Predicción de puntajes de admisión en programas de ciencias de la Universidad Industrial de Santander

Lilia Ester Pérez Suárez

Sergio Antonio López Calvachi

Asesor

Jorge Eliecer Ospino Portillo

Universidad Nacional Abierta y a Distancia UNAD

La Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencias de Datos y Analítica

2025

Nota de Aceptación

Jorge Eliecer Ospino Portillo

Director de Trabajo de Grado

_____ _____

Jurado Jurado

2025

Resumen

Este proyecto de investigación busca desarrollar un modelo predictivo para los puntajes de admisión a los programas de pregrado de la Facultad de Ciencias de la Universidad Industrial de Santander (UIS) utilizando algoritmos de aprendizaje automático. Esta investigación surge de la necesidad de abordar la falta de orientación vocacional e información para la toma de decisiones académicas entre los estudiantes de último año de bachillerato. Mediante la integración de datos demográficos y académicos, el modelo de machine learning (ML) desarrollado proporcionará una herramienta que ayudará a los estudiantes a tomar decisiones informadas sobre sus futuras trayectorias académicas. Este enfoque innovador no solo mejorará la orientación académica, sino que también contribuirá a la satisfacción y éxito profesional de los estudiantes. Los resultados esperados incluyen un modelo de aprendizaje automático para la predicción de los puntajes de admisión en programas de Ciencias de la Universidad Industrial de Santander.

Palabras claves: Predicción, UIS, ICFES, Saber11, Aprendizaje.

Abstract

This research project aims to develop a predictive model for admission scores to the undergraduate programs of the Faculty of Sciences at the Universidad Industrial de Santander (UIS) using machine learning algorithms. This investigation is motivated by the lack of vocational guidance and adequate information for academic decision-making among final-year high school students. By integrating demographic and academic data, the developed machine learning (ML) model will provide a tool to help students make informed decisions about their future academic paths. This innovative approach will not only enhance academic guidance but also contribute to students' professional satisfaction and success. Expected outcomes include a machine learning model for the prediction of admission scores in Science programs at the Universidad Industrial de Santander.

Keywords: Prediction, UIS, ICFES, Saber11, Learning.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El paso de la educación media a la educación superior marca un momento crucial en la vida de los jóvenes, definido por decisiones que influirán en su futuro académico y profesional. En Colombia, este proceso se ve afectado por la falta de información clara y una orientación vocacional insuficiente, dificultando que los estudiantes de último grado seleccionen programas de pregrado acordes a sus intereses y habilidades. Esta situación puede generar decisiones inadecuadas, afectando su trayectoria educativa y su inserción laboral.

El desafío es especialmente evidente entre quienes buscan ingresar a programas de alta demanda, como los ofrecidos por la Facultad de Ciencias de la Universidad Industrial de Santander (UIS), reconocida como una de las mejores instituciones públicas del país. Aunque la admisión en la UIS se basa en el Examen de Estado Saber 11, cada programa aplica ponderaciones específicas sobre las áreas evaluadas, incrementando la complejidad para los aspirantes al intentar estimar sus probabilidades de ingreso. Aunque existen estudios que modelan el puntaje global del Saber 11, se identificó la ausencia de investigaciones centradas en la predicción del puntaje ponderado de admisión en la Facultad de Ciencias de la UIS, dejando a los estudiantes sin herramientas de apoyo cuantitativo.

Frente a este vacío, el uso de técnicas de aprendizaje automático (Machine Learning) emerge como una opción valiosa, permitiendo analizar grandes volúmenes de datos y descubrir patrones complejos que facilitan predicciones más acertadas. En este contexto, la pregunta que guía esta investigación es: ¿cómo pueden los modelos de aprendizaje automático contribuir a predecir los puntajes de admisión a los programas de la Facultad de Ciencias de la UIS, apoyando la toma de decisiones de los aspirantes?

El objetivo general de este trabajo es desarrollar modelos predictivos que estimen los puntajes de admisión utilizando algoritmos de Machine Learning. Para ello, se adoptó una metodología cuantitativa y predictiva, basada en datos públicos del ICFES del Examen Saber 11 de 2020, complementados con información histórica del desempeño de las instituciones educativas. Los objetivos específicos fueron: 1) Revisar la literatura sobre modelos predictivos y factores asociados al rendimiento académico; 2) Estimar y comparar distintos modelos predictivos (Regresión Lineal con Selección Hacia Atrás, Lasso y Ridge), evaluando su desempeño; y 3) Analizar las variables predictoras más relevantes según los modelos generados.

El documento se organiza en varias secciones: primero, se presenta el planteamiento del problema y la justificación del estudio; luego, el marco teórico y conceptual. A continuación, se describe la metodología, incluyendo las fuentes de datos, el preprocesamiento de la información y los modelos implementados. Posteriormente, se exponen los resultados obtenidos de la evaluación de los modelos predictivos y el análisis de las variables relevantes. Finalmente, se formulan las conclusiones principales y recomendaciones para futuras aplicaciones e investigaciones.

Se espera que este estudio aporte un modelo útil para estimar los puntajes de admisión, contribuya a mejorar la orientación académica de los estudiantes y fortalezca el conocimiento sobre los factores que inciden en el acceso a la educación superior en Colombia.

Justificación

La toma de decisiones sobre la elección de una carrera universitaria es un proceso crucial para los estudiantes de grado 11, que a menudo se enfrentan a una falta de información adecuada. Esta carencia puede llevar a decisiones subóptimas y, en última instancia, a insatisfacción profesional. En el contexto específico de los programas de ciencias de la Universidad Industrial de Santander (UIS), esta falta de orientación se convierte en un obstáculo aún más relevante. Este proyecto busca abordar esta necesidad imperante de información al proporcionar una herramienta de predicción de puntajes de admisión, lo que ayudaría a los estudiantes a tomar decisiones más informadas sobre sus futuros académicos y profesionales.

Se ha observado una falta de orientación vocacional para los jóvenes colombianos, como señalan Argüelles y Meléndez (2023), lo que subraya aún más la importancia de este estudio. Al emplear modelos de aprendizaje automático, este proyecto no solo busca predecir los puntajes de admisión, sino también ofrecer una guía vocacional a los estudiantes. Esto no solo aumentaría la probabilidad de que los estudiantes encuentren satisfacción en sus futuras carreras, sino que también contribuiría a su éxito y realización personal.

El impacto de esta investigación radica en que prestaría un servicio a la población estudiantil que está pensando en aplicar a los programas de ciencias de esta universidad, lo que la convierte en una de las instituciones más apetecidas por los estudiantes, según el más reciente informe del Center for World University Rankings 2022 2023 de acuerdo a Master (2023).

Según la revisión bibliográfica realizada, este trabajo sería el primero en abordar directamente la predicción del puntaje de admisión a los programas de la facultad de ciencias de esta universidad y por lo tanto sería pionera en ofrecer una referencia concreta para futuras investigaciones y desarrollos en este campo de la orientación educativa en Colombia.

Objetivos

Objetivo General

Predecir los puntajes de admisión a los programas de pregrado de la Facultad de Ciencias de la Universidad Industrial de Santander UIS, haciendo uso de algoritmos de Machine Learning.

Objetivos Específicos

Revisar en la literatura los modelos de Machine Learning y los factores asociados a la predicción del rendimiento académico en pruebas estandarizadas como SABER 11.

Estimar y comparar diferentes modelos predictivos para los puntajes de admisión a los programas de pregrado de la Facultad de Ciencias de la UIS, evaluando su capacidad predictiva mediante métricas de error.

Analizar la influencia de las variables predictoras en los puntajes de admisión, a partir de los modelos de Machine Learning estimados.

Descripción del Problema

La elección de carrera universitaria en Colombia se complica por la falta de información y orientación vocacional, un desafío particularmente agudo para los aspirantes a programas competitivos como los de la Facultad de Ciencias de la Universidad Industrial de Santander (UIS). La admisión a la UIS depende de los resultados del Saber 11, pero con ponderaciones específicas por programa que dificultan a los estudiantes estimar sus posibilidades reales de ingreso. A pesar de estudios que predicen el puntaje general Saber 11, no existen herramientas específicas para predecir el puntaje ponderado de admisión a Ciencias en la UIS. Esta brecha informativa incrementa el riesgo de decisiones de postulación subóptimas y la consecuente insatisfacción académica o profesional.

Planteamiento del Problema

Uno de los problemas más relevantes para los estudiantes de grado 11 que están pensando en aplicar a un programa universitario es la deficiencia en la información disponible para la toma de decisiones para la postulación universitaria. Los estudiantes que planean aplicar a los programas de ciencias de la Universidad Industrial de Santander (UIS) no son ajenos a esta problemática.

Este proyecto está enfocado en responder a la necesidad de información para la toma de decisiones de la comunidad de estudiantes de últimos grados de bachillerato que están en el proceso de orientación profesional y están pensando en postularse a un programa de pregrado de la Facultad de Ciencias de la UIS. Arguelles y Meléndez (2023) mencionan ‘la carencia de orientación vocacional para los jóvenes colombianos acerca de la elección de una carrera de pregrado acorde a sus aptitudes’. Además, los modelos de aprendizaje automático podrían ofrecer una orientación vocacional a los estudiantes que aspiran a iniciar su educación

profesional. Esto los ayudaría a tomar decisiones informadas sobre sus trayectorias profesionales, aumentando la probabilidad de satisfacción laboral, realización y éxito.

La UIS basa sus decisiones de admisión en los componentes del Examen de Estado Saber 11, aplicando ponderaciones específicas. El documento titulado “Criterios de Admisión” de la Universidad Industrial de Santander UIS (2022), actualizado el 17 de mayo de 2022, establece los requisitos y ponderaciones para la admisión a los programas de pregrado. Teniendo en cuenta lo anterior, la universidad calcula los puntajes de admisión.

Tabla 1

Ponderación para los Programas de la Facultad de Ciencias de la UIS

Programas Académicos	Ciencias Naturales	Lectura Crítica	Matemáticas	Sociales y Ciudadanas	Inglés
Matemáticas	10%	20%	50%	10%	10%
Lic. en Matemáticas	10%	20%	50%	10%	10%
Física	20%	25%	25%	20%	10%
Química	40%	15%	25%	10%	10%
Biología	35%	20%	25%	10%	10%

Nota. Criterios de Admisión Universidad Industrial de Santander UIS (2022). Tomado de la Universidad Industrial de Santander.

En la revisión de la literatura, no se encontró ningún trabajo que abordara directamente la modelación y predicción del puntaje de admisión a los programas de la UIS y menos aún para la Facultad de Ciencias de la UIS. Si bien en la literatura revisada se encuentran varios trabajos relacionados con la modelación y predicción del puntaje del examen de estado Saber 11, como se evidencia en los trabajos de Martínez et al. (2020), Peña y González (2022) y Soto et al. (2023).

