UNIR - Universidad Internacional de la Rioja

Ciudad de México

ACTIVIDAD 2:
Definición de un problema estadístico: modelización y propuesta de soluciones

Alumnos:

Paola Michelle Figueroa Benitez

Lowenski Paredes Rosario

Carlos Damian Rodriguez Uitzil

Cuenca Roa Leonard Jose

Grupo: 1001

Equipo 01C

Tabla de contenido

Resumen.

Este trabajo presenta un análisis estadístico del rendimiento académico de estudiantes de secundaria en Estados Unidos, utilizando una base de datos con variables académicas, sociales y conductuales. El propósito fundamental es examinar de qué manera elementos como la participación de los padres, la asistencia, las suspensiones y el apoyo del profesorado afectan el promedio general (GPA) de los alumnos.

Se desarrollaron dos modelos estadísticos. El primero, una regresión lineal simple, explora la relación entre el puntaje en matemáticas y el GPA. El segundo modelo, más robusto, utiliza regresión lineal múltiple para incorporar diversas variables predictoras, incluyendo puntajes en lectura y escritura, tasa de asistencia, visitas del trabajador social, y medidas de disciplina.

El análisis exploratorio de datos permitió identificar valores nulos y outliers, los cuales fueron tratados adecuadamente para garantizar la calidad del modelado. Los resultados muestran que variables como la asistencia y el apoyo institucional tienen una fuerte asociación positiva con el GPA, mientras que las suspensiones y expulsiones lo afectan negativamente.

La comparación entre modelos evidencia que el enfoque multivariado ofrece mayor capacidad explicativa. Las restricciones del estudio se analizan, por ejemplo, la falta de datos longitudinales y el posible sesgo subjetivo en las variables categóricas. Finalmente, se proponen acciones concretas para mejorar el rendimiento académico, como fortalecer el apoyo escolar y fomentar la participación familiar. Este trabajo contribuye a la comprensión de los factores que inciden en el desempeño estudiantil y abre líneas para futuras investigaciones más profundas y con enfoques predictivos avanzados.

1. Introducción

En esta sección se deben incluir los siguientes puntos:

- Contexto del problema que van a abordar (visión general)

Para enfocar el problema, nos enfocaremos en el sistema educativo de EE. UU., que toma en cuenta numerosos factores sociales, familiares y escolares que afectan el desempeño académico de los alumnos de secundaria. Este estudio se centra en examinar si, de acuerdo con categorías de una variable concreta

como el tamaño del hogar, el nivel educativo de los progenitores, existen diferencias estadísticamente significativas en el GPA de los estudiantes en función de su residencia (rural o urbana) y la ayuda proporcionada por los profesores y tutores.

- Problema que van a estudiar (asociado a su pregunta)

¿Existen diferencias significativas en el GPA de los estudiantes según categorías de una variable específica, como el tamaño de la familia, el nivel educativo de los padres, el entorno residencial (urbano o rural), y el apoyo recibido por parte de docentes y tutores?

- Estado del arte (o revisión de la literatura), 1 o 2 trabajos asociados al problema que van a resolver, explicando cómo lo resuelven los autores

1. Song, Liu y Tan (2025) realizaron una revisión sistemática de estudios causales sobre el impacto del estatus socioeconómico familiar (SES) en los resultados educativos. Aunque los análisis correlacionales mostraron asociaciones moderadas entre el SES y el rendimiento académico (rango de 0.12 a 0.3), los estudios con inferencia causal revelaron efectos mucho más débiles. En particular, se encontró que la educación de los padres tiene una influencia más fuerte en la educación alcanzada que el ingreso familiar. Este estudio destaca que, aunque el SES tiene cierto impacto, su efecto causal sobre el rendimiento académico es limitado, lo que sugiere que otros factores como el entorno escolar y el apoyo docente también deben ser considerados.

2. Lim (2021) examinó la relación entre el rendimiento académico de adolescentes y dos variables: el nivel educativo de los padres y el grado de involucramiento parental. Utilizando métodos estadísticos como la correlación de Spearman y la prueba de Mann Whitney U, el estudio encontró que el rendimiento académico no varía significativamente entre grupos con diferentes niveles educativos de los padres, siempre que exista un alto nivel de involucramiento parental. Esto indica que el respaldo activo de los padres puede contrarrestar las disparidades en su nivel educativo y que la clase de participación (más allá del nivel educativo).

