

Universidad de Costa Rica

Escuela de Matemática

Herramientas de Ciencia de Datos I

CA-0204

Tema de investigación:

“Evaluación del rendimiento académico en estudiantes de educación superior.”

Estudiantes:

Daniel Jesús Cascante Espinoza C4D841

Samuel Jiménez Zamora C4G334

Greivin Isaac Arroyo Arrieta C4C796

Alejandro Matías Bianchini Quirós C4D170

Profesor:

Joshua Cervantes Artavia

Ciudad Universitaria Rodrigo Facio

II Ciclo, 2025

## Tabla de Contenidos

<b>BITÁCORA 1.....</b>	<b>3</b>
Justificación del tema escogido:.....	3
Pregunta de investigación:.....	3
Objeto de estudio:.....	3
Conceptos que fundamentan la pregunta de investigación:.....	3
Tres teorías que fundamentan la pregunta de investigación:.....	4
Referencias:.....	6
<b>BITÁCORA 2.....</b>	<b>8</b>
Tabla con las primeras 5 filas.....	9
Tabla de información de variables cuantitativas.....	12
<b>BITÁCORA 3.....</b>	<b>18</b>
Técnicas utilizadas para responder a la pregunta de investigación:.....	18
Estadística descriptiva:.....	18
Visualización de datos:.....	18
Análisis de correlación:.....	18
Modelos estadísticos básicos:.....	19
Marco Metodológico.....	20
Resultados:.....	21
Conclusiones:.....	27
Limitaciones:.....	27

## **BITÁCORA 1**

### **Justificación del tema escogido:**

Se eligió este tema porque permite analizar múltiples variables (personales, familiares y hábitos de estudio) que influyen en el rendimiento académico. Es un tema actual y relevante, ya que entender estos factores puede ayudar a mejorar estrategias de aprendizaje.

### **Pregunta de investigación:**

¿Cuáles son los factores personales, familiares y de hábitos de estudio más influyentes en el rendimiento académico final de los estudiantes de educación superior?

### **Objeto de estudio:**

Estudiantes de educación superior incluidos en el dataset (características familiares, hábitos de estudio, entre otras) y su nota y rendimiento académico final.

### **Conceptos que fundamentan la pregunta de investigación:**

#### **1. Rendimiento académico:**

El rendimiento académico se define como la medida del logro del estudiante en áreas educativas, expresado a través de indicadores como calificaciones, promedios de notas y resultados en pruebas estandarizadas (OCDE, 2019). Este constructo es fundamental para evaluar la eficacia de los programas educativos y para identificar las necesidades de los estudiantes.

## 2. Factores personales y familiares:

Los factores personales y familiares son elementos contextuales que influyen de manera significativa en el desarrollo y en el desempeño académico del estudiante. Estos factores incluyen variables personales como la autoeficacia y la motivación, así como variables familiares como el nivel socioeconómico, el nivel educativo de los padres y el apoyo familiar (Coleman et al., 1966).

## 3. Hábitos de estudio:

Los hábitos de estudio se refieren a las estrategias autorreguladoras que los estudiantes emplean de manera sistemática para adquirir, procesar y retener información. Estos comportamientos incluyen la gestión del tiempo, la organización del espacio de estudio, la toma de notas y el uso de técnicas de memorización (Zimmerman, 2000).

### **Tres teorías que fundamentan la pregunta de investigación:**

#### 1. Teoría del Aprendizaje Autorregulado (Zimmerman, 2000) — Educativa:

La teoría del aprendizaje autorregulado plantea que los estudiantes que planifican, supervisan y evalúan activamente sus métodos de estudio obtienen mejores resultados académicos. Según Zimmerman (2000), este proceso ocurre en un ciclo de tres fases: preparación, ejecución y autorreflexión, donde los alumnos ajustan sus estrategias de acuerdo con sus metas y el feedback recibido.

#### 2. Teoría de la Autoeficacia (Bandura, 1977) — Psicológica:

La teoría de la autoeficacia sostiene que el rendimiento de un individuo depende en gran medida de la creencia en su propia capacidad para organizar y ejecutar acciones que le permitan alcanzar un objetivo (Bandura, 1977). En el ámbito académico, los estudiantes con una autoeficacia elevada tienden a perseverar más, mostrar mayor motivación y enfrentar de mejor manera los retos del aprendizaje.

### 3. Teoría del Capital Humano (Becker, 1964) — Económica/Social:

La teoría del capital humano, propuesta por Becker (1964), interpreta la educación como una inversión que incrementa las capacidades productivas de las personas. Bajo esta perspectiva, el rendimiento académico está influido por los recursos económicos, sociales y culturales disponibles para el estudiante, ya que estos determinan su acceso a materiales, apoyo familiar, calidad educativa y oportunidades de aprendizaje.

