráfico 11. Rendimiento académico estimado por el modelo Grandient Boosting según tipo de institución
Fuente: INEVAL (2022)

Evaluación y clasificación del rendimiento académico

En esta sección, se analiza y compara la capacidad de clasificación de diversos modelos de aprendizaje automático para identificar estudiantes con rendimiento elemental o insuficiente. Los modelos evaluados incluyen el modelo Logit, KNN, Árbol de decisión, Random Forest, Naive Bayes, Gradient Boosting y MLP. Se examinan y discuten diversos indicadores de rendimiento, como la precisión, el área bajo la curva ROC (AUC), el recall y el F1-score, para determinar las fortalezas y debilidades de cada modelo en función de estos criterios.

De los resultados de la Tabla 5, se destaca que el modelo Gradient Boosting alcanza la mayor precisión y AUC, si bien presenta valores de recall y F1-score inferiores en comparación con otros modelos. Tanto el modelo Random Forest como el modelo Logit exhiben un desempeño sólido en términos de precisión, recall, F1-score, accuracy y AUC, situándose como los segundos mejores en la mayoría de estos indicadores.

Por otro lado, el modelo Naive Bayes muestra el rendimiento más bajo en cuanto a precisión y accuracy, aunque sobresale en recall y F1-score. Los modelos MLP y Árbol de decisión presentan resultados intermedios en la mayoría de los indicadores evaluados. No obstante, en el caso del modelo MLP, su AUC resulta ser el valor más bajo entre todos los modelos analizados.

Tabla 5. Promedio de los indicadores de evaluación según la técnica utilizada para la clasificación del tipo de rendimiento

Indicador	Logit	KNN	Árbol de decisión	Random Forest	Naive Bayes	Gradient Boosting	MLP
Precisión	0.6259	0.5754	0.613	0.628	0.5267	0.6613	0.6197
Recall	0.4317	0.3899	0.4606	0.4135	0.554	0.1742	0.4079
F1-score	0.511	0.4647	0.4631	0.4968	0.539	0.2741	0.4907
Accuracy	0.7504	0.7287	0.7384	0.7451	0.7135	0.7232	0.7449
AUC	0.7902	0.7569	0.767	0.792	0.7552	0.7782	0.6494

Elaboración: propia
Fuente: INEVAL (2022)

El modelo logit fue seleccionado para llevar a cabo la clasificación, debido a su rendimiento equilibrado en los indicadores evaluados. A pesar de tener un desempeño intermedio en cuanto a precisión, se destaca como uno de los mejores modelos en términos de accuracy y AUC, sin descuidar su desempeño en recall, por lo que se considera que tiene un rendimiento equilibrado en los indicadores analizados.

De acuerdo con este modelo, la probabilidad de lograr un rendimiento elemental varía de manera notable según el quintil socioeconómico del estudiante (ver Gráfico 12). Por ejemplo, en el quintil socioeconómico uno, menos del 1% de los estudiantes presenta una probabilidad del 50% o más de alcanzar un rendimiento elemental. Este porcentaje se incrementa conforme los estudiantes se encuentran en quintiles socioeconómicos más altos. Para aquellos ubicados en los quintiles dos y tres, la proporción de estudiantes con una probabilidad del 50% o más de obtener un rendimiento elemental es del 4% y 13%, respectivamente. En el caso de los quintiles cuatro y cinco, dicha proporción aumenta al 29% y 55%, respectivamente.

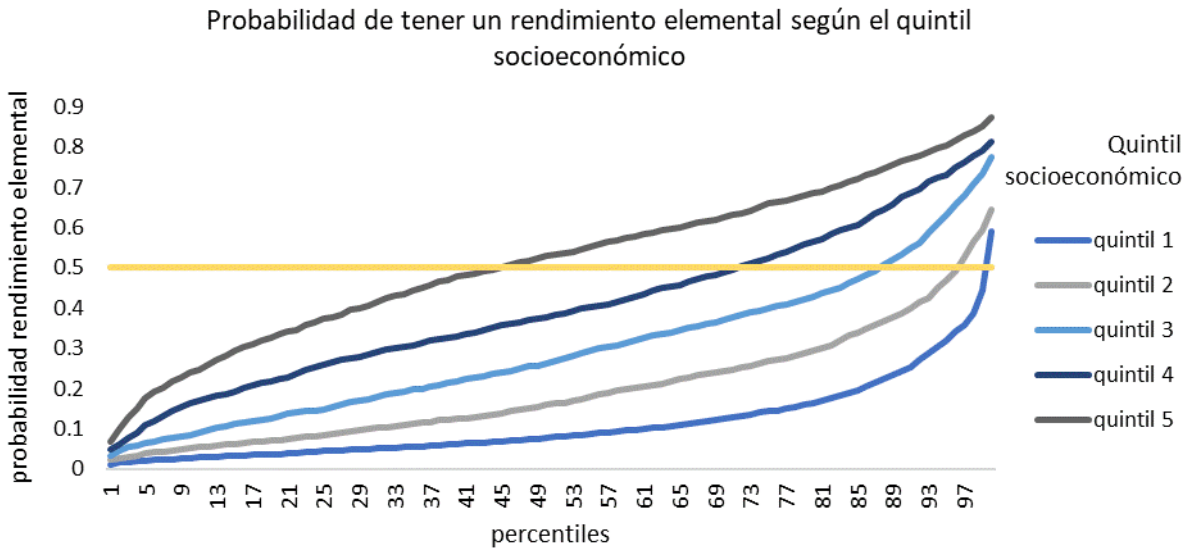


Gráfico 12. Probabilidad de obtener un rendimiento elemental según el quintil socioeconómico del estudiante
Fuente: INEVAL (2022)

4.3. Determinantes del rendimiento académico

Para analizar las relaciones entre las características y la variable objetivo, se emplean modelos de regresión múltiple y regresión logística con el objetivo de identificar los determinantes del rendimiento académico, medidos por la nota promedio y el tipo de rendimiento. Estos modelos ofrecen la ventaja de revelar las relaciones entre las variables. Al evaluar la significancia global de los modelos, se constata que, las variables explicativas en conjunto explican la nota promedio y el tipo de rendimiento con un nivel de significancia del 1% (consultar Tabla 6, Test F y Test Chi2).

Las variables explicativas se clasifican en tres grupos: características demográficas,

características educativas y percepción del estado emocional del estudiante. En general, los resultados muestran congruencia en la relación entre las características y el rendimiento académico en ambos modelos de regresión.

Factores socioeconómicos

En lo que respecta a las características demográficas, se encuentra que el rendimiento académico no está relacionado con el género del estudiante. Sin embargo, la edad sí tiene un efecto positivo y significativo en el rendimiento académico: por cada año adicional, el rendimiento académico se incrementa en 2 puntos ($p < 0.01$) manteniendo todo lo demás constante. La variable mestizo muestra un efecto positivo, aunque marginalmente significativo, en el rendimiento académico ($p < 0.1$). Los resultados indican que aquellos estudiantes que se autoidentifican como mestizos poseen un 25.5% de probabilidad de alcanzar un rendimiento elemental en comparación con los grupos que se autoidentifican de otra manera (ver Tabla 6, columna 3), ceteris paribus. Estos hallazgos son coherentes con el estudio de Sanchez (2011), quien demostró que los estudiantes con autoidentificación indígena presentan un rendimiento menor en comparación con los no indígenas, confirmando brechas en función de la etnia.

