Informe Integral del Proyecto: Análisis del Rendimiento Estudiantil

1. Introducción

Este informe detalla todo el proceso de análisis de un conjunto de datos estudiantiles, con el fin de comprender cómo distintos hábitos diarios, condiciones de vida y salud mental influyen sobre el rendimiento académico. El análisis incluyó desde la exploración inicial del dataset (EDA inicial), transformación y enriquecimiento de los datos (ETL), hasta un análisis exploratorio final (EDA final) y modelos de machine learning para segmentar y predecir comportamientos estudiantiles.

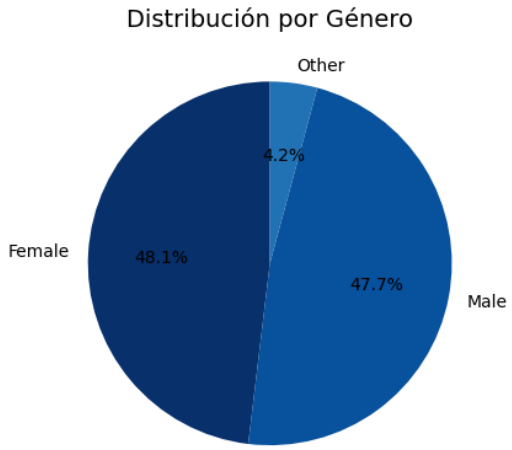
2. Análisis Exploratorio de Datos Inicial (EDA)

El EDA inicial nos permitió comprender la estructura y distribución de los datos originales. Incluyó pasos como:

- Inspección de tipos de variables y valores nulos.
- Distribución de variables clave: género, horas de estudio, sueño, uso de redes sociales y puntaje en exámenes.
- Identificación de outliers mediante boxplots.
- Correlaciones iniciales entre variables numéricas mediante mapas de calor.

1. Distribución por Género

Resultado: Se observa una proporción bastante equilibrada entre géneros, con una ligera mayoría de mujeres (aproximadamente 48.1%).

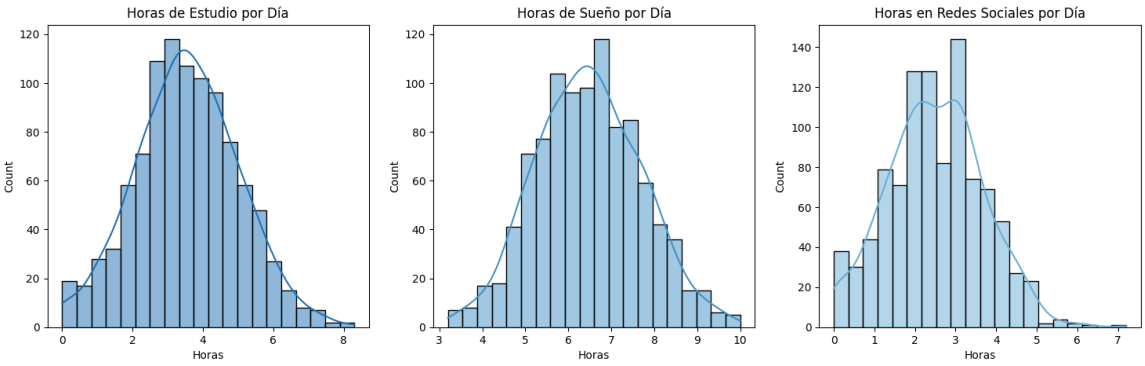


2. Histogramas de Hábitos Diarios

Horas de Estudio: Distribución simétrica centrada en torno a las 3.5 horas diarias. Se identifican algunos valores extremos, desde estudiantes que no estudian (0 h) hasta aquellos que superan las 8 h diarias.

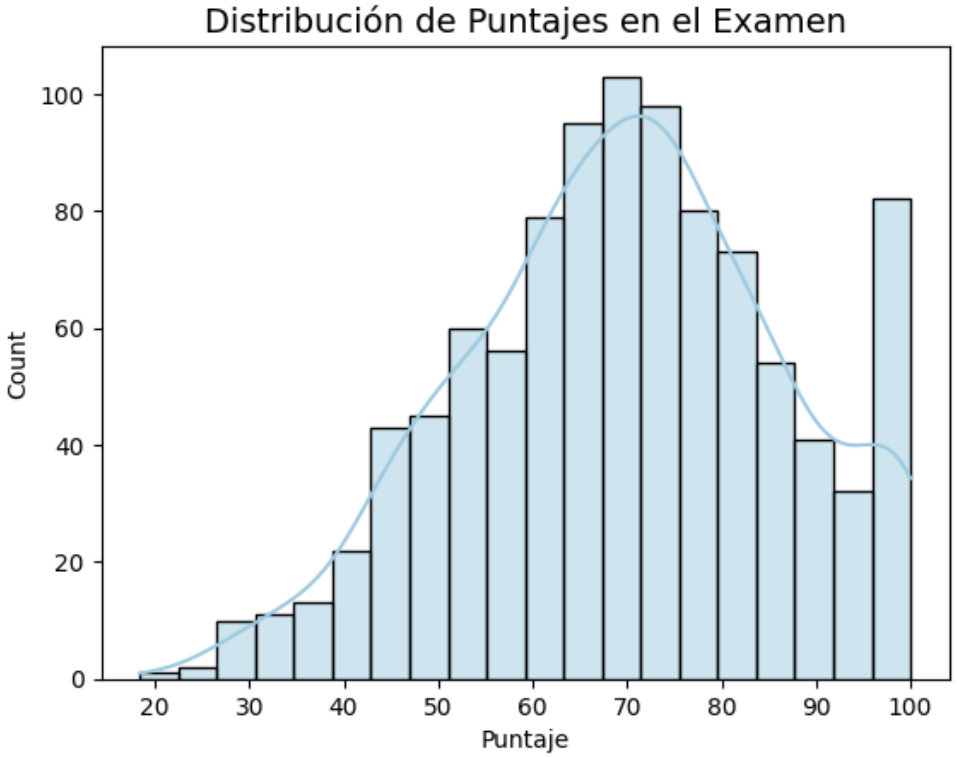
Horas de Sueño: La mayoría de los estudiantes duerme entre 6 y 7 horas. Se detectan posibles casos de insomnio con menos de 5 horas de sueño.

Redes Sociales: La mayoría invierte entre 2 y 3 horas por día en redes sociales. Pocos estudiantes presentan valores extremos por encima de 6 horas.



