Fecha de presentación: julio, 2024, **Fecha de Aceptación**: agosto, 2024, **Fecha de publicación**: septiembre, 2024

51

TÉCNICAS DE SOFT COMPUTING PARA DETERMINAR LOS FACTORES QUE INFLUYEN EN LA DESERCIÓN ESTUDIANTIL

SOFT COMPUTING TECHNIQUES TO DETERMINE THE FACTORS THAT INFLUENCE STUDENT DROPOUT

Byron Oviedo-Bayas[1*]
E-mail: boviedo@uteq.edu.ec
ORCID: https://orcid.org/ 0000-0002-5366-5917
Joseph Espinoza-Astudillo[1]
E-mail: espinozaa5@uteq.edu.ec
ORCID: https://orcid.org/0009-0009-5945-4393
Efraín Díaz-Macías[1]
E-mail: efraindiaz@uteq.edu.ec
ORCID: https://orcid.org/0000-0003-4087-029X
Jorge Guanín-Fajardo[1]
E-mail: jorgeguanin@uteq.edu.ec
ORCID: https://orcid.org/0000-0001-9150-4009
[1] Universidad Técnica Estatal de Quevedo. Ecuador.
* Autor para correspondencia

Cita sugerida (APA, séptima edición)

Oviedo–Bayas, B., Espinoza-Astudillo, J., Díaz-Macías, E., y. Guanín-Fajardo, J. (2024). Técnicas de soft computing para determinar los factores que influyen en la deserción estudiantil. ***Revista Conrado, 20***(100). 449-465.

RESUMEN

El estudio realizado en la Universidad Técnica Estatal de Quevedo profundizó en la deserción estudiantil examinando datos académicos, socioeconómicos y personales a través de técnicas de Soft Computing, utilizando 225 variables y 160,000 registros. Se compararon algoritmos como Random Forest, Gradient Boosting y SVM, destacando a Random Forest por su precisión en predecir factores críticos que influyen en la deserción. Entre los factores determinantes se encuentran la nota de corte 2, la ocupación del jefe del hogar y la edad del estudiante, señalados como los más significativos para el rendimiento estudiantil. Los hallazgos del estudio subrayan la complejidad de la deserción estudiantil, destacando que los estudiantes con mayores riesgos de reprobar tienden a tener notas bajas a mitad de período, provienen de hogares con jefes de hogar desempleados o en trabajos no calificados, y están en el rango de edad de 40-45 años. Por otro lado, los perfiles de estudiantes con mayores probabilidades de aprobar muestran patrones opuestos en estas variables. Estos resultados resaltan la necesidad de adoptar enfoques integrales y personalizados para abordar la deserción, considerando tanto factores académicos como contextos socioeconómicos y personales.

Palabras clave:

Predicción, Educación Superior, Redes neuronales artificiales, Modelos predictivos.

ABSTRACT

The study conducted at the Technical State University of Quevedo delved into student dropout by examining academic, socioeconomic, and personal data through Soft Computing techniques, utilizing 225 variables and 160,000 records. Algorithms such as Random Forest, Gradient Boosting, and SVM were compared, with Random Forest standing out for its precision in predicting critical factors influencing dropout. Among the determining factors are the cut-off grade 2, the head of household's occupation, and the student's age, identified as the most significant for student performance. The findings of the study highlight the complexity of student dropout, emphasizing that students at greater risk of failing tend to have low mid-term grades, come from households with unemployed heads or in unskilled jobs, and are in the age range of 40-45 years. On the other hand, student profiles with higher chances of passing show opposite patterns in these variables. These results underline the need for adopting comprehensive and personalized approaches to address dropout, considering both academic factors and socioeconomic and personal contexts.

Keywords:

Prediction, University Education, Artificial Neural Networks, Predictive Models.

INTRODUCCIÓN

El presente proyecto buscó identificar los factores que intervienen en la deserción de los estudiantes de las Instituciones de Educación Superior (IES) para tomar acciones que nos permita proponer soluciones. Esto debido a que las Instituciones de Educación Superior cada 5 años están expuestas a un periodo de evaluación y acreditación.

En este contexto las Instituciones de Educación Superior (IES) en Ecuador han reconocido la importancia de identificar y abordar los factores que afectan el desarrollo académico de los estudiantes y obstaculizan una adecuada gestión de su formación. Entre estos factores, las tasas de retención y deserción estudiantil han sido identificadas como áreas de preocupación prioritarias (Tinto, 2010). Las tasas de retención estudiantil se refieren a la capacidad de las IES para mantener a los estudiantes matriculados y comprometidos con su educación. Una baja tasa de retención puede indicar dificultades en el compromiso de los estudiantes, falta de apoyo académico o problemas en la calidad de la enseñanza.

Por otro lado, las bajas tasas de deserción estudiantil son un indicador importante de la eficiencia y efectividad de los programas educativos en las IES (Tinto, 2010). Una alta tasa de deserción estudiantil puede señalar obstáculos en el proceso de formación, como la falta de orientación académica adecuada. En Ecuador, existen diversos problemas que afectan la continuidad y el progreso de las actividades académicas de los estudiantes en las IES, incluyendo el entorno familiar, la situación laboral de los estudiantes, la elección de carrera, problemas de salud, barreras socioeconómicas y otros factores personales y contextuales (Fernández et al., 2019; Cortés et al., 2019; Fernández Villafañez, 2022). Detectar y comprender estos factores se vuelve fundamental para tomar medidas efectivas y diseñar estrategias que fomenten la retención estudiantil y nos ayude a reducir la deserción estudiantil.

El tema de deserción estudiantil ha sido estudiado desde hace muchos años atrás; es así como en un estudio realizado por (Murillo Aida y Jurado de los Santos Pedro, Zabala et al., 2021, hacen uso de un enfoque metodológico descriptivo, donde los datos se recopilaron mediante un cuestionario ad-hoc con una muestra de 1487 estudiantes. Los resultados revelan la importancia de factores motivacionales y la satisfacción personal en la relación con la persistencia de los estudiantes en el programa académico. Se considero que la permanencia y la persistencia son dos procesos independientes que interactúan y tienen un impacto directo en la retención de los estudiantes de la educación superior. Con este trabajo, se busca confirmar que un modelo predictivo servirá para determinar los factores con mayor impacto en la deserción y retención estudiantil que un cuestionario ad-hoc.

En otro estudio realizado por Oviedo et al. (2022), se se plantea la implementación de un método de agrupación que permita integrar de manera efectiva los datos educativos (socioeconómicos, rendimiento académico y deserción) en la Facultad de Ingeniería de la Universidad Técnica Estatal de Quevedo. Los resultados obtenidos a partir de esta investigación permitieron obtener una compresión más profunda de los diversos factores que influyen en el rendimiento de los estudiantes permitiendo así ofrecer a las autoridades institucionales la posibilidad de identificar estrategias que contribuyan a mejorar la retención y el rendimiento académico de los estudiantes. Tomando en cuenta esta investigación, se compararó si un método de agrupación es la mejor forma de obtener los factores más influyentes con respecto a la deserción y retención estudiantil.

Por otro lado, la investigación realizada por Donoso et al. (2010), se evidenció que la retención de estudiantes en la educación superior es un aspecto crucial que ha ganado relevancia en los últimos años, ya que antes no se le otorgaba la importancia debida. Esto debido a diversas razones como la competencia, las presiones de las familias, las presiones de las instituciones de financiamiento y las políticas de desarrollo de los países basadas en el fortalecimiento del capital humano avanzado. Se realizó tanto desarrollo teórico como desarrollo practico para presentar propuestas de iniciativas dirigidas al sistema educativo y a las instituciones de educación en Chile. Con respecto a este estudio, se buscó determinar cuáles son las diversas razones causantes de la deserción y retención estudiantil presentes en las instituciones de educación en Ecuador.

De igual manera, en la tesis realizada por Mayancela (2016), se eligió a los estudiantes de la carrera de Pedagogía debido al bajo índice de titulación de los estudiantes en la modalidad a distancia. Se hizo uso de una metodología mixta que permitió determinar el número de graduados, desertores y aquellos que aún no concluyen su carrera universitaria. Los resultados obtenidos permitieron ayudar a las autoridades y a la carrera a elaborar planes de mejora para aumentar el índice de titulación. Teniendo en cuenta la investigación se amplió el estudio recopilando información relacionada al tema no solo con una carrera en específico, sino con toda la Universidad Técnica Estatal de Quevedo.

