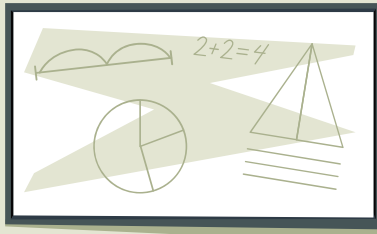
Alumno: Martín, Rodrigo G.

Factores del Rendimiento Estudiantil

Un enfoque basado en datos para mejorar la educación



Comenzar

Profesor: Benitez, Gustavo

Curso: Data Science II: Machine Learning para la Ciencia de Datos

Comisión 61680 – *CODERHOUSE – 05/04/2025*

¿Cuáles son los factores clave más influyentes?

Objetivo General

Desarrollar un modelo predictivo para Exam_Score basado en factores socioeconómicos, demográficos y académicos

Objetivos Específicos

01.

Identificar los factores más influyentes

Determinar cuáles de las variables (categóricas y numéricas) tienen el mayor impacto en el rendimiento académico de los estudiantes.

02.

Desarrollar un modelo predictivo preciso

Crear un modelo de Machine Learning que pueda predecir con precisión el Exam_Score de un estudiante en función de las variables disponibles.

03.

Evaluar el rendimiento del modelo

Medir la precisión y la capacidad de generalización del modelo predictivo utilizando métricas adecuadas.

04.

Proporcionar información útil

Generar informes y visualizaciones que permitan a las partes interesadas (educadores, padres, estudiantes) comprender mejor los factores que influyen en el rendimiento académico y tomar decisiones informadas.

Análisis Exploratorio de Datos (EDA)

Recopilación de datos

Obtener los datos desde archivo CSV

Limpieza de datos

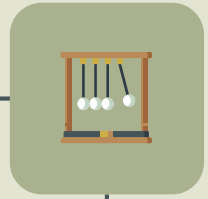
Detectar y manejar valores faltantes, duplicados o inconsistentes.

Visualización de datos

Crear gráficos para identificar patrones, tendencias.

Resumen estadístico

Generar descripciones estadísticas y análisis de correlación.



Detección de outliers

Detectar y manejar valores atípicos.

Factores Categóricos

Parental_Involvement: Nivel de participación de los padres en la educación del estudiante.	**Teacher_Quality:** Nivel de efectividad y competencia del profesor en la enseñanza.
Access_to_Resources: Disponibilidad de materiales y recursos educativos.	**Peer_Influence:** Impacto del grupo de compañeros en el comportamiento y desempeño académico
Extracurricular_Activities: Participación en actividades fuera del programa académico regular.	**Physical_Activity:** Cantidad de horas de actividad física que realiza el estudiante en la semana.
Sleep_Hours: Promedio de horas que el estudiante duerme por noche.	**Learning_Disabilities:** Presencia de dificultades específicas de aprendizaje en el estudiante.
Motivation_Level: Grado de interés y compromiso del estudiante hacia sus estudios.	**Parental_Education_Level:** Nivel máximo de educación alcanzado por los padres del estudiante.
Internet_Access: Disponibilidad de conexión a internet en el hogar del estudiante.	**Distance_from_Home:** Distancia entre el hogar del estudiante y la institución educativa.
Tutoring_Sessions: Número de sesiones de tutoría adicional que recibe el estudiante.	**Gender:** Género con el que se identifica el estudiante.
Family_Income: Ingreso económico total del núcleo familiar.	**School_Type:** Categoría o tipo de institución educativa

