Aplicación de técnicas de selección y algoritmos de clasificación en el análisis de resultados pruebas saber pro de la escuela de ingeniería de sistemas UNAD Bogotá

Andrés Eduardo Combariza Monroy

Asesor

Mireya García García

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Ciencia de Datos y Analítica

2024

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi familia, quienes han sido mi pilar fundamental a lo largo de este camino. En especial, a mis padres, quienes, con su amor incondicional, paciencia y apoyo constante me han guiado y motivado en cada etapa de mi vida académica. A mi madre, por ser mi inspiración diaria, por creer en mí incluso en los momentos más difíciles. Su fortaleza y dedicación son un ejemplo que llevo siempre conmigo. A mi padre, por su incansable esfuerzo y por enseñarme el valor del trabajo duro y la perseverancia. Su confianza en mis capacidades me ha impulsado a alcanzar mis metas y superar mis propios límites. A ambos, les debo cada logro y éxito obtenido. Sin su apoyo incondicional y sacrificios, nada de esto habría sido posible. Este trabajo es un reflejo del amor y la dedicación que me han brindado a lo largo de los años.

Gracias por ser mi luz guía y mi sostén en todo momento.

Resumen

Este documento presenta la creación de un modelo que aplica un algoritmo de clasificación utilizando datos públicos del ICFES para analizar los factores demográficos y socioeconómicos que influyeron en los resultados de las pruebas Saber Pro de la Escuela de Ingeniería de la UNAD, Bogotá, entre 2019 y 2022. Además, destaca la importancia de la ciencia de datos y su contribución a la toma de decisiones organizacionales mediante el desarrollo de modelos y herramientas analíticas

***Palabras claves*:** Machine Learning, algoritmo de clasificación, árbol de decisión, factores demográficos y socio económicos, pruebas saber Pro.

Abstract

This document presents the creation of a model that applies a classification algorithm using public data from ICFES to analyze the demographic and socioeconomic factors that influenced the results of the Saber Pro tests of the School of Engineering at UNAD, Bogotá, between 2019 and 2022. Additionally, it highlights the importance of data science and its contribution to organizational decision-making through the development of models and analytical tools.

***Keywords*:** Machine Learning, classification algorithm, decision tree, demographic and socioeconomic factors, Saber Pro tests**.**

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

Utilizar los conocimientos adquiridos para dar respuesta o soluciones a preguntas o problemáticas existentes, podría decirse que es la razón de ser de una propuesta educativa; las nuevas disciplinas en este caso la ciencia de datos debe empezar a darse a conocer y más aún quien tiene el manejo de estos conocimientos, el científico de datos, especialidad que está tomando fuerza y creciendo de manera considerable, por ello es que en este trabajo se denotará un especial interés en poner de manifiesto la relevancia de la disciplina mencionada y de quien la utiliza con rigurosidad.

Para ello se inicia determinando una situación, resultados de las pruebas Saber Pro que acorde al (Decreto No. 1781 de 2003) donde se afirma que los resultados son utilizados para evaluar la calidad del servicio de educación superior, los datos de los resultados de las pruebas Saber Pro se encuentran en la base de datos abiertos del Instituto Colombiano para la Evaluación de la Educación (ICFES). Estos se usarán para determinar los factores demográficos y socio económicos que inciden en los resultados. De esta manera se genera una condición de interés para la presente investigación ¿Cómo hacer uso de algoritmos de clasificación para determinar los factores que incidieron en los resultados de las pruebas Saber Pro de los estudiantes de Ingeniería de sistemas de la UNAD sede Bogotá en los años 2019 al 2022? Para ello se plantea generar un modelo con el fin de aplicar algoritmos de Machine Learning de clasificación en los factores demográficos y socio económicos que incidieron en los resultados de las pruebas saber pro; realizando los siguientes procesos: 1. Obtener información de la base de datos abiertos del ICFES (Data ICFES) de los factores demográficos y socioeconómicos de los años 2019 al 2022.

Luego realizar el análisis exploratorio de datos demográficos y socioeconómicos de los años 2019 al 2022; posteriormente aplicar limpieza y transformación de los datos demográficos y

socio económicos para finalmente presentar el modelo de datos para aplicar el algoritmo de Machine Learning de clasificación.

Para la sustentación teórica se realizó la revisión de antecedentes o trabajos que retomaran tanto problemática, contexto, factores y uso de ciencia de datos, posteriormente se fundamenta con las teorías y términos propios iniciando con lo concernientes a que se entenderá por factores demográficos y socio económicos, continuando conceptos de la ciencia de datos, Machine Learning, lo cual permitirá a los lectores mayor claridad y apropiación de los términos en que se habla, ejemplo de ello se encontrará, Scikit-learn, Análisis de PCA, K-fold, cross-validation, árbol de decisión, entre otros; se continúa con la metodología ,especificando procesos, códigos y demás procedimientos, se prosigue con los datos, el modelo y se finaliza con la presentación de las conclusiones; reiterando el interés de que se evidencie el aporte que tanto el científico de datos como la ciencia de datos aportan para que se genere información tan relevante que va más allá del registro de datos, para dar explicaciones cuantitativas que propicien la argumentación y apoyo en la toma decisiones en las organizaciones.

Planteamiento del Problema

En Colombia, para obtener el título de pregrado, es necesario presentar las pruebas Saber Pro como requisito obligatorio desde el año 2009. Estas pruebas, administradas por el Instituto Colombiano para la Evaluación de la Educación Superior (ICFES), están conformadas por módulos de competencias genéricas y específicas de cada área profesional. Los resultados de estas pruebas son fundamentales para medir la calidad educativa y evaluar los objetivos de formación de las instituciones.

Las universidades cumplen un papel crucial en la formación académica, investigación y desarrollo social. Sin embargo, para que puedan cumplir eficientemente su función, es necesario analizar los resultados de las pruebas Saber Pro. Factores demográficos y socioeconómicos, como la situación financiera o el entorno social de los estudiantes, podrían estar relacionados con los resultados obtenidos, lo cual genera interrogantes importantes como: ¿Qué factores influyen en el rendimiento de los estudiantes? ¿Cómo inciden estos factores en los resultados de las pruebas? ¿Es posible predecir el puntaje global usando información socioeconómica y demográfica? ¿Cómo pueden estas respuestas orientar la gestión curricular para mejorar la formación académica?

Este trabajo de investigación se centra en los estudiantes de ingeniería de sistemas de la Universidad Nacional Abierta y a Distancia (UNAD), sede Bogotá, durante el periodo 2019-2022. A través del análisis de los datos proporcionados por el ICFES, se busca identificar los factores que han repercutido en los resultados de las pruebas Saber Pro. Para lograrlo, se propone aplicar tecnologías de ciencia de datos, específicamente algoritmos de clasificación de Machine Learning, que permiten analizar grandes volúmenes de datos, identificar patrones y generar información relevante para la toma de decisiones.

La ciencia de datos, que combina estadística, matemáticas, inteligencia artificial e ingeniería informática, ofrece herramientas para responder la pregunta clave: ¿Cómo hacer uso de algoritmos de clasificación para determinar los factores que incidieron en los resultados de las pruebas Saber Pro de los estudiantes del programa de ingeniería de sistemas de la UNAD sede Bogotá, en los años 2019 al 2022?

Al identificar los elementos que inciden en los resultados, se podrán implementar estrategias que permitan optimizar el proceso de enseñanza-aprendizaje y, por lo tanto, los resultados de las pruebas Saber Pro.

El análisis planteado tiene implicaciones tanto a nivel institucional como a nivel nacional. En lo micro, podrá contribuir a la gestión curricular y pedagógica de la UNAD; a nivel macro, permitirá generar políticas y estrategias de mejora en la educación superior, incidiendo en el rendimiento académico de los estudiantes y en su proyección profesional. Así, el presente trabajo no solo busca comprender los factores que afectan el desempeño en las pruebas Saber Pro, sino también proporcionar un modelo predictivo que facilite la toma de decisiones orientadas al mejoramiento continuo de la educación en el país.

Justificación

La ciencia de datos es una disciplina que está tomando auge e importancia en las organizaciones, puesto que una de sus finalidades es dar funcionalidad a datos existentes y el científico de datos en su rol, identificar patrones, tendencias y relaciones que se pueden utilizar para la toma de decisiones en una organización; por ello como miembro de la comunidad educativa de la Universidad Nacional Abierta y a Distancia, en adelante se mencionará como UNAD, de la escuela de ingeniería y la especialidad de ciencia de datos, se genera el interés de determinar haciendo uso del algoritmo de clasificación los factores que inciden en los resultados de las pruebas Saber Pro de los estudiantes del programa de ingeniería de sistemas de la UNAD Bogotá; teniendo en cuenta que la UNAD, desde sus inicios ha tenido como objetivo favorecer la educación de las clases sociales menos favorecidas, siendo una de sus metas, formar a comunidades populares, se rige actualmente por el lema "una educación para todos", lo cual genera accesibilidad y mayor cobertura; por esto en las bases de datos abiertos del ICFES, (Data ICFES), se puede encontrar que la población estudiantil de la UNAD está formada en un alto porcentaje por personas de estratos 1,2, 3 y en un bajo porcentaje por población de estratos 4, 5 y 6; investigaciones como la de (Vargas, 2014), realizada con estudiantes de la universidad de Costa Rica, han concluido que las variables demográficas y socio económicas, tienen afectación positiva para un bajo desempeño académico.

Es el interés del presente trabajo de investigación determinar las variables demográficas y socioeconómicas que brinden información pertinente, que permita tener de forma verificable, para la toma de decisiones que mejoren los resultados de las pruebas y la calidad educativa acorde a los requerimientos del tipo de población estudiantil de la UNAD. Además al implemen-

tar los conocimientos adquiridos tales como el proceso de aplicación de algoritmos de clasificación utilizados en Machine Learning y generar modelos, se profundiza en la utilidad de lo aprendido y pone de manifiesto la pertinencia de la implementación de disciplinas como la ciencia de datos, de quienes, siendo entonces importante que se empiece a conocer la función de un analista de datos y un científico de datos, estos contribuyan con información que permiten a las instituciones conocer y comprender las variables que inciden, en este caso, en los procesos educativos y realicen cambios en pro de la calidad de la educación.

Objetivos

General

Desarrollar un modelo de análisis predictivo que utilice algoritmos de clasificación para identificar los factores que influyen en los resultados de las pruebas Saber Pro de estudiantes de Ingeniería de Sistemas de la UNAD, sede Bogotá, durante los años 2019 a 2022.

Específicos

Realizar un análisis exploratorio de los datos demográficos y socioeconómicos de los años 2019 al 2022 en los resultados de las pruebas Saber Pro de estudiantes de Ingeniería de Sistemas de la UNAD, sede Bogotá, para identificar patrones y tendencias relevantes que puedan influir en el desempeño académico.

Aplicar limpieza y transformación de los datos demográficos y socioeconómicos de los años 2019 al 2022, con el fin de garantizar su calidad, homogeneidad y preparación para el análisis predictivo.

Generar un modelo de análisis predictivo que utilice algoritmos de clasificación avanzados, como la regresión logística y los árboles de decisión, para determinar los factores demográficos y socioeconómicos que tienen mayor impacto en los resultados de las pruebas Saber Pro de los estudiantes de ingeniería de sistemas de la UNAD sede Bogotá, durante el periodo 2019-2022.

Marco de Referencia

Estado del Arte

Al hacer la revisión se pudo observar que la problemática, el contexto y la población son temas de interés que han generado varias investigaciones. Se encuentran desde el uso de modelos de regresión y clasificación para analizar y tomar decisiones basadas en datos, de organizaciones y a nivel educativo ha ido en creciente demanda, ya que al utilizar la información existente se puede determinar acciones de mejora; para el interés y sustento del presente trabajo se referencia el trabajo de Cano (2024), en el cual se utilizan técnicas de Machine Learning, emplea regresión lineal, regresión logística y árbol de decisiones, para predecir el rendimiento académico, plantea cómo determinar, recopilar y analizar datos significativos y seleccionar el algoritmo pertinente, se consideró porque emplea procedimientos de Machine Learning, hace análisis de factores demográficos y socioeconómicos, además pertenece al repositorio de la UNAD.

Se tomó en cuenta que las investigaciones tengan relación tanto con la temática, pruebas Saber pro, el contexto que en este caso son las universidades y los factores demográficos y socioeconómicos de los estudiantes, que es de donde se obtienen los datos. Con el trabajo de Timarán Pereira, Caicedo Zambrano, & Hidalgo Troya (2021) da aportes conceptuales, metodológicos y con posibilidad de corroborar los datos obtenidos con los de la presente investigación.

El siguiente trabajo, de Cifuentes Medina, Chacón Benavides, & Fonseca Correa (2020), donde se realiza el análisis de los resultados de las Pruebas Saber Pro en estudiantes de la licenciatura en Educación Básica de la Universidad pedagógica y Tecnológica de Colombia (UPTC)", en él se estudia el efecto de los resultados obtenidos en las pruebas Saber Pro de los estudiantes de una carrera universitaria, hace uso de la base de datos abiertos del ICFES, aunque no especifica que se hubiera utilizado un proceso de Machine Learning, por lo cual no hay sustento, tiene

puntos de coincidencia que aportaron. Fundamenta la idea del presente trabajo de la importancia que tiene tanto la ciencia de datos como el científico de datos, ya que, aunque se utilicen las técnicas de la disciplina se generan datos, pero no los pertinentes e incidentes para generar toma de decisiones.

Oviedo Carrascal & Jiménez Giraldo (2019), se referencia porque hace un estudio similar al propuesto, aplica la metodología de minería de datos CRISP-DM, en la universidad de Antioquía, tomando factores demográficos y socioeconómicos, encontrando que algunos de estos influyen en los resultados de las pruebas saber Pro, es decir, fundamenta lo propuesto en la presente investigación.

Franco (2017), de él se puede tomar como referencia el sustento teórico de dos aspectos que tienen en común, resultados de las pruebas Saber Pro y los factores demográficos y socioeconómicos, igualmente la metodología utilizada para predicción tiene aportes al presente trabajo.

