Educación en Colombia: Predicción del desempeño en las pruebas Saber 11 mediante técnicas de Machine Learning

Programa de Maestría en Inteligencia Artificial y Ciencia de Datos

Aprendizaje Automático

Proyecto

Martín Alonso Herrera Vargas - Código 22501540 – martin.herrera@uao.edu.co

José Luis Santamaria Andrade – Código 22502265 – jose.santamaria@uao.edu.co

Luisa María Candelo Angulo – Código 22500699 - luisa_mar.candelo@uao.edu.co

Profesor

Francisco Mercado

Junio 5, 2025

1. Descripción del fenómeno y problema a abordar

Colombia enfrenta profundas desigualdades territoriales y sociales que afectan el acceso y la calidad de la educación, especialmente en los niveles medio y superior. Según el Banco Mundial (2024), más de 16 millones de colombianos viven en pobreza, particularmente en La Guajira y Chocó, donde dos de cada tres personas están en esa condición y el empleo formal no supera el 20%. Estas condiciones estructurales impactan directamente en la posibilidad de acceder a una educación de calidad y, por ende, a la educación superior.

Los resultados de las pruebas Saber 11, administradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), reflejan una marcada brecha regional. Departamentos como Chocó, La Guajira, Cauca, Sucre y Nariño presentan consistentemente puntajes muy por debajo del promedio nacional, en contraste con regiones como Bogotá, Cundinamarca y Boyacá (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2024). Esta desigualdad educativa se traduce en una baja tasa de ingreso a la universidad: solo el 55 % de los jóvenes entre 17 y 21 años acceden a la educación superior, muy por debajo del promedio de la OCDE (70 %). Además, la Universidad Nacional de Colombia, la institución pública más prestigiosa del país ha experimentado una caída del 47 % en el número de aspirantes entre 2019 y 2025, lo que evidencia un cuello de botella en el sistema de admisión y una posible desmotivación o falta de preparación de los estudiantes para acceder a la educación superior (Mazo, 2025).

En este contexto, el uso de técnicas de Machine Learning ha emergido como una alternativa prometedora para abordar estas desigualdades. Diversas investigaciones han demostrado su efectividad para predecir el rendimiento académico y orientar intervenciones educativas. Por ejemplo, García (2024) desarrolló modelos de regresión lineal, logística y árboles de decisión para predecir el rendimiento en las pruebas Saber Pro, alcanzando precisiones del 89%, 91% y 77% respectivamente. De manera similar, Giraldo y Mira (2023) aplicaron regresión lineal para predecir el puntaje global en la prueba Saber 11 en Antioquia, obteniendo un coeficiente de determinación de 0.43, lo que indica una capacidad predictiva moderada con margen de mejora. Por su parte, Vargas y Ardila (2024) utilizaron técnicas avanzadas como XGBoost y SHAP para

predecir el desempeño en Saber 11 a partir de variables socioeconómicas, logrando una alta precisión y robustez en los resultados.

Estas investigaciones respaldan la pertinencia de desarrollar y entrenar modelos de aprendizaje automático para predecir de manera precisa la puntuación global en la prueba Saber 11, lo cual permitiría identificar tempranamente a estudiantes con alto potencial o en riesgo, orientar la asignación de recursos educativos y diseñar políticas públicas que reduzcan las brechas socio-regionales de acceso a la educación superior.

2. Obtención y generación del conjunto de datos

El conjunto de datos utilizado en esta investigación corresponde a los resultados históricos de las pruebas Saber 11 aplicadas en Colombia entre los años 2011 y 2022, disponibles públicamente a través del portal de datos abiertos del gobierno colombiano(2024). Esta fuente provee información detallada sobre características sociodemográficas del estudiante (género, estrato socioeconómico, número de personas en el hogar, nivel educativo de los padres), atributos del colegio (naturaleza, ubicación, calendario), ubicación geográfica (departamento y municipio de residencia y presentación), y los resultados obtenidos en las distintas áreas evaluadas (lectura crítica, matemáticas, ciencias naturales, sociales y ciudadanía, inglés), así como el puntaje global.

El dataset original contiene aproximadamente 7,11 millones de registros y 51 variables, con un tamaño de alrededor de 3 GB en formato CSV. Dada esta magnitud, el procesamiento completo excede las capacidades de memoria de Google Colab en su versión gratuita. Por tal motivo, se definieron dos etapas de tratamiento de datos: una en entorno local y otra en la nube.