Rodríguez et al. (2021) menciona que los estudios indican un marcado incremento en el empleo de técnicas de aprendizaje automático para predecir el rendimiento académico. No obstante, los métodos empleados para aplicar estos algoritmos implican múltiples puntos de decisión y la ajustada configuración de los parámetros del modelo, los cuales se aplican de manera individualizada, determinados más por elecciones personales que por metodologías uniformes.

En este contexto, surge una pregunta relevante: ¿Cómo pueden los modelos de aprendizaje automático ser utilizados para predecir de manera precisa los puntajes de admisión a los programas de pregrado de la Facultad de Ciencias de la Universidad Industrial de Santander, facilitando así una toma de decisiones más informada para los estudiantes?

Para dar respuesta al problema planteado, la presente investigación plantea una estrategia metodológica basada en el desarrollo de cinco modelos de predicción, uno para cada área evaluada en las pruebas Saber 11: Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales y Ciudadanas, e inglés. Para cada una de estas áreas se construirá un modelo predictivo independiente que permita estimar el puntaje que obtendría un estudiante, con base en las variables más correlacionadas previamente identificadas a través de análisis exploratorios.

Posteriormente, los puntajes predichos por área se utilizarán para calcular un puntaje de admisión estimado, aplicando las ponderaciones específicas que utiliza la Universidad Industrial de Santander (UIS) para cada uno de los programas académicos de la Facultad de Ciencias. La UIS asigna distintos pesos porcentuales a cada componente del examen Saber 11, dependiendo del enfoque formativo de cada programa, tal como se observa en la tabla de criterios oficiales de admisión.

Por ejemplo: Para el programa de Matemáticas, la ponderación asignada es del 50% para el área de Matemáticas, 20% para Lectura Crítica, y 10% para cada una de las otras áreas.

En contraste, el programa de Química otorga una ponderación del 40% a Ciencias Naturales, 25% a Matemáticas, y solo 15% a Lectura Crítica.

Con esta información, se obtendrá un puntaje total de admisión estimado por programa, lo que permitirá a los estudiantes simular y visualizar su posible resultado en procesos de selección.

Este enfoque tiene como finalidad apoyar la orientación vocacional y la toma de decisiones informada de los estudiantes de grado 11 que aspiran a ingresar a programas de ciencias, ofreciendo una herramienta basada en datos y en los criterios reales de admisión institucional.

Marco de Referencia

El desempeño académico y la predicción de resultados en pruebas estandarizadas constituyen un campo de investigación que ha evolucionado significativamente con la incorporación de modelos avanzados de análisis de datos. En particular, el uso de algoritmos de aprendizaje automático para predecir resultados de exámenes como el Saber 11 ha sido objeto de múltiples estudios (Martínez et al., 2020; Rodríguez Hernández et al., 2021; Peña y González, 2022). Estos estudios han contribuido a un mejor entendimiento de los factores que influyen en el rendimiento académico, permitiendo desarrollar modelos predictivos más precisos y adaptados a las necesidades educativas específicas.

Un aspecto central en el análisis del rendimiento académico es el enfoque en las variables socioeconómicas y demográficas que pueden influir en los resultados de los estudiantes. Investigaciones como la de Peña y González (2022) han demostrado la importancia de incorporar estas variables para mejorar la precisión de los modelos predictivos. Estos estudios han ayudado a entender cómo factores externos al desempeño académico, como el estrato socioeconómico y las condiciones de acceso a recursos tecnológicos, afectan significativamente los resultados en pruebas estandarizadas (Timarán et al., 2020; Fernández, 2017). Por ejemplo, los estudios de Timarán et al. (2020) han mostrado que estudiantes de estratos socioeconómicos altos y aquellos que asisten a colegios con mejores condiciones TIC tienen un mejor desempeño en lectura crítica.

Además, la adaptación de técnicas de minería de datos y aprendizaje automático en la educación ha permitido explorar nuevas metodologías para la evaluación y mejora continua del proceso educativo. Por ejemplo, Soto-Acevedo et al. (2023) desarrollaron un modelo que utiliza técnicas de machine learning para predecir el rendimiento en pruebas estandarizadas, lo cual

destaca la transición hacia enfoques más analíticos y basados en datos en el ámbito educativo (Soto y Acevedo, 2023; Burgos, 2021). Este cambio metodológico ha permitido la identificación de patrones complejos en los datos educativos que no eran visibles con técnicas estadísticas tradicionales, lo que ha llevado a una mejora en la precisión de las predicciones.

El marco teórico de esta investigación se basa en la premisa de que los modelos predictivos pueden ser herramientas esenciales para la orientación académica y vocacional de los estudiantes. La literatura revisada sugiere que un enfoque integrado que combine variables académicas, socioeconómicas y demográficas puede ofrecer predicciones más robustas y útiles para los procesos de admisión universitaria (Rodríguez Hernández et al., 2021; Moncayo, 2016). Además, el enfoque metodológico de esta investigación se alinea con el modelo CRISP DM, que ha demostrado ser efectivo en proyectos de minería de datos y ciencia de datos para proporcionar un marco estructurado y eficaz desde la comprensión del problema hasta la aplicación práctica del modelo predictivo (Soto-Acevedo et al., 2023; Timarán et al., 2020). CRISP DM facilita la sistematización de los proyectos de minería de datos, lo que es crucial para asegurar la replicabilidad y la validez de los resultados obtenidos.

Por último, es crucial considerar los avances tecnológicos y metodológicos en el campo de la ciencia de datos para adaptar y mejorar continuamente las herramientas de predicción en educación. El enfoque adaptativo permitirá no solo prever el rendimiento en pruebas estandarizadas sino también apoyar a los estudiantes en la toma de decisiones más informadas respecto a su futuro académico y profesional (Arguelles y Meléndez, 2023; Chica et al., 2010). Este enfoque también contribuirá a una mayor satisfacción profesional y éxito académico, alineándose con los objetivos de las instituciones educativas de mejorar la calidad y eficacia de sus programas académicos (Viana, 2016; Ardila, 2017). La integración de tecnologías como las

redes neuronales artificiales y el aprendizaje profundo (deep learning) ha mostrado ser particularmente prometedora para capturar relaciones no lineales y complejas en los datos educativos (Rodríguez, 2021).

Además, investigaciones recientes han explorado cómo la infraestructura tecnológica y las condiciones de aprendizaje, como el acceso a internet y recursos tecnológicos en el hogar, impactan el rendimiento académico. Estudios como los de Serna y Gómez (2021) han utilizado métricas como Chi Square y SVM para caracterizar a los estudiantes y mejorar la precisión de las predicciones. También, los trabajos de Burgos (2021) y Soto-Acevedo et al. (2023) han demostrado que las técnicas de deep learning pueden superar a los métodos tradicionales en la predicción del rendimiento académico.

En el contexto de la Universidad Industrial de Santander (UIS) utiliza las pruebas Saber 11 como criterio principal para la admisión a sus programas de pregrado. Sin embargo, los estudios previos no han abordado de manera específica la modelación y predicción de los puntajes de admisión en esta universidad. Este proyecto de investigación busca llenar ese vacío, aplicando técnicas avanzadas de machine learning para desarrollar un modelo predictivo que ayude a los estudiantes a tomar decisiones informadas sobre su futuro académico (Universidad Industrial de Santander, 2022).

Finalmente, la revisión de la literatura muestra un creciente interés en la aplicación de técnicas de machine learning para la mejora educativa. Investigaciones como las de Martínez et al. (2020), Peña y González (2022), y Soto-Acevedo et al. (2023) han establecido una base sólida sobre la cual se puede construir. Este proyecto no solo busca aplicar estos conocimientos, sino también expandirlos, proporcionando un modelo práctico y accesible para la predicción de puntajes de admisión en programas de ciencias en Colombia.

Metodología

El presente estudio adoptó un enfoque cuantitativo, predictivo y analítico para estimar los puntajes de admisión a los programas de la Facultad de Ciencias de la Universidad Industrial de Santander (UIS), haciendo uso de algoritmos de aprendizaje automático (Machine Learning). El objetivo fue construir modelos que permitan anticipar con un nivel aceptable de precisión el puntaje que un estudiante podría obtener en las pruebas de admisión, tomando como insumo tanto variables individuales como contextuales.

Para el desarrollo del modelo se utilizaron bases de datos públicas suministradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), correspondientes al Examen Saber 11 del año 2020. En una primera fase, se consolidaron los registros de estudiantes que presentaron el examen en los calendarios A y B, generando un conjunto inicial con 520.307 observaciones y 78 variables. Esta base fue posteriormente enriquecida mediante la integración de información histórica correspondiente al desempeño académico por institución educativa en los años 2015 a 2019, lo cual se logró a partir de una segunda base con más de 7.1 millones de registros. Estas variables históricas fueron estandarizadas.

Figura 1

Mapa Mental de Variables Usadas para los Modelos por Temas

El mapa mental presenta una clasificación conceptual de las variables utilizadas en los modelos predictivos, organizadas en categorías temáticas como hábitos y recursos del hogar, características del estudiante, familia, colegio y puntajes previos. Esta representación facilita la comprensión de la estructura y relaciones entre las variables analizadas.

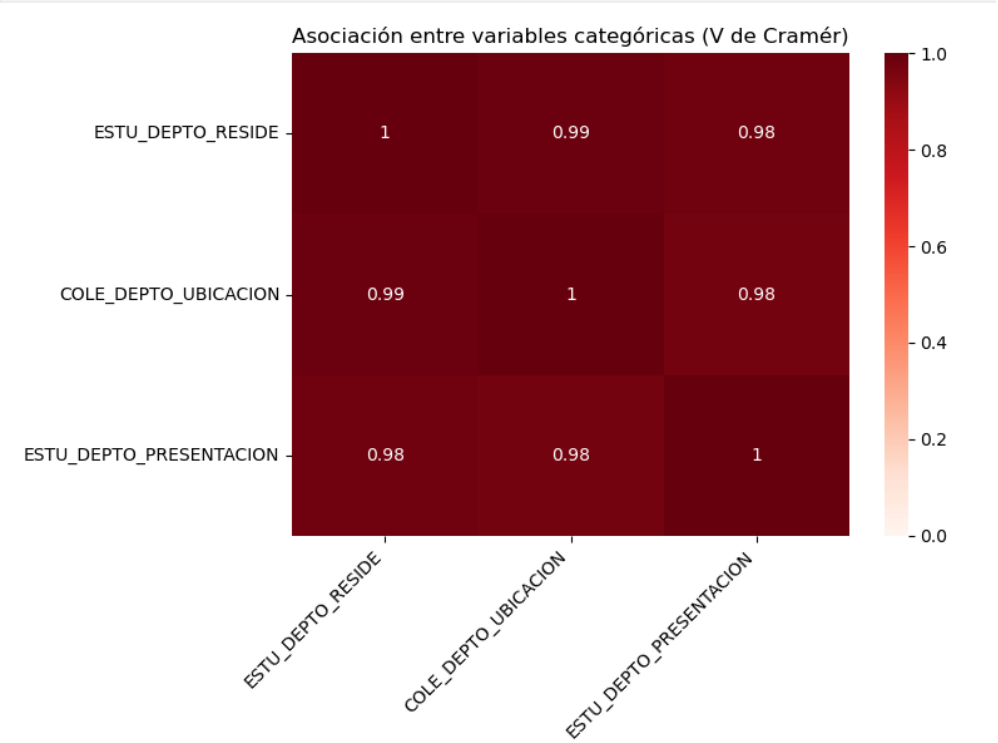
Esta integración permitió calcular promedios institucionales históricos por área (Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales y Ciudadanas, e Inglés) para cada colegio identificado por código DANE. Esta incorporación del comportamiento histórico institucional constituye una contribución metodológica central, ya que introduce una dimensión contextual robusta que permite modelar patrones estructurales en el rendimiento educativo. Mientras la mayoría de los estudios se limitan a analizar variables individuales reportadas por el estudiante, el presente trabajo incorpora el legado institucional como predictor clave.

