Técnicas de Machine Learning para la predicción del rendimiento académico en las pruebas Saber Pro en Colombia.

Miguel Angel Garcia Cano

Asesor

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Dedicatoria

A Dios, por ser mi apoyo incondicional en este proceso de adquisición de conocimiento para encontrar el propósito que tanto he buscado.

A mi esposa, por el amor, la confianza, el apoyo, el tiempo de dedicación, por cada consejo que me brindas al momento de decaer para levantarme y seguir adelante.

A mis padres, por enseñarme a nunca rendirme, por enseñarme a trabajar y a ser un hombre íntegro, correcto y responsable.

A mis profesores, por dedicarme el tiempo, el conocimiento, las asesorías y enseñarme que con pasión y empeño se pueden logras las metas propuestas.

Este trabajo está dedicado a todas las personas que realizar su aporte al cumplimiento de los objetivos planteados.

Agradecimientos

Me gustaría extender mi más sincero agradecimiento a todos los que ayudaron a terminar este trabajo:

A mi director de tesis Luis Angel Anillo Arrieta, por sus asesorías expertas, aporte de conocimiento y correcciones a tiempo para que este trabajo cumpla con todas las expectativas propuestas.

A mi tutora Dayana Alejandra Barrera Buitrago, pero el acompañamiento incondicional, apoyo de conocimiento desde los inicios de la especialización hasta el momento, también quiero agradecer por su apoyo experto con sus retroalimentaciones precisas y concretas.

A mis compañeros Darwin Mercado y Diego Vásquez, que en los inicios del proyecto construimos significativamente las bases sólidas del proyecto, para su futuro desarrollo.

A la UNAD, por ser mi institución mentora desde el pregrado con la carrera de Ingeniería Electrónica y ahora con la especialización de Ciencias de Datos y Analítica.

Resumen

Este estudio emplea técnicas de machine learning, que incluyen regresión lineal, regresión logística y árbol de decisión, para predecir el rendimiento académico en Colombia para las evaluaciones Sabre Pro-periodo 2022. Las pruebas denominadas Sabre Pro son esenciales para evaluar el rendimiento de los estudiantes de pregrado del país. Sin embargo, debido a que muchos factores, como la escolaridad previa, el entorno socioeconómico y los rasgos personales, influyen en este desempeño, predecirlo con precisión es difícil.

El proyecto plantea preocupaciones de como determinar variables importantes, recopilar y analizar datos pertinentes y seleccionar algoritmos apropiados. Los modelos de predicción se crean utilizando árbol de decisión, regresión logística y regresión lineal. El objetivo del estudio es identificar cuál de estos algoritmos es más preciso a la hora de pronosticar el rendimiento académico de los estudiantes en los exámenes Sabre Pro.

Este asunto es importante porque tiene el potencial de mejorar la educación superior colombiana al permitir que las instituciones identifiquen a los estudiantes con falencias para prestarles una mayor asistencia y desarrollen programas personalizados. Los resultados de la investigación pueden ayudar al gobierno colombiano y a los sectores educativos a desarrollar aplicaciones valiosas y así poder ser más selectivo a la hora de tomar decisiones.

***Palabras clave**:* Predicción, Regresión Lineal, Regresión Logística, Árbol de decisión

Abstract

This study employs machine learning approaches, including linear regression, logistic regression, and decision tree, to predict academic performance in Colombia for the Saber Pro-2022 assessments. Tests called Saber Pro are essential to assess student performance. in the country's higher education system. However, because many factors, such as prior schooling, socioeconomic background, and personal traits, influence this performance, predicting it accurately is difficult.

The project raises concerns about how to determine important variables, collect and analyze relevant data, and select appropriate algorithms. Prediction models are created using decision tree, logistic regression and linear regression. The goal of the study is to identify which of these algorithms is most accurate in predicting students' academic performance on Saber Pro exams.

This matter is important because it has the potential to improve Colombian higher education by enabling institutions to identify students who require the greatest assistance and develop customized programs to improve academic achievement. The research findings can help Colombian government and education sectors by providing a valuable tool for data-driven decision-making. They can also improve student performance on Saber Pro tests.

Keywords: Prediction, Linear Regression, Logistic Regression, Decision tree.

Tabla de contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Glosario

Predicción: Proceso de estimar o pronosticar un resultado futuro o desconocido basado en datos históricos o información disponible en el presente.

Machine Learning (ML): Se ocupa de la creación de modelos y algoritmos que permitan a las computadoras aprender de los datos y realizar tareas particulares sin necesidad de ser programadas explícitamente.

Rendimiento Académico: Medida del éxito de un estudiante en un entorno educativo, generalmente evaluado mediante calificaciones, puntajes en exámenes estandarizados u otros indicadores de logro académico.

Regresión Lineal: Una técnica estadística que supone una relación lineal entre una variable dependiente y una o más variables independientes para simular esa relación.

Regresión Logística: Un método de modelado estadístico que utiliza una o más variables independientes para predecir la probabilidad de un evento binario (como éxito o fracaso, sí o no).

Árbol de Decisión: Método de aprendizaje automático que utiliza una estructura de árbol para representar y dividir el espacio de características en segmentos más pequeños, permitiendo la toma de decisiones basada en reglas lógicas.

Impugnar: Cuestionar o poner en duda la validez, precisión o exactitud de algo, como un resultado, una afirmación o una decisión.

Pruebas Saber Pro: Exámenes estandarizados administrados por el Instituto Colombiano para la Evaluación de la Educación (ICFES) en Colombia, diseñados para evaluar el nivel de competencia y el rendimiento de los estudiantes universitarios en áreas específicas del conocimiento y habilidades.

Introducción

En el dinámico escenario de la educación, la búsqueda continua por mejorar la calidad y el nivel académico de los estudiantes de pregrado ha tenido siempre una importancia constante. En este contexto, el presente proyecto adquiere una relevancia significativa tanto en el ámbito académico como social en Colombia y específicamente en la Universidad Nacional Abierta y a Distancia (UNAD). Abordando el desafío de la predicción del rendimiento académico con las pruebas Saber Pro de sus estudiantes.

La aplicación de modelos de Machine Learning para seleccionar a los estudiantes con bajo rendimiento académico representa un paso crucial hacia la identificación de variables determinantes que ofrecen información valiosa. Esta información, a su vez, posibilita la implementación de medidas de apoyo más personalizadas y oportunas, mejorando de manera sustancial la calidad de vida estudiantil y, por ende, el desempeño en las pruebas Saber Pro.

Diversos estudios realizados por expertos en el campo educativo, como Barrera, Hanushek, Coleman, Romero y Ventura, han arrojado luz sobre la importancia de la educación en el desarrollo individual y colectivo. Estos análisis han demostrado que la calidad educativa es un factor determinante en áreas tan diversas como los salarios, la salud y la innovación.

En el ámbito profesional, la implementación de modelos de Machine Learning abre nuevas puertas para la intervención efectiva en problemáticas actuales y la identificación de oportunidades de mejora para los estudiantes. Esta intervención, basada en datos y análisis predictivos, impacta directamente en la forma de estudio, el aprovechamiento del tiempo libre y el futuro académico de los individuos.

La integración de tecnologías avanzadas como el Machine Learning no solo optimiza los recursos disponibles, sino que también permite una distribución más eficiente de estos, focalizándose en los

grupos estudiantiles que requieren un acompañamiento más cercano. Además, estas prácticas innovadoras en la gestión del rendimiento académico ayudan a las instituciones educativas, como la UNAD, a mantenerse a la vanguardia y competitivas en el ámbito global.

En este sentido, la iniciativa no sólo marca un avance sustancial en nuestro conocimiento de las variables que influyen en el desempeño de los estudiantes, sino que también allana el camino para una educación superior más eficaz e inclusiva que se adapte a las necesidades únicas de cada estudiante.

Planteamiento del problema

En Colombia, las pruebas Saber Pro son un instrumento fundamental para evaluar el rendimiento de los estudiantes de pregrado. La precisión en la predicción del rendimiento académico en estas pruebas es esencial para instituciones educativas y organismos gubernamentales, ya que les permite distinguir áreas de mejora e implementar planes efectivos para el éxito estudiantil.

Sin embargo, predecir el rendimiento en las pruebas Saber Pro es un desafío debido a la multitud de factores que pueden afectar los resultados. Factores como la calidad de la educación previa, el entorno socioeconómico de los estudiantes y sus características personales pueden tener un impacto significativo. Por lo tanto, se plantea la necesidad de utilizar técnicas de Machine Learning para desarrollar modelos de predicción que tomen en cuenta estas variables y brinden una estimación precisa del rendimiento académico.

Justificación

Esta iniciativa aborda la dificultad de pronosticar el rendimiento académico en las pruebas Sabre Pro de sus estudiantes, lo que lleva a incrementar la calidad de la educación superior, haciéndola relevante para el contexto académico y social de la UNAD y Colombia. Al utilizar modelos de ML (Machine Learning) para detectar a los estudiantes que tienen un desempeño académico deficiente, se pueden identificar variables que brindan información pertinente, lo que permite brindar un apoyo más oportuno y personalizado a los estudiantes actuales. Esto mejorará tanto el rendimiento académico en las próximas pruebas Sabre Pro como la calidad de vida estudiantil.

Barrera, F. (2016). Ha realizado diferentes análisis sobre como la educación es de vital importancia para generar mayor nivel y crecimiento, en salarios, salud, innovación, etc.

Hanushek, E. (2010, 2011) y Colemann, J. (1966). Han realizado estudios en colegios y estudios superiores (Universidad) se determinan los factores que más afectan al estudiante, ya sea de manera positiva o negativa

En el marco del desarrollo profesional, al aplicar estos modelos de Machine Learning, se podrá intervenir sobre problemáticas actuales y se aportarán nuevas oportunidades de mejora para los estudiantes, influyendo directamente en su forma de estudio, uso del tiempo libre y mejora en el futuro académico. Tiene como objetivo elevar el nivel académico de los estudiantes, el machine learning permite comprender mejor las variables que influyen en la educación del estudiante, para que las instituciones educativas, como la UNAD, puedan tomar medidas y mejorar su nivel académico. Con un enfoque predictivo, los algoritmos de machine learning podrían detectar variables tempranas para tomar decisiones a tiempo con el fin de apoyar a los estudiantes que presentan bajos rendimientos académicos.

Predecir el desempeño de los estudiantes es una de las aplicaciones más antiguas y útiles de EDM (Minería de datos Educativos), y su objetivo es estimar el valor desconocido del desempeño, conocimiento, puntaje o calificación de los estudiantes a partir de otra información, aspectos o comportamiento de esos estudiantes. (Romero & Ventura, 2013, p. 12).

En el contexto de la implementación de tecnologías de avanzada como Machine Learning, se pueden optimizar recursos para darle un mejor enfoque y aplicarlos a los estudiantes que verdaderamente necesitan el acompañamiento, con el propósito de tomar las mejores decisiones, enfocados en grupos que las necesiten; además, la implementación de tecnologías de avanzada como Machine Learning ayuda a las instituciones educativas, como la UNAD, a mantenerse en competencia globales al adoptar prácticas innovadoras en la gestión del rendimiento académico.

Objetivos

Objetivo general

El propósito de este proyecto es desarrollar un modelo de Machine Learning que realice la predicción sobre el rendimiento académico de los estudiantes en las pruebas Saber Pro en instituciones de educación superior en Colombia. Nuestro principal enfoque será dirigido a los estudiantes de la UNAD, con el objetivo de identificar necesidades de estudiantes actuales, con esto poder ofrecer un apoyo oportuno, mejorando así los resultados de las pruebas futuras.

Objetivos específicos

Procesar la base de datos de las pruebas Saber Pro presentadas por los estudiantes en el año 2022, que fueron publicadas por el ICFES.

Seleccionar algoritmos de Machine Learning adecuados con la meta de predecir e identificar el nivel académico.

Evaluar el modelo de predicción, identificando las variables con mayor peso que influyen en el rendimiento académico.

Desarrollar un sistema de alerta temprana para estudiantes en riesgo de caer en un rendimiento académico bajo.

Marco conceptual

Teniendo en cuenta la problemática de deserción de estudiantes de pregrado en la Universidad Nacional Abierta y a Distancia, se va a tomar el estado del arte haciendo referencia a los siguientes conceptos; ICFES Saber Pro, Rendimiento Académico, Machine Learning, Deserción Estudiantil, para así tener más claridad que la problemática presentada tiene solución utilizando diferente metodología que conlleve a una predicción con altos estándares para poder así actuar frente a la deserción estudiantil.

ICFES Saber Pro

En Colombia se tienen unas pruebas específicas las cuales son utilizadas para realizar un seguimiento del rendimiento académico, estas dependen del tipo de educación en la cual esté cada individuo, en este proyecto aplicado se va a utilizar las pruebas saber pro como base del estudio. Según el Decreto 3963 de 2009: Evaluar el desarrollo de habilidades de los estudiantes en programas que casi han finalizado en instituciones de educación superior; desarrollar indicadores del valor agregado de la educación basados en el nivel de habilidades de quienes se preparan para matricularse en este nivel de educación; y servir como fuente de información necesaria para desarrollar indicadores de evaluación de alta calidad para servicios de educación pública, programas de educación superior e instituciones. Estos indicadores apoyan el desarrollo de políticas, la certificación de prácticas institucionales y el proceso de toma de decisiones en todos los ámbitos, componentes y jerarquías que componen el sistema educativo. (Quintero, 2019, p. 28).

Se puede entender que este tipo de pruebas basadas en criterios generales pueden tener ciertos desaciertos en cuanto a la evaluación del rendimiento académico o posible desempeño profesional de cierto individuo, dado que factores como nervios, ansiedad, bloqueos mentales, etc., pueden salir a flor

de piel al momento del desarrollo de estas pruebas (Madaus, Russell, & Higgins, 2009). Igualmente, con el uso de los modelos de machine learning se podrá visualizar claramente cuáles fueron los factores que tuvieron un impacto directo sobre el rendimiento académico, dado que se podrán evaluar muchas variables que se incluyen en la base de datos del ICFES 2022, con la cual se estará trabajando en este proyecto aplicado, y poder así determinar cuáles son los factores externos a lo solo numérico que afectan a la hora de presentar estas pruebas.

Rendimiento académico

El rendimiento académico se puede ver como un valor numérico que se obtiene al momento de desarrollar una prueba o asignatura, pero también se debe de tener en cuenta otras formas de definirlo dado que también podría ser que el valor aritmético obtenido no sea consecuencia de la cantidad de aprendizaje y desarrollo que obtuvo un individuo. (Leonardo E. Contreras1, 2020)

Deserción Estudiantil

El fenómeno de la deserción estudiantil en el nivel de pregrado en Colombia puede conceptualizarse como el abandono prematuro de los estudios universitarios por parte de los estudiantes matriculados en programas de educación superior. Este fenómeno involucra una serie de factores multidimensionales que pueden influir en la decisión de un estudiante de dejar sus estudios antes de completar su formación académica. Estos factores pueden incluir variables individuales, familiares, académicas, socioeconómicas, institucionales y culturales, que interactúan de manera compleja y afectan la permanencia y éxito de los estudiantes en el sistema educativo. Un marco conceptual para comprender la deserción estudiantil en pregrado en Colombia debe considerar tanto los factores internos y externos que contribuyen a este fenómeno, así como las posibles estrategias y políticas que pueden implementarse para prevenir y abordar esta problemática, fomentando así la retención y graduación de los estudiantes universitarios.