3. Distribución de Puntajes en el Examen

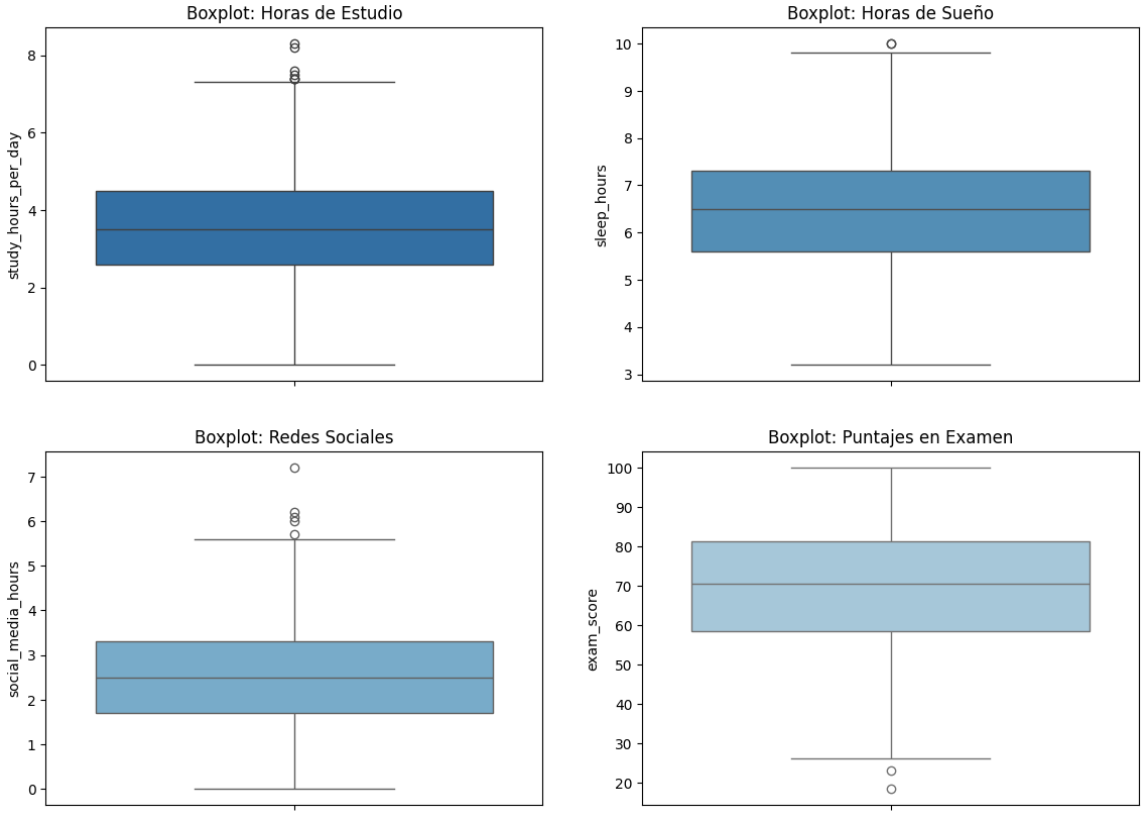
Observación: La concentración principal de los puntajes está entre los 60 y 90 puntos. Hay algunos estudiantes con puntajes bajos (menores a 40) y pocos alcanzan el valor máximo (100).



4. Boxplots: Detección de Outliers

Se identificaron valores atípicos en las siguientes variables:

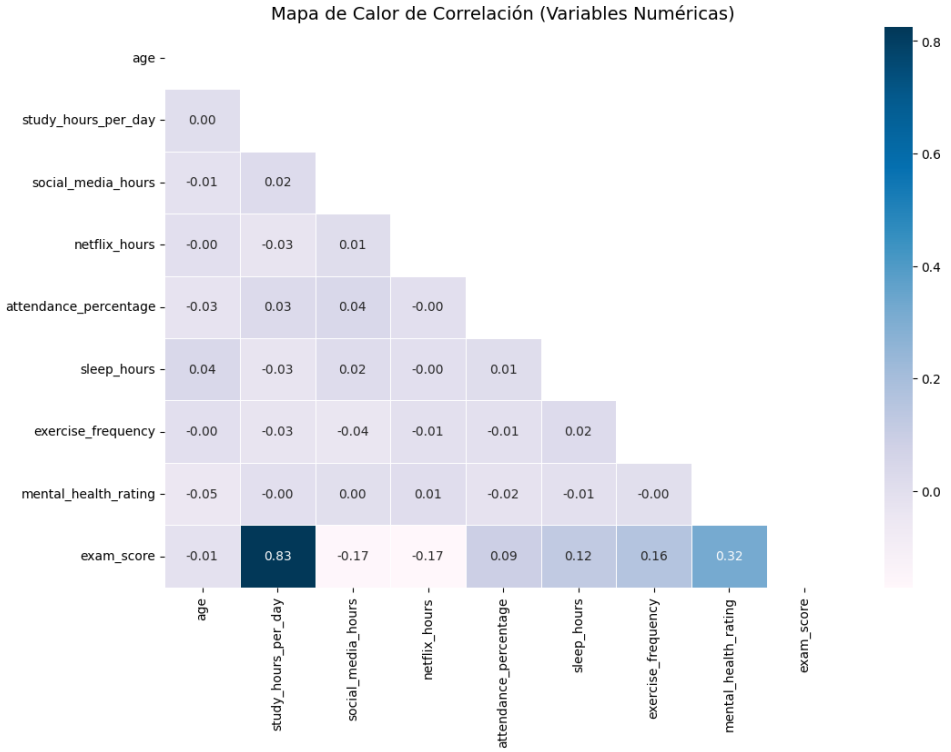
- **Estudio**: Estudiantes que reportan 0 horas diarias.
- **Sueño**: Casos con menos de 4 horas o más de 9 horas por día.
- **Redes Sociales**: Algunos estudiantes dedican más de 6 horas diarias.
- **Examen**: Presencia de puntajes muy bajos, incluso por debajo de 30.