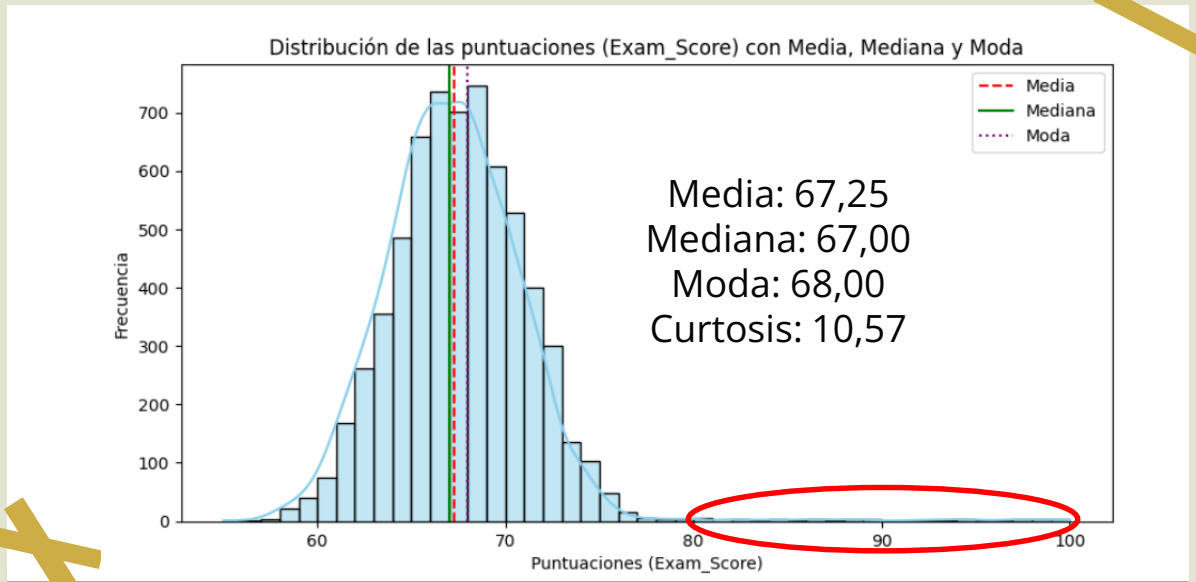
Numéricos

- ✓ **Hours_Studied:** Número de horas que el estudiante dedica al estudio por semana.
- ✓ **Attendance:** Porcentaje de asistencia del estudiante a las clases programadas.
- ✓ **Previous_Scores:** Calificaciones obtenidas por el estudiante en evaluaciones anteriores.
- ✓ **Exam_Score:** Puntuación obtenida por el estudiante en la evaluación actual.

Limpieza de datos

Duplicados	Nulos	Cero	Anómalos
0 valores	Teacher_Quality (78) Parental_Education_Level (90) Distance_from_Home (67)	Tutoring_Sessions (1471) Physical_Activity (44)	1 valor
	Se reemplazó *Teacher Quality* por la moda y *Parental Education Level* y *Distance from Home* se decidió eliminar las instancias.	Son consistentes respectivamente y no corresponden a valores nulos o mal tomados en el dataset.	La variable objetivo Exam_Score presenta una puntuación de 101y se reemplazó por la puntuación de 100.