No se encontraron trabajos de investigación, artículos que traten directamente de los factores demográficos y socioeconómicos en la UNAD en el programa de ingeniería de sistemas, en el repertorio revisado como ya se mencionó. Se encuentra gran número de investigaciones donde se utiliza el análisis de datos, más no la ciencia de datos, la diferencia radica en que el análisis de datos se centra en interpretar información para responder preguntas específicas, mientras que la ciencia de datos combina estadísticas, programación y modelos avanzados para predecir, automatizar y optimizar decisiones basadas en datos. Se emplean procedimientos (algoritmos, árbol de decisiones) mediante el uso de software de minería de datos como es el caso de Yachay -

DTM; la UNAD ha sido centro de interés de estudios realizados referentes a los resultados obtenidos por los estudiantes en las pruebas Saber Pro, sin embargo, la escuela de Ciencias básicas, tecnología e ingeniería, en el programa de Ingeniería de sistemas, no han sido analizadas.

Marco Contextual

Esta investigación se desarrolla en la UNAD, específicamente en la sede Bogotá, dentro de la Escuela de Ciencias Básicas, Tecnología e Ingeniería, con énfasis en el programa de Ingeniería de Sistemas. El estudio se centra en los estudiantes que presentaron las pruebas Saber Pro entre los años 2019 y 2022.

Resultados de la población se obtuvieron de la base de datos abiertos del ICFES de los resultados de las pruebas Saber Pro, obteniéndose una muestra de 2834 estudiantes de ingeniería de sistemas de la UNAD, sede Bogotá.

Marco Teórico

Para el presente trabajo de investigación, se realiza una revisión de los conceptos, teorías y herramientas tecnológicas utilizados en el tema de investigación. Inicialmente se define los conceptos claves que forman base del estudio, seguido de un análisis de teorías que sustentan el procedimiento aplicado, posteriormente se presentan las herramientas tecnológicas y analíticas empleadas para abordar el problema planteado.

Factores socioeconómicos: El nivel socioeconómico es un indicador que surge a partir del análisis del salario o del dinero que obtiene un individuo; de sus condiciones de empleo; y de su formación educativa. (Julián Pérez Porto & Ana Gardey, 2021)

Factores demográficos: Los factores o indicadores demográficos son estadísticas que proporcionan información sobre la estructura, la dinámica y las características de una población. Estos indicadores son utilizados para analizar diferentes aspectos de las poblaciones humanas,

como su tamaño, distribución, crecimiento, composición por edad e identidad de género, entre otros. (Sposob, 2024).

Algoritmos supervisados: Para analizar los factores demográficos y socioeconómicos que impactaron los resultados de las pruebas Saber Pro, se seleccionaron algoritmos de clasificación, como la regresión logística y los árboles de decisión. Para esta investigación se destacan las ventajas de estos métodos en comparación con otros algoritmos: La regresión logística se destaca por su simplicidad e interpretabilidad, eficiencia computacional y capacidad para manejar probabilidades, permitiendo clasificaciones precisas y la evaluación de incertidumbres. Por otro lado, los árboles de decisión ofrecen visualizaciones claras, manejan variables mixtas, es decir numéricas y categóricas, sin transformaciones complejas y capturan interacciones no lineales, lo que los hace versátiles y fáciles de entender, incluso para audiencias no técnicas.

Como se describe en Ramasubramanian & Moolayil (2019), los algoritmos supervisados, son uno de los algoritmos más comunes en aprendizaje automático. Se denomina supervisado porque el proceso de aprendizaje de los algoritmos a partir del conjunto de datos de entrenamiento puede compararse con un maestro que guía el aprendizaje. En este tipo de algoritmos, los resultados esperados ya se conocen y los datos de entrenamiento están etiquetados con las respuestas correctas.

Los problemas tratados principalmente con algoritmos supervisados se pueden clasificar en los siguientes dos tipos:

Clasificación: un problema se considera de clasificación cuando el resultado es categórico, como "negro", "enseñanza" o "no enseñanza".

Regresión: un problema se denomina de regresión cuando la salida es un valor continuo o real, como "distancia" o "peso".

Los algoritmos que trabajan con variables continuas se denominan algoritmos de regresión, mientras que aquellos que manejan variables categóricas son conocidos como algoritmos de clasificación. En los algoritmos de clasificación, la variable objetivo, dependiente o criterio es categórica. Según el número de clases, estos problemas pueden dividirse en diferentes grupos.

Clasificación binaria

Clasificación multinomial

Clasificación multi-etiqueta

Regresión logística: De acuerdo con Madhavan (2015, p. 139), la regresión logística es un método de análisis que permite estimar la probabilidad de que un evento ocurra en función de ciertos parámetros dados. Es comúnmente empleada como una técnica de clasificación con resultados binarios. Las probabilidades que representan los posibles resultados de una prueba se modelan utilizando una función logística en relación con las variables explicativas o predictoras.

Árbol de decisión: Los Árboles de Decisión, como se plantea en Ramasubramanian et al. (2019, p. 247), al igual que la regresión logística, representan otra técnica de clasificación ampliamente utilizada debido a su simplicidad y su carácter transparente. Un árbol de decisión es un diagrama de flujo sencillo que adopta la forma de un árbol invertido. Comienza con un nodo raíz, que se divide en múltiples ramas hacia otros nodos, los cuales se recorren según una decisión, y finaliza en un nodo hoja donde se determina el resultado final. Los árboles de decisión pueden aplicarse tanto a problemas de clasificación como de regresión. Existen diversas variantes de esta técnica implementadas en el aprendizaje automático. Entre las más populares se encuentran:

Iterative Dichotomiser 3 (ID3)

Sucesor de ID3 (C4.5)

Árbol de Clasificación y Regresión (CART)

Detector Automático de Interacciones por Chi-cuadrado (CHAID)

Árboles de Inferencia Condicional (C Trees)

La lista anterior no es exhaustiva. Hay otras alternativas, y cada una de ellas tiene pequeñas variaciones en la forma en que abordan el proceso de creación del árbol.

En la presente investigación se realizó el proceso para la aplicación de algoritmos de clasificación, en este caso regresión logística y árbol de decisión, con el cual se determinaron los factores demográficos y socioeconómicos de los estudiantes que presentaron las pruebas Saber Pro en la UNAD Bogotá, del programa de ingeniería de sistemas entre los años 2019 a 2022. Se buscó la aplicación de estos dos algoritmos de clasificación para comparar su funcionamiento y resultados; aunque los dos son para clasificación presentan diferencias como: la regresión logística asume que existe una relación lineal entre variables independientes y la probabilidad de ocurrencia de una categoría, el resultado que retorna son coeficientes indicando la influencia de cada variable predictora sobre la probabilidad de ocurrencia de una categoría en específico. Mientras que el árbol de decisión no asume relaciones lineales, captura relaciones un poco más complejas, ofrece una mejor interpretación, ya que presenta una estructura jerárquica de decisiones, lo cual permite observar un conjunto de reglas que utiliza para clasificar observaciones, en lugar de coeficientes, proporciona medidas de importancia para los predictores basadas en cuántas veces se usan para dividir datos. Contrastando la teoría con la práctica, corrobora su uso en la presente investigación.

Análisis de correlación: Tal como lo señala Madhavan (2015, p. 48), en estadística, la correlación mide el grado de similitud o relación entre dos variables aleatorias. La forma más común de correlación es la correlación de Pearson, la cual se expresa de la siguiente manera:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

La fórmula presentada describe la correlación de Pearson como la covarianza entre X e Y, dividida entre las desviaciones estándar de X e Y. También puede interpretarse como el valor esperado de la suma del producto de las diferencias de las variables aleatorias respecto a sus medias, dividido por las desviaciones estándar de X e Y.

Hiperparámetros: Nyuytiymbiy (2020) indica que los hiperparámetros son valores que controlan el proceso de aprendizaje y determinan los parámetros del modelo que un algoritmo de aprendizaje ajustará. El prefijo "hiper_" indica que estos son parámetros de "alto nivel" que dirigen tanto el proceso de aprendizaje como los parámetros del modelo que se derivan de este.

Los valores de los hiperparámetros se eligen y configuran antes de iniciar el entrenamiento del modelo. Por lo tanto, se considera que los hiperparámetros son externos al modelo, ya que no pueden ser modificados por el modelo durante su aprendizaje o entrenamiento. Aunque los hiperparámetros son utilizados por el algoritmo de aprendizaje durante el proceso, no forman parte del modelo final.

Cuando se requiere aplicar un algoritmo de Machine Learning como en este caso de clasificación, es necesario hacer un procesamiento previo de los datos, con el fin de obtener el mejor funcionamiento del modelo que se va a generar. La forma en que inicialmente se puede hacer este ajuste es través de los hiperparámetros, los cuales son valores que se pueden ajustar para cambiar la forma en que trabaja el modelo.

Precisión: “Precisión o Valor Predictivo Positivo (PPV): La precisión se define como la proporción de etiquetas positivas correctamente predichas respecto al número total de etiquetas positivas predichas”. (Ramasubramanian et al., 2019, p. 236).

$$precision(PPV) = \frac{TP}{TP+FP} \quad (2)$$

Recall: o Sensibilidad: El recall mide cuán sensible es tu modelo representando la proporción del número de etiquetas positivas correctamente predichas respecto al número total de etiquetas positivas reales. (Ramasubramanian et al., 2019, p. 237).

$$recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score: Puntuación F1: es la media armónica entre la precisión y el recall. Es una mejor métrica para considerar que la precisión general en la mayoría de los casos". (Ramasubramanian et al., 2019, p. 237).

$$F1 - score = \frac{2*precision*recall}{precision+recall} \quad (4)$$

Accuracy: La exactitud o accuracy se define como: la proporción del total de predicciones correctas respecto al número total de predicciones en toda la muestra de prueba. Por lo tanto, esto sería simplemente la suma de Verdaderos Positivos y Verdaderos Negativos dividida por todas las métricas en la matriz de confusión. (Ramasubramanian et al., 2019, p. 236).

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (5)$$

Herramientas Tecnológicas

Python: es un lenguaje de programación potente y fácil de aprender. Tiene estructuras de datos de alto nivel eficientes y un enfoque simple pero efectivo para la programación orientada a objetos. La sintaxis elegante y la tipificación dinámica de Python, junto con su naturaleza interpretada, lo convierten en un lenguaje ideal para la creación de scripts y el desarrollo rápido de aplicaciones en muchas áreas en la mayoría de las plataformas. (Python.org, 2024)

Anaconda: Es una distribución de ciencia de datos Python/R que contiene conda. Conda es un entorno y un administrador de paquetes que ayuda a los usuarios a administrar los otros 7500 paquetes de código abierto que están disponibles a través de la distribución. (Anaconda, 2024)

Al hacer uso del lenguaje de programación Python, que se despliega un entorno de Anaconda, una vez los datos están cargados, se inicia una exploración de estos, lo cual permite detectar diferentes anomalías tales como valores nulos o faltantes en los registros del conjunto de datos, valores atípicos, el tipo de variables que están contenidas en los datos, datos inconsistentes y la frecuencia de las clases de las variables categóricas, siendo posible identificar inconsistencias en las clases de las variables categóricas.

Pandas: es un paquete de Python que ofrece estructuras de datos rápidas, flexibles y expresivas diseñadas para que trabajar con datos "relacionales" o "etiquetados" sea fácil e intuitivo. Su objetivo es ser el componente básico de alto nivel para realizar análisis prácticos de datos del mundo real en Python. Además, tiene el objetivo más amplio de convertirse en la herramienta de análisis y manipulación de datos de código abierto más potente y flexible disponible en cualquier lenguaje. (Pandas, 2024)

Numpy: (Numerical Python) es una biblioteca de código abierto de Python que se utiliza en casi todos los campos de la ciencia y la ingeniería. Es el estándar universal para trabajar con datos numéricos en Python y se encuentra en el núcleo de los ecosistemas científicos de Python y PyData. (Numpy developers, 2024)

Scikit-learn: es una biblioteca de aprendizaje automático de código abierto que admite el aprendizaje supervisado y no supervisado. También proporciona varias herramientas para el ajuste de modelos, el preprocesamiento de datos, la selección de modelos, la evaluación de modelos y muchas otras utilidades. (scikit-learn developers, 2024)

One-hot encoder: codifica las características categóricas como un arreglo numérico.

De acuerdo con scikit-learn developers (2020), La entrada para este transformador debe ser una lista de números enteros o de cadenas de texto, que representan los valores de la variable

categórica. Se transforman mediante un esquema one-hot (one-of-K o dummy), otorgándole a cada categoría una columna binaria. Debido a esto, habrá una matriz dispersa (solamente ceros) o un arreglo denso que la función devolverá según el valor del parámetro “sparse”.

Análisis de PCA: el análisis de componentes principales (PCA, por sus siglas en inglés) como se define en Dangeti (2017, p. 320), es una técnica utilizada para reducir la dimensionalidad de los datos y tiene diversas aplicaciones. Este método disminuye las dimensiones de un conjunto de datos al proyectarlos en un subespacio de menor dimensión. Por ejemplo, un conjunto de datos bidimensional puede reducirse al proyectar los puntos sobre una línea, lo que permite representar cada instancia con un solo valor en lugar de un par de valores. De manera análoga, un conjunto de datos tridimensional podría reducirse a dos dimensiones proyectando las variables en un plano. Las principales utilidades de PCA incluyen:

Mitigar el curso de la dimensionalidad

Comprimir los datos mientras se minimiza la pérdida de información al mismo tiempo.

Entender la estructura de datos con cientos de dimensiones puede ser difícil, por lo tanto, al reducir las dimensiones a 2D o 3D, las observaciones pueden visualizarse fácilmente.

Prueba Selectkbest: es un tipo de selección de características univariadas, tal como señala (scikit-learn developers, 2024), opera seleccionando las mejores características en función de pruebas estadísticas univariadas. SelectKBest elimina todo excepto las características con la puntuación más alta.

Con el método *SelectKBes*t, se aplica una prueba estadística para cada variable de manera individual generando la selección de acuerdo con la mejor puntuación acorde al objetivo

Prueba RFE: eliminación Recursiva de Características, como se explica en (scikit-learn developers, 2024), la eliminación recursiva de características (RFE) utiliza un estimador externo

que asigna pesos a las características, como los coeficientes de un modelo lineal, con el objetivo de seleccionar las más relevantes reduciendo gradualmente el conjunto de características. En primer lugar, el estimador se entrena con el conjunto completo de características, y su importancia se evalúa mediante un atributo específico o invocable. Posteriormente, se eliminan las características menos relevantes del conjunto actual. Este proceso se repite de forma recursiva sobre el conjunto reducido hasta alcanzar la cantidad deseada de características seleccionadas.