Machine Learning

Se ha observado una amplia gama de metodologías de Machine Learning (ML) aplicadas al campo del rendimiento académico, que incluyen tanto modelos supervisados como no supervisados. Sin embargo, para este estudio en particular, se ha optado por centrarse exclusivamente en modelos supervisados.

Los modelos supervisados son aquellos que aprenden de conjuntos de datos etiquetados, donde ya se conoce la relación entre las variables de entrada y la variable de salida. Esto les permite generar salidas precisas para nuevas entradas, basadas en el aprendizaje de patrones en los datos de entrenamiento.

Para la predicción del rendimiento académico en este estudio, se han seleccionado específicamente tres modelos supervisados:

Regresión Lineal: Este modelo establece una relación lineal entre las variables de entrada y la variable de salida, lo que lo hace adecuado para predecir valores numéricos, como puntajes en pruebas estandarizadas.

Regresión Logística: Aunque su nombre sugiere regresión, la regresión logística se utiliza para problemas de clasificación binaria. En este contexto, podría predecir, por ejemplo, si un estudiante pasa o no una prueba basada en sus características académicas y demográficas.

Arboles de decisiones: Un modelo de aprendizaje automático supervisado para regresión y clasificación se denomina árbol de decisión. Para categorizar o pronosticar la variable objetivo, operan dividiendo continuamente el conjunto de datos en subgrupos más pequeños de acuerdo con criterios particulares.

Al construir un árbol de decisión, cada nodo interno representa una característica (o atributo) del conjunto de datos, cada rama representa el resultado potencial de una característica y cada hoja representa la etiqueta de clasificación o el resultado final de la regresión. Seleccionar las características que mejor dividen las clases objetivo en cada nodo es un paso en el proceso de construcción del árbol.

Esta elección de modelos permite una evaluación exhaustiva de las proyecciones de rendimiento académico. Se anticipa que estos modelos ofrecerán información importante sobre las variables que influyen en el desempeño de los estudiantes en el examen Saber Pro colombiano.

Marco teórico

Ministerio de Educación Nacional de Colombia (2022). Las Pruebas Saber Pro, son evaluaciones estandarizadas aplicadas en Colombia para medir el nivel académico de los estudiantes de pregrado en Colombia. Estas pruebas cubren áreas fundamentales como razonamiento cuantitativo, lectura crítica, competencias ciudadanas y competencias específicas según la carrera.

Importancia del Rendimiento Académico

Leonardo E. Contreras (2020) menciona que el nivel académico de los estudiantes es un indicador crucial de la calidad de la educación y puede influir en diversas áreas, como la empleabilidad, la movilidad social y el desarrollo económico de un país. Por lo tanto, comprender y predecir el rendimiento académico es de gran interés para educadores, instituciones educativas, políticos y empleadores.

Enfoque del Aprendizaje Automático para la Predicción del Rendimiento Académico

El aprendizaje automático ofrece un enfoque sistemático y basado en datos para analizar y predecir el rendimiento académico en las pruebas Saber Pro. Al aprovechar técnicas y algoritmos avanzados, se pueden identificar patrones complejos en los datos que pueden ser difíciles de descubrir mediante métodos tradicionales. Según Gómez y Rodríguez (2020), el uso del aprendizaje automático ha demostrado ser efectivo para predecir el rendimiento académico en estas pruebas, proporcionando una visión más profunda y precisa de los factores que influyen en los resultados de los estudiantes.

Técnicas de Aprendizaje Automático Aplicables

Regresión Lineal y Logística: Estas técnicas son adecuadas para predecir variables continuas (como puntajes en las pruebas) o variables binarias (como el éxito o fracaso en la prueba). (Smith & Jones, 2021, p. 45).

Arboles de decisiones: Los árboles de decisiones ofrecen flexibilidad para modelar relaciones complejas y no lineales entre las variables predictoras y el rendimiento académico.

Factores Predictores del Rendimiento Académico

- Historial académico previo (calificaciones, promedio académico).
- Variables demográficas (edad, género, ubicación geográfica).
- Recursos socioeconómicos (nivel educativo de los padres, ingresos familiares).
- Participación en programas de preparación para exámenes.
- Motivación y actitudes hacia el aprendizaje.
- Uso de tecnología educativa y recursos de aprendizaje.

Evaluación del Modelo

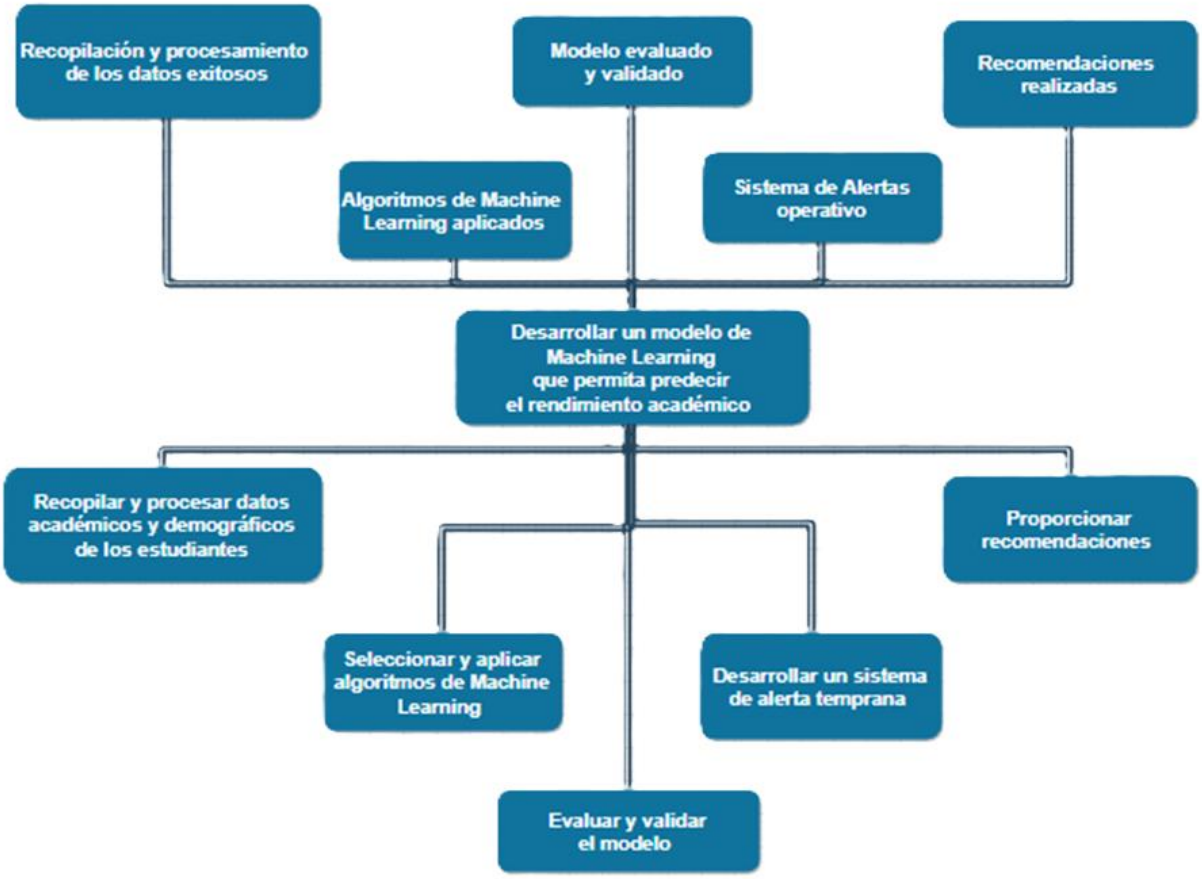
Contreras y Hernandez (2020) explica que la evaluación del modelo de predicción debe realizarse mediante técnicas adecuadas de validación cruzada y métricas de desempeño, como el error cuadrático medio, la precisión, el área bajo la curva ROC (Característica Operativa del Receptor), entre otras, para determinar la eficacia y generalización del modelo.

Consideraciones Éticas y de Privacidad

Es crucial abordar consideraciones éticas y de privacidad al trabajar con datos educativos, garantizando la confidencialidad de la información del estudiante y obteniendo el consentimiento adecuado para el uso de los datos en investigación.

Figura 1.

Técnica de árbol causa y efecto.



Fuente. Elaboración propia.

Metodología

Paso 1. Recopilación de información a partir de los datos:

Adquirir conjuntos de datos históricos de las evaluaciones de Sabre Pro a partir de 2022; estos conjuntos de datos incluyen factores demográficos, socioeconómicos y educativos, junto con información sobre las puntuaciones individuales en cada área evaluada.

Paso 2. Exploración de Datos:

Para el conjunto de datos de prueba Sabre Pro-2022, se utilizó un análisis estadístico descriptivo para comprender la distribución de las variables. Utilizando dos clases de variables (categóricas y numéricas), caracterizaremos los datos obtenidos en este paso. Es importante identificar las variables que pueden usarse para análisis posteriores y la selección de un modelo predictivo o clasificación de algoritmos supervisados, así como la aplicación de algoritmos no supervisados para identificar patrones en los datos.

Paso 3. Elección de las diferentes características:

Utilice técnicas de selección de características para encontrar las variables más importantes que afectan el rendimiento académico. Para elegir las variables correctas, se debe realizar un análisis exploratorio de los datos para encontrar estas variables. Es posible que se descubran variables durante el proceso de exploración. pertinentes que no ofrecen parámetros cruciales para determinar el éxito académico de los estudiantes cuando se modelan utilizando algoritmos de aprendizaje automático.

En las características relevantes de las variables elegidas deben estar presentes los siguientes elementos: factores socioeconómicos y demográficos.

Las variables relacionadas con la educación.

- Variables específicas del programa para su clasificación.

- Los factores de calificación de la prueba Sabre Pro-2022.

Paso 4. Limpieza, transformación, procesamiento de los datos obtenidos:

El paso 4 implica limpiar los datos adquiridos de las selecciones de variables del paso 3. Para limpiar adecuadamente los datos, usaremos Jupyter Notebook y el lenguaje de programación Python. Luego codificaremos los datos en el DataFrame producido en Jupyter Notebook de acuerdo con la siguiente secuencia: nulo, vacío, repetido, sin formato y reemplazable.

Una vez que se ha limpiado el marco de datos en Jupyter, las variables de categoría seleccionadas deben procesarse o cambiarse para prepararlo para análisis adicionales utilizando diferentes modelos de predicción de aprendizaje automático. La biblioteca Sklearn también se utilizará en Jupyter Notebook para realizar la correspondiente transformación o procesamiento de datos categóricos utilizando LabelEncoder o OneHotEncoder de Sklearn.

Paso 5. Desarrollo de Modelos:

Implementación de modelos de Machine Learning, incluyendo Árboles de Decisión, Regresión Lineal, Regresión Logística y arboles de decisiones, utilizando un 80% en datos de entrenamiento y 20% en datos de prueba, para predecir el rendimiento académico y factores clasificatorios que podrían afectar el nivel académico.

Paso 6. Entrenamiento y Evaluación de Modelos:

Entrenamiento de los modelos con los datos definidos en el paso 5 80% y 20%.

Evaluación de la precisión de cada modelo utilizando métricas como el error cuadrático medio (MSE), el coeficiente de determinación (R^2) y la precisión (accuracy), dependiendo del tipo de algoritmo. (recall, F1-score)

Paso 7. Selección y validación del Modelo:

Comparación de la precisión de los modelos y selección del algoritmo que ofrezca el rendimiento más alto, dependiendo especialmente de la precisión (accuracy), coeficiente de determinación y error medio cuadrático.

Luego aplicamos validaciones cruzadas para evaluar la capacidad generalizada de los modelos.

Paso 8 Interpretación de Resultados:

Una vez realizado el análisis de las variables que afectan el rendimiento académico con la modelación del mejor modelo predictivo y clasificatorio de machine learning, se realiza la interpretación de los resultados obtenidos por el algoritmo seleccionado, esto nos lleva a los tres últimos objetivos específicos del proyecto de grado; desarrollar a partir de los resultados obtenidos alertas tempranas para evitar que los estudiantes en riesgo caigan en bajo rendimiento académico, proporcionar recomendaciones específicas para mitigar el bajo rendimiento y evitar la deserción de los estudiantes que se encuentren en zonas de riesgo con bajo rendimiento.

Clasificación, limpieza y análisis de los datos obtenidos.

De acuerdo con los datos obtenidos de las pruebas Saber Pro 2022 correspondiente a los estudiantes de pregrado, es necesario realizar el análisis exploratorio de los datos para así poder escoger que variables tiene importancia para su posterior modelamiento con los algoritmos escogidos, y así poder hacer la predicciones y encontrar las posibles causas futuras del porque existe bajo rendimiento en los estudiantes de la UNAD, para poder enfocar esfuerzos necesarios en puntos clave que puedan nivelar el nivel académico.

Enlace base de datos:

ICFES. (2024). Data ICFES. ICFES. https://www.icfes.gov.co/data-icfes

Tabla 1.

Clasificación de datos obtenidos.

Base de datos	Icfes Saber Profesional genéricas 2022
Columnas	109
Registros	127342

Fuente. Elaboración propia.

Es de importancia definir que la base de datos disponible tiene 127342 registros y 109 columnas o variables en las cuales podemos encontrar los siguientes aspectos importantes:

- Resultados ICFES por asignatura.
- Registros socioeconómicos: departamento, municipio, área donde reside, tiene internet, tiene computador, horas de la semana que trabaja y método como estudia.
- Datos de la institución donde cursa la carrera: Fecha de ingreso, código de institución donde estudia, programa académico, método, departamento de la institución.

Como observamos a continuación en la “tabla 2” tenemos 56 columnas o variables aprovechables de los 109 originales y las cuales fueron clasificada según sus datos en **categóricas o numéricas.**

Tabla 2.

Variables seleccionadas.

Icfes Saber Profesional genéricas 2022	Cantidad
Total, Columnas	109
Aprovechables	56

Fuente. Elaboración propia.

De las variables seleccionadas podemos mencionar las siguientes:

Características personales:

- Género del estudiante: (ESTU_GENERO).
- Fecha de nacimiento del estudiante: (ESTU_FECHANACIMIENTO).
- Indica si el estudiante pertenece a una etnia: (ESTU_TIENEETNIA).
- Tipo de documento del estudiante: (ESTU_TIPODOCUMENTOSB11).
- Semestre que el estudiante está cursando: (ESTU_SEMESTRECURSA).
- Edad del estudiante: (EDAD_ESTUDIANTE).

