

ANÁLISIS DE FACTORES INFLUYENTES EN LA PERFORMANCE ACADÉMICA DE ESTUDIANTES

Camila Spinelli

Fundamentos de la Ciencia de Datos



1. Título del Proyecto

- “Análisis de factores influyentes en la performance de estudiantes aplicando metodología CRISP-DM”.

2. Introducción – El Problema de Negocio

- Se desea desarrollar un modelo que ayude a identificar los patrones en el entorno y comportamiento de los estudiantes, de forma que se permita a la institución educativa y sus interesados a implementar estrategias para ayudar a mejorar el rendimiento académico.
- Algunas preguntas que se desean responder con el proyecto de ciencia de datos:
 - ¿Cuáles son las variables que tienen mayor nivel de influencia en el rendimiento académico?
 - ¿Cómo impacta el entorno familiar en los resultados de los estudiantes?
 - ¿Existen patrones en los hábitos de estudio que se relacionen con la performance del estudiante?
 - ¿Existe una relación entre los factores personales y la performance académica?

3. Objetivos

- Generar un modelo de datos que permita identificar y analizar que factores influyen en el rendimiento académico de los estudiantes, con el objetivo de obtener patrones que ayuden a poder proporcionar diferentes accionables para optimizar su performance en los exámenes.
- Se desea:
 - Evaluar el impacto de las variables de hábito de estudio (pe: asistencia a clase, horas de estudio, sesiones de tutoría) del estudiantes en el rendimiento académico.
 - Analizar la influencia del contexto familiar en los resultados.
 - Monitorizar y visualizar mediante un dashboard el impacto de las diferentes variables clave en los resultados académicos.
 - Entregar un dashboard de PowerBI con un reporte de las conclusiones finales.

4. Fuentes de datos

- El DataSet que se va a utilizar en este proyecto contiene información de **6607 estudiantes con 20 variables** que representan sus hábitos de estudio, entorno familiar, datos personales y resultados académicos.
 - Datos personales: género, problemas de aprendizaje, horas de sueño, nivel de motivación, participación en actividades extracurriculares, influencia de pares.
 - Hábito de estudio: Cantidad de horas dedicadas al estudio, asistencia a clase, asistencia a tutorías.
 - Entorno familiar: nivel de involucramiento familiar, nivel de educación de los padres, nivel económico familiar, acceso a internet.
 - Entorno académico: Acceso a recursos, calidad de profesor, tipo de educación.
- Las columnas incluyen tanto datos cuantitativos como cualitativos y los datos son de tipo numérico o string.
- En la siguiente pagina se adjunta dos tablas:
 - La primera con una descripción de los datos: nombre de columnas, tipo de datos, clasificación de medida y una breve descripción del dato.
 - La segunda con una descripción estadística de los datos cuantitativos.
- Estos datos son de tipo cross sectional ya que representan las características de los estudiantes en un único punto en el tiempo, de origen estructurado y de fuente primaria.
- Se puede encontrar el jupyter notebook y markdown del análisis inicial de datos en el repositorio de [GitHub bajo la carpeta de informes](#).

Descripción general de los datos:

Columna	Tipo de dato	Clasificación	Breve descripción
Hours_studied	Int64	Cuantitativo	Horas dedicadas al estudio
Attendance	Int64	Cuantitativo	Porcentaje de asistencia a clase
Parental_involvement	String	Cualitativo	Nivel de involucramiento de los padres (bajo, medio, alto)
Access_to_Resources	String	Cualitativo	Nivel de acceso a recursos (bajo, medio, alto)
Extracurricular_Activities	String	Cualitativo	Estudiante realiza actividades extracurriculares (si, no)
Sleep_Hours	Int64	Cuantitativo	Horas de sueño
Previous_Scores	Int64	Cuantitativo	Resultados anteriores
Motivation_Level	String	Cualitativo	Nivel de motivación de estudio (bajo, medio, alto)
Internet_Access	String	Cualitativo	Estudiante cuenta con acceso a internet (si, no)
Tutoring_Sessions	Int64	Cuantitativo	Sesiones de tutoría que el estudiante asistió
Family_Income	String	Cualitativo	Ingreso económico del núcleo familiar (bajo, medio, alto)
Teacher_Quality	String	Cualitativo	Calidad de profesor del estudiante (bajo, medio, alto)
School_Type	String	Cualitativo	Tipo de institución al que asiste (pública, privada)
Peer_Influence	String	Cualitativo	Nivel de influencia de sus pares (negativa, neutra, positiva)
Physical_Activity	Int64	Cuantitativo	Horas de actividad física
Learning_Disabilities	String	Cualitativo	Dificultad de aprendizaje (si, no)
Parental_Education_Level	String	Cualitativo	Nivel de educación de los padres (secundario, universitario, posgrado)
Distance_from_Home	String	Cualitativo	Distancia de la casa al centro educativo (lejos, cerca, moderado)
Gender	String	Cualitativo	Genero del estudiantes (femenino, masculino)
Exam_Score	Int64	Cuantitativo	Resultado final del exámen

Descripción estadística de los datos cuantitativos:

Columna	Media	Desviación Estándar	Min.	25 %	Mediana (50%)	75 %	Máx.
Hours_studied	19.98	5.99	1	16	20	24	44
Attendance	79.98	11.55	60	70	80	90	100
Sleep_Hours	7.03	1.47	4	6	7	8	10
Previous_Scores	75.07	14.40	50	63	75	88	100
Tutoring_Sessions	1.49	1.23	0	1	1	2	8
Physical_Activity	2.97	1.03	0	2	3	4	6
Exam_Score	67.24	3.89	55	65	67	69	101

5. Plan de Trabajo

Fase del Proyecto	Semana									
	01	02	03	04	05	06	07	08	09	10
Fase 1: Comprensión del negocio	■									
Fase 2: Comprensión de los datos		■	■	■						
Fase 3: Preparación de los datos		■	■	■	■					
Fase 4: Modelado			■	■	■	■				
Fase 5: Evaluación				■	■	■				
Fase 5: Implementación							■	■		
Entrega del Proyecto									■	

6. Metodología

- La metodología que se utilizará en este proyecto será CRISP-DM.
- Para responder a las preguntas plantadas se utilizarán modelos de regresión por la naturaleza de los datos a analizar.
- Para el entregable del proyecto se va a realizar un reporte final y un dashboard en PowerBI con gráficos y métricas clave para entender la influencia de cada variable.

7. Tecnología

- Gestión de datos: MySQL.

- Lenguaje de programación librerías: Python con Pandas.
- Plataforma de desarrollo del proyecto: Jupyter Notebooks.
- Control de versiones: [GitHub](#).

8. Análisis Exploratorio de Datos

Introducción:

El objetivo principal de este proyecto es entender cuales son los factores que mas influyen en el rendimiento académico (medido en los datos como Exam_Score).

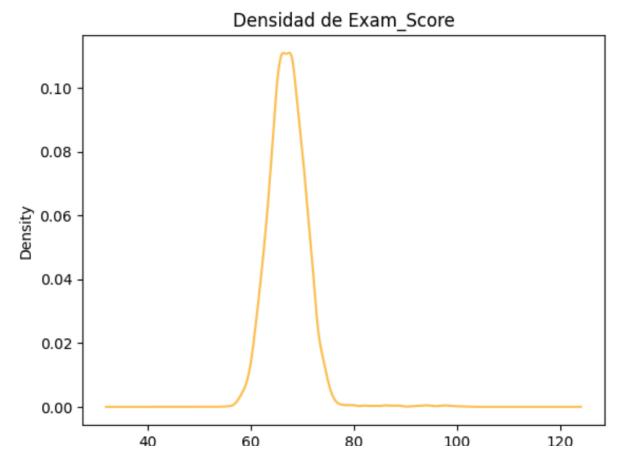
Para el total de 6607 datos de alumnos que vamos a estudiar, encontramos que existen datos faltantes en las siguientes variables:

Variable	Missing Values	Percentage
Teacher_Quality	78	1.180566
Parental_Education_Level	90	1.362192
Distance_from_Home	67	1.014076

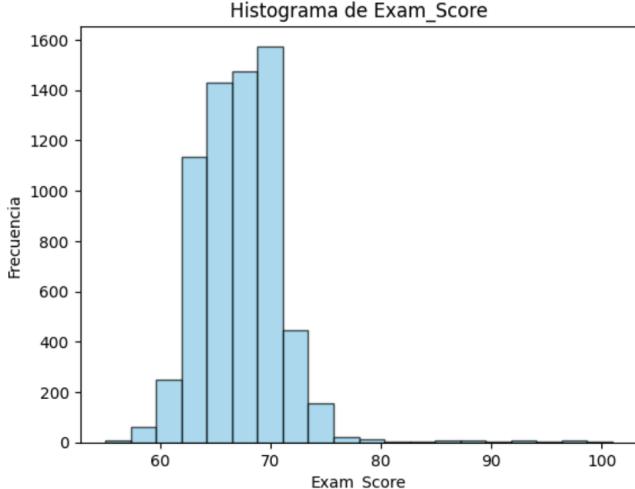
Dado que no se logró identificar una tendencia clara entre los valores faltantes y otras variables (como, por ejemplo, si Teacher_Quality tenía una moda diferente para escuelas publicas y privadas o si Parental_Education_Level o Distance_from_Home variaban su moda según los diferentes rango de ingreso familiar) y como el porcentaje de valores faltantes es muy bajo, se decidió imputar los datos utilizando los valores de la moda para todos los casos. De forma que los diferentes valores faltantes fueron cambiados por los siguientes valores:

Variable	Moda
Teacher_Quality	Medium
Parental_Education_Level	High School
Distance_from_Home	Near

Comencemos analizando nuestra variable objetivo, **Exam_Score**:

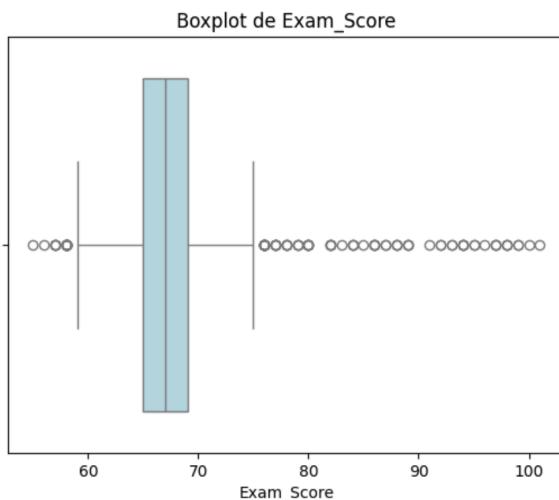


El gráfico muestra una distribución asimétrica sesgada hacia la derecha. El pico de densidad más alta está alrededor de la media (67.24). La mayoría de los estudiantes presentaron un rendimiento cercano al promedio, pero la cola derecha muestra que hay una proporción de estos que obtiene puntajes más altos.



El histograma reafirma lo que ya vimos en el gráfico anterior: una distribución con mayor densidad en los valores cercanos a la media, con una baja significativa en las frecuencias de valores mayores a 80. Hay pocos estudiantes con valores muy altos (mayores a 90) o muy bajos (menores a 60).

Este gráfico también presenta una distribución sesgada a la derecha, confirmando que la mayoría de los estudiantes obtienen resultados cercanos a la media, con algunas excepciones con valores altos.



En el gráfico de boxplot se observa que existen outliers para ambos extremos pero predominan los valores hacia la derecha (puntajes más altos).

Para el análisis exploratorio del resto de las variables, las agruparemos según los diferentes objetivos planteados en este proyecto. Cada variable independiente se ha asignado a un objetivo específico, dando como resultado el siguientes listado:

Objetivos	Variables
1: ¿Cómo impacta el entorno familiar en los resultados de los estudiantes?	Parental_Involvement, Access_to_Resources, Family_Income, Internet_Access, Parental_Education_Level
2: ¿Existe una relación entre los factores personales y la performance académica?	Extracurricular_Activities, Motivation_Level, Peer_Influence, Learning_Disabilities, Sleep_Hours, Physical_Activity
3: ¿Existen patrones en los hábitos de estudio que se relacionen con la performance del estudiante?	Hours_Studied, Attendance, Previous_Scores, Tutoring_Sessions

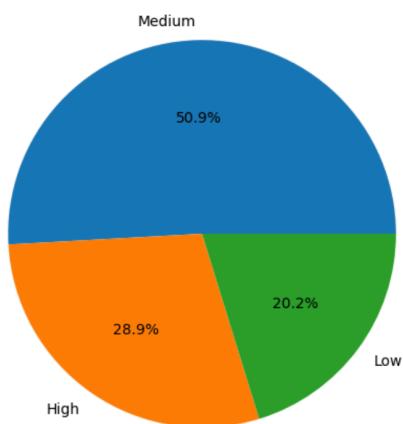
1. Entorno Familiar

1.1 Análisis univariado

Todas las variables que se incluyen en este objetivo son de tipo categóricas, por lo que haremos un estudio de la moda de cada una e interpretaremos gráficos circulares para ver como se distribuyen los datos.