Los estudiantes de la región Sierra y aquellos ubicados en áreas urbanas también demuestran un rendimiento académico significativamente más alto en comparación con sus contrapartes ($p<0.01$). De esta manera, quienes estudian en la región Sierra y en áreas urbanas tienen una probabilidad 2.65 y 1.39 veces más de obtener una nota elemental que sus contrapartes, respectivamente (ver Tabla 6, columna 3).

Los hallazgos se encuentran en línea con el estudio de Muelle (2018) quien encontró que existen brechas regionales en el desempeño académico de los estudiantes. En el Ecuador existen una serie de diferencias regionales no solo socioeconómicas sino también en ámbitos de educación, así por ejemplo las tasas de deserción estudiantil son más altas en el oriente y la costa en comparación con la sierra (INEVAL, 2018), resultado que a largo plazo influye en el rendimiento académico.

Adicionalmente, se evidencia que el uso del computador para realizar tareas, así como, pertenecer a un quintil socioeconómico más alto, aumenta significativamente la posibilidad de obtener una calificación más alta ($p < 0.01$). En este contexto, quienes tienen un nivel socioeconómico mayor

disponen de recursos adicionales que les permite tener un mejor rendimiento (Sirin, 2005; Villa Lever et al., 2017).

No obstante, se espera que un estudiante que trabaje tenga, en promedio, un 17.3% menos de probabilidad de alcanzar al menos una nota elemental en comparación con aquellos que no trabajan. Este resultado concuerda con el análisis realizado por Carrillo & Ríos (2013), quienes proponen que una explicación posible a este fenómeno radica en la limitación de tiempo que enfrentan los estudiantes trabajadores para dedicar a sus tareas, estudiar o participar en actividades complementarias que enriquezcan su formación académica.

Características educacionales

En relación con las características educativas, se ha observado que los estudiantes que asisten a colegios privados o fiscomisionales presentan un rendimiento académico significativamente superior al de aquellos que cursan en colegios públicos ($p < 0.01$), siendo 4.88 y 1.66 veces más probable, respectivamente, que alcancen una calificación elemental en comparación con estudiantes de instituciones públicas (ver Tabla 6, columna 3). Si bien algunos investigadores, como Moreno Treviño & Cortez Soto (2020), atribuyen este efecto a la infraestructura y los recursos disponibles en los centros educativos con financiamiento privado, en el caso de Ecuador, la situación parece estar más vinculada con el nivel socioeconómico de los estudiantes. Benalcázar (2017) señala que en este país, estudiar en instituciones privadas se asocia con un mayor nivel socioeconómico, lo que, a su vez, afecta positivamente el rendimiento académico.

Por otro lado, un mayor número de días y horas de estudio está positivamente relacionado con el rendimiento académico ($p < 0.01$). Resultado que puede explicarse conforme a la teoría del aprendizaje basada en el tiempo y la práctica, según el cual, la cantidad de tiempo que un estudiante dedica al estudio y la práctica de una materia específica influye directamente en su capacidad para aprender y retener la información (Plant et al., 2005), fortaleciendo sus habilidades y comprensión del contenido, lo que, en última instancia, se refleja en un mejor rendimiento académico (Corno, 2000).

Por el contrario, los estudiantes que han perdido un año escolar presentan un rendimiento académico significativamente más bajo ($p < 0.01$), con un 57% menos de probabilidad de alcanzar

una calificación elemental en comparación con aquellos que no han perdido ningún año escolar. Este hallazgo puede atribuirse a diversos factores, como la desmotivación que afecta el compromiso y esfuerzo en el aprendizaje (Rumberger, 1995), así como la falta de apoyo adecuado durante el proceso de reincorporación escolar. Estos estudiantes pueden enfrentar barreras adicionales, como el estigma social y la ausencia de programas específicos diseñados para ayudarles a superar dificultades académicas y adaptarse a su nueva situación (Higgins & Simpson, 2011).

Los hallazgos indican que aquellos estudiantes que desean seguir una carrera tienen un 53.3% más de probabilidad de obtener una calificación elemental superior en comparación con aquellos que no desean continuar con sus estudios. Este factor puede considerarse como un indicador de motivación. Según Ryan & Deci (2000), la motivación desempeña un papel crucial en el aprendizaje y puede llevar a un mayor compromiso y esfuerzo en el ámbito académico. En consecuencia, esta mayor motivación se traduce en un rendimiento académico superior. Por último, no se encontró una relación significativa entre las faltas y el rendimiento.

Percepción del estado emocional del estudiante

En relación con la percepción del estado emocional del estudiante, aquellos que reportan mayor satisfacción con las personas con las que vive tienden a tener un rendimiento académico ligeramente más alto ($p < 0.01$) (ver Tabla 6 columna 1). Este resultado podría atribuirse al apoyo emocional y social que el estudiante recibe en su entorno familiar o de convivencia, como sugieren Wang & Eccles (2012). Un ambiente de apoyo y comprensión permite a los estudiantes enfrentar de manera más efectiva los desafíos asociados con la educación (Amato, 2001), lo que podría contribuir a un mejor rendimiento académico.

Sin embargo, los estudiantes que se sienten contentos con su situación actual no muestran una relación significativa con el rendimiento académico. Por último, los estudiantes que perciben que están contentos con el colegio tienen un rendimiento académico significativamente más bajo ($p<0.01$). Este hallazgo puede deberse a que estos estudiantes estén satisfechos con aspectos no académicos del colegio, como actividades extracurriculares, amistades o ambiente escolar, o a que no enfrentan suficientes desafíos académicos, lo cual les hace sentir cómodos pero no les motiva a mejorar su rendimiento (Dweck, 2006).

Tabla 6. Determinantes del rendimiento y de la probabilidad de obtener un rendimiento elemental

	Nota			Elemental				
Variable	**Coef**	**std err robustos**		**dy/dx**	**std err**		**odds ratio**	**std err robustos**	
	(1)			**(2)**			**(3)**		
constante	625.162	10.046	***				0.000	1.155	***
Hombre	0.021	0.862		0.015	0.015		1.082	0.082	
Edad	2.031	0.560	***	0.036	0.012	***	1.216	0.065	***
Mestizo	1.063	0.969		0.042	0.024	*	1.255	0.129	*
Sierra	9.203	0.877	***	0.180	0.017	***	2.651	0.094	***
Urbano	2.405	0.843	***	0.061	0.015	***	1.390	0.081	***
Trabaja	-1.719	0.877	**	-0.035	0.017	**	0.827	0.092	**
computador	0.849	0.190	***	0.013	0.003	***	1.071	0.018	***
Quintil	1.876	0.328	***	0.045	0.006	***	1.277	0.033	***
Características de educación									
Días	0.467	0.254	*	0.015	0.005	***	1.085	0.027	***
Horas	2.044	0.904	**	0.047	0.016	***	1.288	0.087	***
Privado	19.244	1.138	***	0.293	0.019	***	4.881	0.103	***
fiscomisional	7.575	0.982	***	0.094	0.018	***	1.662	0.100	***
Faltas	0.366	0.842		0.020	0.015		1.117	0.080	
Perdido	-8.158	1.196	***	-0.156	0.029	***	0.430	0.160	***
Carrera	3.180	1.227	**	0.079	0.031	**	1.533	0.165	***
Percepción de estado del estudiante									
satisfacción	1.133	0.433	***	0.011	0.008		1.061	0.043	
contento	-0.583	0.437		-0.011	0.008		0.943	0.042	
Estado	-2.862	0.865	***	-0.068	0.016	***	0.694	0.089	***
N. obs	4055			4055					
R2	0.212								
Pseudo R2				0.200					
Test F	59.74								
p-value test F	0.000								
Chi2				995.78					
p-value Chi2				0.000					

Nota: p-value<0.01***; <0.05**; <0.1*

Elaboración: propia
Fuente: INEVAL (2022)

4.4. Caracterización de los grupos de estudiantes

Para identificar clústeres de estudiantes con características y necesidades similares, se aplicaron tres algoritmos de aprendizaje no supervisado: K-means, DBSCAN y Agglomerative Clustering.