## Referencias:

- Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.  
<https://www.nber.org/books-and-chapters/human-capital-theoretical-and-empirical-analysis-special-reference-education-first-edition>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>  
<https://educational-innovation.sydney.edu.au/news/pdfs/Bandura%201977.pdf>
- B. J. Zimmerman & D. H. Schunk (1990), Self-regulated learning and academic achievement: Theoretical perspectives (pp. 1–37). Lawrence Erlbaum Associates Publishers.  
[https://www.researchgate.net/publication/243775466\\_Self-Regulated\\_Learning\\_and\\_Academic\\_Achievement\\_An\\_Overview](https://www.researchgate.net/publication/243775466_Self-Regulated_Learning_and_Academic_Achievement_An_Overview)
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of educational opportunity. U.S. Government Printing Office. <https://files.eric.ed.gov/fulltext/ED012275.pdf>
- Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2019). PISA 2018: Resultados de la prueba de lectura. Volumen I. Ediciones OCDE.  
[https://www.oecd.org/en/publications/pisa-2018-results-volume-i\\_5f07c754-en.html](https://www.oecd.org/en/publications/pisa-2018-results-volume-i_5f07c754-en.html)

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91. <https://doi.org/10.1006/ceps.1999.1016>

## **BITÁCORA 2**

Este apartado de la investigación se concentra en el estudio e interpretación de las variables que están relacionadas a los estudiantes listados en el DataFrame. Las variables incluidas exploran diferentes ámbitos, por ejemplo: características individuales, factores socioeconómicos y factores contextuales. La interpretación de las mismas permitirá desglosar su potencial influencia en los resultados académicos de la muestra.

Las características que tienen los datos del DataFrame incluyen: categóricas, binarias, integers. Los datos están concentrados en la población estudiantil de una universidad (no citada) de la Facultad de Ingeniería y la Facultad de Ciencias de la Educación en 2019. En total se consideraron 145 estudiantes y en el data frame se distribuyen por edad, de 18 a 21 años, de 22 a 25 años y de 26 en adelante.

Las variables que se incluyen en el DataFrame son sobre los estudiantes: edad, sexo, tipo de educación colegial, beca (tiene o no tiene y de cuanto porcentaje), trabaja, extracurriculares, pareja, salario (si tiene trabajo), transporte a la universidad, modalidad de vivienda, educación de la madre, educación del padre, cantidad de hermanos/as, estado marital de los papás, ocupación de la madre, ocupación del padre, horas de estudio semanales, frecuencia de lectura, asistencia a coloquios o seminarios, impacto de proyectos en su éxito, asistencia a clases, preparacion a primera ronda de exámenes, preparacion a segunda ronda de exámenes, toma notas en clases, pone atención en clase, discutir aumenta interés en el curso, aula invertida, ponderado del



semestre anterior, ponderado esperado para graduación, siglas del curso, nota del curso, identificación personal.

En el análisis de datos, los valores faltantes (NA o null) son aquellos registros donde no se dispone de información, y los outliers son observaciones que se encuentran muy alejadas del rango esperado de la variable y pueden afectar los resultados del análisis.

### Tabla con las primeras 5 filas

Variable*	STUDENT1	STUDENT2	STUDENT3	STUDENT4	STUDENT5
Edad del estudiante	2	2	2	1	2
Sexo	2	2	2	1	2
Tipo de Instituto de Procedencia	3	3	2	1	1
Tipo de beca	3	3	3	3	3
Trabajo adicional	1	1	2	1	2
Actividad artística o deportiva	2	2	2	2	2
Tiene pareja	2	2	2	1	1
Rango salarial	1	1	2	2	3
Medio de Transporte a la universidad	1	1	4	1	1
Tipo de alojamiento	1	1	2	2	4
Nivel educativo de la madre	1	2	2	1	3
Nivel educativo	2	3	2	2	3

del padre					
Número de hermanos	3	2	2	5	2
Estado civil de los padres	1	1	1	1	1
Ocupación de la madre	2	2	2	2	2
Ocupación del padre	5	1	1	1	4
Horas de estudio semanales	3	2	2	3	2
Frecuencia de lectura no científica	2	2	1	1	1
Frecuencia de lectura científica	2	2	2	2	1
Asistencia a seminarios	1	1	1	1	1
Percepción del impacto de proyectos	1	1	1	1	1
Asistencia a clases	1	1	1	1	1
Metodo de preparacion para exámenes parciales	1	1	1	1	2
Estrategia de preparación para exámenes parciales	1	1	1	2	1
Toma de apuntes en clases	3	3	2	3	2
Nivel de atención en clase	2	2	2	2	2
Percepción de la	1	3	1	2	2

utilidad del debate					
Percepción de la metodología clase invertida	2	2	1	1	1
Promedio de calificaciones de último semestre	1	2	2	3	2
Promedio de calificaciones esperado al graduarse	1	3	2	2	2
ID del curso	1	1	1	1	1
Calificación	1	1	1	1	1

\* : Columna agregada para mejor claridad

Fuente: Elaboración propia (columnas intercambiadas por filas para mejor estética)