Características familiares:

- Nivel de educación del padre: (FAMI_EDUCACIONPADRE).
- Nivel de educación de la madre: (FAMI_EDUCACIONMADRE).
- Estrato socioeconómico de la vivienda del estudiante: (FAMI_ESTRATOVIVIENDA).
- Indica si la familia del estudiante tiene internet: (FAMI_TIENEINTERNET).
- Indica si la familia del estudiante tiene computador: (FAMI_TIENECOMPUTADOR).

Características socioeconómicas/geográficas:

- Departamento de residencia del estudiante: (ESTU_DEPTO_RESIDE).
- Código del departamento de residencia del estudiante: (ESTU_COD_RESIDE_DEPTO).
- Municipio de residencia del estudiante: (ESTU_MCPIO_RESIDE).
- Código del municipio de residencia del estudiante: (ESTU_COD_RESIDE_MCPIO).
- Área de residencia del estudiante (urbano/rural): (ESTU_AREARESIDE).
- Título de bachiller obtenido por el estudiante:(ESTU_TITULOOBTENIDOBACHILLER).
- Indica si el estudiante paga la matrícula con una beca: (ESTU_PAGOMATRICULABECA).
- Indica si el estudiante paga la matrícula con un crédito: (ESTU_PAGOMATRICULACREDITO).
- Indica si los padres pagan la matrícula del estudiante: (ESTU_PAGOMATRICULAPADRES).

- Indica si el estudiante paga su propia matrícula: (ESTU_PAGOMATRICULAPROPIO).
- Horas semanales que trabaja el estudiante: (ESTU_HORASSEMANATRABAJA).
- Presentación de la casa del estudiante: (ESTU_PRESENTACIONCASA).
- Código de la institución educativa: (INST_COD_INSTITUCION).
- Nombre de la institución educativa: (INST_NOMBRE_INSTITUCION).
- Programa académico del estudiante: (ESTU_PRGM_ACADEMICO).
- Código SNIES del programa académico: (ESTU_SNIES_PRGMACADEMICO).
- Método del programa académico: (ESTU_METODO_PRGM).
- Municipio de la institución educativa: (ESTU_INST_MUNICIPIO).
- Departamento de la institución educativa: (ESTU_INST_DEPARTAMENTO).
- Origen de la institución (pública/privada): (INST_ORIGEN).

Limpieza de datos

En el filtrado de los datos, en primer lugar, ya definido el nuevo dataframe con las variables descartables nos queda una base de datos con una extensión de 127342 instancias con 56 variables aprovechables, para luego ver cuántos datos nulos hay en ella aplicamos la siguiente función:

Tabla 3.

Tamaño de la nueva base de datos.

Se descartaron datos nulos	icfes_final = icfes_final. dropna ()
Registros totales	RangeIndex: **127342** entries
Registros finales	Index: **107370** entries
Data columns	(total 56 columns)

Fuente. Elaboración propia.

Parámetros y métricas descriptivas de los datos.

En este paso tenemos que seleccionar correctamente qué variables principales vamos a utilizar en los algoritmos supervisados, para realizar la respectiva predicción de los datos de acuerdo con el desempeño académico de los estudiantes; a continuación, realizamos mediante la función “describe” la interpretación de las métricas descriptivas de las variables para así conocer más a fondo desde el punto de vista estadístico cómo su comportamiento:

Tabla 4.

Métricas descriptivas de las variables numéricas.

	Frecuencia	Media	Desviación Estándar	Mínimo	25%	Mediana 50%	75%	Máximo
ESTU_COD_RESIDE_DEPTO	107370	34,51	682,67	5	11	15	52	99999
ESTU_COD_RESIDE_MCPIO	107370	30000,47	25664,91	5001	11001	15790	52001	99999
INST_COD_INSTITUCION	107370	2303,91	1502,15	1101	1218	1818	2744	9930
ESTU_SNIES_PRGMACADEMICO	107370	40474,46	42781,94	1	2692	15422	91236	110700
MOD_RAZONA_CUANTITAT_PUNT	107370	146,54	32,33	0	123	146	169	300
MOD_RAZONA_CUANTITAT_DESEM	107370	2,18	0,89	1	1	2	3	4
MOD_RAZONA_CUANTITATIVO_PNAL	107370	50,94	28,66	1	26	51	76	100
MOD_RAZONA_CUANTITATIVO_PNBC	107370	50,74	28,63	1	26	51	75	100
MOD_LECTURA_CRITICA_PUNT	107370	152,18	30,56	0	130	152	174	300
MOD_LECTURA_CRITICA_DESEM	107370	2,30	0,85	1	2	2	3	4
MOD_LECTURA_CRITICA_PNAL	107370	50,94	28,60	1	26	51	76	100
MOD_LECTURA_CRITICA_PNBC	107370	50,89	28,61	1	26	51	76	100
MOD_COMPETEN_CIUDADA_PUNT	107370	144,94	33,38	0	121	145	169	300
MOD_COMPETEN_CIUDADA_DESEM	107370	2,13	0,89	1	1	2	3	4
MOD_COMPETEN_CIUDADA_PNAL	107370	51,67	28,56	1	27	52	76	100
MOD_COMPETEN_CIUDADA_PNBC	107370	51,71	28,54	1	27	52	76	100

MOD_INGLES_PUNT	107370	159,23	31,52	0	136	157	179	300
MOD_INGLES_PNAL	107370	51,80	28,53	1	27	52	77	100
MOD_INGLES_PNBC	107370	51,66	28,51	1	27	52	76	100
MOD_COMUNI_ESCRITA_PUNT	107370	142,66	32,70	65	122	144	163	300
MOD_COMUNI_ESCRITA_DESEM	107370	2,34	0,78	1	2	2	3	4
MOD_COMUNI_ESCRITA_PNAL	107370	50,25	28,81	1	25	50	75	100
MOD_COMUNI_ESCRITA_PNBC	107370	50,36	28,81	1	25	50	75	100
PUNT_GLOBAL	107370	149,11	24,36	48	132	148	165	278
PERCENTIL_GLOBAL	107370	53,45	27,48	1	30	54	77	100
PERCENTIL_NBC	107370	53,54	27,46	1	30	54	77	100
ESTU_INSE_INDIVIDUAL	107370	54,18	6,50	14,00165	49,65572	54,13272	58,61208	79,59835
ESTU_NSE_INDIVIDUAL	107370	2,56	0,95	1	2	2	3	4
ESTU_NSE_IES	107370	2,64	0,82	2	2	2	3	4

Fuente. Elaboración propia.

En este caso se realizó el análisis solamente en las variables numéricas que contuvieran las clasificación general y puntuación de cada campo en específico evaluado en las pruebas ICFES.

Para lo cual en la siguiente “tabla 5” describimos cada variable elegida para así poder entender mejor cómo están conformada estas variables y en el momento de utilizar el algoritmo nos arroje una mejor predicción de los datos:

Tabla 5.

Elección de las variables numéricas.

Variable numérica	Descripción
PERCENTIL_GLOBAL	Percentil global en que se encuentra el evaluado Rango: 0 - 100
PUNT_GLOBA	Puntaje total obtenido Rango: 0 - 300
PERCENTIL_NBC	Percentil núcleo básico de Conocimiento Rango: 0-100
MOD_COMUNI_ESCRITA_PUNT	Puntaje comunicación escrita Rango:0-300
MOD_INGLES_PUNT	Puntaje inglés Rango:0-300
MOD_COMPETEN_CIUDADA_PUNT	Puntaje competencias ciudadanas Rango:0-300
MOD_LECTURA_CRITICA_PUNT	Puntaje lectura crítica 0 - 300
MOD_RAZONA_CUANTITAT_PUNT	Puntaje razonamiento cuantitativo Rango:0-300

Fuente. Elaboración propia.

Ahora para las variables categóricas vamos a realizar un análisis descriptivo para tener en cuenta qué variable podemos utilizar que nos ayude en la implementación de los modelos.

Tabla 6.

Descripción de las variables categóricas.

ESTU_GENERO	Genero
ESTU_FECHANACIMIENTO	Fecha de nacimiento
ESTU_TIENEETNIA	Etnia
ESTU_DEPTO_RESIDE	Departamento de residencia
ESTU_MCPIO_RESIDE	Municipio de residencia
ESTU_AREARESIDE	Área de residencia
ESTU_TITULOOBTENIDOBACHILLER	Título de bachillerato obtenido
ESTU_PAGOMATRICULABECA	Pago de matrícula por beca
ESTU_PAGOMATRICULACREDITO	Pago de la matricula mediante crédito
ESTU_PAGOMATRICULAPADRES	Pago de la matricula mediante padres
ESTU_PAGOMATRICULAPROPIO	Pago de la matricula recursos propios
ESTU_SEMESTRECURSA	Semestre que cursa actualmente el estudiante
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de su vivienda según el recibo de energía eléctrica
FAMI_TIENEINTERNET	Su hogar cuenta con conexión a internet
FAMI_TIENECOMPUTADOR	Su hogar cuenta con computador
ESTU_HORASSEMANATRABAJA	Cuantas horas trabaja a la semana

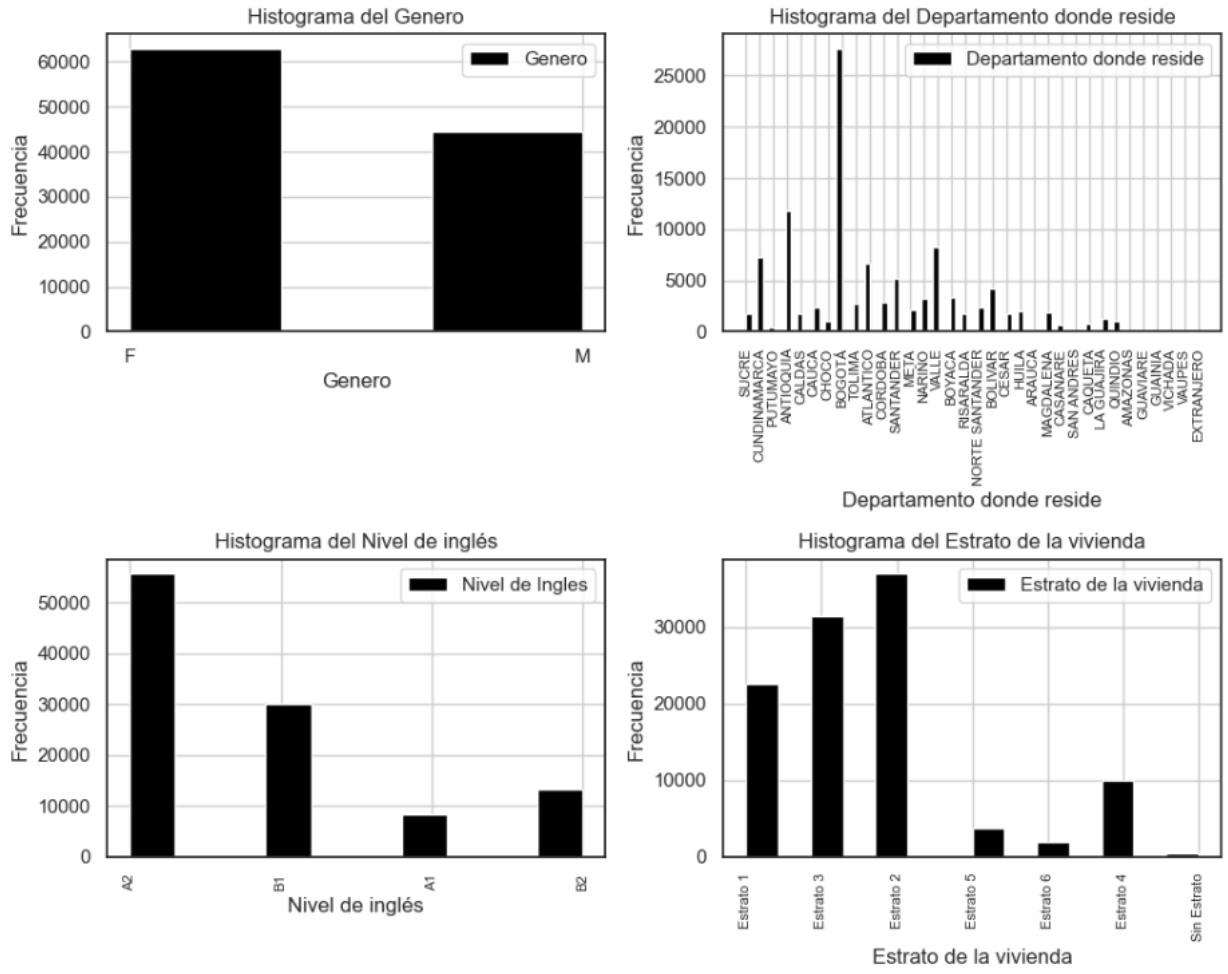
ESTU_PRESENTACIONCASA	Presento la prueba en casa
INST_NOMBRE_INSTITUCION	Nombre de la institución
ESTU_PRGM_ACADEMICO	Nombre del programa académico que cursa
ESTU_METODO_PRGM	Método del programa académico
ESTU_INST_MUNICIPIO	Municipio de la institución
ESTU_INST_DEPARTAMENTO	Departamento de la institución
INST_ORIGEN	Naturaleza u origen de la IES
MOD_INGLES_DESEM	Nivel de ingles

Fuente. Elaboración propia.

Es importante resaltar que las variables categóricas son importantes para nuestro análisis de predicción por tal motivo a continuación realizaremos dicho análisis mediante histogramas para observar el comportamiento de algunas variables categóricas de importancia:

Figura 2.

