TRABAJO DE GRADO
Opción Seminario-Diplomado.

Predicción del puntaje global en la prueba Saber 11 mediante técnicas de Machine learning

Corporación Universitaria Remington.
Facultad de ingenierías
Ingeniería de Sistemas

Diego Fernando Giraldo Osorio
John Fredy Mira Mejía
Opción de Trabajo de grado Seminario-Diplomado.
2023

Dedicatoria

A Mercedes, Anyela, Elena y Elizabeth, cuatro mujeres excepcionales en mi vida. A ustedes, mi más profundo agradecimiento por estar siempre ahí, iluminando mi camino con su apoyo y amor incondicional.

Agradecimientos

A todos los profesores que han cruzado mi camino, les expreso mi más sincero agradecimiento por su invaluable vocación. Su dedicación ha dejado una huella duradera en mi aprendizaje y crecimiento personal.

Tabla de contenido

Lista de Tablas

Lista de Figuras

Resumen

Este trabajo de grado se centra en la aplicación de técnicas de machine learning, específicamente regresión lineal, para predecir el puntaje global en la prueba Saber 11 con base en una extensa base de datos que abarca el período 2019-2022 y comprende 296,343 registros de estudiantes del departamento de Antioquia.

La base de datos incluye diversas variables explicativas, como el área del colegio (urbano, rural), la condición bilingüe del colegio, el municipio del colegio, la edad del estudiante, la cantidad de cuartos en el hogar, el nivel educativo de los padres, el estrato socioeconómico, la posesión de automóvil, la disponibilidad de computador e internet, entre otras. Estas variables proporcionan un panorama completo de las condiciones socioeconómicas y educativas de los estudiantes, permitiendo una evaluación integral de su entorno.

La metodología se basa en la implementación de un modelo de regresión lineal que utiliza estas variables como predictores del puntaje global en la prueba Saber 11. La elección de la regresión lineal se fundamenta en su capacidad para modelar relaciones lineales entre variables, lo que facilita la interpretación de la contribución relativa de cada factor al rendimiento estudiantil.

Palabras clave

Machine learning, regresión lineal, prueba saber 11, predicción, educación.

Introducción

La evaluación de la calidad educativa desempeña un papel fundamental en la comprensión y mejora del sistema educativo. En este contexto, la prueba Saber 11 emerge como un indicador clave, proporcionando una evaluación estandarizada de las competencias básicas de los estudiantes. Este estudio se enfoca en el departamento de Antioquia, explorando la posibilidad de utilizar técnicas avanzadas de machine learning, específicamente la regresión lineal, para predecir de manera precisa el puntaje global en la prueba Saber 11.

Con una base de datos robusta que abarca un período significativo de cuatro años (2019-2022) y comprende 296,343 registros de estudiantes, se busca identificar los factores socioeconómicos y educativos que influyen en el rendimiento académico. Variables como el área del colegio, condición bilingüe, municipio del colegio, edad del estudiante, nivel educativo de los padres, estrato socioeconómico, y la disponibilidad de recursos tecnológicos, entre otras, se considerarán como predictores potenciales.

1. Objetivos

1.1 General

Desarrollar y entrenar un modelo de machine learning con una extensa base de datos para predecir de manera precisa la puntuación global en la prueba Saber 11.

1.2 Específicos

- Analizar la base de datos con el fin de entender cómo está estructurada la información.
- Preprocesar la base de datos mediante técnicas de limpieza y transformación, abordando posibles valores atípicos, datos faltantes y normalizando variables para garantizar la calidad y coherencia de la información utilizada en el modelo de machine learning.
- Implementar un modelo de regresión lineal utilizando las variables sociodemográficas y educativas seleccionadas como predictores, entrenando el modelo con una porción significativa de la base de datos y evaluando su desempeño mediante métricas pertinentes.
- Evaluar la generalización del modelo mediante la aplicación de conjuntos de datos independientes, verificando su capacidad para predecir con precisión el puntaje global en contextos educativos similares.