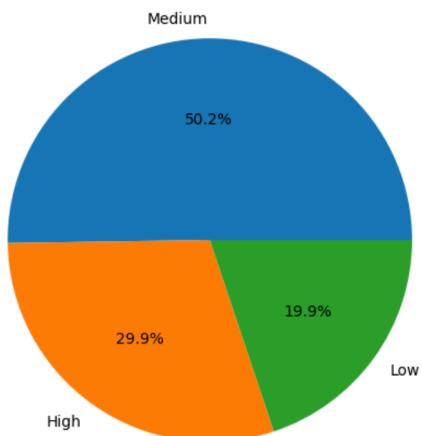
Variable	Moda
Parental_Involvement	Medium
Access_to_Resources	Medium
Family_Income	Low
Internet_Access	Yes
Parental_Education_Level	High School

Distribución de Parental_Involvement



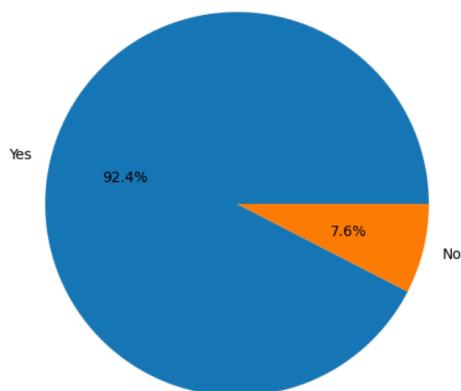
La categoría "Medium" (con el 51% de los datos) es la moda para el involucramiento de los padres. Las categorías "High" y "Low" están distribuidas equitativamente, cada una abarcando cerca de un cuarto de los datos.

Distribución de Access_to_Resources



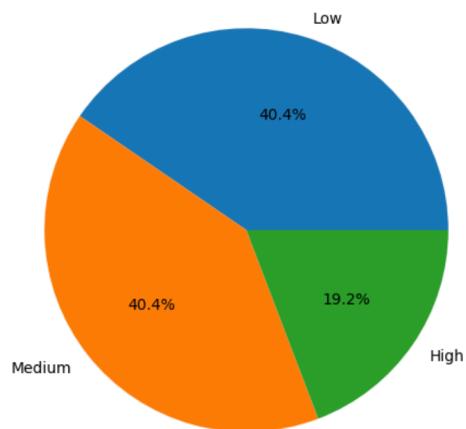
Similar a "Parental Involvement", la categoría "Medium" domina con el 50.2% de los datos. "High" y "Low" representan cada una cerca del 25%, sugiriendo un acceso equilibrado, aunque la mayoría de los estudiantes disponen de recursos de manera moderada.

Distribución de Internet_Access



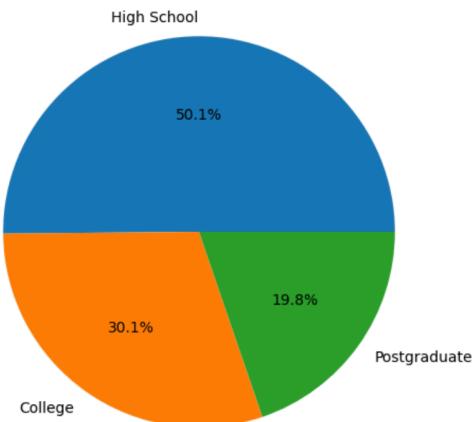
El acceso a Internet es ampliamente predominante, con un 92.4% de los estudiantes indicando "Sí". La categoría "No" constituye solo el 7.6%.

Distribución de Family_Income



Las categorías "Low" y "Medium" están equilibradas, cada una con un 40.4% de los datos, mientras que "High" constituye solo el 19.2%.

Distribución de Parental_Education_Level



"High School" es la categoría más frecuente, representando el 50.1%. "College" y "Postgraduate" se dividen casi equitativamente entre el resto de los datos, con un 30.1% y un 19.8%, respectivamente.

1.1 Análisis bivariado

En esta sección se presenta un análisis bivariado de las variables seleccionadas en relación con nuestra variable objetivo, *Exam_Score*. Este análisis incluye visualizaciones como boxplots y un heatmap que se construyó tras aplicar el método de *one-hot encoding* a las variables categóricas:

El heatmap refleja las correlaciones entre las variables categóricas transformadas mediante *one-hot encoding* y la variable objetivo (*Exam_Score*). Las observaciones principales son las siguientes:

1. Correlaciones bajas:

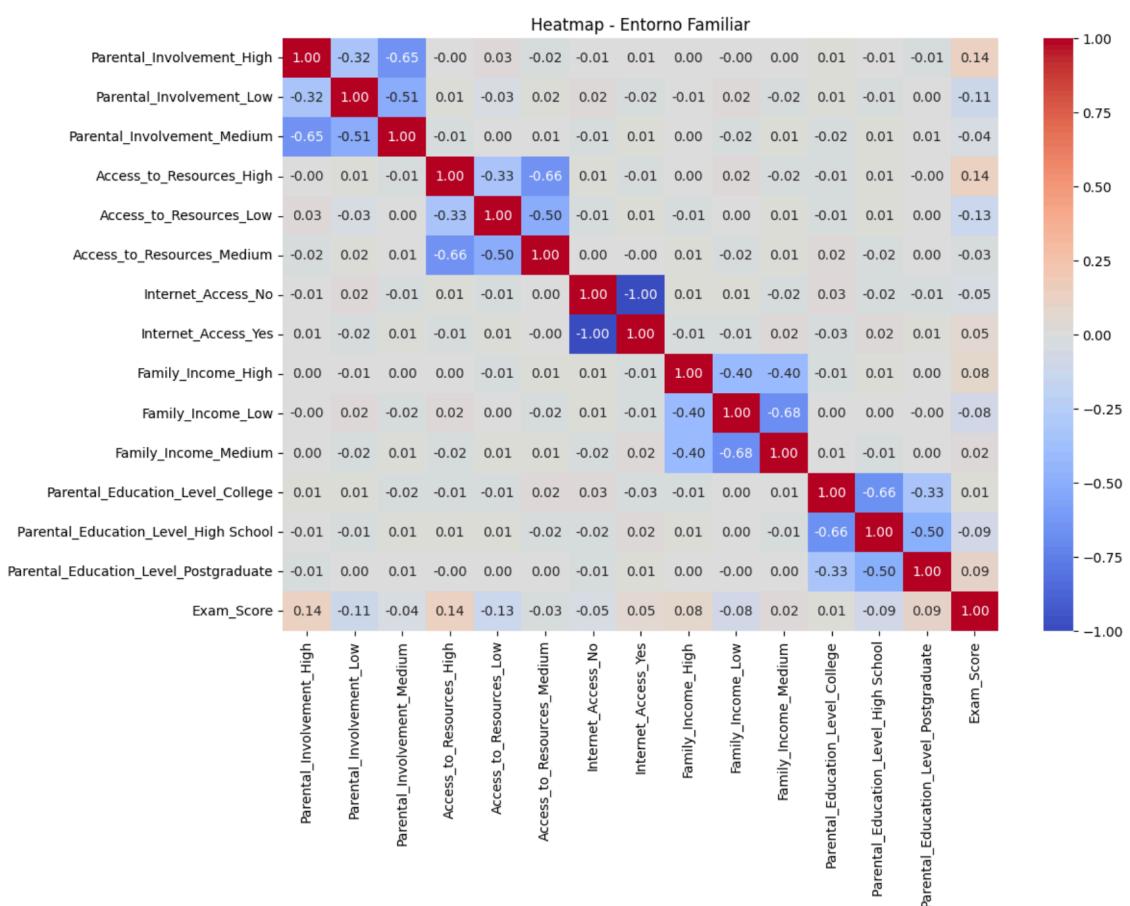
Las correlaciones con la variable *Exam_Score* son generalmente débiles, con valores que no superan ± 0.14 . Esto indica que las variables categóricas del entorno familiar, como *Parental_Involvement*, *Access_to_Resources* y *Parental_Education_Level*, tienen una relación limitada con el rendimiento académico.

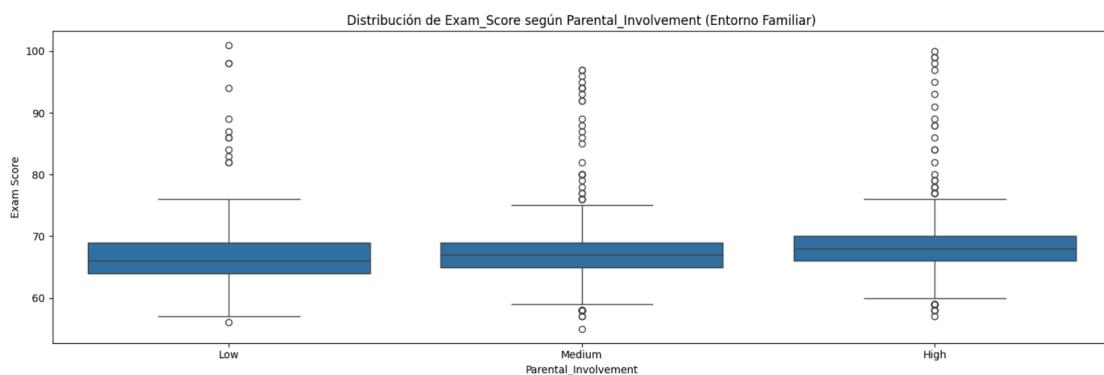
2. Impacto del nivel educativo de los padres:

1. La categoría *High School* dentro de *Parental_Education_Level* tiene una ligera correlación negativa con *Exam_Score* (-0.05)
2. *Postgraduate* muestra una correlación positiva muy baja (+0.03)

3. Acceso a recursos e implicación parental:

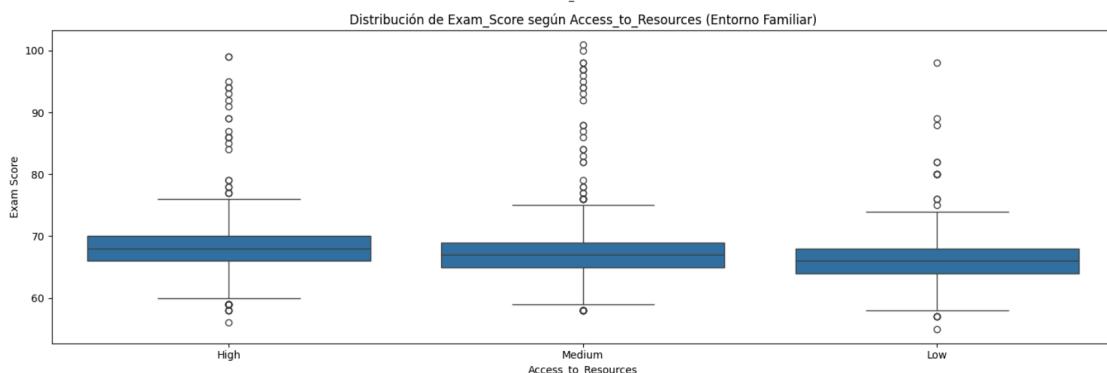
1. Los estudiantes con *High* en *Access_to_Resources* y *Parental_Involvement* tienden a mostrar una correlación positiva muy leve con el rendimiento académico (+0.14)
2. Las categorías *Low* en ambas variables presentan correlaciones negativas mínimas (-0.11 y -0.13, respectivamente)





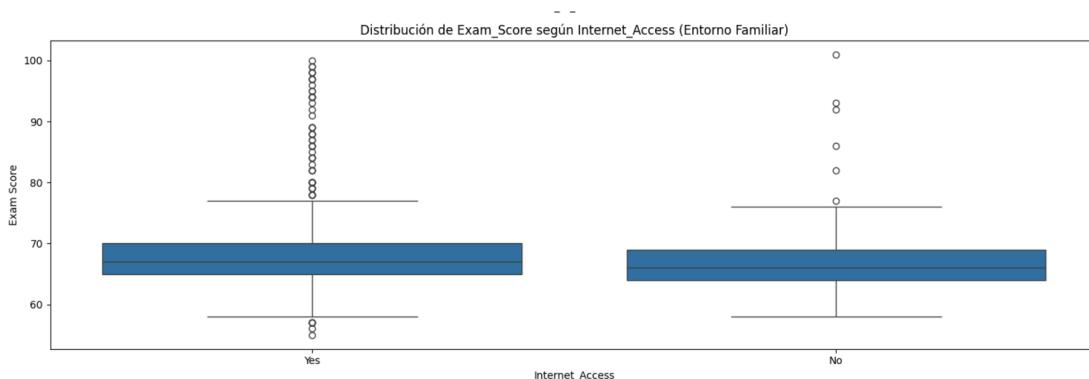
Los estudiantes con un nivel alto de implicación parental (*High*) muestran una mediana ligeramente superior en *Exam_Score* en comparación con los niveles “*Medium*” y “*Low*”. Sin embargo, las diferencias no son significativas, ya que las distribuciones son similares en todos los grupos.