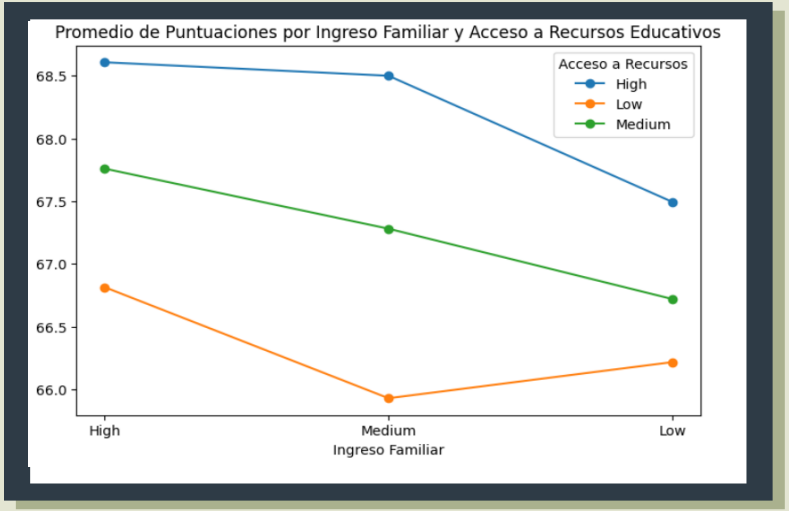
Visualización Exam_Score



Distribución

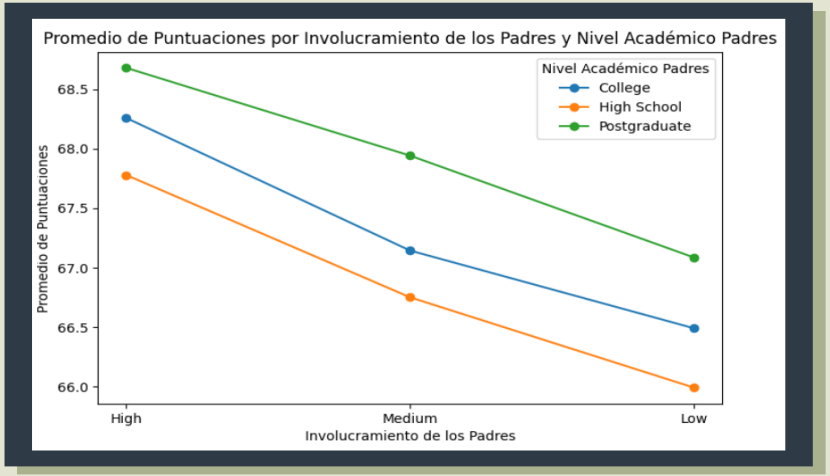
- Simétrica
- Leptocúrtica
- Valores atípicos para puntuaciones mayores a 80