En la etapa de preprocesamiento de datos, se llevaron a cabo tareas de limpieza y transformación. La edad de los estudiantes fue calculada a partir de la fecha de nacimiento. Para las variables categóricas, se realizó imputación por moda para valores en blanco.

Para la estimación de los modelos, no se tuvieron en cuenta algunas variables como ESTU_ESTUDIANTE ya que esta variable no presentaba variabilidad porque todos los valores en la base eran “Estudiante”, las variables asociadas con identificación de municipios e identificación de establecimientos se eliminaron debido al gran número de variables dicotómicas que se requerían para su tratamiento, lo cual representaba una gran carga computacional y no permitía que los modelos fueran estimados. Las tres variables por departamentos asociadas a ubicación del colegio, residencia del estudiante, y departamento de presentación se escogió la primera, las tres variables muestran una alta asociación según la medida V de Cramér. El listado de todas las variables involucradas en el estudio se puede visualizar en el anexo 1.

Figura 2

Medida V de Cramér para Asociación de Variables Geográficas por Departamento



El mapa de calor de la Figura 2, muestra los valores de asociación entre las variables categóricas ESTU_DEPTO_RESIDE, COLE_DEPTO_UBICACION y ESTU_DEPTO_PRESENTACION, calculados mediante la medida V de Cramér. Los valores cercanos a 1 indican una alta asociación entre las variables, lo cual sugiere redundancia de información geográfica en el modelo.

Con el conjunto de datos limpio y enriquecido, se procedió a la fase de modelado predictivo. Un aspecto crucial de esta etapa es que los modelos utilizados no solo buscan ajustar relaciones estadísticas entre las variables independientes y la variable objetivo, sino que también incorporan mecanismos automáticos de selección de variables.

Esta funcionalidad permite que el modelo descarte, de forma algorítmica y sin intervención manual, aquellos predictores que no aportan valor explicativo significativo. Como resultado, se obtiene un modelo más parsimonioso, interpretable y robusto frente al sobreajuste, que facilita la toma de decisiones basada en evidencia y mejora su capacidad de generalización a nuevos datos.

A continuación, se describen en detalle los tres enfoques de regresión implementados sobre la base de datos, esta se particionó entre base de entrenamiento y base de prueba con proporciones 80/20.

Métodos

Regresión Lineal con Eliminación hacia Atrás (OLS - Backward Elimination)

La regresión lineal múltiple es una técnica estadística clásica que modela la relación entre una variable dependiente continua Y y un conjunto de p variables independientes $X_1, X_2, \dots, X_p$. El modelo lineal se expresa como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

donde:

β_0 es el intercepto,

β_j son los coeficientes de regresión asociados a cada variable independiente X_j,

ε es el término de error aleatorio, que se asume con media cero y varianza constante.

En la técnica de eliminación hacia atrás (backward elimination), se inicia el proceso con todos los predictores disponibles en el modelo. Luego, se evalúa la significancia estadística de cada uno mediante su valor p obtenido del contraste de hipótesis:

$$H_0: \beta_j = 0 \quad (2)$$

Se elimina iterativamente la variable con el mayor valor p (es decir, la menos significativa), y el modelo se reestima en cada paso, hasta que todas las variables remanentes sean estadísticamente significativas (generalmente con $p < 0.05$).

Esta técnica se apoya en el criterio de significancia estadística y, aunque es determinista y dependiente del orden de entrada de las variables, tiene la ventaja de generar modelos interpretables y explicativos, preservando solo los predictores que aportan evidencia sólida de asociación con la variable dependiente. En el contexto de esta tesis, el modelo OLS sirve como referencia base para evaluar la efectividad de métodos más modernos como Lasso y Ridge.

Modelo Lasso (Least Absolute Shrinkage and Selection Operator)

El modelo Lasso es una extensión del modelo de regresión lineal que incorpora un mecanismo de penalización para la selección de variables y el control de sobreajuste. La regresión Lasso optimiza la siguiente función objetivo:

$$\min \beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (3)$$

donde:

y_i es el valor observado de la variable dependiente para el individuo i,

x_{ij} es el valor del predictor j para el individuo i,

β_j son los coeficientes a estimar,

λ es el parámetro de regularización que controla la penalización.

El término $\sum_{j=1}^{p} |\beta_j|$ corresponde a la norma L_1, que al ser incorporada en la función de pérdida, tiene el efecto deseado de forzar a cero algunos coeficientes. Es decir, Lasso realiza selección automática de variables al mismo tiempo que ajusta el modelo, eliminando por completo aquellos predictores con bajo poder explicativo.

La principal ventaja de Lasso radica en su capacidad para manejar situaciones donde existe un gran número de variables, muchas de las cuales podrían estar correlacionadas. En el ámbito educativo, donde los factores que influyen en el rendimiento académico pueden ser numerosos y heterogéneos (socioeconómicos, familiares, institucionales, etc.), Lasso permite identificar los más relevantes, mejorando la interpretabilidad del modelo sin sacrificar precisión predictiva.

Además, el parámetro λ se puede calibrar mediante validación cruzada, lo que permite seleccionar el grado óptimo de penalización que maximiza la capacidad de predicción del modelo sobre datos no observados.

Modelo Ridge

El modelo Ridge, al igual que Lasso, incorpora una penalización en la estimación de los coeficientes, pero en este caso utiliza la norma L_2, lo que genera una regularización suave sobre los coeficientes. La función de pérdida a minimizar en Ridge es:

$$\min \beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \qquad (4)$$

donde:

y_i es el valor observado de la variable dependiente para el individuo i,

x_{ij} es el valor del predictor j para el individuo i,

β_j son los coeficientes a estimar,

λ es el parámetro de regularización.

A diferencia de Lasso, Ridge no fuerza a cero los coeficientes, sino que los reduce en magnitud, por lo cual es especialmente útil cuando todas las variables aportan al modelo pero podrían presentar multicolinealidad. En consecuencia, Ridge no elimina variables

explícitamente, pero estabiliza sus coeficientes en presencia de multicolinealidad, es decir, cuando dos o más predictores están altamente correlacionados.

Este modelo es particularmente útil cuando se enfrentan conjuntos de datos con muchas variables predictoras que aportan información redundante pero aún relevante. En este trabajo, Ridge permite conservar el efecto combinado de múltiples predictores correlacionados (por ejemplo, condiciones socioeconómicas, infraestructura del colegio, historial institucional), sin amplificar artificialmente su impacto individual.

Ridge mejora la robustez del modelo, reduce la varianza de los coeficientes estimados y ofrece una solución más estable, especialmente en contextos donde el número de predictores es elevado en relación con el tamaño de la muestra.

Cada uno de estos modelos fue calibrado sobre la misma base de datos y con los mismos parámetros de entrada. Posteriormente, se evaluaron comparativamente a través de un conjunto de métricas de desempeño que permitieron establecer cuál técnica ofrecía el mejor balance entre capacidad explicativa, error de predicción y complejidad del modelo.

Métricas de Evaluación del Desempeño Predictivo

Para cuantificar la precisión y la bondad de ajuste de los modelos de regresión desarrollados (Regresión Lineal con Selección Hacia Atrás, Lasso y Ridge), se emplearon las siguientes métricas estándar de evaluación, calculadas sobre el conjunto de datos de prueba:

Error Cuadrático Medio (MSE - Mean Squared Error): Representa el promedio de los errores al cuadrado entre los valores predichos por el modelo $(\hat{y}_i)$ y los valores reales observados (y_i). Su fórmula es:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5)$$

Al elevar al cuadrado las diferencias, el MSE penaliza en mayor medida los errores grandes. Sus unidades son el cuadrado de las unidades de la variable objetivo. Un valor más bajo indica un mejor ajuste.

Raíz del Error Cuadrático Medio (RMSE - Root Mean Squared Error): Es la raíz cuadrada del MSE,

$$RMSE = \sqrt{MSE}\ . \qquad (6)$$

Su principal ventaja es que se expresa en las mismas unidades que la variable objetivo, lo que facilita la interpretación de la magnitud típica del error de predicción del modelo. Al igual que el MSE, valores más bajos indican un mejor rendimiento.

Error Absoluto Medio (MAE - Mean Absolute Error): Corresponde al promedio de las diferencias absolutas entre los valores predichos y los reales:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (7)$$

Mide la magnitud promedio de los errores sin considerar su dirección y es menos sensible a valores atípicos que el MSE o RMSE. Se expresa en las mismas unidades que la variable objetivo y valores más bajos son preferibles.

Mediana del Error Absoluto (Median Absolute Error): Es la mediana de los errores absolutos individuales:

$$Mediana(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \dots, |y_n - \hat{y}_n|) \qquad (8)$$

Esta métrica es particularmente útil por su robustez frente a valores atípicos (outliers) en los errores de predicción, ofreciendo una medida de la magnitud del error para la mitad central de las predicciones.

Coeficiente de Determinación (R^2): Mide la proporción de la varianza total de la variable dependiente (y) que es explicada por el modelo de regresión.

Se calcula como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \qquad (9)$$

Donde $\overline{y}$ es la media de los valores observados. El R^2 varía típicamente entre 0 y 1, donde valores más cercanos a 1 indican que el modelo explica una mayor proporción de la variabilidad de los datos, sugiriendo un mejor ajuste predictivo en relación con un modelo que simplemente predice la media.

La evaluación comparativa de los modelos OLS-Backward, Lasso y Ridge se basará en el análisis conjunto de estas métricas, buscando identificar el modelo que ofrezca el mejor equilibrio entre precisión predictiva y robustez.

Tipo de Estudio

La presente investigación se enmarca dentro de un enfoque cuantitativo, dado que se fundamenta en la recolección y análisis de datos numéricos, como los puntajes obtenidos en las pruebas Saber 11 y variables demográficas, socioeconómicas e institucionales susceptibles de medición. Se utilizaron técnicas estadísticas y de aprendizaje automático para procesar esta información y evaluar los resultados mediante métricas numéricas.