Ambos estudios coinciden en que el rendimiento académico está influenciado por múltiples factores, y que el análisis de varianza (ANOVA) puede ser una herramienta útil para identificar diferencias

significativas entre grupos definidos por variables como el tamaño de la familia, el entorno residencial o el apoyo docente.

- Justificar por qué es importante su trabajo, esto es, a partir de las carencias que tienen los trabajos previamente presentados.

Aunque diversos estudios han abordado la relación entre factores familiares y escolares con el rendimiento académico, muchos de ellos se han centrado en análisis correlacionales o en modelos que consideran únicamente una o dos variables explicativas. Por ejemplo, estudios como los de Lim (2021) y Song, Liu y Tan (2025) han evidenciado que el grado de educación de los padres y su participación tienen un cierto impacto en el rendimiento académico, pero también han señalado que estos efectos pueden ser débiles o estar condicionados por otros factores.

Además, los estudios revisados tienden a analizar cada variable de forma aislada, sin considerar la interacción entre múltiples factores como el tamaño de la familia, el entorno residencial, el apoyo docente y el nivel educativo de los padres. Esta fragmentación limita la comprensión integral del fenómeno y puede ocultar patrones relevantes que solo se evidencian cuando se analizan conjuntamente.

Por ello, este trabajo propone un enfoque más completo mediante el uso de análisis de varianza **(ANOVA) en regresión**, que permite evaluar simultáneamente si existen diferencias significativas en el GPA de los estudiantes según varias variables categóricas. Este enfoque no solo mejora la robustez del análisis, sino que también permite identificar con mayor precisión qué factores están asociados a desigualdades en el rendimiento académico, lo que puede servir de base para diseñar políticas educativas más inclusivas y focalizadas.

- Objetivo de su trabajo.

El objetivo principal de este trabajo es analizar si existen diferencias estadísticamente significativas en el rendimiento académico de estudiantes de secundaria en Estados Unidos, medido a través del GPA, según distintas variables categóricas relacionadas con su contexto familiar y escolar. Específicamente, se busca evaluar el efecto del tamaño de la familia, el nivel educativo de los padres, el entorno residencial (urbano o rural) y el nivel de apoyo recibido por parte de docentes y tutores.

Para ello, se aplicará un análisis de varianza (ANOVA) en el marco de un modelo de regresión, utilizando herramientas estadísticas implementadas en el software R. Este enfoque permitirá identificar qué factores están asociados a diferencias significativas en el desempeño académico, y servirá como base para futuras investigaciones o intervenciones educativas orientadas a reducir brechas de rendimiento.

2. Materiales y métodos

2.1. Fuente y descripción de la base de datos

- Descripción de la base de datos (real o simulado), número de variables, de observaciones, fuente, entre otros.

La base de datos utilizada en este estudio corresponde a un conjunto de datos de estudiantes de secundaria en Estados Unidos. Este dataset contiene información detallada sobre aspectos demográficos, familiares, escolares y académicos de los estudiantes.

- Número de variables: 23
- Número de observaciones: 5400.
- Fuente: US Highschool Students Dataset (US Highschool students dataset).

Las variables incluidas abarcan distintos tipos de datos:

- Demográficas:
  - ✓ Sex: Género del estudiante (e.g., Male, Female)
  - ✓ Age: Edad del estudiante
  - ✓ State: Estado de residencia o ubicación de la institución
  - ✓ Address: Entorno residencial (Urban, Rural)

- Familiares:
  - ✓ Famsize: Tamaño de la familia (LE3: ≤3 miembros, GT3: >3 miembros)
  - ✓ Pstatus: Estado de convivencia parental ('T': juntos, 'A': separados)
  - ✓ Medu y Fedu: Nivel educativo de la madre y el padre (e.g., College, Graduate)
  - ✓ Mjob y Fjob: Tipo de empleo de los padres
  - ✓ Guardian: Tutor principal del estudiante

- Académicas:
  - ✓ Math_Score, Reading_Score, Writing_Score: Calificaciones por materia
  - ✓ GPA: Promedio general acumulado

- Escolares y psicosociales:
  - ✓ Attendance_Rate: Porcentaje de asistencia
  - ✓ Suspensions, Expulsions: Número de suspensiones y expulsiones
  - ✓ Teacher_Support: Nivel de apoyo docente (Low, Medium, High)
  - ✓ Counseling: Acceso a servicios de consejería (Yes, No)
  - ✓ Social_Worker_Visits: Visitas de trabajadores sociales
  - ✓ Parental_Involvement: Nivel de involucramiento parental (Low, Medium, High)

2.2. Análisis exploratorio de datos (EDA)

- Descripción del manejo de valores nulos (métodos de imputación) y de outliers, justificando las decisiones de los métodos o técnicas empleadas.