5. Mapa de Calor de Correlaciones

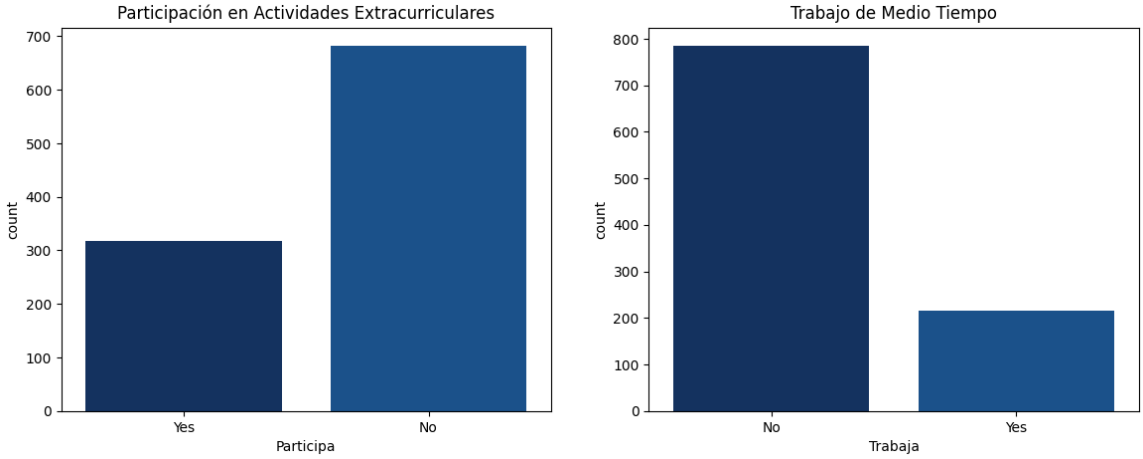
Relaciones destacadas entre variables:

- **Horas de Estudio y Puntaje de Examen**: Correlación positiva moderada (alrededor de 0.54).
- **Horas de Sueño y Salud Mental**: Relación leve y positiva.
- **Horas en Redes Sociales y Puntaje de Examen**: Correlación leve y negativa.



6. Participación Extracurricular y Trabajo de Medio Tiempo

La mayoría de los estudiantes no participa en actividades extracurriculares ni tiene un trabajo de medio tiempo. Esto sugiere que el tiempo diario está destinado principalmente al estudio, descanso y ocio.



3. Proceso ETL (Extracción, Transformación y Carga)

Se aplicó una transformación exhaustiva al dataset para preparar los datos para el análisis avanzado:

• Codificación de variables categóricas: género, participación extracurricular, trabajo, dieta, calidad de internet.
• Creación de nuevas variables:
- `tiempo_libre`: 24h menos todas las actividades medidas.
- `ratio_estudio_sueno`: proporción entre horas de estudio y sueño.
- `rendimiento_academico`: categorización de puntaje en bajo (0-50), medio (50-75), alto (75-100).
• Limpieza general, manejo de valores faltantes, y exportación del dataset limpio.

4. EDA Final

Con el dataset limpio y enriquecido, realizamos un análisis visual y relacional más profundo:
• Se exploraron las distribuciones de nuevas variables.
• Se evaluaron correlaciones más precisas, mostrando fuerte asociación entre horas de estudio y puntaje.
• Se agregaron variables derivadas para entender el equilibrio entre vida académica y personal.

1. Distribución de Rendimiento Académico

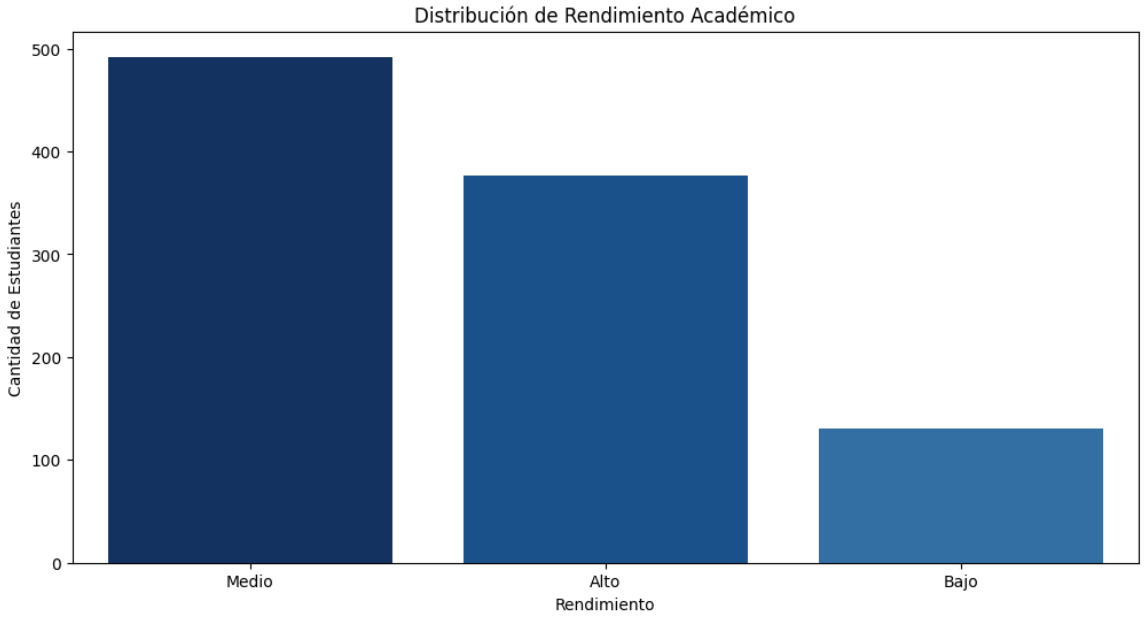
¿Qué muestra?
Se categorizó a los estudiantes en tres grupos: Alto, Medio y Bajo rendimiento académico.