Los modelos se evaluaron mediante los coeficientes de silueta, Calinski-Harabasz y Davies-Bouldin para cada algoritmo (ver Tabla 7). K-means obtuvo el coeficiente de silueta más alto (0.5741), lo que indica que los clústeres formados por este algoritmo son más cohesivos y mejor separados que los generados por los otros dos algoritmos. Además, los coeficientes de Calinski-Harabasz y Davies-Bouldin también respaldan el mejor rendimiento de K-means en comparación con DBSCAN y Agglomerative Clustering. Debido a que K-means presenta los mejores coeficientes, se considera el algoritmo más adecuado para los datos entre los tres métodos de agrupamiento. En el Anexo 2, se muestra una representación gráfica de los clústeres generados por cada método, donde se observa que K-means logra una adecuada separación e identificación de los clústeres y DBSCAN presenta el peor rendimiento entre las tres técnicas aplicadas.

Una posible explicación del peor rendimiento de DBSCAN en este caso puede estar relacionada con la naturaleza del algoritmo y la estructura de los datos. DBSCAN, un algoritmo de clustering basado en densidad, podría tener dificultades para separar adecuadamente los grupos si los datos no presentan clústeres con densidades notablemente diferentes y si los parámetros de radio y la cantidad mínima de puntos no se ajustan correctamente (Ester et al., 1996). Por otro lado, K-means, que se basa en la minimización de las distancias dentro de los clústeres (MacQueen, 1967), puede ser más adecuado para este conjunto de datos si los clústeres presentan una estructura más uniforme y están separados por distancias relativamente similares.

Tabla 7. Indicadores de evaluación de las técnicas de clustering

Indicador	k-means	DBSCAN	Agglomerative Clustering
Silhouette score	0.5741	0.4860	0.5371
Calinski-Harabasz score	6321.9466	98.9296	5843.5711
Davies-Bouldin score	0.5452	2.8890	0.5746

Elaboración: propia
Fuente: INEVAL (2022)

De acuerdo con la técnica K-means, se generaron 3 clústeres. Los Gráficos 13 y 14 ofrecen un resumen de las características de cada clúster. El primer clúster comprende el 46.4% de los estudiantes, mientras que el segundo y tercer clúster representan el 24.4% y 29.2%, respectivamente. A continuación, se presenta un resumen de los perfiles de cada clúster, basado en los hallazgos obtenidos:

Clúster 1: Este grupo está compuesto principalmente por estudiantes que muestran un rendimiento académico más alto, ya que poseen una mayor proporción de estudiantes con una alta probabilidad de obtener calificaciones elementales, lo que resulta en un promedio de notas más elevado. El clúster presenta una proporción equilibrada entre hombres y mujeres, con una mayor proporción de estudiantes que se autoidentifican como mestizos y viven predominantemente en la región Sierra en comparación con los otros clústeres. Además, estos estudiantes dedican más tiempo al estudio, tanto en horas como en el uso del computador, en comparación con los otros dos grupos. Este clúster se caracteriza por tener una menor proporción de estudiantes que trabajan y una mayor proporción que estudian en institutos privados. También se distingue por contar con una mayor proporción de estudiantes en los quintiles socioeconómicos más altos y una menor proporción que han perdido un año académico. En general, este clúster representa a estudiantes con un perfil académico más sólido y con mayores recursos socioeconómicos disponibles.

Clúster 2: El segundo grupo está compuesto por estudiantes con un rendimiento académico moderado, con una proporción de estudiantes con promedios de notas similares al clúster 3, pero inferiores al clúster 1. Los estudiantes de este clúster dedican menos tiempo al estudio en comparación con los otros grupos y cuentan con una mayor proporción de estudiantes que trabajan. Asimismo, presentan una mayor proporción de estudiantes en los quintiles socioeconómicos más bajos y una mayor cantidad de faltas y estudiantes que han perdido un año académico. Además, estos estudiantes se caracterizan por autoidentificarse como mestizos en menor proporción que el clúster uno y estudian en regiones de la Sierra como el Oriente y la Costa. Una menor proporción de estudiantes de este clúster se siente satisfecho con las personas con las que vive en comparación con el resto de clústeres. En resumen, este clúster describe a estudiantes con un perfil académico más modesto y enfrentando mayores desafíos socioeconómicos, de educación y familiar.

Clúster 3: Este grupo está compuesto principalmente por estudiantes con un rendimiento académico similar al del clúster 2, pero con una dedicación al estudio más alta y una proporción ligeramente menor de estudiantes que han perdido un año académico. Además, tiene una proporción de estudiantes que trabaja relativamente similar al grupo dos. Aunque enfrentan desafíos socioeconómicos mayores al clúster 2, estos estudiantes parecen tener un enfoque más claro en sus objetivos de carrera y un nivel de satisfacción con las personas con las que viven similar al clúster uno. Este grupo está caracterizado por estar conformado por una mayor

proporción de mujeres, y provienen de otras regiones en comparación a los otros dos clústeres. En general, este clúster representa a estudiantes con un perfil académico con mayores problemas socioeconómicos, pero con una mayor dedicación al estudio y un enfoque más definido en sus metas profesionales.

De acuerdo con los hallazgos, es necesario desarrollar estrategias específicas para cada clúster con el objetivo de incrementar su rendimiento académico y facilitar una transición exitosa hacia la educación superior. Es importante tener en cuenta que, aunque el primer grupo tenga un rendimiento elemental (700 a 800 puntos), este grupo debe esforzarse aún más para alcanzar un rendimiento satisfactorio (800-950 puntos). La Tabla 8 resume algunas estrategias que se podrían implementar para cada grupo con el fin último de mejorar el rendimiento académico en cada grupo, de forma que sus probabilidades de ingresar a una institución de educación superior incrementen.