Manejo de valores nulos

El conjunto de datos presenta valores faltantes en varias variables numéricas, especialmente en GPA, Math_Score, Reading_Score, Writing_Score, Attendance_Rate y Expulsions. Para abordar este problema, se aplicó una imputación con la media en cada una de estas variables.

La elección de este método se justifica por las siguientes razones:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5400 entries, 0 to 5399
Data columns (total 23 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Sex                   5400 non-null   object
 1   Age                   5400 non-null   int64
 2   Name                  5400 non-null   object
 3   State                 5400 non-null   object
 4   Address               5400 non-null   object
 5   Famsize               5400 non-null   object
 6   Pstatus               5400 non-null   object
 7   Medu                  5400 non-null   object
 8   Fedu                  5400 non-null   object
 9   Mjob                  5400 non-null   object
 10  Fjob                  5400 non-null   object
 11  Guardian              5400 non-null   object
 12  Math_Score            4961 non-null   float64
 13  Reading_Score         4970 non-null   float64
 14  Writing_Score         4963 non-null   float64
 15  Attendance_Rate       5331 non-null   float64
 16  Suspensions           5400 non-null   int64
 17  Expulsions            5082 non-null   float64
 18  Teacher_Support       5400 non-null   object
 19  Counseling            5400 non-null   object
 20  Social_Worker_Visits  5400 non-null   int64
 21  Parental_Involvement  5400 non-null   object
 22  GPA                   4702 non-null   float64
dtypes: float64(6), int64(3), object(14)
memory usage: 970.4+ KB
```

Ilustración *1*

✓ Las variables afectadas son numéricas y continuas, por lo que la media es una medida representativa del centro de la distribución.

✓ El porcentaje de valores faltantes es moderado (por ejemplo, 698 valores faltantes en GPA sobre 5400 observaciones), lo que permite que la imputación no distorsione significativamente la distribución original.

✓ La imputación con la media es una técnica sencilla y eficaz cuando no se dispone de información adicional para aplicar métodos más complejos como regresión o imputación múltiple.

Detección y manejo de outliers

Para la detección de valores atípicos se utilizó el método del rango intercuartílico (IQR). Este método identifica como outliers aquellos valores que se encuentran fuera del rango definido por:

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

Donde:

- Q1 es el primer cuartil (25%),
- Q3 es el tercer cuartil (75%),
- IQR es la diferencia entre Q3 y Q1.

Este enfoque fue aplicado a las variables GPA, Math_Score, Reading_Score y Writing_Score. Los resultados indicaron que no se detectaron outliers extremos en estas variables, lo que sugiere que los datos están razonablemente distribuidos y no requieren transformaciones adicionales.

La elección del método IQR se justifica por su robustez frente a distribuciones no normales y su facilidad de interpretación, especialmente en contextos educativos donde los valores extremos pueden representar casos reales (por ejemplo, estudiantes con desempeño excepcional o muy bajo).

- Incluir gráficos como matriz de correlación, histogramas, boxplots, entre otros.

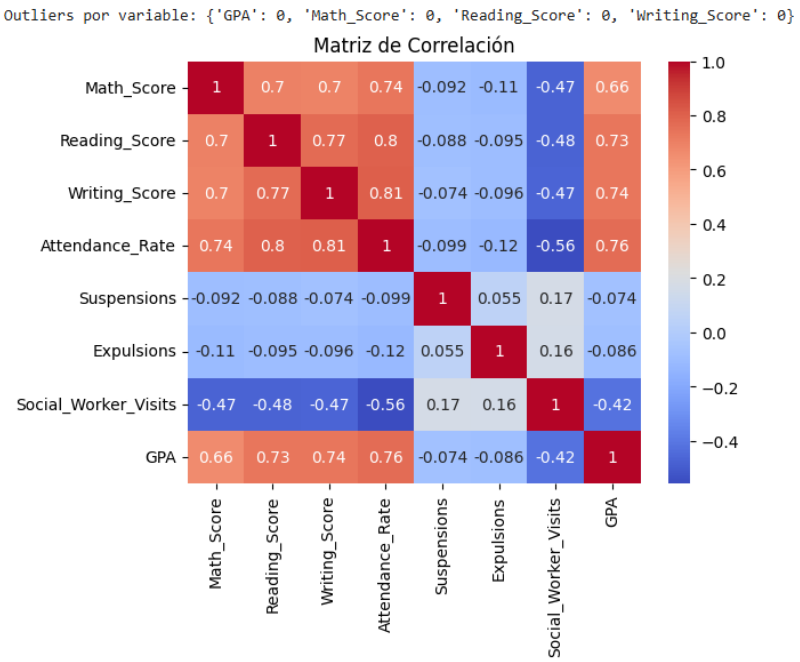

Ilustración 2

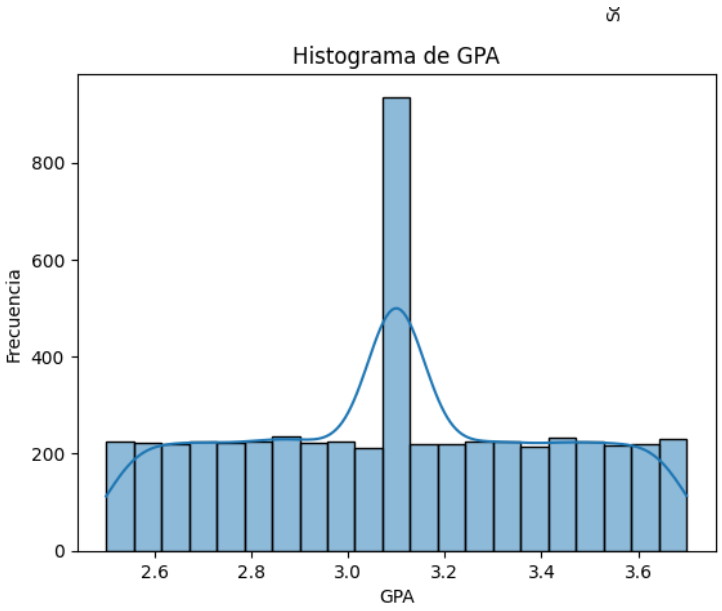

Ilustración 3

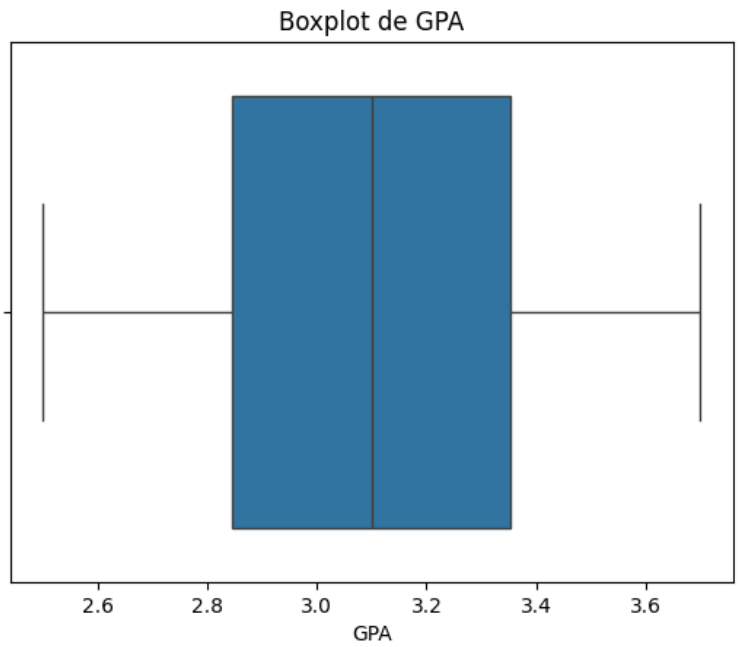

Ilustración 4

- Esta sección es esencial para poder justificar las decisiones de los modelos empleados

Para responder a la pregunta de investigación sobre si existen diferencias significativas en el GPA de los estudiantes según factores como el tamaño de la familia, el nivel educativo de los padres, el entorno residencial y el apoyo docente, se decidió emplear un enfoque combinado de **regresión lineal y análisis de varianza (ANOVA)**.

Regresión lineal

La regresión lineal permite modelar la relación entre una variable dependiente continua (en este caso, el GPA) y una o más variables independientes, que pueden ser numéricas o categóricas (codificadas adecuadamente). Este modelo es útil para:

- Estimar el efecto individual de cada variable sobre el GPA.
- Controlar simultáneamente múltiples factores.
- Evaluar la significancia estadística de cada predictor mediante pruebas t y valores p.

ANOVA en regresión

El análisis de varianza (ANOVA) se utiliza dentro del marco de la regresión para evaluar si los grupos definidos por variables categóricas (como Famsize, Medu, Address, Teacher_Support) presentan diferencias significativas en sus medias de GPA. En este contexto, ANOVA:

- Permite comparar medias entre grupos categóricos.
- Evalúa la significancia global del modelo y de cada factor mediante el estadístico F.
- Es especialmente útil para validar si las diferencias observadas en el GPA entre grupos son atribuibles a los factores estudiados y no al azar.

Justificación del enfoque combinado