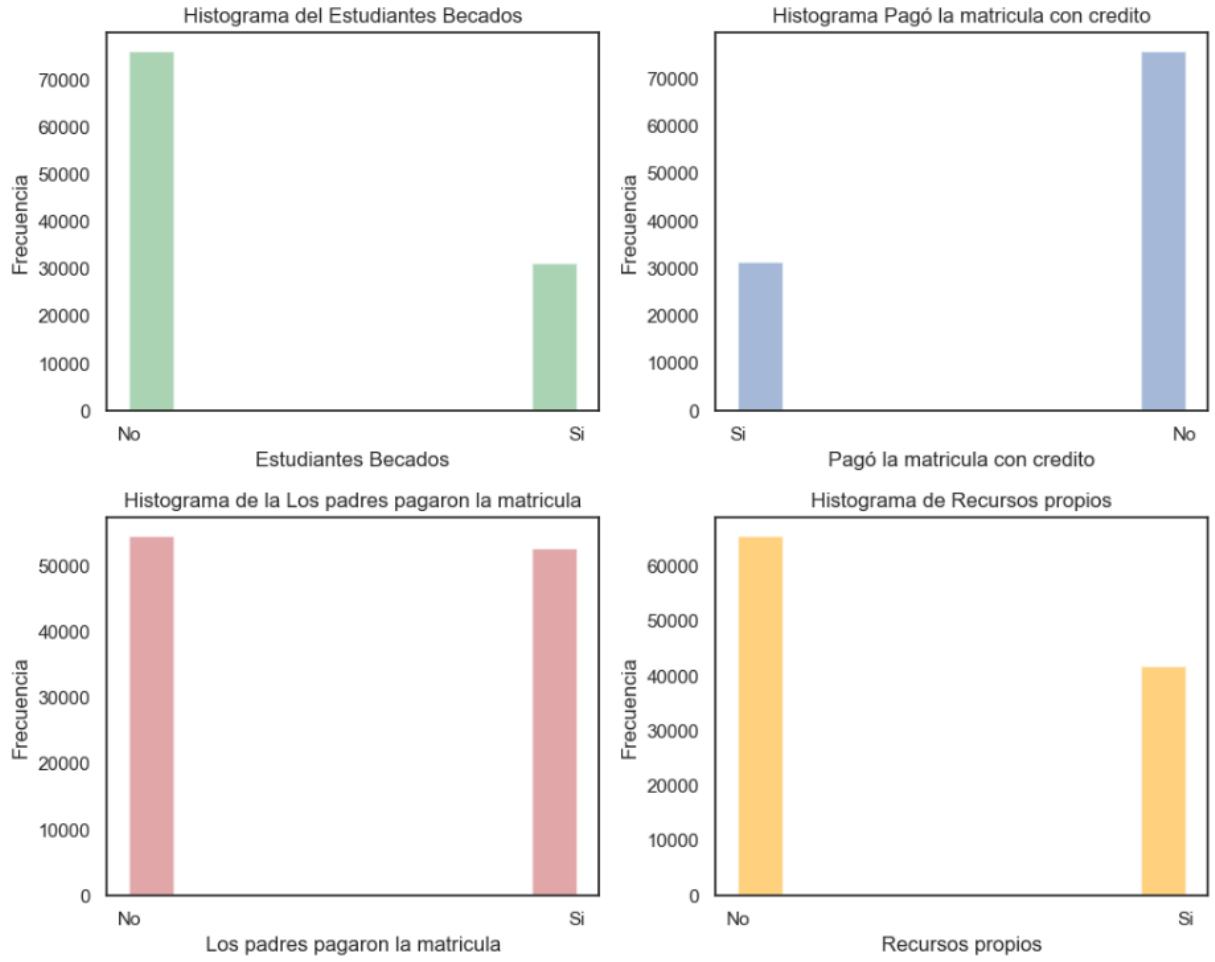
Perfiles de los estudiantes.



Fuente. Elaboración propia.

Figura 3.

Método de pago de la matricula.

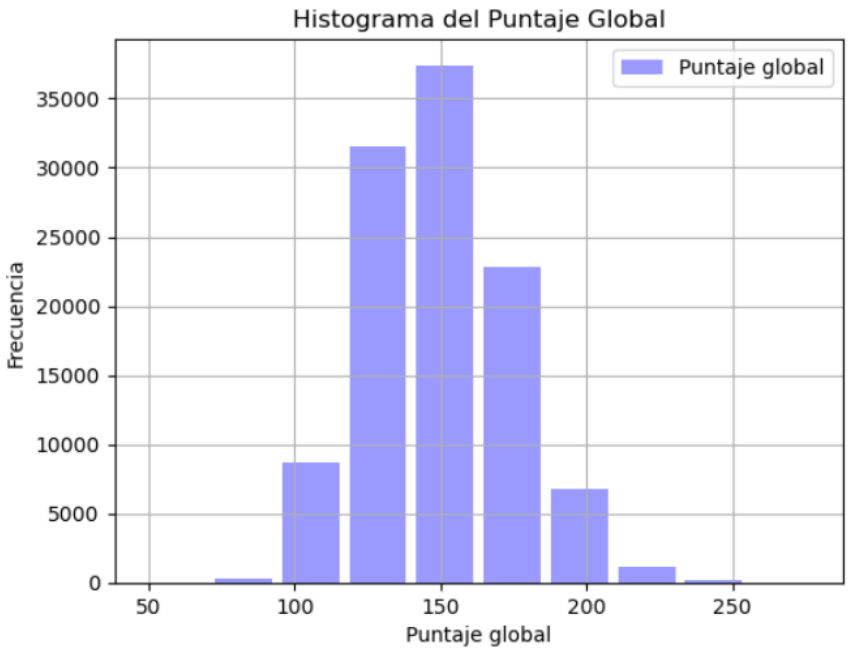


Fuente. Elaboración propia.

Para un análisis más completo de las variables socioeconómicas, demográficas y los diferentes puntajes en las áreas del conocimiento evaluado en la Prueba saber Pro-2022, categorizar y tomar como referencia la variable puntaje global, ya que está es el resultado de las pruebas y con la cual podemos tomar como variable de salida para modelar los diferentes algoritmos supervisado-escogidos:

Figura 4.

Histograma puntaje global.



Fuente. Elaboración propia.

El histograma nos muestra que una gran cantidad de estudiantes se encuentran con un puntaje de 150, distribuidos de la siguiente manera: entre 0 y 100 la frecuencia de los estudiantes llega hasta 10000, luego entre 100 y 200 es donde se encuentran la mayor cantidad de estudiantes con ese puntaje, y de 200 en adelante encontramos un mínimo de estudiantes con ese rango de puntuación.

En cuanto a la prueba Saber Pro y TyT, esta posee los mismos 5 ejes que comprende la prueba saber 11, además de una prueba específica acorde a la carrera profesional de cada estudiante.

Las competencias básicas: lectura, razonamiento cuantitativo, competencias ciudadanas, cuentan con 30 preguntas e inglés con 45, en cuanto a comunicación escrita el puntaje se obtiene de la evaluación del cumplimiento de las especificaciones del tema asignado.

El puntaje de la prueba se da en un rango de 0 a 300 puntos, ahora bien, un puntaje bueno se encuentra entre 145 a 149, pero un resultado sobresaliente se encuentra entre 165 a 180 puntos.

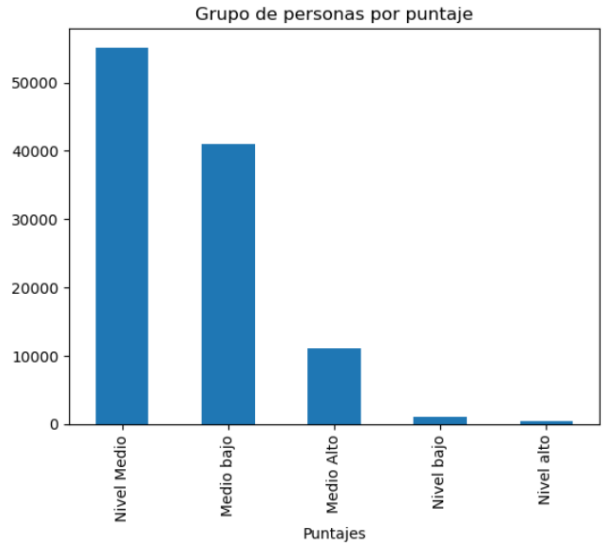
Ahora bien, también se puede considerar bueno un puntaje que esté en el rango del promedio nacional, en la prueba 11 se puede observar que, en los últimos 5 años, los puntajes oscilaron entre 247 a 252 puntos.

En cuanto a la prueba Saber Pro, el promedio de los puntajes de los últimos 5 años tienen un rango de entre 145 a 156 puntos.

Según lo mencionado anteriormente propongo esas etiquetas ('Nivel Bajo, Medio Bajo, Nivel Medio, Medio Alto y Nivel Alto'), para poder agrupar a los estudiantes según su puntaje en las pruebas saber pro, esto me serviría para realizar el análisis clasificatorio en cualquier modelo de algoritmo de machine learning, para luego así poder analizar que estudiantes tendrán un bajo rendimiento y sus posibles causas.

Figura 5.

Grupo de estudiantes por puntaje global.



Fuente. Elaboración propia.

Analizando este diagrama de barras nos damos cuenta que el nivel académico demostrado en esta prueba saber pro 2022 en los estudiantes de educación superior es nivel medio, más de 60 mil

estudiantes están en este nivel con una puntuación de 140 a 180 del puntaje global, en el nivel medio bajo se encuentran 50000 estudiantes con un puntaje de 100 a 140, en el nivel medio alto se están más de 10000 estudiante con un puntaje entre 180 a 220, en el nivel bajo tenemos alrededor de 5000 estudiante con un puntaje mejor a 100, y por ultimo existen una pequeña cantidad de estudiantes con puntaje alto mayor a 220.

Figura 6.

Diagrama de flujo limpieza, análisis y exploración de los datos.

Fuente. Elaboración propia.

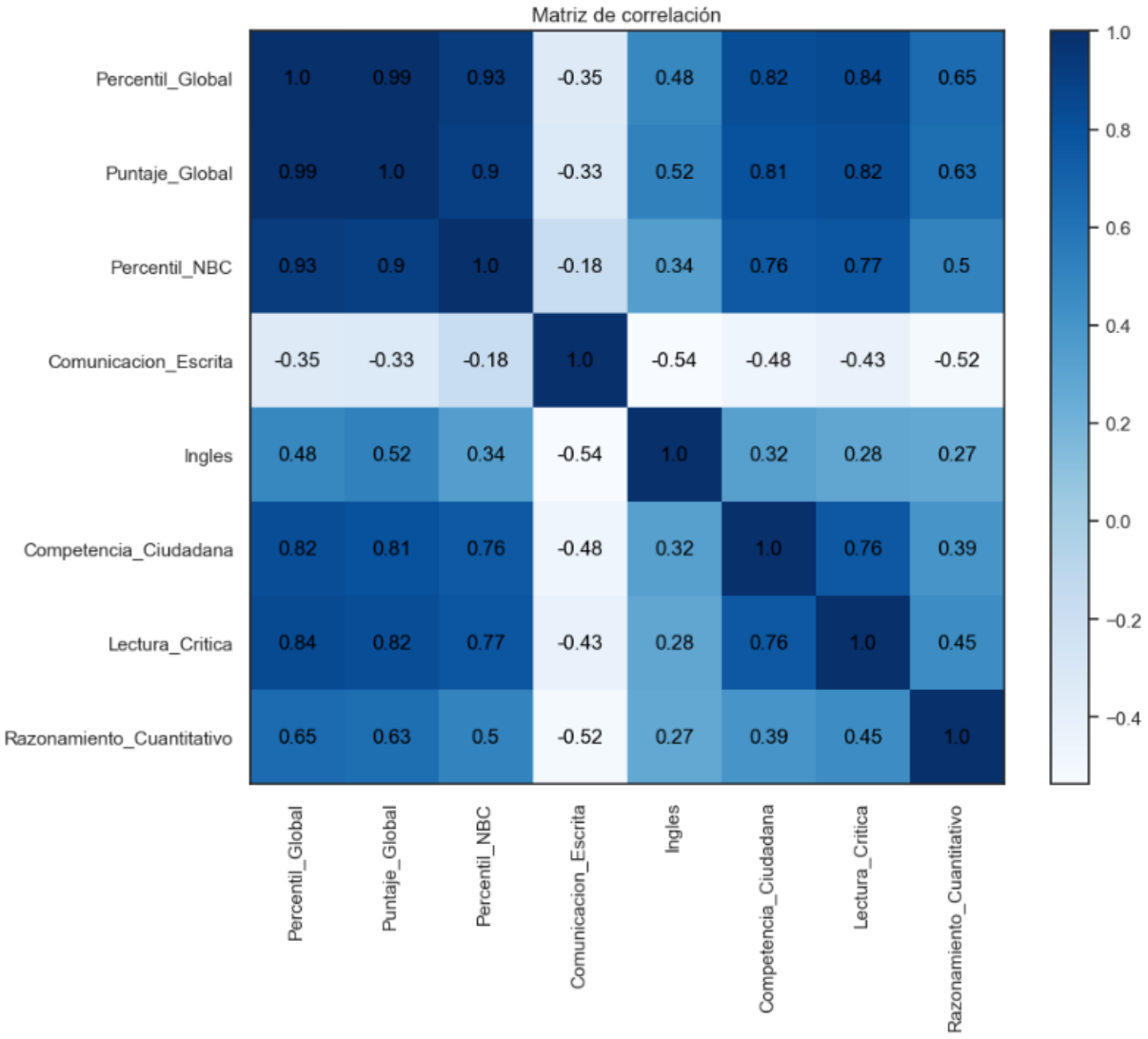
Para pronosticar el nivel académico de los estudiantes universitarios, utilizamos un modelo de regresión lineal múltiple.

En el paso 4 de la metodología del proyecto aplicado, se refiere a la limpieza, transformación y procesamiento de los datos obtenidos de repositorios pruebas Saber Pro 2022, en este caso como ya realizamos la limpieza necesaria e interpretación de los datos, seguimos con la transformación y procesamiento de los datos; una de las herramientas utilizada fue OneHotEncoder de la librería de Sklearn para la variable dependiente Puntaje global y para las variables independientes utilizamos get.dumies para el modelamiento supervisado de Regresión Lineal múltiple, a continuación observamos los diferentes pasos realizados.

De acuerdo con la siguiente matriz de correlación realizaremos la selección de las variables numéricas que más tenga correlación para así poder modelarla en el modelo de RL múltiple:

Figura 7.

Matriz de correlación variables numéricas.



Fuente. Elaboración propia.