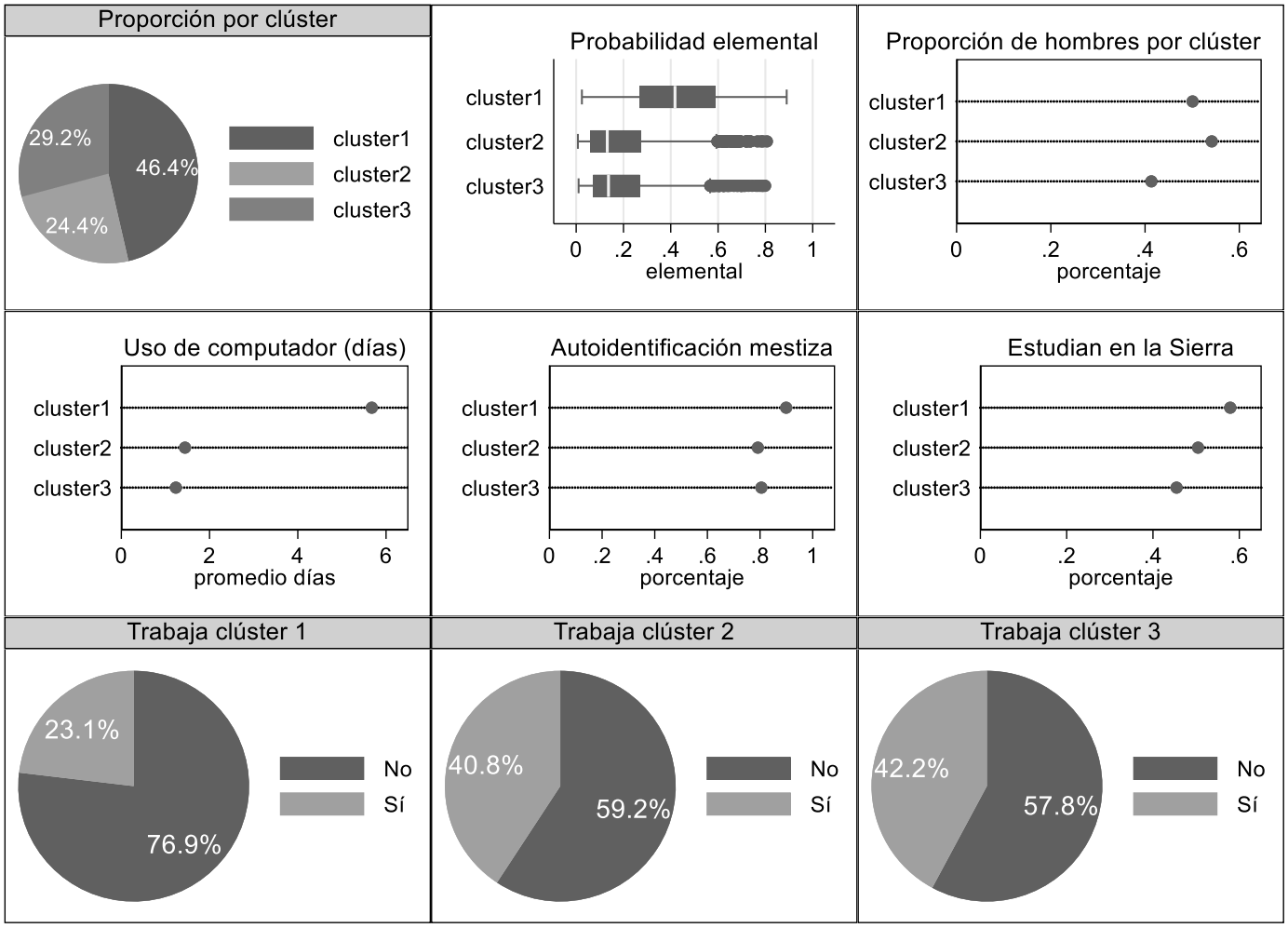


Gráfico 13. Características de los clústeres según las variables socioeconómicas del estudiante
Fuente: INEVAL (2022)

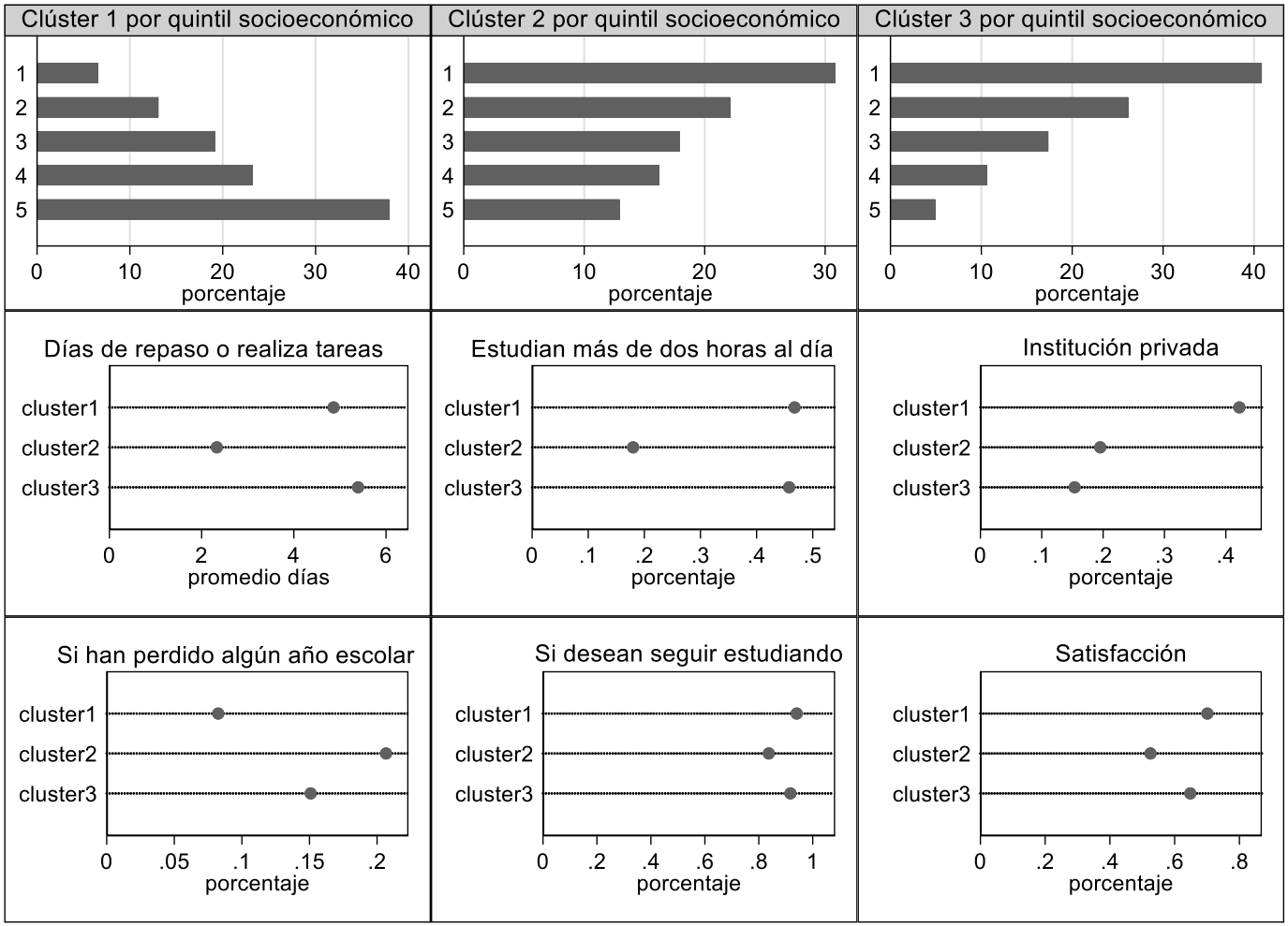


Gráfico 14. Características de los clústeres según el quintil socioeconómico, variables de educación y percepción del estado emocional del estudiante
Fuente: INEVAL (2022)