2. Marco conceptual y contextual

La educación es un pilar fundamental para el desarrollo tanto individualmente como colectivamente. En este contexto, la evaluación académica, se convierte en un medio para medir las competencias y habilidades de los estudiantes. El análisis de los resultados de estas evaluaciones no solo es esencial para la rendición de cuentas en el sistema educativo, sino también para identificar áreas de mejora y optimizar estrategias pedagógicas.

De acuerdo con el ICFES la evaluación académica se realiza a través de la prueba Saber 11, una herramienta que proporciona una visión integral del rendimiento estudiantil. Esta evaluación estandarizada no solo mide el conocimiento adquirido por los estudiantes, sino que también sirve como un indicador clave para evaluar la calidad de la educación.

2.1 Base de datos

Una base de datos es un conjunto de datos que representa entidades (Paré,2002). La base de datos utilizada en este trabajo fue obtenida de la plataforma https://www.datos.gov.co, un repositorio de datos abiertos que proporciona acceso público a diversas fuentes de información gubernamental en Colombia. La base de datos en cuestión contiene información detallada sobre los resultados de la prueba Saber 11 en el departamento de Antioquia, abarcando un período que comprende desde el año 2019 hasta el 2022.

2.2 Preparación de los datos

La preparación de los datos se refiere a una serie de técnicas utilizadas para garantizar la calidad, coherencia y adecuación de la información recopilada de la base de datos antes de ser empleada en el modelo de machine learning. Estos procesos son esenciales para asegurar que los datos sean aptos para su análisis y modelado, mejorando la eficacia y confiabilidad del modelo predictivo.

a. **Selección de variables:** en esta etapa se seleccionan las variables independientes que pueden ser usadas para predecir la variable dependiente.
b. **Limpieza de datos:** implica la identificación y tratamiento de valores faltantes, así como la corrección de posibles errores en los datos para garantizar la coherencia y confiabilidad del conjunto.
c. **Transformación de variables:** en esta fase, las variables pueden someterse a transformaciones, como la normalización o estandarización, para asegurar que todas estén en una escala comparable.
d. **División del conjunto de datos:** Para evaluar la eficacia del modelo, se divide el conjunto de datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba se reserva para evaluar su rendimiento en datos no vistos, proporcionando una medida objetiva de la capacidad predictiva.

2.3 Regresión lineal

La regresión lineal posibilita la anticipación del comportamiento de una variable, ya sea dependiente o predictora, a partir de otra que se considera independiente o predictora. (Dagnino, 2014).

En el contexto de este trabajo, la regresión lineal se refiere a una técnica de modelado estadístico utilizada para comprender y prever la relación entre la variable dependiente, en este caso, el puntaje global en la prueba Saber 11, y un conjunto de variables independientes o predictoras, que abarcan aspectos socioeconómicos y educativos de los estudiantes.

2.4 Coeficiente de correlación

Se puede definir el coeficiente de correlación de manera sencilla como una medida estadística que brinda detalles acerca de la relación lineal que existe entre dos variables arbitrarias. (Lahura, 2003).

El coeficiente de correlación es una herramienta crucial para comprender las relaciones entre las variables presentes en la base de datos de los resultados de la prueba Saber 11. Su utilidad se manifiesta de las siguientes maneras:

a. **Selección de variables predictivas:** permite identificar qué variables explicativas están correlacionadas de manera significativa con la variable objetivo (puntaje global en la prueba). Esto es esencial para seleccionar las características más relevantes en la construcción de modelos predictivos.

b. **Entendimiento de influencias:** proporciona información sobre la dirección y fuerza de las relaciones entre las variables, lo que ayuda a entender cómo diferentes factores, como la educación de los padres o la ubicación del colegio, pueden influir en el rendimiento académico de los estudiantes.

c. **Optimización del modelo:** facilita la optimización del modelo de machine learning al identificar qué variables podrían tener un impacto más significativo en la predicción de los puntajes globales. Esto contribuye a la eficiencia y precisión del modelo.

d. **Validación de suposiciones:** permite validar suposiciones sobre la relevancia de ciertos factores en el contexto educativo, respaldando o desafiando las hipótesis iniciales.