Según Bonilla (2019), evidenció una problemática en la carrera de Publicidad de la Universidad Laica Vicente Rocafuerte de Guayaquil, misma que está relacionada

con la falta de conocimiento por parte de los estudiantes en los últimos ciclos de colegio. Con el objetivo de abordar esta situación, se decidió implementar una campaña de comunicación que aproveche la popularidad de las redes sociales, con el propósito de atraer a estudiantes nuevos y fomentar una forma diferente y llamativa de las diferentes plazas de trabajo que ofrece una agencia de publicidad para contribuir a la retención de los estudiantes que ya estén inscritos en la carrera de Publicidad. Con respecto a esta tesis se demostró que un modelo predictivo es una herramienta muy poderosa con respecto a el uso de redes sociales para determinar los factores de la deserción y retención estudiantil y presentar soluciones.

Los autores Suárez y Díaz (2015), en su publicación realizan una revisión sistemática con el objetivo de analizar las características del estrés académico y su impacto en la salud mental de la población universitaria. Se presentaron los resultados, incluyendo las definiciones de estrés académico, deserción estudiantil y estrategias de retención y otros aspectos relacionados tomando como referencia los programas de retención en Colombia, desde la perspectiva del Ministerio de Educación Nacional. Teniendo en cuenta esta investigación, se agruparon los factores presentes e influyentes en la deserción y retención estudiantil que tengan que ver con problemas socioeconómicos, problemas psicológicos y problemas de rendimiento académico.

Se recopilaron los datos mediante un cuestionario ad hoc con 51 ítems y se utilizaron técnicas descriptivas y de contraste en el análisis. Como resultado se detectó la presencia de problemáticas socio-familiares que influyen en los estudiantes provocando el abandono definitivo o transitorio, teniendo como muestra que el 72% de los estudiantes enfrentaron dificultades socio-familiares durante su trayectoria académica, mientras que el 26% manifestó desmotivación hacia los estudios que estaban cursando y una preferencia por otra titulación o institución. La investigación realizada, demostró que una metodología ex post-facto puede otorgar resultados válidos, pero no precisos debido a que los encuestados podrían falsificar sus respuestas.

Los autores Cortés et al. (2019), demuestran que principalmente existen 5 factores clave relacionados con las causas de la deserción estudiantil: factor individual, factor económico, factor académico y factor institucional. Se evidencio un crecimiento en el índice de deserción y la mayor disponibilidad de datos con que trabajar para poder desarrollar modelos. Esto llevo a que cada vez más instituciones de la educación superior de Chile e investigadores estudien el fenómeno de la deserción estudiantil. Tomando en cuenta los resultados obtenidos aquí, se comparó la realidad entre las instituciones de educación superior de Chile y Ecuador.

En ese sentido Fernández et al. (2019), utilizaron modelos explicativos y modelos predictivos para determinar algoritmos en las causas que influyen en la deserción estudiantil en el Instituto Tecnológico de Costa Rica con el uso de las variables registradas en el Sistema de indicadores de Gestión Institucional (SIGI). Se evidencio que el mejor algoritmo fue “Random Forest” ya que fue capaz de obtener el porcentaje más alto en la captación de deserción real, dando los primeros pasos hacia la construcción de un modelo predictivo más robusto que se espera contribuya a la toma de decisiones en el Instituto Tecnológico de Costa Rica. Teniendo en cuenta este artículo, se buscó una herramienta diferente a Random Forest para analizar y tratar los datos que permita la evaluación de los factores con mayor impacto a la deserción y retención estudiantil.

Por otro lado, González et al. (2020), utilizaron las cadenas de Márkov en una muestra de 5700 estudiantes de 8 facultades de una universidad pública y regional de Chile, donde se demostró que el porcentaje más alto 39% de deserción se da en los primeros 2 semestres de las carreras universitarias que luego se reduce en los semestres posteriores. Esto permitió una fuerte inversión focalizada en el primer año de las carreras universitarias no solo para el caso de estudio, sino para todo el sistema de educación superior. Con esta investigación se evidenció que el uso de cadenas de Márkov es una herramienta capaz de entregar resultados solidos al identificar los factores con mayor impacto en la deserción y retención estudiantil, en este proyecto se busca llegar a resultados precisos haciendo uso de un modelo predictivo escrito en lenguaje programación Python.

En resumen, los estudios e investigaciones analizados revelan la importancia de abordar diversas problemáticas relacionadas con la educación superior, como la retención estudiantil, el estrés académico y la falta de conocimiento en determinadas áreas. Estos aspectos influyen directamente en el rendimiento académico y en la satisfacción de los estudiantes. Los hallazgos de estos estudios aportan conocimientos importantes para tomar medidas que permitan mejorar la calidad de la educación superior, fomentar la retención estudiantil y promover el bienestar de los estudiantes en su trayectoria académica. Estas acciones son fundamentales para el desarrollo de una educación inclusiva y de calidad.

Mediante el uso de técnicas de Soft Computing, como la inteligencia artificial y el aprendizaje automático, se analizó la información recopilada para identificar patrones, correlaciones y tendencias que ayudan a comprender

los factores clave que influyen en la deserción estudiantil. Estas técnicas permitieron un análisis profundo de los datos, incluyendo el procesamiento de grandes volúmenes de información y la detección de patrones.

MATERIALES Y MÉTODOS

Esta investigación se desarrolló en la Universidad Técnica Estatal de Quevedo, ya que es el lugar donde se encontraron las muestras necesarias y donde se obtuvo los datos que fueron objeto de análisis y tratamiento, para presentar posibles soluciones en contra de la retención y deserción del estudiantado de la Universidad Técnica Estatal de Quevedo.

Este trabajo maneja una investigación de tipo descriptivo, en virtud de que se llevó a cabo la recolección de datos de la plataforma académica de la Universidad Técnica Estatal de Quevedo (SGA), del cual se obtuvo el dataset que contenía todos los factores que generan un impacto a la retención y deserción estudiantil. En el proceso de la aplicación del modelo predictivo se pudo encontrar patrones y tendencias que inciden en el problema planteado. Luego, tocó elaborar el marco teórico, mismo que proporcionó la estructura conceptual esencial para entender, analizar y abordar el problema, guiando el enfoque y estableciendo una base sólida para el análisis y la discusión de los resultados. Llevando a cabo una búsqueda literatura relevante para el tema.

La metodología analítica para abordar el estudio de la deserción estudiantil y el tiempo de titulación implicó una serie de pasos fundamentales. En primer lugar, se definió claramente el problema y se recopilaron datos relevantes. Después, se realizó un análisis exploratorio para identificar tendencias y patrones. Luego, se analizaron los factores de impacto utilizando técnicas estadísticas para determinar las relaciones entre variables. A partir de estos análisis, se establecieron patrones y se construyeron modelos predictivos para anticipar situaciones de deserción estudiantil. Finalmente, se propusieron acciones preventivas y recomendaciones basadas en los resultados obtenidos. Esta metodología analítica proporcionó una estructura sólida para abordar de manera efectiva los desafíos de deserción, permitiendo tomar medidas proactivas y mejorar el éxito académico de los estudiantes.

Etapa 1: Revisión bibliográfica:

Se llevó a cabo una revisión bibliográfica exhaustiva sobre la deserción estudiantil, centrándose en los motivos que más comúnmente contribuyen a esta situación. La revisión abordó una variedad de fuentes y estudios académicos para identificar los factores y circunstancias que pueden llevar a la deserción estudiantil, como dificultades académicas, problemas económicos, falta de apoyo social, desmotivación, entre otros. Esta revisión permitió obtener una comprensión más profunda de los motivos subyacentes a la deserción estudiantil, proporcionando una base sólida para el desarrollo de estrategias y acciones preventivas orientadas a reducir los índices de deserción.

Etapa 2: Identificación de requerimientos:

En este proyecto, se especificaron los requerimientos y funcionalidades esenciales que Tableau facilitó para cumplir con los objetivos planteados. Esto conllevó reconocer las características y capacidades particulares de Tableau que se emplearon en el marco del proyecto. Entre los requerimientos se destacaron la habilidad para importar y manipular datos, efectuar análisis y visualizaciones, así como la generación de informes y presentaciones detalladas.

Etapa 3: Pruebas y validación:

En el proyecto se llevaron a cabo pruebas exhaustivas para evaluar la capacidad del modelo predictivo en establecer patrones. Estas pruebas implicaron la utilización de conjuntos de datos representativos y diversos, que abarcaron diferentes escenarios y condiciones relevantes al problema que se estaba abordando. Se analizaron los resultados obtenidos por el modelo para determinar si era capaz de identificar patrones significativos y hacer predicciones precisas.

Etapa 4: Evaluación y trabajos futuros:

Tras obtener los resultados del modelo predictivo, se recomendó realizar una evaluación exhaustiva para identificar soluciones y proponer campañas contra la deserción estudiantil. Se analizaron patrones, factores influyentes y áreas problemáticas de las cuales se podrían sugerir intervenciones como programas de apoyo académico y emocional, becas, tutorías y mejoras en la calidad educativa. El objetivo fue reducir la deserción estudiantil, proporcionando una experiencia educativa satisfactoria.