Tabla 8. Estrategias para mejorar el rendimiento en cada clúster

Clúster 1	**Clúster 2**	**Clúster 3**
Fomentar el desarrollo de habilidades blandas: Aunque los estudiantes de este clúster tienen un buen rendimiento académico, es importante trabajar en habilidades como la comunicación, el trabajo en equipo y la resiliencia para enfrentar futuros desafíos tanto en la educación superior como en el ámbito laboral (Heckman & Kautz, 2012) Promover la tutoría y el apoyo entre pares: Los estudiantes de este grupo pueden beneficiarse de compartir sus conocimientos y habilidades con estudiantes de otros clústeres, promoviendo una cultura de colaboración y aprendizaje mutuo (Topping, 2005). Estimular la participación en actividades extracurriculares: Para un desarrollo integral, se recomienda incentivar a estos estudiantes a participar en actividades de liderazgo que complementen su formación académica (Fredricks & Eccles, 2006).	Implementar programas de refuerzo académico: Es fundamental diseñar e implementar programas de apoyo y refuerzo en las áreas donde estos estudiantes presentan dificultades, con el objetivo de mejorar sus habilidades y conocimientos (Fuchs & Fuchs, 2006). Ofrecer orientación vocacional y acompañamiento: Para ayudar a estos estudiantes a establecer metas y objetivos claros, se recomienda dar orientación vocacional y acompañamiento en su proceso educativo (Gati & Saka, 2001). Facilitar el acceso a recursos y becas: Dado que los estudiantes de este clúster enfrentan mayores desafíos de tipo socioeconómico, es crucial identificar oportunidades de becas y recursos para reducir las barreras económicas y facilitar su acceso y permanencia en la educación superior (Bettinger, 2004). Promover cursos adicionales: para aquellos estudiantes que han perdido algún año escolar, se debe impartir cursos adicionales de acuerdo a las necesidades de este grupo de estudiantes para fortalecer algún posible vacío existente (Jimerson et al., 2002). Fomentar la motivación y el compromiso con el aprendizaje: Es importante trabajar con estos estudiantes en su actitud y compromiso hacia el	Facilitar acceso a recursos: este grupo al igual que el grupo dos requiere de becas y recursos digitales que le permita facilitar el aprendizaje (Bettinger, 2004). Promover la colaboración entre docentes y estudiantes: La creación de espacios de colaboración entre docentes y estudiantes puede mejorar el rendimiento académico de este clúster, al permitir un aprendizaje más dinámico y personalizado (Cornelius-White, 2007). Impartir técnicas de estudio: este grupo presenta dedicación para realizar tareas en un número de días similar al clúster uno, pero no logran obtener un rendimiento similar, por lo que se requiere apoyo en técnicas de estudio para que los días dedicados sean efectivos (Weinstein et al., 1986). Establecer estrategias de estudio: considerando que parte de este clúster trabaja, se puede generar opciones de estudio para equilibrar sus responsabilidades laborales y académicas por ejemplo grupos de estudio y trabajo en equipo. Además se podría ofrecer capacitaciones extras en días flexibles al horario escolar. Considerando que este grupo también se compone de una proporción de estudiantes que han perdido algún año escolar requieren de capacitaciones para nivelar los conocimientos que sientan vacíos y dar

	aprendizaje, incentivándolos a participar activamente en sus clases y a desarrollar habilidades de estudio efectivas (Zimmerman & Schunk, 2011) Establecer programas de mentoría y tutoría: puede ayudar a estos estudiantes a recibir apoyo académico y emocional personalizado, lo que les permitirá enfrentar mejor sus desafíos (Crisp & Cruz, 2009).	acompañamiento para evitar la desmotivación.

Elaboración: propia basada en los estudios de Bettinger (2004); Cornelius-White (2007); Crisp & Cruz (2009); Fredricks & Eccles (2006); Fuchs & Fuchs (2006); Gati & Saka (2001); Heckman & Kautz (2012); Jimerson et al. (2002); Topping (2005); Weinstein et al. (1986); Zimmerman & Schunk (2011).

5. Conclusión, recomendaciones y trabajos futuros

El estudio realizado analizó el rendimiento académico de los estudiantes de bachillerato en Ecuador durante el ciclo académico 2021-2022, utilizando el método CRISP-DM y datos de INEVAL. Conforme a las preguntas de investigación realizadas, los hallazgos del estudio permiten entender cuáles son las características de los estudiantes que influyen en el rendimiento académico para identificar los factores sobre los cuales se podría reforzar desde las instituciones de educación secundaria, así como establecer grupos de estudiantes con características similares para proporcionar estrategias educativas para cada grupo.

Respecto a la primera pregunta de investigación planteada, los hallazgos del estudio destacan la importancia de factores socioeconómicos, geográficos, educativos y de percepción en el desempeño académico de los estudiantes. Entre los factores más relevantes se encuentran el nivel socioeconómico, el tipo de institución, si el estudiante trabaja, la cantidad de horas y días dedicados a las tareas, la pérdida de años escolares y la motivación a continuar estudiando luego de culminar el bachillerato. Además, se determinó que la satisfacción del estudiante con su entorno familiar también incide en su desempeño académico. En relación con la segunda

pregunta de investigación, se identificaron tres grupos distintos de estudiantes, cada uno con necesidades específicas y propuestas de estrategias de mejora adaptadas a sus características.

A partir de estos hallazgos, se genera las siguientes recomendaciones. En primer lugar, se debe diseñar programas de apoyo académico y socioemocional dirigidos a estudiantes identificados como en situación de vulnerabilidad o riesgo académico con el fin de dar acompañamiento y que puedan mejorar su rendimiento y así aumentara sus posibilidades de acceso a la educación superior. En segundo lugar, generar colaboración entre instituciones educativas, autoridades tanto de educación secundaria como universitaria, por ejemplo organizar visitas guiadas a campus universitarios y talleres informativos para familiarizar a los estudiantes con la vida universitaria y las expectativas académicas. De forma que, permita mejorar la conexión entre la educación secundaria y superior, asegurando una transición exitosa y abordando las brechas en la preparación de los estudiantes.

En tercer lugar, se recomienda generar estrategias en los que se involucre a la comunidad educativa, incluidos padres y madres de familia, en el diseño y ejecución de estrategias de mejora del rendimiento académico, fomentando la participación y corresponsabilidad en el proceso educativo. En cuarto lugar, se recomienda que se pueda introducir actividades, programas correctivas o intervenciones preventivas para aquellas instituciones en los que el promedio del rendimiento se encuentre por debajo del promedio nacional. Por último, con el fin de cerrar las brechas de rendimiento se recomienda que se realice capacitaciones o fomento al uso de los avances tecnológicos y de las herramientas digitales especialmente para aquellos grupos con mayores brechas en rendimiento como colegios públicos, niveles o estratos socioeconómicos más bajos.

Dados los hallazgos, el estudio contribuye a la literatura empírica sobre el rendimiento académico en estudiantes de bachillerato y da a conocer los determinantes del mismo, además permitió identificar grupos de estudiantes con características similares evidenciando la necesidad de aplicar estrategias diferenciadas para cada grupo. Las Instituciones educativas y responsables políticos pueden utilizar estos hallazgos para desarrollar intervenciones y estrategias personalizadas que mejoren el acceso y el éxito en la educación superior. Por ejemplo, pueden diseñar programas de tutoría y apoyo académico dirigidos a estudiantes con

necesidades específicas, adaptar los planes de estudio y métodos de enseñanza para abordar las áreas identificadas como problemáticas, y establecer alianzas con organizaciones comunitarias y gubernamentales para proporcionar recursos adicionales a los estudiantes en situación de vulnerabilidad. Como resultado, el estudio contribuye a la toma de decisiones en el ámbito educativo.