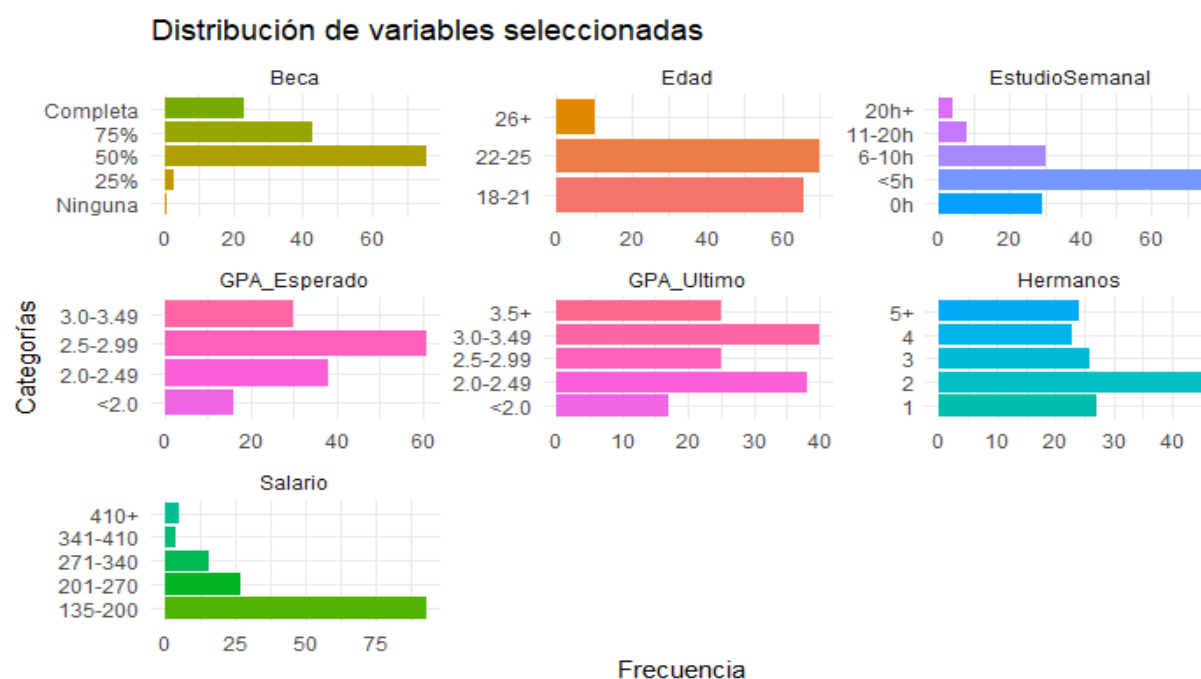
La función `summary()` en R proporciona un resumen estadístico de los datos. Para variables numéricas, incluye el valor mínimo (Min.), el primer cuartil (1st Qu.), que indica que el 25% de los datos se encuentran por debajo de este valor, la mediana (Median), la media (Mean), el tercer cuartil (3rd Qu.), que indica que el 75% de los datos se encuentran por debajo de este valor, y el valor máximo (Max.). Estos indicadores permiten tener una visión rápida de la distribución y los rangos de los datos, así como detectar posibles valores extremos o atípicos.

**Tabla de información de variables cuantitativas**

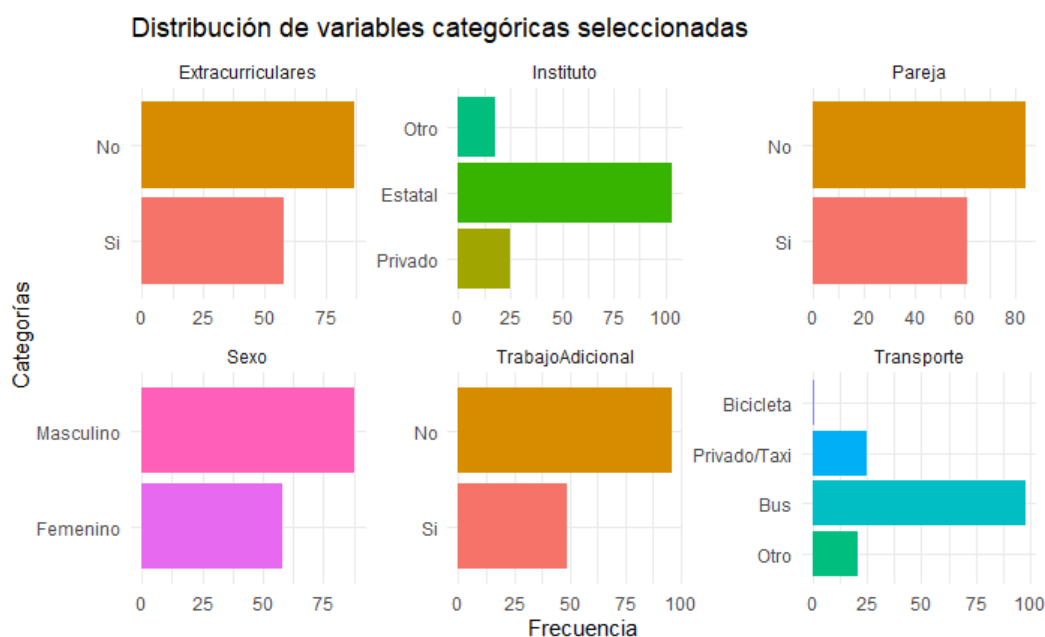
Variable	X1	X4	X8	X13	X17	X29	X30
Mínimo	1	1	1	1	1	1	1
1 cuartil	1	3	1	2	2	2	2
Mediana	2	3	1	3	2	3	3
Media	1.621	3.572	1.628	2.807	2.2	3.124	2.724
3 cuartil	2	4	2	4	3	4	3
Máximo	3	5	5	5	5	5	4

Fuente: Elaboración propia

A continuación, un gráfico que representa la distribución de las variables cuantitativas que fueron seleccionadas en la parte anterior por ser consideradas las más representativas.