Datos

El conjunto de datos para este estudio se extrajo del Sistema de Gestión Académica de la Universidad Técnica Estatal de Quevedo. Este conjunto de datos comprende 225 variables y 160,000 registros, que encapsulan una amplia gama de información sobre los estudiantes la cual será esencial para identificar patrones y correlaciones sobre la deserción estudiantil.

RESULTADOS Y DISCUSIÓN

Recopilación de datos

Para desarrollar y afinar el modelo predictivo en cuestión, se recurrió a un conjunto de datos integrales y minuciosamente recopilados de la plataforma académica de la Universidad Técnica Estatal de Quevedo. Esta base de datos no solo es rica en información académica, sino que también es una fuente invaluable sobre la trayectoria y el progreso de los estudiantes, documentando meticulosamente sus calificaciones y las asignaturas que han cursado. Aún más importante, esta plataforma ofrece una perspectiva holística del perfil estudiantil, ya que, durante el proceso de matriculación y admisión, se solicita a los estudiantes que aporten una amplia gama de información personal.

Esta información abarca mucho más que el mero rendimiento académico; incluye aspectos fundamentales como el contexto socioeconómico, detalles familiares y otros factores personales que podrían tener un impacto significativo en su desempeño académico. La recopilación de estos datos es un paso crucial, pues proporciona al modelo una base sólida para entender y predecir cómo estas variadas influencias pueden afectar los resultados académicos de los estudiantes. Por lo tanto, el modelo no solo se basa en métricas académicas, sino que también incorpora una dimensión más personalizada y contextual, mejorando significativamente la precisión y relevancia de sus predicciones. Esto permite adaptar las intervenciones y apoyos educativos de una manera más efectiva y enfocada, atendiendo a las necesidades y circunstancias únicas de cada estudiante.

Tratamiento de los datos

El conjunto de datos analizado comprende un total de 225 variables, proporcionando una amplia gama de información sobre los estudiantes. Sin embargo, no todas estas variables son pertinentes para el enfoque específico de este estudio. Por lo tanto, se hace imprescindible realizar un proceso de limpieza y filtrado de los datos. Este proceso involucra identificar y descartar aquellas variables que no contribuyen significativamente al tema de estudio del proyecto, con el objetivo de mejorar la precisión y relevancia del análisis.

En el conjunto de datos existen variables relacionadas con los familiares del estudiante, su ocupación, nivel educativo y tipo de trabajo. Estas variables se multiplican hasta alcanzar un total de 33 posibles familiares que el estudiante puede registrar en la plataforma académica. Sin embargo, en la práctica, la mayoría de los estudiantes solo registra entre 2 y 3 familiares, por lo tanto, es necesario implementar estrategias de manejo de datos faltantes para garantizar la integridad y la fiabilidad del análisis realizado. Esto incluye técnicas como la imputación de datos, eliminación de variables con excesivos valores nulos o la utilización de métodos estadísticos que se ajusten a la presencia de estos datos incompletos.

Para crear el modelo predictivo enfocado en la deserción y retención estudiantil, se consideraron variables que tengan una relación directa con el rendimiento y la experiencia estudiantil. Variables como «Asignatura», «Nota final», «Asistencia %», y «Estado» son cruciales, ya que reflejan directamente el rendimiento académico del estudiante y su compromiso con los estudios. La «Edad» y «Etnia» ofrecen perspectivas importantes sobre la diversidad y los desafíos específicos que enfrentan los estudiantes. Variables como «¿El estudiante es cabeza de familia?», «¿El estudiante depende económicamente de sus padres u otras personas?», y «¿Quién cubre los gastos del estudiante?» proporcionan información valiosa sobre las responsabilidades y presiones económicas que podrían influir en la decisión de un estudiante de continuar o abandonar sus estudios. Además, el «Nivel de instrucción del Jefe del Hogar» y el «Tipo de vivienda» podrían indicar el nivel socioeconómico y el entorno de apoyo del estudiante, factores que pueden afectar la permanencia en la educación.

Las variables analizadas proporcionan una visión holística y exhaustiva del estudiante, integrando dimensiones académicas, personales y socioeconómicas aportando una comprensión profunda y precisa de los factores que influyen en la retención y deserción estudiantil. Las variables seleccionadas para la investigación se detallan en la Tabla 1.

Tabla 1: Variables Agrupadas.

Variables Académicas	Variables Socioeconómicas	Variables Personales
Periodo	Familiar parentesco F1	Edad
Carrera	¿Convive F1?	Etnia
Nivel	Nivel de titulación F1	Región

Matrículas	Familiar parentesco F2	
Corte 1	¿Convive F2?	
Corte 2	Nivel de titulación F2	
Examen Final	¿El estudiante es cabeza de familia?	
	¿El estudiante depende económicamente de sus padres u otras personas?	
	¿Quién cubre los gastos del estudiante?	
	¿Cuál es el nivel de instrucción del jefe del Hogar?	
	¿Cuál es la ocupación del jefe del Hogar?	
	¿Cuál es el tipo de vivienda?	
	¿Su vivienda es?	

Fuente: Elaboración de autores

Variable dependiente y variables independientes

La variable dependiente en este caso de estudio es la variable "Estado" la cual es la que brinda información sobre si el estudiante aprueba o reprueba la materia, y las demás variables que conforman el dataset actúan como variables independientes.

Proceso de limpieza en Tableau

Tableau ofrece una representación visual intuitiva y accesible de los datos contenidos en nuestro conjunto de datos, facilitando su comprensión. El proceso comienza con la limpieza de los datos, específicamente abordando los valores nulos, utilizando la eficiente herramienta de filtrado que Tableau proporciona para este propósito. Esta funcionalidad permite una identificación y manejo efectivos de los valores nulos, asegurando que el análisis posterior se base en un conjunto de datos.

Tras la eliminación de las variables que no son relevantes para el proyecto, el enfoque se centra en mejorar la visualización de aquellas variables que sí tendrán un impacto significativo en el estudio del problema. Este proceso es esencial para destacar los datos más pertinentes, facilitando un análisis más efectivo y una comprensión clara de los factores críticos relacionados con la investigación.

Finalmente, la limpieza de datos se completa, proporcionando una visibilidad clara y detallada sobre las variables específicamente seleccionadas para este estudio. Este paso crucial asegura que solo los datos más relevantes y útiles sean utilizados en el análisis posterior, permitiendo así una evaluación precisa y enfocada de los aspectos clave del estudio.

Categorización de variables

Esta etapa reviste una importancia fundamental en nuestra investigación, ya que la adecuada categorización de las variables desempeña un papel crucial en la preparación de los datos para su posterior análisis y entrenamiento del modelo. A través de esta categorización, logramos la conversión de datos cualitativos en datos numéricos o códigos, lo cual es esencial para su incorporación efectiva en el proceso de entrenamiento del modelo. Esta transformación facilita la interpretación y procesamiento de los datos, permitiendo un análisis más estructurado y preciso. Además, esta categorización nos permite evaluar y comprender mejor la relación entre las variables y su impacto en el fenómeno de estudio, la deserción y retención estudiantil.

El resultado de este proceso es un archivo en el que se refleja cómo se han codificado las variables, lo que brinda la posibilidad de realizar pruebas y experimentos con el modelo desarrollado. Esta visualización de la codificación es esencial para comprender cómo se han transformado los datos y cómo afecta al rendimiento del modelo en la

predicción de la deserción y retención estudiantil. Este enfoque flexible y controlado nos permite ajustar y mejorar el modelo a medida que avanzamos en nuestra investigación.

Balanceo de datos

El balanceo de los datos de un dataset se considera una práctica esencial en el campo del procesamiento de datos y análisis predictivo, especialmente relevante en áreas como el aprendizaje automático y el modelado estadístico. La necesidad de balancear los datos surge de varios factores críticos que impactan directamente en la eficacia y fiabilidad de los modelos predictivos.

En primer lugar, el balanceo de datos es fundamental para prevenir el sesgo en los modelos predictivos. Un dataset desbalanceado puede llevar a que el modelo desarrolle una preferencia hacia la clase más representativa, resultando en una capacidad reducida para identificar correctamente los casos pertenecientes a la clase minoritaria. Esto es particularmente problemático en aplicaciones donde la identificación precisa de las clases menos comunes es crucial, como en la detección de transacciones fraudulentas o el diagnóstico de condiciones médicas raras.

Además, el equilibrio de los datos contribuye a la mejora de la precisión de las predicciones del modelo. Al asegurar que las clases minoritarias sean adecuadamente representadas, los modelos son capaces de aprender de manera más efectiva las características distintivas de todas las clases, lo que lleva a un mejor rendimiento general. Otro punto para destacar es el impacto del balanceo de datos en la sensibilidad y especificidad de los modelos. En contextos donde es vital acertar en la detección de los verdaderos positivos, como en aplicaciones médicas, el balanceo de los datos puede incrementar significativamente la sensibilidad del modelo, es decir, su capacidad para identificar correctamente los casos positivos, sin comprometer la especificidad, o su habilidad para reconocer los negativos.

