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

Este proyecto busca desarrollar una herramienta de Machine Learning que ayude a predecir el desempeño de los estudiantes en las pruebas de estado Saber 11, con base en los principales factores socio-demográficos que impactan los resultados.

Luisa María Candelo - 22500699

José Luis Santamaría - 22502265

Martin Alonso Herrera - 22501540

Definición del problema

Desigualdades territoriales y sociales

Colombia enfrenta profundas desigualdades sociales que dificultan el acceso a educación de calidad.

Amplia diferencia de resultados entre departamentos

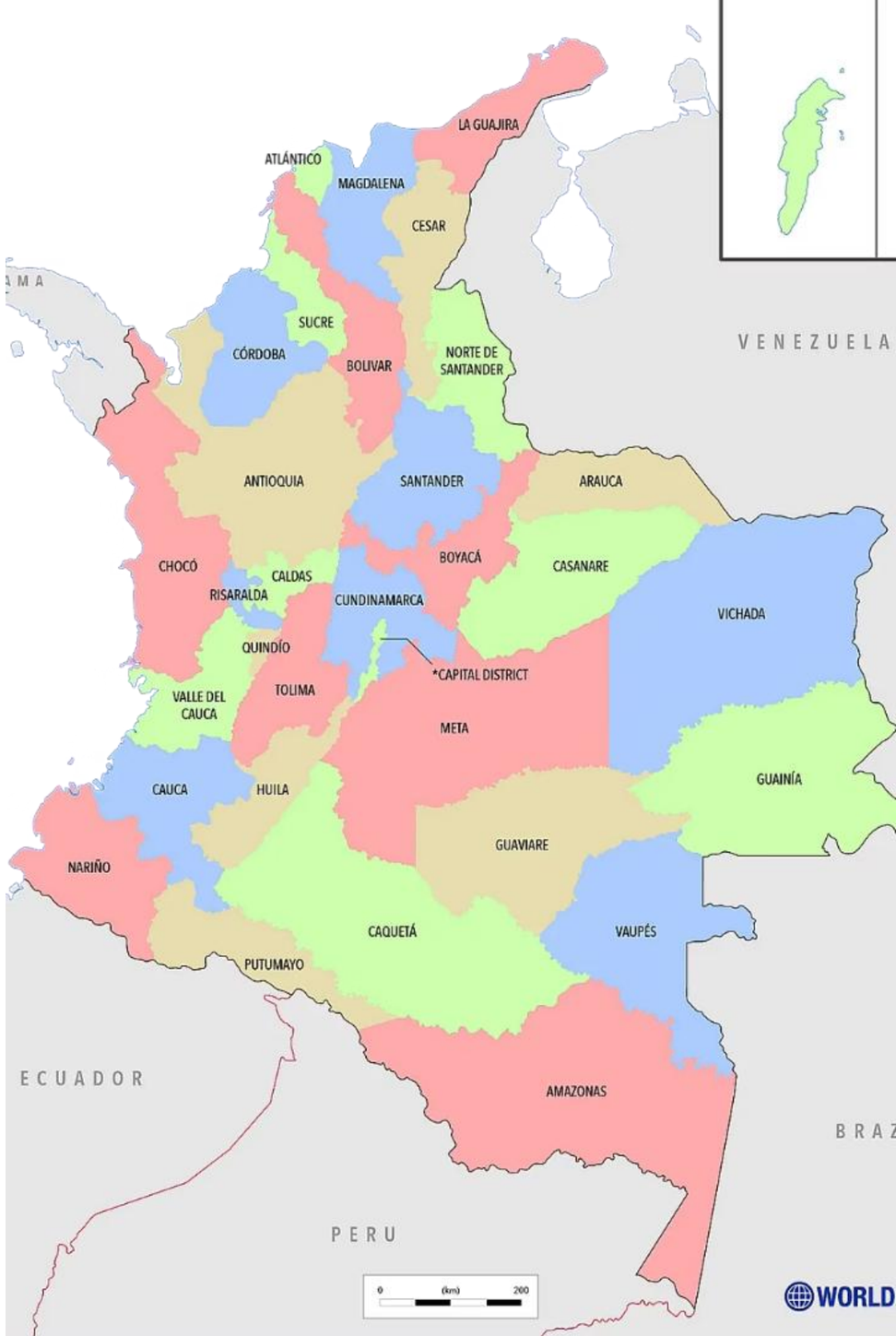
Departamentos como Chocó y La Guajira, quienes tienen índices de pobreza muy altos, presentan puntajes bajos en Saber 11.