Los outliers hacia puntajes altos se observan en todas las categorías, destacando que algunos estudiantes logran buenos resultados independientemente del nivel de implicación parental.



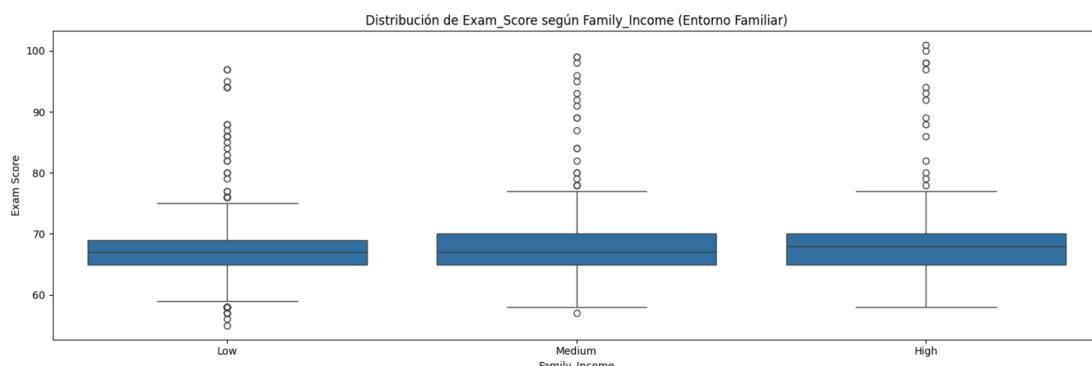
Los estudiantes clasificados como *High* en esta variable tienen una mediana levemente superior en comparación con *Medium* y *Low*. Este patrón refuerza la ligera correlación positiva observada en el heatmap.

La dispersión es similar entre los grupos, indicando que el acceso a recursos no es un factor determinante.



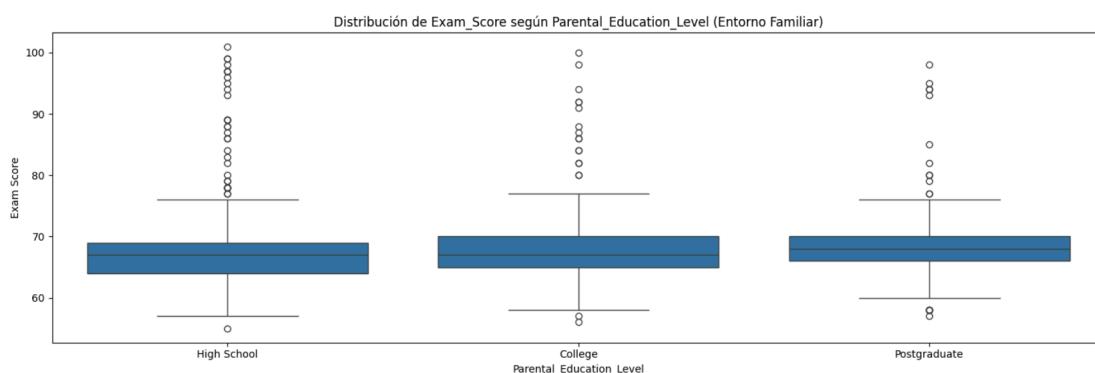
No se observan diferencias significativas en la mediana de *Exam_Score* entre los estudiantes con y sin acceso a internet. Esto coincide con la correlación insignificante detectada en el heatmap.

Ambos grupos presentan una distribución homogénea, con outliers en el extremo superior.



Los estudiantes provenientes de familias con ingresos altos (*High*) tienen una mediana ligeramente superior en comparación con *Low* y *Medium*. Este patrón refuerza la correlación positiva leve observada en el heatmap.

La dispersión es similar entre los grupos, con presencia de outliers en los valores más altos de *Exam_Score*.



Los estudiantes con padres que alcanzaron un nivel educativo de *Postgraduate* muestran una mediana comparable a los de *High School* y *College*. Esto coincide con las correlaciones bajas identificadas en el heatmap.

Las distribuciones son similares entre los grupos, con outliers que reflejan altos puntajes independientemente del nivel educativo de los padres.

En general, aunque las correlaciones son débiles, el análisis sugiere que un mejor acceso a recursos y una mayor implicación parental están ligeramente asociados con un mejor rendimiento académico.

Las observaciones previas nos hacen tomar como hipótesis que las variables del entorno familiar tienen poco impacto en el rendimiento académico. Aunque existen algunos patrones que podrían llegar a considerarse, estos son muy débiles por lo que concluiremos que el entorno familiar no tiene un gran impacto en el rendimiento académico de los estudiantes.

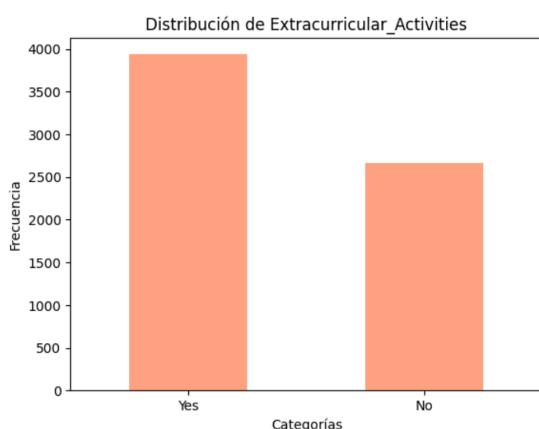
2. Factores Personales

1. Análisis Univariado

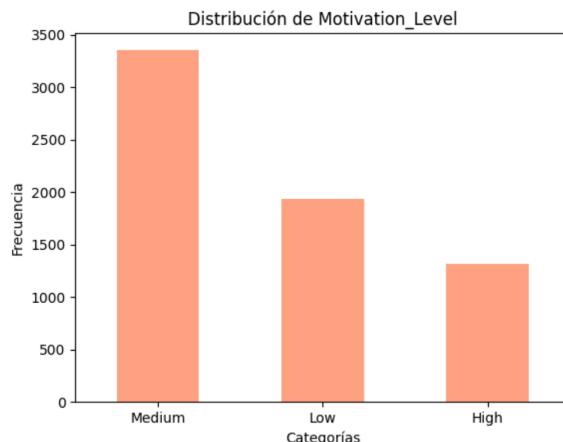
Las variables englobadas en este objetivo son de tipo categórico y cuantitativas. De esta forma, incluiremos un estudio de la moda e interpretaremos gráficos de barras para las primeras, e histogramas y boxplot para el resto.

Variable	Moda
Extracurricular_Activities	Yes
Motivation_Level	Medium
Peer_Influence	Positive
Learning_Disabilities	No

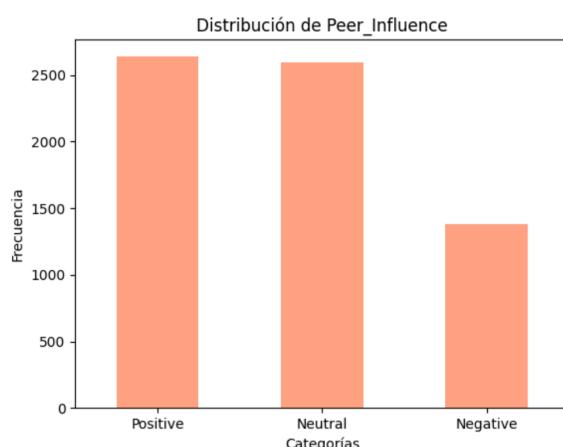
Análisis de Gráficos:



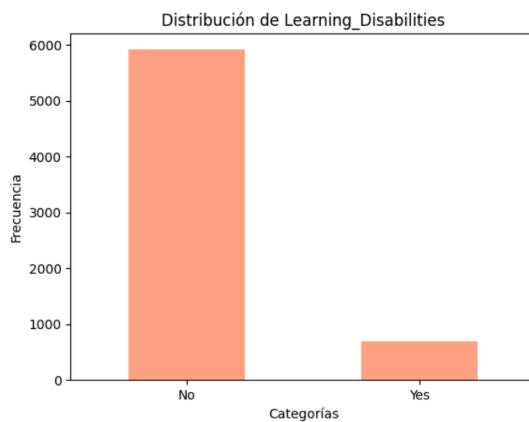
El gráfico muestra que más de la mitad de los estudiantes participan en actividades extracurriculares (“Yes”). Por otra parte, la categoría “No” también representa una porción significativa de los datos, aunque aun menor.



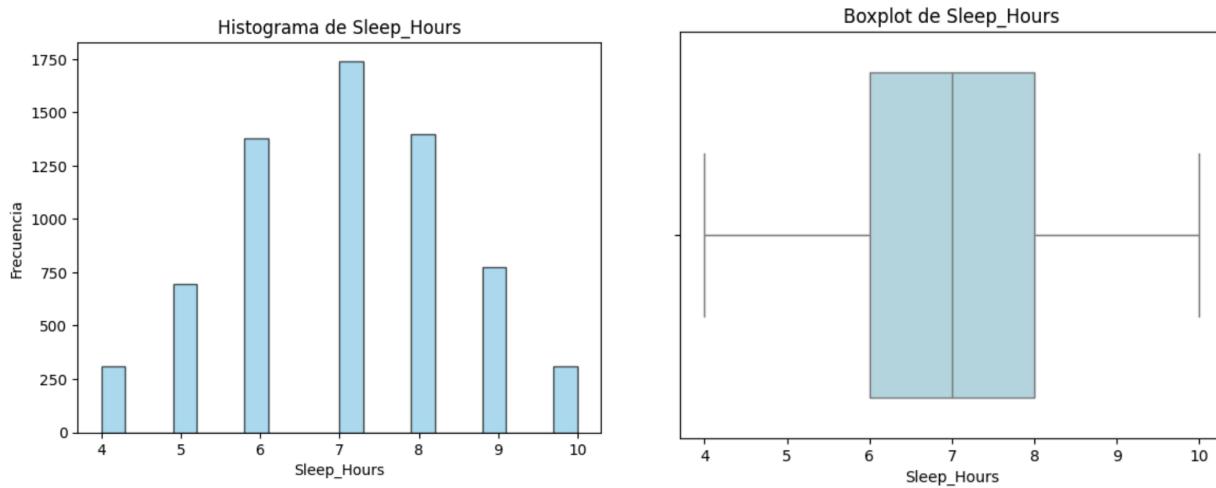
En este gráfico se observa que la mayoría de los estudiantes reportan un nivel de motivación "Medium". Las categorías "Low" y "High" están menos representadas, siendo "Low" más frecuente que "High". Esto sugiere que, aunque los estudiantes están moderadamente motivados en su mayoría, hay un grupo significativo con baja motivación que podría requerir atención.



El gráfico revela que la influencia positiva ("Positive") y la influencia neutral ("Neutral") son las más frecuentes y están equilibradas. Por otro lado, la influencia negativa ("Negative") tiene una menor proporción, lo cual es un indicador positivo del ambiente social en el que se desarrollan los estudiantes.

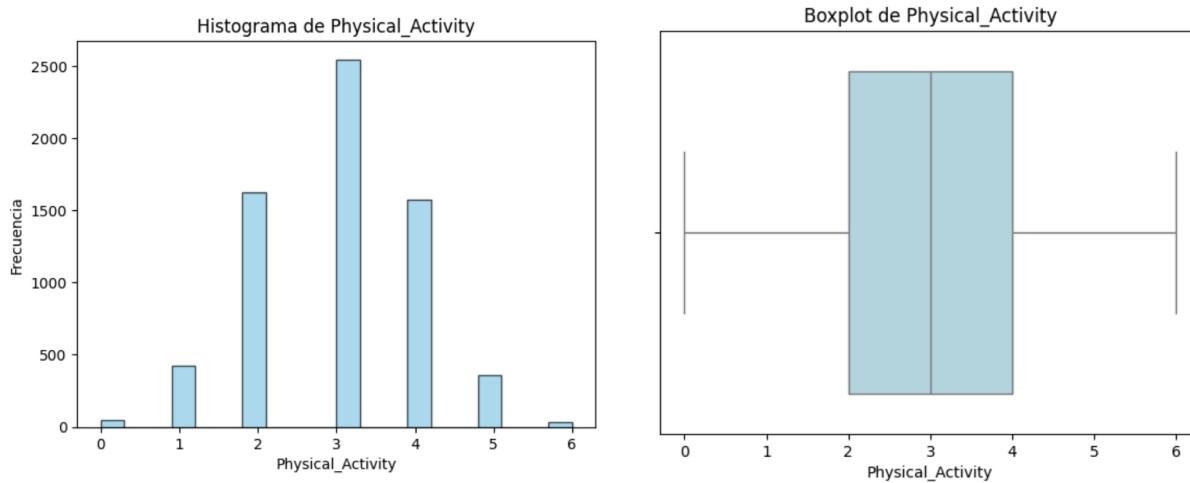


La categoría "No" agrupa significativamente la mayoría de los datos en este caso, lo que indica que la mayoría de los estudiantes no reportan discapacidades de aprendizaje. La categoría "Yes" es mínima para esta muestra.