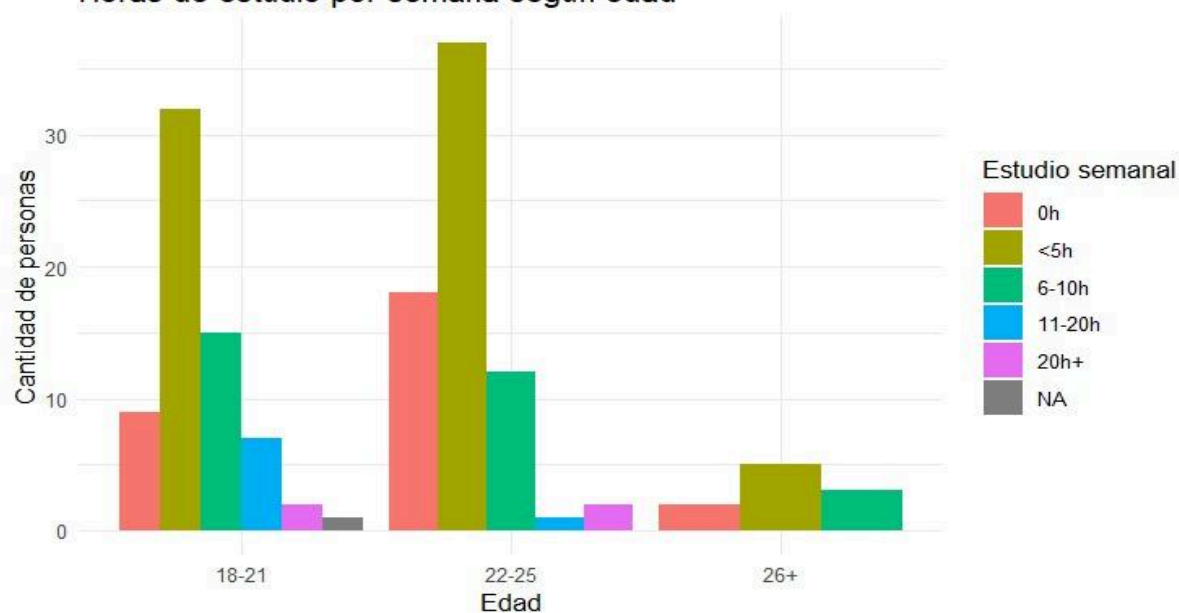


De la misma forma, también se presenta un gráfico para la distribución de las variables categóricas más relevantes.

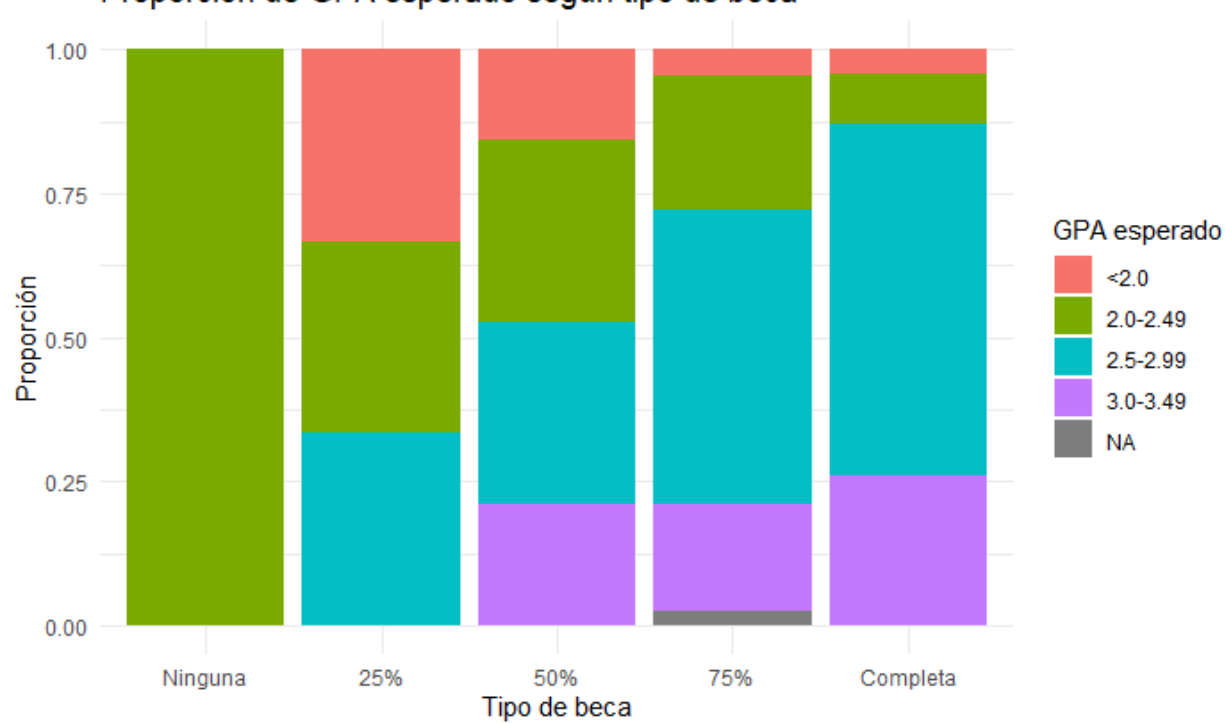


Finalmente, se realizaron dos gráficos para representar la relación entre algunas de las variables seleccionadas. El primero muestra la cantidad de horas que dedican los estudiantes al estudio independiente según su edad. El segundo muestra la relación entre la beca asignada a cada estudiante con su respectivo GPA esperado.