Asimismo, el estudio posee un alcance predictivo, ya que su objetivo principal es desarrollar y evaluar modelos capaces de pronosticar o estimar un valor futuro (el puntaje ponderado de admisión a la UIS) basándose en un conjunto de variables predictoras conocidas. Se busca anticipar el desempeño probable de los aspirantes en el proceso de admisión.

Finalmente, la investigación tiene un componente analítico, pues no se limita a describir o predecir, sino que también busca comprender las relaciones entre las variables predictoras y la variable objetivo mediante la aplicación y comparación de diferentes algoritmos de Machine

Learning (Regresión Lineal con Selección Hacia Atrás, Lasso y Ridge), así como el análisis de la influencia de dichas variables en los resultados de los modelos.

Recolección de Datos

La base fundamental para el desarrollo de los modelos predictivos en este estudio provino de fuentes de datos públicas suministradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES). Específicamente, se utilizaron las bases de datos correspondientes a los resultados del Examen Saber 11 presentados durante el año 2020, consolidando los registros de los calendarios A y B. Este conjunto de datos inicial comprendía 520.307 observaciones (estudiantes) y 78 variables originales, incluyendo puntajes por área, información demográfica y socioeconómica del estudiante, y datos del establecimiento educativo.

Para enriquecer el análisis y añadir una dimensión contextual relevante, esta base inicial fue integrada con información histórica del desempeño académico promedio por institución educativa. Se utilizó una segunda base de datos del ICFES que contenía más de 7.1 millones de registros históricos correspondientes a los años 2015 a 2019. A partir de esta información, se calcularon y estandarizaron los promedios históricos de puntaje por área (Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales y Ciudadanas, e Inglés) para cada colegio, identificado por su código DANE. La incorporación de este desempeño histórico institucional se considera una contribución metodológica clave, al introducir predictores que capturan patrones estructurales del entorno educativo del estudiante.

Posteriormente, se realizó un proceso de preprocesamiento y limpieza de datos. Esto incluyó el cálculo de la edad de los estudiantes a partir de su fecha de nacimiento y la imputación por moda para tratar valores faltantes en variables categóricas. Se llevó a cabo una selección de variables, descartando aquellas sin variabilidad (p. ej., ESTU_ESTUDIANTE) y

otras que implicaban una alta carga computacional debido a su gran número de categorías (p. ej., identificadores de municipios y establecimientos específicos). Dada la alta correlación entre las variables de ubicación departamental (residencia, lugar de presentación del examen, ubicación del colegio), se optó por conservar únicamente la variable COLE_DEPTO_UBICACION. El listado detallado de todas las variables finalmente incluidas en el modelado se presenta en el Anexo 1.

Resultados

En esta sección se presenta el análisis comparativo de tres técnicas de regresión — regresión lineal con eliminación hacia atrás (OLS_Backward), regresión Lasso y regresión Ridge aplicadas a cinco programas de pregrado de la Facultad de Ciencias de la Universidad Industrial de Santander: Biología, Física, Licenciatura en Matemáticas, Matemáticas y Química. El objetivo fue identificar cuál de estos modelos ofrece el mejor desempeño predictivo para cada carrera, y al mismo tiempo evaluar su comportamiento general con base en criterios de error y capacidad explicativa.

Comparación entre Modelos

De manera general, el modelo Ridge mostró un desempeño sistemáticamente superior en relación con las demás técnicas evaluadas. Aunque las diferencias entre Ridge, y OLS_Backward fueron en muchos casos mínimas a menudo en el cuarto decimal, Ridge presentó una ventaja consistente tanto en la reducción de error como en la estabilidad de los coeficientes. Su capacidad para manejar multicolinealidad y evitar el sobreajuste lo consolidó como el modelo más robusto del conjunto.

El modelo OLS_Backward, por su parte, ofreció un rendimiento muy similar al de Ridge. Este modelo se basa en la exclusión secuencial de variables no significativas y resulta especialmente útil cuando se busca una estructura interpretativa clara y basada en la inferencia estadística. Sin embargo, su falta de regularización lo hace más vulnerable a la varianza en conjuntos de datos con variables correlacionadas.

En contraste, el modelo Lasso, aunque metodológicamente interesante por su capacidad para seleccionar automáticamente un subconjunto de variables relevantes mostró un rendimiento

ligeramente inferior. En todos los casos analizados, Lasso presentó los mayores valores de error (MSE, RMSE y MAE) y los menores valores de R^2.

Resultados por Carrera

Los resultados por programa académico se resumen en el siguiente cuadro:

Tabla 2

Resultados por Programa Académico

Carrera	Modelo	MSE	RMSE	MAE	Mediana Error	R2
Biología	Lasso	51,600	7,183	5,709	4,854	0,462
Biología	OLS Backward	49,710	7,051	5,589	4,724	0,482
Biología	Ridge	49,706	7,050	5,589	4,723	0,482
Física	Lasso	52,561	7,250	5,775	4,927	0,459
Física	OLS Backward	50,647	7,117	5,653	4,790	0,479
Física	Ridge	50,646	7,117	5,653	4,792	0,479
Lic. en Matemáticas	Lasso	56,603	7,523	5,963	5,035	0,451
Lic. en Matemáticas	OLS Backward	54,697	7,396	5,849	4,930	0,470
Lic. en Matemáticas	Ridge	54,691	7,395	5,849	4,932	0,470
Matemáticas	Lasso	56,603	7,523	5,963	5,035	0,451
Matemáticas	OLS Backward	54,697	7,396	5,849	4,930	0,470
Matemáticas	Ridge	54,691	7,395	5,849	4,932	0,470
Química	Lasso	52,358	7,236	5,750	4,886	0,460
Química	OLS Backward	50,446	7,103	5,628	4,755	0,480
Química	Ridge	50,443	7,102	5,628	4,756	0,480

Nota. Comparación de los modelos OLS Backward, Lasso y Ridge para los diferentes programas.

Específicamente, para cada carrera se puede observar que:

Biología: El modelo Ridge obtuvo resultados ligeramente superiores, con un MSE de 49,651, RMSE de 7,0463, MAE de 5,5855 y un R^2 de 0,4822. Esto indica que Ridge predice con menor dispersión y mejor precisión los puntajes de admisión al programa de Biología, ofreciendo

a los aspirantes una estimación más confiable. Aunque los valores obtenidos fueron muy cercanos a los del modelo OLS Backward, la capacidad superior de Ridge para gestionar la multicolinealidad le permite ofrecer una mayor robustez y estabilidad en sus predicciones.

Física: El modelo Ridge presentó un desempeño óptimo con un MSE de 50,5917, RMSE de 7,1128, MAE de 5,6495 y un R^2 de 0,4795, superando ligeramente al modelo OLS Backward. Esta superioridad marginal es una clara indicación de la efectividad de Ridge para manejar variables altamente correlacionadas, ofreciendo una mayor confiabilidad en escenarios predictivos complejos.

Licenciatura en Matemáticas: Ridge mostró los mejores resultados con un MSE de 44,9042, RMSE de 6,7011, MAE de 5,2867 y un R^2 de 0,4855, aunque empatado en R^2 con OLS Backward. La ventaja del modelo Ridge reside principalmente en sus menores valores de error, lo que implica una mayor exactitud en la estimación del puntaje de admisión para este programa.

Matemáticas: El modelo Ridge obtuvo el mejor rendimiento, con un MSE de 54,624, RMSE de 7,3908, MAE de 5,845 y un R^2 de 0,4704. Aunque el modelo OLS Backward estuvo cerca en términos de desempeño, Ridge mantuvo una ventaja significativa en términos de estabilidad predictiva.

Química: Ridge nuevamente mostró superioridad, registrando un MSE de 50,386, RMSE de 7,0983, MAE de 5,625 y un R^2 de 0,4808, empatado en R^2 con OLS Backward. Sin embargo, Ridge destaca nuevamente por ofrecer errores ligeramente menores. La robustez del modelo Ridge lo convierte en la opción más adecuada para predecir puntajes de admisión en este programa.

Por el análisis anterior, el modelo Ridge se considera superior y se escoge como el mejor modelo.

Variables Destacadas del Modelo Escogido

El análisis de los coeficientes más significativos identificados por el modelo Ridge destaca:

Tabla 3

Las 10 Variables más Influyentes en el Modelo Ridge

10 variables más influyentes	Comparación	Interpretación	Relevancia
ESTU PRIVADO LIBERTAD S (Coef.= –5.73)	Estudiantes privados de libertad (S=Si) vs. categoría base (N = No privado de libertad).	Ser privado de libertad está asociado con una reducción significativa de –5.73 puntos en el puntaje promedio de admisión, en comparación con quienes no lo están.	Esta es la variable podría reflejar una desigualdad en oportunidades de aprendizaje para poblaciones carcelarias o institucionalizadas.
COLE JORNADA NOCHE (Coef. = –2.92)	Jornada noche frente a jornada mañana (categoría base)	Los estudiantes que estudian en la noche presentan una reducción de casi –3 puntos en su puntaje de admisión, en comparación con quienes lo hacen en la mañana.	La jornada nocturna podría estar relacionada con estudiantes que trabajan durante el día o que tienen trayectorias no tradicionales, lo cual puede influir en menor