2.5 Machine learning

Bobadilla, (2021) menciona: “Machine learning es la ciencia que hace que los ordenadores aprendan a partir de los datos” (p.1).

Machine learning se refiere a un enfoque de la inteligencia artificial utilizado para desarrollar y entrenar modelos predictivos capaces de aprender patrones complejos a partir de datos. Se aplicará el machine learning, junto con regresión lineal, para analizar y prever el rendimiento académico de los estudiantes en la prueba Saber 11 en el departamento de Antioquia.

3. Desarrollo e implementación del aprendizaje

La implementación del modelo de machine learning para la predicción del puntaje global de la prueba Saber 11 se llevará a cabo mediante el uso de Python, aprovechando las poderosas bibliotecas y herramientas disponibles en este lenguaje de programación. Python ofrecerá una plataforma flexible y eficiente para desarrollar, entrenar y evaluar el modelo, permitiendo un análisis preciso y riguroso de los factores que influyen en el desempeño académico de los estudiantes.

3.1 Entendimiento de la base de datos

La base de datos utilizada es un subconjunto que se obtuvo de https://www.datos.gov.co. Este subconjunto abarca el período entre los años 2019 y 2022 en el departamento de Antioquia, Colombia, y contiene información detallada de 296,343 registros de estudiantes que participaron en la prueba Saber 11, a continuación, se puede visualizar cada una de las columnas que componen la base de datos:

Tabla 1 Total columnas base de datos.

Fuente: propia.

Nombre de la Columna	Tipo de Dato	Valor de Ejemplo
PERIODO	int64	20194
ESTU_TIPODOCUMENTO	object	TI
ESTU_CONSECUTIVO	object	SB11201940215679
COLE_AREA_UBICACION	object	URBANO
COLE_BILINGUE	object	N
COLE_CALENDARIO	object	A
COLE_CARACTER	object	TÉCNICO/ACADÉMICO
COLE_COD_DANE_ESTABLECIMIENTO	int64	105360001055
COLE_COD_DANE_SEDE	int64	105360001055
COLE_COD_DEPTO_UBICACION	int64	5
COLE_COD_MCPIO_UBICACION	int64	5360
COLE_CODIGO_ICFES	int64	36020
COLE_DEPTO_UBICACION	object	ANTIOQUIA
COLE_GENERO	object	MIXTO
COLE_JORNADA	object	MAÑANA
COLE_MCPIO_UBICACION	object	ITAGÜÍ
COLE_NATURALEZA	object	OFICIAL
COLE_NOMBRE_ESTABLECIMIENTO	object	INSTITUCION EDUCATIVA DIEGO ECHAVARRIA MISAS
COLE_NOMBRE_SEDE	object	INSTITUCION EDUCATIVA DIEGO ECHAVARRIA MISAS
COLE_SEDE_PRINCIPAL	object	S
ESTU_COD_DEPTO_PRESENTACION	int64	5
ESTU_COD_MCPIO_PRESENTACION	int64	5360
ESTU_COD_RESIDE_DEPTO	float64	5
ESTU_COD_RESIDE_MCPIO	float64	5360
ESTU_DEPTO_PRESENTACION	object	ANTIOQUIA
ESTU_DEPTO_RESIDE	object	ANTIOQUIA
ESTU_ESTADOINVESTIGACION	object	PUBLICAR
ESTU_ESTUDIANTE	object	ESTUDIANTE
ESTU_FECHANACIMIENTO	object	11/01/2003
ESTU_GENERO	object	M