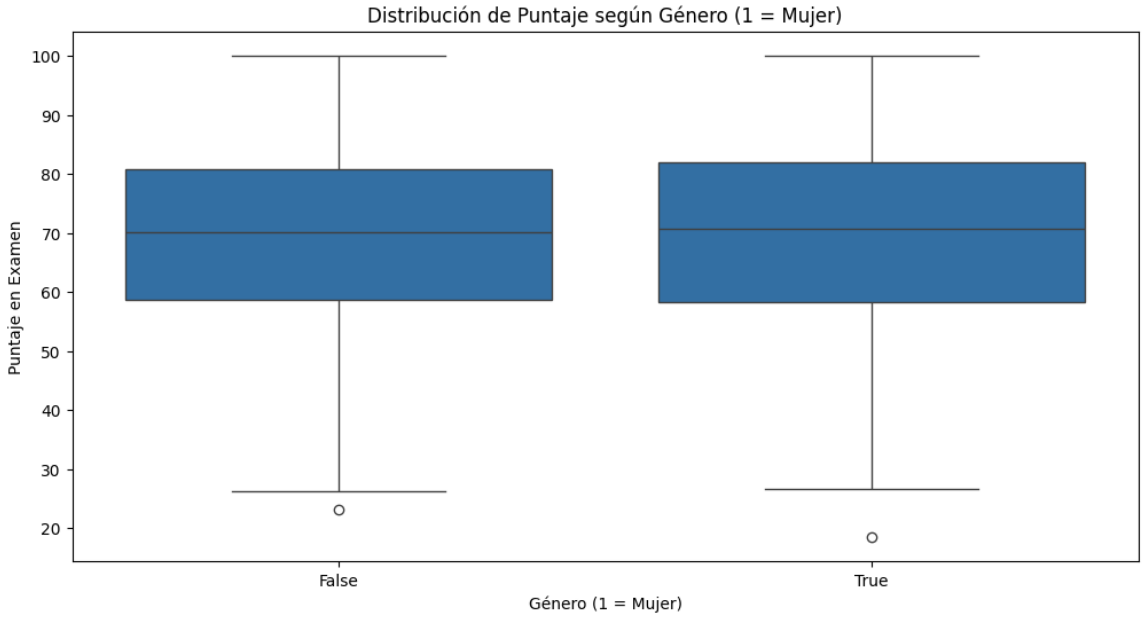
Observaciones:

- La mayoría de los estudiantes pertenece al grupo de rendimiento Medio (aproximadamente 50%).
- Le siguen aquellos con rendimiento Alto (~38%).
- Solo una minoría presenta un rendimiento Bajo (~13%).

Interpretación:

- La distribución es relativamente normal.
- Existe una oportunidad clara de intervención sobre el grupo con bajo rendimiento.
- Una proporción importante de estudiantes se mantiene por encima del nivel mínimo aprobado, lo que indica que los hábitos promedio del conjunto son relativamente adecuados.





2. Participación Extracurricular vs Puntaje Académico (Violin Plot)

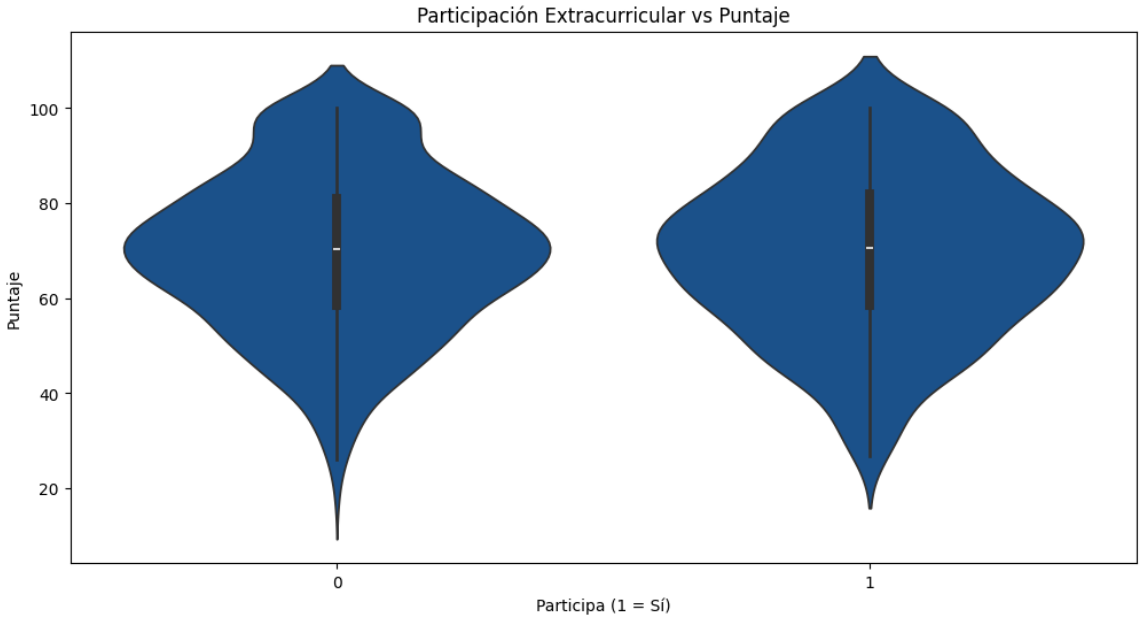
¿Qué muestra?
Comparación de la distribución de puntajes (`exam_score`) entre estudiantes que participan o no en actividades extracurriculares.

Observaciones:

- Ambos grupos muestran una amplia distribución de puntajes.
- Sin embargo, la mediana y el rango superior son ligeramente mayores en quienes participan en actividades extracurriculares.

Interpretación:

- El involucramiento extracurricular podría estar relacionado con un mejor rendimiento académico.
- La relación no es estricta, pero sí sugerente de un posible efecto positivo.



