# Documentación del Proyecto: Análisis del Rendimiento Académico

**Estudiante: Camilo Farina Valencia**

## Descripción del Problema

El rendimiento académico de los estudiantes depende de múltiples factores, como el número de horas de estudio, el uso de herramientas educativas, la asistencia a clases, el nivel de estrés y las horas de sueño. Sin embargo, no siempre es evidente qué factores tienen un impacto significativo en el éxito académico.  
  
Este análisis busca responder preguntas clave como:  
- ¿Las **horas de estudio** influyen directamente en el rendimiento?  
- ¿La **participación en discusiones** mejora las calificaciones?  
- ¿El **uso de tecnología educativa** ayuda a obtener mejores resultados?  
- ¿Factores como el **estrés**, el **sueño** o el **tiempo** en **redes sociales** afectan el puntaje en exámenes?  
  
A través de este estudio, se espera identificar patrones en el aprendizaje de los estudiantes y con base a ello proponer estrategias para mejorar su desempeño académico.

## Contexto

El aprendizaje es un proceso complejo en el que influyen tanto factores internos (hábitos de estudio, motivación, estrés) como externos (acceso a recursos educativos, influencia de compañeros, tecnología).  
  
Con el auge del aprendizaje en línea y el uso de herramientas digitales, es fundamental comprender cómo estos factores afectan el rendimiento estudiantil. El análisis de datos en educación puede proporcionar información valiosa para mejorar metodologías de enseñanza y personalizar estrategias de aprendizaje.  
  
Este estudio se basa en datos recopilados de estudiantes de distintos contextos, con información detallada sobre sus hábitos, nivel de participación, rendimiento en exámenes y uso de tecnología educativa.

## Descripción del Dataset

* Nombre del dataset: `**student\_performance\_large\_dataset** `
* Número de registros: 10,000 estudiantes
* Número de columnas: 15

**Variables principales**:

**Datos personales:** -Student\_ID` → Identificador único del estudiante.  
 - `Age` → Edad del estudiante (18 a 29 años).  
 - `Gender` → Género (Male/Female).  
  
**Hábitos de estudio y aprendizaje:**  
 - `Study\_Hours\_per\_Week` → Horas de estudio semanales.  
 - `Preferred\_Learning\_Style` → Estilo de aprendizaje (Visual, Auditivo, Kinestésico, Lectura/Escritura).  
 - `Online\_Courses\_Completed` → Cursos en línea completados.  
 - `Participation\_in\_Discussions` → Participación en debates académicos (Sí/No).  
  
- - `Assignment\_Completion\_Rate (%)` → Porcentaje de tareas completadas.  
 - `Exam\_Score (%)` → Puntaje en exámenes (40-100%).  
 - `Final\_Grade` → Calificación final (A, B, C, D, F).  
  
**Asistencia y tecnología**:  
 - Attendance\_Rate (%)` → Porcentaje de asistencia a clases.  
 - `Use\_of\_Educational\_Tech` → Uso de tecnología educativa (Sí/No).  
  
**Factores externos:** - `Self\_Reported\_Stress\_Level` → Nivel de estrés reportado (Alto, Medio, Bajo).  
 - `Time\_Spent\_on\_Social\_Media (hours/week)` → Horas semanales en redes sociales.  
 - `Sleep\_Hours\_per\_Night` → Horas promedio de sueño por noche.  
  
**Objetivo**: Explorar la relación entre estas variables para entender qué factores influyen en el rendimiento académico.

## Proceso de Análisis

El análisis se llevará a cabo en varias etapas:  
  
1️. **Exploración de los datos**  
 - Identificar valores faltantes o inconsistencias.  
 - Verificar la distribución de las variables y eliminar valores atípicos.  
  
2️**. Limpieza y transformación de datos** - Manejo de datos nulos.  
 - Normalización de variables para el análisis.  
  
3️**. Análisis exploratorio** - Visualización de tendencias y correlaciones entre variables.  
 - Comparación de grupos según hábitos de estudio y rendimiento.  
  
5️. **Conclusiones y recomendaciones** - Identificación de patrones en el aprendizaje.  
 - Propuestas de estrategias para mejorar el rendimiento estudiantil.

Nota: Se adjunta la carga de la base de datos:

## EDA (exploratory data analysis)

Introducción

Este documento describe el proceso de Exploratory Data Analysis (EDA) y las transformaciones realizadas en el dataset para prepararlo para análisis y modelado.

**1. Carga y Exploración Inicial**

✔️ Se carga el dataset en `df\_transformedST`.

✔️ Se imprime la cantidad de filas y columnas.

✔️ Se obtiene una muestra aleatoria con `sample(1)`.

✔️ Se revisan valores nulos con `isnull().sum()`.

**2. Análisis de Valores Únicos en Columnas Categóricas**

✔️ Se identifican las columnas categóricas y se cuentan los valores únicos.

✔️ Se genera un gráfico de barras para visualizar las categorías únicas.

**3. Identificación de Duplicados**

✔️ Se cuentan los registros duplicados con `duplicated().sum()`.

✔️ Se genera un gráfico de barras comparando registros duplicados vs. únicos.

4**. Generación de Estadísticas Generales**

✔️ Se generan estadísticas para variables numéricas (`describe().T`).

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

✔️ Se grafican estadísticas para variables categóricas (`describe(include='object').T`).

5. Visualización de Distribución de Datos

✔️ Se generan histogramas para variables numéricas.