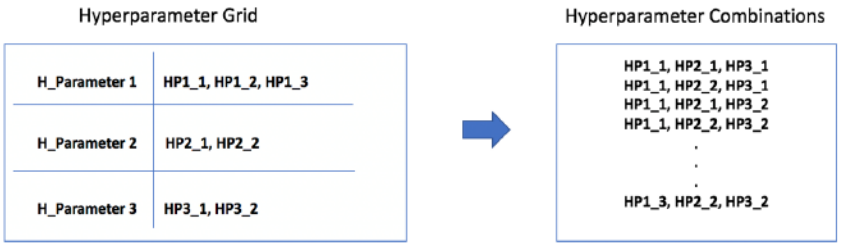
Con el *RFE*, se van eliminando en cada iteración las menos importantes, alcanzando un número de variables relevantes para el modelo, la gráfica que genera debe ser entendida: la más baja indica mayor importancia para el modelo; hasta este paso en el desarrollo de la propuesta se encuentra que los conceptos y teorías revisadas y seleccionadas han propiciado conocimientos que solo se van a entender si se aplicaron, generando mayor apropiación de los procesos para que en posteriores usos ya se omiten pasos o se utiliza la técnica pertinente.

Gridsearch: la búsqueda en cuadrícula se define en Ramasubramanian et al. (2019, p 343) como una técnica comúnmente utilizada en aprendizaje automático para ajustar los hiperparámetros de un modelo y encontrar la combinación más adecuada para obtener el mejor rendimiento. En este enfoque, se define una malla de valores para realizar una búsqueda exhaustiva de las posibles combinaciones de hiperparámetros. El modelo se entrena con todas las combinaciones posibles de estos valores, y luego se selecciona la que muestra el mejor desempeño como la óptima. El siguiente diagrama ilustra cómo funciona la optimización mediante búsqueda en cuadrícula en un conjunto hipotético de parámetros. Al usar esta malla de hiperparámetros, se generan todas las combinaciones posibles y el modelo se entrena para cada una de ellas.

La principal ventaja de la optimización mediante búsqueda en cuadrícula es que reduce considerablemente el tiempo necesario para encontrar el conjunto óptimo de hiperparámetros, dado el número limitado de valores candidatos a evaluar.

Figura 1

Grid Search



Nota. Tomada de Applied supervised Learning with R, (p. 343), Ramasubramanian et al. (2019).

Debido a que no existe una fórmula universal para seleccionar estos valores, se prueban diferentes combinaciones mediante métodos como la búsqueda en cuadrícula (Grid Search), esta, prueba varias combinaciones de hiperparámetros y encuentra la que proporciona el mejor rendimiento del modelo mediante una técnica para evaluar el rendimiento de manera más robusta, ya que en lugar de dividir los datos una sola vez en entrenamiento y prueba, se dividen en K partes y se entrena el modelo K veces con diferentes combinaciones de datos de entrenamiento y prueba. Al final, promedia los resultados para obtener una evaluación más confiable.

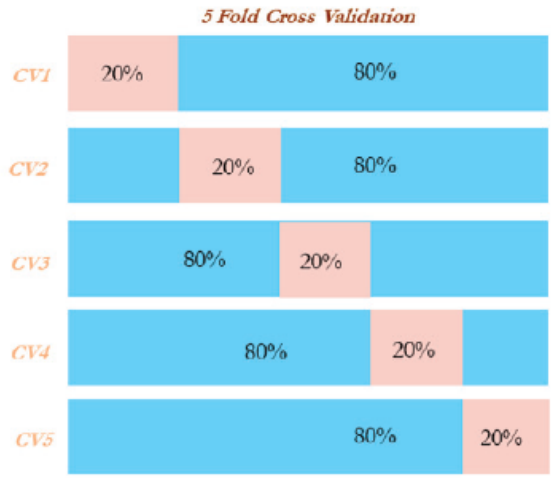
Cross-validation: basado en la definición de validación cruzada según Dangeti (2017, p. 46), es otra técnica utilizada para asegurar la robustez del modelo, aunque implica un mayor coste computacional. En el enfoque tradicional de modelado, un modelo se entrena con un conjunto de datos y se evalúa con otro conjunto de prueba. Sin embargo, en algunos casos, los conjuntos de entrenamiento y prueba podrían no haberse seleccionado de manera homogénea, lo que

podría resultar en la aparición de casos atípicos en los datos de prueba, afectando negativamente el rendimiento del modelo.

En cambio, con la validación cruzada, los datos se dividen en varios subconjuntos iguales, y el modelo se entrena utilizando todos los subconjuntos excepto uno, que se utiliza para evaluar el rendimiento. Este proceso se repite tantas veces como el número de subconjuntos definidos por el usuario, asegurando una evaluación más completa y fiable del modelo.

Figura 2

Cross-Validation



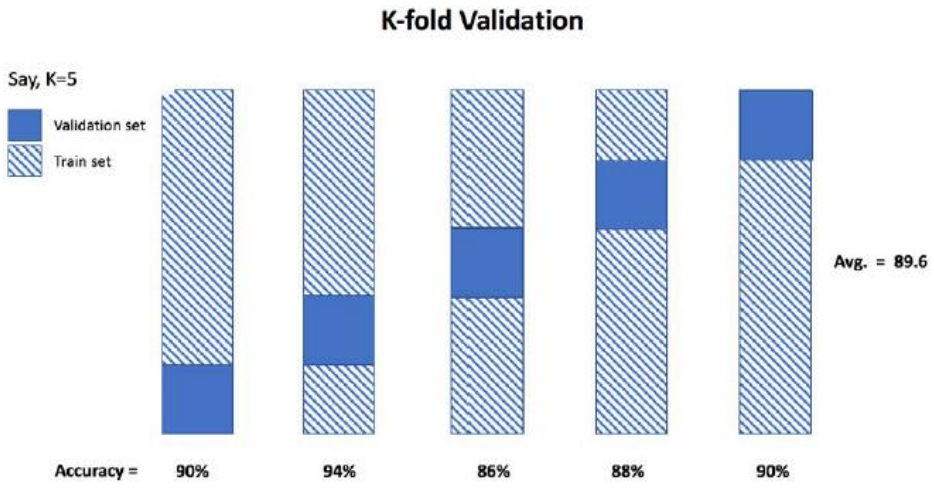
Nota. Tomada de Statistics for Machine Learning, (p. 46), Dangeti (2017)

K-fold cross-validation: como lo define Ramasubramanian et al. (2019, p. 335), esta técnica es ampliamente recomendada para la evaluación de modelos. En ella, los datos se dividen en k grupos, utilizando k-1 grupos para entrenar el modelo y el grupo restante para validarlo. Este proceso se repite k veces, con un grupo diferente utilizado para la validación en cada iteración, lo que asegura que cada grupo se utilice como conjunto de prueba en algún momento. Los resulta-

dos finales se obtienen promediando los errores a lo largo de las k iteraciones. La validación cruzada k-fold soluciona los problemas de la técnica de retención, ya que minimiza los riesgos derivados de cómo se dividen los datos, dado que cada punto de datos se evalúa al menos una vez. A medida que el valor de k aumenta, la varianza del modelo disminuye. Los valores más comunes de k son 5 y 10. El principal inconveniente de este método es que requiere entrenar el modelo k veces, lo que implica un mayor tiempo de cómputo en comparación con el método de retención, ya que se entrena y valida el modelo k veces. La siguiente visualización muestra un ejemplo de validación cruzada con 5 pliegues y los resultados agregados (hipotéticos) de todas las iteraciones.

Figura 3

K-Fold Cross-Validation



Nota. Tomada de Applied supervised Learning with R, (p. 335), Ramasubramanian et al. (2019)

Matriz de confusión: el análisis del rendimiento de un modelo en algoritmos de clasificación comienza con la matriz de confusión. Como menciona Ramasubramanian et al. (2019, p.

236), esta matriz es una herramienta que muestra de manera clara cómo se distribuyen las predicciones de cada clase en comparación con los valores reales de cada clase.

Figura 4

Matriz de Confusión

		Predicted	
		No	Yes
Actual	No	True Negative	False Positive
	Yes	False Negative	True Positive

Nota. Tomada de Applied supervised Learning with R, (p. 236), Ramasubramanian et al. (2019)

TN: valores False que fueron predichos como False

FN: valores True que fueron predichos como False

FP: valores False que fueron predichos como True

TP: valores True que fueron predicho como True

AUC - Area Under the Curve: “Representa la probabilidad de que el modelo, si se da un ejemplo positivo y negativo elegido al azar, clasificará el positivo mayor que el negativo”. (Google Developers, 2024)

Curva ROC (Receiver Operating Characteristic): Google Developers (2024) lo define como: una representación visual del rendimiento del modelo en todos los umbrales. La versión larga del nombre, característica operativa del receptor, es una retención. de la detección de radares de la Segunda Guerra Mundial.

La curva ROC se dibuja calculando la tasa de verdaderos positivos (TPR) por sus siglas en inglés y una tasa de falsos positivos (FPR) por sus siglas en inglés, en cada umbral posible (en la práctica, a los intervalos seleccionados), luego se grafica la TPR sobre la FPR. Un modelo perfecto que en algún umbral tiene una TPR de 1.0 y una FPR de 0.0.

Marco Normativo

Ley 30 de 1992 – Ley General de Educación Superior

Regula el sector de la educación superior en Colombia, estableciendo que esta debe ser un proceso continuo, orientado al desarrollo integral de los individuos y su formación tanto académica como profesional. Además, la ley asegura la autonomía universitaria y la calidad de la educación, fomentando la libertad en los ámbitos de enseñanza, aprendizaje, investigación y cátedra.

Además, resalta el papel fundamental de la educación superior como un servicio público cultural y subraya su contribución al progreso científico, cultural, económico, político y ético del país.

Ley 1753 de 2015 – Plan Nacional de Desarrollo

Conocida como el Plan Nacional de Desarrollo 2014-2018 "Todos por un nuevo país", establece que la educación es un pilar clave para el progreso del país. En el ámbito de la educación superior, el plan subraya la necesidad de reducir las desigualdades en el acceso y la calidad educativa, promoviendo la equidad de oportunidades para todos los ciudadanos. Además, tiene como objetivo fortalecer la competitividad y la infraestructura estratégica, al tiempo que fomenta la movilidad social y el desarrollo sostenible.

Decreto 1075 de 2015 – Decreto Único Reglamentario del Sector Educación

Denominado Decreto Único Reglamentario del Sector Educación en Colombia, tiene como propósito consolidar y simplificar las normativas existentes en el ámbito educativo. Este decreto unifica en un solo cuerpo normativo diversas disposiciones que anteriormente estaban distribuidas en múltiples decretos y resoluciones.

Recopila y actualiza la legislación vigente en el sector educativo, facilitando su consulta y aplicación. Abarca la educación formal, no formal e informal. Reconoce y regula la autonomía

de las instituciones de educación superior, asegurando su derecho a autogobernarse y gestionar sus propios recursos.

El decreto busca eliminar redundancias y conflictos en las normativas, ofreciendo un marco claro y coherente. Establece lineamientos y mecanismos para garantizar y mejorar la calidad educativa en todos los niveles. Promueve la evaluación y acreditación de instituciones y programas educativos, e implementa políticas y programas para fomentar la inclusión educativa, asegurando un acceso equitativo para todos los ciudadanos.

Además, incluye medidas específicas dirigidas a poblaciones vulnerables y personas con necesidades educativas especiales.

Ley 1581 de 2012 – Ley de Protección de Datos Personales

Establece un marco normativo para proteger la información personal de los ciudadanos, asegurando su derecho a la privacidad y un manejo adecuado de sus datos.

Entre los principales aspectos de la Ley se encuentran:

Derechos de los Titulares de Datos: Los ciudadanos tienen el derecho a conocer, actualizar y corregir sus datos personales. También pueden solicitar prueba de la autorización otorgada para el tratamiento de sus datos y ser informados sobre el uso que se les da.

Obligaciones de los responsables del Tratamiento: Las entidades encargadas de recolectar y manejar datos personales deben obtener una autorización previa, explícita e informada del titular. Además, deben garantizar la confidencialidad, integridad y seguridad de los datos recolectados, implementando medidas adecuadas para su protección.

La ley establece sanciones tanto administrativas como económicas para las entidades que incumplan sus disposiciones. La Superintendencia de Industria y Comercio es la entidad encargada de supervisar el cumplimiento de la ley y atender las reclamaciones de los titulares.

Esta es fundamental para proteger la privacidad y los datos personales de los ciudadanos en Colombia. Define un marco claro para el manejo adecuado de la información y asegura derechos esenciales, lo que refuerza la confianza en el uso de los datos personales por parte de las organizaciones.

Ley 1324 de 2009 – Ley Sobre los Resultados de las Pruebas Saber

Establece lineamientos y criterios para la organización del sistema de evaluación de la calidad educativa en Colombia. Su propósito es promover una cultura de evaluación continua y fortalecer la supervisión y control por parte del Estado en el sector educativo. Además, la ley reestructura al ICFES (Instituto Colombiano para el Fomento de la Educación Superior) y establece normativas para la evaluación externa e imparcial de los establecimientos educativos y los programas académicos.

Entre los principales aspectos de la Ley se tienen:

Parámetros y Criterios de Evaluación: La evaluación debe ser objetiva, equitativa, comparable, periódica, confidencial y relevante. Se fomenta la participación de la comunidad educativa en la creación de herramientas y estrategias para la evaluación.

Cultura de Evaluación: Se busca instaurar una cultura de evaluación dentro del sistema educativo, con el fin de facilitar la supervisión estatal. La evaluación externa e imparcial será llevada a cabo por académicos externos coordinados por el ICFES.

Transformación del ICFES: La ley redefine al ICFES, otorgándole la responsabilidad de realizar la evaluación externa e independiente de la calidad educativa.

Metodología

La metodología aplicada en este trabajo de investigación consta de seis fases. Inicialmente, se procederá a la obtención y recolección de los datos necesarios para la investigación. Luego, se explorarán los datos obtenidos para identificar problemas y patrones relevantes. Posteriormente, se realizará la preparación de los datos para asegurar su calidad y adecuación para el análisis. En la fase de modelado, se aplicarán técnicas estadísticas y de aprendizaje automático para desarrollar modelos predictivos. A continuación, los modelos serán ajustados y evaluados según su desempeño y capacidad para cumplir con los objetivos planteados. Finalmente se mostrarán los resultados obtenidos.