Horas de estudio por semana según edad



Proporción de GPA esperado según tipo de beca



En el DataFrame trabajado, todas las columnas presentan valores completos y consistentes, ya que los datos fueron recolectados siguiendo rangos definidos (por ejemplo, edad, becas, horas de estudio). Por lo tanto, no se identifican valores faltantes ni outliers en este conjunto de datos. Esto garantiza que los análisis posteriores sean confiables y no requieran limpieza adicional de estos tipos de anomalías.

Aunque en este caso no se presentan, es importante conocer las técnicas comunes para tratar estos problemas en otros conjuntos de datos:

- Valores faltantes:
  - Eliminación de filas o columnas con datos faltantes.
  - Imputación usando la media, mediana, moda o interpolación.
  - Modelos predictivos para estimar valores perdidos.
- Outliers:
  - Eliminación de los registros que se encuentren fuera de los límites definidos (por ejemplo, usando percentiles o desviaciones estándar).
  - Transformaciones de la variable (como logaritmos) para reducir el efecto de valores extremos.
  - Revisión del origen del dato para determinar si es un error o una observación válida.

Estas técnicas son útiles en casos donde los datos no siguen rangos controlados y pueden contener errores o variaciones extremas.



## Referencias:

*UCI machine learning repository.* (s/f). Uci.edu.

<https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation>

Ferrero, Rosana (2017). *¿Cómo lidiar con los datos atípicos (Outliers)?*. Máxima Formación.

<https://www.maximaformacion.es/blog-dat/como-lidiar-con-los-datos-atipicos-outliers/>

### **BITÁCORA 3**

#### **Técnicas utilizadas para responder a la pregunta de investigación:**

##### **Estadística descriptiva:**

Se utilizarán medidas de tendencia central (media, mediana) y medidas de dispersión (desviación estándar, rango intercuartílico, varianza), con el fin de caracterizar las principales variables del estudio y obtener una primera aproximación a la estructura del conjunto de datos.

##### **Visualización de datos:**

A través de gráficos generados con R (histogramas, diagramas de caja, gráficos de barras o de dispersión), se explorará la distribución de las variables y se identificarán comportamientos atípicos, tendencias generales y posibles relaciones preliminares entre ellas.

##### **Análisis de correlación:**

Utilizando variables cuantitativas, se aplicaron coeficientes de correlación como Pearson para evaluar la fuerza y dirección de las asociaciones entre pares de variables relevantes (numiqo Team; 2025). Esto permite determinar si existen relaciones significativas que puedan aportar evidencia para responder la pregunta de investigación.

**Modelos estadísticos básicos:**

Para el análisis se aplicó un modelo de regresión lineal con el propósito de examinar cómo un conjunto de variables independientes predice o influye sobre una variable dependiente, permitiendo evaluar la magnitud y dirección de dichas relaciones (Wooldridge, 2020). Este enfoque es adecuado cuando se busca identificar patrones lineales entre variables y estimar efectos bajo supuestos estadísticos claramente definidos (James et al., 2021). Para la implementación del modelo se empleó la función `lm()` del software R, la cual facilita el ajuste de modelos lineales mediante mínimos cuadrados y ofrece resúmenes detallados sobre la significancia estadística y el ajuste global del modelo.

**Evaluación de supuestos mediante gráficos de diagnóstico**

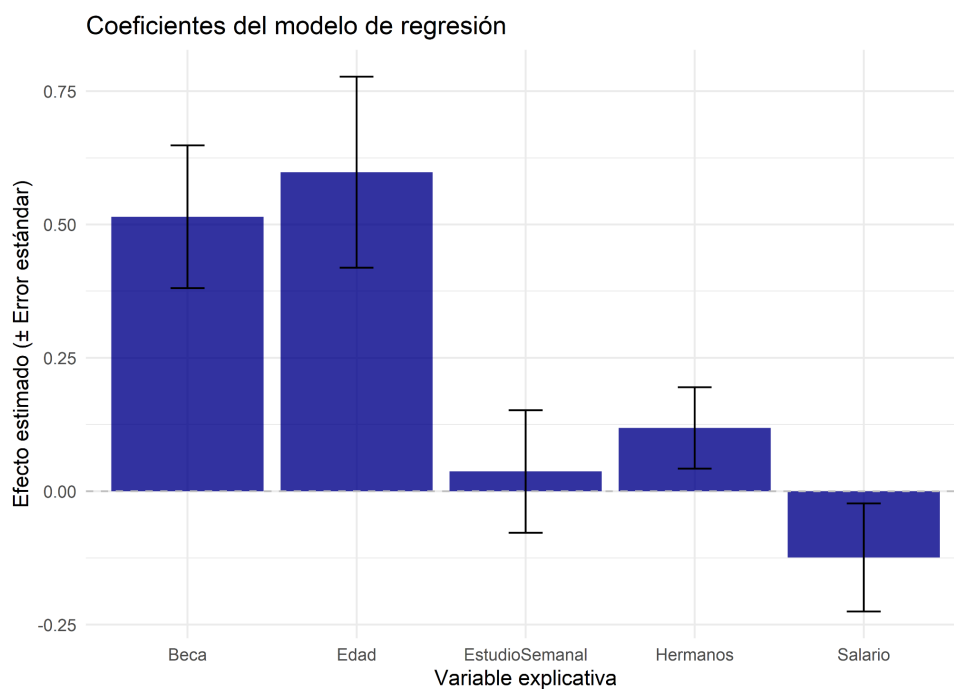
Para evaluar la calidad del modelo de regresión y verificar que cumpliera con los supuestos estadísticos necesarios, se utilizaron las técnicas de diagnóstico que ofrece la función `plot(lm)` en R. Estas herramientas permiten analizar aspectos fundamentales como la linealidad de la relación entre las variables, la normalidad de los residuos, la homocedasticidad o constancia de la varianza, y la identificación de observaciones con alta influencia. Cada uno de los cuatro gráficos generados aporta información clave para determinar si el modelo es adecuado y si las inferencias obtenidas son válidas, por lo que forman parte esencial del procedimiento de evaluación y verificación del ajuste del modelo (Fox, 2016; R Core Team, 2024).

