# 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:

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

## 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)

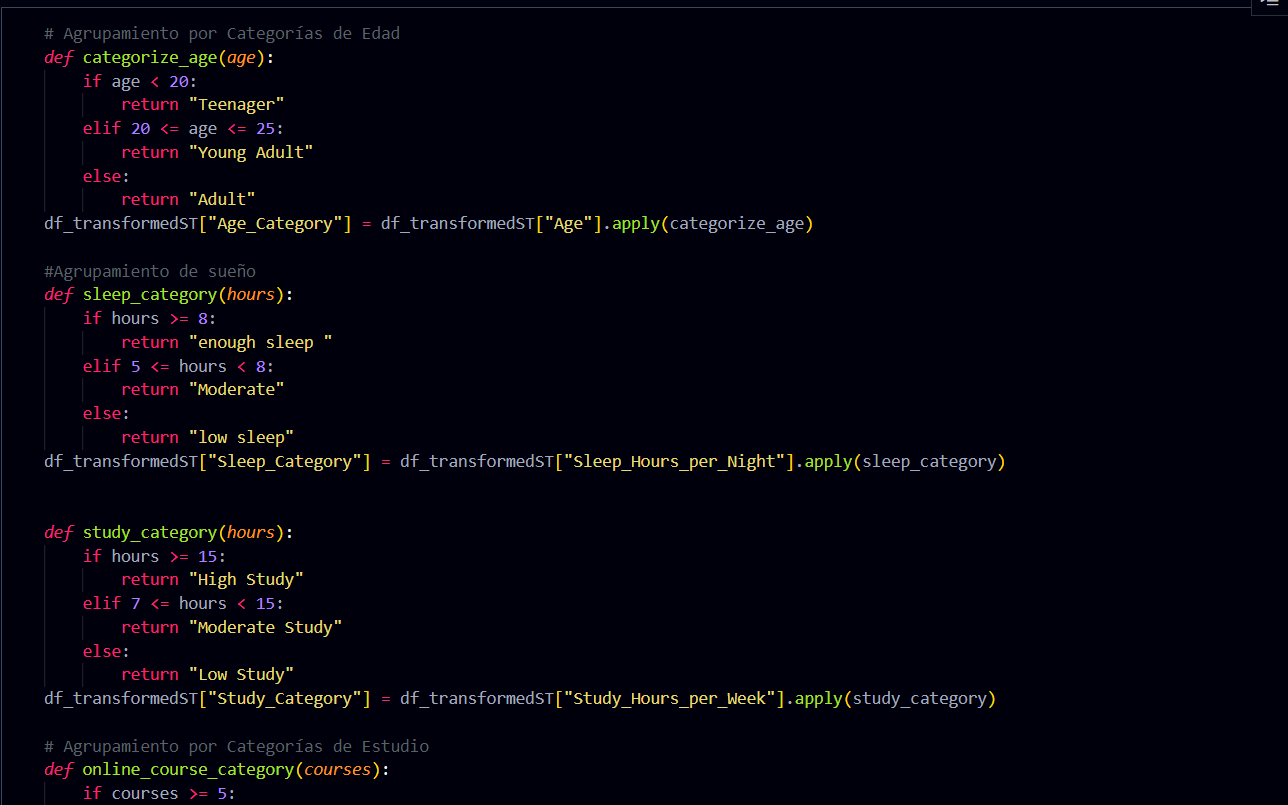
✔️ Study\_Category: High Study (≥15 horas), Moderate Study (7-15 horas), Low Study (<7 horas)

✔️ 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

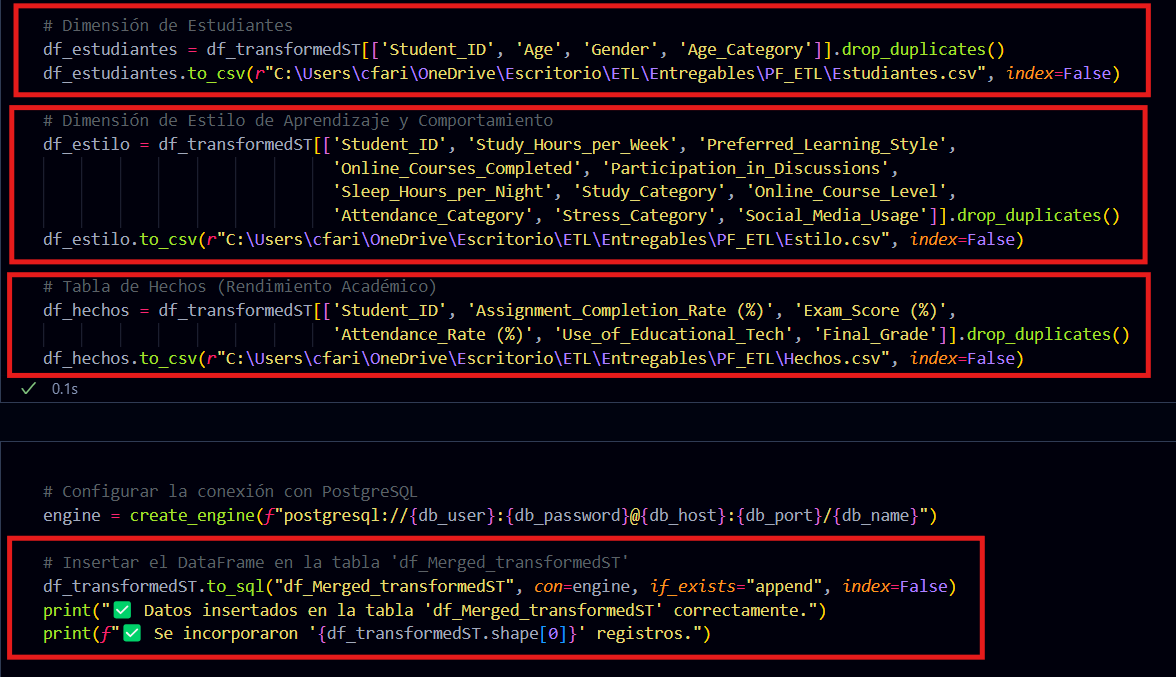
El proceso de EDA y transformación permitió obtener un dataset limpio, categorizado y preparado para modelado.