Obtención de Datos

Inicialmente los datos se obtuvieron del sitio web del ICFES, datos abiertos, filtrando estos datos para dar alcance y contexto dentro de la investigación, obteniendo los que coincidieron con el criterio de búsqueda, es decir, estudiantes de la UNAD del programa ingeniería de sistemas, de la ciudad de Bogotá, entre los años 2019 a 2022. El resultado obtenido es un archivo de tipo texto (csv), por cada uno de los años. Cabe resaltar que no en todos los periodos se encontró la misma información, ya que algunas variables que se manejan en el año 2021 y 2022, no están presentes en la información de los años 2019 y 2020, por los cual se hizo necesario realizar una depuración de estas variables para poder contar con conjuntos de datos con información estandarizada. Los campos son los siguientes:

Tabla 1

Variables Categóricas del Conjunto de Datos

Campo	Descripción	Opciones de Respuesta
ESTU_GENERO	Género del estudiante	F – femenino
		M - masculino
PERIODO	Periodo de aplicación de la prueba	20183 – Profesionales
		20182 - Profesionales en exterior
		20195 - Profesionales
		20194 - Profesionales exterior
		20202 - Profesionales exterior
		20203 – Profesionales
		20212 - Profesionales exterior
		20213 – Profesionales
		20222 - Profesionales exterior
		20223 – Profesionales
ESTU_PAGOMATRICULABECA	Variable que define si el pago de matrícula es por beca	Si
		No
ESTU_CURSODOCENTESIES	Se preparó para el examen Saber Pro en su IES con docentes de la institución (número de horas)	No
		Menos de 20 horas
		Entre 20 y 30 horas
		Más de 30 horas
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de su vivienda según recibo de energía eléctrica	Estrato 1
		Estrato 2
		Estrato 3
		Estrato 4
		Estrato 5
		Estrato 6
		Sin Estrato
		Si
FAMI_TIENEINTERNET	¿Su hogar cuenta con servicio o conexión a internet?	No
		Si
FAMI_TIENECOMPUTADOR	¿Cuáles de los siguientes bienes posee su hogar?: Computador	No
		No

ESTU_HORASSEMANATRABAJA	¿Trabaja usted actualmente?	Más de 30 horas
		Entre 11 y 20 horas
		Entre 21 y 30 horas
		Menos de 10 horas
		0

Tabla 2

Variables Numéricas del Conjunto de Datos

Campo	Descripción	Opciones de Respuesta
MOD_RAZONA_CUANTITAT_PUNT	Puntaje razonamiento cuantitativo	Numérica - Rango [0, 300]
MOD_RAZONA_CUANTITATIVO_PNAL	Percentil nacional razonamiento cuantitativo	Numérica - Rango [1, 100]
MOD_RAZONA_CUANTITATIVO_PGREF	Percentil por grupo de referencia razonamiento cuantitativo	Numérica - Rango [1, 100]
MOD_RAZONA_CUANTITAT_DESEM	Nivel de desempeño razonamiento cuantitativo	Numérica - Rango [1, 4]
MOD_LECTURA_CRITICA_PUNT	Puntaje lectura crítica	Numérica - Rango [0, 300]
MOD_LECTURA_CRITICA_PNAL	Percentil nacional lectura crítica	Numérica - Rango [1, 100]
MOD_LECTURA_CRITICA_PGREF	Percentil por grupo de referencia lectura crítica	Numérica - Rango [1, 100]
MOD_LECTURA_CRITICA_DESEM	Nivel de desempeño lectura crítica	Numérica - Rango [1, 4]
MOD_COMPETEN_CIUDADA_PUNT	Puntaje competencias ciudadanas	Numérica - Rango [0, 300]
MOD_COMPETEN_CIUDADA_PNAL	Percentil nacional competencias ciudadanas	Numérica - Rango [1, 100]
MOD_COMPETEN_CIUDADA_PGREF	Percentil por grupo de referencia competencias ciudadanas	Numérica - Rango [1, 100]
MOD_COMPETEN_CIUDADA_DESEM	Nivel de desempeño competencias ciudadanas	Numérica - Rango [1, 4]
MOD_INGLES_PUNT	Puntaje inglés	Numérica - Rango [0, 300]
MOD_INGLES_DESEM	Nivel de desempeño inglés	"-A1"

		"A1"
		"A2"
		"B1"
		"B2"
MOD_INGLES_PNAL	Percentil nacional inglés	Numérica - Rango [1, 100]
MOD_INGLES_PGREF	Percentil por grupo de referencia inglés	Numérica - Rango [1, 100]
MOD_COMUNI_ESCRITA_PUNT	Puntaje comunicación escrita	Numérica - Rango [0, 300]
MOD_COMUNI_ESCRITA_DESEM	Nivel de desempeño comunicación escrita	Numérica - Rango [1, 4]
MOD_COMUNI_ESCRITA_PNAL	Percentil nacional comunicación escrita	Numérica - Rango [1, 100]
MOD_COMUNI_ESCRITA_PGREF	Percentil por grupo de referencia comunicación escrita	Numérica - Rango [1, 100]
PUNT_GLOBAL	Puntaje total obtenido	Numérica - Rango [0, 300]

El conjunto de datos finalmente se compone por doce variables categóricas y veintiuna variables numéricas.

Exploración de Datos

Para llevar a cabo esta investigación, se hace uso del lenguaje de programación Python, el cual se despliega desde un entorno de Anaconda haciendo uso de la diversidad de paquetes y librerías óptimos para ciencia de datos que esta ofrece.

Los archivos de tipo texto, se cargan en el entorno de Anaconda, mediante la librería Pandas, para su posterior manipulación.

Una vez los datos están cargados, se inicia una exploración de estos. Inicialmente se obtienen las medias del puntaje global de los estudiantes por cada año, la frecuencia de género de los estudiantes que presentaron las pruebas Saber Pro del programa de ingeniería de sistemas de

la UNAD sede Bogotá, en los años 2019 a 2022, inconsistencias en las clases de las variables categóricas, datos nulos o faltantes.

Análisis de Correlación

Para este análisis se toman las variables numéricas del conjunto de datos, las cuales indican el puntaje por cada módulo de la prueba Saber Pro y el puntaje global, para identificar y medir la asociación entre variables.

Preparación y Limpieza de Datos

Es posible identificar inconsistencias en las clases de las variables categóricas, se evidencian errores de escritura en las clases, esto hace que se tome como clases independientes cuando realmente se quiere indicar un registro de la misma clase.

Se encontraron 19 registros vacíos, los cuales se eliminaron del conjunto de datos junto con las clases inconsistentes, la variable "ESTU_HORASSEMANATRABAJA", muestra errores de escritura en sus clases como es el caso de "No Especifica" y "No especifica", las cuales se corrigen y se unen dentro de una sola clase. En el caso de la variable "FAMI_ESTRATOVIVIENDA", clases que no son significativas dentro de la variable como es el caso de ("Estrato 6", "Estrato 5", "Sin Estrato"), estas se eliminan de la variable. Al finalizar esta limpieza el conjunto de datos cuenta con un total de 2677 registros.

Transformación de Datos

La primera transformación que se realiza a los datos es la codificación, es decir, tomar las variables categóricas cuyas clases son texto como por ejemplo la variable "FAMI_ESTRATOVIVIENDA", que tiene diferentes clases como (Estrato 1, Estrato 2, Estrato 3, Estrato 4, Estrato 5, Sin estrato), este tipo de representación de datos es inadecuado para un algoritmo de Machine Learning, ya que generalmente estos trabajan con variables numéricas.

Para la codificación se hace uso de la librería One-Hot-Encoding de Python. Esta técnica convierte cada clase de la variable categórica en una columna con valores de 1 y 0.

La segunda transformación que se realiza es tomar la variable “PUNT_GLOBAL”, y dividirla en dos clases: “Por encima de la media” la cual tomará un valor numérico de 1 y la otra clase es “Por debajo de la media”, la cual tomará un valor numérico de 0. El punto de referencia que se toma para realizar la división es el cálculo de la media de la variable “PUNT_GLOBAL”, la cual tiene un valor de 143.83.

Una vez realizadas las transformaciones correspondientes, se unen los conjuntos de datos, es decir, las variables codificadas y la variable resultante de “PUNT_GLOBAL”, para formar un nuevo conjunto de datos el cual se va a utilizar para posteriormente aplicar el modelo de Machine Learning.

Selección de Variables

Esta etapa es de gran importancia para la construcción de modelos de Machine Learning por varias razones:

Mejora la precisión del modelo, ya que incluir variables irrelevantes o redundantes puede afectar de una manera negativa la capacidad de un modelo para predecir nuevos datos.

Ayuda a reducir la complejidad de un modelo, cuando menos variables se tengan y que están sean las que realmente aportan información significativa, el modelo será más fácil de interpretar y entender.

Reduce el tiempo y costos asociados a recopilación y procesamiento de datos.

En el proceso de construcción del modelo se hace uso de tres técnicas de selección de variables:

PCA (Análisis de componentes principales), transforma las variables originales en un conjunto de variables que no están correlacionadas las cuales reciben el nombre de componentes principales, cada uno de los componentes principales explican una parte de la variabilidad de los datos, lo que permite reducir la dimensionalidad sin perder información relevante.

Selectkbest, esta técnica aplica una prueba estadística para cada variable de manera individual, y selecciona un número (k) de variables de acuerdo con la mejor puntuación. Las variables que se seleccionan son las que muestran una mejor relación significativa con la variable objetivo.

RFE (Eliminación recursiva de características), recorre de manera iterativa el conjunto de variables, eliminando en cada iteración las menos importantes, hasta alcanzar un número óptimo de variables relevantes para el modelo.

Construcción de Modelos

Con las variables seleccionadas, se generan los conjuntos de entrenamiento y de prueba, para poder entrenar los modelos. Para el conjunto de entrenamiento se toma el 80% del total de los datos y el 20% restante para el conjunto de prueba.

Tamaño del conjunto de entrenamiento: 2141 registros

Tamaño del conjunto de prueba: 536 registros

Ajuste de los Modelos

Validación Cruzada y Optimización de Hiperparámetros para el modelo de regresión logística: Lo primero es establecer el tipo de algoritmo, en este caso el primero que se va a realizar es el de regresión logística, para ello, se hace uso de la librería LogisticRegression del paquete scikit-learn de Python.

Seguido se establece la cuadrilla de hiperparámetros con los hiperparámetros que se van a evaluar. Para este caso en el modelo de regresion logística se evaluarán los siguientes parámetros:

Parámetro C: permite controlar la regularización del modelo, esto es útil para evitar el sobreajuste (overfitting), es decir evita que el modelo se ajuste demasiado a los datos de entrenamiento, ya que aprendería solo los patrones de los datos de entrenamiento, pero conservaría los patrones específicos de estos datos, lo cual llevaría a fallas al utilizar datos nuevos. Un valor pequeño de C indica una regularización fuerte (más penalización) y un valor grande indica una regularización débil (menos penalización). Se establecen los siguientes valores: 0.01, 0.1, 1, 10, 100.

Parámetro penalty: especifica el tipo de penalización o regularización que se aplicará al modelo para evitar el sobreajuste. Los valores que se establecen son: l1 regularización Lasso, l2 regularización Bridge,

Parámetro Solver: indica el algoritmo utilizado para optimizar la función de pérdida de la regresión logística. La función que se establece es 'liblinear', ya que es compatible con el tipo de penalización l1 y l2.

Una vez se establecen los hiperparámetros a evaluar para el modelo de regresión logística, se utiliza la técnica de Búsqueda de cuadrícula (grid search), donde se prueba múltiples combinaciones de hiperparámetros definidos en la cuadrilla de parámetros para encontrar la combinación que optimiza el rendimiento del modelo según una métrica específica (como 'accuracy').

Los parámetros que se establecen en la cuadrícula de búsqueda son los siguientes:

Estimator: acá se define el modelo que en este caso es el de regresión logística que se definió anteriormente.

Param_grid: es la lista de hiperparámetros que se desean evaluar, ya se definió anteriormente en la cuadrilla de hiperparámetros.

Cv: es el número de particiones (folds) que se van a utilizar en la validación cruzada (cross-validation), ayuda a evaluar el modelo de forma más robusta, reduciendo el riesgo de depender de un solo conjunto de validación.

Scoring: define la métrica que se va a utilizar para evaluar el modelo, en este caso se utiliza ‘accuracy’ (mide el porcentaje de predicciones correctas).

Teniendo los parámetros definidos, se entrena el modelo de regresion lineal utilizando los datos de entrenamiento tanto las entradas (variables predictoras), como la salida (variable dependiente).

GridSearch ajustará el modelo (estimator) usando estos datos y evaluará cada combinación de parámetros en las particiones de validación de la siguiente manera:

Prueba cada combinación de parámetros en param_grid.

Para cada combinación, realiza validación cruzada (con cv=5).

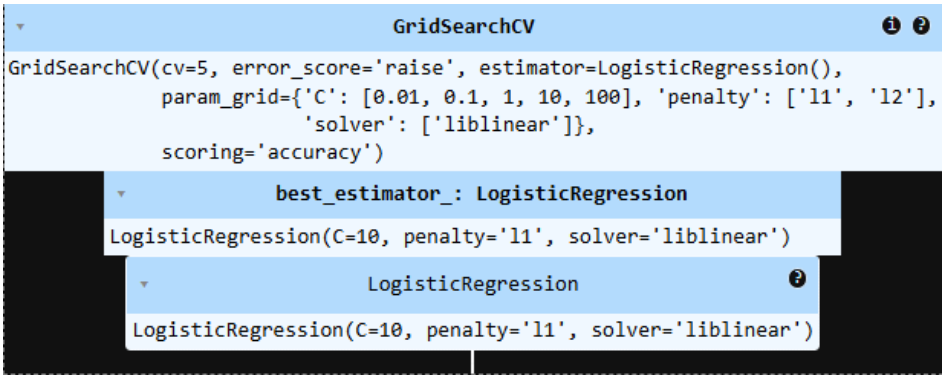
Calcula la métrica accuracy en cada pliegue y promedia los resultados.

Almacena los resultados de todas las combinaciones.

