Predicción del Rendimiento Estudiantil

Análisis y Modelado de Machine Learning del Student Exam Performance Dataset

Contexto del Dataset

Sobre el Contenido

Este conjunto de datos captura factores académicos, personales, familiares y relacionados con la escuela que influyen en el rendimiento de los exámenes de los estudiantes.
Combina de manera integral los hábitos de estudio, la asistencia, la motivación, los recursos y detalles socioeconómicos para analizar resultados reales.

Contexto de Aplicación

Diseñado específicamente para **el análisis de datos educativos y Machine Learning**. Ayuda a explorar cómo el estilo de vida, el entorno y el esfuerzo individual impactan la calificación.
Es sumamente útil para la predicción del rendimiento, el análisis de factores correlacionados y la investigación educativa moderna.

Dimensiones Analizadas

Hábitos Académicos

Métricas clave que reflejan la dedicación, incluyendo horas de estudio semanales, porcentaje de asistencia escolar y tutorías.

Factores Personales

Rutinas de bienestar que pueden influir en el desarrollo cognitivo, tales como la calidad del sueño y la frecuencia de actividad física.

Contexto y Entorno

Variables socioeconómicas como el nivel de ingresos familiares, el nivel de motivación, la calidad de los maestros y acceso a internet.

Resumen del Conjunto de Datos

6,607

Estudiantes Evaluados

Base de datos robusta libre de registros duplicados.

20

Atributos Analizados

Compuesto por 7 variables numéricas y 13 categóricas.

235

Valores Nulos Detectados

Localizados en 3 variables categóricas específicas.

Estructura Detallada (Data Schema)

Columna (Atributo)	No-Nulos	Dtype
0. Hours_Studied	6607	int64
1. Attendance	6607	int64
2. Parental_Involvement	6607	object
3. Access_to_Resources	6607	object
4. Extracurricular_Activities	6607	object
5. Sleep_Hours	6607	int64
6. Previous_Scores	6607	int64
7. Motivation_Level	6607	object
8. Internet_Access	6607	object
9. Tutoring_Sessions	6607	int64

Columna (Atributo)	No-Nulos	Dtype
10. Family_Income	6607	object
11. Teacher_Quality	6529	object
12. School_Type	6607	object
13. Peer_Influence	6607	object
14. Physical_Activity	6607	int64
15. Learning_Disabilities	6607	object
16. Parental_Education_Level	6517	object
17. Distance_from_Home	6540	object
18. Gender	6607	object
19. Exam_Score (Target)	**6607**	**int64**

Naranja: Atributos con valores nulos | **Azul:** Variable objetivo (Target)

Desglose Exacto de Valores Nulos

Variable Categórica	Cantidad de Nulos	Porcentaje del Total (%)
Nivel Educativo de los Padres (Parental_Education_Level)	90	1.36 %
Calidad del Maestro (Teacher_Quality)	78	1.18 %
Distancia a Casa (Distance_from_Home)	67	1.01 %

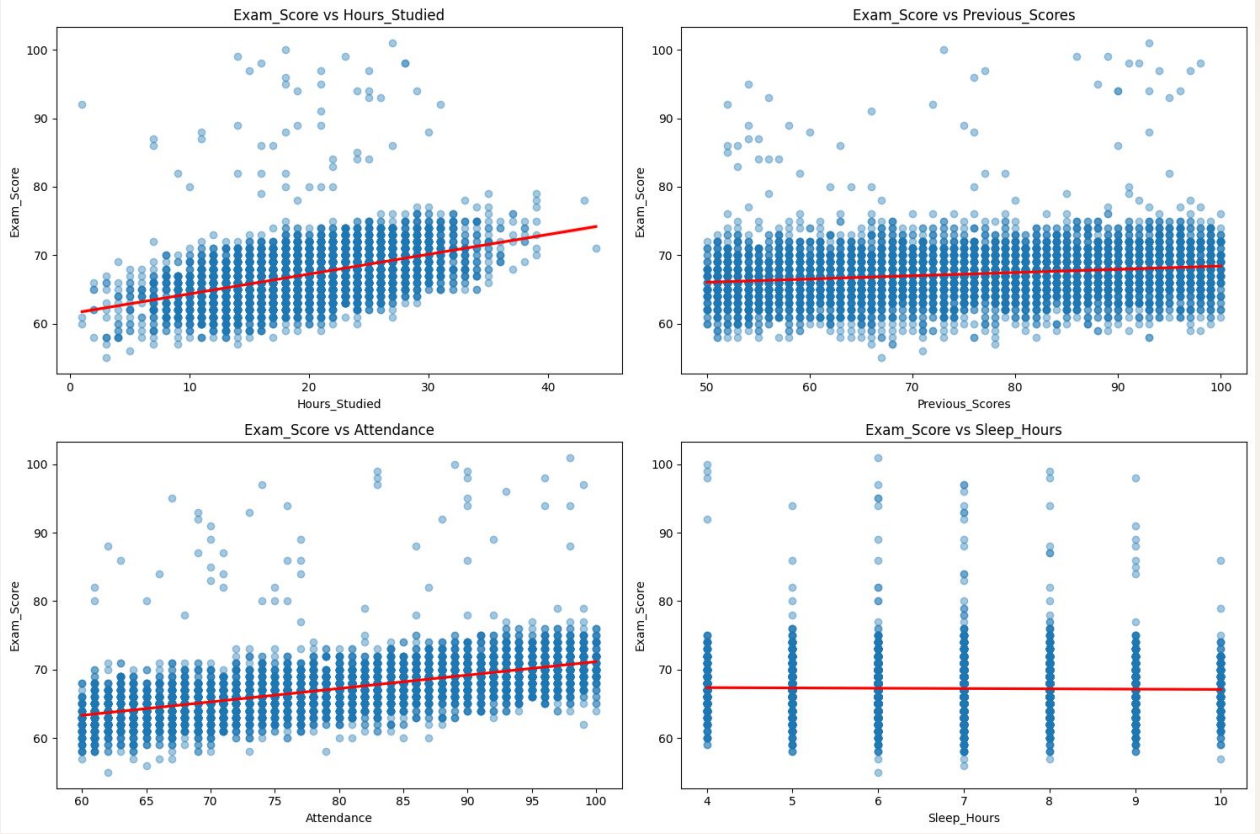
Impacto Mínimo: Al representar apenas el ~1% de los datos, la estrategia óptima para el preprocesamiento de estas variables categóricas es la **imputación por la moda** (el valor más frecuente), preservando la integridad del dataset sin distorsionar el modelo.