Finalmente, el balanceo de datos facilita una evaluación más precisa y equitativa del rendimiento de los modelos predictivos. Los datasets equilibrados permiten que las métricas de evaluación reflejen de manera más fidedigna la capacidad del modelo para generalizar y predecir con precisión sobre datos nuevos y no vistos anteriormente, evitando así evaluaciones sesgadas que podrían sobreestimar el rendimiento real del modelo en situaciones prácticas.

En el proceso de balanceo de los datos de un dataset, se emplearon técnicas avanzadas como Random UnderSampling, SMOTE (Synthetic Minority Over-sampling Technique) y SMOTE Tomek para abordar el problema del desbalance entre clases. Cada una de estas técnicas tiene características únicas que contribuyen de manera significativa al mejoramiento de la calidad y equidad de los datos, lo que a su vez impacta positivamente en el rendimiento de los modelos predictivos.

El uso de estas técnicas de balanceo de datos en conjunto ofrece una solución más completa y efectiva para el desafío del desbalance de clases en datasets. Al equilibrar adecuadamente las clases, se facilita el desarrollo de modelos predictivos más precisos, justos y capaces de generalizar mejor sobre datos no vistos, mejorando significativamente la toma de decisiones basada en datos en diversos campos de aplicación. Como resultados de los algoritmos mencionados se obtuvo:

Tabla 2: Resultados de algoritmos de balanceo.

	Dropout	Pass	Overall	IR
Original	3,039	101,795	104,834	33.50
RandomUnderSampler	3,039	3,039	6,078	1
SMOTE	101,795	101,795	203,590	1
SMOTETomek	101,756	101,756	203,512	1

Fuente: Elaboración de autores

La tabla 2 muestra un resumen de la distribución de dos clases, 'Dropout' y 'Pass', en un dataset antes y después de aplicar distintas técnicas de balanceo de datos. El 'IR' (Imbalance Ratio) mide la proporción entre las clases; un IR de 1 indica que hay un balance perfecto entre las clases.

La elección de SMOTETomek está basada en su capacidad de balancear el dataset (como lo indican los IR de 1 en las técnicas de balanceo) al tiempo que mejora la calidad de los datos. Al eliminar los enlaces Tomek, que son pares de datos cercanos pero de clases opuestas, SMOTETomek puede reducir el ruido y las ambigüedades en la frontera

de decisión entre las clases. Esto es beneficioso para el modelo predictivo, ya que ayuda a mejorar la generalización y evitar el sobreajuste que podría resultar de la creación de demasiados datos sintéticos con SMOTE. Por lo tanto, se prefirió SMOTETomek sobre las otras técnicas por mantener un balance entre las clases y al mismo tiempo limpiar el dataset para una mejor calidad de los datos de entrenamiento.

Comparación entre algoritmos

Se realizó una comparativa exhaustiva entre tres algoritmos de aprendizaje automático ampliamente reconocidos: Random Forest, Gradient Boosting y Support Vector Machine (SVM), con el fin de determinar cuál de ellos muestra un rendimiento superior en términos de precisión predictiva, capacidad de generalización y eficiencia computacional. Cada uno de estos algoritmos tiene sus propias fortalezas y es idóneo para diferentes tipos de tareas y conjuntos de datos (Tabla 3).

Tabla 3: Tabla comparativa entre algoritmos.

Criterio	Random Forest	Gradient Boosting	Support Vector Machine (SVM)
Complejidad del Modelo	Media-Alta	Alta	Media-Alta
Capacidad para Datos No Lineales	Excelente	Excelente	Excelente
Sensibilidad a Características No Escaladas	Baja	Media	Alta
Velocidad de Entrenamiento	Rápida en paralelo	Más lenta, secuencial	Variable, depende del tamaño del dataset
Uso de Memoria	Moderado-Alto	Moderado-Alto	Moderado-Alto
Interpretabilidad	Moderada	Moderada-Baja	Baja

Fuente: Elaboración de autores

La tabla proporciona una comparativa de tres populares algoritmos de aprendizaje automático: Random Forest, Gradient Boosting y Support Vector Machine (SVM), según diversos criterios de rendimiento (Learn, s.f.).

Rendimiento de modelos

Tras haber procesado meticulosamente los datos, que incluyó su limpieza, codificación y balanceo mediante técnicas especializadas, se avanzó hacia la fase crucial de evaluación del rendimiento de varios algoritmos de aprendizaje automático, con el objetivo de entrenar un modelo predictivo óptimo. Esta evaluación fue sistemática y metódica, tomando en consideración una variedad de métricas de rendimiento para asegurar una comparación exhaustiva y objetiva. Los algoritmos sometidos a prueba incluyeron Random Forest, Gradient Boosting y Support Vector Machine (SVM), cada uno con sus propias particularidades y potencialidades en el manejo de datos complejos y su habilidad para generalizar predicciones a partir de nuevas entradas. El proceso de selección del algoritmo más eficaz se basó en una combinación de precisión predictiva, eficiencia computacional y capacidad de generalización, buscando así el equilibrio ideal entre rendimiento y practicidad para aplicaciones del mundo real.

- Rendimiento del algoritmo Random forest

Codificando en Python se utiliza la función **accuracy_score** para calcular la precisión del modelo, la cual devolvió un resultado de 98.90%. Esto indica que el modelo fue capaz de predecir correctamente el 98.90% de los casos en el conjunto de datos de prueba. El soporte para cada clase es casi igual, con 20,397 para la clase 0 y 20,306 para la clase 1, lo que indica un dataset balanceado, y el soporte total están consideradas 40,703 instancias de ambas clases.

- Rendimiento del algoritmo Gradient Boosting

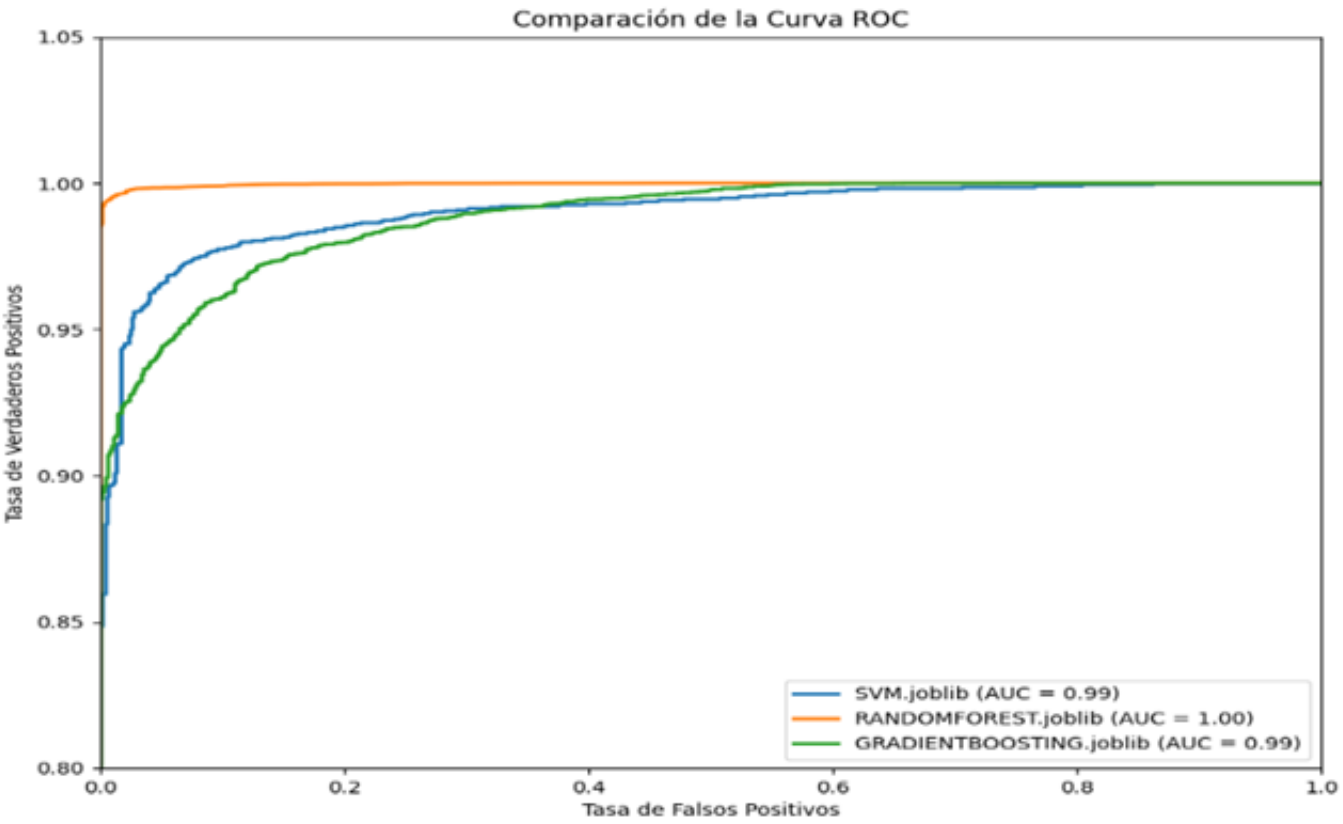
Al aplicar la precisión del modelo de clasificación se obtiene el valor de 98.23%, lo que indica que el modelo fue capaz de predecir correctamente el 98.23% de las instancias en el conjunto de datos de prueba. El «Reporte de Clasificación» de la evaluación del modelo muestra una precisión de la clase '0' del 98%, mientras que de la clase '1' el 100%, lo que podría enmascarar el hecho de que su capacidad para identificar correctamente la clase '0' es significativamente más baja que para la clase '1', lo que sugiere que el modelo podría estar sesgado hacia la clase más representada.