Análisis Multivariado Exam_Score



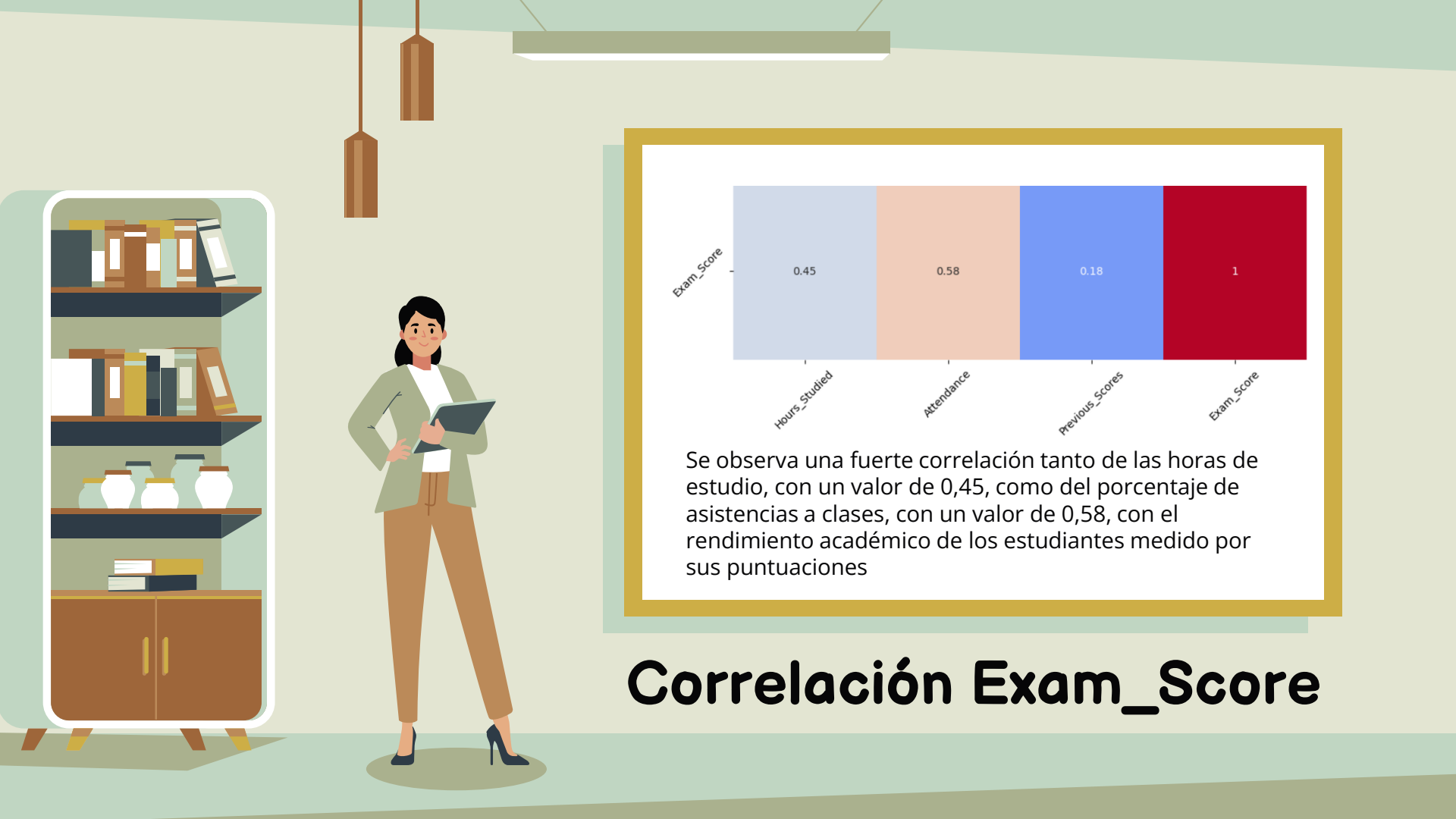
Income_Family y Access_to_Resources

Tendencia descendente general al disminuir el ingreso económico del núcleo familiar y a mayor disponibilidad mejores resultados académicos.



Parental_Education_Level y Parental_Involvement

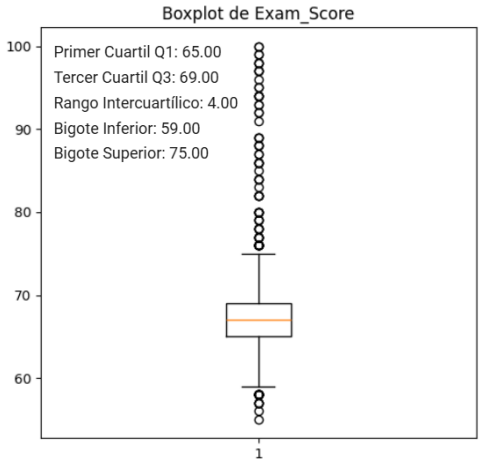
Tendencia descendente en las puntuaciones a medida que disminuye la participación de los padres y un mayor nivel educativo de los padres se correlaciona con mejores puntuaciones.

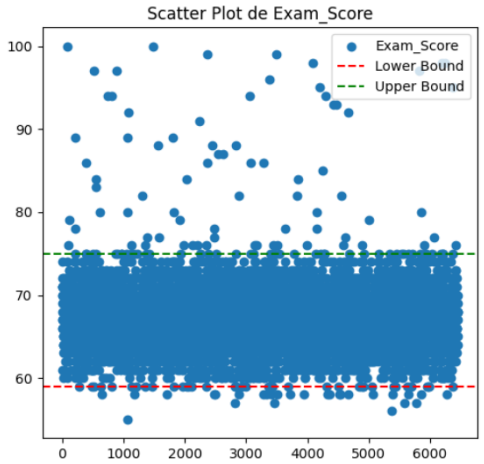


Se observa una fuerte correlación tanto de las horas de estudio, con un valor de 0,45, como del porcentaje de asistencias a clases, con un valor de 0,58, con el rendimiento académico de los estudiantes medido por sus puntuaciones

Correlación Exam_Score

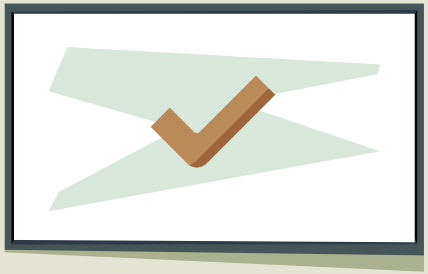
Outliers: 104

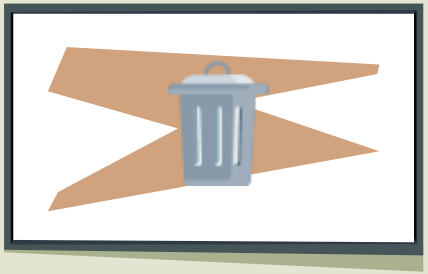




Estos **outliers** contienen información valiosa sobre el rendimiento académico de ciertos estudiantes, en su mayoría, aquellos que han obtenido puntuaciones superiores a 75 puntos.