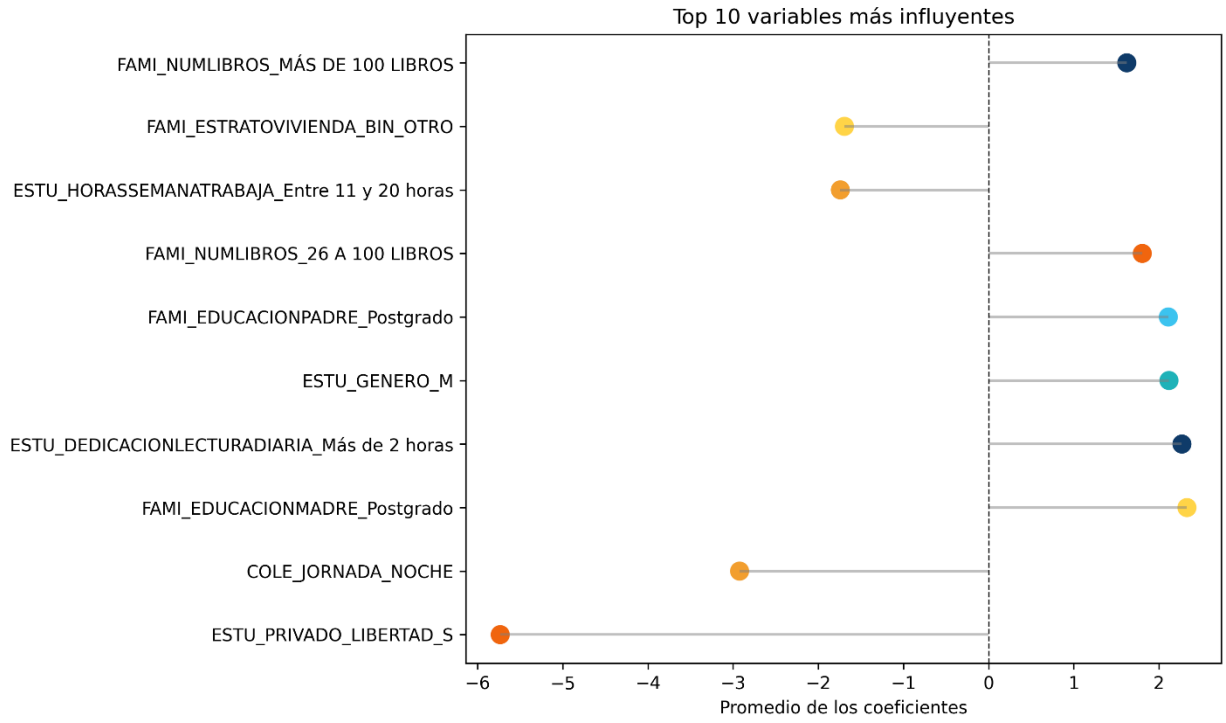
10 variables más influyentes	Comparación	Interpretación	Relevancia
			dedicación al estudio.
FAMI EDUCACION MADRE Postgrado (Coef. = +2.33)	Educación posgrado de la madre frente a la categoría base secundaria completa.	Se observa un incremento promedio de +2.33 puntos en los puntajes de admisión para estudiantes cuya madre tiene estudios de posgrado.	Podría reflejar la importancia del capital cultural familiar en este caso de la madre en el rendimiento académico.
ESTU DEDICACION LECTURA DIARIA Más de 2 horas (Coef. = +2.27)	Más de 2 horas de lectura frente a la base “30 minutos o menos”.	La dedicación prolongada a la lectura está asociada a un aumento significativo en el puntaje, cerca de +2.27 puntos.	La lectura regular y prolongada podría ester reforzando la comprensión y habilidades cognitivas clave para el éxito académico.
ESTU GENERO M (Coef. = +2.12)	Estudiantes hombres frente a la base mujeres.	Ser hombre está asociado con un incremento de +2.12 puntos en el puntaje, en promedio.	Este resultado puede estar reflejando sesgos estructurales en el sistema educativo.
FAMI EDUCACION PADRE_Postgrado (Coef. = +2.11)	Posgrado frente a la base secundaria completa.	Al igual que con la madre, el nivel educativo del padre	Este resultado podría estar reforzando la

10 variables más influyentes	Comparación	Interpretación	Relevancia
		tiene un efecto positivo similar (+2.11).	importancia del rol del entorno familiar como predictor del desempeño académico.
FAMI NUMLIBROS 26 A 100 LIBROS (Coef. = +1.80)	26 a 100 libros frente a la base 0 a 10 libros.	Tener una biblioteca doméstica moderadamente abastecida se asocia con +1.80 puntos adicionales.	Podría tratarse de un indicador indirecto del capital cultural y acceso a recursos educativos.
ESTU HORAS SEMANA TRABAJA Entre 11 y 20 horas (Coef. = –1.74)	Trabajo entre 11 y 20 horas frente a la base: no trabajar (0 horas).	La carga laboral media podría comprometer el rendimiento académico, reduciendo el puntaje en –1.74 puntos.	Este hallazgo podría sugerir que trabajar mientras se estudia tiene efectos adversos notables sobre el desempeño académico.
FAMI ESTRATOVIVIENDA BIN OTRO (Coef. = –1.69)	Agrupación de Sin estrato versus la base estrato 3 o menos.	Reportarse como Sin Estrato, podría comprometer el rendimiento académico, reduciendo el puntaje en –1.69 puntos.	Las viviendas “Sin estrato” podrían estar ubicadas en zonas rurales dispersas, territorios sin planificación urbana, resguardos

10 variables más influyentes	Comparación	Interpretación	Relevancia
			indígenas o zonas con informalidad estructural. Esto podría implicar limitado acceso a servicios públicos, conectividad y recursos educativos. En este contexto, la ausencia de estrato es más grave que pertenecer a un estrato bajo.
FAMI NUMLIBROS MÁS DE 100 LIBROS (Coef. = +1.62)	Más de 100 libros frente a la base, 0 a 10 libros	Tener una biblioteca amplia se traduce en +1.62 puntos, reforzando el impacto positivo del entorno educativo en casa.	Este resultado podría afianzar en mayor medida que la riqueza cultural del hogar influye directamente en las oportunidades educativas.

Figura 3

Coeficientes Promedio de las Variables más Influyentes en el Modelo Ridge



Nota. El gráfico muestra las 10 variables con mayor peso predictivo, considerando su efecto relativo frente a su categoría base.

Este gráfico muestra el impacto promedio (coeficiente) de las 10 variables más influyentes del modelo. Las líneas horizontales representan la dirección e intensidad del efecto. Los puntos de colores indican los valores del coeficiente para cada variable.

El eje X representa el promedio del coeficiente estimado por el modelo Ridge, el eje Y muestra los nombres de las variables categóricas.

La línea vertical punteada en cero indica el punto de referencia: coeficientes negativos indican efecto perjudicial sobre el puntaje, y positivos, efecto beneficioso.

El gráfico deja en evidencia que, las condiciones estructurales vulnerables como privación de libertad, trabajo, informalidad en vivienda, podrían afectar negativamente el rendimiento.

Por otro lado, los factores familiares y académicos como el nivel educativo de los padres, la lectura habitual y el capital cultural (número de libros) tienen un impacto marcadamente positivo. Esto es consistente con hallazgos en la literatura, como los de Idris et al. (2020), quienes concluyeron que «la alta educación del padre y de la madre contribuye positivamente al rendimiento académico de sus hijos» (p. 82).

También se encontraron diferencias por género, con una ventaja promedio para los hombres en este contexto.

Conclusiones

A partir del desarrollo de la presente investigación, orientada a la predicción de puntajes de admisión en programas de la Facultad de Ciencias de la Universidad Industrial de Santander (UIS) mediante técnicas de Machine Learning, se extraen las siguientes conclusiones principales, alineadas con los objetivos propuestos:

La revisión de la literatura confirmó la creciente aplicación de modelos de Machine Learning para predecir el rendimiento académico, pero también evidenció un vacío específico en la predicción de puntajes de admisión ponderados para la UIS, particularmente para su Facultad de Ciencias, validando la pertinencia y originalidad de este estudio. Además, se constató la relevancia de incluir factores socioeconómicos y contextuales, además de los resultados académicos previos.

Se estimaron y compararon exitosamente tres modelos predictivos (Regresión Lineal con Selección Hacia Atrás, Lasso y Ridge) para los puntajes de admisión. La evaluación rigurosa mediante métricas de error (MSE, RMSE, MAE) y el coeficiente de determinación (R^2) demostró que, si bien todos los enfoques ofrecieron resultados razonables, el modelo de Regresión Ridge presentó consistentemente el desempeño más robusto y estable a través de los cinco programas analizados (Biología, Física, Lic. en Matemáticas, Matemáticas y Química). Su capacidad para gestionar la multicolinealidad sin eliminar variables resultó ventajosa frente a OLS Backward (competitivo pero más sensible a la colinealidad) y Lasso (que mostró un rendimiento predictivo inferior en este caso particular).

El análisis de los coeficientes del modelo Ridge, el de mejor desempeño, permitió identificar variables predictoras con influencia significativa en los puntajes de admisión estimados. Factores como el género del estudiante, la dedicación diaria a la lectura, el nivel

educativo de los padres, el promedio histórico del colegio en áreas clave como matemáticas, y condiciones específicas como ser estudiante privado de libertad, mostraron asociaciones relevantes. Estos hallazgos subrayan la importancia de considerar el contexto socioeconómico e institucional, más allá de las notas individuales, y cumplen con el objetivo de comprender qué factores impulsan las predicciones del modelo.

En síntesis, la investigación valida el uso de regresión penalizada (Ridge) como una herramienta efectiva y estable para predecir puntajes de admisión en este contexto, aporta un modelo específico ajustado a los criterios de la UIS y ofrece información valiosa sobre los factores determinantes, sentando una base sólida para futuras aplicaciones de orientación académica.

Recomendaciones

Basado en los resultados y conclusiones de este estudio, se proponen las siguientes recomendaciones:

Aplicación Práctica: Se recomienda considerar el desarrollo y la implementación de una herramienta de software basada en el modelo de Regresión Ridge validado. Esta herramienta podría ser de utilidad para los futuros estudiantes que aspiran a ingresar a los programas de la Facultad de Ciencias de la UIS, ofreciéndoles una estimación de sus posibles puntajes de admisión y apoyando así una toma de decisiones más informada.

Limitaciones y Líneas de Investigación Futura

Este estudio usa principalmente la base de datos del Examen Saber 11 de 2020, lo que restringe la generalización a otros años, y depende de la calidad de los datos históricos de desempeño académico. Metodológicamente, se centra en modelos de regresión específicos. Además, el modelo predice puntajes solo para la Facultad de Ciencias de la UIS y no realiza un seguimiento del desempeño estudiantil a largo plazo para validar su capacidad predictiva del éxito académico futuro.

Se recomienda extender esta línea de investigación aplicando metodologías similares para predecir puntajes de admisión en otras facultades de la UIS o en diferentes instituciones de educación superior colombianas, adaptando los modelos a los criterios de ponderación específicos de cada caso.

Se recomienda explorar el uso de técnicas de Machine Learning más avanzadas (p.ej., modelos no lineales o ensambles) para determinar si es posible alcanzar mejoras sustanciales en la capacidad predictiva de los modelos.

Si existiera la posibilidad de acceder a los datos, se recomienda realizar estudios longitudinales que hagan seguimiento al desempeño académico de los estudiantes admitidos, con el fin de validar la relación entre los puntajes de admisión (reales y predichos) y el éxito académico a largo plazo dentro de la universidad.