## **Marco Metodológico**

La presente investigación adopta un enfoque cuantitativo, dado que se basa en el estudio sistemático de un conjunto de datos recopilados (Samaniego, 2022). Este trabajo se centra en medir y analizar variables numéricas, por lo que este enfoque permite examinar patrones, tendencias y relaciones dentro del dataset seleccionado con el fin de obtener evidencia objetiva que contribuya a responder la pregunta de investigación. El análisis se realiza mediante métodos implementados en el lenguaje R, lo cual garantiza una interpretación sustentada en datos.

El estudio es de tipo descriptivo, ya que busca caracterizar las principales variables del dataset y comprender su comportamiento sin manipularlas (Samaniego, 2022). El propósito es documentar y presentar posibles asociaciones presentes en los datos, ofreciendo una visión general y contextualizada del fenómeno analizado. Este tipo de estudio permite interpretar cómo se manifiestan ciertas tendencias dentro del conjunto de información disponible.

El diseño de la investigación es no experimental y documental (Samaniego, 2022). Es no experimental porque no se interviene ni se modifican los valores de las variables del dataset; únicamente se observan y analizan tal como fueron registradas originalmente. Asimismo, es documental porque el trabajo se fundamenta en la revisión y análisis de una base de datos existente, la cual constituye la fuente primaria de información. Este diseño permite desarrollar un estudio basado en evidencia empírica sin alterarla, apoyándose en herramientas computacionales para describir los patrones encontrados.

**Resultados:**

*Gráfico de coeficientes del modelo de regresión - Elaboración propia*

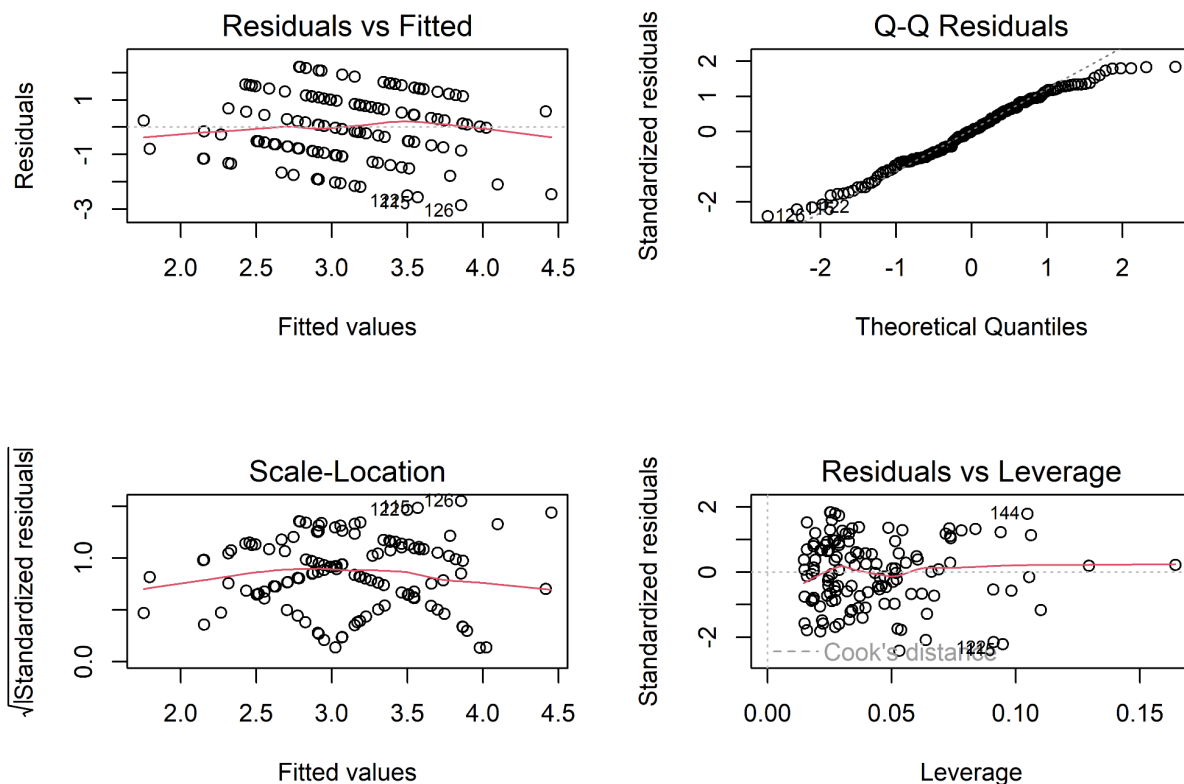
El gráfico anterior muestra los coeficientes estimados del modelo de regresión lineal para varias variables explicativas, junto con sus errores estándar representados por barras verticales. Esto permite identificar la dirección, magnitud y precisión estadística del efecto de cada variable.