Enfoque Dual



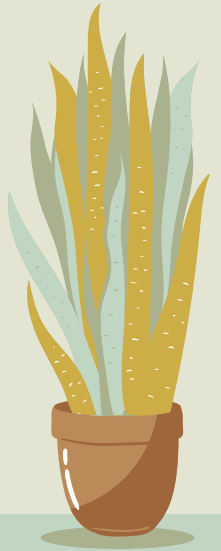


Modelado 1
Incluyendo Outliers

Su inclusión enriquece la comprensión de los patrones de rendimiento académico y contribuye a la robustez del modelo predictivo.

Modelado 2
Eliminando Outliers

Se eliminarán del dataset original los 104 outliers (1,61% del total de instancias) que se encuentran por debajo del Bigote Inferior 59 y por encima del Bigote Superior 75.



Incluyendo Outliers

Todas las Variables

Regresión Lineal

R^2 = 0,5911
RMSE = 2,5112
MSE = 6,3064
MAE = 0,5659

XGBoost Regressor

R^2 = 0,5521
RMSE = 2,628
MSE = 6,9066
MAE = 0,9085

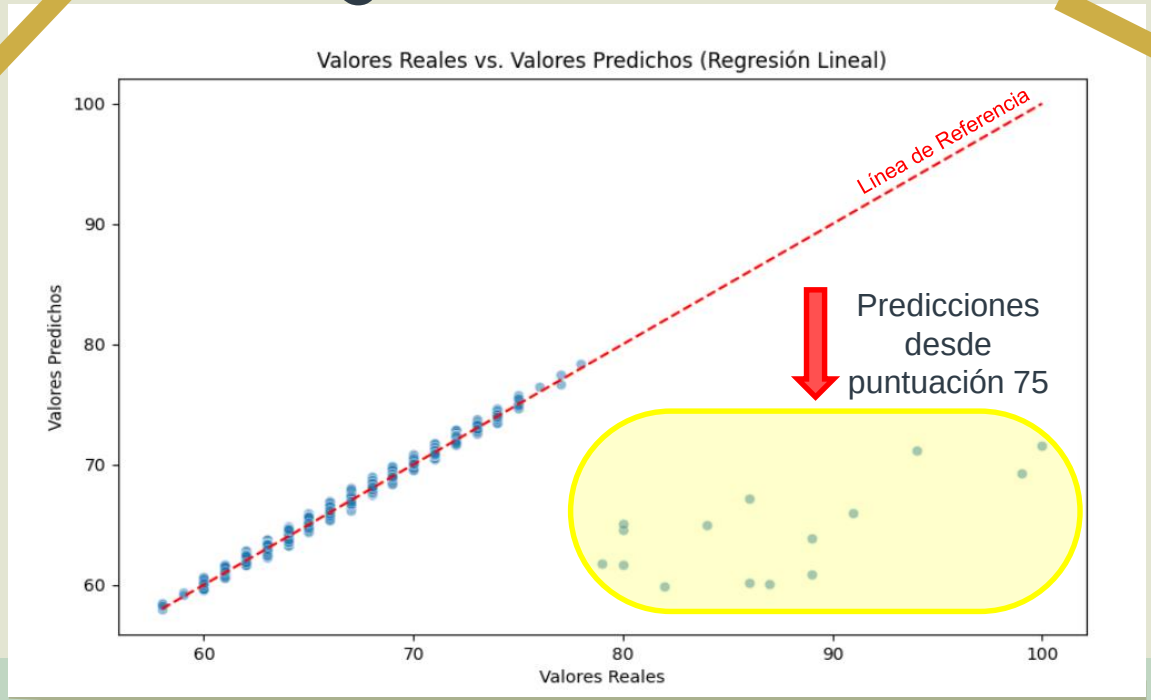
Bosque Aleatorio

R^2 = 0,5352
RMSE = 2,6773
MSE = 7,1679
MAE = 1,1407

Regularización Lasso (L1)