De acuerdo con los valores de los hiperparámetros establecidos, se obtienen los siguientes resultados como se observa en la siguiente gráfica:

Figura 5

Parámetros de la Cuadrícula de Búsqueda



Nota. resultados con el mejor estimador para el modelo de Regresión Logística

La información que entrega la gráfica es la siguiente: para el modelo de regresión logística, se establece que los mejores hiperparámetros son: C = 10, penalty = `l1' y solver = 'liblinear'. Se obtiene un 'accuracy' de 0.592

El resultado 0.592 es el mejor puntaje promedio de validación cruzada obtenido por el modelo durante el proceso de búsqueda de hiperparámetros con GridSearchCV.

Validación Cruzada y Optimización de Hiperparámetros para el modelo de árbol de decisión: Establecer el modelo de árbol de decisión para obtener el estimador, se usa la función DecisionTreeClassifier del paquete sklearn.tree.

Se establece la cuadrilla de hiperparámetros con los hiperparámetros que se van a evaluar. Para este caso en el modelo de árbol de decisión se evaluarán los siguientes parámetros:

Parámetro 'max_depth': determina la profundidad máxima del árbol, es decir la cantidad máxima de nodos que va a contener el árbol. Los valores que se evaluarán son: 3, 4, 5, 6, 7, 10, None.

Parámetro 'min_samples_split': determina la cantidad mínima de muestras que debe contener un nodo para dividirlo. Los valores que se evaluarán son: 2, 5, 10.

Parámetro 'min_samples_leaf': determina la cantidad mínima de muestras por hoja. Los valores que se evaluarán son: 1, 5, 10.

Con la cuadrilla de hiperparámetros establecida para el modelo de árbol de decisión, se utiliza la técnica de Búsqueda de cuadricula (grid search), donde se prueba múltiples combinaciones de hiperparámetros definidos en la cuadrilla de parámetros para encontrar la combinación que optimiza el rendimiento de un modelo según una métrica específica (como 'accuracy').

Estimator: acá se define el modelo que en este caso es el de árbol de decisión que se definió anteriormente.,

param_grid: es la lista de hiperparámetros que se desean evaluar, ya se definió anteriormente en la cuadrilla de hiperparámetros del árbol de decisión.

Cv: es el número de particiones (folds) que se van a utilizar en la validación cruzada (cross-validation), ayuda a evaluar el modelo de forma más robusta, reduciendo el riesgo de depender de un solo conjunto de validación.

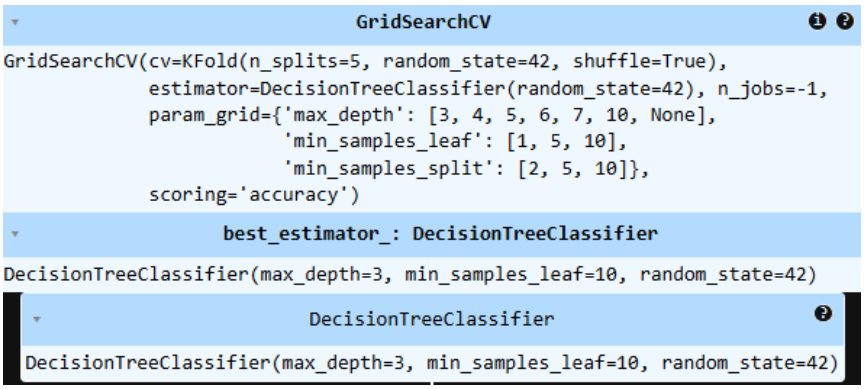
Scoring: define la métrica que se va a utilizar para evaluar el modelo, en este caso se utiliza ‘accuracy’ (mide el porcentaje de predicciones correctas).

n_jobs=-1: se utiliza para especificar el número de trabajos (procesos) que se ejecutarán en paralelo al realizar la búsqueda en cuadrícula (grid search). Esto permite aprovechar múltiples núcleos de CPU para acelerar el proceso de entrenamiento y evaluación de modelos. (-1) utiliza todos los núcleos disponibles en la CPU. Esto significa que GridSearchCV ejecutará tantas tareas en paralelo como núcleos tenga la CPU de la máquina en la que se está ejecutando el código.

De acuerdo con los valores de los hiperparámetros establecidos, se obtienen los siguientes resultados como se observa en la siguiente gráfica:

Figura 6

Parámetros de la Cuadrícula de Búsqueda



Nota. resultados con el mejor estimador para el modelo de Árbol de decisión

La información que entrega la gráfica es la siguiente: para el modelo de regresión logística, se establece que los mejores hiperparámetros son: max_depth = 3, min_samples_leaf = 10. Se obtiene un ‘accuracy’ de 0.5894

El resultado 0.5894 es el mejor puntaje promedio de validación cruzada obtenido por el modelo durante el proceso de búsqueda de hiperparámetros con GridSearchCV.

Evaluación de los Modelos

Modelo de Regresión Logística

Después de obtener los mejores hiperparámetros se realizan predicciones con el conjunto de datos de prueba. Al realizar un reporte de clasificación de los datos predichos se obtienen los siguientes resultados:

Figura 7

Reporte de Clasificación del Modelo de Regresión Logística

```
Exactitud en el conjunto de prueba: 0.6380597014925373
Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.61      0.70      0.65       262
           1       0.67      0.58      0.62       274

    accuracy                           0.64       536
   macro avg       0.64      0.64      0.64       536
weighted avg       0.64      0.64      0.64       536

Matriz de confusión:
 [[183  79]
 [115 159]]
```

La exactitud determina la proporción de clasificaciones correctas sobre el total de clasificaciones realizadas. En este caso, aproximadamente el 63.8% de las clasificaciones fueron correctas en el conjunto de prueba. Las clases que se indican se toman de la siguiente manera: clase 1, puntajes que están por encima de la media del puntaje global, clase 0, puntajes que están por debajo de la media del puntaje global

Desempeño por Clase. Clase 0: (Precisión: 0.61, Recall (Recuperación): 0.70, F1-Score: 0.65):

La precisión del 61% indica que, de todas las predicciones positivas para la Clase 0, el 61% fueron correctas.

La recuperación (recall) del 70% indica que el modelo pudo identificar correctamente el 70% de las instancias de Clase 0.

La F1-score de 0.65 representa un equilibrio entre precisión y recuperación.

Clase 1: (Precisión: 0.67, Recuperación: 0.58, F1-Score: 0.62):

La precisión del 67% indica que, de todas las predicciones positivas para la Clase 1, el 67% fueron correctas.

La recuperación del 58% indica que el modelo pudo identificar correctamente el 58% de las instancias de Clase 1.

La F1-score de 0.62 representa un equilibrio entre precisión y recuperación.

Promedio General. Macro Avg y Weighted Avg: Ambas métricas tienen valores similares de 0.64 para precisión, recuperación y F1-score, lo que representa un desempeño equilibrado entre las dos clases.

Matriz de Confusión. Clase 0: 183 instancias fueron correctamente clasificadas como Clase 0.

79 fueron incorrectamente clasificadas como Clase 1.

Clase 1: 159 instancias fueron correctamente clasificadas como Clase 1.

115 fueron incorrectamente clasificadas como Clase 0.

Curva ROC y AUC. El AUC (Área bajo la curva ROC) se mide en una escala de 0 a 1, resume el desempeño del modelo en una cifra única.

0.5: El modelo es equivalente a un clasificador aleatorio (sin capacidad para distinguir clases).

>0.5 y <0.7: Desempeño bajo, pero mejor que el azar.

0.7 a 0.8: Desempeño aceptable/moderado.

0.8 a 0.9: Buen desempeño.

>0.9: Desempeño excelente (puede indicar overfitting si las métricas en validación son mucho más bajas).

Modelo de Árbol de Decisión

Después de obtener los mejores hiperparámetros se realizan predicciones con el conjunto de datos de prueba. Al realizar un reporte de clasificación de los datos predichos se obtienen los siguientes resultados:

Figura 8

Reporte de Clasificación del Modelo de Árbol de decisión

	precision	recall	f1-score	support
0	0.60	0.53	0.57	268
1	0.58	0.65	0.61	268
accuracy			0.59	536
macro avg	0.59	0.59	0.59	536
weighted avg	0.59	0.59	0.59	536

Figura 9

Matriz de Confusión del Àrbol de Decisión

```
[143 125]
[ 95 173]
```

Desempeño por Clase. Clase 0: (Precisión: 0.60, Recall (Recuperación): 0.53, F1-Score: 0.57):

La precisión del 60% indica que, de todas las predicciones positivas para la Clase 0, el 60% fueron correctas.

La recuperación (recall) del 53% indica que el modelo pudo identificar correctamente el 53% de las instancias de Clase 0.

La F1-score de 0.57 representa un equilibrio entre precisión y recuperación.

Clase 1: (Precisión: 0.58, Recuperación: 0.65, F1-Score: 0.61):

La precisión del 57% indica que, de todas las predicciones positivas para la Clase 1, el 57% fueron correctas.

La recuperación del 65% indica que el modelo pudo identificar correctamente el 65% de las instancias de Clase 1.

La F1-score de 0.61 representa un equilibrio entre precisión y recuperación.

Promedio General. Macro Avg y Weighted Avg: Ambas métricas tienen valores similares de 0.59 para precisión, recuperación y F1-score, lo que representa un desempeño equilibrado entre las dos clases.

Matriz de Confusión. Clase 0: 143 instancias fueron correctamente clasificadas como Clase 0, 125 fueron incorrectamente clasificadas como Clase 1.

Clase 1: 95 instancias fueron correctamente clasificadas como Clase 1, 173 fueron incorrectamente clasificadas como Clase 0.

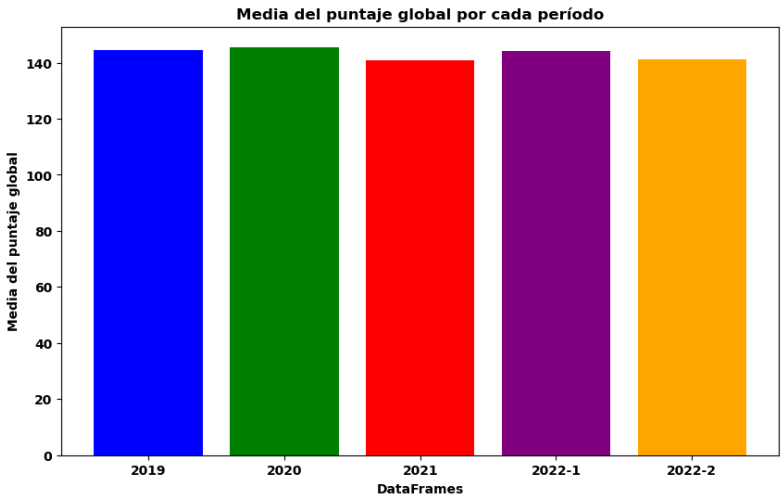
Resultados

Resultados de la Exploración de Datos

Se obtienen las medias del puntaje global de los estudiantes por cada año.

Figura 10

Promedio Puntaje Global, Discriminado Por Años



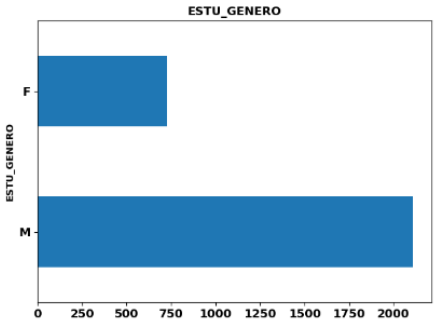
Como en la gráfica no se evidencia una diferencia significativa entre la media de los resultados globales de los estudiantes, entonces, se va a trabajar con el conjunto de datos de todos los años unidos en uno solo.

La cantidad total inicial de registros es de 2834. Ahora se procede a hacer una exploración al conjunto de datos, lo cual permite detectar diferentes anomalías tales como valores nulos o faltantes en los registros del conjunto de datos, valores atípicos, el tipo de variables que están contenidas en los datos, datos inconsistentes y la frecuencia de las clases de las variables categóricas.

Por ejemplo, se puede ver la frecuencia de género de los estudiantes que presentaron las pruebas Saber Pro en los años 2019 a 2022.

Figura 11

Frecuencia de Género



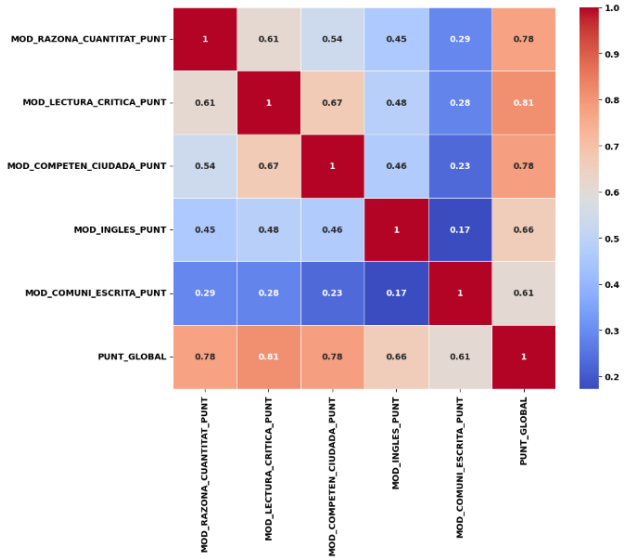
Nota. Estudiantes que presentaron las pruebas Saber Pro en los años 2019 a 2022

Resultados del Análisis de Correlación

Se puede observar una buena correlación entre cada uno de los módulos con el puntaje global, y una baja correlación entre los módulos.