Investigaciones futuras pueden abordar el impacto de las tecnologías de la información dentro de la educación secundaria y cómo éstas pueden mejorar o no el rendimiento académico. Además sería beneficioso realizar investigaciones donde se pueda evaluar el impacto de diferentes estrategias de pedagogías implementadas en la educación secundaria y cómo estas a su vez influyen en el rendimiento. También sería interesante analizar variables de éxito de la educación secundaria de más largo plazo como por ejemplo, las tasas de inserción y deserción universitaria conforme al rendimiento que han obtenido en la educación secundaria.

Los métodos de aprendizaje automático aplicados en este estudio han demostrado ser herramientas valiosas para analizar y predecir el rendimiento académico de los estudiantes de bachillerato en Ecuador. Cada modelo implementado tiene sus ventajas y limitaciones, y la elección del método más adecuado dependerá del contexto y los objetivos específicos de cada investigación. En este estudio, una vez seleccionado los hiperparámetros que optimizan los indicadores de evaluación, se encontró que el modelo Gradient Boosting mostró mejores resultados de predicción y el modelo Logit proveyó de resultados equilibrados para la clasificación. Mientras que para la agrupación se utilizó K means por su mejor rendimiento con respecto al resto de técnicas.

Para trabajos futuros en este campo, se recomienda explorar la aplicación de otros métodos de aprendizaje automático, como las redes neuronales convolucionales y recurrentes, para abordar problemas más complejos y específicos relacionados con el rendimiento académico. Además, sería beneficioso realizar investigaciones longitudinales para evaluar el impacto de las intervenciones y políticas educativas en el rendimiento académico a lo largo del tiempo.

6. Referencias

Adewale, A. M., Bamidele, A. O., & Lateef, U. O. (2018). Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach. *International Journal of Science and Technology Education Research*, *9*(1), 1–8.

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, *9*(129), 6415–6426.

Ahmad, M. S., Asad, A. H., & Mohammed, A. (2021). A Machine Learning Based Approach for Student Performance Evaluation in Educational Data Mining. *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 187–192.

Amato, P. R. (2001). Children of divorce in the 1990s: an update of the Amato and Keith (1991) meta-analysis. *Journal of Family Psychology*, *15*(3), 355.

Astin, A. W. (1997). *What matters in college?* JB.

Austria-Carlos, M. A., Venegas-Martínez, F., & Pérez Lechuga, G. (2018). Diferencias por género en la tasa de ganancia salarial de la educación superior y posgrado en México. *Papeles de Población*, *24*(96), 157–186.

Benalcázar, M. (2017). Educación privada versus educación pública en el Ecuador. *Revista Publicando*, *4*(11(1)), 484–498. https://revistapublicando.org/revista/index.php/crv/article/view/577

Bettinger, E. (2004). How financial aid affects persistence. In *College choices: The economics of where to go, when to go, and how to pay for it* (pp. 207–238). University of Chicago Press.

Brahim, G. Ben. (2022). Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*, *47*(8), 10225–10243.

Bressane, A., Spalding, M., Zwirn, D., Loureiro, A. I. S., Bankole, A. O., Negri, R. G., de Brito Junior, I., Formiga, J. K. S., Medeiros, L. C. de C., & Pampuch Bortolozo, L. A. (2022). Fuzzy artificial intelligence—based model proposal to forecast student performance and retention risk in engineering education: An alternative for handling with small data. *Sustainability*, *14*(21), 14071.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, *3*(1), 1–27.

Carrillo Regalado, S., & Ríos Almodóvar, J. G. (2013). Trabajo y rendimiento escolar de los estudiantes universitarios. El caso de la Universidad de Guadalajara, México. *Revista de La Educación Superior*, *42*(166), 9–34.

Chigbu, B. I., & Nekhwevha, F. H. (2021). High school training outcome and academic performance of first-year tertiary institution learners-Taking'Input-Environment-Outcomes model'into account. *Heliyon*, *7*(7), e07700.

Chollet, F., & Allaire, J. J. (2018). Deep Learning with R Manning Publications Co. *Shelter Island, NY*.

Conley, D. T. (2007). Redefining college readiness. *Educational Policy Improvement Center (NJ1)*.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, *77*(1), 113–143.

Corno, L. (2000). Looking at homework differently. *The Elementary School Journal*, *100*(5), 529–548.
Crisp, G., & Cruz, I. (2009). Mentoring college students: A critical review of the literature between 1990 and 2007. *Research in Higher Education*, *50*, 525–545.
Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*, 224–227.
Dubuc, M.-M., Aubertin-Leheudre, M., & Karelis, A. D. (2020). Lifestyle habits predict academic performance in high school students: the adolescent student academic performance longitudinal study (ASAP). *International Journal of Environmental Research and Public Health*, *17*(1), 243.
Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random house.
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*(34), 226–231.
Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance--A Critical Literature Review.* ERIC.
Fredricks, J. A., & Eccles, J. S. (2006). Is extracurricular participation associated with beneficial outcomes? Concurrent and longitudinal relations. *Developmental Psychology*, *42*(4), 698.
Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*(1), 93–99.
Gati, I., & Saka, N. (2001). High school students' career-related decision-making difficulties. *Journal of Counseling & Development*, *79*(3), 331–340.
Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."
Gershenson, S. (2013). The causal effect of commute time on labor supply: Evidence from a natural experiment involving substitute teachers. *Transportation Research Part A: Policy and Practice*, *54*, 127–140.
Greene, W. H. (2003). Econometric analysis, 4th edn Prentice-Hall. *Upper Saddle*.
Han, C., Farruggia, S. P., & Solomon, B. J. (2022). Effects of high school students' noncognitive factors on their success at college. *Studies in Higher Education*, *47*(3), 572–586.
Hanke, J. E., & Wichern, D. W. (2009). Business forecasting 9th ed. *New Jersey*.
Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, *4*(1), 131–157.
Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464.
Higgins, S., & Simpson, A. (2011). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. By John AC Hattie: Pp 392. London: Routledge. 2008.£ 90 (hbk),£ 27.99 (pbk),£ 35.37 (e-book). ISBN-13 978-0415476171 (hbk), ISBN-13 978-0415476188 (pbk), ASIN: B001OLRMHS (e-book).* Taylor & Francis.
INEVAL. (2018). *La educación en Ecuador, logros alcanzados y nuevos desafíos*. https://www.evaluacion.gob.ec/wp-content/uploads/downloads/2019/02/CIE_ResultadosEducativos18_20190109.pdf

INEVAL. (2020). *Informe de resultados Evaluación costa 2019-2020*. https://www.evaluacion.gob.ec/wp-content/uploads/downloads/2020/06/24.1.-DAGI_SBAC20-InformeCosta2019-2020_20200618.pdf

INEVAL. (2021). *Rendición de cuentas*. https://www.evaluacion.gob.ec/wp-content/uploads/downloads/2022/03/INEVAL_informePreliminar_rendicion_de_cuentas_2021.pdf

INEVAL. (2022). *Banco de Información*. http://evaluaciones.evaluacion.gob.ec/BI/nacional/

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jawad, K., Shah, M. A., & Tahir, M. (2022). Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing. *Sustainability*, *14*(22), 14795.

Jimerson, S. R., Anderson, G. E., & Whipple, A. D. (2002). Winning the battle and losing the war: Examining the relation between grade retention and dropping out of high school. *Psychology in the Schools*, *39*(4), 441–457.

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, *11*(7), 3130.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms. *Worked Examples, and Case Studies*.