En la primera fase, se trabajó en un entorno local para realizar una filtración inicial por periodo, seleccionando únicamente los registros comprendidos entre los años 2020 y 2022. Esta decisión respondió tanto a la viabilidad técnica como a la necesidad de contar con datos recientes y homogéneos, posterior a los cambios curriculares implementados en 2019. Posteriormente, se aplicó un ordenamiento ascendente por año y estudiante, y se eliminaron columnas consideradas poco relevantes para el objetivo del estudio, como los códigos identificadores del estudiante, colegio o municipio, al no aportar información explicativa directa. El resultado fue un nuevo archivo con un tamaño optimizado de 150 MB y aproximadamente 1,1 millones de registros.

En la segunda fase, el archivo filtrado fue cargado en Google Colab. Una vez allí, se renombraron las variables para facilitar su lectura y manipulación, se eliminaron los registros cuyo puntaje global era nulo o cero, y se realizaron ajustes mínimos a campos con valores faltantes, en particular el número de personas en el hogar. Esta variable fue preservada para análisis posteriores, dada su posible relevancia como proxy de condiciones de hacinamiento o estructura familiar. En esta etapa no se eliminaron registros duplicados ni se realizó imputación completa de valores anómalos, con el propósito de mantener un análisis fiel al comportamiento real de los datos reportados por el ICFES.

Se realizó también un análisis exploratorio preliminar para conocer la distribución de variables, identificar valores atípicos y preparar el conjunto para su transformación. Finalmente, se redujo la dimensionalidad del conjunto seleccionando únicamente las variables con mayor relevancia explicativa, basándose en un análisis de cardinalidad, correlación y resultados preliminares de modelos basados en árboles. Variables de alta cardinalidad o escaso valor predictivo, como nombres propios de instituciones y municipios, fueron excluidas. Además, se aplicó un Análisis de Componentes Principales (PCA) con fines de visualización, que permitió reducir el conjunto de características a dos componentes principales. Esta proyección también mostró una separación parcial entre clases (alto/bajo rendimiento), lo que valida su utilidad para explorar la estructura del espacio de datos.

3. Definición del problema de aprendizaje automático

El problema abordado en esta investigación corresponde al campo del aprendizaje supervisado, en el que se dispone de un conjunto de datos etiquetados con una variable objetivo conocida, con el fin de entrenar modelos capaces de hacer predicciones. En este caso, se busca modelar y predecir el puntaje global obtenido por los estudiantes en las pruebas Saber 11, utilizando como insumos sus características personales, familiares y educativas. Esta variable es de naturaleza numérica continua, lo que plantea inicialmente un problema de regresión. No obstante, también se aborda como un problema de clasificación binaria, estableciendo un umbral mínimo de puntaje como criterio de admisión hipotético a la educación superior, con el fin de identificar tempranamente estudiantes con mayor o menor probabilidad de alcanzar ese nivel.

Con base en estos dos enfoques, se definen las siguientes dos tareas de aprendizaje automático:

- **Predicción del puntaje global (regresión)**

En esta tarea, el objetivo es estimar el puntaje global en una escala de 0 a 500, a partir de un conjunto de variables explicativas que incluyen aspectos como el estrato de vivienda, el nivel educativo de los padres, la naturaleza y ubicación del colegio, el género del estudiante, entre otras. Para abordar este problema de regresión se implementaron dos modelos:

Regresión lineal: Este modelo busca establecer una relación lineal entre las variables predictoras y el puntaje global. Su simplicidad y carácter explicativo permiten interpretar los coeficientes asociados a cada variable, lo que resulta útil para analizar el peso relativo de cada factor en el desempeño académico. Se probó también una regresión regularizada (ElasticNet) para verificar si se producen cambios significativos.

Árbol de decisión (regresor): Este modelo no paramétrico permite capturar relaciones no lineales y segmentar los datos en subgrupos homogéneos mediante reglas lógicas, facilitando la interpretación de las decisiones del modelo.

- **Clasificación de estudiantes según umbral mínimo (clasificación binaria)**

En este segundo enfoque, el problema se redefine como una tarea de clasificación binaria. Para ello, se creó una variable categórica que indica si un estudiante obtuvo un puntaje global superior al promedio nacional. Esta transformación del problema permite identificar de forma temprana a los estudiantes que presentan mayor probabilidad de superar el umbral necesario para acceder a programas universitarios, y constituye una herramienta útil para focalizar intervenciones académicas o becas. Para esta tarea se entrenó un modelo de regresión logística. Este modelo es utilizado para clasificación binaria, permitiendo estimar la probabilidad de que un estudiante supere el umbral. Su principal ventaja es la facilidad de interpretación y el bajo riesgo de sobreajuste.