Figura 12

Matriz de Correlación



Nota. puntaje de los módulos de las pruebas Saber Pro con el puntaje global

Resultados de la Preparación y Limpieza de Datos

Figura 13

Inconsistencias en las Clases de las Variables Categóricas

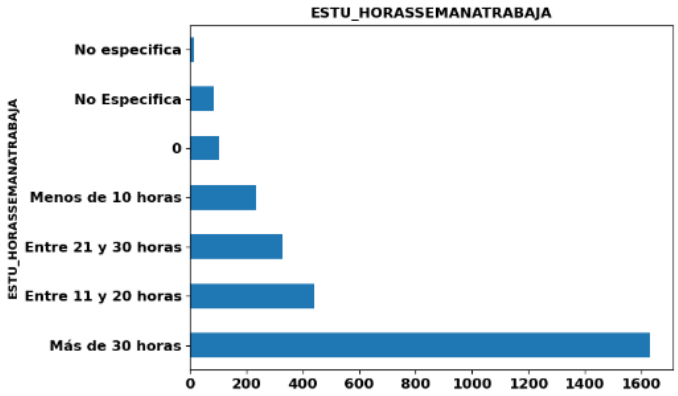
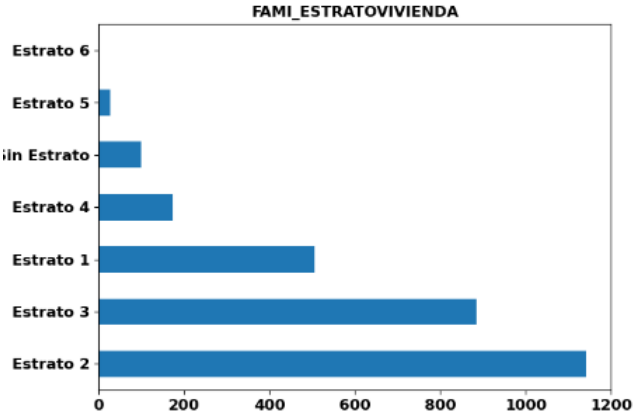


Figura 14

Frecuencia



Nota. Estrato de los estudiantes que presentaron las pruebas Saber Pro en los años 2019 a 2022

Resultados de la Transformación de Datos

Figura 15

Fragmento de la Tabla Generada con la Codificación One-Hot

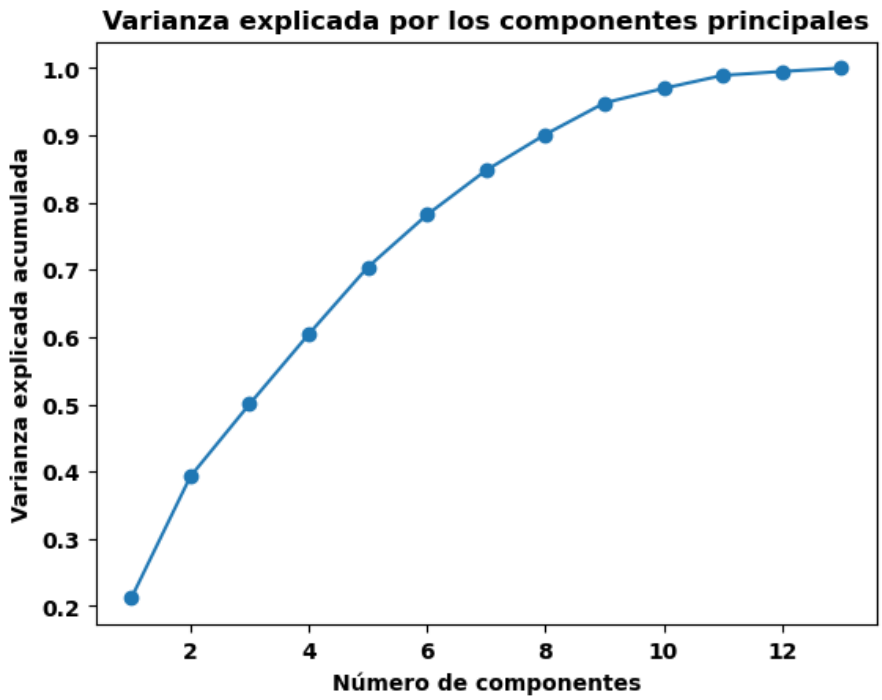
	ESTU_COMOCAPACITOEXAMENSB11_Repasó por cuenta propia	ESTU_COMOCAPACITOEXAMENSB11_Tomó un curso de preparación	FAMI_ESTRATOVIVIENDA_Estrato 2	FAMI_ESTRATOVIVIENDA_Estrato 3
0	1.0	0.0	1.0	0.0
1	1.0	0.0	1.0	0.0
2	1.0	0.0	0.0	1.0
3	1.0	0.0	0.0	0.0
4	1.0	0.0	0.0	1.0

Resultados de la Aplicación de Técnicas de Selección de Datos

La siguiente gráfica muestra la cantidad de variabilidad de los datos que explican los componentes de PCA, lo que permite establecer el número adecuado de componentes principales.

Figura 16

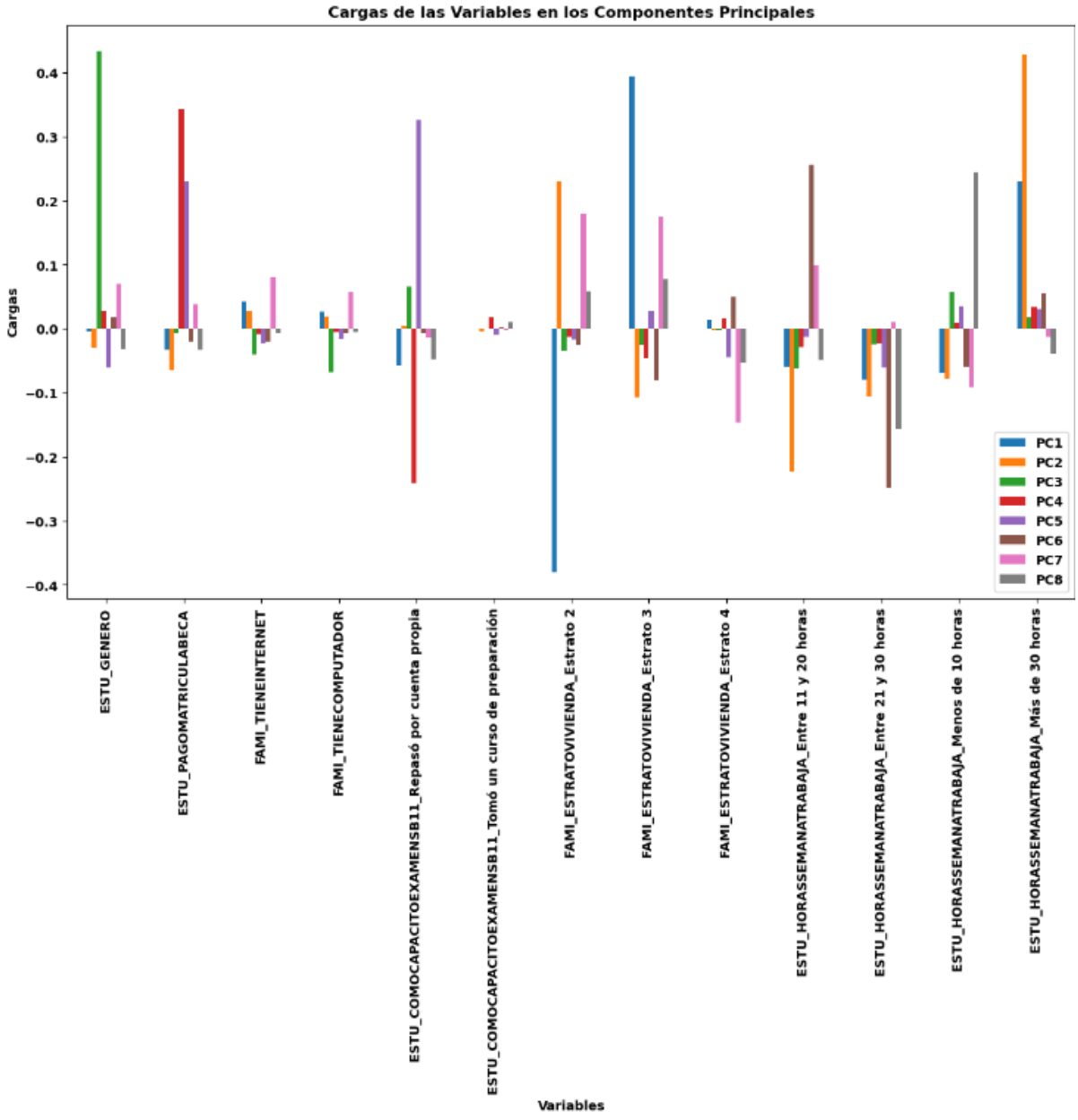
Varianza Explicada Acumulada por los Componentes Principales



La siguiente gráfica permite visualizar las cargas de las variables en los 8 componentes principales.

Figura 17

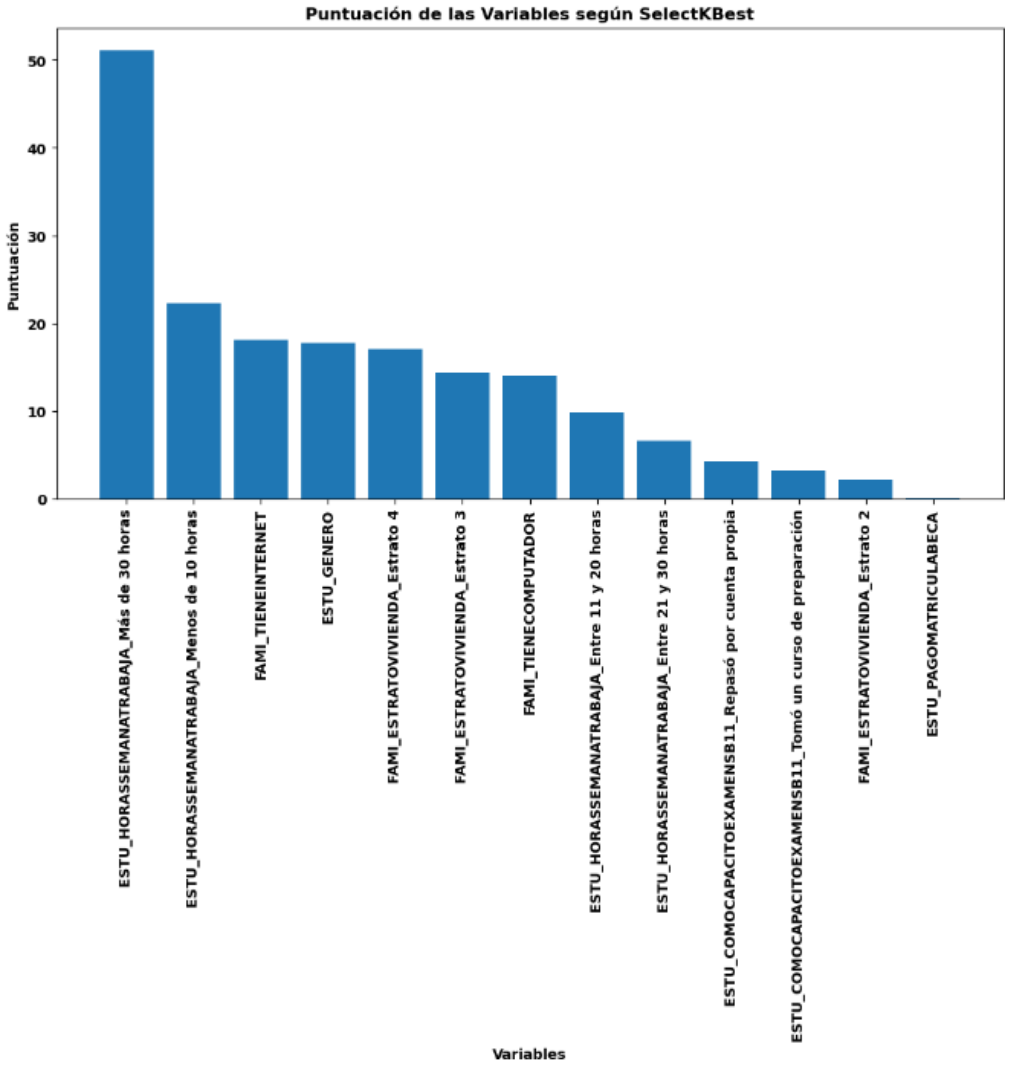
Carga de las Variables del Conjunto de Datos en Cada Componente Principal



La siguiente gráfica permite visualizar qué variables tienen mayor relevancia según el método SelectKBest

Figura 18

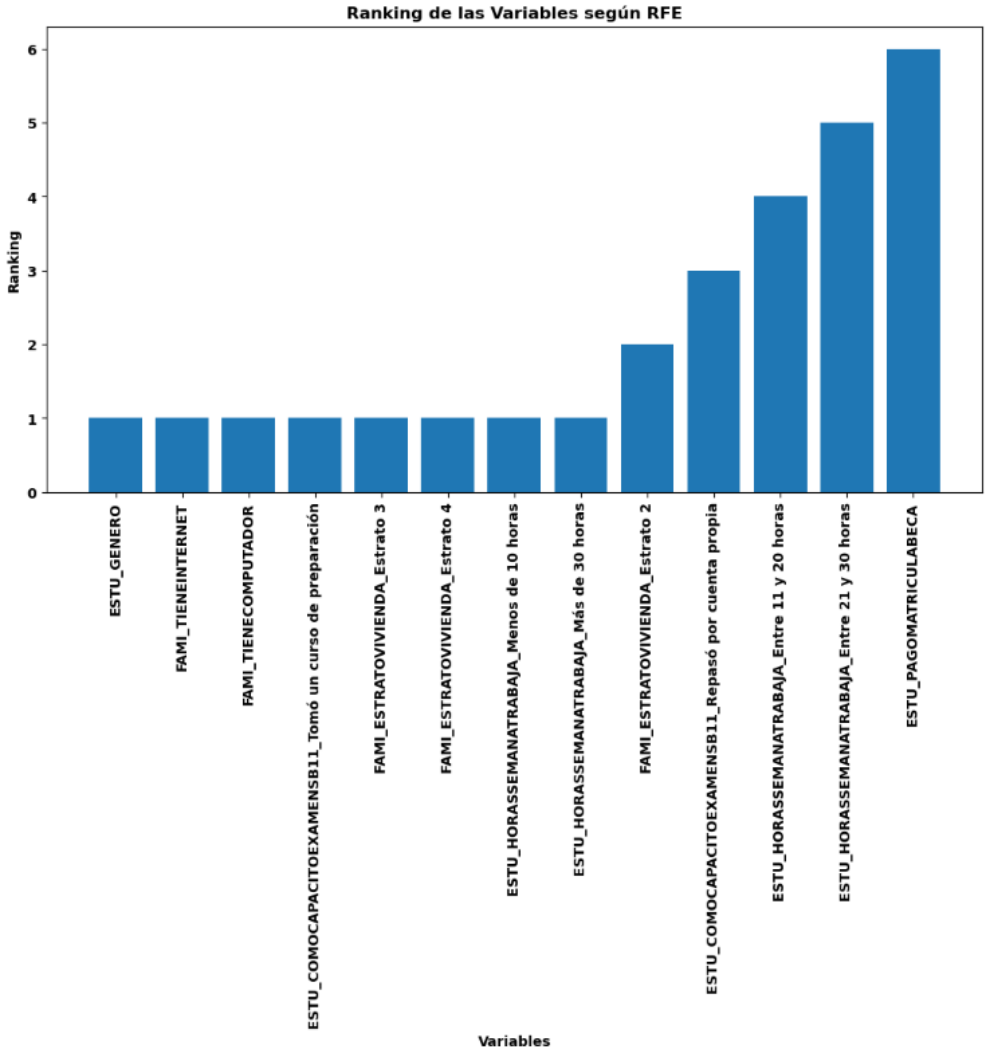
Importancia de las Variables Según el Método SelectKBest



En la siguiente gráfica, un ranking más bajo indica mayor importancia. Por lo tanto, las variables con barras más bajas son consideradas más importantes para el modelo.