Acceso a la Educación Superior

Solo el 55% de los jóvenes accede a la universidad, una cifra muy baja en comparación con los países de la OCDE (70%).

Desinterés en la Educación Superior

La Universidad Nacional ha visto una reducción del 47% en aspirantes entre 2019 y 2025.



Recopilación y Generación de Datos

Fuente de Datos

Resultados de pruebas Saber 11 entre 2011 y 2022.

Características Demográficas

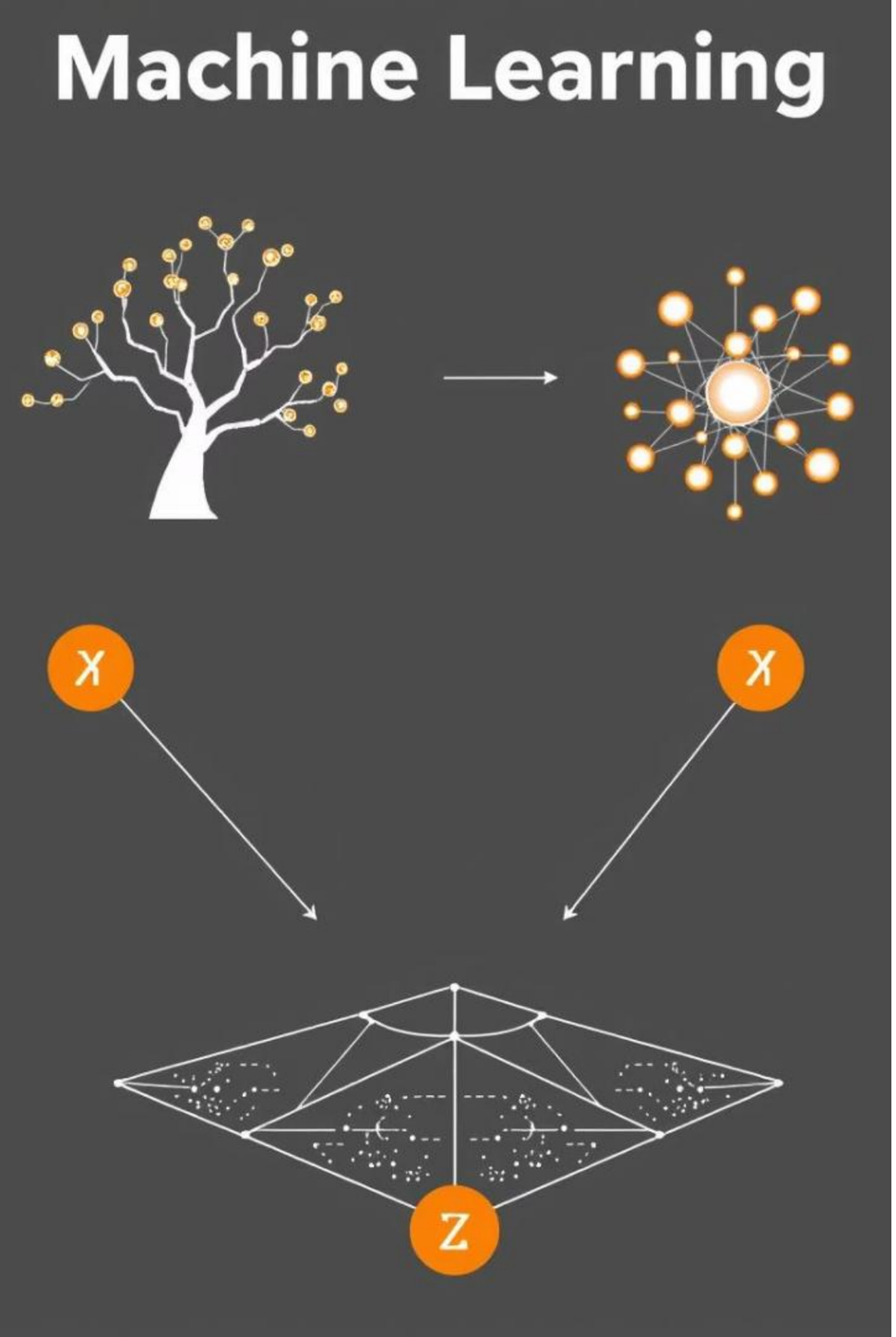
Inclusión de variables demográficas relevantes.