1. Beca

- Coeficiente positivo, aproximadamente 0.50.
- Recibir una beca se asocia con un aumento en el rendimiento académico.
- El error estándar es moderado. Existe incertidumbre, pero el efecto es positivo.

2. Edad

- a. Coeficiente positivo, aproximadamente 0.60.
  - b. La edad tiene un impacto positivo en el rendimiento académico, asociándose fuertemente.
  - c. El error estándar es amplio: sin embargo, el intervalo se mantiene por encima de cero.
3. Estudio Semanal
- a. Coeficiente cercano a cero.
  - b. El número de horas de estudio semanal genera un efecto mínimo.
  - c. El error estándar es relativamente grande; tiene alta incertidumbre, nos dice que es estadísticamente significativo.
4. Hermanos
- a. El coeficiente es moderado, aproximándose a 0.15 positivamente.
  - b. Tener más hermanos incrementa ligeramente el rendimiento académico.
  - c. El error estándar no es grande.
5. Salario
- a. Coeficiente negativo, aproximadamente -0.10.
  - b. Un mayor salario se asocia a una disminución del rendimiento académico.
  - c. El error estándar es amplio; el efecto podría no ser significativo.



*Gráficos de diagnóstico del modelo - Elaboración propia*

### **Residuals vs Fitted**

El primer gráfico, Residuals vs Fitted, se utiliza para evaluar dos supuestos fundamentales del modelo de regresión: la linealidad y la homocedasticidad (indica que la varianza de los errores (residuos) se mantiene constante a lo largo de todos los valores de la variable independiente o de los valores ajustados del modelo). En este gráfico se espera que los residuos se dispersen de manera aleatoria alrededor de la línea horizontal, sin formar patrones

visibles. En el modelo analizado se observa una ligera curvatura en la línea de suavizamiento (LOESS), lo cual sugiere que la relación entre las variables no es lineal. Además, aunque la dispersión no muestra cambios drásticos, sí se percibe una variación un poco mayor en los extremos, lo que puede indicar una leve violación del supuesto de linealidad. A pesar de ello, el patrón general no representa una desviación severa del comportamiento esperado.

### **Q-Q Residuals**

Permite examinar si los residuos siguen una distribución normal. Este supuesto es clave para garantizar la validez de las pruebas de los intervalos del modelo. En este caso, los puntos se alinean adecuadamente sobre la recta teórica en la mayoría de los cuantiles, lo que indica que los residuos son aproximadamente normales. Sin embargo, existe un ligero desvío en los extremos, especialmente en las colas, lo que refleja una normalidad imperfecta pero aceptable para fines prácticos. Estas desviaciones no suelen afectar la interpretación del modelo cuando el resto del comportamiento es estable.

### **Scale-Location**

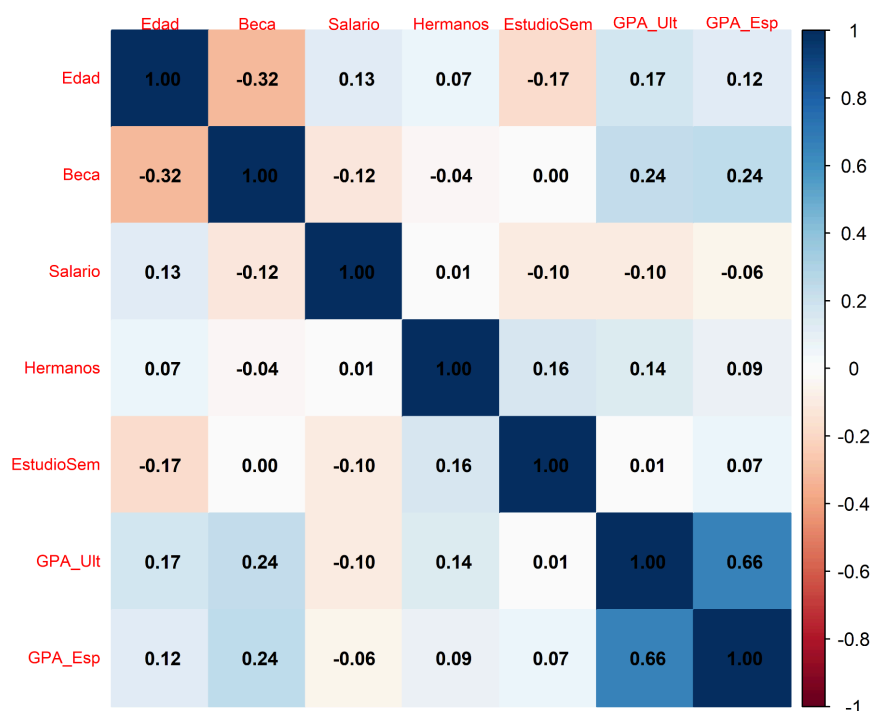
Permite evaluar el supuesto de homocedasticidad, es decir, que la varianza de los residuos sea constante a lo largo de todos los valores ajustados. En un modelo ideal, la línea roja debería mantenerse horizontal y los puntos deberían dispersarse de forma uniforme. En el gráfico analizado, se observa cierta curvatura y cambios en la dispersión, particularmente alrededor de los valores ajustados medios y altos. Esto sugiere la presencia de una ligera heterocedasticidad.