Kirst, M., & Venezia, A. (2004). *From high school to college: Improving opportunities for success*. San Francisco: Jossey-Bass.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

Kuh, G. D., Kinzie, J. L., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What matters to student success: A review of the literature* (Vol. 8). National Postsecondary Education Cooperative Washington, DC.

Li, S., & Liu, T. (2021). Performance prediction for higher education students using deep learning. *Complexity*, *2021*, 1–10.

Lucero, M. F. (2019). Rendimiento de la educación en Ecuador. *Estudios de La Gestión: Revista Internacional de Administración*, *6*, 37–90.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

Maimon, O., & Rokach, L. (2005). Data Mining and Knowledge Discovery Handbook . In *Springer*. https://books.google.com.ec/books?hl=en&lr=&id=jizrAIWUJ6UC&oi=fnd&pg=PA321&dq=Rokach,+L.,+%26+Maimon,+O.+(2005).+Clustering+methods.+In+Data+mining+and+knowledge+discovery+handbook+(pp.+321-352).+Springer,+Boston,+MA.&ots=mSmP5C8sbH&sig=d12wTda6ZIvtP8JdnhSBlR55rHo&redir_esc=y#v=onepage&q&f=false

Martínez, A. C. (2019). La teoría del capital humano, fundamento del programa Beca 18. *Investigaciones Sociales*, *22*(40), 319–332.

Masud, S., Mufarrih, S. H., Qureshi, N. Q., Khan, F., Khan, S., & Khan, M. N. (2019). Academic performance in adolescent students: the role of parenting styles and socio-

demographic factors–a cross sectional study from peshawar, Pakistan. *Frontiers in Psychology*, *10*, 2497.

Moreno Treviño, J. O., & Cortez Soto, S. N. (2020). Rendimiento académico y habilidades de estudiantes en escuelas públicas y privadas: evidencia de los determinantes de las brechas en aprendizaje para México. *Revista de Economía*, *37*(95), 73–106.

Muelle, L. (2018). Desigualdades regionales y sociales del rendimiento escolar al término de la educación primaria en el Perú. *Revista Peruana de Investigación Educativa*, *10*(10), 127–157.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, *31*, 274–295.

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, 389–394.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237.

Ñiquen Lasteros, O. (2019). El impacto del nivel educativo alcanzado en el índice de calidad del empleo en el Perú, 2016. *Revista Peruana de Investigación Educativa*, *11*(11), 5–38. https://doi.org/10.34236/RPIE.V11I11.91

Oreopoulos, P., & Salvanes, K. G. (2011). Priceless: The Nonpecuniary Benefits of Schooling. *Journal of Economic Perspectives*, *25*(1), 159–184. https://doi.org/10.1257/JEP.25.1.159

Park, H., Byun, S., & Kim, K. (2011). Parental involvement and students' cognitive outcomes in Korea: Focusing on private tutoring. *Sociology of Education*, *84*(1), 3–22.

Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research. Volume 2*. ERIC.

Pérez, A. B. D., Quispe, F. M. P., Aguilar, O. A. G., & Cortez, L. C. C. (2020). Transición secundaria-universidad y la adaptación a la vida universitaria. *Revista de Ciencias Sociales*, *26*(3), 244–258.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, *95*(4), 667.

Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, *30*(1), 96–116.

Quintero Montaño, W. J. (2020). La formación en la teoría del capital humano: una crítica sobre el problema de agregación. *Análisis Económico*, *35*(88), 239–265.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, *10*(3), 1042.

Reardon, S. F., & Owens, A. (2014). 60 years after Brown: Trends and consequences of school segregation. *Annual Review of Sociology*, *40*, 199–218.

Rodríguez Rosero, D. D., Ordoñez Ortega, R. E., & Hidalgo Villota, M. E. (2021). Academic Performance Determinants of High School Students in the Department of Nariño, Colombia. *Lecturas de Economia*, *94*, 87–126.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.
Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*(3), 583–625.
Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67.
Sanchez-Jabba, A. (2011). Etnia y rendimiento académico en Colombia. *Revista de Economía Del Rosario*, *14*(2), 189–227.
Sandoval Vega, A. E. (2022). *Meritocracia y desigualdad de oportunidades. Análisis del rendimiento en el examen de acceso al Sistema de Educación Superior Ecuatoriano para el período 2018-2019.* http://bibdigital.epn.edu.ec/handle/15000/22868
Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1–21.
Schunk, D. H., & Greene, J. A. (2017). Historical, contemporary, and future perspectives on self-regulated learning and performance. In *Handbook of self-regulation of learning and performance* (pp. 1–15). Routledge.
SENESCYT. (2021). *Sistema Ecuatoriano de Acceso a la Educación Superior Octubre 2021 ESTRUCTURA GENERAL PARA LA PRESENTACIÓN DE PROGRAMAS Y PROYECTOS DE INVERSIÓN*. https://www.educacionsuperior.gob.ec/wp-content/uploads/2023/02/PROYECTO_SEAES.pdf
SENESCYT. (2022). *PLAN ESTRATÉGICO INSTITUCIONAL 2021 - 2025* . https://www.educacionsuperior.gob.ec/wp-content/uploads/2022/03/Plan-Estrate%CC%81gico-Institucional-2021-2025-Senescyt.pdf
Silva, P. L., Nunes, L. C., Seabra, C., Balcao Reis, A., & Alves, M. (2020). Student selection and performance in higher education: admission exams vs. high school scores. *Education Economics*, *28*(5), 437–454.
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.
Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to data mining Pearson Education India*. Indian Nursing Council New Delhi, India.
Theobald, M. (2021). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology*, *66*, 101976.
Tinto, V. (2012). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago press.
Topping, K. J. (2005). Trends in peer learning. *Educational Psychology*, *25*(6), 631–645.
Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179.
Venezia, A., & Jaeger, L. (2013). Transitions from high school to college. *The Future of Children*, 117–136.
Vieira, C., Vieira, I., & Raposo, L. (2018). Distance and academic performance in higher education. *Spatial Economic Analysis*, *13*(1), 60–79.

Villa Lever, L., Canales Sánchez, A., & Hamui Sutton, M. (2017). Expresiones de las desigualdades sociales en espacios universitarios asimétricos. *Universidad Nacional Autónoma de México Instituto de Investigaciones Sociales Consejo Nacional de Ciencia y Tecnología (Conacyt)*. http://ru.iis.sociales.unam.mx:8080/jspui/bitstream/IIS/5290/1/expresiones_desigualdades.pdf#page=55

Villarruel-Meythaler, R. E., Tapia-Morales, K. I., & Cárdenas-García, J. K. (2020). Determinantes del rendimiento académico de la educación media en Ecuador. *Revista Economía y Política*, *32*, 212–234.

Wang, C., Chang, L., & Liu, T. (2022). Predicting student performance in online learning using a highly efficient gradient boosting decision tree. *Intelligent Information Processing XI: 12th IFIP TC 12 International Conference, IIP 2022, Qingdao, China, May 27–30, 2022, Proceedings*, 508–521.

Wang, M., & Degol, J. (2014). Staying engaged: Knowledge and research needs in student engagement. *Child Development Perspectives*, *8*(3), 137–143.

Wang, M., & Eccles, J. S. (2012). Social support matters: Longitudinal effects of social support on three dimensions of school engagement from middle to high school. *Child Development*, *83*(3), 877–895.

Wang, Z. (2022). Higher Education Management and Student Achievement Assessment Method Based on Clustering Algorithm. *Computational Intelligence and Neuroscience*, *2022*.