El histograma muestra una distribución unimodal con un pico en las 7 horas, lo que indica que la mayoría de los estudiantes tienen un tiempo de sueño considerado adecuado. Los valores se distribuyen entre 4 y 10 horas, con una disminución gradual en los extremos, sugiriendo patrones de sueño bastante homogéneos.

Por otra parte, el boxplot confirma la distribución observada en el histograma, con la mediana cerca de las 7 horas. El rango intercuartílico (entre 6 y 8 horas) refuerza que la mayoría de los datos se concentran en un intervalo saludable. No se identifican valores atípicos significativos.



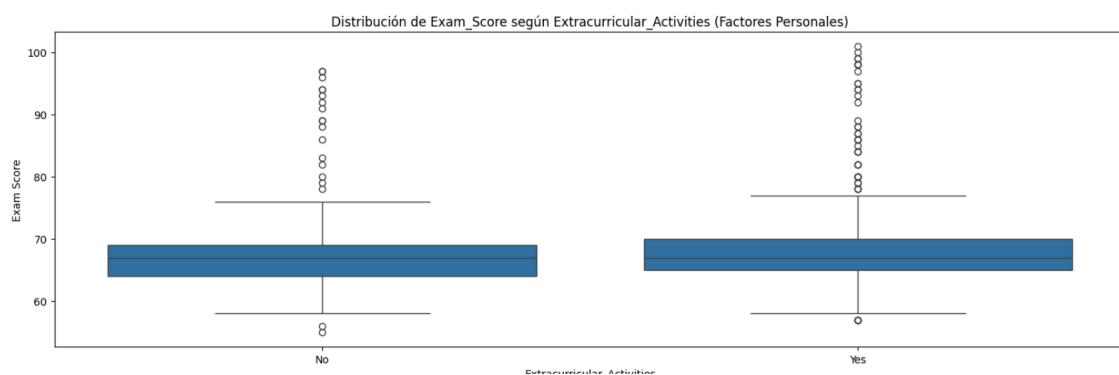
La mayoría de los estudiantes realiza entre 3 y 4 horas de actividad física semanal. Los valores extremos (0 o 6 horas) son menos frecuentes, lo que indica que la actividad física regular es común en esta población estudiada.

El boxplot respalda el histograma, mostrando que los valores centrales están concentrados entre 3 y 4 horas. Los valores atípicos son mínimos y no afectan la interpretación general de la distribución.

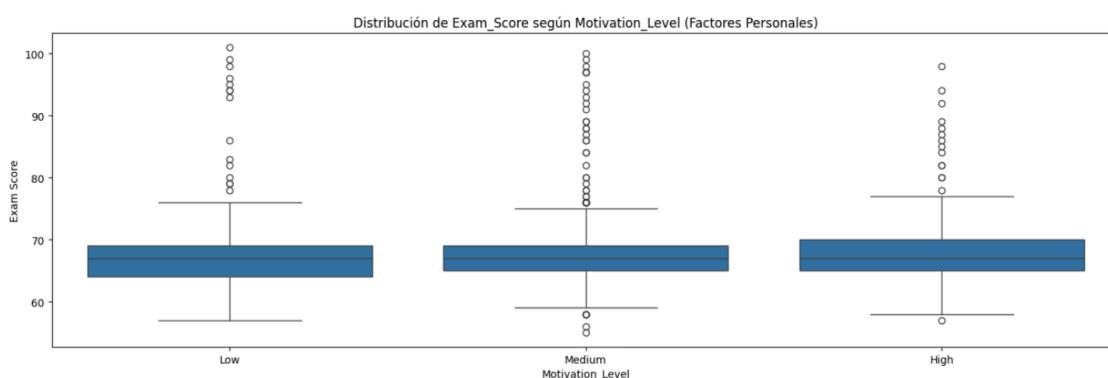
Se destaca que más de la mitad de los estudiantes participan de actividades extracurriculares y la influencia positiva de los compañeros. Además, los datos de sueño y actividad física reflejan hábitos saludables en la mayoría de los estudiantes.

2. Análisis Bivariado

Al igual que para el objetivo anterior, en esta sección se presenta un análisis bivariado de las variables seleccionadas en relación con nuestra variable objetivo, *Exam_Score*. Este análisis incluye visualizaciones como boxplots, pairplots y un heatmap que se construyó tras aplicar el método de *one-hot encoding* a las variables categóricas:

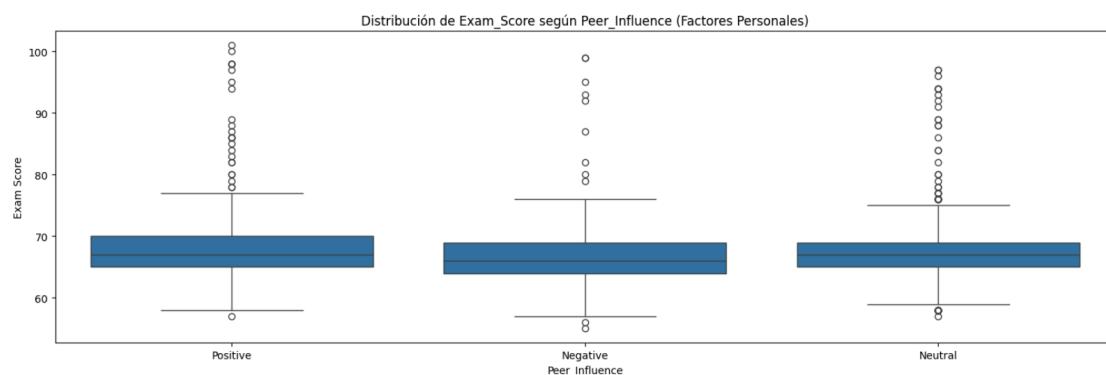


La mediana de los puntajes de "Exam_Score" es similar para estudiantes que participan en actividades extracurriculares (Yes) y los que no (No). Sin embargo, hay mayor dispersión en los puntajes altos para aquellos que participan en actividades extracurriculares. Esto podría indicar que la participación en actividades no afecta significativamente el rendimiento promedio, pero puede estar asociada con estudiantes que logran puntajes muy altos.

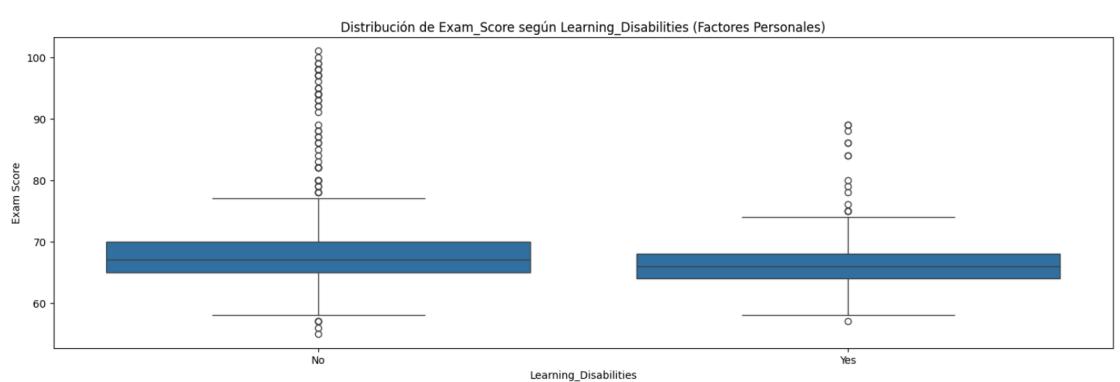


Los puntajes de "Exam_Score" presentan una mediana muy similar entre los tres niveles de motivación (alrededor de 70). Esto sugiere que la puntuación típica no varía significativamente según el nivel de motivación.

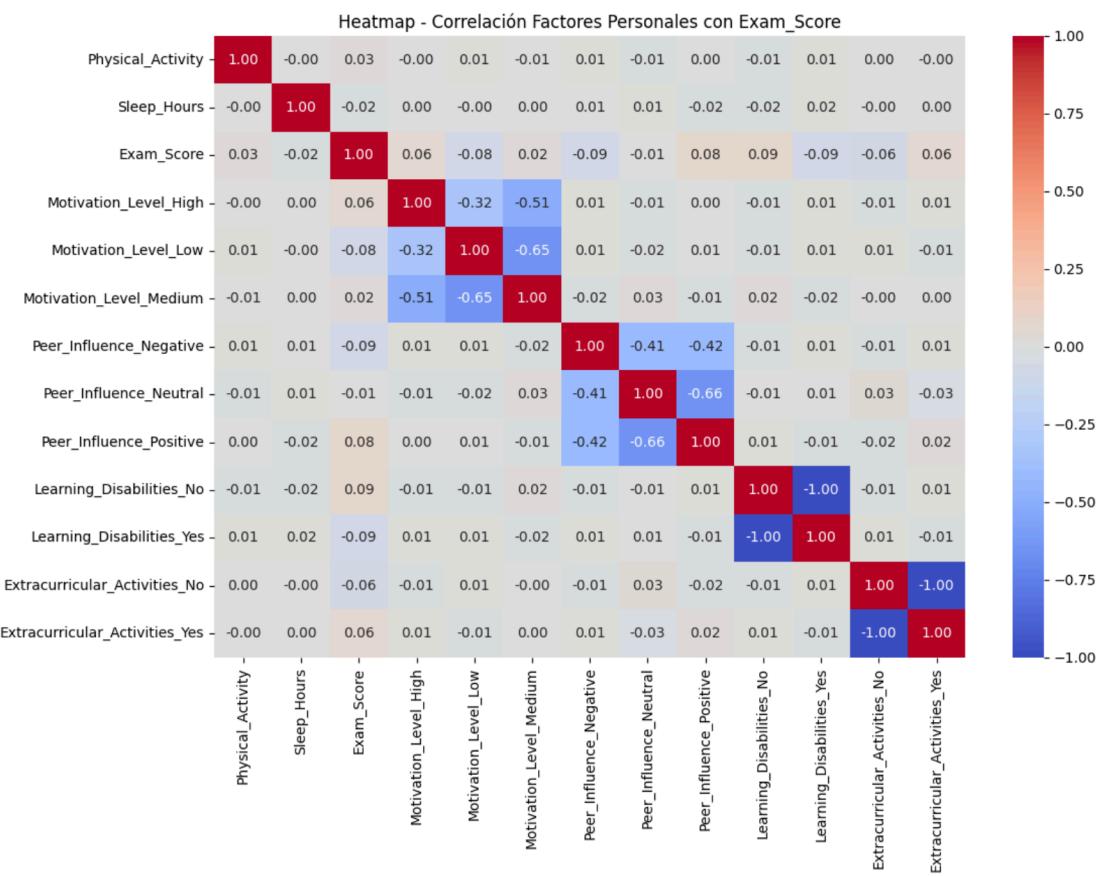
Para todas las categorías existen valores atípicos, tanto por la parte superior como por la parte inferior (menos para la categoría "Low" en este caso). Los outliers no son especialmente más frecuentes para un nivel de motivación en particular.



La influencia positiva de los compañeros ("Positive") parece estar asociada con una mediana ligeramente más alta en los puntajes de "Exam_Score" comparada con las influencias "Neutral" y "Negative". Sin embargo, la diferencia no es drástica, y la existencia de outliers en ambos extremos de las categorías sugiere que otros factores podrían tener más peso en el rendimiento.