ESTU_MCPIO_PRESENTACION	object	ITAGÜÍ
ESTU_MCPIO_RESIDE	object	ITAGÜÍ
ESTU_NACIONALIDAD	object	COLOMBIA
ESTU_PAIS_RESIDE	object	COLOMBIA
ESTU_PRIVADO_LIBERTAD	object	N
FAMI_CUARTOSHOGAR	object	Uno
FAMI_EDUCACIONMADRE	object	Secundaria (Bachillerato) completa
FAMI_EDUCACIONPADRE	object	Primaria incompleta
FAMI_ESTRATOVIVIENDA	object	Estrato 3
FAMI_PERSONASHOGAR	object	3 a 4
FAMI_TIENEAUTOMOVIL	object	No
FAMI_TIENECOMPUTADOR	object	Si
FAMI_TIENEINTERNET	object	Si
FAMI_TIENELAVADORA	object	Si
DESEMP_INGLES	object	A1
PUNT_INGLES	float64	52
PUNT_MATEMATICAS	int64	66
PUNT_SOCIALES_CIUDADANAS	int64	57
PUNT_C_NATURALES	int64	58
PUNT_LECTURA_CRITICA	int64	48
PUNT_GLOBAL	int64	284

En total, la base de datos está compuesta por 51 columnas. Más adelante se llevará a cabo un análisis para descartar aquellas variables que no contribuirían a predecir la variable de interés. Durante la exploración de la base de datos, se observó que las columnas tienen un prefijo que indica si el valor está relacionado con el colegio, el estudiante o la familia del estudiante.

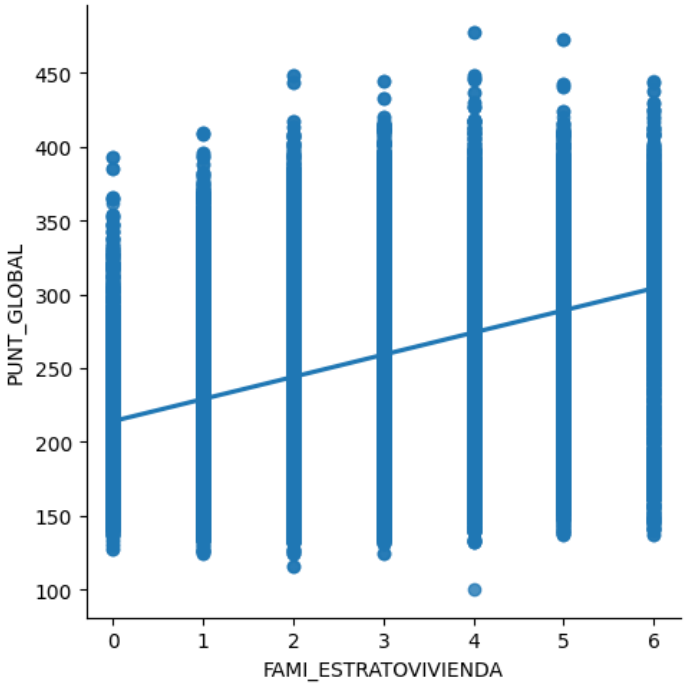
3.2 Preprocesamiento de la base de datos

3.2.1 Selección de variables

Tras un exhaustivo análisis de la base de datos original, se procedió a descartar aquellas variables que no aportan información significativa para predecir la variable dependiente (puntaje global). Este proceso de selección cuidadosa es muy importante porque garantiza que el modelo de machine learning se base únicamente en las variables relevantes, optimizando así la precisión y la capacidad de generalización del modelo. El conjunto resultante de variables predictoras se ha refinado estratégicamente para centrarse en aquellas que ofrecen la mayor contribución al objetivo de predecir el rendimiento académico de los estudiantes, a continuación, se pueden visualizar algunas gráficas que ayudaron a tomar la decisión de qué variable podían explicar mejor la puntuación global:

Figura 1 Relación entre puntaje global y estrato.