Weinstein, C. E., Mayer, R. E., & Wittrock, M. C. (1986). Handbook of research on teaching. *Handbook of Research on Teaching*.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, *4*(5), e12482.

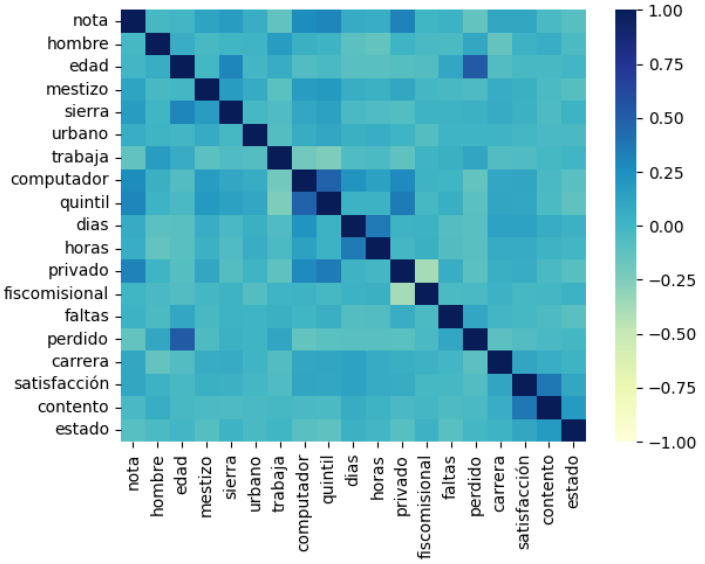
York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation*, *20*(1), 5.

Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of self-regulation of learning and performance.* Routledge/Taylor & Francis Group.

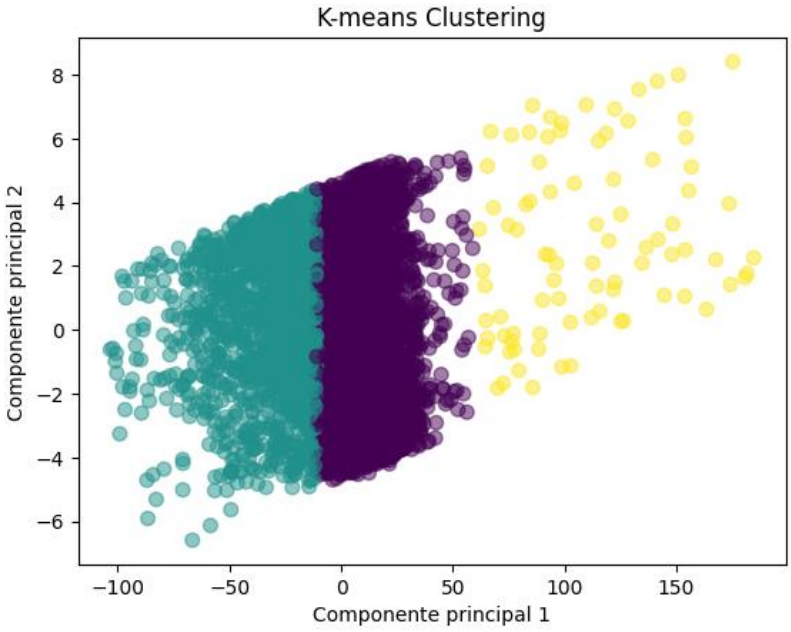
Zins, J. E., Bloodworth, M. R., Weissberg, R. P., & Walberg, H. J. (2007). The scientific base linking social and emotional learning to school success. *Journal of Educational and Psychological Consultation, 17*(2–3), 191–210.

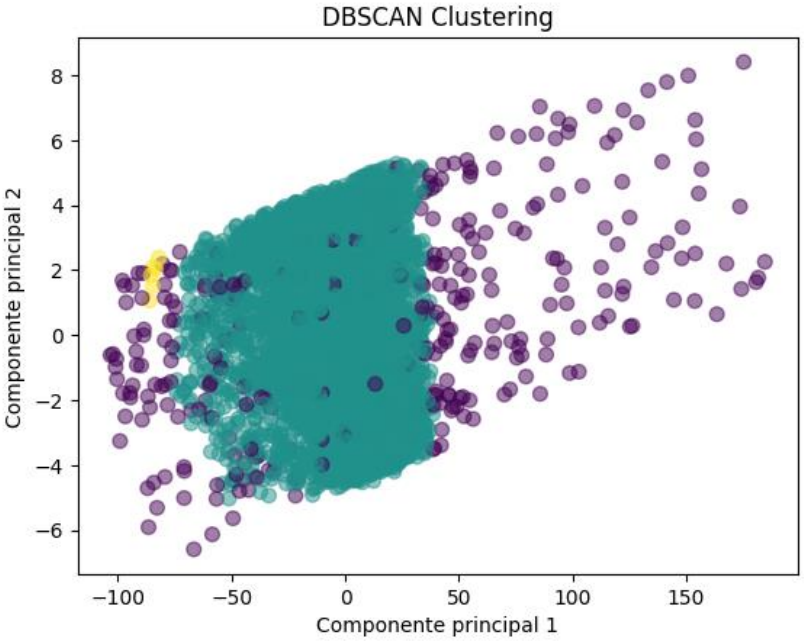
7. Anexos

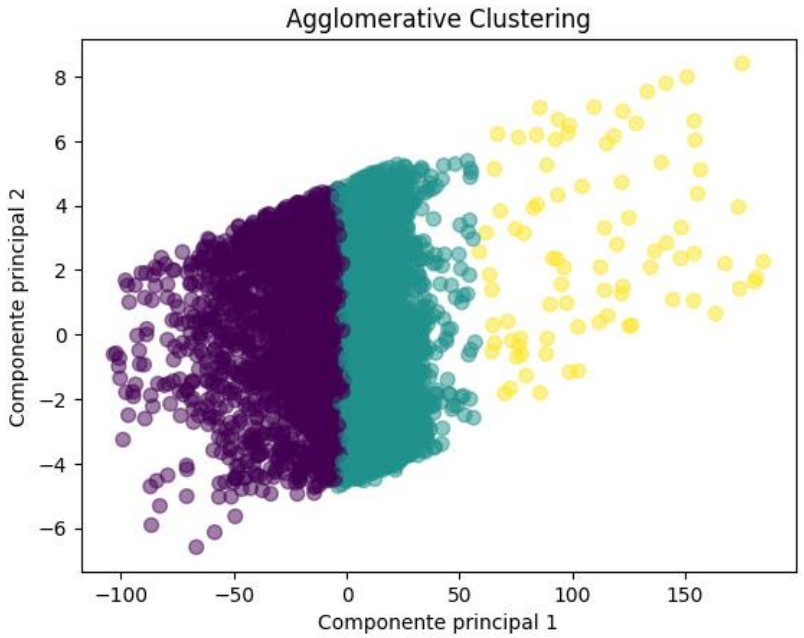
Anexo 1. Mapa de calor de la correlación de Pearson entre las variables



Anexo 2. Evaluación gráfica de los clústeres generados por K-means, DBSCAN y Agglomerative Clustering




DBSCAN Clustering
Componente principal 2
Componente principal 1
8
6
4
2
0
−2
−4
−6
−100
−50
0
50
100
150


Agglomerative Clustering
Componente principal 2
Componente principal 1
8
6
4
2
0
−2
−4
−6
−100
−50
0
50
100
150