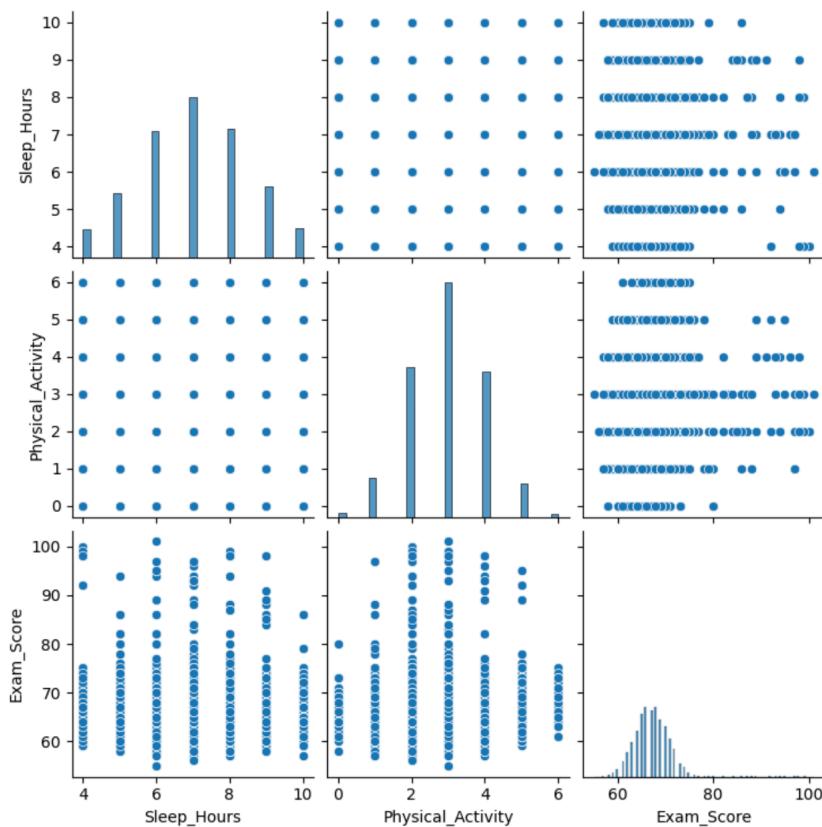
Los estudiantes con discapacidades de aprendizaje ("Yes") tienen una distribución más limitada y centrada alrededor de la mediana, aunque los outliers sugieren que algunos logran puntajes excepcionalmente altos. Los estudiantes sin discapacidades ("No") tienen una mayor variabilidad en sus puntajes, pero la mediana de ambos grupos es muy similar, indicando que este factor no genera diferencias sustanciales en el promedio general.



El heatmap muestra correlaciones débiles entre las variables elegidas en este objetivo y el "Exam_Score" (todas son por debajo de 0.1).

"Motivation_Level_High" y "Peer_Influence_Positive" tienen ligeras correlaciones positivas con "Exam_Score", aunque el impacto es pequeño.

La correlación negativa con "Motivation_Level_Low" y "Peer_Influence_Negative" refuerza la idea de que estos factores podrían estar asociados con menores puntajes, pero sigue siendo muy débil la correlación en este caso.



La dispersión de los puntajes de "Exam_Score" no parece estar significativamente influenciada por "Sleep_Hours" o "Physical_Activity".

"Sleep_Hours" muestra un comportamiento más uniforme entre los diferentes rangos, sin patrones claros de aumento o disminución del puntaje.

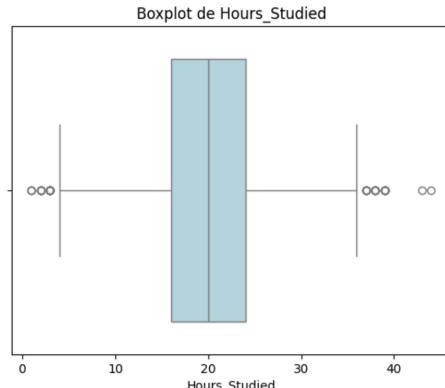
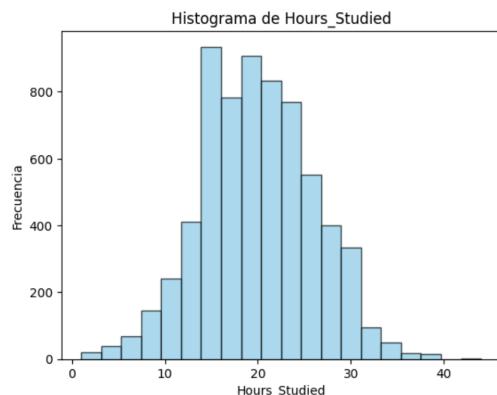
"Physical_Activity" tiene una distribución más segmentada, pero tampoco se observan tendencias claras que vinculen esta variable con el "Exam_Score".

Este análisis sugiere que los factores personales tienen una influencia muy limitada en el "Exam_Score" cuando se analizan de forma independiente. Aunque algunos factores muestran ligeras tendencias, los efectos no son lo suficientemente fuertes como para considerarlos determinantes

3. Hábitos de Estudio

1. Análisis Univariado

Las variables seleccionadas para este objetivo son todas de tipo cuantitativas. Por lo que el análisis que realizaremos en esta sección estará basado en histogramas y boxplots.

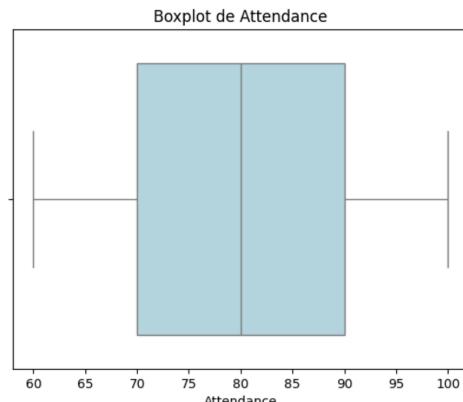
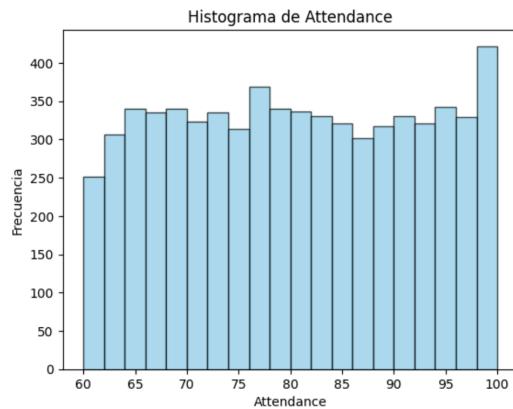


El histograma muestra una distribución simétrica centrada alrededor de las 20 horas de estudio semanales. La mayoría de los estudiantes estudian entre 10 y 30 horas semanales.

El boxplot refuerza que la mediana está cerca de las 20 horas, con algunos pocos valores atípicos cercanos a 40 horas en el extremo superior y menores de 5 horas en el otro caso.

Ambos gráficos indican que la mayoría de los estudiantes tienen hábitos de estudio moderados (10-30 horas), con una distribución bien centrada y pocos valores extremos.

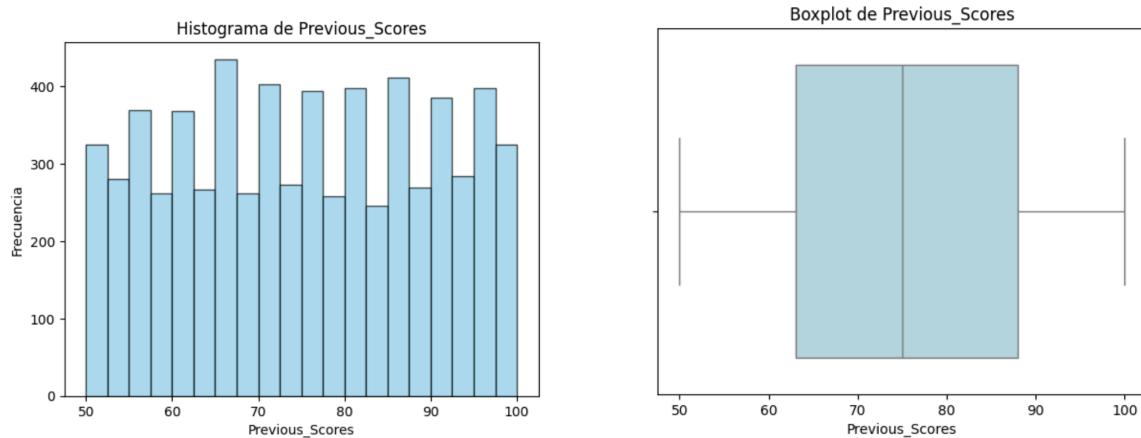
Podemos concluir de ambos gráficos que los estudiantes tienen un patrón de estudio consistente y equilibrado, donde la mayoría de los estudiantes dedican entre 15 y 25 horas semanales, que corresponde aproximadamente al rango intercuartílico (IQR).



En este caso el histograma nos muestra una distribución relativamente uniforme con una ligera concentración hacia el 100% de asistencia.

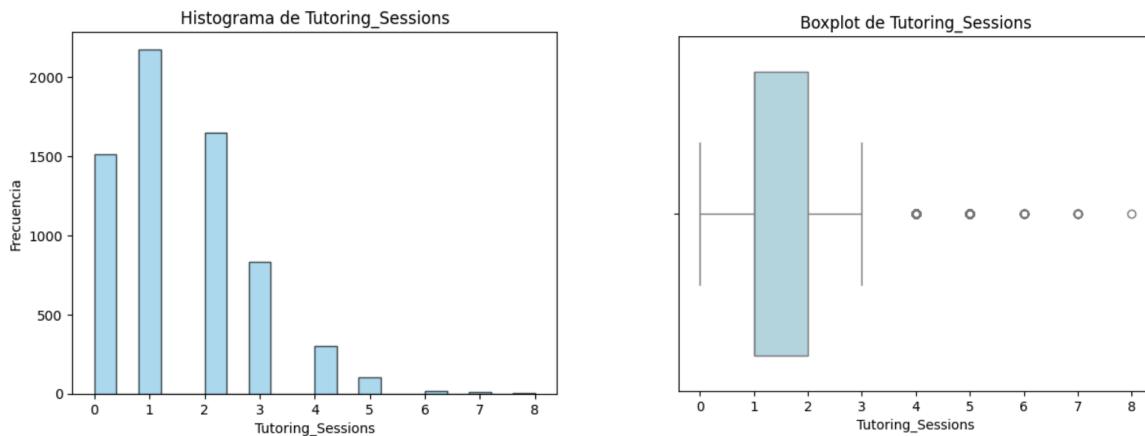
El boxplot nos muestra que la mediana se encuentra en un rango cercano al 80% con valores equilibrados. Por lo que al menos la mitad de los estudiantes tienen una asistencia superior a este valor.

Ambos gráficos sugieren que asistencia a clase está dentro de un buen promedio, con la mayoría de los estudiantes cumpliendo más del 75%.



El histograma muestra una distribución cercana a uniforme, con valores de puntajes previos bien distribuidos entre 50 y 100. Esto indica que no hay rangos específicos donde se concentre un grupo considerable de estudiantes, lo que refleja una diversidad en los niveles de rendimiento académico previo.

El boxplot complementa este análisis al mostrar que la mediana se encuentra alrededor de 75, con una muy ligera inclinación hacia puntajes más altos. No se observan valores extremos significativos, lo que sugiere que la mayoría de los estudiantes se encuentran dentro de un rango esperado.



El histograma de sesiones de tutoría muestra una fuerte asimetría hacia la izquierda, con la mayoría de los estudiantes asistiendo entre 0 y 2 sesiones. Las frecuencias disminuyen rápidamente para valores mayores a 2, y las sesiones superiores a 4 son casos muy raros.

El boxplot refuerza esta observación, con una mediana de 1 sesión de tutoría. El rango intercuartílico está entre 1 y 2 sesiones, por lo que el 50% de los estudiantes se encuentran en

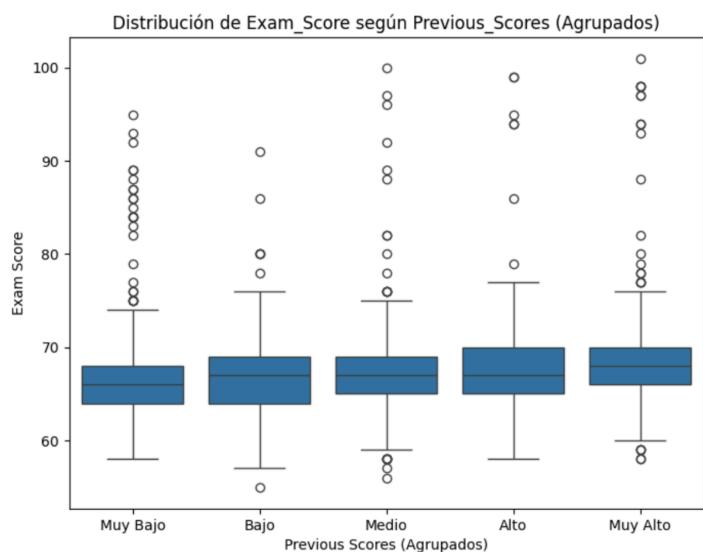
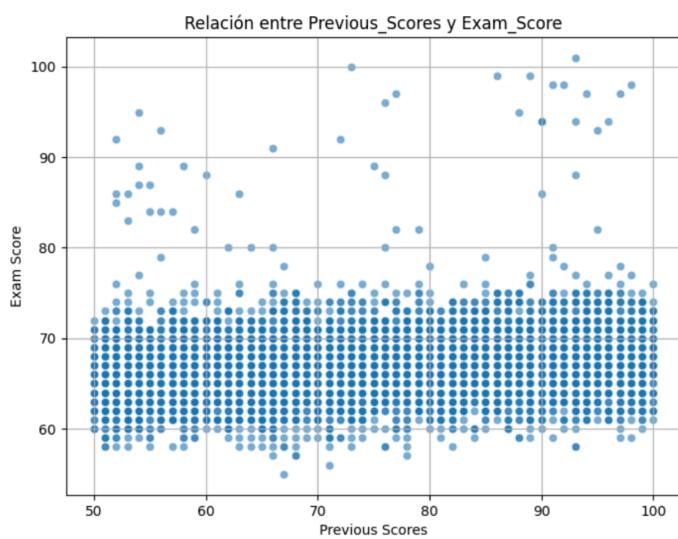
este intervalo. Los valores entre 5 y 8 sesiones son claramente excepciones, marcados como atípicos.