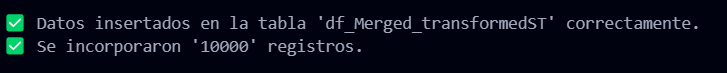
Se eliminaron valores nulos y duplicados, se normalizaron datos categóricos y numéricos, y se generaron agrupaciones relevantes.

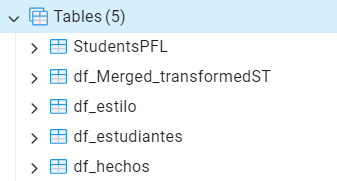
Este dataset está listo para análisis predictivo y modelado de Machine Learning.

## Proceso de Guardado y Generación del Dashboard

En primer lugar, se anexa muestra del Merged en donde se almacena en la nueva tabla, pero donde la información ya está ajustada para consumir.

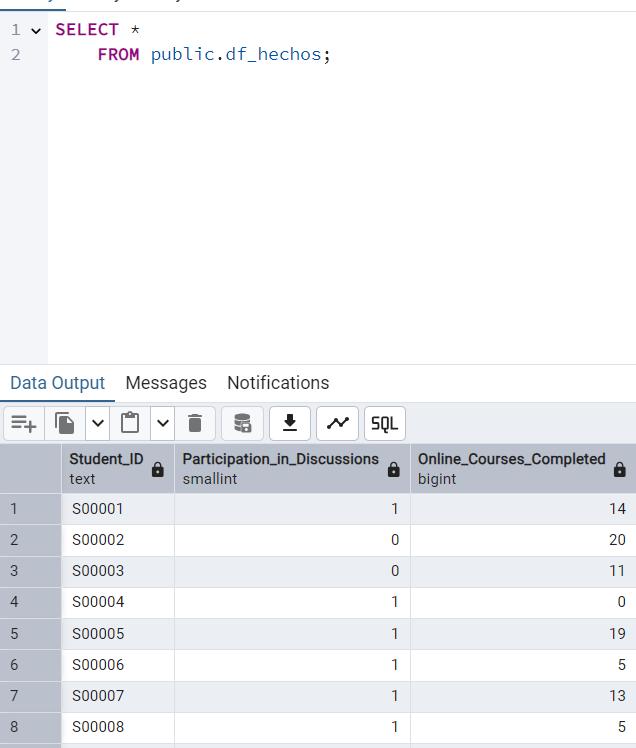






Tabla

El contenido generado por IA puede ser incorrecto.



Tabla

El contenido generado por IA puede ser incorrecto.