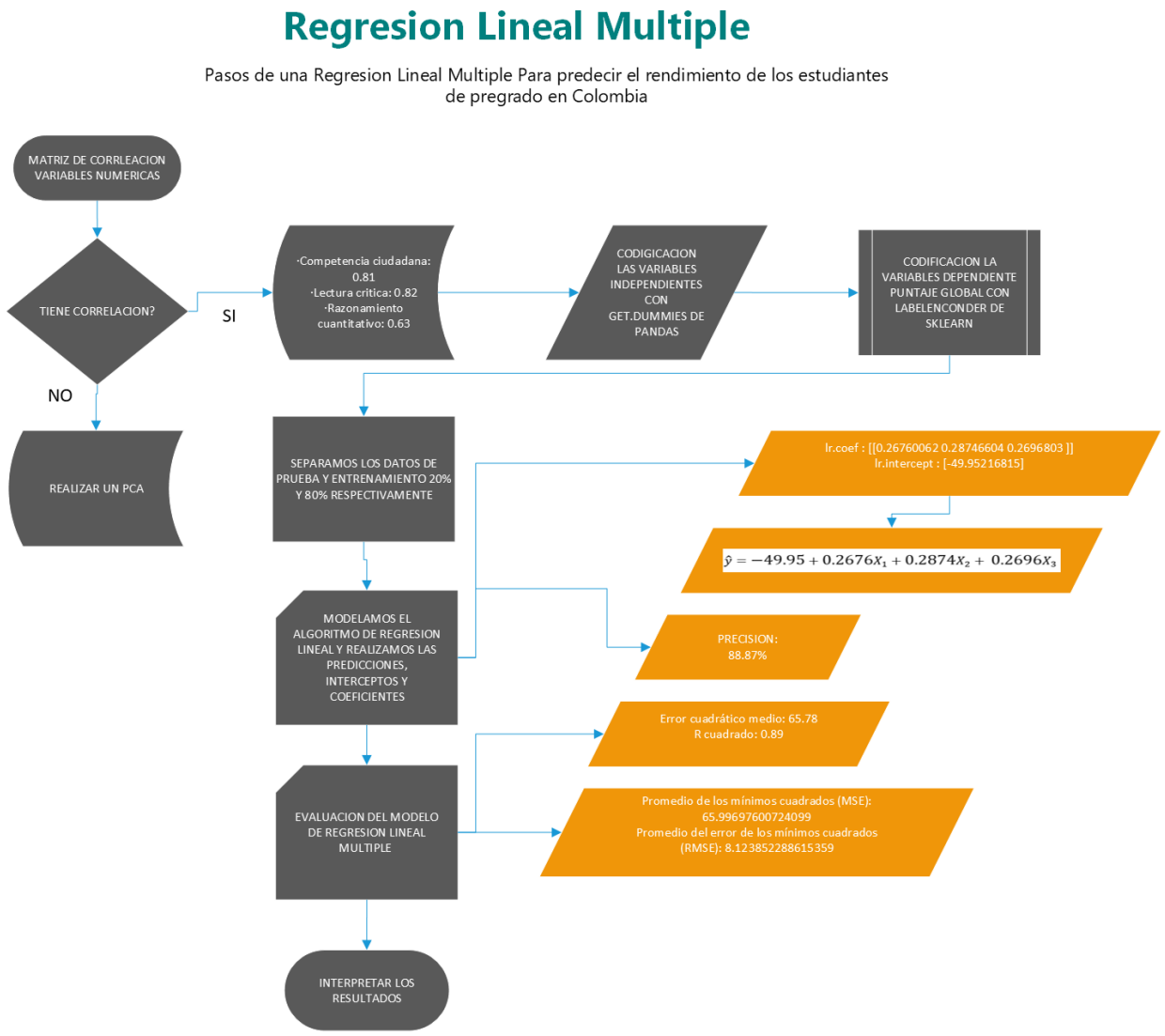
Observamos que las variables seleccionadas según su relación con la variable puntaje global son:

- Competencia ciudadana: 0.81
- Lectura critica: 0.82
- Razonamiento cuantitativo: 0.63

Ahora procedemos a codificar dichas variables para luego realizar el modelamiento de los datos:

Figura 8.

Diagrama de flujo del modelo en Python de REGRESION LINEAL MULTIPLE



Fuente. Elaboración propia.

$$\hat{y} = -49.95 + 0.2676X_1 + 0.2874X_2 + 0.2696X_3$$

Donde $\hat{y}$: $Variable\ de\ salida$

$X_1,\ X_2,\ X_3$: $Coeficientes\ de\ regresión$

En la ilustración 8 podemos analizar diferentes aspectos; se hallaron los coeficientes de las variables con respecto al puntaje global, esto nos indica que las pendientes de cada variable independiente se puede observar en este resultado, al analizar estos datos vemos algunas pendientes positivas y otras negativas un ejemplo de ello es si analizamos la primera pendiente 0.26 nos podría indicar que si esta pendiente aumenta y las otras quedan igual, aumentaría en la misma proporción respecto al puntaje global.

Con respecto al intercepto en este caso arrojó un valor de -49,95 lo que significa que cuando todas las variables independientes sean 0 el puntaje global va a tomar el valor de -49.95.

La precisión del modelo de regresión lineal múltiple en este caso fue del 88.87%, este porcentaje de precisión nos podría predecir la variable puntaje global a partir de las variables independientes, competencias ciudadanas, lectura crítica y razonamiento cuantitativo en gran medida, por tal motivo estas variables tienen una significancia alta en estas predicciones.

El coeficiente de determinación R^2, arrojo un valor de 0.89 los cual nos indica que el modelo tiene una precisión del 89%, este es un valor moderadamente bueno para realizar las predicciones de variabilidad para el puntaje global, aunque seguiremos realizando la evaluación con otros modelos supervisados.

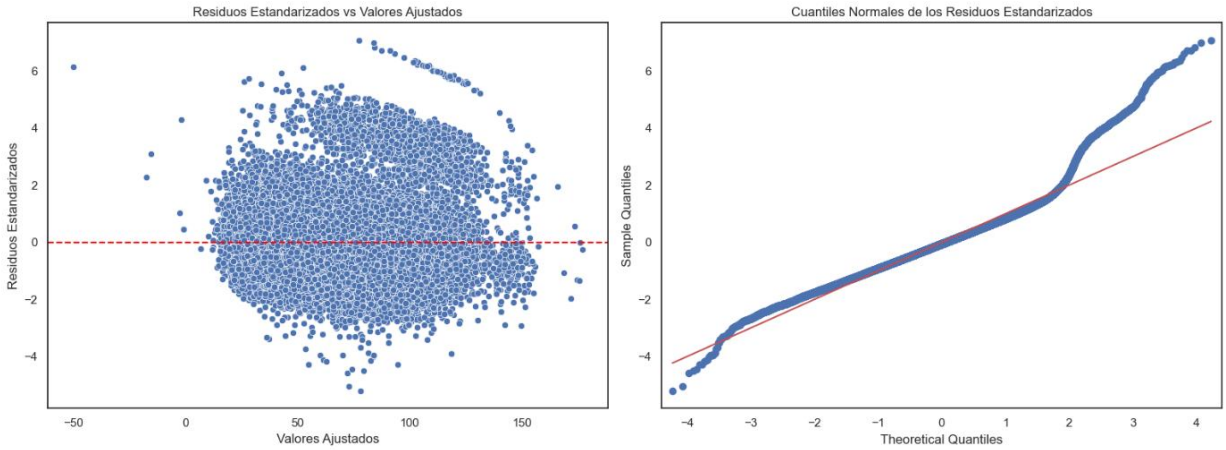
- Promedio de los mínimos cuadrados (MSE): 65.99697600724099
- Promedio del error de los mínimos cuadrados (RMSE): 8.123852288615359

En cuanto al promedio de los mínimos cuadrados (MSE) y el promedio del error de los mínimos cuadrados (RMSE), tiene una alta relación ya que con estos valore podemos hallar la diferencia promedio entre los valore predichos y los valores reales, en este caso un RMSE de 8.12 para el tamaño de los datos nos arroja un error relativo promedio muy bajo alrededor de 0.038% . Aunque estas cifras son útiles para evaluar el rendimiento del modelo, es importante compararlas con otras métricas y

considerar el contexto específico del problema para una evaluación más completa. En general, cuanto más cercanos a cero sean estos valores, mejor será el rendimiento del modelo.

Figura 9.

Grafica de Residuos Estandarizados Vs Valores ajustados del modelo de regresión lineal y grafica de cuantiles normales del modelo de Regresión Logística.



Fuente. Elaboración propia.

En la figura 9 nos muestra el grafico de dispersión de los Residuos estandarizados Vs valores ajustados, observamos que existe homocedasticidad en los datos con algunos valores atípicos observables, pero en su mayoría los datos se encuentran centrados hacia la línea roja del medio, lo cual nos dicta que el modelo tiene una buena capacidad predictiva, en la otra grafica de la figura 9 correspondiente a los cuantiles normales de los residuos estandarizados, en los extremos de la gráfica datos atípicos, pero en general existe una linealidad de los residuos estandarizados lo cual nos indica que existe una relación directa entre los residuos estandarizados y los cuantiles normales del modelo de regresión lineal múltiple.

Modelo Regresión Logística Múltiple para clasificar y predecir el nivel académico de los estudiantes.

En este caso vamos a realizar la clasificación de los estudiantes por el puntaje global, el puntaje de la prueba se da en un rango de 0 a 300 puntos, ahora bien, un puntaje nivel medio se encuentra entre 145 a 149.

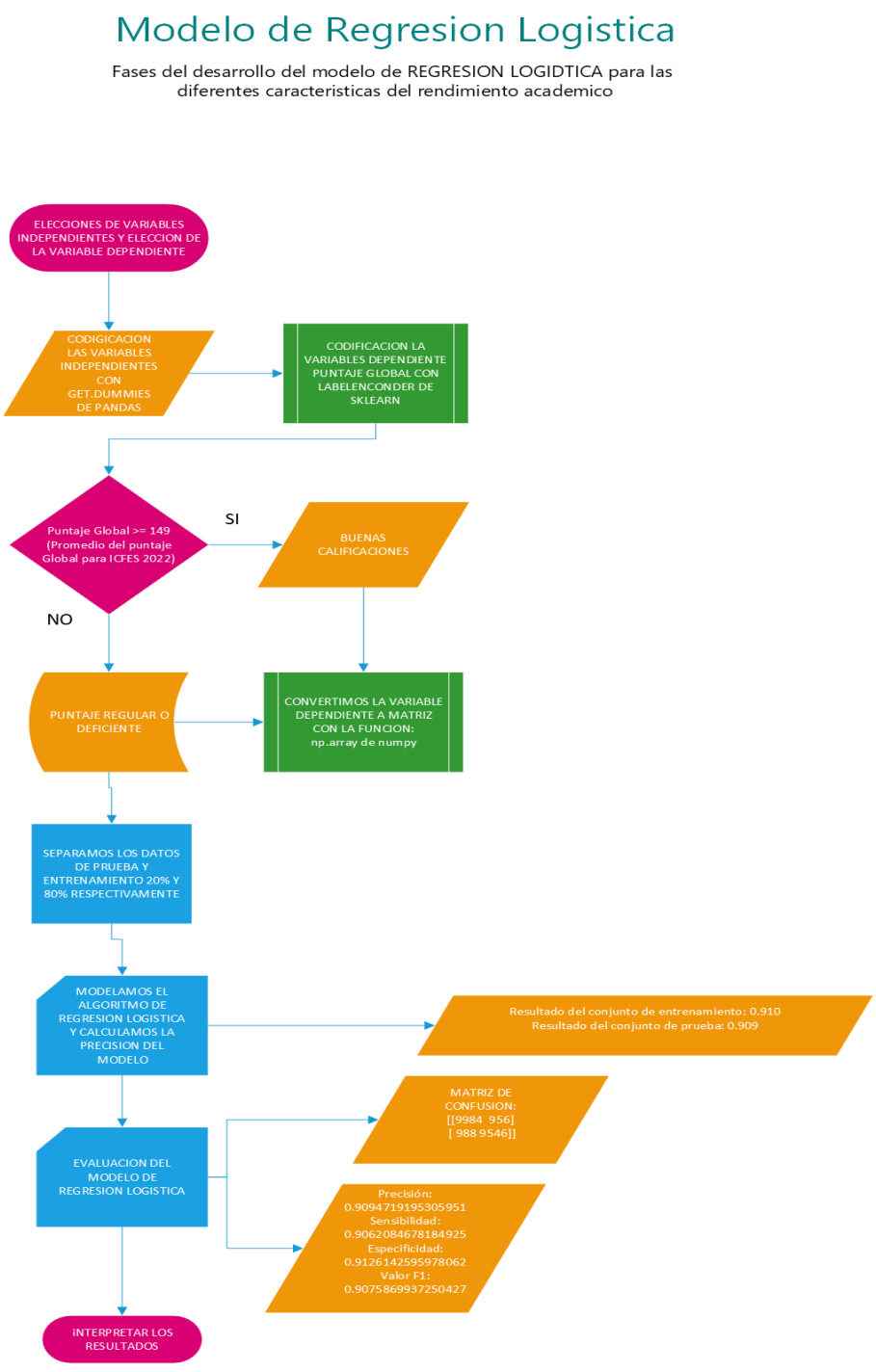
Con lo anterior nuestra variable dummie sería el puntaje global tomando como referencia el puntaje nivel medio en adelante, para nuestro caso 149 hacia arriba seria 1 y por lo contrario 149 hacia abajo 0. Se elige punto de partida el puntaje 149 ya que se encuentra dentro del nivel medio y además es el promedio de las calificaciones de los estudiantes que presentaron la pruebas Saber Pro.

El paso a seguir es evaluar el modelo de regresión logística para ver qué tan bueno se comporta haciendo predicciones en base a los valores introducidos, en general con este modelo lo que se busca es que de acuerdo a las variable independiente usadas en el modelos de regresión lineal múltiple que incluye variables socioeconómicas, demográficas, sean clasificadas por 1 y 0 para ver en predicción como será el rendimiento académico de acuerdo a la variable dependiente puntaje global, esto lo evaluaremos con una matriz de confusión.

Con la ayuda de estos resultados, podemos demostrar que la regresión logística es un modelo útil para predecir el rendimiento académico de los estudiantes de pregrado con base en las pruebas Sabre Pro-2022. La precisión del 90% del modelo en los datos de prueba y del 91% en el conjunto de datos de entrenamiento indican que no está sobrevalorando los datos de entrenamiento y está generalizando efectivamente nuevos datos.

Figura 10.

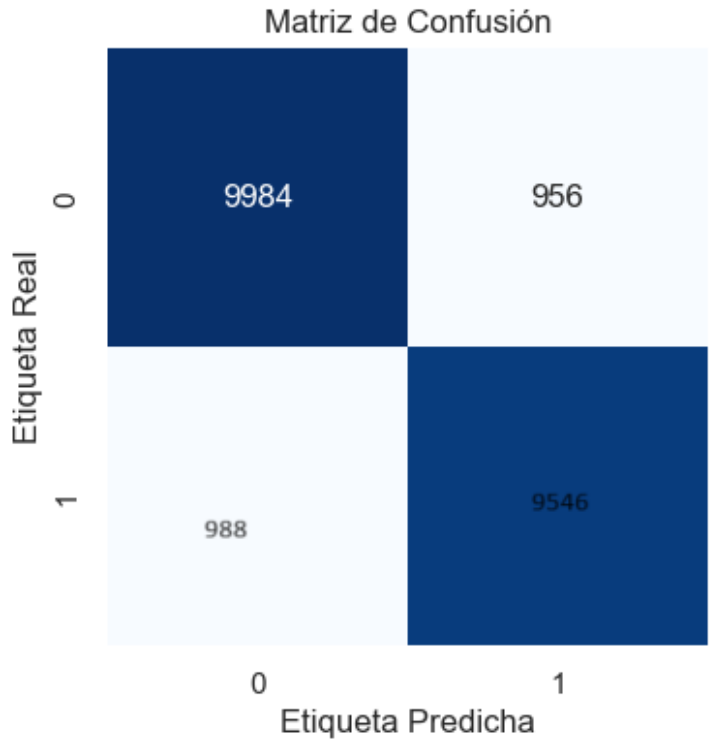
Diagrama de flujo del modelo en Python de RL.