✔️ Se utilizan Boxplots para detectar outliers.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

6. Manejo de Valores Nulos

✔️ Se eliminan registros con valores nulos (`dropna()`).

7. Conversión de Variables Categóricas a Numéricas

✔️ Se convierten a valores numéricos las siguientes columnas:

- `Gender`

- `Preferred\_Learning\_Style`

- `Participation\_in\_Discussions`

- `Use\_of\_Educational\_Tech`

- `Final\_Grade`

✔️ Se usa `astype('category').cat.codes` para asignar un número a cada categoría.

✔️ Se elimina `Student\_ID` por no aportar al análisis.

8. Verificación de Duplicados (Post Transformación)

✔️ Se verifica nuevamente la existencia de registros duplicados.

9. Creación de Nuevas Categorías de Agrupamiento

✔️ Age\_Category: Teenager (<20 años), Young Adult (20-25 años), Adult (>25 años)

✔️ Sleep\_Category: Enough Sleep (≥8 horas), Moderate (5-8 horas), Low Sleep (<5 horas)

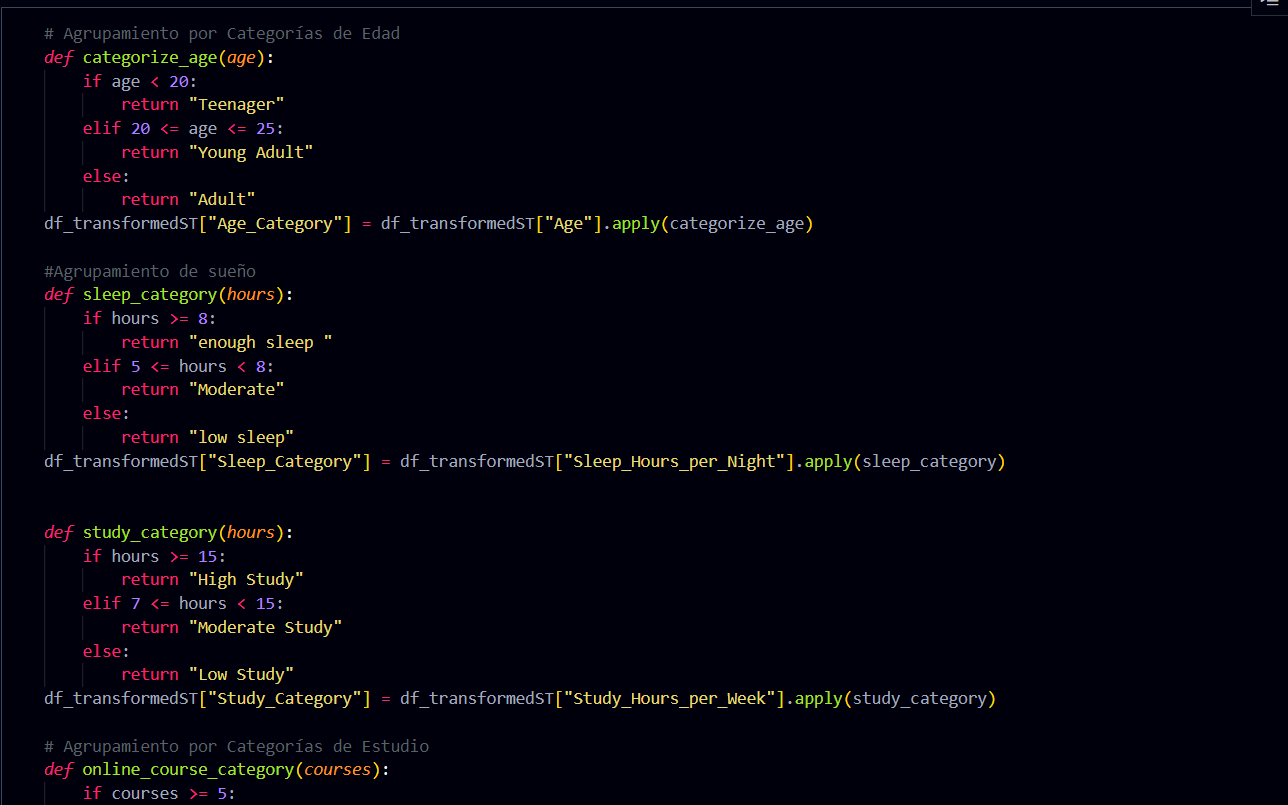
✔️ Study\_Category: High Study (≥15 horas), Moderate Study (7-15 horas), Low Study (<7 horas)

✔️ Online\_Course\_Level: Expert (≥5 cursos), Intermediate (2-4 cursos), Beginner (<2 cursos)

✔️ Attendance\_Category: High Attendance (≥85%), Moderate Attendance (60-85%), Low Attendance (<60%)

✔️ Stress\_Category: High Stress (Nivel = 2), Moderate Stress (Nivel = 1), Low Stress (Nivel = 0)

✔️ Social\_Media\_Usage: High Social Media Use (≥10 horas/semana), Medium Social Media Use (5-10 horas/semana), Low Social Media Use (<5 horas/semana)



Conclusión

El proceso de EDA y transformación permitió obtener un dataset limpio, categorizado y preparado para modelado.

Se eliminaron valores nulos y duplicados, se normalizaron datos categóricos y numéricos, y se generaron agrupaciones relevantes.

Este dataset está listo para análisis predictivo y modelado de Machine Learning.