Figura 19

Importancia de las Variables Según el Método RFE



En la siguiente tabla se muestran las variables que seleccionó cada uno de los métodos:

Tabla 3

Variables Seleccionadas por Cada Uno de los Métodos

PCA	RFE	SelectKBest
ESTU_GENERO	ESTU_GENERO	ESTU_HORASSEMANATRA-BAJA_Mas de 30 hrs
FAMI_TIENEINTERNET	FAMI_TIENEINTERNET	ESTU_HORASSEMANATRA-BAJA_Menos 10 hrs
FAMI_TIENECOMPUTADOR	FAMI_TIENECOMPUTADOR	FAMI_TIENEINTERNET
ESTU_COMOCAPACITOEXA-MENSB11_Repaso cuenta propia	FAMI_ESTRATOVI-VIENDA_Estrato 3	ESTU_GENERO
FAMI_ESTRATOVIVIENDA_Estrato 3	FAMI_ESTRATOVI-VIENDA_Estrato 4	FAMI_ESTRATOVI-VIENDA_Estrato 4
ESTU_HORASSEMANATRA-BAJA_Menos 10 hrs	ESTU_HORASSEMANATRA-BAJA_Menos 10 hrs	FAMI_ESTRATOVI-VIENDA_Estrato 3
ESTU_HORASSEMANATRA-BAJA_Mas de 30 hrs	ESTU_HORASSEMANATRA-BAJA_Mas de 30 hrs	FAMI_TIENECOMPUTADOR

De acuerdo con la tabla 3, solo difieren en una variable, por lo cual esta no se tendrá en cuenta. Después de aplicar las técnicas de selección de variables y comparar los resultados de cada una de ellas, se seleccionaron las variables que se identificó eran consistentes y comunes en las tres técnicas de selección utilizadas, las cuales se muestran a continuación:

Tabla 4

Variables Finales

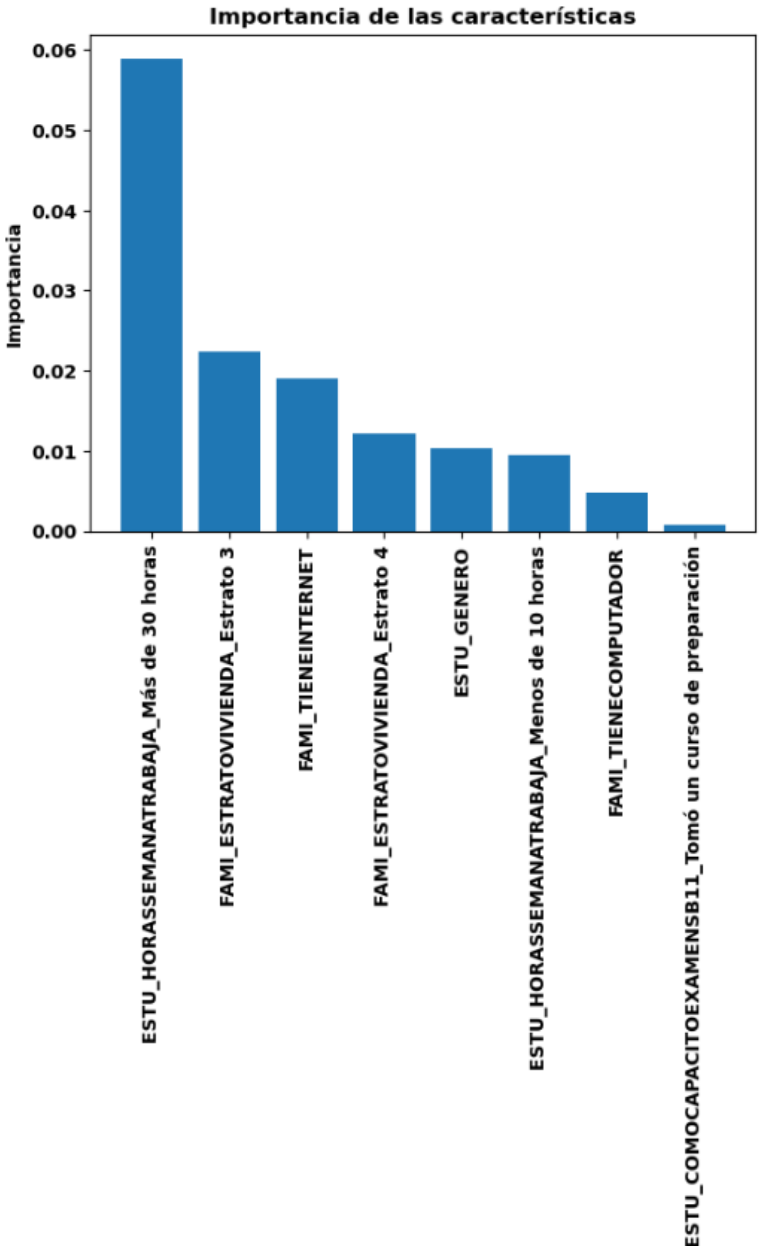
Variables Seleccionadas
ESTU_HORASSEMANATRABAJA_Mas de 30 hrs
ESTU_HORASSEMANATRABAJA_Menos 10 hrs
FAMI_TIENEINTERNET
ESTU_GENERO
FAMI_ESTRATOVIVIENDA_Estrato 4
FAMI_ESTRATOVIVIENDA_Estrato 3
FAMI_TIENECOMPUTADOR

Resultados del Modelo de Regresión Logística

La siguiente gráfica muestra las variables con coeficientes más grandes en valor absoluto. Estas variables tienen un mayor impacto en la predicción del modelo de Regresión logística.

Figura 20

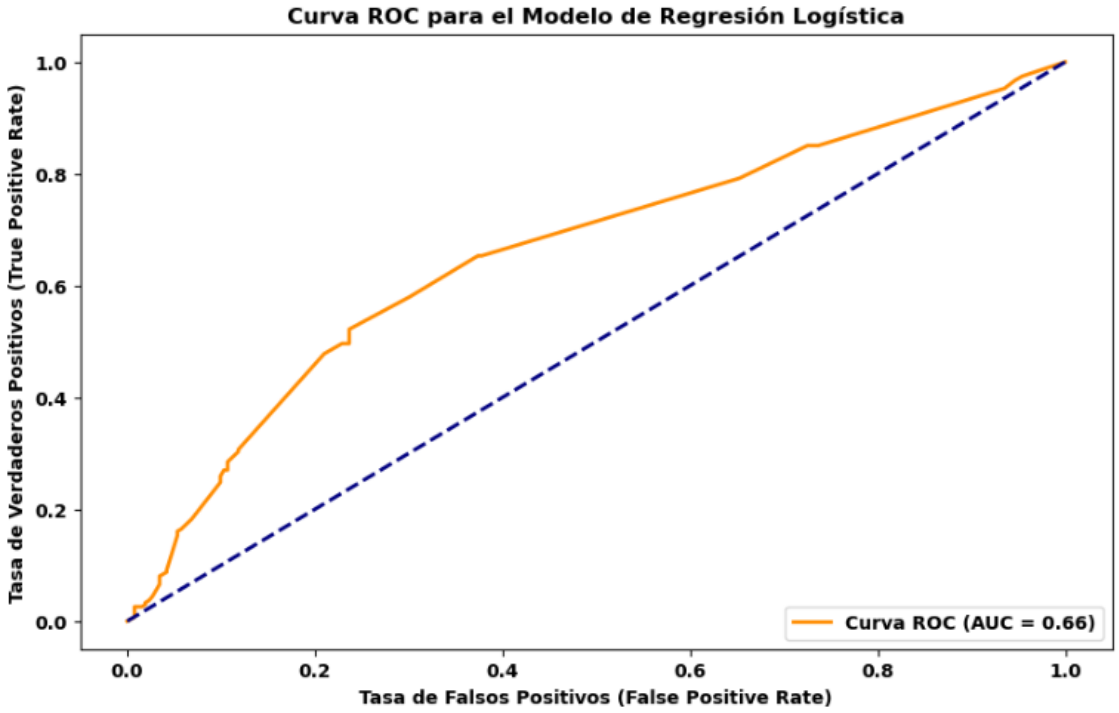
Variables Más Significativas Para el Modelo de Regresión Logística



Curva ROC: La curva muestra la capacidad del modelo para distinguir entre las clases positivas y negativas a diferentes umbrales. Una curva más cercana a la esquina superior izquierda indica un mejor desempeño del modelo.

Figura 21

Curva ROC Modelo de Regresión Logística.



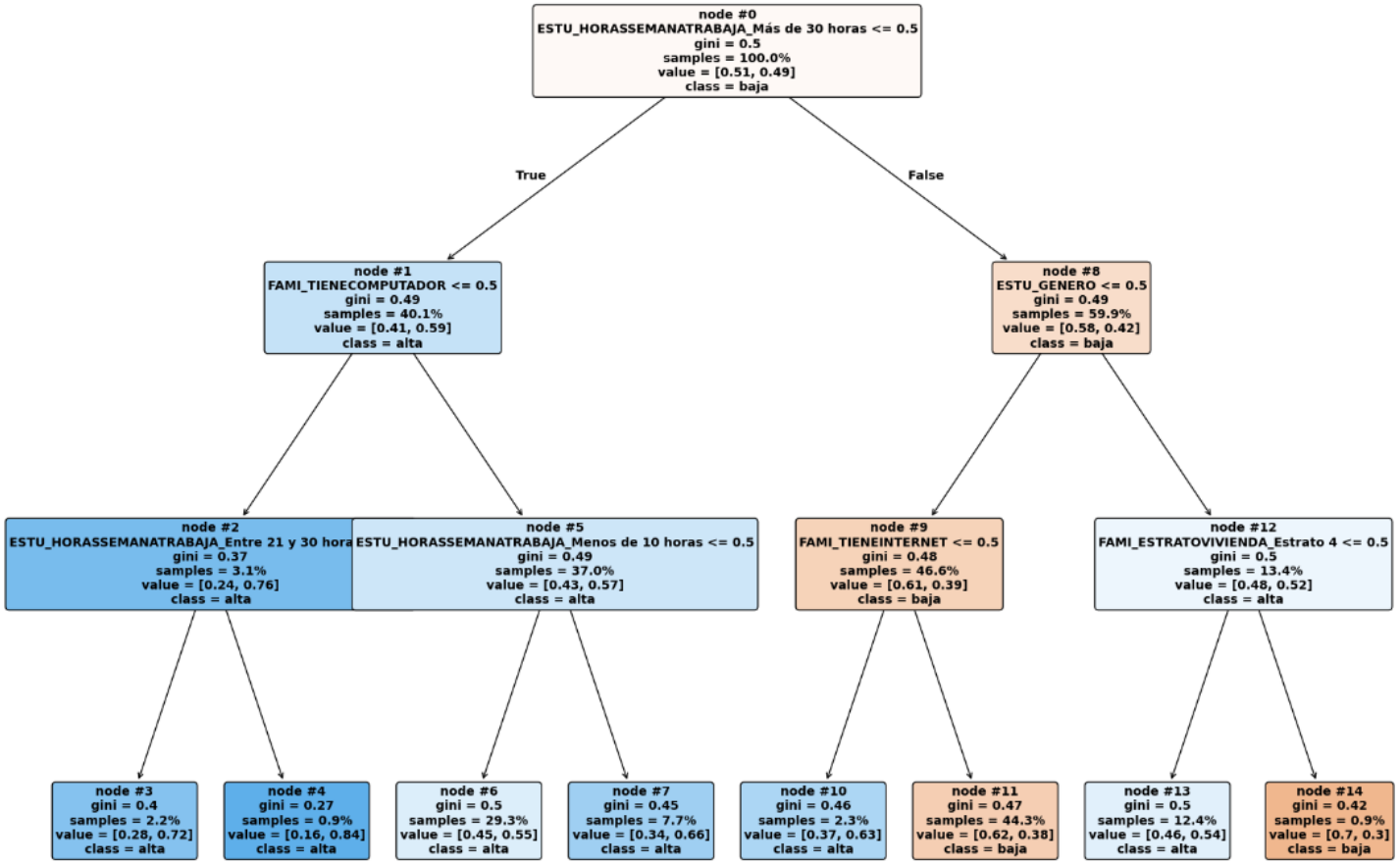
Un AUC de 0.66, indica que hay un 66% de probabilidad de que el modelo asigne una puntuación más alta a una instancia positiva que a una negativa. Esto significa que el modelo tiene una capacidad moderada para identificar las clases.

Resultados del Modelo de Árbol de Decisión

La gráfica del árbol de decisión es una representación visual que facilita la comprensión del proceso de toma de decisiones del modelo. Facilita la interpretación de las condiciones, la pureza de los nodos, la distribución de las clases y la estructura general del árbol.