4. Preprocesamiento de los datos

Tras la carga del conjunto de datos (correspondiente a los años 2020 a 2022), se procedió al preprocesamiento de la información para adecuarla al entrenamiento de modelos de aprendizaje automático. En esta etapa se aplicaron transformaciones y técnicas

específicas a las variables tanto numéricas como categóricas, con el objetivo de mejorar la calidad y estructura de los datos.

En primer lugar, se estandarizaron los nombres de las variables para facilitar su manipulación. Se eliminaron registros con valores nulos o cero en la variable objetivo *puntaje global*, ya que estos no representan un resultado válido de evaluación. Asimismo, se redujo la dimensionalidad del conjunto de variables mediante una selección basada en criterios de densidad y relevancia, eliminando aquellas con alta cardinalidad o bajo poder explicativo, como los nombres de municipios o sedes educativas. Las variables seleccionadas fueron transformadas mediante OneHotEncoding, lo que permitió convertirlas en variables binarias para facilitar su uso en modelos lineales.

Por otra parte, se realizó la selección de características y reducción de dimensionalidad dado que el conjunto de datos original contenía 51 variables, se aplicó una combinación de técnicas para reducir la dimensionalidad y evitar el sobreajuste. En primer lugar, se eliminaron variables con alta cardinalidad, como nombres de municipios y sedes educativas, que no aportaban valor explicativo directo y dificultaban el procesamiento debido a su codificación categórica dispersa. Luego, se realizó un análisis de densidad (proporción de valores únicos) y una matriz de correlación entre variables numéricas, identificando redundancias o atributos sin relación con la variable objetivo. Como tercera estrategia, se utilizaron modelos preliminares de árboles de decisión para calcular la importancia relativa de cada variable. Esta técnica permitió priorizar aquellas variables que contribuían más a la reducción de la impureza en los nodos del árbol. Finalmente, se seleccionó un subconjunto de variables con base en estos criterios, manteniendo un equilibrio entre capacidad predictiva y simplicidad del modelo. Este proceso permitió reducir el número de variables a menos de la mitad, facilitando el entrenamiento eficiente de los modelos y mejorando su capacidad de generalización.

Adicionalmente, se construyó una nueva variable binaria *puntaje global*, que indica si el estudiante obtuvo un puntaje global superior a la mediana nacional. Esta variable permitió replantear el problema como una tarea de clasificación binaria, útil para modelos como la regresión logística y los árboles de decisión clasificadores.

5. Entrenamiento y evaluación de modelos

Para el entrenamiento de los modelos, se dividió el conjunto de datos en 80% para entrenamiento y 20% para prueba, garantizando una distribución proporcional de la

variable objetivo en ambas particiones. Se aplicaron tres modelos distintos: regresión lineal, regresión logística y árboles de decisión, cada uno evaluado con las métricas correspondientes al tipo de tarea.

- **Regresión lineal**

Se implementó un modelo de regresión lineal ordinaria utilizando las variables categóricas transformadas y las variables numéricas previamente seleccionadas. El modelo mostró un desempeño moderado en las métricas de regresión, con un coeficiente de determinación (R^2) de 46%, lo que indica que el modelo es capaz de explicar casi la mitad de la variabilidad del puntaje global. El error cuadrático medio (MSE) obtenido fue de 1.487, y su raíz (RMSE) fue de 38 puntos, sobre una escala total de 500. Adicionalmente, se exploró el uso del modelo ElasticNet, el cual combina las penalizaciones L1 y L2 para mejorar la generalización del modelo. Para ello, se aplicó un preprocesamiento que incluyó codificación one-hot para las variables categóricas y paso directo para las variables numéricas, integrando el modelo en un pipeline. Se configuraron los hiperparámetros con alpha = 0,5, l1_ratio = 0,5 y max_iter = 10000. Sin embargo, esta alternativa no arrojó mejoras significativas respecto a la regresión lineal simple, manteniéndose el RMSE en 38. Por esta razón, se optó por conservar el modelo base por su simplicidad y capacidad explicativa.