- Rendimiento del algoritmo Maquina de soporte vectorial

En la evaluación del modelo de clasificación utilizando el algoritmo de Support Vector Machine (SVM) devuelve una precisión del 98.19%. De las 40,703 muestras evaluadas este modelo indica que existen 20,397 muestras para la clase '0' y 20,306 para la clase '1'. En el análisis indicó que la sensibilidad de la clase '0' es de 0.99 y para la clase '1' de 0.97; mientras que la media armónica entre precisión y la sensibilidad de ambas clases se ubicó en 0.98.

Gráfico de la curva ROC (Receiver Operating Characteristic)

Fig. 1: Curva de Roc.



Fuente: Elaboración de autores

La figura 1 muestra una gráfica de la Curva ROC (Receiver Operating Characteristic) que compara el rendimiento de tres modelos de clasificación diferentes: SVM (Support Vector Machine), Random Forest y Gradient Boosting. La Curva ROC es una herramienta gráfica utilizada para evaluar la capacidad de diagnóstico de un sistema clasificador binario.

En la gráfica, el eje x representa la Tasa de Falsos Positivos (1 - Especificidad), y el eje y representa la Tasa de Verdaderos Positivos (Sensibilidad). Una curva más cercana a la esquina superior izquierda indica un mejor rendimiento. La línea punteada representa un clasificador aleatorio, con un Área Bajo la Curva (AUC) de 0.5, lo que significa que no tiene capacidad de clasificación mejor que el azar.

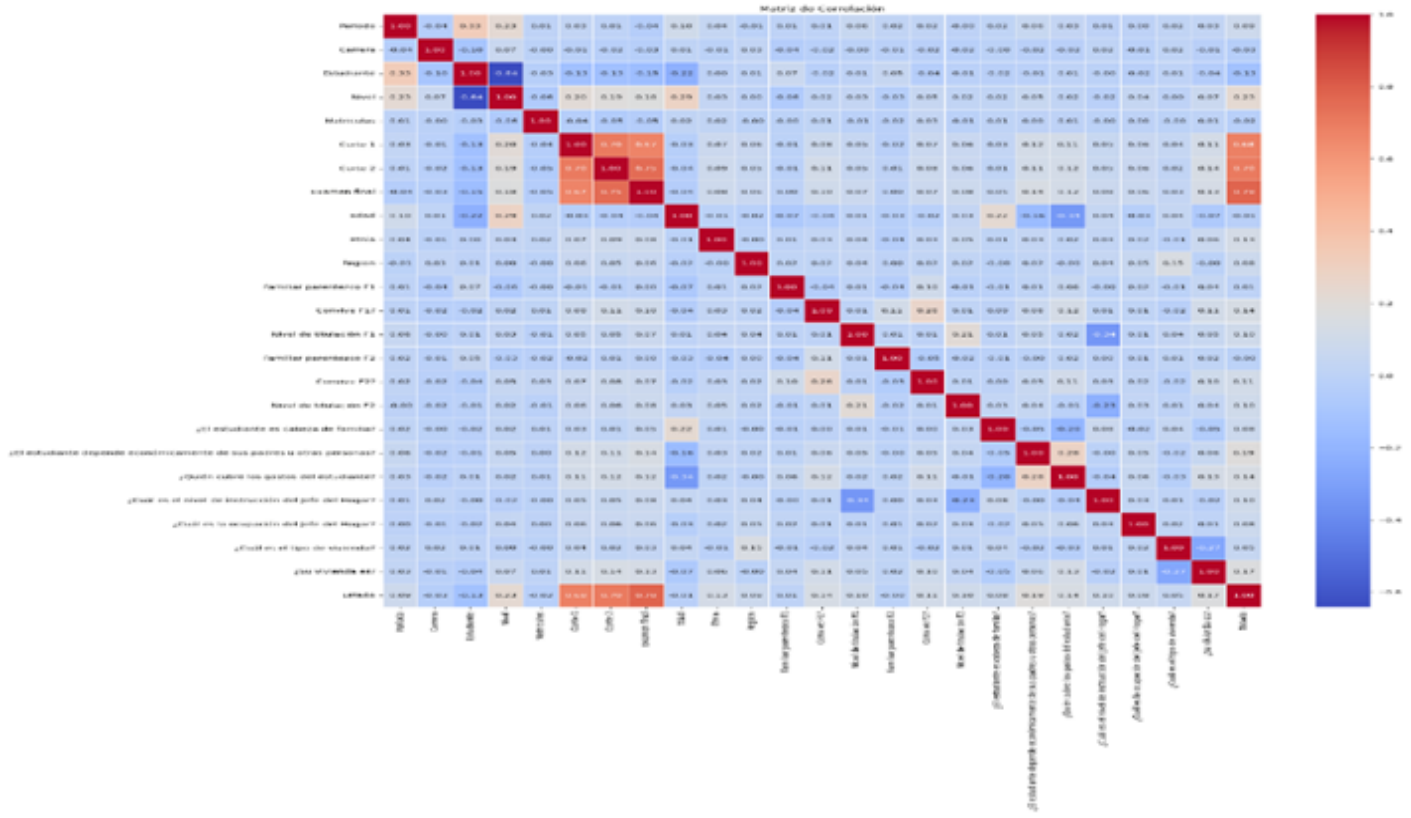
Selección de algoritmo

Tras una cuidadosa evaluación de las métricas de rendimiento, se seleccionó el algoritmo Random Forest para el entrenamiento del modelo predictivo debido a su destacada precisión y la solidez general demostrada en los resultados obtenidos. Esta decisión se basó en un análisis comprensivo que ponderó no solo la precisión, sino también otros

indicadores clave de rendimiento que apuntan a la eficacia del algoritmo en el manejo de los datos y su capacidad predictiva.

Matriz de correlación de modelo predictivo de la totalidad de las variables

Fig. 2: Matriz de correlación.



Fuente: Elaboración de autores

La figura 2 muestra una matriz de correlación, que es una tabla que ilustra la correlación entre múltiples variables. Cada celda de la matriz muestra el coeficiente de correlación entre dos variables. Los valores de correlación varían entre -1 y +1:

Esta herramienta es útil para identificar relaciones entre variables, lo que puede ser fundamental para la selección de características en el modelado estadístico y en el aprendizaje automático. Las variables con alta correlación pueden llevar a problemas de multicolinealidad en modelos lineales, mientras que en otros modelos pueden indicar características redundantes que podrían ser eliminadas para simplificar el modelo.

Siendo las variables Nota corte 1, Nota corte 2 y examen final las que mayor coeficiente de correlación positiva tienen denotando como son las variables más representativas dentro del modelo.

Entrenamiento del Modelo Random Forest con variables agrupadas

Con el propósito de realizar un análisis más detallado, las variables del estudio se clasificaron en tres categorías principales: socioeconómicas, académicas y personales. Para cada categoría, se empleó el algoritmo Random Forest, elegido por su eficacia probada, para entrenar un modelo predictivo independiente. A través de este enfoque, se entrenaron tres modelos distintos, correspondientes a cada grupo de variables. Posteriormente, se identificó la variable más significativa dentro de cada modelo, lo cual permitió destacar la característica individual con el mayor impacto predictivo en su respectivo conjunto. Este proceso resultó en una comprensión más rica de la influencia específica de las variables dentro de sus dominios categorizados.

- Modelo variables socioeconómicas (Precisión del modelo = 94.84%) (Tabla 4).