Tamaño del Dataset

7.11 millones de registros y 51 variables.

Formato y Volumen

Aproximadamente 3 GB en formato CSV.



Desarrollo de Modelos de Machine Learning

1 Análisis Exploratorio

Comprender la estructura y distribución de los datos.

Tratamiento de valores nulos.

2 Regresión Lineal

3 Regresión Logística

4 Árboles de Decisión

Modelo de Regresión Lineal

46%

R^2 Coeficiente

Desempeño moderado del modelo.

1.487

MSE

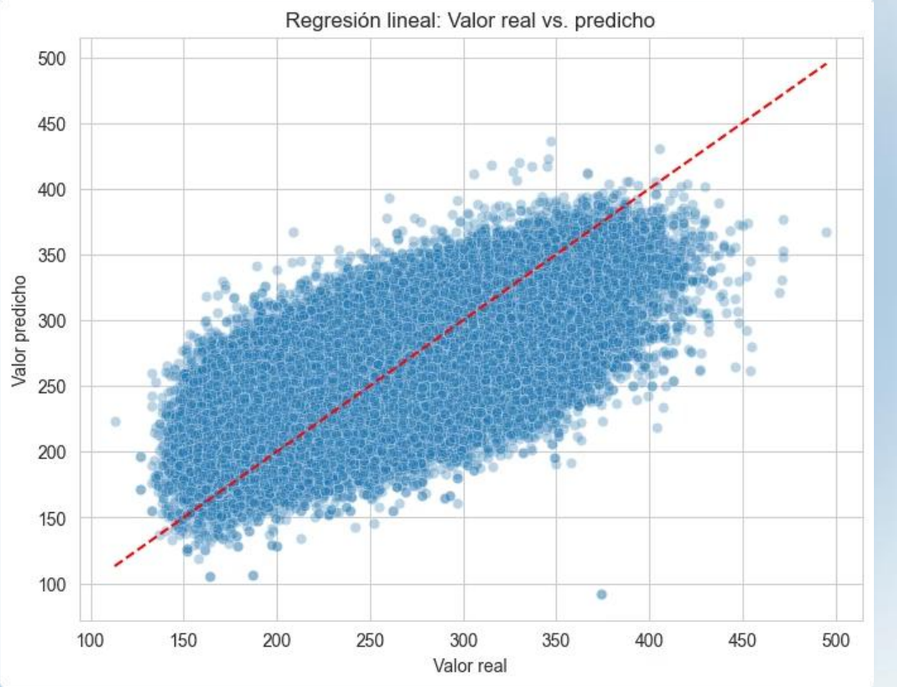
Error Cuadrático Medio.

38

RMSE

Raíz del error cuadrático medio.

Se implementó un modelo de regresión lineal ordinaria. Este modelo mostró un desempeño moderado. Se exploraron alternativas como ElasticNet, sin mejoras significativas.



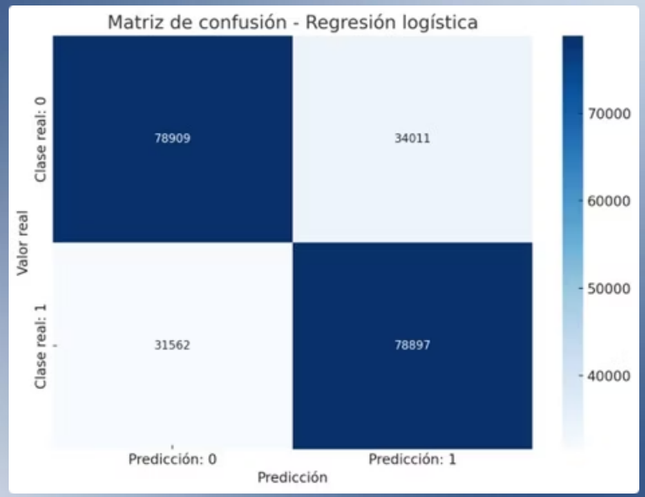
Modelo de Regresión Logística

71%

Accuracy

Desempeño moderado del modelo.

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. El modelo fue entrenado con 1000 iteraciones, obteniendo resultados consistentes con los esperados



Modelo de Árbol de Decisión

Precisión Inicial

El modelo obtuvo una precisión del 67% en clasificación binaria.