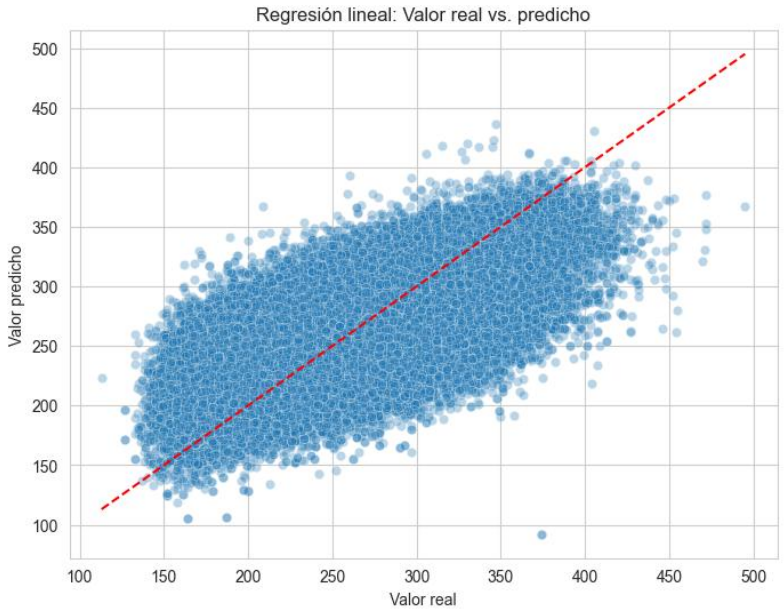

Figura 1. Gráfico de relación valores reales vs valores estimados

El gráfico anterior permite tener una visual del ajuste de los datos reales a la línea de tendencia de la regresión. Idealmente, si el modelo fuera perfecto, todos los puntos caerían sobre la línea roja discontinua (la línea de identidad, donde valor real = valor predicho). Sin embargo, los puntos están algo dispersos alrededor de la línea de tendencia, lo que sugiere que el modelo no captura muy bien la variabilidad de los datos.

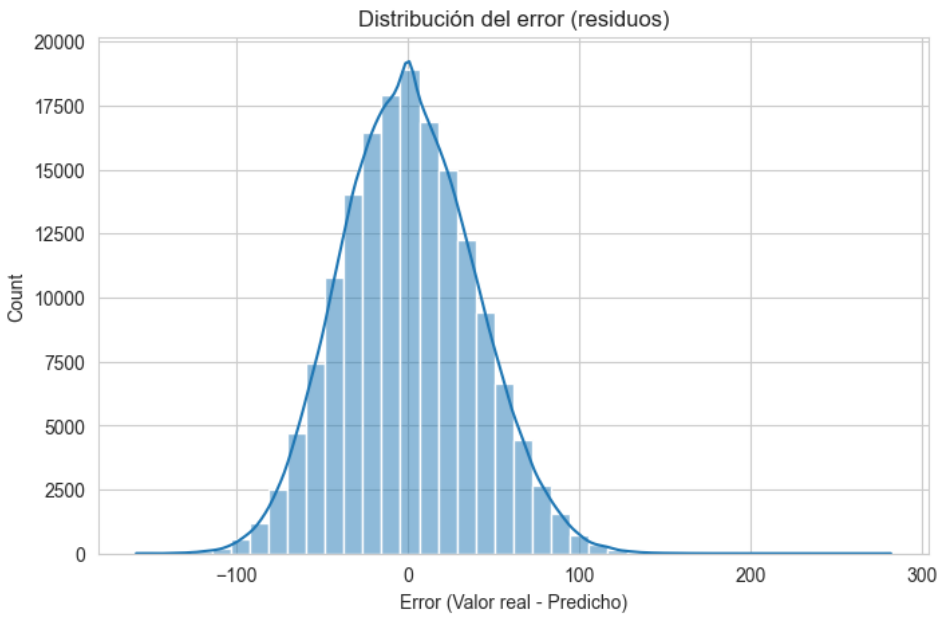

Figura 2. Gráfico de distribución de los valores de los errores del modelo

El gráfico anterior muestra la distribución de los valores de los errores del modelo, los cuales tienen una distribución que se aproxima a una normal.

- **Regresión logística**

Con el modelo de regresión logística se abordó la tarea de clasificación binaria. Para reducir la complejidad, se seleccionaron únicamente variables sociodemográficas de baja cardinalidad, eliminando atributos como nombres de municipios y códigos institucionales. El modelo fue entrenado con 1000 iteraciones, obteniendo resultados consistentes con los esperados. Se generó una matriz de confusión (Figura 3) y un reporte de clasificación que evidenciaron una precisión global (accuracy) del 71%, lo indica que el modelo acierta en aproximadamente 7 de cada 10 predicciones.