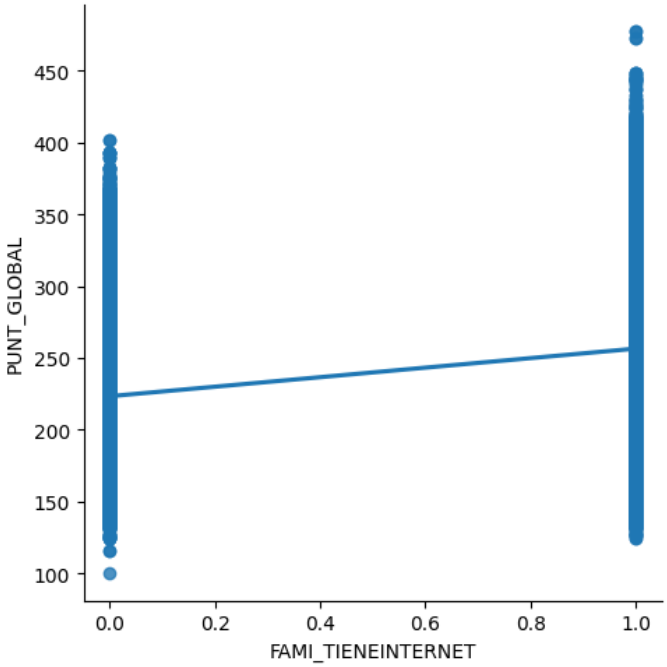
Fuente: propia.



En la Figura 1, se destaca una correlación positiva entre el puntaje global y la variable "estrato". Se observa un aumento progresivo en el puntaje global a medida que el estrato socioeconómico aumenta, sugiriendo una relación significativa entre este factor y el rendimiento académico.

Figura 2 Relación entre puntaje global y acceso a internet.

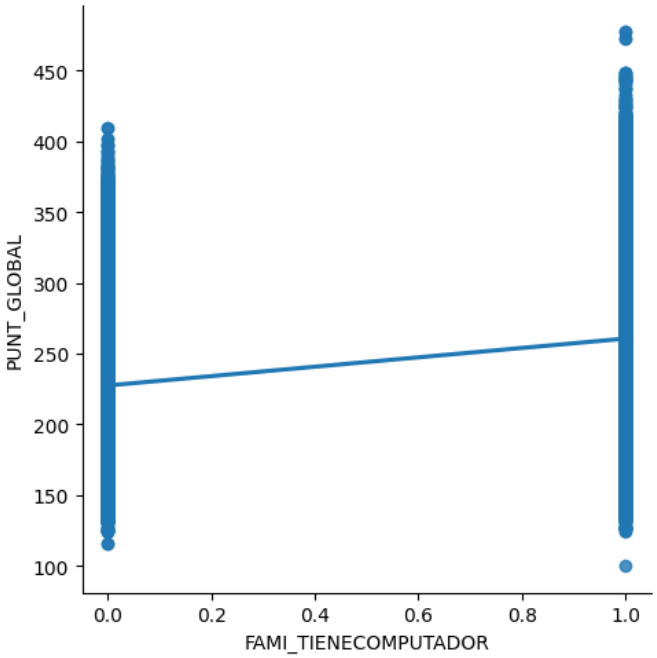
Fuente: propia.



La Figura 2 proporciona una visualización clara de la relación positiva entre el puntaje global obtenido en la prueba Saber 11 y la variable "familia tiene internet". Este hallazgo sugiere que los estudiantes cuyas familias cuentan con acceso a internet tienden a obtener puntajes globales más altos en la prueba. La correlación positiva indica una posible influencia de la conectividad a internet en el rendimiento académico, aspecto que podría explorarse más a fondo en el análisis de variables explicativas en el contexto de este estudio.

Figura 3 Relación entre puntaje global y acceso a computador.

Fuente: propia.



En la Figura 3, se evidencia una correlación positiva entre el puntaje global y la variable "acceso a computador". Los resultados revelan que a medida que la presencia de acceso a computadora se acerca a 1, se observa un incremento en los puntajes globales, indicando una posible influencia positiva del acceso a computadoras en el rendimiento académico.

Tabla 2 Columnas explicativas de la base de datos.

Fuente: propia.