La mayoría de los estudiantes realiza entre 1 y 2 sesiones de tutoría, lo que refleja un acceso o interés limitado en este recurso académico. Los valores extremos sugieren que solo unos pocos estudiantes recurren intensivamente a estas sesiones, lo cual podría estar relacionado con necesidades académicas específicas.

2. Análisis Bivariado

En esta sección se presenta un análisis bivariado de las variables seleccionadas en relación con nuestra variable objetivo, *Exam_Score*. Este análisis incluye visualizaciones scatterplots, boxplots y un heatmap:

Relación entre Previous_scores y Exam_Score:



El cálculo de la Correlación entre Previous_Scores y Exam_Score dio como resultado 0.18.

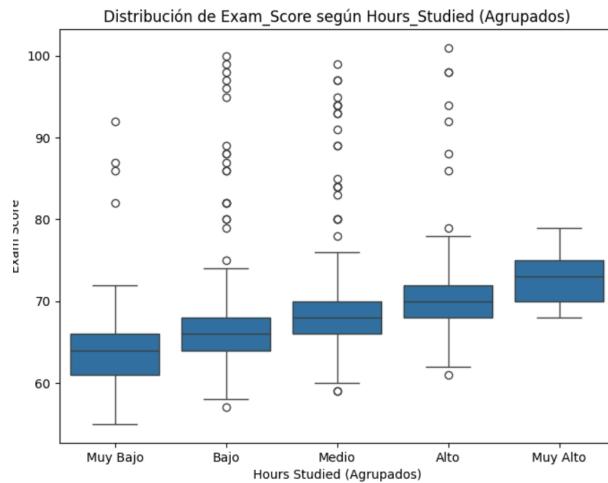
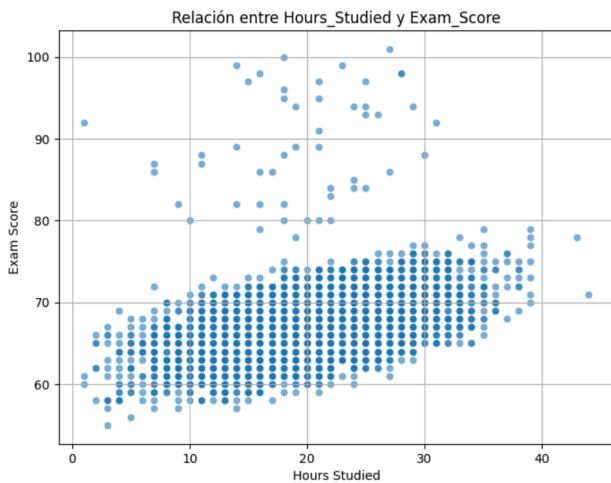
El gráfico de dispersión muestra una ligera correlación positiva entre ambas variables, aunque la relación no es fuerte.

Esto indica que los estudiantes con puntajes previos más altos tienden a tener puntajes ligeramente más altos en los exámenes, pero no es determinante.

Los boxplots revelan que a medida que los puntajes previos aumentan (de Muy Bajo a Muy Alto), la mediana y el rango intercuartil de Exam_Score también tienden a incrementarse ligeramente.

Sin embargo, hay una gran cantidad de variabilidad en cada categoría, lo que sugiere que otros factores también están influyendo en los resultados del examen.

Relación entre Hours_Studied y Exam_Score:

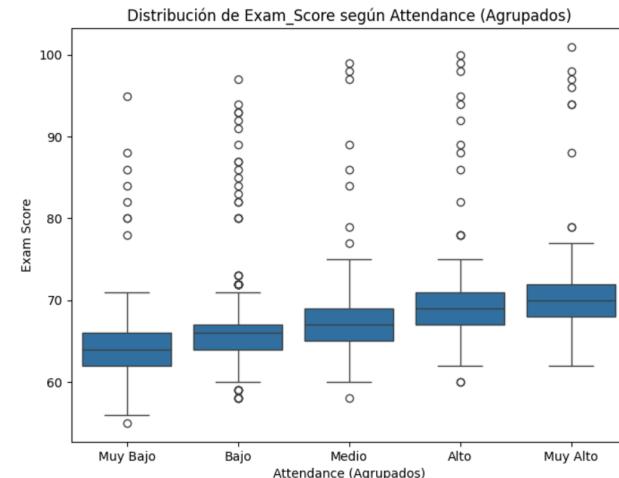
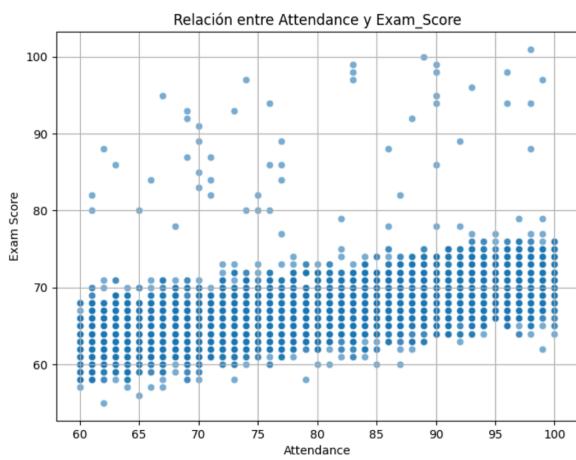


El cálculo de la Correlación entre Hours_Studied y Exam_Score dio como resultado 0.45, lo que indica una correlación positiva moderada.

El gráfico de dispersión muestra que a medida que aumentan las horas de estudio los puntajes de examen tienden a incrementarse. El patrón general sugiere que los estudiantes que dedican más tiempo al estudio logran mejores resultados en los exámenes.

Los boxplots por categorías agrupadas de horas de estudio refuerzan esta observación. A medida que se avanza de la categoría "Muy Bajo" a "Muy Alto", la mediana de Exam_Score aumenta, indicando un impacto positivo de las horas de estudio en el rendimiento. Sin embargo, los outliers dentro de cada categoría demuestran que otros factores también podrían influir en los resultados.

Relación entre Attendance y Exam_Score:

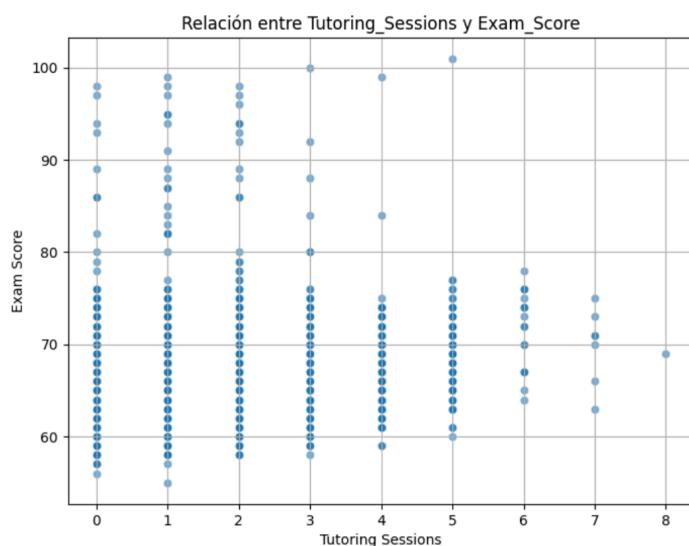


El cálculo de la Correlación entre Attendance y Exam_Score dio como resultado 0.58, lo que representa una relación positiva significativa.

El gráfico de dispersión revela que los estudiantes con mayores niveles de asistencia tienden a obtener puntajes más altos en los exámenes.

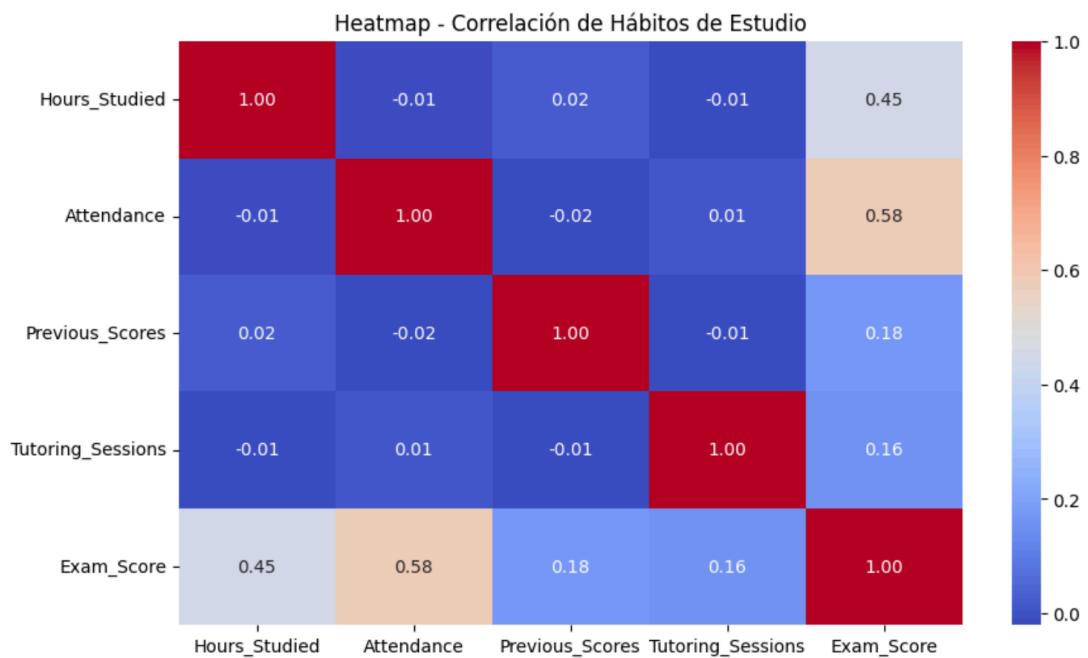
En los boxplots agrupados, se observa un aumento en la mediana de Exam_Score a medida que se pasa de la categoría "Muy Bajo" a "Muy Alto" en Attendance. Esto respalda la idea de que una mayor asistencia está asociada con mejores puntajes. Aunque hay outliers en todas las categorías.

Relación entre Tutoring_Sessions y Exam_Score:



La correlación entre Tutoring_Sessions y Exam_Score es más débil, con un valor de 0.16, lo que indica una relación positiva pero limitada.

El gráfico de dispersión muestra que, aunque la mayoría de los estudiantes tienen pocas sesiones de tutoría (entre 0 y 2), los puntajes de examen tienden a aumentar ligeramente con un mayor número de sesiones. La variabilidad dentro de cada nivel es alta, lo que sugiere que el impacto de las tutorías no es tan significativo como el de otras variables.



El heatmap ofrece una representación visual de las correlaciones entre las variables estudiadas, complementando el análisis numérico previo. La asistencia presenta la correlación más alta con los puntajes de examen (0.58), indicando que los estudiantes con mayor regularidad en clase tienden a obtener mejores resultados.

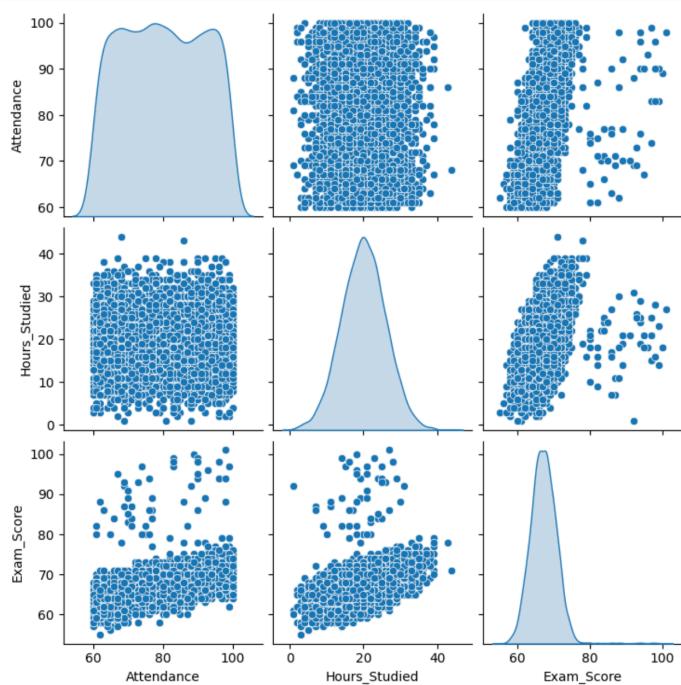
Las horas de estudio tienen una correlación moderada con el rendimiento (0.45), sugiriendo que un mayor tiempo dedicado al estudio está asociado con mejores puntajes, aunque con cierta variabilidad.