Fuente. Elaboración propia.

Figura 11.

Matriz de confusión del modelo de Regresión Logística.



Fuente. Elaboración propia.

Aquí está la interpretación de esta matriz:

- Verdaderos positivos (TP): 9984
- Falsos positivos (FP): 956
- Falsos negativos (FN): 988
- Verdaderos negativos (TN): 9546

Verdaderos negativos TN: en este caso los verdaderos negativos son unos del número más alto que tenemos en la matriz de confusión, estos pertenecen a número de estudiantes que superaron el nivel medio escogido, esto nos indica que en las predicciones existen 9546 estudiantes que obtuvieron más del puntaje promedio y son predicciones correctas.

Falsos positivos FP: en este caso los falsos positivos son los que en la matriz de confusión dieron como estudiante con puntaje por debajo del promedio, pero son predicciones falsas, este equivale a 956 estudiantes.

Falsos negativos FN: en este caso los falsos negativos son los estudiantes que superaron el puntaje promedio según el modelo pero que la predicción fue equivocada, este valor equivale a 988 estudiantes, pero fue una predicción falsa.

Verdaderos positivos TP: en este caso los verdaderos positivos son la cantidad de estudiantes en donde el modelo predice correctamente que están por debajo del puntaje promedio, para nuestro caso es de 9984 estudiantes según el modelo.

Ahora vamos a ver las métricas del modelo de regresión logística:

- **Precisión:** 0.9094719195305951
- **Sensibilidad:** 0.9062084678184925
- **Especificidad:** 0.9126142595978062
- **Valor F1:** 0.9075869937250427

Análisis de las métricas calculadas del modelo de regresión logística basándonos en la matriz de confusión:

Precisión: En este modelo obtuvimos una precisión de 0.9094, lo cual significa que nuestro modelo tiene una capacidad de explicar el 90.94% de los datos observados, es un valor alto por encima del 80% que es lo ideal.

Sensibilidad: tasa positiva verdadera, en este caso fue del 0.9062, es decir que el modelo puede explicar el 90.62% de los casos positivos verdaderos del modelo.

Especificidad: La especificidad corresponde a los casos negativos verdadero, 0.9126, analizando más a fondo por que la especificidad de nuestro modelo arrojo un resultado del 91.26% observamos que los casos negativos verdaderos son una de las proporciones más elevadas que el resto de los datos observados.

F1 score: El F1score nos indica la relación que existe entre la precisión del modelo y la sensibilidad es útil cuando hay desequilibrio entre las clases predichas, en nuestro caso fue del 90.95%.

Modelo de Árbol de decisión para buscar características del rendimiento académico.

Con este modelo de clasificación y de regresión buscamos principalmente encontrar las características que afectan la variable objetivo, en este caso el puntaje global, para ello hemos seleccionado mediante una serie de pruebas de significancia de variable las más importantes que puede afectar o no la variable dependiente. Con la elección y clasificación de estas variables buscamos realizar la predicción del rendimiento de los estudiantes de pregrado, encontrando así el camino hacia la optimización de recursos y una educación más inclusiva con capacidad de ser equitativa para los estudiantes.

Después de discretizar la variable dependiente puntaje global teniendo en cuenta dos rangos nivel bajo y nivel medio tomando como punto de referencia el promedio del puntaje (149), realizamos las tablas de contingencia de cada variable con respecto al puntaje global y luego aplicamos la prueba de chi cuadrado y así verificar y elegir las variables con mayor significancia dando como resultado los siguientes valores:

Género (Genero):

Chi-cuadrado: 697.86

P-valor: 8.73e-154

Existe una asociación significativa entre el género del estudiante y su puntaje global en las pruebas ICFES. El bajo valor de p-valor indica que la diferencia en los puntajes no es probablemente debido al azar.

Estrato:

Chi-cuadrado: 173.86

P-valor: 5.19e-36

Hay una asociación significativa entre el estrato socioeconómico del estudiante y su puntaje en las pruebas. Los estudiantes de diferentes estratos pueden tener puntajes distintos, y esta diferencia no es probablemente aleatoria.

Tipo de bachillerato:

Chi-cuadrado: 110.95

P-valor: 8.07e-25

Hay una relación significativa entre el tipo de bachillerato cursado por el estudiante y su rendimiento en las pruebas. Los estudiantes que cursaron diferentes tipos de bachillerato pueden tener puntajes diferentes de manera no aleatoria.

Nivel de inglés:

Chi-cuadrado: 37407.02

P-valor: 5.15e-07

Existe una relación altamente significativa entre el nivel de desempeño en inglés y el puntaje global en las pruebas ICFES. Esto sugiere que los estudiantes con diferentes niveles de competencia en inglés tienen puntajes diferentes de manera muy significativa.

Estudios del padre:

Chi-cuadrado: 6929.35

P-valor: 8.40e-05

Hay una asociación significativa entre el nivel educativo del padre y el puntaje global del estudiante en las pruebas. Esto indica que el nivel educativo del padre puede influir en el desempeño académico del estudiante.

Estudios de la madre:

Chi-cuadrado: 7686.38

P-valor: 4.30e-09

Similar al nivel educativo del padre, el nivel educativo de la madre también muestra una asociación significativa con el puntaje global del estudiante en las pruebas. Esto sugiere que el nivel educativo de la madre puede ser un factor importante en el rendimiento académico del estudiante.

Variables demográficas y socioeconómicas para la modelación del árbol de decisión son:

- Genero.
- Estrato.

- Tipo de bachiller.
- Nivel de inglés.
- Estudios de padre.
- Estudios de la madre.

Como estas variables son categóricas es necesario llevarlas a variables de entrada numéricas para su procesamiento en el algoritmo para ello utilizamos la función preprocessing. LabelEcoder (), esta función de la librería de Sklearn nos permite convertir estas variables a numéricas según su asignación.

Cabe mencionar que las características nuevas de las variables fueron convertidas y agrupadas en una vector para su respectivo procesamiento.

Con respecto a la variable objetivo en este caso puntaje global, también es procesada mediante cierto tipo de condiciones como se mencionó anteriormente, esa variable fue clasificada de acuerdo con el promedio de la calificación de los estudiantes:

- Puntaje global <= 149 es igual a 0, bajas calificaciones
- Puntaje global > 149 es igual a 1, buenas calificaciones

Para los datos de entrenamiento y prueba se utilizó la relación 80% entrenamiento y 20% prueba.

Utilizando de la librería de Sklearn DecisionTreeClassifier, después de un preprocesamiento de los datos y un correcto modelamiento del algoritmo los arroja la siguiente precisión:

- ❖ Precisión del modelo de entrenamiento: 0.777
- ❖ Precisión del modelo de prueba: 0.772

La precisión reportada de 0.777 para los datos de entrenamiento y 0.772 para los datos de prueba indica la proporción de predicciones correctas del modelo de árbol de decisión en ambos conjuntos de datos.

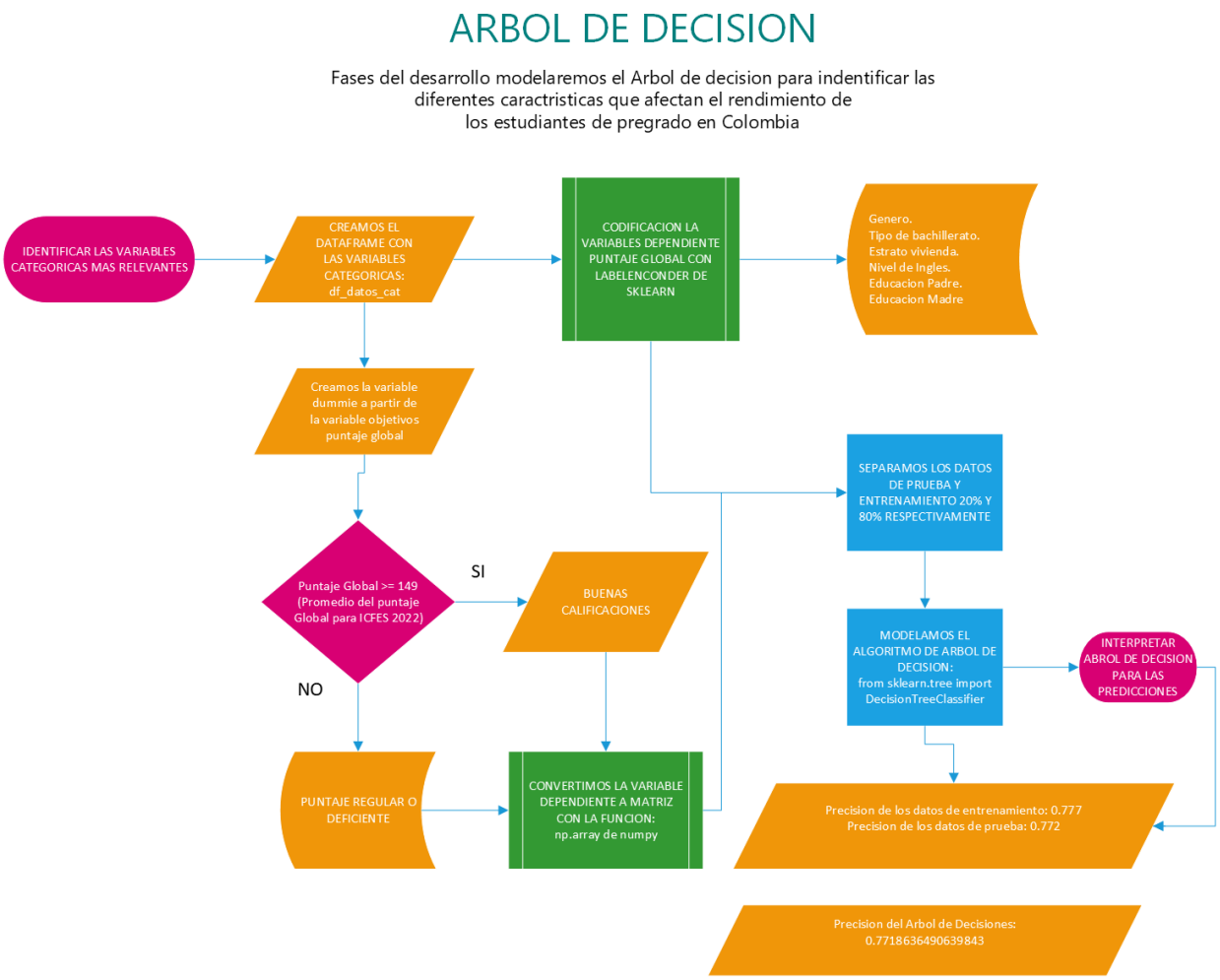
Una tasa de precisión del 77,7 % en los datos de entrenamiento implica que aproximadamente el 77,7 % de los casos en el conjunto de datos de entrenamiento fueron clasificados adecuadamente por el modelo.

Mientras tanto, el modelo clasificó con precisión aproximadamente el 77,2% de los casos en el conjunto de datos de prueba, de acuerdo con su precisión del 77,2% en los datos de prueba. Dado que

la precisión de los datos de prueba es comparable a la precisión de los datos de entrenamiento, esto indica que el modelo se generaliza bien a datos nuevos y desconocidos.

Figura 12.

Diagrama de Flujo del modelo en Python del ARBOL DE DECISION.

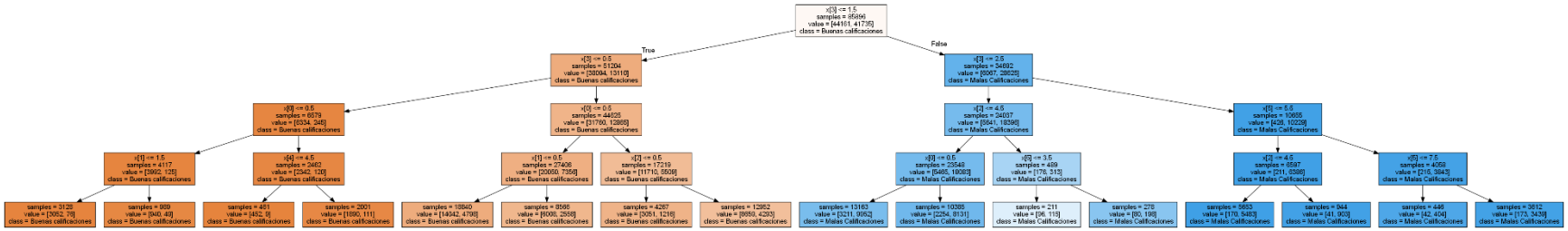


Fuente. Elaboración propia.

Visualización del árbol de decisión:

Figura 13.

Árbol de decisión que define características del rendimiento académico.



Fuente. Elaboración propia.

Identificación de las variables independientes o características dentro del árbol de decisión:

- X0 = Genero
- X1 = Tipo de bachiller.
- X2 = Estrato
- X3 = Nivel de inglés.
- X4 = Estudio del padre.
- X5 = Estudio de la madre.

Observamos que los primeros clasificadores y predictores de los nodos del árbol de decisión es la característica nivel de inglés, esta tiene un rango de clasificación <= 1.5 esto equivale que si el estudiante tiene un nivel de A2 y es verdadero pasaría al siguiente nodo donde estaría en las clases de malas calificaciones, en el siguiente nodo analizamos que la siguiente característica clasificadora es el género si es femenina y su nivel de inglés es bajo entonces es probable que tenga malas calificaciones, en el siguiente nodo observamos que la características predominantes son nivel de estudios de los padres, estrato y tipo de bachiller, en los cuales obtenemos resultados clasificatorios como: si el estudiante tiene nivel de inglés A1, es bachiller técnico y es femenina, obtendrá bajas calificaciones.

Por el contrario, si es falsa la afirmación de nivel de inglés A2, seria nivel de inglés B1, el estudiante fuera de un nivel socioeconómico estrato 4, de género femenino, obtuvieron en su mayoría buenas calificaciones.

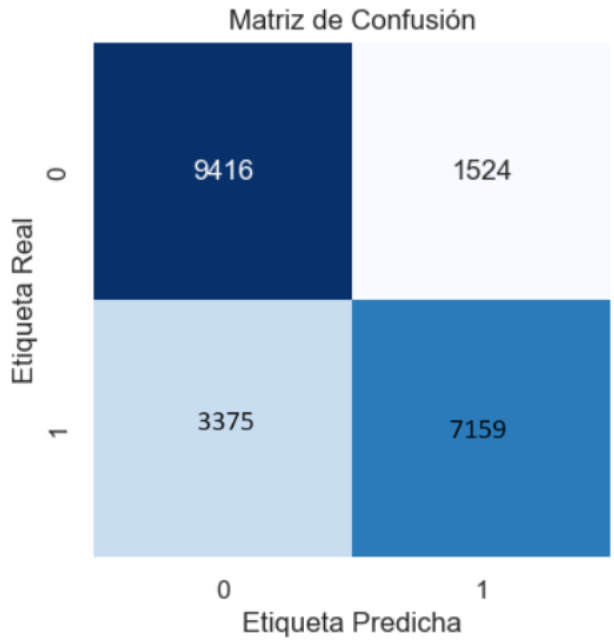
En otro caso si su estrato fuera por debajo del 4, el nivel de estudio de su madre fuera educación profesional completa estaría en buenas calificaciones, pero no tan numeroso como el anterior.