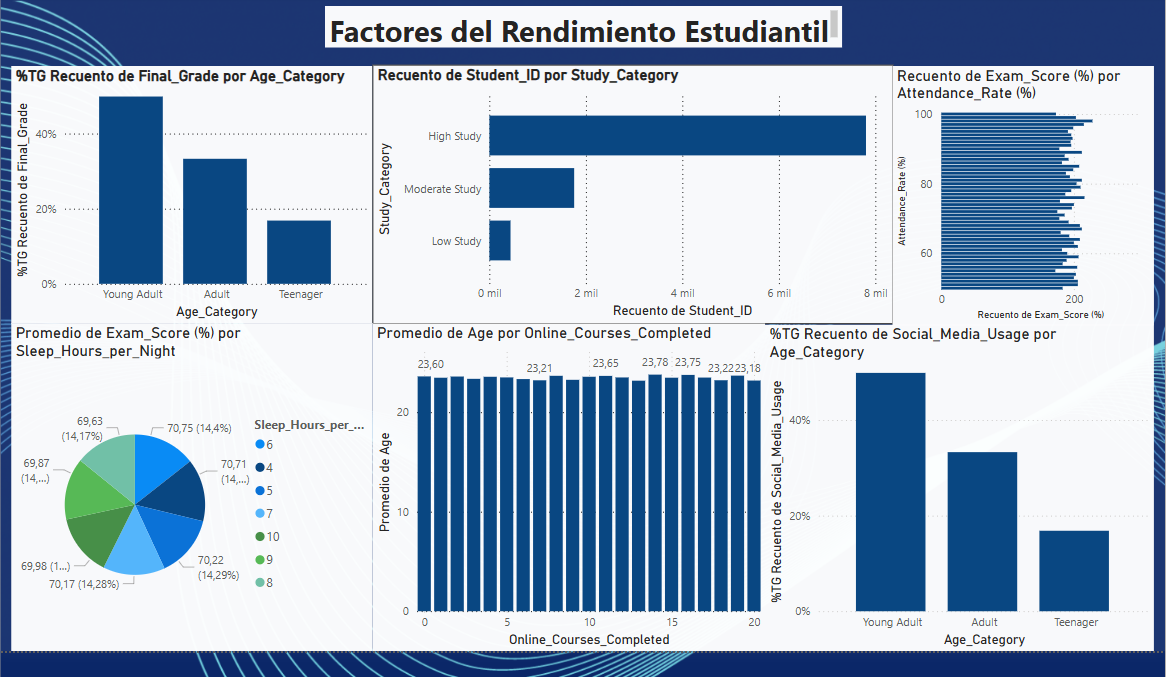
Finalmente, una vez se almacena en la base de datos, se va a consumir usando power bi, donde se hace todo el respectivo. En este apartado es donde se establecen las relaciones y creación del modelo dimensional para ir anexando los respectivos tipos de gráficos.

Establecemos las relaciones:

Interfaz de usuario gráfica, Aplicación, Word

El contenido generado por IA puede ser incorrecto.

**Dashboard:**



A partir de las diferentes gráficas presentadas en el tablero, se pueden extraer las siguientes **conclusiones** sobre los factores de rendimiento estudiantil:

1. **Final Grade por Categoría de Edad**
   * El **mayor recuento** de calificaciones finales corresponde a la categoría **Young Adult**, seguido de Adult y, en menor medida, Teenager.
   * Esto sugiere que la mayoría de los estudiantes evaluados pertenecen al grupo de jóvenes adultos o que estos están más representados en el dataset. Podría indicar que tienen más participación o se registran más sus calificaciones.
2. **Study\_Category (High, Moderate, Low)**
   * Se observa una **diferencia notable** en el recuento de estudiantes entre las categorías de estudio. Por ejemplo, si la barra de High Study es la más alta, se concluye que un gran número de estudiantes invierte muchas horas de estudio.
   * Esta variable es clave para analizar la relación entre la dedicación al estudio y las calificaciones finales.
3. **Exam\_Score vs Attendance\_Rate**
   * El gráfico relacionado con el **Exam\_Score** y la **Attendance\_Rate** muestra una tendencia donde la **mayor asistencia** podría correlacionarse con **mejores puntajes de examen**.
   * Esto coincide con la idea de que la asistencia regular es un fuerte predictor de rendimiento académico.
4. **Edad y Online\_Courses\_Completed**
   * Hay un gráfico que relaciona la **edad** o la categoría de edad con los **cursos en línea completados**. Si se observa que los adultos jóvenes completan más cursos en línea, podría indicar una mayor afinidad por el aprendizaje autodidacta en ese grupo.
5. **Sleep\_Hours\_per\_Night**
   * Existe un gráfico que compara las horas de sueño con la categoría de edad o el rendimiento. Si, por ejemplo, se ve que los estudiantes con más horas de sueño tienen un mejor desempeño, sugeriría la importancia del descanso en el rendimiento académico.
   * También puede mostrar que los adolescentes duermen más, pero no necesariamente obtienen las mejores calificaciones, lo que abriría nuevas hipótesis sobre la calidad del sueño o la gestión del tiempo.
6. **Social\_Media\_Usage por Age\_Category**
   * El gráfico que compara el uso de redes sociales con la edad indica cómo diferentes grupos etarios utilizan las redes sociales. Por ejemplo, si **Young Adult** presenta mayor uso de redes sociales, se podría explorar si esto afecta o no sus calificaciones.

**Conclusión General**

* **Asistencia y horas de estudio**: Se destacan como factores que influyen positivamente en los puntajes de exámenes y las calificaciones finales.
* **Edad y categoría de estudio**: Los adultos jóvenes tienen mayor representación y, en muchos casos, aparecen con mayor dedicación a cursos en línea o mayor uso de redes sociales.
* **Descanso y uso de redes**: Son variables que podrían moderar o impactar el rendimiento. Un alto uso de redes sociales en ciertos grupos de edad podría o no tener efectos negativos, dependiendo de otros factores como horas de estudio o asistencia.

En conjunto, los gráficos sugieren que **la constancia en el estudio (High Study), una alta tasa de asistencia y un equilibrio en hábitos de descanso** podrían correlacionarse con mejores resultados académicos. De todas formas, siempre es recomendable profundizar con correlaciones específicas y analizar cada variable en mayor detalle para confirmar las hipótesis que surgen de estas visualizaciones. (Esto para futuros proyectos)