Por otra parte, los puntajes previos y las sesiones de tutoría muestran correlaciones débiles con Exam_Score (0.18 y 0.16, respectivamente). Esto indica que, aunque tienen una influencia leve, no son determinantes en los resultados de los exámenes.

En general, el análisis refuerza que la asistencia a clases y las horas de estudio son los hábitos de estudio más relevantes para el desempeño académico, mientras que los puntajes previos y las tutorías tienen un impacto más limitado.

3. Análisis Multivariado

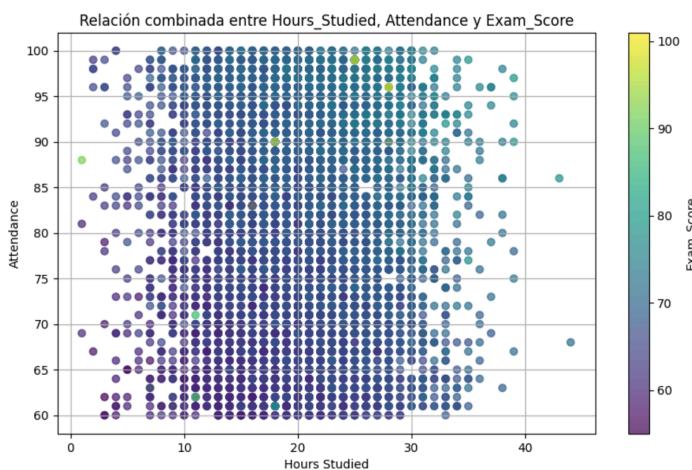
Aquí analizaremos el comportamiento de los datos tomando las variables que mayor correlación tienen con Exam_Score: Hours_Studyed y Attendance. Para esto utilizamos un gráfico pairplot y otro scatterplot



entre las horas de estudio y la asistencia a clase, lo que nos permite interpretar que estas variables influyen de manera independiente en el puntaje del examen.

El pairplot permite observar las relaciones entre las variables que:

- Existe una correlación positiva moderada entre Hours_Studied y Exam_Score, indicando que los estudiantes que dedican mas tiempo al estudio tienden a obtener puntajes mas altos en el examen. Aunque tambien existe cierta variabilidad en los resultados.
- La correlación entre Attendance y Exam_Score es mas fuerte, lo que nos muestra que una mayor asistencia a clase esta asociada con una mejor performance.
- No se identifican patrones claros



El gráfico representa cómo se relacionan las horas de estudio y la asistencia con el puntaje de examen, utilizando un mapa de color para indicar los valores de Exam_Score. No se observa un patrón claro entre las tres variables, pero sí se pueden identificar ciertas tendencias generales:

- Los estudiantes con mayores horas de estudio (por encima de 20 horas semanales) y una asistencia superior al 85% tienden a tener puntajes de examen más altos, aunque no de forma consistente.
- Hay algunos puntos atípicos destacados por colores más claros, correspondientes a puntajes altos (cerca de 100), que no parecen depender exclusivamente de niveles extremos de asistencia o estudio.
- Aunque no se identifica un patrón evidente, este gráfico refuerza la idea de que tanto las horas de estudio como la asistencia tienen un impacto moderado en el desempeño académico.

9. ¿Qué nos dicen los datos?

- La asistencia y las horas de estudio son las variables que más impactan el desempeño en los exámenes.
- Los puntajes previos también tienen influencia, pero no tan marcada como las dos variables anteriores.
- Las sesiones de tutoría tienen un impacto positivo pero limitado, probablemente debido a su baja frecuencia.
- A nivel multivariado, el desempeño en los exámenes parece depender de una combinación de esfuerzos previos (Previous_Scores), hábitos actuales (Hours_Studied), y compromiso con el aprendizaje (Attendance).
- El entorno familiar y los factores personales tienen poca influencia en el rendimiento académico.

10. Análisis de Regresión

En esta sección trabajaremos en crear un modelo de regresión que nos ayude a predecir la performance de un estudiante basándonos en algunas variables relacionadas con sus hábitos de estudio. En particular, intentaremos responder la pregunta **“¿Cómo se relacionan las horas de estudio, el puntaje de exámenes y la asistencia a clase con el puntaje del examen final?”**.

La nota de examen (exam_score) es nuestro indicador sobre del rendimiento académico de los estudiantes. Este resultado está influido por varios factores, de los cuales nos enfocaremos en los que están relacionados con los hábitos de estudio. En este caso:

- Asistencia a clase (Attendance):

La asistencia refleja la exposición del estudiante a los contenidos y a explicaciones directas del profesor. Lo que podría impactar de forma positiva en la comprensión de los conceptos y, como consecuencia, en su rendimiento.

- Horas de estudio (Hours_Studied):

Las horas dedicadas al estudio nos muestran el esfuerzo individual y la preparación del alumno para el examen.

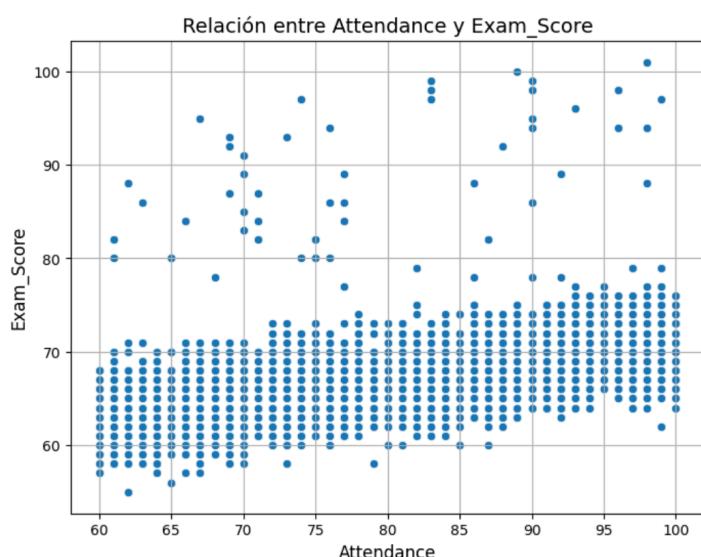
- Nota previa (Previous_scores):

Los puntajes previos indican el nivel de conocimiento base de un estudiante y pueden ser un predictor de su próxima performance.

De esta forma exam_score será nuestra variable dependiente y Attendance, Hours_Studied y Previous_scores las variables independientes que utilizaremos para el modelo de regresión.

1. Análisis bivariado

Gráficos de dispersión:

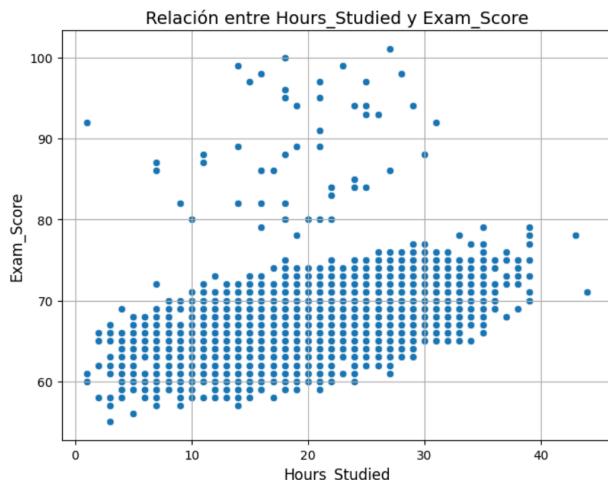


dispersión

- En este gráfico se puede observar una relación positiva entre la asistencia a clase y los puntajes de examen, lo que indica que los estudiantes con mayor asistencia tienden a obtener mejores resultados en sus exámenes.

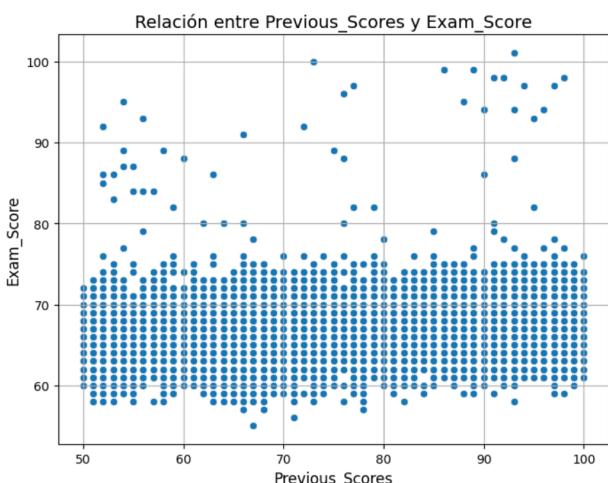
- Aunque existe una relación positiva, los puntajes no están estrictamente alineados a lo largo de una línea recta. Esto

sugiere que hay otros factores, además de la asistencia, que influyen en el desempeño de los estudiantes.



- Para el gráfico de horas de estudio y nota de examen se puede observar una tendencia general de aumento de nota a mayores horas de estudio. Por lo que mayores horas de estudio tienden a relacionarse con mejores puntajes.

- A pesar de la tendencia positiva, hay una considerable dispersión de los datos, lo que indica que las horas de estudio no son el único factor que influye en el resultado final.



- Por último, para los puntajes previos con el puntaje del examen actual, se puede observar una relación positiva mucho más débil que las relaciones anteriores analizadas.

- También se puede observar una dispersión alta en los datos.

Cálculo de correlación y P-value:

Variable	Correlación	P-Valor
Attendance	0.581072	0.000000e+00
Hours_Studied	0.445455	1.286349e-319
Previous_Scores	0.175079	1.244380e-46

- Attendance: presenta la correlación positiva más alta de las variables seleccionadas y un p-value extremadamente bajo, indicando que la correlación es estadísticamente significativa. Probablemente sea la variable independiente más influyente en el modelo.

- Hours_studied: tiene la segunda correlación positiva más significativa y un p-value muy cercano al 0, lo que muestra que esta correlación también es significativa.
- Previous_scores: es la variable con el valor de correlación más bajo. No parece ser un factor determinante en la nota pero igualmente existe una relación positiva. El p-value es muy bajo (al igual que para las otras variables), por lo que esta relación es estadísticamente significativa.

2. Modelo de Regresión

Después del análisis bivariado de la sección anterior, en este punto se optó por ajustar un modelo de regresión lineal múltiple con la siguiente ecuación:

$$\text{Exam_Score} = \beta_0 + \beta_1 \times \text{Attendance} + \beta_2 \times \text{Hours_Studied} + \beta_3 \times \text{Previous_Scores} + \epsilon$$

Como las variables presentan escalas razonables y no hay outliers extremos que distorsionen las relaciones, se ha optado por no realizar transformaciones a los datos.

El modelo ajustado nos permite identificar cómo cada variable independiente influye en el puntaje del examen. A continuación, se presentan los coeficientes estimados y su interpretación:

Variable	Coeficiente	P-Valor	Intervalo de confianza
Intercepto	41.999	0,000	[41.422, 42.576]
Attendance	0.1985	0,000	[0.193, 0.204]
Hours_Studied	0.2902	0,000	[0.280, 0.300]
Previous_Scores	0.0475	0,000	[0.043, 0.052]

- Intercepto (β_0): 41.9990
 - Cuando todas las variables independientes son cero, se espera que el puntaje del examen sea **41.999**.
- Attendance (β_1): 0.1985
 - Por cada incremento de 1 unidad en **Attendance**, se espera un aumento de **0.1985 puntos** en **Exam_Score**, manteniendo constantes las demás variables.
 - El coeficiente asociado a Attendance sugiere que la asistencia tiene un impacto significativo en el puntaje del examen. Esto refuerza la importancia de la regularidad en clase para un mejor desempeño académico.
- Hours_Studied (β_2): 0.2902

- Por cada hora adicional de estudio, se espera un incremento de **0.2902 puntos** en el puntaje del examen, manteniendo constantes las demás variables.
- Las horas de estudio muestran un impacto ligeramente mayor
- Previous_Scores (β_3): 0.0475
 - Por cada punto adicional en el puntaje previo, se espera un aumento de **0.0475 puntos** en el puntaje del examen, manteniendo constantes las demás variables.
 - Los puntajes previos también tienen una influencia positiva, aunque más leve como habíamos visto en el punto anterior.