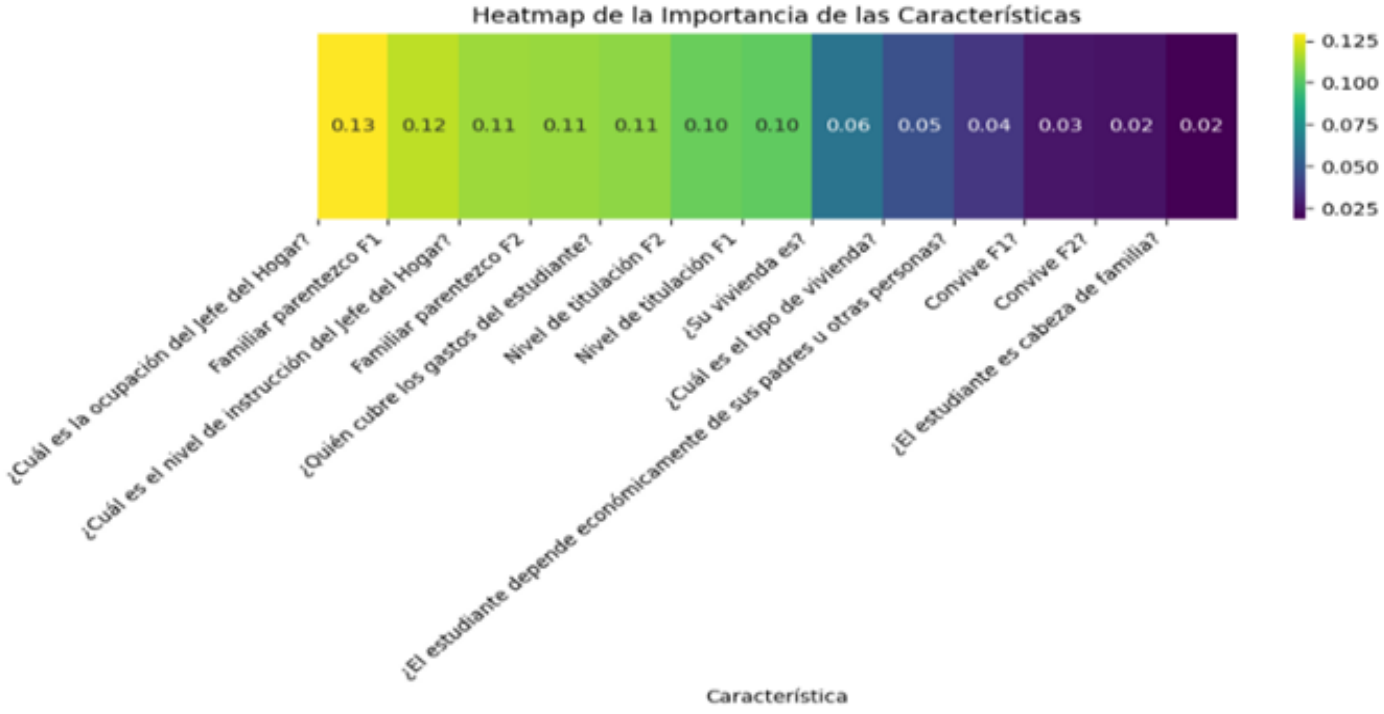
Tabla 4: Rendimiento de modelo socioeconómico.

	Precisión	Recall	F1-Score	Support
0	0.95	0.94	0.95	20397
1	0.94	0.95	0.95	20306

Fuente: Elaboración de autores

Se obtuvo como F1-score 0.95 y 0.95 mostrando un alto rendimiento del modelo del conjunto de variables socioeconómicas para predecir los dos estados, “Aprobado” y “Reprobado” (Figura 3).

Fig. 3: Heatmap característica más importante modelo socioeconómico.



Fuente: Elaboración de autores

Y como variable más significative “¿Cuál es la ocupación del jefe del hogar?”, demostrando que dicha variable influye en el rendimiento del estudiante. Esto indica que la ocupación del jefe de familia tiene una influencia notable en el rendimiento académico del estudiante. Esta influencia puede atribuirse a diversos factores socioeconómicos que están intrínsecamente relacionados con la ocupación, tales como el nivel de ingresos del hogar, la estabilidad económica, el acceso a recursos educativos, y el entorno socio-cultural que proporciona el jefe del hogar.

- Modelo variables académicas (Precisión del modelo = 98.34%) (Tabla 5).

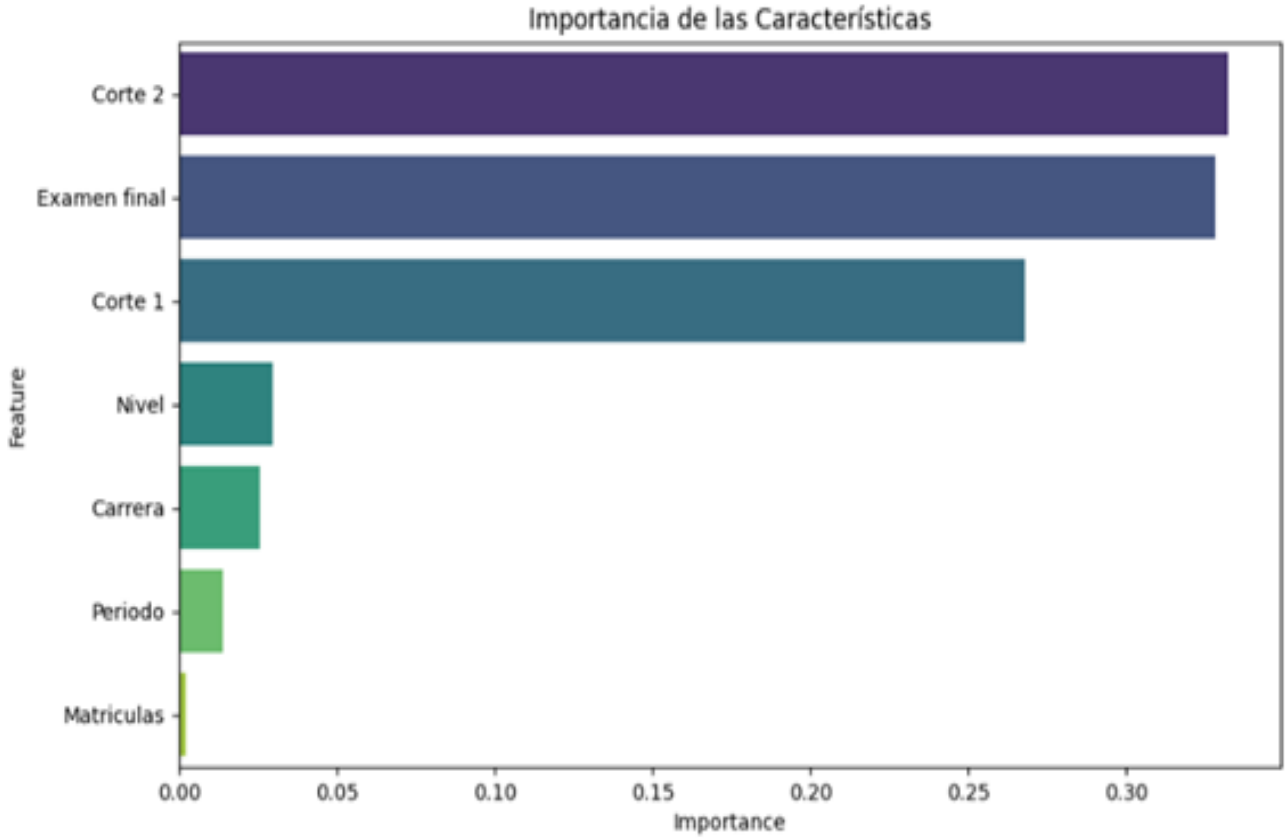
Tabla 5. Rendimiento de modelo académico.

	Precisión	Recall	F1-Score	Support
0	0.97	1.00	0.98	20397
1	1.00	0.97	0.98	20306

Fuente: Elaboración de autores

Se obtuvo como F1-score 0.98 y 0.98 mostrando un alto rendimiento del modelo del conjunto de variables académicas para predecir los dos estados, “Aprobado” y “Reprobado” (Figura 4).

Fig. 4: Característica más importante en modelo académico.



Fuente: Elaboración de autores

Y como variable más significative “Nota Corte 2” demostrando que la nota del segundo corte evaluativo tiene un impacto critico en el rendimiento del estudiante. Este hallazgo sugiere que el desempeño del estudiante en ese punto del periodo académico puede ser un predictor clave de su éxito o dificultades futuras. El hecho de que la «Nota Corte 2» sea un indicador tan decisivo puede reflejar varios factores. Por ejemplo, puede ser un reflejo de la capacidad del estudiante para adaptarse y responder a los contenidos del curso, su habilidad para manejar el ritmo y las demandas académicas, o puede ser un indicador de la efectividad de las estrategias de estudio empleadas.

- Modelo variables personales (Precisión del modelo = 57.91%) (Tabla 6).