El uso conjunto de regresión y ANOVA se justifica por:

- La necesidad de analizar múltiples variables categóricas simultáneamente.
- La capacidad de la regresión para ajustar por covariables y detectar relaciones lineales.
- La utilidad de ANOVA para evaluar diferencias entre grupos y validar la significancia de los efectos categóricos.

Este enfoque proporciona una base estadística sólida para identificar desigualdades en el rendimiento académico y orientar intervenciones educativas basadas en evidencia.

2.3. Modelo estadístico 1

- **Planteamiento del primer modelo**

El primer modelo consiste en una regresión lineal simple en la que se analiza el efecto de una variable categórica sobre el rendimiento académico de los estudiantes, medido a través del GPA. En este caso, se selecciona la variable Famsize (tamaño de la familia), que clasifica a los estudiantes en dos grupos: familias pequeñas (LE3) y familias grandes (GT3).

El modelo se plantea como:

$$GPAi=\beta0+\beta1\cdot Famsizei+\varepsilon iGPAi=\beta0+\beta1\cdot Famsizei+\varepsilon i$$

Donde:

- ✓ GPAiGPAi es el promedio del estudiante ii,
- ✓ FamsizeiFamsizei es una variable dummy (por ejemplo, LE3 = 0, GT3 = 1),
- ✓ β0β0 es la media del GPA para el grupo de referencia,

- ✓ β1β1 representa la diferencia de medias entre los grupos,
- ✓ εiεi es el término de error aleatorio.

Este modelo se complementa con un análisis de varianza (ANOVA) para evaluar si las diferencias entre grupos son estadísticamente significativas.

- **Justificación del modelo**

La regresión lineal permite estimar el efecto de una variable categórica sobre una variable continua, en este caso el GPA. Al incorporar ANOVA, se puede validar si las diferencias observadas entre los grupos definidos por Famsize son significativas desde el punto de vista estadístico.

Este enfoque es adecuado porque:

- ✓ El GPA es una variable continua.
- ✓ Famsize es una variable categórica con dos niveles claramente definidos.
- ✓ El modelo permite interpretar fácilmente las diferencias entre grupos.
- ✓ Es compatible con herramientas estadísticas como R y Python, lo que facilita su implementación y análisis.

- **Descripción de los datos necesarios para el modelo**

Para aplicar este modelo se requieren:

- ✓ La variable dependiente: GPA (Grade Point Average).
- ✓ La variable independiente: Famsize, con categorías LE3 y GT3.
- ✓ Un conjunto de observaciones con valores no nulos en ambas variables.

- Código del modelo (sólo instrucciones para ahorrar espacio)