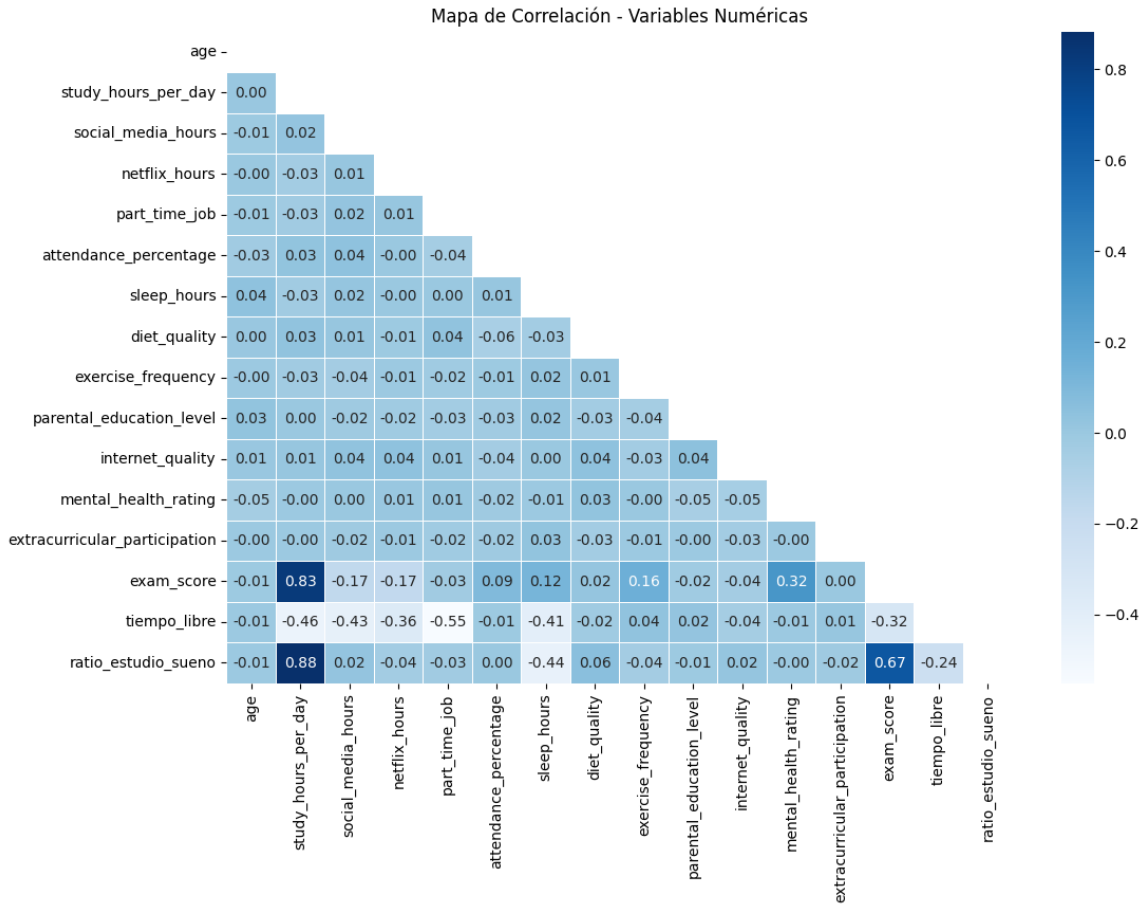
---

3. Mapa de Calor de Correlaciones Numéricas

¿Qué muestra?
Visualiza las correlaciones entre todas las variables numéricas del dataset. La intensidad y el color indican la dirección y fuerza de la relación.

Hallazgos clave:

- **Horas de Estudio vs Puntaje**: Correlación muy fuerte (0.83). Indica una relación directa significativa.
- **Ratio Estudio/Sueño vs Puntaje**: Correlación fuerte (0.67). Mantener un buen equilibrio entre estudio y descanso es altamente relevante.
- **Salud Mental vs Puntaje**: Correlación positiva moderada (0.32). Un mejor estado de bienestar mental tiende a asociarse con mayor rendimiento.
- **Tiempo Libre vs Puntaje**: Correlación negativa moderada (-0.32). Mayor cantidad de tiempo libre parece asociarse con menor rendimiento, posiblemente por menor dedicación.

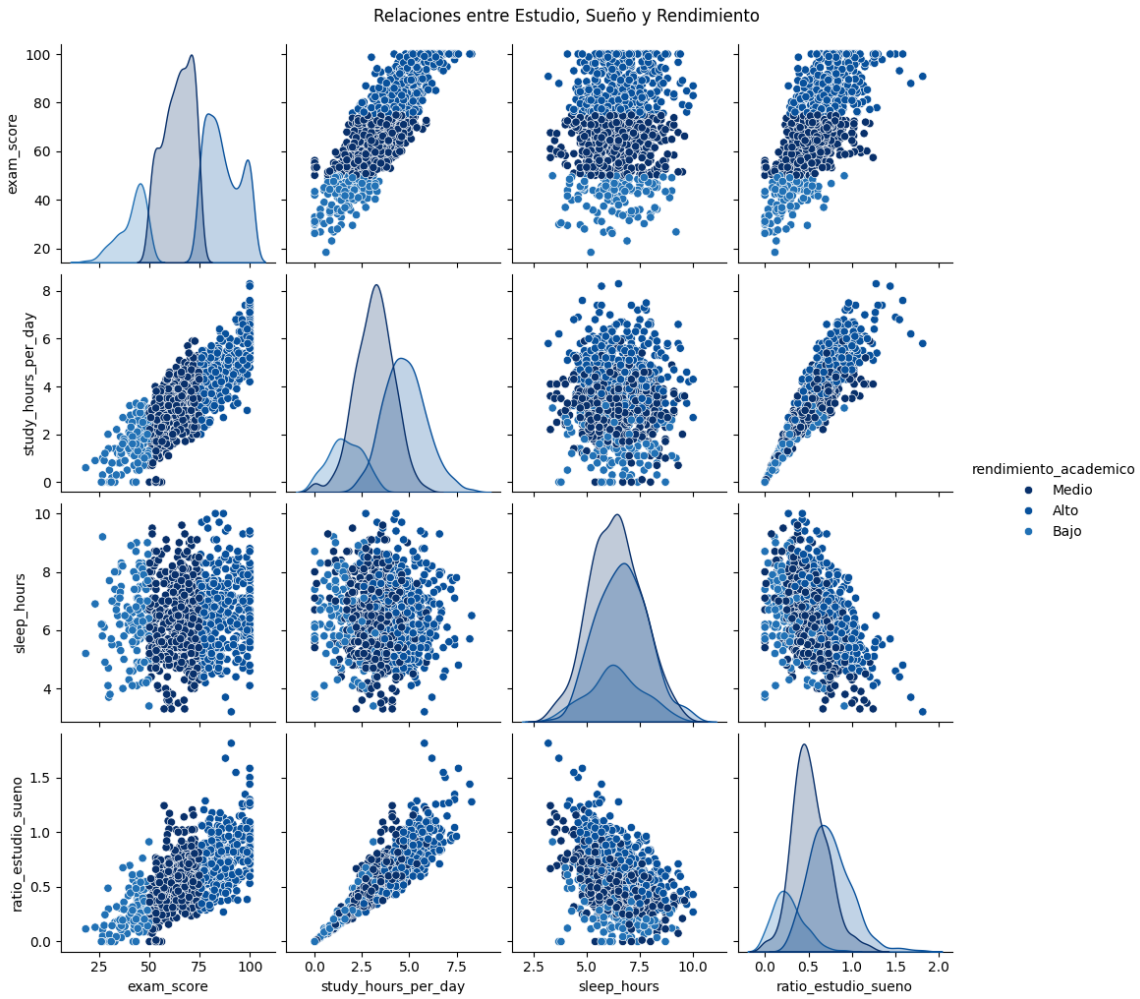


4. Pairplot de Estudio, Sueño y Rendimiento

¿Qué muestra?
Relaciones entre `exam_score`, `study_hours_per_day`, `sleep_hours` y `ratio_estudio_sueno`, con color de puntos representando el rendimiento académico (Alto, Medio, Bajo).