Por último, tendiendo en cuenta los estudios de la madre donde observamos que, si la madre posee un nivel académico de postgrado, y tiene estrato socioeconómico alto tiene altas probabilidades de obtener altas calificaciones y por último el mayor grado de calificación para los estudiantes tiene como condición que la madre tenga un nivel académico hasta maestría.

Matriz de confusión para el modelo de Árbol de decisión y sus respectivas métricas:

Figura 14.

Matriz de confusión para el modelo de Árbol de Decisión.



Fuente. Elaboración propia.

Aquí está la interpretación de esta matriz de confusión para el Árbol de Decisión:

- Verdaderos positivos (TP): 9416
- Falsos positivos (FP): 1524
- Falsos negativos (FN): 3375
- Verdaderos negativos (TN): 7169

Verdaderos negativos TN: en este caso los verdaderos negativos son unos del número más alto que tenemos en la matriz de confusión, estos pertenecen a número de estudiantes que superaron el nivel medio escogido, esto nos indica que en las predicciones existen 7169 estudiantes que obtuvieron más del puntaje promedio y son predicciones correctas, dependiendo de las características de las variables introducidas en el modelo.

Falsos positivos FP: en este caso los falsos positivos son los que en la matriz de confusión dieron como estudiante con puntaje por debajo del promedio, pero son predicciones falsas, este equivale a 1524 estudiantes.

Falsos negativos FN: en este caso los falsos negativos son los estudiantes que superaron el puntaje promedio según el modelo pero que la predicción fue equivocada, este valor equivale a 3375 estudiantes, pero fue una predicción falsa.

Verdaderos positivos TP: en este caso los verdaderos positivos son la cantidad de estudiantes en donde el modelo predice correctamente que están por debajo del puntaje promedio, para nuestro caso es de 9416 estudiantes según el modelo.

Ahora vamos a ver las métricas del modelo de regresión logística:

- Precisión: 0.7718636490639843
- Precisión: 0.7794784897095466
- Sensibilidad: 0.7718636490639843
- F1 Score: 0.7697737234001659

Análisis de las métricas calculadas del modelo árbol de decisión basándonos en la matriz de co nfusión:

Precisión: En este modelo obtuvimos una precisión de 0.7718, lo cual significa que nuestro mo delo tiene una capacidad de explicar el 77.18% de los datos observados.

Sensibilidad: tasa positiva verdadera, en este caso fue del 0.7794, es decir que el modelo pue de explicar el 77.94% de los casos positivos verdaderos del modelo.

Especificidad: La especificidad corresponde a los casos negativos verdadero, 0.7718, analizand o más a fondo por que la especificidad de nuestro modelo arrojo un resultado del 77.18% observamos q ue los casos negativos verdaderos son una de las proporciones más elevadas que el resto de los datos ob servados.

F1 score: El F1score nos indica la relación que existe entre la precisión del modelo y la sensibili dad es útil cuando hay desequilibrio entre las clases predichas, en nuestro caso fue del 79.69%.

Sistema de alerta temprana (SAT) para estudiantes en riesgo de caer en un rendimiento académico bajo.

Podemos generar recomendaciones sobre el puntaje obtenido por estudiantes de pregrado de universidades colombianas en función de su preparación mediante el uso del análisis predictivo de regresión lineal múltiple, que tiene una precisión de predicción del 88,89% de los datos observados. Este podría estar enfocado a la preparación de las pruebas de saber pro, enfocándose principalmente en la preparación de habilidades ciudadanas, lectura crítica y razonamiento cuantitativo, reforzando áreas como la lectura en los estudiantes, procurar que en las áreas impartidas de sus carreras especificas pueda tener la facilidad de leer un libro relacionado con sus competencias, con respecto a las habilidades ciudadanas en las universidad se podría reforzar los derechos y deberes de cada ciudadano, maneras de actuar y ofrecer como lectiva cursos como la cultura ciudadana y por último el razonamiento cuántico, es de saber que en el transcurso de las carreras de pregrado se refuerza en gran nivel esta medida, se podría dictar cursos intersemestrales para aquellos estudiantes que tengan falencias en el campo de los cálculos y físicas para así tener presente estos conocimientos a la hora de enfrentarse a las pruebas saber pro y al salir al campo laboral.

Esta alerta temprana sugiere que en los estudiantes con menor preparación en estos campos o materias tendrían que enfocar esfuerzos y profundizar ya sea en lectura critica, razonamiento cuantitativo o habilidades ciudadanas para tener una mayor preparación y así poder obtener mejor nivel académico.

El umbral escogido para el modelamiento es del 149 en el puntaje global, este umbral defina cuando un estudiante podría estar en riesgo de bajo rendimiento, de allí podríamos aplicar la ecuación que nos arroja estos algoritmos para discriminar cuál específicamente es el campo para reforzar.

Siguiendo con el umbral definido en la variable objetico puntaje global, con la regresión logística podríamos clasificar qué estudiantes estaría en riesgo de acuerdo con las diferentes características arrojadas por el modelo.

Este algoritmo nos arroja la mayor precisión de los tres modelos, pero la elección de un solo modelo para predecir el rendimiento académico no sería conveniente, ya que cada modelo aporta información valiosa para desarrollar una alerta temprana, en este caso el modelo de regresión logística nos dio una precisión del 90.94%, y analizando la matriz de confusión la variable objetivo tiene en la

diagonal de verdaderos los valores o instancia con mayor número y equilibrado, por este motivo las métricas del modelo son altas y confiables por encima del 80% y tiene buena capacidad para predecir.

El modelo de árbol de decisión nos logra mostrar a partir de las variable categóricas seleccionadas para el modelo una visión demográfica y socioeconómica para la clasificación predictiva de los niveles de calificación de los estudiantes de pregrado, características como género, estrato socioeconómico, tipo de bachillerado, nivel de inglés, estudio de los padres nos darían una perspectiva de cómo serán sus calificaciones en las pruebas saber pro, con esta información se podría realizar un sondeo, a los estudiantes con estas características y enfocar esfuerzos sobre esta población identificada, este modelo nos arroja una precisión del 71.18%.

Conclusiones

Con énfasis en los estudiantes de la UNAD en particular, hemos desarrollado un eficiente SAT para identificar los factores que pueden generar un bajo nivel académico en las pruebas Saber Pro en instituciones de educación superior en Colombia. Este sistema se basa en el análisis de modelos de regresión lineal múltiple, regresión logística y árboles de decisión, así como de los objetivos específicos del proyecto.

A través del modelo de regresión lineal múltiple, hemos podido predecir con una precisión del 88.89% el rendimiento académico de los estudiantes en las pruebas Saber Pro. Esto nos ha permitido identificar áreas específicas que pueden influir directamente en el puntaje global como variable dependiente, estas variables independientes (lectura critica, razonamiento cuantitativo y competencia ciudadana) afectan directamente los resultados de esta prueba y se puede tomar como referencia al momento de predecir el rendimiento académico de los estudiantes ya sea reforzando un área en específico o las tres simultáneamente.

La regresión logística, con una precisión del 90.94%, nos ha brindado una herramienta valiosa para clasificar a los estudiantes en riesgo de bajo rendimiento académico. Al analizar las características específicas de los estudiantes, como el género, el estrato socioeconómico y el nivel de educación de los padres, podemos identificar con precisión aquellos que pueden necesitar intervención y apoyo adicionales.

Por último, el árbol de decisiones nos ha proporcionado una visión demográfica y socioeconómica única para la clasificación predictiva del rendimiento académico de los estudiantes. A través de la identificación de características clave, como el género, el estrato socioeconómico y el nivel de inglés, podemos dirigir nuestros esfuerzos de apoyo hacia las poblaciones específicas que enfrentan un mayor riesgo de bajo rendimiento académico.

En conjunto, estos modelos y análisis nos han permitido desarrollar un sistema de alerta temprana integral que puede identificar de manera proactiva a los estudiantes en riesgo y proporcionarles el apoyo necesario para mejorar su rendimiento académico. Al implementar este sistema, esperamos no solo mejorar los resultados de las pruebas Saber Pro, sino también brindar a los estudiantes una base sólida para su éxito académico y profesional futuro.

Recomendaciones

Evaluación del modelo

Se deben utilizar métodos de validación cruzada para evaluar el rendimiento de los modelos, ya que esto puede ayudar a evitar el sobreajuste.

Métricas de evaluación: al evaluar estos modelos, utilice métricas de evaluación como precisión más sensibilidad y especificidad; otros incluyen AUC-ROC y puntuación F1.

Configurando Técnicamente el Sistema SAT

Prepárate tecnológicamente: procura que tu infraestructura técnica esté lista para el soporte que debe brindarle al SAT; Es necesario contar con servidores, bases de datos y plataformas de integración.

Interfaz de usuario: para garantizar que los educadores puedan recibir alertas e informes fácilmente, la interfaz de usuario debe ser lo más fácil de usar posible.

Consideraciones éticas y de privacidad

Protección de datos: implementar medidas estrictas para garantizar que la privacidad y confidencialidad de los datos de los estudiantes estén protegidas a toda costa.

Transparencia de datos del SAT: que se sepa cómo se utilizan y qué objetivos se supone que se deben alcanzar.

Referencias

Contreras, L. E. (2020). Rendimiento académico: Conceptualización, medición y factores de influencia. Revista Innovación y Desarrollo, 13(1), 67-78.

Contreras, L. E. (2020). Rendimiento académico: Definición y concepto. Revista de Investigación en Educación, 22-35.

Contreras, L. E., & Hernández, J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. SciELO.

García, R., & José, L. (2003). Cómo elaborar un proyecto de investigación. Instituto Mexicano de Contadores Públicos. (pp. 23-49). Recuperado de https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=318111&lang=es&site=ehost-live&ebv=EB&ppid=pp_Cover

Hernández, R., & Mendoza, C. (2018). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta (pp. 42-64). Recuperado de https://www-ebooks7-24-com.bibliotecavirtual.unad.edu.co/?il=6443&pg=42

Madaus, G. F., Russell, M. K., & Higgins, J. W. (2009). The paradoxes of high-stakes testing. Journal of Education Policy, 149-160.

Madaus, G., Russell, M., & Higgins, J. (2009). The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society. Information Age Publishing.

Madaus, S., Russell, M., & Higgins, J. (2009). Limitations and criticisms of standardized testing. Journal of Educational Measurement, 46, 304-319.

Ministerio de Educación Nacional de Colombia. (2022). Fenómeno de la deserción estudiantil en el nivel de pregrado en Colombia. Recuperado de https://www.mineducacion.gov.co/

Quincho, R., Cárdenas, J., Inga, V., Bada, W., Espinoza, G., & Yangali, H. (2022). Metodología de la investigación científica: Técnicas e instrumentos de investigación. Editorial Maipue. (pp. 7-60). Recuperado de https://editorial.inudi.edu.pe/index.php/editorialinudi/catalog/view/90/133/157

Quintero, J. (2019). Decreto 3963 de 2009: Por el cual se reglamenta el artículo 54 de la Ley 1324 de 2009, en lo relacionado con la evaluación del desempeño institucional de la educación superior. Diario Oficial, No. 47.518, 7 de diciembre de 2009.

Quintero, J. (2019). Pruebas Saber Pro en Colombia: Objetivos y funciones. Revista Internacional de Investigación en Educación, 45-58.

Quintero, L. N. (2019). Estudio del desempeño académico de estudiantes colombianos en las pruebas Saber 11 y Saber Pro para la elección vocacional y permanencia universitaria. Facultad de Ingeniería en Tecnologías de Información y Comunicación, 28-29.

Smith, A. B., & Jones, C. D. (2021). Machine learning applications in academic performance prediction: A review. Journal of Educational Data Mining, 45-63.

Apéndices

Apéndice A

Exploración de los datos.

Base de datos	Icfes Saber Profesional genéricas 2022
Columnas	109
Registros	127342
	listado
1	**ESTU_TIPODOCUMENTO**
2	**ESTU_NACIONALIDAD**
3	**ESTU_GENERO**
4	**ESTU_FECHANACIMIENTO**
5	**ESTU_EXTERIOR**
6	**PERIODO**
7	**ESTU_CONSECUTIVO**
8	**ESTU_ESTUDIANTE**
9	**ESTU_PAIS_RESIDE**
10	**ESTU_TIENEETNIA**
11	**ESTU_DEPTO_RESIDE**
12	**ESTU_COD_RESIDE_DEPTO**
13	**ESTU_MCPIO_RESIDE**
14	**ESTU_COD_RESIDE_MCPIO**
15	**ESTU_AREARESIDE**
16	**ESTU_ESTADOCIVIL**
17	**ESTU_COLE_TERMINO**
18	**ESTU_CODDANE_COLE_TERMINO**
19	**ESTU_COD_COLE_MCPIO_TERMINO**
20	**ESTU_OTROCOLE_TERMINO**
21	**ESTU_TITULOOBTENIDOBACHILLER**
22	**ESTU_VALORMATRICULAUNIVERSIDAD**
23	**ESTU_PAGOMATRICULABECA**
24	**ESTU_PAGOMATRICULACREDITO**
25	**ESTU_PAGOMATRICULAPADRES**
26	**ESTU_PAGOMATRICULAPROPIO**
27	**ESTU_COMOCAPACITOEXAMENSB11**
28	**ESTU_CURSODOCENTESIES**
29	**ESTU_CURSOIESAPOYOEXTERNO**
30	**ESTU_CURSOIESEXTERNA**
31	**ESTU_SIMULACROTIPOICFES**
32	**ESTU_ACTIVIDADREFUERZOAREAS**
33	**ESTU_ACTIVIDADREFUERZOGENERIC**
34	**ESTU_TIPODOCUMENTOSB11**
35	**ESTU_SEMESTRECURSA**
36	**FAMI_EDUCACIONPADRE**
37	**FAMI_EDUCACIONMADRE**
38	**FAMI_OCUPACIONPADRE**