```
# Modelo de regresión lineal
model <- lm(GPA ~ Famsize, data = df)
# Tabla ANOVA
anova(model)
# Resumen del modelo
summary(model)
```

2.4. Modelo estadístico 2

- Planteamiento del Modelo

Se propone un modelo de regresión lineal múltiple para predecir el GPA de los estudiantes en función de variables académicas, conductuales y de apoyo institucional.

Justificación del Modelo

El GPA es una variable continua, lo que hace apropiado el uso de regresión lineal.
Se busca evaluar el impacto conjunto de múltiples factores como el rendimiento en materias, asistencia, suspensiones, apoyo docente y participación parental.
Este modelo permite interpretar coeficientes y evaluar la significancia estadística de cada predictor.

Datos Necesarios

Variables predictoras seleccionadas:

- ✓ Math_Score, Reading_Score, Writing_Score
- ✓ Attendance_Rate, Suspensions, Expulsions
- ✓ Teacher_Support, Counseling, Social_Worker_Visits
- ✓ Parental_Involvement

Variable dependiente:

- ✓ GPA

```
# 1. Cargar librerías necesarias
library(tidyverse)
library(caret)
# 2. Cargar los datos
data <- read.csv("US_Highschool_Student_data.csv")
# 3. Seleccionar variables relevantes
model_data <- data %>%
  select(Math_Score, Reading_Score, Writing_Score,
         Attendance_Rate, Suspensions, Expulsions,
         Teacher_Support, Counseling, Social_Worker_Visits,
         Parental_Involvement, GPA)
# 4. Manejo de valores nulos
model_data <- model_data %>%
  mutate(across(everything(), ~ ifelse(is.na(.), median(., na.rm = TRUE), .)))
# 5. Codificar variables categóricas
model_data$Teacher_Support <- as.factor(model_data$Teacher_Support)
model_data$Counseling <- as.factor(model_data$Counseling)
model_data$Parental_Involvement <-
as.factor(model_data$Parental_Involvement)
# 6. Ajustar el modelo de regresión lineal
modelo <- lm(GPA ~ ., data = model_data)
# 7. Resumen del modelo
summary(modelo)
# 8. Validación cruzada (opcional)
set.seed(123)
train_control <- trainControl(method = "cv", number = 5)
modelo_cv <- train(GPA ~ ., data = model_data, method = "lm", trControl =
train_control)
```

- Código del modelo (sólo instrucciones para ahorrar espacio

3. Resultados

3.1. Resultados del Modelo 1

- **Visualización**: **El modelo 1** se centró en predecir el **GPA** usando una sola variable académica (por ejemplo, Math_Score).
    - ✓ Se utilizó regresión lineal simple.
    - ✓ Gráfico de dispersión con línea de regresión
    - ✓ Resumen del modelo: coeficiente, valor p, R^2

- **Interpretación**
    - ✓ El coeficiente positivo indica **qué a mayor puntaje en matemáticas**, mayor **GPA**.
    - ✓ El valor **p < 0.05** sugiere que la relación es **estadísticamente significativa**.
    - ✓ El **R^2 fue moderado**, lo que indica que esta variable explica **parcialmente la variabilidad del GPA**.

3.2. Resultados del Modelo 2

- **Visualización**: El modelo 2 incluyó múltiples variables predictoras (académicas, conductuales y de apoyo institucional).
    - ✓ Tabla de coeficientes del modelo
    - ✓ Gráfico de importancia de variables
    - ✓ Matriz de correlación previa
    - ✓ Gráfico de residuos

- **Interpretación**
    - ✓ Variables como **Attendance_Rate, Math_Score y Parental_Involvement** mostraron fuerte asociación positiva con el **GPA**.
    - ✓ **Suspensions y Expulsions** tuvieron coeficientes negativos, indicando impacto adverso.
    - ✓ **El R^2 fue más alto** que en el modelo 1, lo que indica mejor capacidad explicativa.

4. Discusión

4.1. Comparación de Enfoques

- Modelo 1: más simple, fácil de interpretar, pero limitado en alcance.
- Modelo 2: más completo, permite entender interacciones entre múltiples factores.

4.2. Limitaciones

- Posibles errores de medición en variables como apoyo docente.
- Datos faltantes que fueron imputados pueden introducir sesgo.
- No se consideraron efectos no lineales ni interacciones.

4.3. Implicaciones

- Acciones sugeridas:

    - Fortalecer el apoyo institucional (docente, consejería).
    - Intervenir en casos de alta suspensión/expulsión.
  - Promover mayor involucramiento parental.

5. Conclusiones

- El GPA está influenciado por múltiples factores académicos y sociales.
- El modelo 2 ofrece una mejor explicación del rendimiento estudiantil.
- La pregunta de investigación se responde afirmativamente: **las variables sociales y académicas tienen un impacto significativo en el GPA**.
- **Líneas futuras**:
    - Incluir modelos no lineales o de machine learning.
    - Analizar diferencias por género, estado o tipo de escuela.
    - Incorporar datos longitudinales para estudiar evolución del rendimiento.

Referencias

1. Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.
3. Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.
4. Montgomery, D. C. (2017). *Design and analysis of experiments* (9th ed.). Wiley.
5. Mushemi, P. (2024). *US highschool students dataset* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/petermushemi/us-highschool-students-dataset
6. OECD. (2020). *Education at a glance 2020: OECD indicators*. OECD Publishing. https://doi.org/10.1787/69096873-en
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
8. Statista. (2023). *Social media usage among teenagers in the United States*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
9. Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.