Análisis Exploratorio de Datos

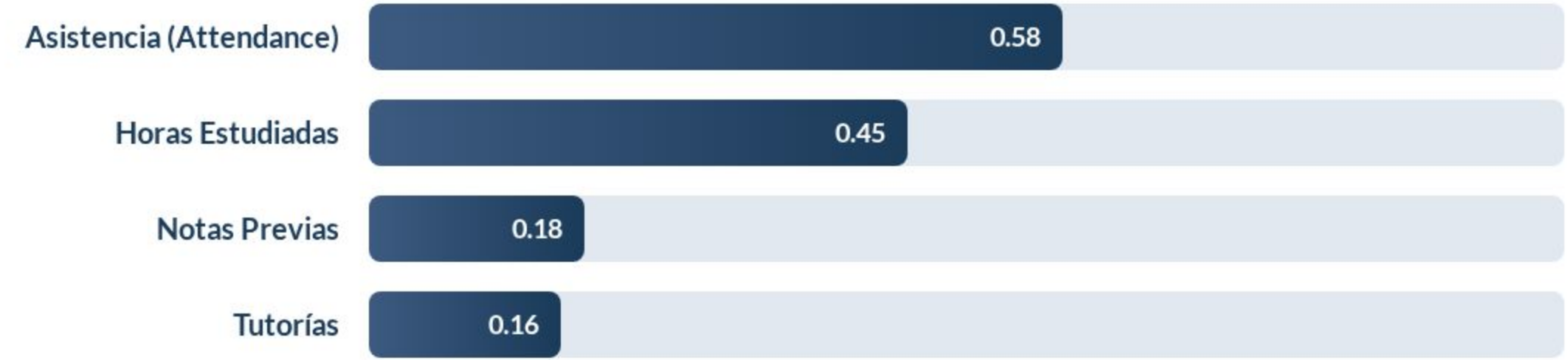
Descubriendo los patrones detrás del rendimiento

Comportamiento de las Variables

- **Horas de Estudio:** Presentan una distribución con sesgo izquierdo, con un pico predominante entre 20 y 22 horas semanales.
- **Asistencia Escolar:** Mayor concentración en el rango alto (90-100%), con casos mínimos extremos alcanzando el 60%.
- **Calificación Final:** Distribución cercana a la normal con promedio de 67-68 puntos, con límite lógico de 100.
- **Hábitos de Sueño:** Tendencia bimodal que se concentra fuertemente en estudiantes que duermen entre 6 y 8 horas diarias.



Impacto en la Nota Final (Correlación)



Hallazgo Principal: La asistencia a clases y el tiempo de estudio son los predictores positivos absolutos del éxito. Paradójicamente, factores físicos (sueño y actividad física) demostraron tener un impacto lineal casi nulo en la nota final.

Preprocesamiento y Modelado

Preparando los datos para la inferencia algorítmica

Pipeline de Transformación

Features Numéricas

Procesadas mediante un SimpleImputer con estrategia de mediana para proteger contra valores atípicos, seguido de estandarización empleando StandardScaler.

Features Ordinales

Tratadas con imputación por moda y transformadas usando OrdinalEncoder, mapeando respetando su jerarquía intrínseca (Low < Medium < High).

Features Nominales

Imputación del valor más frecuente y codificación utilizando OneHotEncoder (eliminando la primera categoría para evitar trampas de colinealidad).

Comparación de Modelos (Cross-Validation)

Algoritmo Evaluado	Mejor RMSE (CV) *	Hiperparámetros Óptimos Encontrados
Ridge Regression (Lineal)	2.049	alpha: 10.0
Gradient Boosting Regressor	2.149	learning_rate: 0.1, max_depth: 3
Random Forest Regressor	2.367	max_depth: None, min_samples_split: 5

* El Error Cuadrático Medio (RMSE) penaliza los errores grandes. Un valor más bajo indica mayor precisión.

Rendimiento del Modelo Ridge (Prueba)

1.800

RMSE (Margen de Error)

El modelo final tiene un error promedio de apenas 1.8 puntos sobre la calificación real del estudiante. Una inferencia sumamente precisa.

77.1%

R-Cuadrado (R^2)

El modelo logra explicar exitosamente más de tres cuartas partes de la varianza en las calificaciones de los alumnos.

Aplicación de Inferencia (GUI)

Para trasladar el modelo del laboratorio a un entorno práctico para docentes, desarrollamos un programa ejecutable con validación de entradas.

joblib

Librería encargada de la **carga y serialización**. Permite importar el pipeline pre-entrenado (model.pkl) instantáneamente.

pandas

Motor de estructuración. Transforma las variables ingresadas en la interfaz en un DataFrame ordenado, el formato exacto requerido.

tkinter / ttk

Construcción nativa de la **Interfaz Gráfica** interactiva. Validando límites y mostrando la proyección matemática al educador.