Nombre de la Columna	Tipo de Dato	Valor de Ejemplo
PERIODO	int64	20194
COLE_AREA_UBICACION	object	URBANO
COLE_BILINGUE	object	N
COLE_CARACTER	object	TÉCNICO/ACADÉMICO
COLE_JORNADA	object	MAÑANA
COLE_MCPIO_UBICACION	object	ITAGÜÍ
ESTU_FECHANACIMIENTO	object	11/01/2003
FAMI_CUARTOSHOGAR	object	Uno
FAMI_EDUCACIONMADRE	object	Secundaria (Bachillerato) completa
FAMI_EDUCACIONPADRE	object	Primaria incompleta
FAMI_ESTRATOVIVIENDA	object	Estrato 3
FAMI_PERSONASHOGAR	object	3 a 4
FAMI_TIENEAUTOMOVIL	object	No
FAMI_TIENECOMPUTADOR	object	Si
FAMI_TIENEINTERNET	object	Si
FAMI_TIENELAVADORA	object	Si
PUNT_GLOBAL	int64	284

3.2.2 Transformación de variables y limpieza de datos

A continuación, se analizan una a una las variables explicativas, identificando sus posibles valores y, de acuerdo con estos, se llevan a cabo las transformaciones o limpiezas necesarias.

Tabla 3 Diccionario de datos y transformaciones necesarias.

Fuente: propia

Nombre de la Columna	Posibles valores	¿Transformación necesaria?
PERIODO	Año de presentación de la prueba	No, esta columna solo se usará para calcular la edad del estudiante al momento de presentar la prueba.
COLE_AREA_UBICACION	Urbano, Rural	Convertir variable categórica en varias columnas binarias utilizando la estrategia de codificación dummy encoding.
COLE_BILINGUE	S, N	Transformar sus valores a valores numéricos, donde S=1 y N=0.
COLE_CARACTER	Académico, Técnico/Académico, Técnico	Eliminar los registros con valor "NO APLICA". Convertir variable categórica en varias columnas binarias utilizando la estrategia de codificación dummy encoding.
COLE_JORNADA	Mañana, Completa, única, Sabatina, Tarde, Noche	Convertir variable categórica en varias columnas binarias utilizando la estrategia de codificación dummy encoding.
COLE_MCPIO_UBICACION	Municipios de Antioquia	Convertir variable categórica en varias columnas binarias utilizando la estrategia de codificación dummy encoding.
ESTU_FECHANACIMIENTO	Fecha de nacimiento del estudiante	Calcular edad usando la columna "PERIODO"
FAMI_CUARTOSHOGAR	Uno, Dos, Tres, Cuatro, Cinco, Seis o más	Convertir el valor a entero

FAMI_EDUCACIONMADRE	Mayor grado académico cursado	Eliminar los registros con valor "No aplica" y "No sabe". Convertir variable categórica en varias columnas binarias utilizando la estrategia de codificación dummy encoding.
FAMI_EDUCACIONPADRE	Mayor grado académico cursado	Eliminar los registros con valor "No aplica" y "No sabe". Convertir variable categórica en varias columnas binarias utilizando la estrategia de codificación dummy encoding.
FAMI_ESTRATOVIVIENDA	Estrato 1, Estrato 2, Estrato 3, Estrato 4, Estrato 5, Estrato 6, Sin Estrato	Convertir el valor a valores numéricos, donde "Sin Estrato" = 0
FAMI_PERSONASHOGAR	1 a 2, 3 a 4, 5 a 6, 7 a 8, 9 o más	Convertir el valor a valores numéricos, usando el primer dígito de las categorías.
FAMI_TIENEAUTOMOVIL	Si, No	Transformar sus valores a valores numéricos, donde Si=1 y No=0.
FAMI_TIENECOMPUTADOR	Si, No	Transformar sus valores a valores numéricos, donde Si=1 y No=0.
FAMI_TIENEINTERNET	Si, No	Transformar sus valores a valores numéricos, donde Si=1 y No=0.
FAMI_TIENELAVADORA	Si, No	Transformar sus valores a valores numéricos, donde Si=1 y No=0.

Después de realizar la transformación, se identificaron registros inválidos (NaN), los cuales fueron eliminados.
La siguiente tabla presenta los coeficientes de correlación entre la variable de interés y diversas variables explicativas. Los coeficientes oscilan entre -0,28 y 0,41, indicando la fuerza y dirección de la relación entre cada variable y el puntaje global obtenido en la prueba Saber 11.