Figura 22

Esquema del Árbol de Decisión

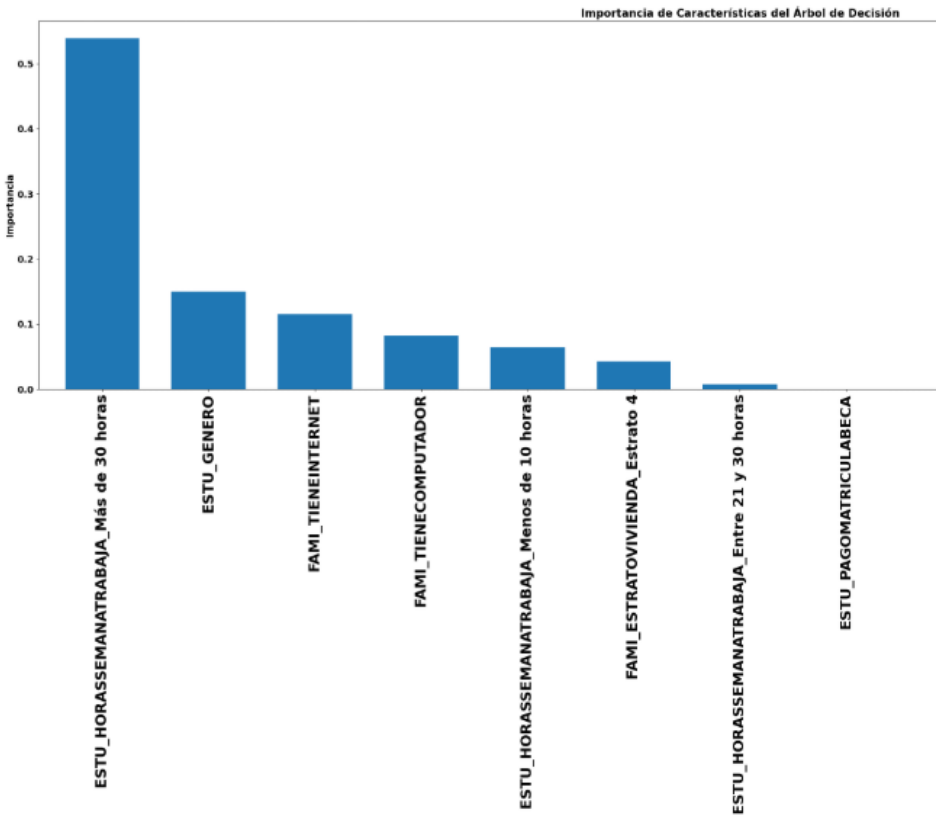


Nota. Generado por el modelo, tomando las variables que considera más significativas.

La gráfica de barras muestra la importancia relativa de cada variable en la toma de decisiones del modelo de árbol de decisión. La importancia de una característica indica cuánto contribuye esa variable a la predicción del modelo. Las barras más altas a la izquierda representan las características que tienen mayor influencia en las decisiones del modelo. Estas variables son consideradas las más importantes para la predicción de la variable objetivo. Las barras más bajas a la derecha representan las características que tienen menor influencia en las decisiones del modelo. La gráfica permite identificar de manera visual cuáles son las variables más importantes, cuáles se deben priorizar en el análisis, ingeniería de características, o para recolectar datos adicionales.

Figura 23

Variables que el Modelo Considera más Relevantes.

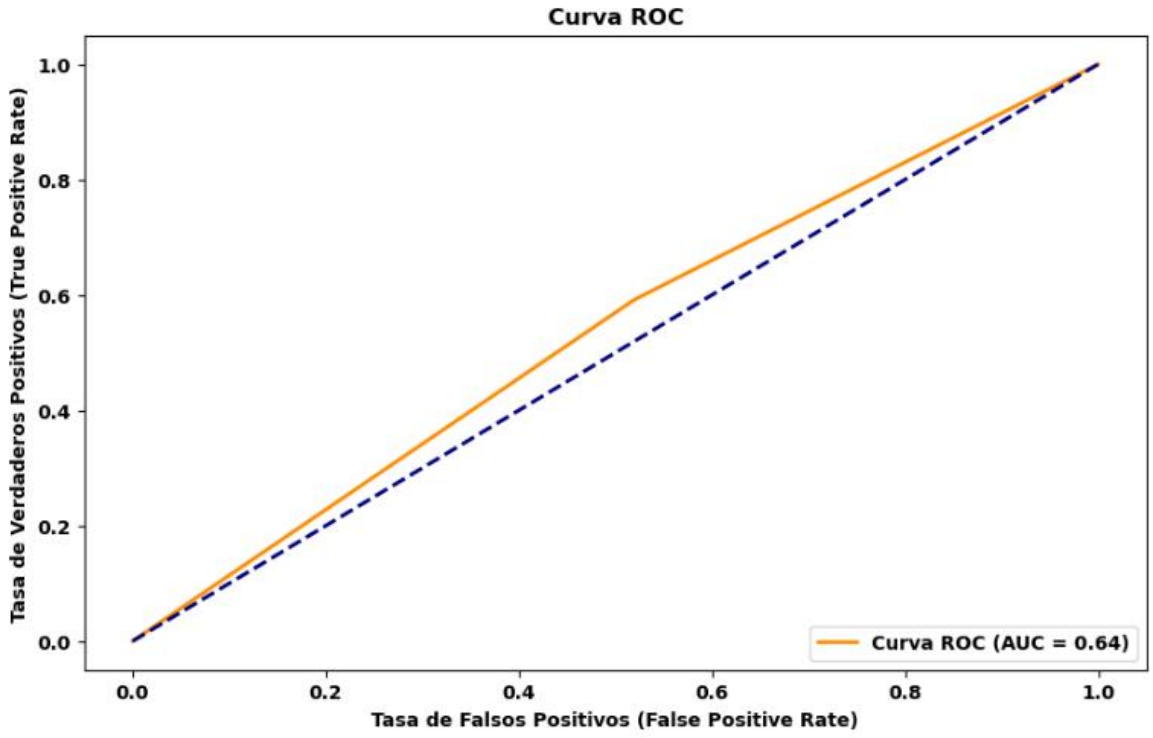


Curva ROC y AUC

Se utilizan otras métricas adicionales para la evaluación del modelo: la curva ROC y AUC (Área bajo la curva). Las dos se muestran en la siguiente gráfica.

Figura 24

Curva ROC Modelo del Árbol de Decisión.



Un AUC de 0.64, indica que hay un 64% de probabilidad de que el modelo asigne una puntuación más alta a una instancia positiva que a una negativa. Esto significa que el modelo tiene una capacidad moderada para identificar las clases.

Conclusiones

Con la aplicación de técnicas para seleccionar variables y el uso de algoritmos de regresión logística y árboles de decisiones, se logró identificar un conjunto de factores demográficos y socioeconómicos que afectan la calificación de los estudiantes. Esta calificación se puede establecer en relación con la media, es decir, si está por encima o por debajo del puntaje global obtenido en las pruebas Saber Pro de los estudiantes del programa de Ingeniería de Sistemas de la UNAD en Bogotá entre 2019 y 2022. Estos resultados ofrecen información significativa para comprender cómo estos factores influyen en el rendimiento académico de los estudiantes en estas pruebas estandarizadas.

Los factores demográficos y socio económicos que más relevancia tuvieron de acuerdo con el algoritmo de regresión logística son: el estudiante trabaja más de treinta horas a la semana, el estudiante es estrato tres o cuatro, el estudiante cuenta con internet y el género del estudiante. Con el algoritmo de árbol de decisión: el estudiante trabaja más de treinta horas a la semana, el género del estudiante, el estudiante cuenta con internet, el estudiante cuenta con computador.

Según las métricas analizadas, se determina que la efectividad del modelo no alcanza niveles superiores, dado que todas las variables son categóricas. Esto obligó a realizar un proceso de codificación para transformarlas en variables numéricas, creando columnas con valores de 1 y 0. Como resultado, el modelo no consigue identificar con mayor exactitud una correlación entre las variables predictoras y la variable objetivo a predecir.

Recomendaciones

Para realizar estudios más precisos, se recomienda a las organizaciones mejorar la calidad y cantidad de los datos, en lugar de agrupar en categorías amplias, considerar subcategorías más específicas que capturen matices importantes. Los algoritmos requieren variables numéricas, ya sean discretas o continuas, para poder tener un rendimiento óptimo, por lo que se recomienda a las organizaciones en este caso al ICFES, modificar la presentación de los datos en los casos donde una variable que es numérica se da en forma de rangos, esto se puede evidenciar en variables como la cantidad de horas que trabaja un estudiante a la semana.

De acuerdo con los resultados obtenidos del modelo de árbol de decisión y el modelo de regresión lgística, se identificó que el factor "ESTU_HORASSEMANATRABAJA_Más de 30 horas" es el principal determinante en el rendimiento académico de los estudiantes. Este factor es el nodo raíz del modelo, lo que significa que trabajar más de 30 horas a la semana afecta significativamente el desempeño, clasificando a la mayoría de estos estudiantes en la categoría de bajo rendimiento. Para el caso de los estudiantes que trabajan más de treinta horas a la semana, se recomienda ofrecer programas de orientación que les ayude a manejar su tiempo de manera eficiente y a priorizar su educación, para el desarrollo de habilidades útiles en el contexto académico y profesional, como la disciplina, la gestión del tiempo y la responsabilidad. Además, proveer acompañamiento psicoeducativo para asegurarse de que el equilibrio entre trabajo y estudios no afecte su bienestar.

De acuerdo con los resultados obtenidos en el modelo de regresion logística y del árbol de decisión, se identificó que factores relacionados con la infraestructura tecnológica como FAMI_TIENEINTERNET (nodo #9) y FAMI_TIENEPC (nodo #1) tienen una influencia significativa en el rendimiento académico de los estudiantes. Aquellos estudiantes que no tienen acceso

a internet o a un computador en su vivienda muestran una mayor tendencia a clasificarse en la categoría de bajo rendimiento. Para aquellos estudiantes que tengan dificultades en adaptar la infraestructura tecnológica a la modalidad virtual, es decir, que no cuenten con un plan de internet en su vivienda o con un dispositivo de cómputo, la institución puede considerar, establecer acuerdos con empresas de telecomunicaciones para ofrecer planes de internet de bajo costo o gratuitos a estudiantes con dificultades económicas, implementar un programa de préstamo de computadores, tablets o dispositivos móviles para aquellos que no tengan acceso a tecnología adecuada y habilitar puntos de acceso en municipios o zonas rurales, como bibliotecas públicas o centros de aprendizaje, equipados con internet y computadores, para que los estudiantes puedan conectarse.

La metodología utilizada en este trabajo de investigación es útil para aplicar como base en futuras investigaciones donde se quiera obtener información más especifica referente a cada uno de los módulos de las pruebas Saber Pro, y determinar qué factores demográficos y socioeconómicos influyen en los resultados de cada módulo, ya que esta investigación se centra en el puntaje global de los resultados de los estudiantes del programa de ingeniería de sistemas de la UNAD sede Bogotá.

Referencias

Anaconda. (2024). *Anaconda Documentation*. https://docs.anaconda.com/

Cano, M. A. (2024). *Técnicas de Machine Learning para la predicción del rendimiento académico en las pruebas 1 Saber Pro en Colombia*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://repository.unad.edu.co/bitstream/handle/10596/62951/mgarcia.pdf?sequence=1&isAllowed=y

Cifuentes Medina, J., Chacón Benavides, J., & Fonseca Correa, L. (2020). *Análisis de los resultados de las Pruebas Saber Pro en estudiantes de la licenciatura en Educación Básica de la Universidad Pedagógica y Tecnológica de Colombia (UPTC).* https://revistasum.umanizales.edu.co/ojs/index.php/plumillaeducativa/article/view/3833

Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing.

Franco, J. A. (2017). *Factores demográficos, académicos y socio económicos que influyen en los resultados del componente genérico de las pruebas Saber Pro.* chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://core.ac.uk/download/pdf/217558165.pdf

Función Pública. (2024). *Decreto 1075 de 2015 Sector Educación*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=77913

Función Pública. (2024). *Ley 1324 de 2009*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=36838

Función Pública. (2024). *Ley 1581 de 2012*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Función Pública. (2024). *Ley 1753 de 2015 – Plan Nacional de Desarrollo*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=61933

Función Pública. (2024). *Ley 30 de 1992 - Ley General de Educación Superior*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=253

Google Developers. (septiembre de 2024). *Conceptos de AA.* https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419

Gregoria Castañeda, Miguel Ruiz, Olga Viloria, Rosa Castañeda, & Yajaira Quevedo. (noviembre de 2007). *El rol de las universidades en el contexto de la responsabilidad social empresarial*. 100-132. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.redalyc.org/pdf/782/78230805.pdf

Julián Pérez Porto, & Ana Gardey. (mayo de 2021). *Definición de*. https://definicion.de/socioeconomico/

Karthik Ramasubramanian, & Jojo Moolayil. (2019). *Applied supervised Learning with R*.

Madhavan, S. (2015). *Mastering Python for Datascience*. Packt Publishing Ltd.

Numpy developers. (2024*). NumPy Documentation*. https://numpy.org/doc/1.24/user/absolute_beginners.html

Nyuytiymbiy, K. (diciembre de 2020). *Medium*. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

Oviedo Carrascal, A., & Jiménez Giraldo, J. (2019). *Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO.* https://revistas.elpoli.edu.co/index.php/pol/article/view/1499

Pandas. (2024). *Pandas documentation*. https://pandas.pydata.org/docs/getting_started/overview.html

Python.org. (2024). *Python documentation*. https://docs.python.org/3/tutorial/index.html

Real Academia Española. (2024). *Diccionario de la lengua española*. https://dle.rae.es/universidad

scikit-learn developers. (2020). *sklearn.preprocessing.OneHotEncoder*. https://qu4nt.github.io/sklearn-doc-es/modules/generated/sklearn.preprocessing.OneHotEncoder.html

scikit-learn developers. (2024). *Scikit-learn*. https://scikit-learn.org/stable/getting_started.html

scikit-learn developers. (2024). *Scikit-learn Recursive feature elimination*. https://scikit-learn.org/1.5/modules/generated/sklearn.feature_selection.RFE.html

scikit-learn developers. (2024) *Scikit-learn Univariate feature selection*. https://scikit-learn.org/1.5/modules/feature_selection.html#univariate-feature-selection

Sposob, G. (octubre de 2024). *Concepto*. https://concepto.de/factores-demograficos/#ixzz8rfpDBV61

Timarán Pereira, S., Caicedo Zambrano, S., & Hidalgo Troya, A. (2021). *Aplicación de la Minería de Datos en la Detección de Patrones de Desempeño Académico en las Pruebas Saber Pro*. https://sired.udenar.edu.co/7024/

Vargas, G. M. (2014). *Factores asociados al rendimiento académico tomando en cuenta el nivel socioeconómico: Estudio de regresión múltiple en estudiantes universitarios*. https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-42582014000100007