Interpretación:

- Los estudiantes con más horas de estudio y un buen balance con las horas de sueño tienden a obtener mejores resultados.
- El rendimiento bajo se concentra en estudiantes que duermen mucho y estudian poco, o con un ratio de estudio/sueño muy bajo.
- No se observa una relación clara entre el sueño por sí solo y el puntaje: el equilibrio entre descanso y estudio es lo determinante.



5. Clustering: Perfiles Estudiantiles

Aplicamos K-Means con estandarización previa y reducción de dimensionalidad con PCA. El método del codo sugirió 3 clústers como óptimo. Estos representaron perfiles distintos de estudiantes según hábitos y rendimiento.

1. Clustering de Perfiles Estudiantiles

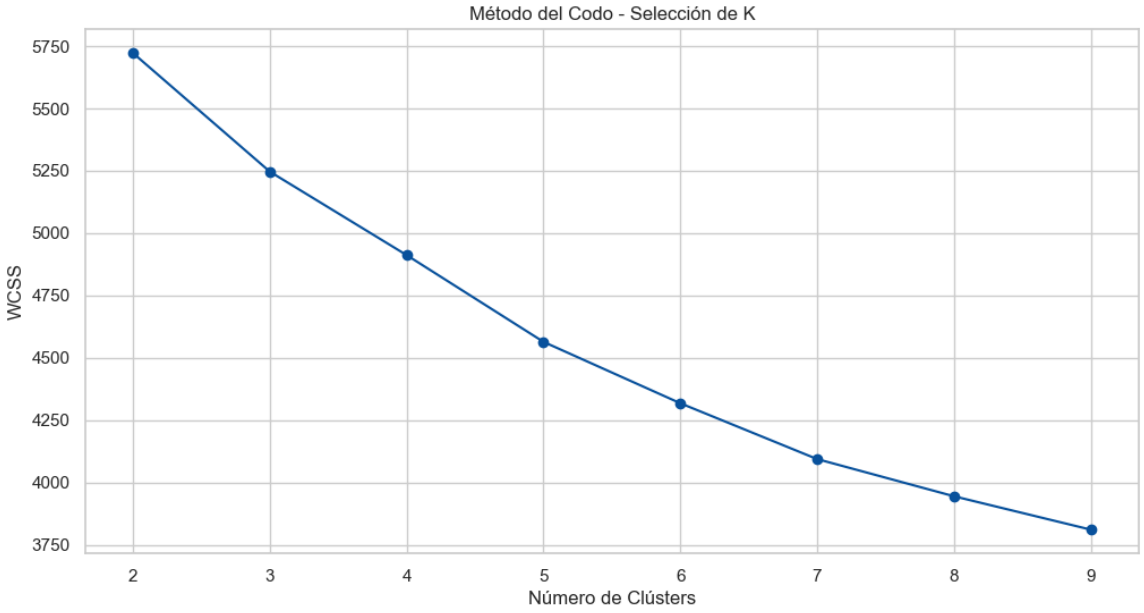
Objetivo:
Agrupar estudiantes en perfiles según similitudes en sus hábitos y estilo de vida.

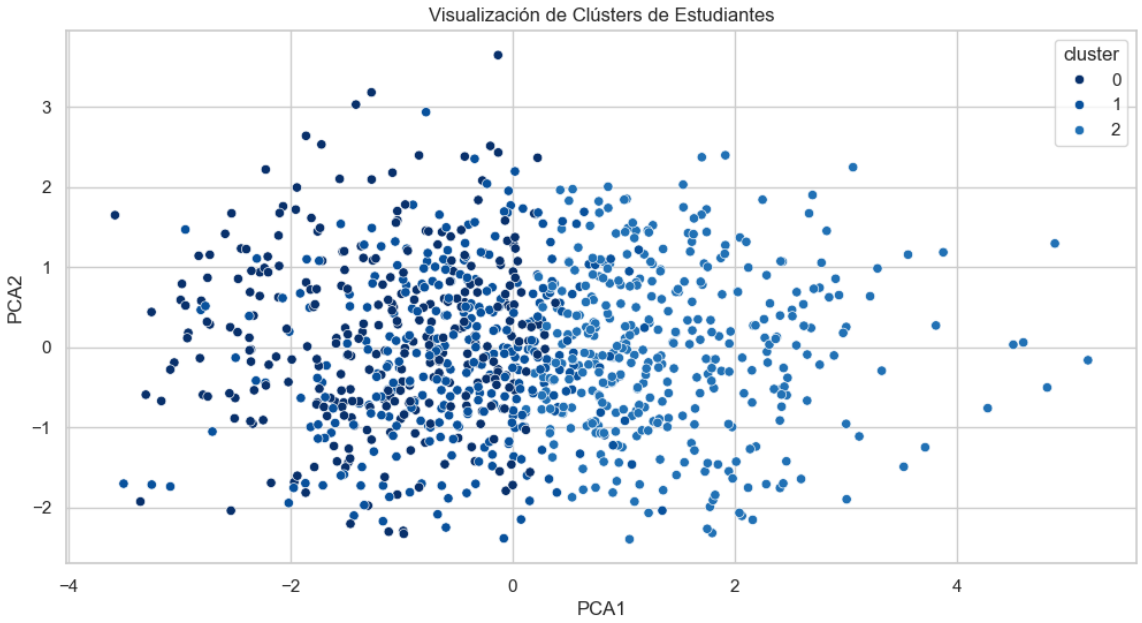
Proceso metodológico:

- **Selección de variables clave**: `study_hours_per_day`, `sleep_hours`, `exercise_frequency`,

`mental_health_rating`, `ratio_estudio_sueno`, `social_media_hours`, `netflix_hours`

- **Estandarización de datos**: Se aplicó `StandardScaler` para normalizar los datos y evitar sesgos por escalas distintas.
- **Determinación del número óptimo de clústers (k)**: Se utilizó el método del codo (Elbow Method) y el coeficiente de Silhouette para encontrar el valor más adecuado de k.
- **Aplicación de K-Means**: Se seleccionó k=3, resultando en tres perfiles principales de estudiantes con patrones diferenciados de comportamiento.
- **Visualización con PCA (Análisis de Componentes Principales)**: Se redujo la dimensionalidad del espacio de datos a dos componentes para representar visualmente la segmentación lograda por K-Means.