Referencias

Ardila Perez, Y. A., y Garcia Duarte, M. V. (2017). *When Standardized Tests Set the Agenda: A Study with School Leavers* [Trabajo de grado, Universidad Distrital Francisco José de Caldas]. Repositorio Institucional de la Universidad Distrital Francisco José de [1] Caldas. http://hdl.handle.net/11349/12879

Arguelles Monterrosa, C. J., y Meléndez Armella, L. V. (2023). *Diseño de un sistema que permite identificar las aptitudes para el estudio de carreras STEM basado en los resultados de las pruebas estandarizadas ICFES Saber 11* [Trabajo de grado de pregrado, Universidad de Córdoba]. Repositorio Institucional de la Universidad de Córdoba. https://repositorio.unicordoba.edu.co/bitstream/handle/unicordoba/39146/e-ee02-407d-0b69-0bc41304a4c7

Burgos-Moreno, A. S. (2021). *A Deep Learning Approach to Predicting Colombian College Entrance Exam SABER 11 Scores* [Tesis de maestría, Tilburg University]. Repositorio Institucional Arno de Tilburg University. http://arno.uvt.nl/show.cgi?fid=160821

Chica-Gómez, S., Galvis-Gutiérrez, D. M., y Ramírez-Hassan, A. (2010). Determinantes del rendimiento académico en Colombia: Pruebas ICFES - Saber 11°. *Revista Universidad EAFIT*, *46*(159), pp. 11-26. https://repository.eafit.edu.co/server/api/core/bitstreams/2441b41c-dd44-4c5e-bb25-6cbc72b67e7c/content

Fernández-García, S. (2017). *Características socioeconómicas de los estudiantes y particulares de los colegios que pueden explicar las diferencias entre los resultados de colegios públicos y privados en las pruebas de estado ICFES Saber 11*. Repositorio Institucional de la Universidad de los Andes.

https://repositorio.uniandes.edu.co/server/api/core/bitstreams/1c1951fea-6eb1-472b-8cb-f4029c977a0c/content

Gómez Silva, C. A. (2016). Clasificación de colegios según las pruebas Saber 11 del ICFES: Un análisis usando modelos marginales (MM). *Sociedad y Economía*, (30), pp. 69-89. http://www.scielo.org.co/scielo.php%3Fpid%3DS1657-75852016000300004%26script%3Dsci_abstract%26tlng%3Dpt

Idris, M., Hussain, S., y Ahmad, N. (2020). Relationship between parents' education and their children's academic achievement. *Journal of Arts and Social Sciences, 7*(2), pp 82-92. https://doi.org/10.46662/jass-vol7-iss2-2020(82-92)

Instituto Colombiano para la Evaluación de la Educación. (2021a). *Saber 11 2020 2* [Base de datos]. Datos.gov.co. https://www.datos.gov.co/Educaci-n/Saber-11-2020-2/rnvb-vnyh/about_data

Instituto Colombiano para la Evaluación de la Educación. (2021b). *Saber 11 2020 1* [Base de datos]. Datos.gov.co. https://www.datos.gov.co/Educaci-n/Saber-11-2020-1/a8xr-en99/about_data

Martínez, D., Salcedo-Parra, O. J., y Aguilera Prado, M. A. (2020). Machine learning applied to ICFES tests to identify measurement patterns in students. *International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), 10*(6), pp. 1-8. https://www.researchgate.net/publication/343088490_Machine_Learning_Applied_to_ICFES_Tests_to_Identify_Measurement_Patterns_in_Students/links/5f0d2a6f92851c8b1f3bb3f6/Machine-Learning-Applied-to-ICFES-Tests-to-Identify-Measurement-Patterns-in-Students.pdf

Martínez-Cernera, D. E., Salcedo Parra, O. J., y Aguilera Prado, M. A. (2021). Forecasting model with machine learning in higher education ICFES exams. *International Journal of Electrical and Computer Engineering (IJECE), 11*(6), pp. 5402-5410. https://www.google.com/search?q=https://doi.org/10.11591/ijece.v11i6.pp5402-5410

Martinez Mateus, W. A., y Turriago Hoyos, Á. (2015). Análisis de distribución geográfica y espacial de los resultados de las Pruebas Saber 11 del Instituto Colombiano para el Fomento de la Educación Superior (ICFES), 2005-2012. *Cuadernos Latinoamericanos de Administración, XI*(21), pp. 38-49. https://www.redalyc.org/articulo.oa%3Fid%3D409640964005

Moncayo-Cabrera, M. A. (2016). *Determinantes que influyen en el rendimiento académico: un estudio aplicado para Colombia a partir de las Pruebas Icfes - Saber 11* [Trabajo de grado de pregrado, Universidad de La Salle]. Repositorio Institucional Ciencia Unisalle. https://ciencia.lasalle.edu.co/economia/232/viewcontent.cgi%3Farticle%3D1231%26context%3Deconomia

Peña-Lozano, Y., y González-Vélez, J. J. F. (2022). Modelo de predicción de los resultados de la prueba ICFES Saber 11 en el área de matemáticas a partir de variables socioeconómicas. *Studies in Engineering and Exact Sciences, 3*(1), pp. 52-68. https://doi.org/10.54021/sees.v3n1-006

Rodríguez-Hernández, C. F., Musso, M., Kyrdt, E., y Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence, 2*, 100018. https://doi.org/10.1016/j.caeai.2021.100018

Serna-Cardona, L. A., Hernández-Gómez, K. A., y Orozco-Gutiérrez, Á. A. (2021). Caracterización de los estudiantes de educación media en el departamento de Risaralda usando la métrica Chi-Square. *Scientia Et Technica*, *26*(2), pp. 119-126. https://doi.org/10.22517/23447214.24587

Soto-Acevedo, M., Abuchar-Curi, A. M., Zuluaga-Ortiz, R. A., y Delahoz-Dominguez, E. J. (2023). A machine learning model to predict standardized tests in engineering programs in Colombia. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, *18*(3), pp. 211-218. https://doi.org/10.1109/RITA.2023.3301396

Timarán-Pereira, R., Hidalgo-Troya, Á., y Caicedo-Zambrano, J. (2020). Factores asociados al desempeño académico en lectura crítica en las pruebas Saber 11° con árboles de decisión. *Investigación e Innovación en Ingenierías*, *8*(3), pp. 29-37. https://doi.org/10.17081/inno.8.3.4701

Timarán-Pereira, R., Caicedo-Zambrano, J., y Hidalgo-Troya, Á. (2019). Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. *Revista Investigación Desarrollo e Innovación*, *9*(2), pp. 363-378. https://doi.org/10.19053/rdid.v9n2.2019.9184

Universidad Industrial de Santander. (2022). *Criterios de admisión*. Universidad Industrial de Santander. https://uis.edu.co/wp-content/uploads/2022/05/CRITERIOS-DE-ADMISION-1.pdf

Master, W. (2023, 13 agosto). *UIS, la tercera mejor universidad pública de Colombia: Center for World University Rankings*. Comunicaciones UIS. https://comunicaciones.uis.edu.co/uis-la-tercera-mejor-universidad-publica-del-pais-center-for-world-university-rankings/

Universidad Nacional Abierta y a Distancia. (2024a). *Líneas de Investigación y Productividad ECBTI*. https://academia.unad.edu.co/investigacion-y-productividad-ecbti/lineas

Viana, R. A., Araraz, J. M., y García-Serrano, C. (2020). Efficiency of university education: A partial frontier analysis. *Latin American Economic Review*, *29*(1), pp. 1-16. https://doi.org/10.47872/laer-2020-29-1s

Viana-Barceló, R. A., y Pinto-Prieto, H. M. (2018). Eficiencia de los estudiantes urbanos y rurales de Santander: 'Saber 11' 2016. *Suma de Negocios*, *9*(20), pp. 111-119. https://doi.org/10.14349/sumneg/2018.V9.N20.A5

Apéndices

Apéndice A

Variables Iniciales

Variable	Descripción
EDAD	Edad del estudiante al momento del examen
PROM_5A_COLE_PUNT_LECTURA_CRITICA	Promedio del puntaje en lectura crítica del colegio en los últimos 5 años
PROM_5A_COLE_PUNT_C_NATURALES	Promedio del puntaje en ciencias naturales del colegio en los últimos 5 años
PROM_5A_COLE_PUNT_SOCIALES_CIUDADANAS	Promedio del puntaje en ciencias sociales y ciudadanas del colegio en los últimos 5 años
PROM_5A_COLE_PUNT_MATEMATICAS	Promedio del puntaje en matemáticas del colegio en los últimos 5 años
PROM_5A_COLE_PUNT_INGLES	Promedio del puntaje en inglés del colegio en los últimos 5 años
ESTU_PRIVADO_LIBERTAD	Indicador si el estudiante está privado de libertad
COLE_DEPTO_UBICACION	Departamento de ubicación del colegio
COLE_JORNADA	Jornada escolar del colegio (mañana, tarde, noche, etc.)
COLE_AREA_UBICACION	Zona de ubicación del colegio (urbana o rural)
COLE_SEDE_PRINCIPAL	Sede principal o secundaria del colegio
COLE_CARACTER	Carácter del colegio (oficial o privado)
COLE_BILINGUE	Indica si el colegio ofrece educación bilingüe
COLE_CALENDARIO	Tipo de calendario académico del colegio (A, B, flexible)
COLE_NATURALEZA	Naturaleza del colegio (mixto o unisex)
COLE_GENERO	Género institucional del colegio
ESTU_TIPOREMUNERACION	Tipo de remuneración que recibe el estudiante si trabaja
ESTU_HORASSEMANATRABAJA	Horas semanales que trabaja el estudiante
ESTU_DEDICACIONINTERNET	Dedicación diaria al uso de internet
ESTU_DEDICACIONLECTURADIARIA	Dedicación diaria a la lectura por fuera del contexto escolar
FAMI_SITUACIONECONOMICA	Autoevaluación de la situación económica del hogar
FAMI_COMECEREALFRUTOSLEGUMBRE	Frecuencia con la que consume cereales, frutas o legumbres
FAMI_COMECARNEPESCADOHUEVO	Frecuencia con la que consume carne, pescado o huevo
FAMI_COMELECHEDERIVADOS	Frecuencia con la que consume leche y derivados
FAMI_NUMLIBROS	Número de libros en casa
FAMI_TIENECONSOLAVIDEOJUEGOS	Disponibilidad de consola de videojuegos en casa
FAMI_TIENEMOTOCICLETA	Disponibilidad de motocicleta en el hogar
FAMI_TIENEAUTOMOVIL	Disponibilidad de automóvil en el hogar
FAMI_TIENEHORNOMICROOGAS	Disponibilidad de horno microondas o a gas
FAMI_TIENELAVADORA	Disponibilidad de lavadora en el hogar
FAMI_TIENECOMPUTADOR	Disponibilidad de computador en el hogar
FAMI_TIENESERVICIOTV	Disponibilidad de servicio de televisión
FAMI_TIENEINTERNET	Disponibilidad de internet en el hogar
FAMI_TRABAJOLABORMADRE	Ocupación de la madre
FAMI_TRABAJOLABORPADRE	Ocupación del padre
FAMI_EDUCACIONMADRE	Nivel educativo de la madre
FAMI_EDUCACIONPADRE	Nivel educativo del padre
FAMI_CUARTOSHOGAR	Número de cuartos del hogar
FAMI_PERSONASHOGAR	Número de personas que viven en el hogar
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de la vivienda
ESTU_TIENEETNIA_Si	Pertenencia étnica (indica si se reconoce como parte de un grupo étnico)
ESTU_GENERO_M	Género del estudiante (masculino)
ESTU_NACIONALIDAD	Nacionalidad del estudiante