Tabla 4 Coeficiente de correlación por variable.

Fuente: propia

Variable	**Coeficiente**
COLE_BILINGUE	0,05
COLE_CARACTER	-0,04
FAMI_CUARTOSHOGAR	0,03
FAMI_EDUCACIONMADRE	0,41
FAMI_EDUCACIONPADRE	0,39
FAMI_ESTRATOVIVIENDA	0,31
FAMI_PERSONASHOGAR	-0,11
FAMI_TIENEAUTOMOVIL	0,23
FAMI_TIENECOMPUTADOR	0,31
FAMI_TIENEINTERNET	0,27
FAMI_TIENELAVADORA	0,17
COLE_AREA_UBICACION_RURAL	-0,09
COLE_AREA_UBICACION_URBANO	0,09
COLE_JORNADA_COMPLETA	0,2
COLE_JORNADA_MAÑANA	0,02
COLE_JORNADA_NOCHE	-0,2
COLE_JORNADA_SABATINA	-0,28
COLE_JORNADA_TARDE	0
COLE_JORNADA_UNICA	0,11
ESTU_EDAD	-0,25
COLE_NATURALEZA_NO OFICIAL	0,2
COLE_NATURALEZA_OFICIAL	-0,2

Esta información permite descartar algunas variables que no fueron eliminadas tras el primer análisis.

Como resultado, la base de datos final quedó con 235242 registros y 1324 columnas.

Figura 4 Muestra de datos posterior a la limpieza.

Fuente: propia.

	FAMI_CUARTOSHOGAR	FAMI_EDUCACIONMADRE	FAMI_EDUCACIONPADRE	FAMI_ESTRATOVIVIENDA	FAMI_PERSONASHOGAR	FAMI_TIENEAUTOMOVIL	FAMI_TIENECOMPUTADOR
0	1.0	4.0	1.0	3.0	3.0	0.0	1.0
1	1.0	4.0	1.0	3.0	3.0	0.0	1.0
2	4.0	2.0	4.0	1.0	5.0	0.0	0.0
3	4.0	2.0	4.0	1.0	5.0	0.0	0.0
4	2.0	6.0	6.0	3.0	3.0	1.0	1.0
...	...	...	...	...	...	...	...
262407	3.0	4.0	4.0	1.0	3.0	1.0	1.0
262408	2.0	9.0	9.0	5.0	3.0	1.0	1.0
262409	5.0	9.0	9.0	5.0	5.0	1.0	1.0
262410	2.0	8.0	8.0	6.0	1.0	1.0	1.0
262411	3.0	8.0	8.0	3.0	3.0	1.0	1.0

235242 rows × 1324 columns

3.3 Implementación del modelo

Se procedió a dividir la base de datos en cuatro conjuntos mediante la función train_test_split de la biblioteca sklearn, asignando un 33% de los datos para representar tanto los conjuntos de entrenamiento como de prueba. Para la construcción del modelo, se empleó la función LinearRegression de sklearn.

Una vez entrenado el modelo, se llevaron a cabo cálculos de métricas de evaluación, específicamente el error cuadrático medio (RMSE) y el coeficiente de determinación (R2), para evaluar el rendimiento del modelo. Los resultados obtenidos fueron un RMSE de 39.43 y un R2 de 0.43.

En este caso, un RMSE de 39.43 indica que, en promedio, las predicciones del modelo tienden a desviarse aproximadamente 39.43 unidades de los valores reales de los puntajes globales en la prueba Saber 11. Por otro lado, el coeficiente de determinación (R2) de 0.43 revela que alrededor del 43% de la variabilidad en los puntajes globales puede ser explicada por las variables utilizadas en el modelo. Este valor indica una capacidad moderada del modelo para explicar la variabilidad observada en los datos.

En conclusión, estos resultados sugieren que el modelo tiene cierta capacidad predictiva, pero hay margen para mejoras. Se debe explorar posibles ajustes en las variables o considerar modelos más complejos para mejorar la precisión de las predicciones.