6. Clasificación de Rendimiento Académico

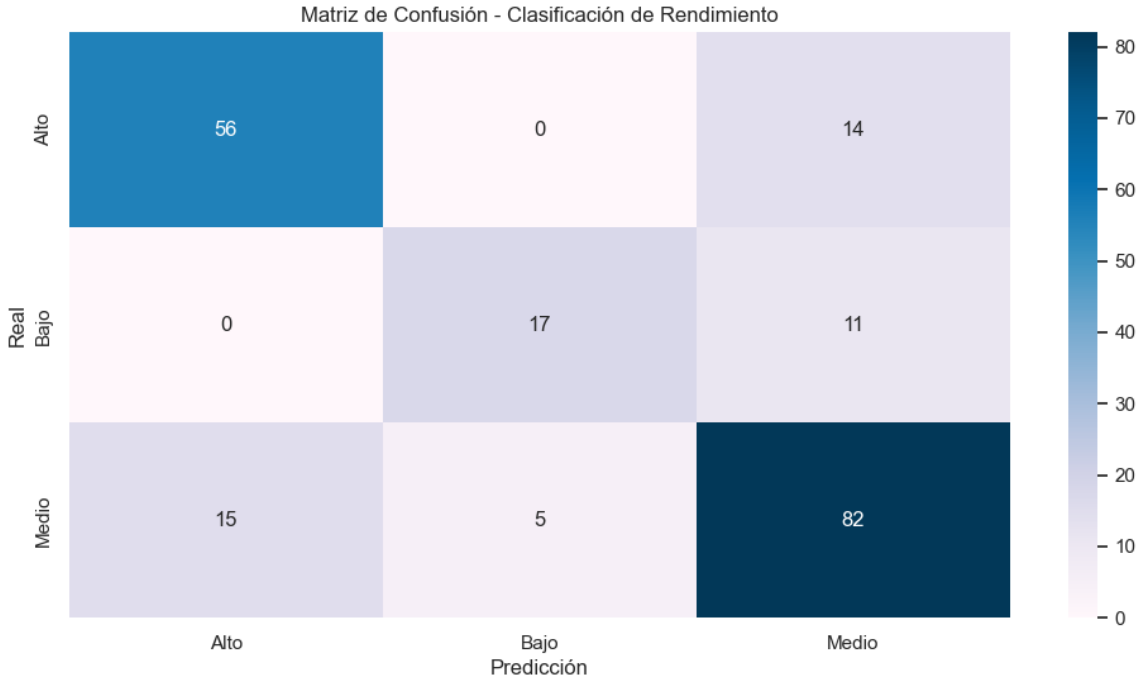
Se entrenó un modelo Random Forest para predecir si un estudiante tendrá rendimiento bajo, medio o alto. Las variables más influyentes fueron las horas de estudio por día, la relación estudio/sueño y la salud mental. El modelo alcanzó una precisión cercana al 80%.

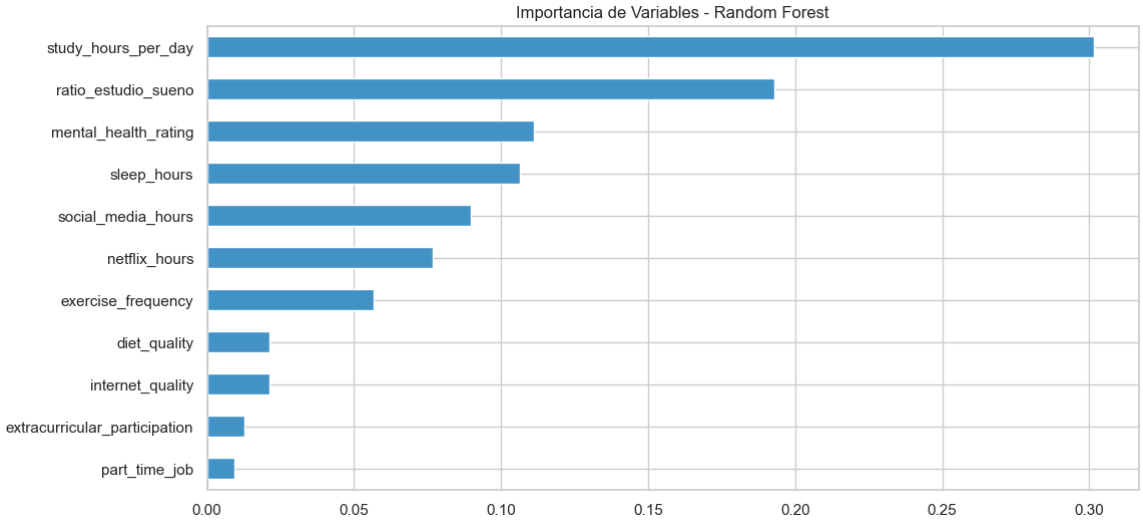
Objetivo:
Predecir si un estudiante presentará un rendimiento académico Bajo, Medio o Alto, a partir de sus características personales y hábitos.

Proceso metodológico:

- **Codificación del target**:
  La variable `rendimiento_academico` fue transformada a valores numéricos utilizando `LabelEncoder`.
- **Selección de variables predictoras**:
  Incluye: horas de estudio, sueño, salud mental, ejercicio, uso de redes, dieta, calidad de internet, participación extracurricular y trabajo.
- **Separación del dataset en entrenamiento y prueba**:
  Se usó un 80% de los datos para entrenamiento y un 20% para prueba (`train_test_split`).
- **Entrenamiento del modelo**:
  Se empleó un `Random Forest Classifier`, elegido por su capacidad para manejar datos mixtos y alta precisión.
- **Evaluación del modelo**:
  Se calcularon métricas de precisión (`precision`), recuperación (`recall`) y puntuación f1 (`f1-score`).

También se graficó una matriz de confusión para visualizar los errores por clase, y se analizó la importancia de cada variable en la predicción.





7. Tiempo Libre vs Salud Mental y Rendimiento

Se segmentó a los estudiantes según su nivel de tiempo libre (bajo, medio, alto) y se analizaron diferencias en salud mental y puntaje.