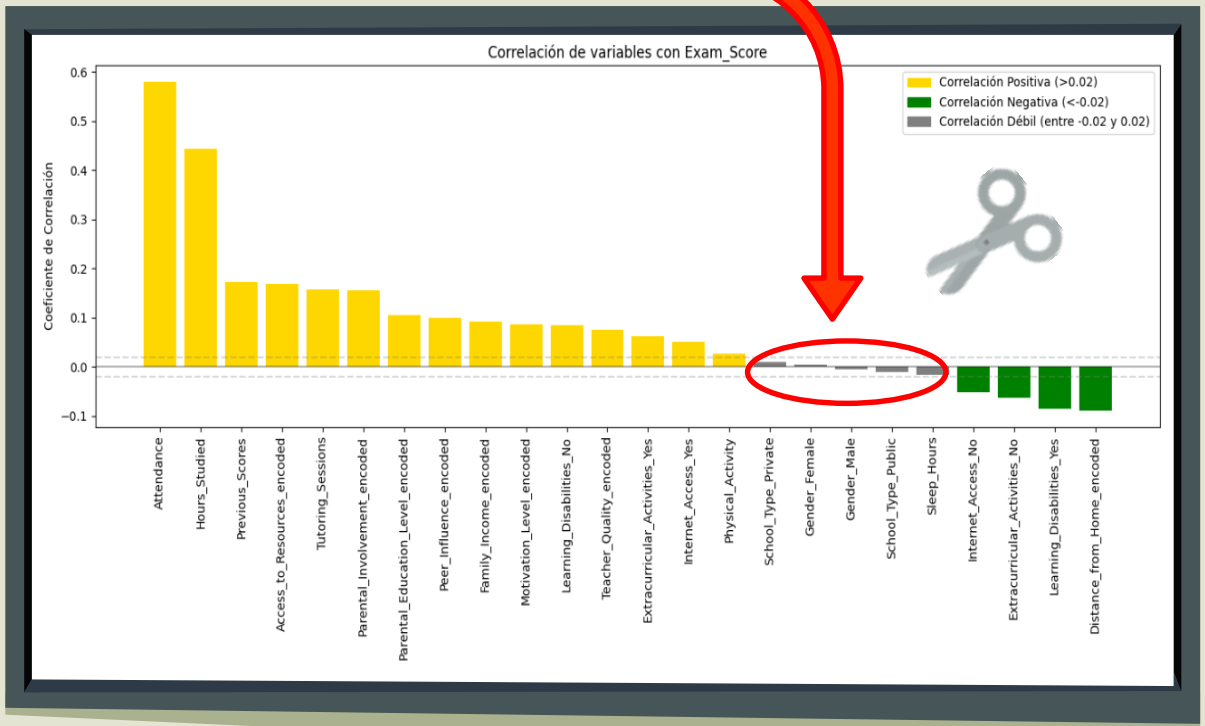
R^2 = 0, 5973
RMSE = 2,4919
MSE = 6,2096
MAE = 0,6033

Regresión Lineal



A partir de la puntuación de 75 en los valores reales, los errores en los valores predichos se acentúan considerablemente. Dicha puntuación coincide con el bigote superior.

Factores Eliminados



Modelado en Dataset reducido

Incluyendo Outliers

Validación Cruzada k-fold

Se buscó producir un conjunto de pruebas como métricas de puntuación usando todo el conjunto de entrenamiento.

MSE Promedio: 4,3271
RMSE Promedio: 2,0436
MAE Promedio: 0,5523
R^2 Promedio: 0,7194

Hiperparámetros

Se utilizó la técnica GridSearchCV sobre el modelo de regresión lineal con regularización Lasso (L1) para encontrar los mejores.