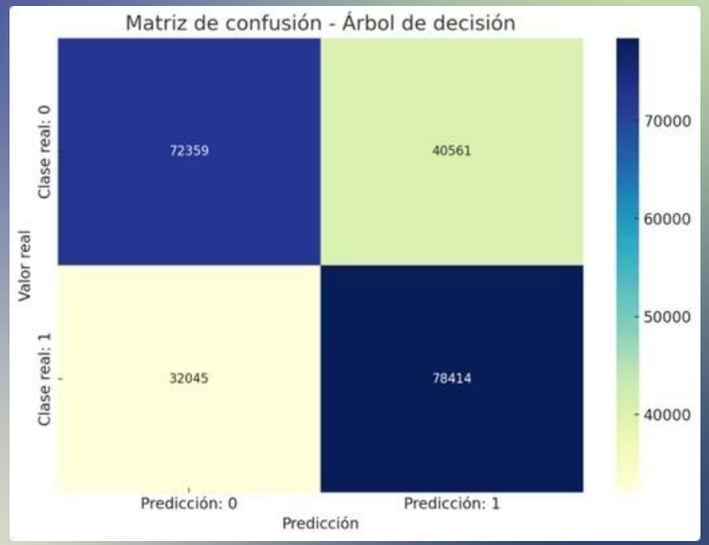
Random Forest

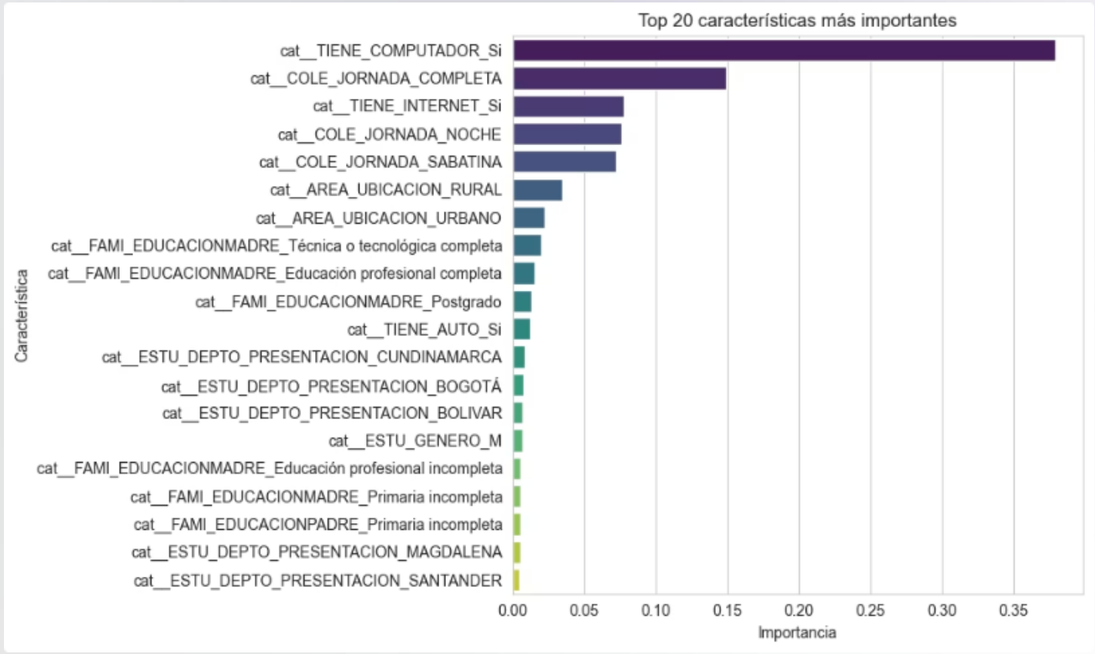
Se exploró el uso de Random Forest.

Accuracy Mejorado

El accuracy inicial fue del 67%, el cual pudo incrementarse hasta 69% mediante iteraciones.

El modelo de árbol de decisión fue entrenado con una profundidad máxima de 10 niveles. Se exploró el uso de Random Forest, con ajustes de hiperparámetros.





Top características más importantes

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.

Conclusiones

Viabilidad

ML puede predecir el desempeño en pruebas Saber 11.

Patrones Relevantes

Identificación de patrones para anticipar rendimiento académico.

Modelos Implementados

Regresión Lineal (46%), Logística (71%), Árbol de Decisión (69%).

Herramienta Complementaria

Apoyo para la toma de decisiones en política educativa.