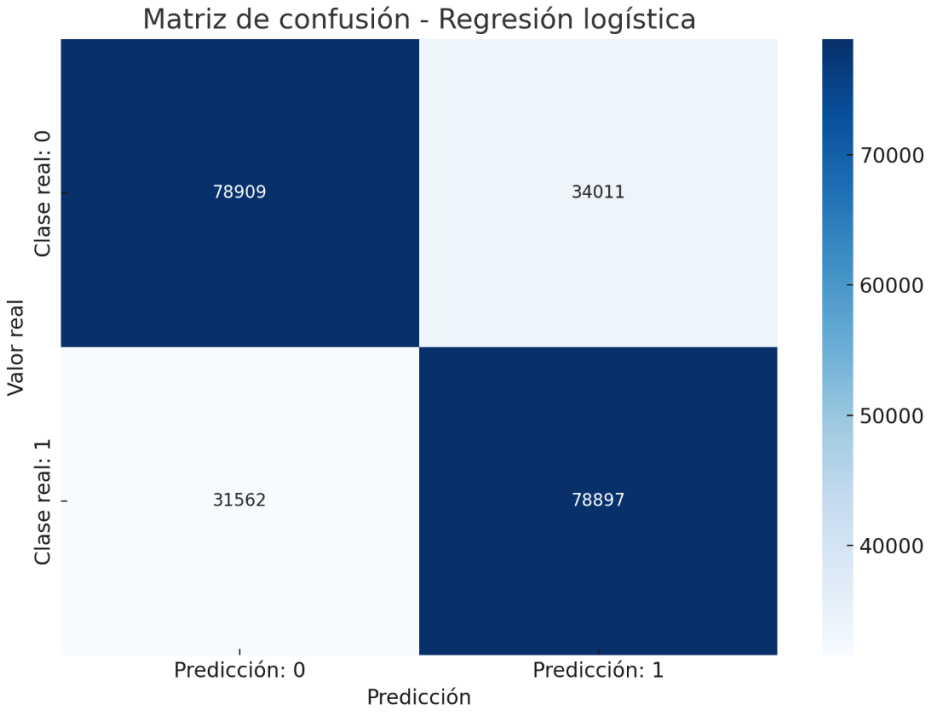

Figura 3. Gráfico de Matriz de confusión de la regresión logística

Como se observa en la matriz de confusión, el modelo clasificó correctamente a 78.909 estudiantes que no superaron el umbral de puntaje, lo que corresponde a los verdaderos negativos (TN). Por otro lado, 34.011 estudiantes que tampoco superaron el umbral fueron clasificados incorrectamente como si lo hubieran hecho, constituyendo los falsos positivos (FP). Asimismo, 31.562 estudiantes que sí superaron el umbral fueron clasificados erróneamente como si no lo hubieran logrado, es decir, falsos negativos (FN). Finalmente, 78.897 estudiantes que efectivamente superaron el umbral fueron correctamente identificados como tales, representando los verdaderos positivos (TP). El modelo muestra un desempeño equilibrado, con una cantidad similar de verdaderos positivos y verdaderos negativos. Sin embargo, también presenta un número considerable de errores (falsos positivos y falsos negativos), lo que sugiere que aún hay margen de mejora, especialmente si se busca reducir el riesgo de clasificar erróneamente a estudiantes con alto potencial académico.

Con el fin de optimizar el rendimiento del modelo de regresión logística, se ejecutó una búsqueda exhaustiva de hiperparámetros mediante validación cruzada con 3 particiones (3-fold cross-validation). Esta búsqueda evaluó un total de 8 combinaciones posibles de parámetros, considerando diferentes valores para la regularización (C), el tipo de penalización (penalty) y el algoritmo de optimización (solver). El criterio de evaluación

utilizado fue el F1-score, una métrica especialmente adecuada para problemas de clasificación binaria con clases potencialmente desbalanceadas, ya que equilibra precisión y exhaustividad. No obstante, tras completar el proceso de validación cruzada, no se observaron mejoras significativas frente al modelo base, por lo que se optó por mantener la configuración original por su simplicidad y eficiencia computacional. Adicionalmente, se realizó un gráfico PCA con el objetivo de reducir las dimensiones de los datos a dos y revisar el comportamiento de su distribución, permitiendo también identificar la separabilidad entre los datos de los estudiantes que superan el umbral o están por debajo:

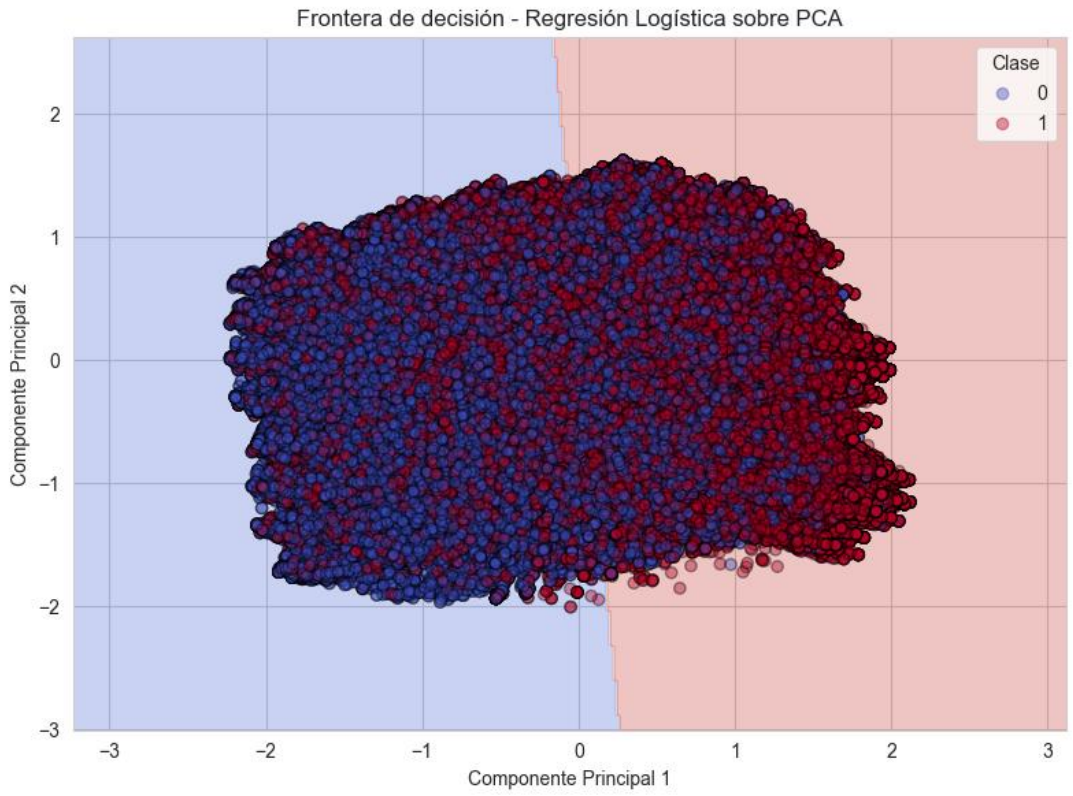

Figura 4. Gráfico de Regresión Logística sobre PCA

- **Árbol de decisión**

El modelo de árbol de decisión fue entrenado inicialmente con una profundidad máxima de 10 niveles, lo cual permitió controlar el sobreajuste y facilitar la interpretación de las reglas generadas. El modelo obtuvo una precisión de 67%. Posteriormente, se implementó un modelo de Random Forest con búsqueda aleatoria de hiperparámetros (RandomizedSearchCV), lo que permitió incrementar la precisión hasta un 69% tras ajustes de hiperparámetros como profundidad y número de árboles. Sin embargo, se

observó que este aumento en la complejidad del modelo y el número de iteraciones no era suficiente para justificar la complejidad computacional adicional, por lo que se priorizó el árbol individual.