• Estudiantes con poco tiempo libre tienden a rendir más, pero con mayor riesgo de estrés.

• Los que tienen más tiempo libre muestran una salud mental más variable y puntajes menores en promedio.

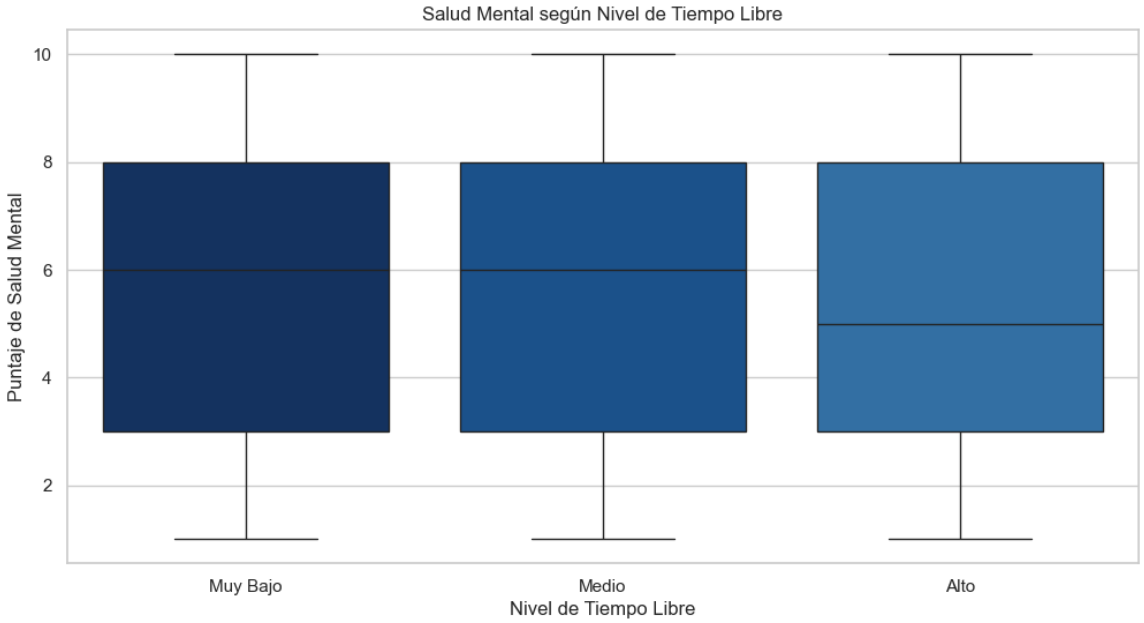
Objetivo:
Explorar cómo el nivel de tiempo libre se relaciona con el estado de salud mental y el rendimiento académico.

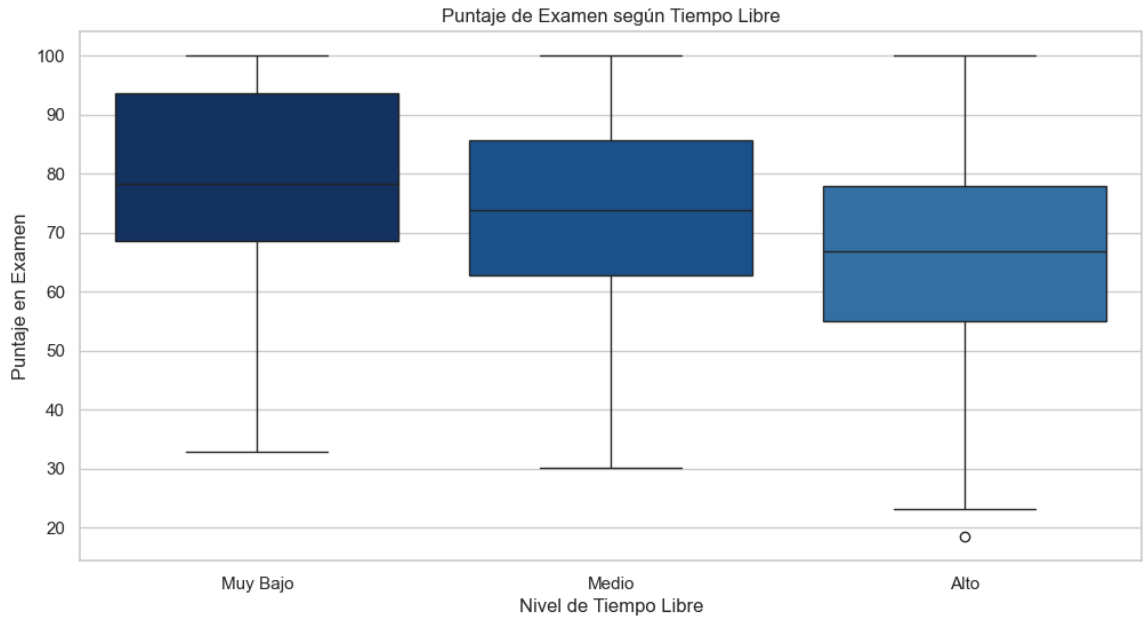
Proceso metodológico:

- **Clasificación de niveles de tiempo libre**:
  Se crearon tres categorías:
  - Muy Bajo: 0 a 4 horas diarias
  - Medio: entre 4 y 8 horas
  - Alto: más de 8 horas
- **Análisis de salud mental**:
  Se utilizó un boxplot para comparar el índice de salud mental (`mental_health_rating`) según el nivel de tiempo libre.
- **Análisis de rendimiento académico**:
  Otro boxplot permitió observar cómo cambia el puntaje de examen (`exam_score`) en función del nivel de tiempo libre.

Principales hallazgos:

- Estudiantes con **muy poco tiempo libre** tienden a tener mejores puntajes, aunque pueden estar más expuestos al estrés.
- En contraste, quienes tienen **mucho tiempo libre** presentan mayor variabilidad emocional y, en promedio, puntajes más bajos.





8. Conclusiones

• Se identificaron 3 perfiles estudiantiles distintos.
• La clave del alto rendimiento no es solo estudiar más, sino tener un balance adecuado entre estudio y descanso.
• La salud mental tiene una correlación positiva significativa con el rendimiento.
• El tiempo libre debe ser gestionado con intención: ni demasiado ni escaso.
• El modelo predictivo puede ser útil para alertar tempranamente sobre riesgo de bajo rendimiento.