Apéndice B

Coeficientes de los Modelos Ridge

Variable	Biología	Física	Lic. En Matemáticas	Matemáticas	Química	prom	abs(prom)
ESTU_PRIVADO_LIBERTAD_S	-5,713	-5,827	-6,030	-6,030	-5,669	-5,854	5,854
COLE_JORNADA_NOCHE	-2,931	-2,936	-3,059	-3,059	-2,945	-2,986	2,986
FAMI_EDUCACIONMADRE_Postgrado	2,380	2,406	2,371	2,371	2,402	2,386	2,386
ESTU_DEDICACIONLECTURADIARIA_Más de 2 horas	2,406	2,553	2,187	2,187	2,387	2,344	2,344
FAMI_EDUCACIONPADRE_Postgrado	2,147	2,187	2,160	2,160	2,166	2,164	2,164
ESTU_GENERO_M	2,052	1,919	2,329	2,329	2,126	2,151	2,151
FAMI_NUMLIBROS_26 A 100 LIBROS	1,843	1,850	1,841	1,841	1,854	1,846	1,846
ESTU_HORASSEMANATRABAJA_Entre 11 y 20 horas	-1,797	-1,878	-1,744	-1,744	-1,791	-1,791	1,791
FAMI_ESTRATOVIVIENDA_BIN_OTRO	-1,679	-1,749	-1,788	-1,788	-1,651	-1,731	1,731
FAMI_NUMLIBROS_MÁS DE 100 LIBROS	1,673	1,680	1,634	1,634	1,687	1,662	1,662
ESTU_HORASSEMANATRABAJA_Menos de 10 horas	-1,621	-1,692	-1,634	-1,634	-1,612	-1,639	1,639
COLE_JORNADA_SABATINA	-1,468	-1,496	-1,508	-1,508	-1,466	-1,489	1,489
FAMI_EDUCACIONMADRE_Técnica o tecnológica completa	1,482	1,505	1,484	1,484	1,488	1,488	1,488
ESTU_DEDICACIONLECTURADIARIA_Entre 1 y 2 horas	1,484	1,558	1,329	1,329	1,482	1,437	1,437
FAMI_CUARTOSHOGAR_Seis o mas	-1,420	-1,378	-1,460	-1,460	-1,437	-1,431	1,431
FAMI_SITUACIONECONOMICA_Peor	1,400	1,439	1,426	1,426	1,396	1,417	1,417
ESTU_DEDICACIONLECTURADIARIA_Entre 30 y 60 minutos	1,368	1,407	1,315	1,315	1,370	1,355	1,355
COLE_DEPTO_UBICACION_SAN ANDRES	-1,294	-1,266	-1,351	-1,351	-1,339	-1,320	1,320
FAMI_TRABAJOLABORMADRE_Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial	-1,352	-1,303	-1,278	-1,278	-1,377	-1,318	1,318
ESTU_DEDICACIONINTERNET_No Navega Internet	-1,267	-1,306	-1,356	-1,356	-1,256	-1,308	1,308
FAMI_ESTRATOVIVIENDA_BIN_estrato4omas	-1,302	-1,286	-1,318	-1,318	-1,313	-1,307	1,307
ESTU_HORASSEMANATRABAJA_Entre 21 y 30 horas	-1,345	-1,408	-1,200	-1,200	-1,344	-1,300	1,300
FAMI_EDUCACIONMADRE_Educación profesional incompleta	1,293	1,344	1,241	1,241	1,292	1,282	1,282
COLE_DEPTO_UBICACION_SANTANDER	1,303	1,236	1,256	1,256	1,325	1,275	1,275
FAMI_EDUCACIONPADRE_Técnica o tecnológica completa	1,260	1,278	1,250	1,250	1,271	1,262	1,262
COLE_DEPTO_UBICACION_AMAZONAS	1,294	1,193	1,203	1,203	1,337	1,246	1,246
FAMI_EDUCACIONPADRE_Educación profesional incompleta	1,238	1,274	1,210	1,210	1,240	1,234	1,234
COLE_DEPTO_UBICACION_HUILA	1,256	1,195	1,216	1,216	1,265	1,229	1,229
FAMI_COMELECHEDERIVADOS_Todos o casi todos los días	1,213	1,187	1,207	1,207	1,233	1,209	1,209
ESTU_HORASSEMANATRABAJA_Más de 30 horas	-1,212	-1,302	-1,030	-1,030	-1,201	-1,155	1,155
COLE_DEPTO_UBICACION_NORTE SANTANDER	1,184	1,063	1,117	1,117	1,224	1,141	1,141
FAMI_COMELECHEDERIVADOS_3 a 5 veces por semana	1,118	1,123	1,143	1,143	1,122	1,130	1,130
FAMI_CUARTOSHOGAR_Cinco	-1,121	-1,088	-1,123	-1,123	-1,136	-1,118	1,118
ESTU_DEDICACIONINTERNET_30 minutos o menos	-1,061	-1,095	-1,131	-1,131	-1,055	-1,094	1,094
COLE_DEPTO_UBICACION_PUTUMAYO	1,067	1,012	1,153	1,153	1,076	1,092	1,092
FAMI_SITUACIONECONOMICA_Mejor	-1,064	-1,085	-1,102	-1,102	-1,054	-1,081	1,081
COLE_DEPTO_UBICACION_CALDAS	1,124	1,076	0,965	0,965	1,145	1,055	1,055
COLE_DEPTO_UBICACION_GUAVIARE	1,008	0,903	1,064	1,064	1,043	1,016	1,016
FAMI_EDUCACIONMADRE_Educación profesional completa	0,998	1,004	0,978	0,978	1,006	0,993	0,993
COLE_DEPTO_UBICACION_CASANARE	1,027	0,938	0,968	0,968	1,050	0,990	0,990
FAMI_TRABAJOLABORPADRE_Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial	-0,974	-0,968	-0,896	-0,896	-0,988	-0,944	0,944
COLE_DEPTO_UBICACION_VAUPES	0,877	0,864	0,968	0,968	0,917	0,919	0,919
COLE_DEPTO_UBICACION_NARIÑO	0,968	0,860	0,844	0,844	0,997	0,903	0,903
COLE_DEPTO_UBICACION_QUINDIO	0,899	0,888	0,808	0,808	0,900	0,861	0,861
FAMI_NUMLIBROS_11 A 25 LIBROS	0,838	0,834	0,803	0,803	0,843	0,824	0,824
ESTU_DEDICACIONLECTURADIARIA_No leo por entretenimiento	-0,811	-0,896	-0,799	-0,799	-0,797	-0,820	0,820
FAMI_TRABAJOLABORPADRE_Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente)	-0,816	-0,830	-0,806	-0,806	-0,823	-0,816	0,816
COLE_DEPTO_UBICACION_RISARALDA	0,817	0,793	0,662	0,662	0,828	0,752	0,752
COLE_DEPTO_UBICACION_ARAUCA	0,799	0,697	0,714	0,714	0,828	0,750	0,750

Variable	Biología	Física	Lic. En Matemáticas	Matemáticas	Química	prom	abs(prom)
ESTU_TIENEETNIA_Si	-0,723	-0,746	-0,760	-0,760	-0,706	-0,739	0,739
FAMI_TRABAJOLABORMADRE_Pensionado	-0,673	-0,722	-0,795	-0,795	-0,658	-0,728	0,728
ESTU_TIPOREMUNERACION_Si, en efectivo y especie	0,648	0,757	0,804	0,804	0,620	0,726	0,726
COLE_DEPTO_UBICACION_ATLANTICO	-0,695	-0,729	-0,703	-0,703	-0,697	-0,705	0,705
FAMI_TRABAJOLABORPADRE_Es vendedor o trabaja en atención al público	-0,682	-0,723	-0,721	-0,721	-0,672	-0,704	0,704
FAMI_EDUCACIONMADRE_Técnica o tecnológica incompleta	0,695	0,735	0,702	0,702	0,683	0,703	0,703
COLE_DEPTO_UBICACION_META	0,710	0,637	0,709	0,709	0,722	0,697	0,697
FAMI_EDUCACIONPADRE_Educación profesional completa	0,695	0,720	0,681	0,681	0,697	0,695	0,695
FAMI_PERSONASHOGAR_1 a 2	-0,663	-0,644	-0,735	-0,735	-0,672	-0,690	0,690
COLE_DEPTO_UBICACION_VICHADA	0,528	0,545	0,910	0,910	0,507	0,680	0,680
FAMI_COMECEREALFRUTOSLEGUMBRE_3 a 5 veces por semana	0,668	0,684	0,678	0,678	0,669	0,676	0,676
FAMI_CUARTOSHOGAR_Cuatro	-0,651	-0,648	-0,692	-0,692	-0,654	-0,668	0,668
COLE_GENERO_FEMENINO	0,621	0,570	0,731	0,731	0,643	0,659	0,659
ESTU_TIPOREMUNERACION_Si, en especie	0,559	0,736	0,738	0,738	0,505	0,655	0,655
ESTU_DEDICACIONINTERNET_Entre 30 y 60 minutos	-0,626	-0,660	-0,669	-0,669	-0,622	-0,649	0,649
COLE_DEPTO_UBICACION_BOLIVAR	-0,616	-0,655	-0,679	-0,679	-0,615	-0,649	0,649
COLE_DEPTO_UBICACION_VALLE	0,658	0,647	0,614	0,614	0,660	0,639	0,639
COLE_CALENDARIO_OTRO	0,687	0,617	0,583	0,583	0,704	0,635	0,635
COLE_DEPTO_UBICACION_CAUCA	0,665	0,633	0,592	0,592	0,674	0,631	0,631
FAMI_TIENECOMPUTADOR_No	-0,573	-0,607	-0,613	-0,613	-0,566	-0,595	0,595
FAMI_TIENEINTERNET_No	-0,564	-0,592	-0,584	-0,584	-0,556	-0,576	0,576
COLE_GENERO_MASCULINO	-0,565	-0,493	-0,616	-0,616	-0,584	-0,575	0,575
COLE_DEPTO_UBICACION_CUNDINAMARCA	0,545	0,527	0,579	0,579	0,552	0,556	0,556
COLE_DEPTO_UBICACION_SUCRE	-0,460	-0,560	-0,631	-0,631	-0,441	-0,544	0,544
FAMI_TRABAJOLABORMADRE_Es agricultor, pesquero o jornalero	-0,498	-0,535	-0,545	-0,545	-0,481	-0,521	0,521
FAMI_TIENEMOTOCICLETA_Si	-0,473	-0,524	-0,482	-0,482	-0,464	-0,485	0,485
FAMI_TRABAJOLABORPADRE_Trabaja en el hogar, no trabaja o estudia	-0,462	-0,480	-0,499	-0,499	-0,450	-0,478	0,478
FAMI_TIENEHORNOMICROOGAS_Si	-0,448	-0,476	-0,509	-0,509	-0,446	-0,478	0,478
FAMI_TIENESERVICIOTV_No	0,460	0,468	0,465	0,465	0,464	0,464	0,464
FAMI_EDUCACIONMADRE_Primaria incompleta	-0,429	-0,424	-0,484	-0,484	-0,427	-0,450	0,450
FAMI_TIENECONSOLAVIDEOJUEGOS_Si	-0,445	-0,449	-0,451	-0,451	-0,448	-0,449	0,449
FAMI_EDUCACIONPADRE_Técnica o tecnológica incompleta	0,414	0,482	0,458	0,458	0,394	0,441	0,441
FAMI_CUARTOSHOGAR_Dos	0,447	0,431	0,433	0,433	0,454	0,439	0,439
FAMI_PERSONASHOGAR_9 o más	-0,428	-0,510	-0,406	-0,406	-0,408	-0,432	0,432
FAMI_TRABAJOLABORPADRE_Es agricultor, pesquero o jornalero	0,463	0,433	0,370	0,370	0,486	0,424	0,424
COLE_DEPTO_UBICACION_MAGDALENA	-0,331	-0,420	-0,525	-0,525	-0,296	-0,419	0,419
COLE_JORNADA_TARDE	-0,412	-0,413	-0,412	-0,412	-0,416	-0,413	0,413
COLE_NATURALEZA_NO OFICIAL	0,402	0,398	0,368	0,368	0,404	0,388	0,388
FAMI_TRABAJOLABORMADRE_Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente)	-0,383	-0,393	-0,384	-0,384	-0,391	-0,387	0,387
COLE_DEPTO_UBICACION_LA GUAJIRA	0,405	0,314	0,376	0,376	0,442	0,383	0,383
COLE_DEPTO_UBICACION_BOYACA	0,420	0,368	0,340	0,340	0,443	0,382	0,382
FAMI_TRABAJOLABORMADRE_Es vendedor o trabaja en atención al público	-0,366	-0,374	-0,375	-0,375	-0,366	-0,371	0,371
COLE_DEPTO_UBICACION_CORDOBA	-0,293	-0,378	-0,444	-0,444	-0,279	-0,368	0,368
FAMI_TRABAJOLABORPADRE_Trabaja por cuenta propia (por ejemplo plomero, electricista)	0,369	0,339	0,338	0,338	0,380	0,353	0,353
ESTU_TIPOREMUNERACION_Si, en efectivo	0,313	0,340	0,364	0,364	0,302	0,336	0,336
FAMI_TRABAJOLABORMADRE_Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo tienda, papelería, etc	-0,338	-0,342	-0,329	-0,329	-0,338	-0,335	0,335
COLE_DEPTO_UBICACION_CAQUETA	0,366	0,294	0,275	0,275	0,396	0,321	0,321
FAMI_TIENEAUTOMOVIL_Si	-0,318	-0,351	-0,310	-0,310	-0,308	-0,319	0,319
EDAD	-0,305	-0,292	-0,339	-0,339	-0,306	-0,316	0,316