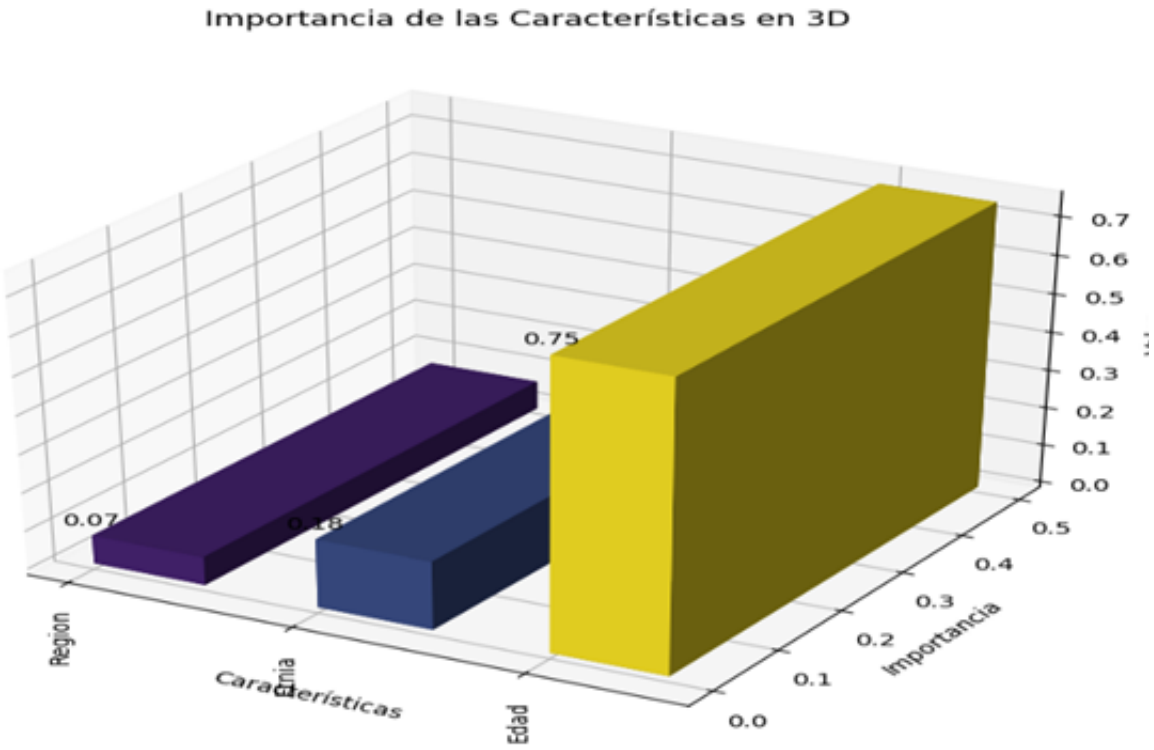
Tabla 6. Rendimiento del modelo personal.

	Precisión	Recall	F1-Score	Support
0	0.59	0.54	0.56	20397
1	0.57	0.62	0.59	20306

Fuente: Elaboración de autores

Se obtuvo como F1-score 0.56 y 0.59 mostrando un rendimiento deficiente del modelo del conjunto de variables personales para predecir los dos estados, “Aprobado” y “Reprobado” (Figura 5).

Fig. 5. Característica más importante en el modelo personal.



Fuente: Elaboración de autores

Y como variable más significative “Edad” demostrando que la edad del estudiante tiene un impacto critico en el rendimiento del estudiante. Este resultado sugiere que los estudiantes de mayor edad tienden a tener más responsabilidades fuera de las instituciones de educación superior perjudicando su rendimiento y conllevando a la deserción de sus carreras universitarias.

Perfiles de estudiantes

Se determinó que las variables más significativas dentro del conjunto de variables académicas, socioeconómicas y personales son, en orden de importancia: la nota de corte 2, la ocupación del jefe del hogar y la edad. Con respecto a los datos se obtuvo (Tabla 7).

Tabla 7. Nota corte 2.

Rango nota corte 2	Casos de Aprobados %	Casos de Reprobados %
0-1	0%	100%
1-2	0%	100%
2-3	7%	93%
3-4	15%	85%
4-5	49%	51%
5-6	72%	28%
6-7	89%	11%
7-8	97%	3%
8-9	> 99%	< 1%
9-10	> 99%	< 1%

Fuente: Elaboración de autores

La tabla 10 ilustra claramente la correlación entre el rango de notas obtenidas en el 'corte 2' y los porcentajes de aprobación y reprobación. Se observa una tendencia donde a medida que el rango de la nota aumenta, el porcentaje de casos de aprobados se incrementa significativamente, mientras que el porcentaje de casos de reprobados disminuye. Específicamente, los rangos de notas más bajos (0-1 y 1-2) tienen una tasa de aprobación del 0%, lo que indica que todos los estudiantes en estos rangos han reprobado. Por otro lado, en los rangos de notas más altos (8-9 y 9-10), más del 99% de los estudiantes han aprobado, con menos del 1% reprobando.

Esta distribución puede ser indicativa de un umbral de rendimiento crítico en torno al rango de 4-5, donde la mayoría de los estudiantes comienza a superar la barrera de la aprobación. Además, los datos sugieren que los estudiantes con notas en el rango de 5-6 o superiores tienen una alta probabilidad de aprobar. Estos resultados pueden ser fundamentales para entender la eficacia de los métodos de enseñanza y evaluación aplicados, así como para identificar áreas de mejora en el soporte educativo a los estudiantes que se encuentran en los rangos de notas inferiores.

Tabla 8. Edad.

Rango de edad	Casos de Aprobados %	Casos de Reprobados %
17-20	95%	5%
20-25	95%	5%
25-30	95%	5%
30-35	94%	6%
35-40	95%	5%
40-45	93%	7%
45-50	98%	2%

Fuente: Elaboración de autores

Los datos mostrados en la tabla 8 reflejan una alta tasa de aprobación generalizada a través de los distintos rangos de edad, manteniéndose consistentemente por encima del 90%. Sin embargo, si se analizan los porcentajes más bajos de aprobación, se observa que el rango de edad de 40-45 años tiene un porcentaje ligeramente mayor de reprobación (7%) en comparación con los otros grupos.

Una explicación hipotética para este aumento relativo en la tasa de reprobación en el grupo de edad de 40-45 años podría estar relacionada con varias transiciones y desafíos que son comunes en esta etapa de la vida. Estos pueden incluir responsabilidades crecientes tanto en el ámbito profesional como en el personal, como cargos de alta dirección que demandan más tiempo o el cuidado de familiares mayores, además de la crianza de los hijos. Estas responsabilidades podrían resultar en una menor disponibilidad de tiempo y recursos para dedicar al estudio y la preparación para evaluaciones, resultando en un impacto negativo en el rendimiento. Además, es posible que existan factores psicosociales, como el estrés o la fatiga acumulada, que podrían afectar a este grupo etario. También podría haber un componente de adaptación a nuevas tecnologías o metodologías de aprendizaje que son más utilizadas en las evaluaciones modernas, lo cual podría suponer una barrera adicional para este grupo de edad en particular.

La Tabla 9 muestra una comparación de las tasas de aprobación y reprobación basadas en la ocupación del jefe de hogar. Es destacable que los 'Inactivos' y 'Trabajadores no calificados' tienen las tasas más altas de aprobación con un 96%, mientras que 'Operadores de instalaciones y máquinas' presentan la tasa de aprobación más baja con un 91%. Aunque las diferencias son relativamente pequeñas, estos resultados podrían sugerir que aquellos jefes de hogar que no participan activamente en la fuerza laboral, o que están empleados en trabajos que no requieren calificación, podrían tener más tiempo o recursos para apoyar a los miembros del hogar en sus procesos educativos, o que podrían vivir en

contextos que de alguna manera favorecen la aprobación en las evaluaciones consideradas.

Por otro lado, la menor tasa de aprobación entre los 'Operadores de instalaciones y máquinas' podría reflejar las exigencias de tiempo y atención asociadas con este tipo de trabajo, que podrían limitar la disponibilidad para el apoyo educativo o involucramiento en actividades de aprendizaje. Basado en los conjuntos de datos proporcionados, se pueden esbozar dos perfiles distintos para los estudiantes con la mayor probabilidad de reprobar y los que tienen la mayor probabilidad de aprobar, respectivamente.