Mejor R^2: 0,7230
Mejores Hiperparámetros:
- '**alpha**': 0,0001
- '**max_iter**': 1000
- '**tol**': 1e-05

2° Enfoque

Eliminando Outliers

El modelo queda más acotado a predecir **Exam_Score** en puntuaciones entre 59 y 75.

Eliminando Outliers

Modelado en Dataset reducido

Regresión Lineal

RMSE = 0,3150
MSE = 0,0993
MAE = 0,2658

$R^2 = 0,9809$

Regularización Lasso (L1)

RMSE = 0,4480
MSE = 0,2007
MAE = 0,3591

$R^2 = 0,9906$

$R^2 = 0,9693$

$R^2 = 0,8919$

XGBoost Regressor

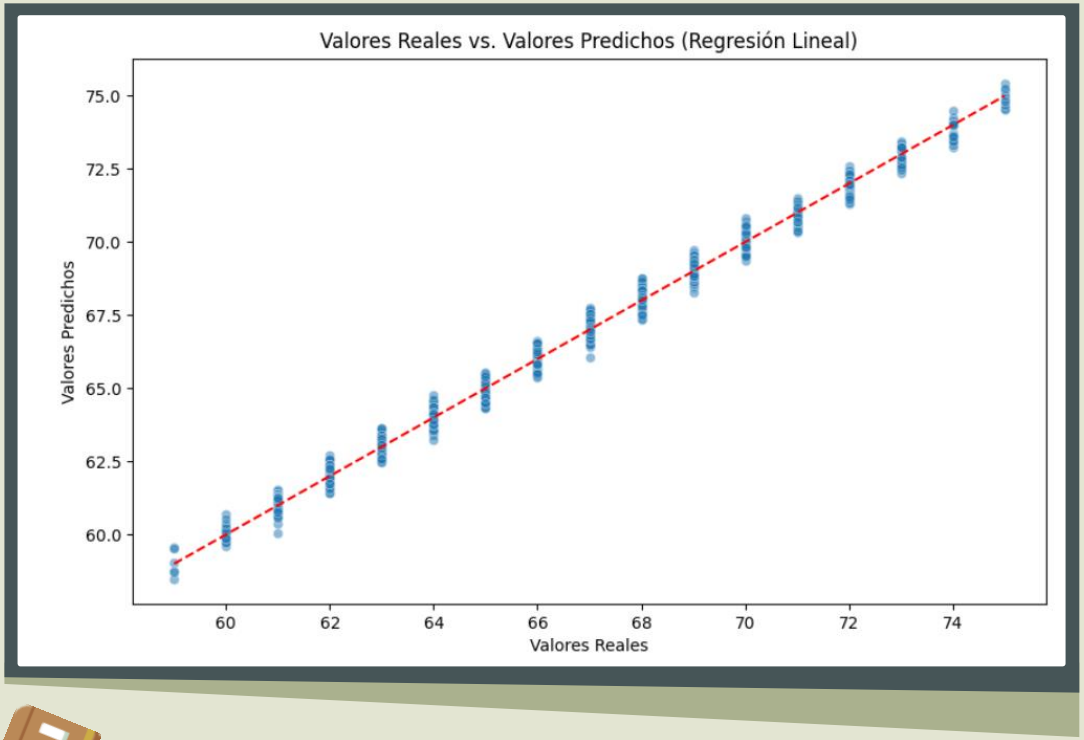
RMSE = 0,5680
MSE = 0,3226
MAE = 0,4474

Bosque Aleatorio

RMSE = 0,5680
MSE = 1,0669
MAE = 0,8485

Regresión Lineal

✓ Dataset Reducido
✓ Sin Outliers



Observación 1

Mayor ajuste a la línea de referencia

Observación 2

Modelo final preciso para puntuaciones entre 59-75

Modelo predictivo con 99,06% de precisión

para puntuaciones entre 59-75.

Impacto significativo: horas de estudio, asistencia, involucración parental, acceso a recursos

Impacto negativo: falta de acceso a internet, ausencia de actividades extracurriculares

Gracias!

Alguna pregunta?