Variable	Biología	Física	Lic. En Matemáticas	Matemáticas	Química	prom	abs(prom)
FAMI_EDUCACIONMADRE_Primaria completa	-0,294	-0,309	-0,329	-0,329	-0,290	-0,310	0,310
FAMI_COMECARNEPESCADOHUEVO_3 a 5 veces por semana	0,302	0,327	0,296	0,296	0,292	0,303	0,303
COLE_SEDE_PRINCIPAL_N	-0,281	-0,282	-0,331	-0,331	-0,277	-0,300	0,300
COLE_DEPTO_UBICACION_ANTIOQUIA	0,287	0,329	0,287	0,287	0,272	0,293	0,293
COLE_CARACTER_TÉCNICO	-0,285	-0,299	-0,283	-0,283	-0,280	-0,286	0,286
PROM_5A_COLE_PUNT_MATEMATICAS	0,235	0,225	0,343	0,343	0,239	0,277	0,277
FAMI_TRABAJOLABORPADRE_Trabaja como personal de limpieza, mantenimiento, seguridad o construcción	-0,230	-0,279	-0,319	-0,319	-0,212	-0,272	0,272
FAMI_CUARTOSHOGAR_Uno	-0,265	-0,283	-0,274	-0,274	-0,261	-0,271	0,271
COLE_CALENDARIO_B	0,323	0,155	0,246	0,246	0,291	0,252	0,252
FAMI_EDUCACIONPADRE_Secundaria (Bachillerato) incompleta	-0,250	-0,267	-0,222	-0,222	-0,249	-0,242	0,242
FAMI_COMECARNEPESCADOHUEVO_1 o 2 veces por semana	-0,207	-0,170	-0,251	-0,251	-0,219	-0,220	0,220
FAMI_EDUCACIONPADRE_Primaria completa	-0,228	-0,253	-0,199	-0,199	-0,220	-0,220	0,220
FAMI_TRABAJOLABORMADRE_Trabaja como personal de limpieza, mantenimiento, seguridad o construcción	-0,194	-0,189	-0,247	-0,247	-0,196	-0,215	0,215
COLE_DEPTO_UBICACION_CHOCO	-0,162	-0,267	-0,247	-0,247	-0,130	-0,211	0,211
PROM_5A_COLE_PUNT_C_NATURALES	0,248	0,181	0,161	0,161	0,270	0,204	0,204
PROM_5A_COLE_PUNT_SOCIALES_CIUDADANAS	0,189	0,230	0,183	0,183	0,186	0,194	0,194
FAMI_PERSONASHOGAR_7 a 8	-0,174	-0,228	-0,160	-0,160	-0,159	-0,176	0,176
FAMI_TRABAJOLABORPADRE_Trabaja como profesional (por ejemplo médico, abogado, ingeniero)	0,189	0,157	0,148	0,148	0,198	0,168	0,168
COLE_DEPTO_UBICACION_TOLIMA	0,219	0,173	0,108	0,108	0,230	0,168	0,168
FAMI_EDUCACIONPADRE_Primaria incompleta	-0,159	-0,168	-0,171	-0,171	-0,154	-0,165	0,165
ESTU_NACIONALIDAD_BIN_EXTRANJERO	-0,052	-0,092	-0,294	-0,294	-0,042	-0,155	0,155
FAMI_TRABAJOLABORPADRE_Es operario de máquinas o conduce vehículos (taxita, chofer)	-0,132	-0,197	-0,161	-0,161	-0,113	-0,153	0,153
COLE_DEPTO_UBICACION_GUAINIA	0,061	-0,031	0,310	0,310	0,101	0,150	0,150
FAMI_TRABAJOLABORMADRE_Trabaja por cuenta propia (por ejemplo plomero, electricista)	0,153	0,147	0,136	0,136	0,151	0,145	0,145
FAMI_EDUCACIONMADRE_Secundaria (Bachillerato) incompleta	-0,125	-0,138	-0,155	-0,155	-0,122	-0,139	0,139
COLE_JORNADA_COMPLETA	-0,102	-0,110	-0,156	-0,156	-0,094	-0,123	0,123
PROM_5A_COLE_PUNT_LECTURA_CRITICA	-0,121	-0,081	-0,133	-0,133	-0,143	-0,122	0,122
FAMI_TIENELAVADORA_No	-0,085	-0,082	-0,107	-0,107	-0,080	-0,093	0,093
COLE_JORNADA_UNICA	0,080	0,096	0,104	0,104	0,074	0,092	0,092
PROM_5A_COLE_PUNT_INGLES	0,087	0,086	0,085	0,085	0,088	0,086	0,086
FAMI_TRABAJOLABORMADRE_Trabaja como profesional (por ejemplo médico, abogado, ingeniero)	0,086	0,121	0,074	0,074	0,076	0,086	0,086
FAMI_TRABAJOLABORPADRE_Pensionado	-0,039	-0,050	-0,137	-0,137	-0,024	-0,077	0,077
ESTU_DEDICACIONINTERNET_Más de 3 horas	0,098	0,088	0,045	0,045	0,102	0,076	0,076
COLE_CARACTER_TÉCNICO/ACADÉMICO	-0,078	-0,079	-0,052	-0,052	-0,082	-0,068	0,068
COLE_AREA_UBICACION_RURAL	-0,020	-0,058	-0,115	-0,115	0,000	-0,062	0,062
FAMI_TRABAJOLABORMADRE_Es operario de máquinas o conduce vehículos (taxita, chofer)	0,051	0,039	0,044	0,044	0,051	0,046	0,046
FAMI_PERSONASHOGAR_5 a 6	0,015	-0,021	0,046	0,046	0,024	0,022	0,022
COLE_DEPTO_UBICACION_CESAR	0,088	0,011	-0,069	-0,069	0,111	0,015	0,015
FAMI_COMECEREALFRUTOSLEGUMBRE_Todos o casi todos los días	0,004	0,033	-0,042	-0,042	-0,001	-0,010	0,010
COLE_BILINGUE_S	-0,009	0,013	0,032	0,032	-0,024	0,009	0,009
Intercepto	21,871	21,855	23,207	23,207	21,559	22,340	22,340

Apéndice C

Categorías Base de las Variables Categóricas

Variable	Categoria_Base
ESTU_NACIONALIDAD_BIN	COLOMBIA
ESTU_GENERO	F
ESTU_TIENEETNIA	No
FAMI_ESTRATOVIVIENDA_BIN	estrato3omenos
FAMI_PERSONASHOGAR	3 a 4
FAMI_CUARTOSHOGAR	Tres
FAMI_EDUCACIONPADRE	Secundaria (Bachillerato) completa
FAMI_EDUCACIONMADRE	Secundaria (Bachillerato) completa
FAMI_TRABAJOLABORPADRE	Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo tienda, papelería, etc
FAMI_TRABAJOLABORMADRE	Trabaja en el hogar, no trabaja o estudia
FAMI_TIENEINTERNET	Si
FAMI_TIENESERVICIOTV	Si
FAMI_TIENECOMPUTADOR	Si
FAMI_TIENELAVADORA	Si
FAMI_TIENEHORNOMICROOGAS	No
FAMI_TIENEAUTOMOVIL	No
FAMI_TIENEMOTOCICLETA	No
FAMI_TIENECONSOLAVIDEOJUEGOS	No
FAMI_NUMLIBROS	0 A 10 LIBROS
FAMI_COMELECHEDERIVADOS	1 o 2 veces por semana
FAMI_COMECARNEPESCADOHUEVO	Todos o casi todos los días
FAMI_COMECEREALFRUTOSLEGUMBRE	1 o 2 veces por semana
FAMI_SITUACIONECONOMICA	Igual
ESTU_DEDICACIONLECTURADIARIA	30 minutos o menos
ESTU_DEDICACIONINTERNET	Entre 1 y 3 horas
ESTU_HORASSEMANATRABAJA	0
ESTU_TIPOREMUNERACION	No
COLE_GENERO	MIXTO
COLE_NATURALEZA	OFICIAL
COLE_CALENDARIO	A
COLE_BILINGUE	N
COLE_CARACTER	ACADÉMICO
COLE_SEDE_PRINCIPAL	S
COLE_AREA_UBICACION	URBANO
COLE_JORNADA	MAÑANA
COLE_DEPTO_UBICACION	BOGOTÁ
ESTU_PRIVADO_LIBERTAD	N