39	**FAMI_OCUPACIONMADRE**
40	**FAMI_ESTRATOVIVIENDA**
41	**FAMI_TIENEINTERNET**
42	**FAMI_TIENECOMPUTADOR**
43	**FAMI_TIENELAVADORA**
44	**FAMI_TIENEHORNOMICROOGAS**
45	**FAMI_TIENESERVICIOTV**
46	**FAMI_TIENEAUTOMOVIL**
47	**FAMI_TIENEMOTOCICLETA**
48	**FAMI_TIENECONSOLAVIDEOJUEGOS**
49	**FAMI_TRABAJOLABORPADRE**
50	**FAMI_TRABAJOLABORMADRE**
51	**ESTU_PLANTELEDUCATIVO**
52	**ESTU_OTROPLANTEL**
53	**ESTU_PREGRADOEXAMENSP**
54	**ESTU_UNPOSTGRADO**
55	**ESTU_HORASSEMANATRABAJA**
56	**FAMI_CUANTOSCOMPARTEBAÑO**
57	**ESTU_PAGOMATRICULA**
58	**ESTU_VLRULTIMOSEMESCURSADO**
59	**ESTU_PRESENTACIONCASA**
60	**ESTU_PAISDOCUMENTOS11**
61	**ESTU_PRESENTACIONSABADO**
62	**INST_COD_INSTITUCION**
63	**INST_NOMBRE_INSTITUCION**
64	**ESTU_PRGM_ACADEMICO**
65	**ESTU_SNIES_PRGMACADEMICO**
66	**GRUPOREFERENCIA**
67	**ESTU_PRGM_CODMUNICIPIO**
68	**ESTU_PRGM_MUNICIPIO**
69	**ESTU_PRGM_DEPARTAMENTO**
70	**ESTU_NIVEL_PRGM_ACADEMICO**
71	**ESTU_METODO_PRGM**
72	**ESTU_NUCLEO_PREGRADO**
73	**ESTU_INST_CODMUNICIPIO**
74	**ESTU_INST_MUNICIPIO**
75	**ESTU_INST_DEPARTAMENTO**
76	**INST_CARACTER_ACADEMICO**
77	**INST_ORIGEN**
78	**ESTU_PRIVADO_LIBERTAD**
79	**ESTU_COD_MCPIO_PRESENTACION**
80	**ESTU_MCPIO_PRESENTACION**
81	**ESTU_DEPTO_PRESENTACION**
82	**ESTU_COD_DEPTO_PRESENTACION**
83	**MOD_RAZONA_CUANTITAT_PUNT**
84	**MOD_RAZONA_CUANTITAT_DESEM**
85	**MOD_RAZONA_CUANTITATIVO_PNAL**
86	**MOD_RAZONA_CUANTITATIVO_PNBC**
87	**MOD_LECTURA_CRITICA_PUNT**

88	**MOD_LECTURA_CRITICA_DESEM**
89	**MOD_LECTURA_CRITICA_PNAL**
90	**MOD_LECTURA_CRITICA_PNBC**
91	**MOD_COMPETEN_CIUDADA_PUNT**
92	**MOD_COMPETEN_CIUDADA_DESEM**
93	**MOD_COMPETEN_CIUDADA_PNAL**
94	**MOD_COMPETEN_CIUDADA_PNBC**
95	**MOD_INGLES_PUNT**
96	**MOD_INGLES_DESEM**
97	**MOD_INGLES_PNAL**
98	**MOD_INGLES_PNBC**
99	**MOD_COMUNI_ESCRITA_PUNT**
100	**MOD_COMUNI_ESCRITA_DESEM**
101	**MOD_COMUNI_ESCRITA_PNAL**
102	**MOD_COMUNI_ESCRITA_PNBC**
103	**PUNT_GLOBAL**
104	**PERCENTIL_GLOBAL**
105	**PERCENTIL_NBC**
106	**ESTU_INSE_INDIVIDUAL**
107	**ESTU_NSE_INDIVIDUAL**
108	**ESTU_NSE_IES**
109	**ESTU_ESTADOINVESTIGACION**

Fuente. Elaboración propia.

Apéndice B

Caracterización y clasificación.

Icfes Saber Profesional genéricas 2022	**Cantidad**
Total, Columnas	109
Aprovechables	53
Caracterizar los datos obtenidos.	
Variables	**Clasificación**
ESTU_TIPODOCUMENTO	Categórica
ESTU_NACIONALIDAD	Categórica
ESTU_GENERO	Categórica
ESTU_FECHANACIMIENTO	Categórica
ESTU_EXTERIOR	Categórica
PERIODO	Categórica
ESTU_CONSECUTIVO	Categórica
ESTU_ESTUDIANTE	Categórica
ESTU_PAIS_RESIDE	Categórica
ESTU_TIENEETNIA	Categórica
ESTU_DEPTO_RESIDE	Categórica
ESTU_COD_RESIDE_DEPTO	Numérica
ESTU_MCPIO_RESIDE	Categórica
ESTU_COD_RESIDE_MCPIO	Numérica

ESTU_AREARESIDE	Categórica
ESTU_ESTADOCIVIL	Categórica
ESTU_COLE_TERMINO	Categórica
ESTU_CODDANE_COLE_TERMINO	Numérica
ESTU_COD_COLE_MCPIO_TERMINO	Numérica
ESTU_OTROCOLE_TERMINO	Categórica
ESTU_TITULOOBTENIDOBACHILLER	Categórica
ESTU_VALORMATRICULAUNIVERSIDAD	Categórica
ESTU_PAGOMATRICULABECA	Categórica
ESTU_PAGOMATRICULACREDITO	Categórica
ESTU_PAGOMATRICULAPADRES	Categórica
ESTU_PAGOMATRICULAPROPIO	Categórica
ESTU_COMOCAPACITOEXAMENSB11	Categórica
ESTU_CURSODOCENTESIES	Categórica
ESTU_CURSOIESAPOYOEXTERNO	Categórica
ESTU_CURSOIESEXTERNA	Categórica
ESTU_SIMULACROTIPOICFES	Categórica
ESTU_ACTIVIDADREFUERZOAREAS	Categórica
ESTU_ACTIVIDADREFUERZOGENERIC	Categórica
ESTU_TIPODOCUMENTOSB11	Categórica
ESTU_SEMESTRECURSA	Numérica
FAMI_EDUCACIONPADRE	Categórica
FAMI_EDUCACIONMADRE	Categórica
FAMI_OCUPACIONPADRE	Categórica
FAMI_OCUPACIONMADRE	Categórica
FAMI_ESTRATOVIVIENDA	Numérica
FAMI_TIENEINTERNET	Categórica
FAMI_TIENECOMPUTADOR	Categórica
FAMI_TIENELAVADORA	Categórica
FAMI_TIENEHORNOMICROOGAS	Categórica
FAMI_TIENESERVICIOTV	Categórica
FAMI_TIENEAUTOMOVIL	Categórica
FAMI_TIENEMOTOCICLETA	Categórica
FAMI_TIENECONSOLAVIDEOJUEGOS	Categórica
FAMI_TRABAJOLABORPADRE	Categórica
FAMI_TRABAJOLABORMADRE	Categórica
ESTU_PLANTELEDUCATIVO	Categórica
ESTU_OTROPLANTEL	Numérica
ESTU_PREGRADOEXAMENSP	NaN
ESTU_UNPOSTGRADO	NaN
ESTU_HORASSEMANATRABAJA	Numérica
FAMI_CUANTOSCOMPARTEBAÑO	Numérica
ESTU_PAGOMATRICULA	Categórica
ESTU_VLRULTIMOSEMESCURSADO	Categórica
ESTU_PRESENTACIONCASA	Categórica
ESTU_PAISDOCUMENTOS11	Categórica
ESTU_PRESENTACIONSABADO	Categórica
INST_COD_INSTITUCION	Numérica
INST_NOMBRE_INSTITUCION	Categórica

ESTU_PRGM_ACADEMICO	Categórica
ESTU_SNIES_PRGMACADEMICO	Numérica
GRUPOREFERENCIA	Categórica
ESTU_PRGM_CODMUNICIPIO	Numérica
ESTU_PRGM_MUNICIPIO	Categórica
ESTU_PRGM_DEPARTAMENTO	Categórica
ESTU_NIVEL_PRGM_ACADEMICO	Categórica
ESTU_METODO_PRGM	Categórica
ESTU_NUCLEO_PREGRADO	Categórica
ESTU_INST_CODMUNICIPIO	Numérica
ESTU_INST_MUNICIPIO	Categórica
ESTU_INST_DEPARTAMENTO	Categórica
INST_CARACTER_ACADEMICO	Categórica
INST_ORIGEN	Categórica
ESTU_PRIVADO_LIBERTAD	Categórica
ESTU_COD_MCPIO_PRESENTACION	Numérica
ESTU_MCPIO_PRESENTACION	Categórica
ESTU_DEPTO_PRESENTACION	Categórica
ESTU_COD_DEPTO_PRESENTACION	Numérica
MOD_RAZONA_CUANTITAT_PUNT	Numérica
MOD_RAZONA_CUANTITAT_DESEM	Numérica
MOD_RAZONA_CUANTITATIVO_PNAL	Numérica
MOD_RAZONA_CUANTITATIVO_PNBC	Numérica
MOD_LECTURA_CRITICA_PUNT	Numérica
MOD_LECTURA_CRITICA_DESEM	Numérica
MOD_LECTURA_CRITICA_PNAL	Numérica
MOD_LECTURA_CRITICA_PNBC	Numérica
MOD_COMPETEN_CIUDADA_PUNT	Numérica
MOD_COMPETEN_CIUDADA_DESEM	Numérica
MOD_COMPETEN_CIUDADA_PNAL	Numérica
MOD_COMPETEN_CIUDADA_PNBC	Numérica
MOD_INGLES_PUNT	Numérica
MOD_INGLES_DESEM	Categórica
MOD_INGLES_PNAL	Numérica
MOD_INGLES_PNBC	Numérica
MOD_COMUNI_ESCRITA_PUNT	Numérica
MOD_COMUNI_ESCRITA_DESEM	Numérica
MOD_COMUNI_ESCRITA_PNAL	Numérica
MOD_COMUNI_ESCRITA_PNBC	Numérica
PUNT_GLOBAL	Numérica
PERCENTIL_GLOBAL	Numérica
PERCENTIL_NBC	Numérica
ESTU_INSE_INDIVIDUAL	Numérica
ESTU_NSE_INDIVIDUAL	Numérica
ESTU_NSE_IES	Numérica
ESTU_ESTADOINVESTIGACION	Categórica

Fuente. Elaboración propia.

Apéndice C

Datos descartables

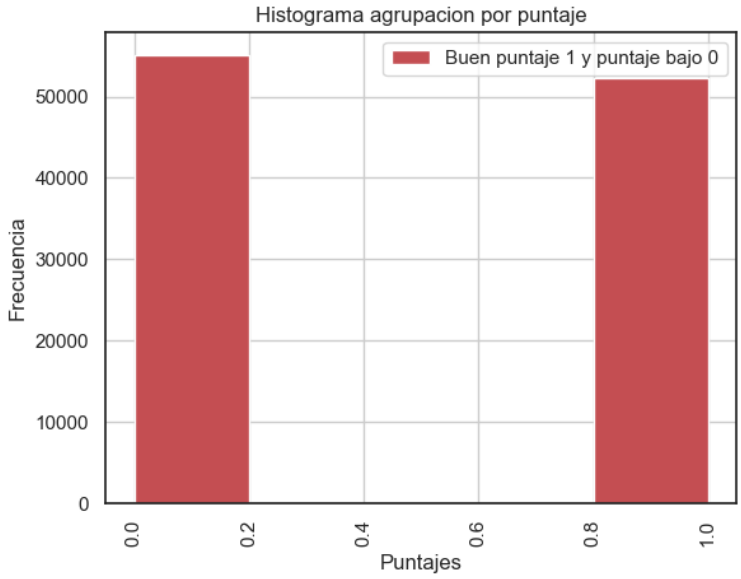
Icfes Saber Profesional genéricas 2022	Cantidad
Total, Columnas	109
Descartables	56
Características descartables	
Variables	**Clasificación**
ESTU_TIPODOCUMENTO	Categórica
ESTU_NACIONALIDAD	Categórica
ESTU_EXTERIOR	Categórica
PERIODO	Categórica
ESTU_CONSECUTIVO	Categórica
ESTU_ESTUDIANTE	Categórica
ESTU_PAIS_RESIDE	Categórica
ESTU_ESTADOCIVIL	Categórica
ESTU_COLE_TERMINO	Categórica
ESTU_CODDANE_COLE_TERMINO	Numérica
ESTU_COD_COLE_MCPIO_TERMINO	Numérica
ESTU_OTROCOLE_TERMINO	Categórica
ESTU_VALORMATRICULAUNIVERSIDAD	Categórica
ESTU_COMOCAPACITOEXAMENSB11	Categórica
ESTU_CURSODOCENTESIES	Categórica
ESTU_CURSOIESAPOYOEXTERNO	Categórica
ESTU_CURSOIESEXTERNA	Categórica
ESTU_SIMULACROTIPOICFES	Categórica
ESTU_ACTIVIDADREFUERZOAREAS	Categórica
ESTU_ACTIVIDADREFUERZOGENERIC	Categórica
ESTU_TIPODOCUMENTOSB11	Categórica
FAMI_EDUCACIONPADRE	Categórica
FAMI_EDUCACIONMADRE	Categórica
FAMI_OCUPACIONPADRE	Categórica
FAMI_OCUPACIONMADRE	Categórica
FAMI_TIENELAVADORA	Categórica
FAMI_TIENEHORNOMICROOGAS	Categórica
FAMI_TIENESERVICIOTV	Categórica
FAMI_TIENEAUTOMOVIL	Categórica
FAMI_TIENEMOTOCICLETA	Categórica
FAMI_TIENECONSOLAVIDEOJUEGOS	Categórica
FAMI_TRABAJOLABORPADRE	Categórica
FAMI_TRABAJOLABORMADRE	Categórica
ESTU_PLANTELEDUCATIVO	Categórica
ESTU_OTROPLANTEL	Numérica
ESTU_PREGRADOEXAMENSP	NaN
ESTU_UNPOSTGRADO	NaN

FAMI_CUANTOSCOMPARTEBAÑO	Numérica
ESTU_PAGOMATRICULA	Categórica
ESTU_VLRULTIMOSEMESCURSADO	Categórica
ESTU_PAISDOCUMENTOS11	Categórica
ESTU_PRESENTACIONSABADO	Categórica
GRUPOREFERENCIA	Categórica
ESTU_PRGM_CODMUNICIPIO	Numérica
ESTU_PRGM_MUNICIPIO	Categórica
ESTU_PRGM_DEPARTAMENTO	Categórica
ESTU_NIVEL_PRGM_ACADEMICO	Categórica
ESTU_NUCLEO_PREGRADO	Categórica
ESTU_INST_CODMUNICIPIO	Numérica
INST_CARACTER_ACADEMICO	Categórica
ESTU_PRIVADO_LIBERTAD	Categórica
ESTU_COD_MCPIO_PRESENTACION	Numérica
ESTU_MCPIO_PRESENTACION	Categórica
ESTU_DEPTO_PRESENTACION	Categórica
ESTU_COD_DEPTO_PRESENTACION	Numérica
ESTU_ESTADOINVESTIGACION	Categórica

Fuente. Elaboración propia.

Apéndice D

Histograma de la categorización de la variable objetivo puntaje global



Fuente. Elaboración propia.