Aunque no parece grave, sí indica que la varianza de los residuos podría no ser completamente constante, lo que puede afectar la eficiencia de las estimaciones.

### **Residuals vs Leverage**

Se utiliza para identificar observaciones influyentes o puntos que podrían afectar de manera significativa la estimación de los coeficientes del modelo. Este gráfico incluye líneas de referencia basadas en la distancia de Cook (puntos que pueden afectar de manera desproporcionada los resultados del modelo; cuantifica cuánto cambiarían los coeficientes del modelo si una observación específica fuera eliminada.), las cuales ayudan a determinar si una observación tiene un nivel inusualmente alto de influencia. En este caso, la mayoría de los puntos se encuentran dentro de los límites esperados y no presentan valores excesivamente altos. Aunque aparecen algunos puntos con mayor leverage, como el identificado alrededor del número 144, su influencia no parece ser tan grande como para comprometer el ajuste general del modelo. En conjunto, este gráfico confirma que no existen observaciones extremadamente influyentes que distorsionen los resultados del modelo.



*Gráfico de correlación de variables numéricas - Elaboración propia*

El gráfico anterior expone las relaciones entre diferentes variables numéricas. Se incluye un gradiente que ayuda a visualizar lo esperado, dado que las relaciones se miden de -1 a 1. Para entender las relaciones, se debe comprender la lógica detrás de los valores que estas brindan: mientras más lejos de 0, más se relacionan entre sí; si está cerca a -1, la relación indica que si una variable sube, la otra baja, y si está cerca a 1, la idea contraria, ambas suben o ambas bajan.

A partir de esto, se puede observar que la diagonal lógicamente está en el azul más oscuro dado que es donde las variables intersecan consigo mismo, sin embargo, la variable *Beca* y la variable *Edad* son las que más se acercan a -1. Esto se puede interpretar con la explicación brindada anteriormente, si la edad sube, la beca baja y viceversa. De manera contraria, la

variable *GPA\_Ult* (GPA del último ciclo lectivo) y *GPA\_Esp* (GPA esperado del ciclo lectivo), son las que más se acercan a 1, lo cual nos indica que si el GPA del ciclo anterior es mayor, el esperado también, lo cual tiene sentido lógico. También se puede resaltar el caso de las variables *EstudioSem* (Estudio Semanal) y *Beca*, las cuales no tienen relación del todo.

### **Conclusiones:**

1. Recibir una beca se asocia de manera positiva con el aumento en el rendimiento académico de los estudiantes, sugiriendo que los apoyos económicos pueden mejorar el desempeño estudiantil.
2. El número de horas de estudio semanal muestra un efecto mínimo sobre el rendimiento académico; sin embargo, variables familiares como el número de hermanos presentan una ligera relación positiva con el desempeño académico.
3. El análisis correlacional y el modelo de regresión permitieron identificar relaciones relevantes pero no determinantes, especialmente entre GPA previo y GPA esperado, que presentan correlación alta. Sin embargo, otras variables no mostraron patrones claros, lo que indica que el rendimiento académico es un fenómeno multifactorial que no puede explicarse únicamente con las variables disponibles en el dataset.

**Limitaciones:**

Algunas de las limitaciones que se nos presentaron fueron:

1. El DataSet: Este dataset estaba muy desordenado con muchas variables que no tenían una implicación sobre nuestra pregunta, algunas de estas como la ocupación de los padres.
2. El tamaño del DataSet: Los datos provienen de una muestra pequeña (145 estudiantes) y de solo dos facultades, lo que reduce la capacidad de generalizar los resultados a toda la población universitaria.
3. Restricciones del modelo estadístico: Algunos supuestos del modelo de regresión (como la linealidad y la homocedasticidad) mostraron desviaciones leves, lo cual puede afectar la precisión del análisis y reducir la confiabilidad de algunas conclusiones.

## Referencias:

Samaniego, G. (2022) *Enfoque, tipo, diseño y método de investigación [Aclarando conceptos]*.

<https://miasasordetesis.com/enfoque-tipo-diseno-metodo-de-investigacion/>

Ortega, C. (s.f) *¿Qué es el coeficiente de correlación de Pearson?*

<https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/>

numiqo Team (2025). *numiqo: Online Statistics Calculator*.

<https://numiqo.es/tutorial/pearson-correlation>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning:*

*With applications in R (2nd ed.). Springer.*

<https://www.casact.org/sites/default/files/2022-12/James-G.-et-al.-2nd-edition-Springer-2021.pdf>

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach (7th ed.)*. Cengage

Learning. [http://lib.ysu.am/disciplines\\_bk/33b6e2aa30379171d53274830843d3f8.pdf](http://lib.ysu.am/disciplines_bk/33b6e2aa30379171d53274830843d3f8.pdf)

Fox, J. (2016). *Applied regression analysis and generalized linear models (3rd ed.)*. SAGE

Publications. <https://www.john-fox.ca/AppliedRegression/chap-25.pdf>

R Core Team. (2024). *R: A language and environment for statistical computing [Computer*

*software]*. R Foundation for Statistical Computing. <https://www.r-project.org/>