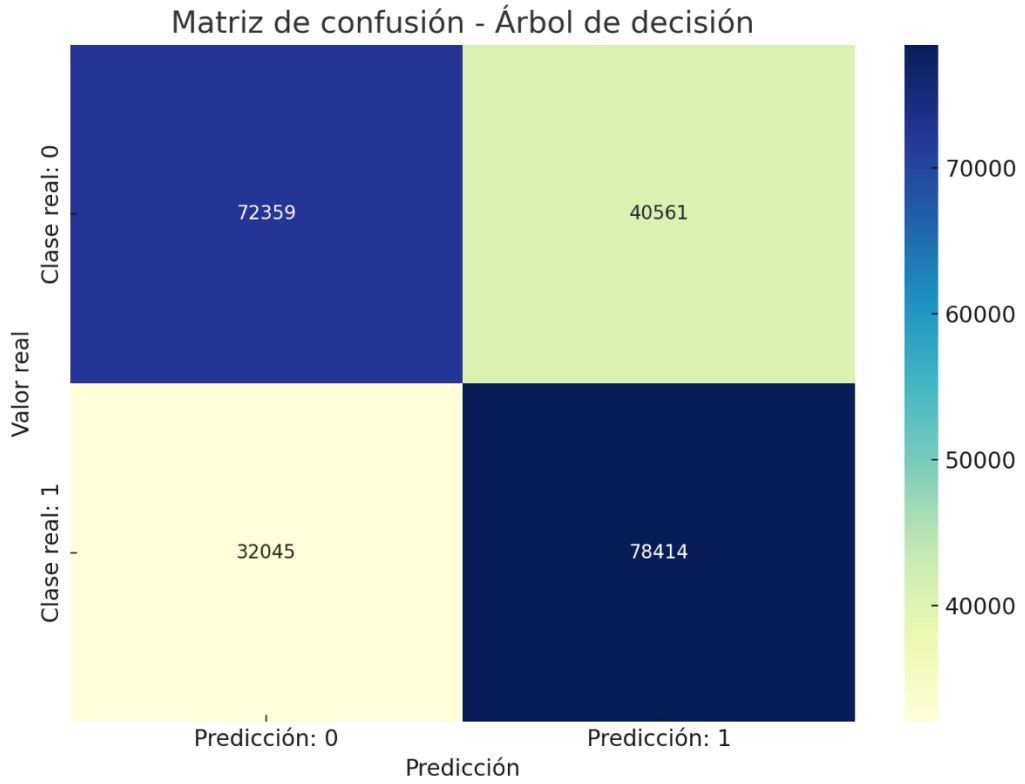

Figura 5. Gráfico de Matriz de confusión del árbol de decisión

Como se observa en la matriz de confusión, el modelo clasificó correctamente a 72.359 estudiantes que no superaron el umbral de puntaje, lo que corresponde a los verdaderos negativos (TN). Por otro lado, 40.561 estudiantes que tampoco superaron el umbral fueron clasificados incorrectamente como si lo hubieran hecho, constituyendo los falsos positivos (FP). Asimismo, 32.045 estudiantes que sí superaron el umbral fueron clasificados erróneamente como si no lo hubieran logrado, es decir, falsos negativos (FN). Finalmente, 78.414 estudiantes que efectivamente superaron el umbral fueron correctamente identificados como tales, representando los verdaderos positivos (TP). Estos resultados reflejan que el modelo presenta un ligero sesgo hacia la clase positiva (puntajes altos) y presenta más falsos positivos (40.561) que falsos negativos (32.045), lo cual indica que tiende a sobreestimar el desempeño de algunos estudiantes, lo que indica que existe margen de mejora.

Por otro lado, dentro del modelo de Árboles de Decisión, se realizó un gráfico con el objetivo de identificar y hacer el ranking de las características que más impactan el resultado de la calificación global con los siguientes resultados:

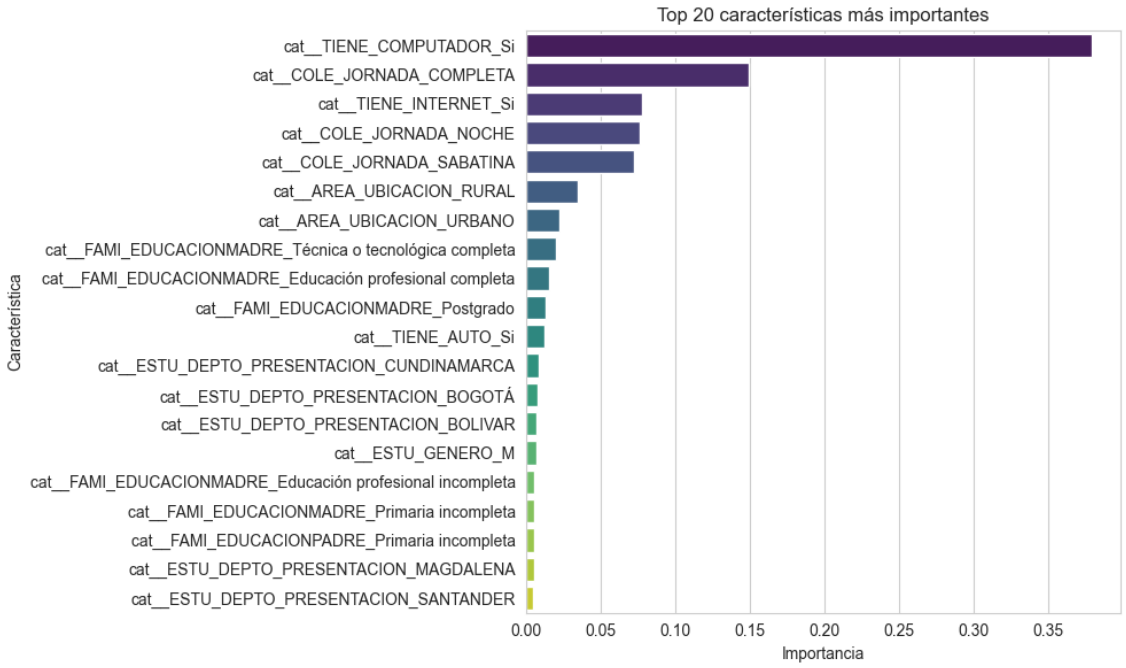

Figura 6. Ranking de las características que más impactan el resultado del modelo

Puede observarse que el acceso a un computador, estudiar en instituciones durante una jornada completa y contar con internet, son los aspectos de mayor impacto para tener un resultado positivo en las pruebas de estado.

6. Conclusiones

Este estudio demuestra la viabilidad de aplicar técnicas de aprendizaje automático para predecir el desempeño de los estudiantes en las pruebas Saber 11, utilizando variables sociodemográficas, educativas y de contexto. A través de modelos de regresión y clasificación, se logró identificar patrones relevantes que permiten anticipar el rendimiento académico con un nivel de precisión aceptable.

Desde el punto de vista técnico, se implementaron y evaluaron tres modelos supervisados: regresión lineal, regresión logística y árboles de decisión. El modelo de regresión lineal mostró una capacidad explicativa moderada, con un coeficiente de determinación (R^2) del 46%, lo que indica que cerca de la mitad de la variabilidad del puntaje global puede ser explicada por las variables seleccionadas. Aunque se exploraron modelos regularizados como ElasticNet, no se observaron mejoras significativas. En el enfoque de clasificación

binaria, la regresión logística alcanzó una precisión del 71%, con un desempeño equilibrado entre clases. El modelo de árbol de decisión, por su parte, logró una precisión del 67%, con una ligera tendencia a sobreestimar el rendimiento de los estudiantes. La visualización de la importancia de las variables reveló que el acceso a computador, internet y la jornada escolar son factores determinantes en el desempeño académico.

Los resultados obtenidos dejan en evidencia que la regresión logística es una herramienta especialmente valiosa para diseñar sistemas de alerta temprana, focalización de ayudas y análisis de equidad educativa a gran escala. Por ejemplo, puede integrarse en sistemas de alerta temprana para identificar estudiantes en riesgo de bajo rendimiento, diseñar programas focalizados de nivelación académica o asignar becas de forma más equitativa.

Estos hallazgos respaldan el uso de modelos de Machine Learning como herramientas complementarias para la toma de decisiones en política educativa, permitiendo identificar estudiantes en riesgo o con alto potencial, y orientar intervenciones focalizadas.

7. Referencias

Banco Mundial. (2024). *Trayectorias: Prosperidad y reducción de la pobreza en el territorio colombiano.* Bogotá: Grupo Banco Mundial.

Garcia, M. A. (2024). *Técnicas de Machine Learning para la predicción del rendimiento académico en las pruebas Saber Pro en Colombia.*

Giraldo, D. F., & Mira, J. F. (2023). *Predicción del puntaje global en la prueba Saber 11 mediante técnicas de Machine Learning.*

Instituto Colombiano para la Evaluación de la Educación (ICFES). (2024). *Puntaje promedio global del examen Saber 11 según departamento.* Bogotá.

Instituto Colombiano para la Evaluación de la Educación (ICFES). (abril de 2024). *Resultados únicos Saber 11*. Obtenido de https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-11/kgxf-xxbe/about_data

Mazo, D. (2025). La Universidad Nacional se queda sola, el desplome histórico de aspirantes prende las alarmas. *infobae*.

Vargas, V., & Ardila, L. F. (2024). *Predicción del desempeño en las pruebas Saber 11 utilizando variables del contexto socio-económico de los aplicantes mediante un análisis estadístico con técnicas de machine learning.*