Figura 5 Código del modelo y resultado obtenido.

Fuente: propia.

```
X,y = tc_data[:,0:1322] , tc_data[:,1323]

X_train, X_test, y_train, y_test = train_test_split(X,
                                   y,
                                   test_size=0.33,
                                   random_state=1)

reg = LinearRegression()

reg.fit(X_train, y_train)

predict = reg.predict(X_train)

rmse = (np.sqrt(mean_squared_error(y_train, predict)))

r2 = round(reg.score(X_train, y_train),2)

print('RMSE: {}'.format(rmse))
print('R2: {}'.format(r2))
```

```
RMSE: 39.43100186555709
R2: 0.43
```

3.4 Evaluar la generalización del modelo

La evaluación del modelo utilizando los datos de prueba (X_test) reveló un rendimiento consistente. Se calculó el error cuadrático medio (RMSE), que se situó en 39.05. Además, se obtuvo un coeficiente de determinación (R2) de 0.44, lo cual sugiere que aproximadamente el 44% de la variabilidad en los puntajes globales de la prueba Saber 11 puede ser explicada por las variables utilizadas en el modelo de regresión lineal.

En conjunto, estos resultados respaldan la utilidad del modelo para predecir los puntajes globales en la prueba Saber 11 y proporcionan una comprensión significativa de la relación entre las variables consideradas. No obstante, es importante reconocer que aún existe un porcentaje considerable de variabilidad no explicada, lo que indica que hay otros factores no considerados en el modelo que pueden influir en los resultados.

Figura 6 Resultado del modelo con los datos de prueba.

Fuente: propia.

```
reg = LinearRegression()

reg.fit(X_test, y_test)

predict = reg.predict(X_test)

rmse = (np.sqrt(mean_squared_error(y_test, predict)))

r2 = round(reg.score(X_test, y_test),2)

print('RMSE: {}'.format(rmse))
print('R2: {}'.format(r2))
```

```
RMSE: 39.0464456175613
R2: 0.44
```

4. Conclusiones

1. La calidad de los resultados del modelo está intrínsecamente ligada a la calidad de los datos. La fase de preparación de datos, que involucra la limpieza y transformación adecuada, resulta esencial para garantizar la eficacia del modelo de machine learning.

2. La identificación y selección cuidadosa de las variables explicativas son fundamentales. En este ejercicio, la correlación entre las variables y el puntaje global en la prueba Saber 11 proporcionó insights valiosos sobre qué factores podrían influir significativamente en el rendimiento académico de los estudiantes.

3. La evaluación del modelo mediante métricas como el RMSE y el coeficiente de determinación R2 es crítica para entender su desempeño en datos no vistos. Estos indicadores ofrecen una medida cuantitativa del grado de ajuste y la capacidad predictiva del modelo. En este caso, los resultados indican que el modelo tiene una capacidad moderada para predecir los puntajes globales en la prueba Saber 11, pero hay margen para mejoras.

4. La distinción entre conjuntos de entrenamiento y prueba en este trabajo es esencial para evaluar la eficacia del modelo en predecir los puntajes globales de la prueba Saber 11. Al emplear datos distintos para entrenamiento y prueba, aseguramos que el modelo pueda generalizar patrones más allá de los datos utilizados inicialmente, validando así su utilidad en situaciones del mundo real.

5. Referencias

Paré, R. C. (2002). *Introducción a las bases de datos*. UOC, la universidad virtual.

Dagnino, J. (2014). *Regresión lineal*. Revista chilena de anestesia.

Bobadilla, J. (2021). *Machine learning y deep learning: usando Python, Scikit y Keras*. Ediciones de la U.

Lahura, E. (2003). *El coeficiente de correlación y correlaciones espúreas (Vol. 218)*. Pontificia Universidad Católica del Perú.

ICFES, *Acerca del examen saber 11.* Sitio web: https://www.icfes.gov.co/acerca-del-examen-saber-11%C2%B0