En cuanto al desempeño global del modelo:

- **R²:** El modelo explica el 57.2% de la variabilidad total observada en los puntajes del examen, lo que indica que una proporción moderada de esta variabilidad puede ser atribuida a las variables seleccionadas.
- **R² ajustado:** También es **57.2%**, lo que indica que este valor no disminuye cuando se ajusta por el número de predictores, confirmando la relevancia de las variables incluidas

En resumen, el modelo ajustado sugiere que la asistencia a clases y las horas de estudio son los factores más determinantes para el puntaje del examen, mientras que los puntajes previos tienen un efecto menor pero aún significativo. Esto refuerza la importancia de hábitos consistentes de estudio y asistencia regular para el éxito académico.

Validación Cruzada

Se aplicó validación cruzada con 5 particiones para evaluar el desempeño del modelo en diferentes subconjuntos de los datos. Los resultados de R² variaron entre 0.418 y 0.628, con un promedio de 0.567. Esto indica que el modelo puede explicar el 56.7% de la variabilidad en los datos de entrenamiento, aunque en algunos subconjuntos el ajuste no es tan fuerte. El error cuadrático medio (MSE) promedio fue de 6.775, lo que sugiere un error moderado en las predicciones.

Evaluación en el Conjunto de Prueba

Al evaluar el modelo en los datos de prueba, se obtuvo un R² de 0.622, lo que indica que el modelo explica el 62.2% de la variabilidad en los datos no vistos. El error cuadrático medio (MSE) fue de 5.338, consistente con los resultados obtenidos en la validación cruzada. Esto sugiere que el modelo tiene un desempeño aceptable y generaliza bien a datos nuevos.

Interpretación de Resultados

El análisis de regresión lineal múltiple realizado nos permite comprender cómo las variables Attendance (asistencia a clase), Hours_Studied (horas de estudio) y Previous_Scores (resultados previos) influyen en el puntaje final del examen (Exam_Score).

Los resultados indican que:

- La **asistencia** (Attendance) tiene el mayor impacto positivo (coeficiente de 0.1985), reflejando que por cada incremento en la asistencia a clases, el puntaje aumenta significativamente. Esto valida la importancia de la regularidad en clase como un factor clave en el desempeño académico.
- Las **horas de estudio** (Hours_Studied) también muestran una relación positiva fuerte (coeficiente de 0.2902), indicando que dedicar más tiempo al estudio impacta de manera considerable el puntaje del examen.
- Los **resultados previos** (Previous_Scores) tienen un impacto positivo menor (coeficiente de 0.0475), lo que sugiere que los alumnos con mejor historial tienden a obtener mejores resultados, aunque este efecto no es tan pronunciado como la asistencia o las horas de estudio.

La **validación cruzada** confirma que el modelo generaliza bien, con un promedio de R^2 de 0.567 y MSE de 6.775 en los subconjuntos. Al probarlo en un conjunto de datos no visto, el modelo explica el 62.2% de la variabilidad en el puntaje del examen, con un MSE de 5.338. Esto sugiere que el modelo tiene un desempeño estable y es útil para hacer predicciones razonables.

Conclusiones

Este análisis proporciona información valiosa para instituciones educativas que buscan optimizar el desempeño académico de sus estudiantes. Las conclusiones clave incluyen:

- **Impacto de la asistencia y el estudio:** La asistencia regular a clases y dedicar tiempo constante al estudio son los factores más influyentes para mejorar los puntajes en exámenes. Esto resalta la importancia de fomentar políticas de asistencia obligatoria y proporcionar recursos para que los estudiantes gestionen su tiempo de estudio de manera eficiente.
- **Influencia de los resultados previos:** Aunque los puntajes previos tienen un impacto más leve, siguen siendo predictivos. Esto sugiere que los estudiantes con buen desempeño histórico mantienen una tendencia positiva, pero las mejoras significativas pueden lograrse mediante intervenciones como tutorías o técnicas de estudio efectivas.
- **Modelo predictivo sólido:** El modelo explica el 62.2% de la variabilidad en los puntajes del examen. Esto indica que, con base en estos factores, es posible anticipar el desempeño futuro de los estudiantes y aplicar medidas correctivas personalizadas para aquellos en riesgo de bajo rendimiento.

En términos prácticos, las instituciones podrían priorizar programas que incentiven la asistencia y el estudio, complementados con análisis regulares de desempeño histórico para ajustar sus estrategias pedagógicas. Este modelo sirve como una herramienta útil para identificar áreas de mejora y enfocar recursos en las intervenciones más impactantes.

11. Análisis de Regresión Logística

En esta sección, se ajustó un modelo de regresión logística para predecir la probabilidad de que un estudiante obtenga un puntaje superior a 70 en el examen. La regresión logística es un método adecuado para este tipo de problemas, ya que permite modelar la relación entre variables independientes y una variable dependiente binaria (en este caso, si el estudiante supera o no el umbral de 70 puntos).

La pregunta que se busca responder es: "¿Cuál es la probabilidad de que un estudiante obtenga un puntaje superior a 80 en el examen, en función de sus horas de estudio, asistencia a clases y puntajes previos?"

La ecuación general del modelo será:

$$\text{logit}(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 \times \text{Attendance} + \beta_2 \times \text{Hours_Studied} + \beta_3 \times \text{Previous_Scores} + \epsilon$$

Donde:

- p es la probabilidad de éxito (puntaje mayor a 70).
- $\text{logit}(p)$ es la función logit que transforma la probabilidad en una escala lineal.

1. Preparación de los datos

Para el análisis de regresión logística, se utilizaron las mismas variables independientes que en el modelo de regresión lineal múltiple: **Attendance** (asistencia a clase), **Hours_Studied** (horas de estudio) y **Previous_Scores** (puntajes previos). La variable dependiente, **Exam_Score**, se recodificó en una variable binaria (**Exam_Score_Binary**), donde:

- **1** representa un puntaje **mayor a 70**
- **0** un puntaje **igual o menor a 70**

Esta transformación permite modelar la probabilidad de que un estudiante supere el umbral de 70 puntos.

2. Ajuste del Modelo Logístico

El modelo logístico se ajustó utilizando las variables explicativas seleccionadas. Los coeficientes estimados para cada variable independiente se presentan en la siguiente tabla, junto con su error estándar, valor-z y p-valor.

Variable	Coeficiente	Error estándar	Z-valor	P-valor	Intervalo de confianza
Intercepto	41.999	823	-33.478	0,000	[-29.155, -25.930]
Attendance	0.1985	6	29.727	0,000	[0.175, 0.200]
Hours_Studied	0.2902	11	27.052	0,000	[0.263, 0.305]
Previous_Scores	0.0475	3	14.013	0,000	[0.040, 0.054]

Los p-valores indican que todos los coeficientes son estadísticamente significativos, lo que sugiere que las variables **Attendance**, **Hours_Studied** y **Previous_Scores** tienen un impacto significativo en la probabilidad de que un estudiante obtenga un puntaje superior a 70.

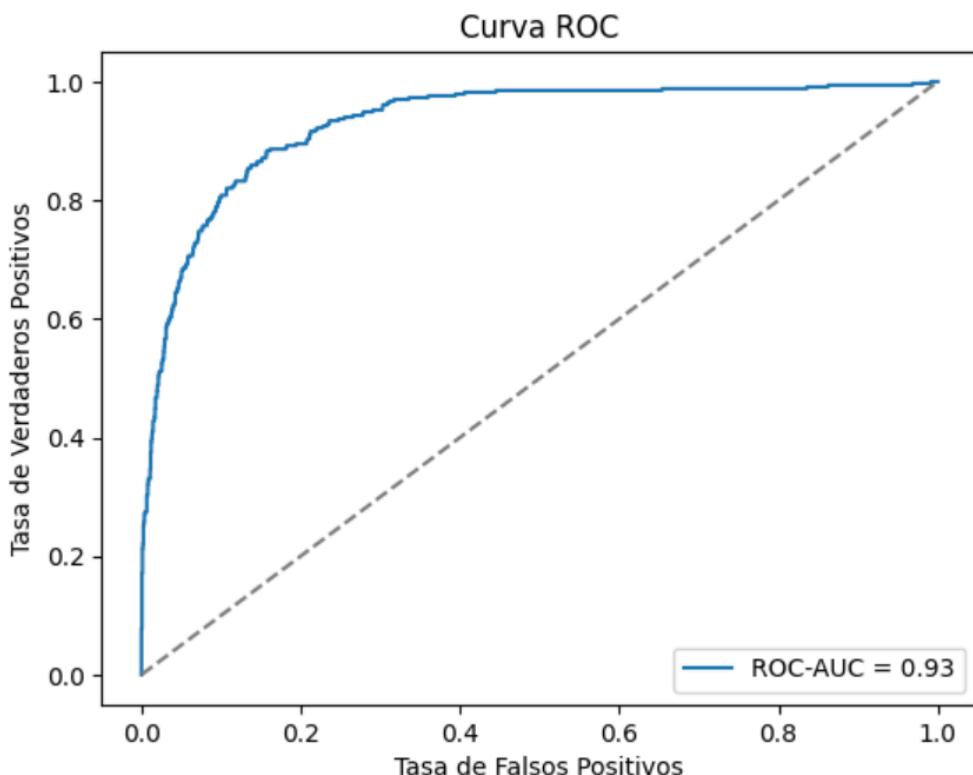
3. Interpretación de los coeficientes

Los coeficientes del modelo logístico permiten interpretar la relación entre cada variable independiente y la probabilidad de obtener un puntaje superior a 70. Por ejemplo, el coeficiente de **Attendance** (0.1877) indica que por cada unidad adicional en la asistencia a clase, el log-odds de superar el umbral de 70 puntos aumenta en 0.1877, manteniendo constantes las demás variables. Esto se traduce en un aumento en la probabilidad de éxito. De manera similar, el coeficiente de **Hours_Studied** (0.2841) sugiere que por cada hora adicional de estudio, el log-odds de obtener un puntaje superior a 70 aumenta en 0.2841.

4. Métricas de desempeño del Modelo

El modelo fue evaluado utilizando un conjunto de prueba, y se obtuvieron las siguientes métricas de desempeño: una precisión del 90%, lo que indica que el modelo clasifica correctamente el 90% de los casos, y un valor de ROC-AUC de 0.93, que refleja una excelente capacidad del modelo para distinguir entre estudiantes que superan o no el umbral de 70 puntos. La curva ROC muestra que el modelo tiene un alto poder de discriminación, lo que lo hace útil para predecir el éxito académico.

Métrica	Valor
Log-Likelihood	-1594.8
Pseudo R-squared	0.4589
Precisión	90 %
ROC-AUC	0.93



Se construyó una matriz de confusión para evaluar la precisión del modelo:

Clase Real \ Clase Predicha	No supera 70	Supera 70
No supera 70	1598	51
Supera 70	140	194

Interpretaciones:

1. Predicciones correctas:

1. Verdaderos Negativos (1598):

El modelo predijo correctamente que 1598 estudiantes no superarían los 70 puntos en el examen. Esto representa un buen desempeño en identificar correctamente la clase mayoritaria.

2. Verdaderos Positivos (194):

El modelo predijo correctamente que 194 estudiantes superarían los 70 puntos. Esto indica que el modelo tiene capacidad para identificar casos positivos a pesar del desbalance de clases.

2. Errores del Modelo:

1. Falsos Positivos (51):

El modelo clasificó erróneamente a 51 estudiantes como si hubieran superado los 70 puntos, cuando en realidad no lo hicieron.

2. Falsos Negativos (140):

El modelo no identificó correctamente a 140 estudiantes que superaron los 70 puntos. Este tipo de error puede ser más relevante, ya que implica perder la oportunidad de identificar y posiblemente brindar apoyo adicional a estudiantes con alto potencial.

3. Desempeño General:

Aunque hay más falsos negativos que falsos positivos, esto es esperable dado el desbalance de clases. Sin embargo, podría ser un área de mejora si el objetivo principal es maximizar la identificación de estudiantes con alto puntaje.

El modelo muestra una capacidad adecuada para diferenciar entre estudiantes que superan y no superan el puntaje de 70, con buenos resultados en ambas clases. Sin embargo, en un contexto práctico, puede ser importante considerar estrategias para reducir los falsos negativos, como ajustar el umbral de clasificación o incluir más variables predictoras.

4. Conclusiones

El modelo de regresión logística sugiere que la asistencia a clases y las horas de estudio son los factores más importantes para predecir la probabilidad de que un estudiante obtenga un puntaje superior a 70. Aunque los puntajes previos también tienen un impacto, este es menor en comparación con la asistencia y las horas de estudio. Estos resultados pueden ayudar a las instituciones educativas a priorizar estrategias que fomenten la asistencia y técnicas de estudio eficaces.

Este modelo puede usarse para identificar estudiantes en riesgo de bajo desempeño y diseñar intervenciones específicas, como tutorías o programas de apoyo, enfocándose en mejorar su asistencia y hábitos de estudio.



Universidad
Alfonso X el Sabio

G R A C I A S