Tabla 9: Comparación de las tasas de aprobación y reprobación

Ocupación	Casos de Aprobados %	Casos de Reprobados %
Desocupados	94%	6%
Empleados de oficina	95%	5%
Fuerzas armadas	95%	5%
Inactivos	96%	4%
Oficiales operarios y artesanos	95%	5%
Operadores de instalaciones y máquinas	91%	9%
Personal directivo de la administración pública y empresas	93%	7%
Profesionales científicos e intelectuales	96%	4%
Técnicos y profesionales de nivel medio	94%	6%
Trabajador calificado agropecuarios y pesqueros	94%	4%
Trabajador de servicios y comerciante	95%	5%
Trabajadores no calificados	96%	4%

Fuente: Elaboración de autores

Perfil del estudiante con mayor probabilidad de reprobar

- **Edad:** En el rango de 40-45 años, por tener un porcentaje ligeramente mayor de reprobación.
- **Ocupación del jefe de hogar:** Hogar con jefe de hogar como 'Operador de instalaciones y máquinas', por mostrar una tasa de aprobación más baja.
- **Nota de corte 2:** Rangos de nota de 0-3, ya que estos rangos tienen los porcentajes más altos de reprobación, con un 0% de aprobados para los rangos de 0-1 y 1-2, y solo un 7% para el rango de 2-3.

Perfil del estudiante con mayor probabilidad de aprobar:

- **Edad:** En el rango de 45-50 años, donde la tasa de aprobación es la más alta.
- **Ocupación del jefe de hogar:** Hogar con jefe de hogar 'Inactivo' o 'Trabajador no calificado', ambos con las tasas más altas de aprobación.
- **Nota de corte 2:** Rangos de nota de 8-10, ya que estos rangos tienen los porcentajes más bajos de reprobación y más del 99% de casos de aprobados.

Tabla 10. Perfil del estudiante

Perfil	Rango de edad	Ocupación del jefe de hogar	Rango de nota corte 2	Casos de Aprobados %	Casos de Reprobados %
Mayor probabilidad de reprobar	40-45	Operadores de instalaciones y máquinas	0-3	91%	9%
Mayor probabilidad de aprobar	45-50	Inactivo o Trabajador no calificado	8-10	>99%	<1%

Fuente: Elaboración de autores

La presente investigación ha revelado que la ocupación del jefe del hogar (Tabla 10), la nota de corte 2 y la edad del estudiante siendo parte de conjuntos de variables académicas, socioeconómicas y personales juegan un papel crucial en el rendimiento académico y la deserción estudiantil en la Universidad Técnica Estatal de Quevedo. Estos resultados coinciden parcialmente con los hallazgos de Oviedo et al. (2022), quienes también identificaron la significativa influencia de factores socioeconómicos, académicos e individuales en el rendimiento estudiantil. Sin embargo, a diferencia Contreras et al. (2020), que reconocen la preeminencia de factores académicos pre-universidad, factores demográficos y factores socioeconómicos. este estudio destaca la complejidad y la naturaleza multifacética de la deserción estudiantil.

La influencia de la ocupación del jefe del hogar sugiere que el contexto socioeconómico del estudiante no puede ser ignorado al diseñar estrategias de intervención educativa. Esta observación es crucial para el desarrollo de políticas inclusivas que aborden las necesidades de una población estudiantil diversa. Por otro lado, la importancia de la nota de corte 2 como un indicador temprano de riesgo de deserción subraya la necesidad de implementar sistemas de apoyo académico temprano para estudiantes en riesgo.

Además, la edad se ha demostrado como un factor diferenciador en el rendimiento académico, lo que plantea preguntas sobre cómo las universidades pueden adaptar sus servicios de apoyo para satisfacer las necesidades de los estudiantes de diferentes grupos de edad. Este hallazgo desafía la noción tradicional de que los factores académicos son los únicos determinantes del éxito estudiantil y abre nuevas líneas de investigación sobre el impacto de la madurez y las responsabilidades externas en la educación superior.

Sin embargo, este estudio no está exento de limitaciones. La investigación se centró exclusivamente en una institución, lo que puede afectar la generalización de los resultados. Futuros estudios deberían explorar la interacción de estos factores en diferentes contextos educativos y culturales para obtener una comprensión más holística de la deserción estudiantil.

CONCLUSIONES

La investigación en la Universidad Técnica Estatal de Quevedo sobre los factores que afectan el rendimiento académico de los estudiantes ha arrojado conclusiones significativas. Se ha identificado que tres elementos clave. La ocupación del jefe del hogar, la nota del segundo corte evaluativo y la edad del estudiante emergen como indicadores cruciales en la influencia del rendimiento académico y la retención estudiantil.

Por un lado, se ha esbozado el perfil de los estudiantes con mayor probabilidad de reprobar: aquellos en el rango de edad de 40-45 años, provenientes de hogares donde el jefe tiene ocupaciones como operador de instalaciones y máquinas, y con notas de corte 2 en los rangos más bajos (0-3). Por otro lado, el perfil de los estudiantes con mayor probabilidad de aprobar incluye a aquellos en el rango de edad de 45-50 años, de hogares con jefes inactivos o trabajadores no calificados, y con notas de corte 2 en los rangos más altos (8-10).

Estos hallazgos enfatizan la necesidad de estrategias educativas integrales que consideren el bienestar socioeconómico y personal de los estudiantes, más allá de los aspectos puramente académicos. La implementación de políticas y programas que aborden estas variables puede contribuir significativamente a mejorar la retención estudiantil y el éxito académico en la Universidad Técnica Estatal de Quevedo.

Los resultados subrayan la importancia de adoptar un enfoque holístico en la educación, reconociendo que el rendimiento académico es el resultado de una compleja interacción de factores académicos, personales, y socioeconómicos, y destacan la importancia de políticas y programas educativos que consideren integralmente el contexto socioeconómico y las etapas de vida de los estudiantes para mejorar su rendimiento y experiencia educativa.

REFERENCIAS BIBLIOGRÁFICAS

Bonilla, R. (2019). ***Youtube como estrategia de contenidos para la atraccion y retencion de estudiantes en la carrera de publicidad de la Universidad Laica Vicente Rocafuerte de Guayaquil***. [Trabajo de titulación. Universidad Laica Vicente Rocafuerte].

Contreras, L. E., Fuentes, H. J., y Rodríguez, J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. ***Formación Universitaria, 13(5).*** https://doi.org/http://dx.doi.org/10.4067/S0718-50062020000500233

Cortés, S., Alvarez, P., Llanos, M., y Castillo, L. (2019). Deserción universitaria: La epidemia que aqueja a los sistemas de eduacion superior. ***Revista Perspectiva***, ***20***(1), 13-25. https://doi.org/ https://doi.org/10.33198/rp.v20i1.00017

Donoso, S., Donoso, G., & Arias, O. (2010). Iniciativas de retención de estudiantes de educación superior. ***Calidad en la Educación, 33, 15-61***. https://doi.org/10.31619/caledu.n33.138

Fernández Villafañez, S. (2022). ***Métodos de regresión y clasificación basados en árboles.*** [Tesis de grado. Universidad de Valladolid].

Fernández, T., Solis, M., Hernández, M. T., y Moreira, T. (2019). Un análisis multinomial y predictivo de los factores asociados a la desercion universitaria. ***Revista Electronica Educare***, ***23***(1), 73-97. https://www.scielo.sa.cr/scielo.php?script=sci_abstract&pid=S1409-42582019000100073&lng=es

González, J., Carvajal, C., y Aspeé, J. (2020). Modelación de la deserción universitaria mediante cadenas de Markov. ***Revista Uniciencia, 34(1),*** 129-146. https://doi.org/ http://dx.doi.org/10.15359/ru.34-1.8

González, T., y Pedraza, I (2017). Variables personales y de contexto que inciden en el abandono universitario. Un estudio a través del análisis de correspondencias simple (ACS). ***Actas XVIII Congreso Internacional de Investigación Educativa: interdisciplinariedad y transferencia (AIDIPE, 2017),*** 1129-1139. https://dialnet.unirioja.es/servlet/articulo?codigo=7715894

Mayancela, L. V. (2016). Causas que inciden en la no titulación de los estudiantes de la carrera de Pedagogía cohorte 2010-2011 de la Universidad Politécnica Salesiana. Ecuador: Universidad Politécnica Salesiana

Murillo-Zavala, A. y Jurado-de los Santos, P. (2021). Permanencia estudiantil: Factores que inciden en el Politécnico Internacional de Bogotá, Colombia. ***Revista Electrónica Educare, 25***(1), 1-25.

Oviedo Bayas, B., Gómez Gómez, J., Zambrano Vega, C., Y Morán Morán, E. R. (2022). Applying bayesian networks in student dropout data. ***Revista Universidad y Sociedad, 14***(2), 297-304. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202022000200297

Suárez, N. y Díaz, L. (2015). Estrés académico, deserción y estrategias de retencion de estudiantes en la educacion superior. ***Revista de Salud Pública, 17(2),*** 300-313. https://doi.org/http://dx.doi.org/10.15446/rsap.v17n2.52891

Tinto, V. (2010). From Theory to Action: Exploring the Institutional Conditions for Student Retention. En, ***Higher Education: Handbook of Theory and Research***. (pp. 51-89). Springer.