Pontificia Universidad
JAVERIANA
Colombia

Reporte 1: Análisis de variables predictivas del rendimiento académico

Juan Pablo Arias Buitrago

Sergio Pardo Hurtado

Facultad de Ciencias Básicas

033520: Análisis de Regresión

Gabriel Camilo Pérez Castañeda, Phd

Pontificia Universidad Javeriana

Facultad de Ciencias

Bogotá D.C.

13 de septiembre de 2025

1. Análisis Previo de los Datos

1.1. Análisis Inicial

A partir de la inspección inicial con la función `str(df)` aplicada sobre el conjunto de datos se obtiene la Tabla 1, donde se observa la estructura general del mismo, incluyendo el tipo de variable y algunos ejemplos de los valores que éste toma.

Variable	**Tipo**	**Ejemplo(s)**
student_id	Caracter (chr)	"S1000", "S1001", "S1002"
age	Numérico	23, 20, 21, 18
gender	Caracter (chr)	Female, Male, ...
study_hours_per_day	Numérico	0, 6.9, 1.4, 5
social_media_hours	Numérico	1.2, 2.8, 3.1, 4.4
netflix_hours	Numérico	1.1, 2.3, 1.3, 0.5
part_time_job	Caracter (chr)	No, Yes
attendance_percentage	Numérico	85, 97.3, 94.8, 71
sleep_hours	Numérico	8, 4.6, 9.2, 7.5
diet_quality	Caracter (chr)	Fair, Good, Poor
exercise_frequency	Numérico	6, 1, 4, 3
parental_education_level	Caracter (chr)	Master, High School, Bachelor, None
internet_quality	Caracter (chr)	Poor, Average, Good
mental_health_rating	Numérico	8, 1, 4, 10
extracurricular_participation	Caracter (chr)	Yes, No
exam_score	Numérico	56.2, 100, 34.3, 78.9

Tabla 1: Descripción de las variables del dataset (n = 1000, p = 16).

Observaciones:

- El dataset contiene 16 variables: 9 numéricas y 7 categóricas, varias de ellas no dicotómicas, lo que implica un esfuerzo adicional en el tratamiento para el modelo.
- Existe un equilibrio entre variables numéricas (9) y categóricas (7), lo cual permitirá explorar diferentes modelos para integrar la información cuantitativa y cualitativa del conjunto de datos.
- Algunas variables categóricas (`diet_quality`, `internet_quality`) presentan una escala ordinal implícita que deberá considerarse en el modelado.
- No se detectaron valores nulos, lo que refleja buena calidad en los datos.

1.2. Relación entre la Variable Objetivo y las Explicativas

Utilizando gráficos de dispersión (para variables numéricas) y gráficos de barras (para variables categóricas), se analiza la relación de cada variable explicativa con la variable objetivo `exam_score`.

1.2.1. exam_score vs. age

De la Figura 1 se puede interpretar que.

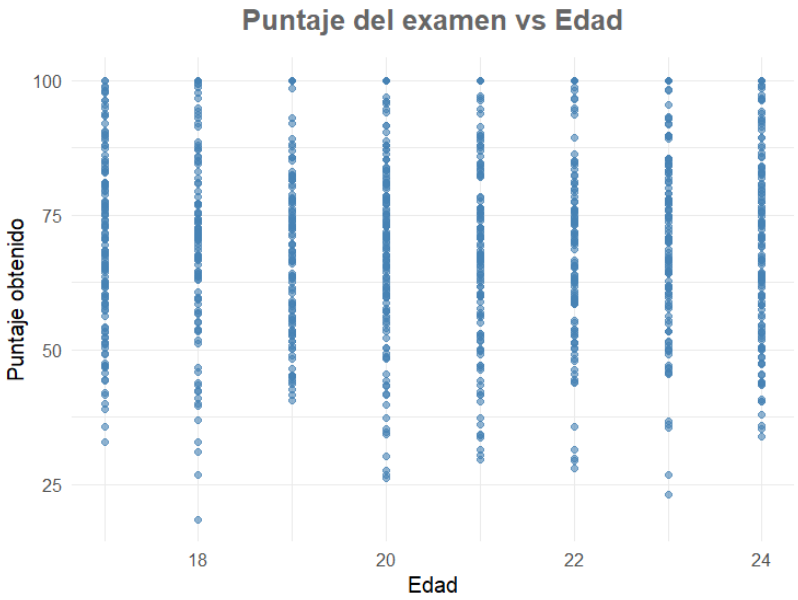

Figura 1: Relación entre `exam_score` y `age`

1.2.2. exam_score vs. study_hours_per_day

De la Figura 2 se puede interpretar que.

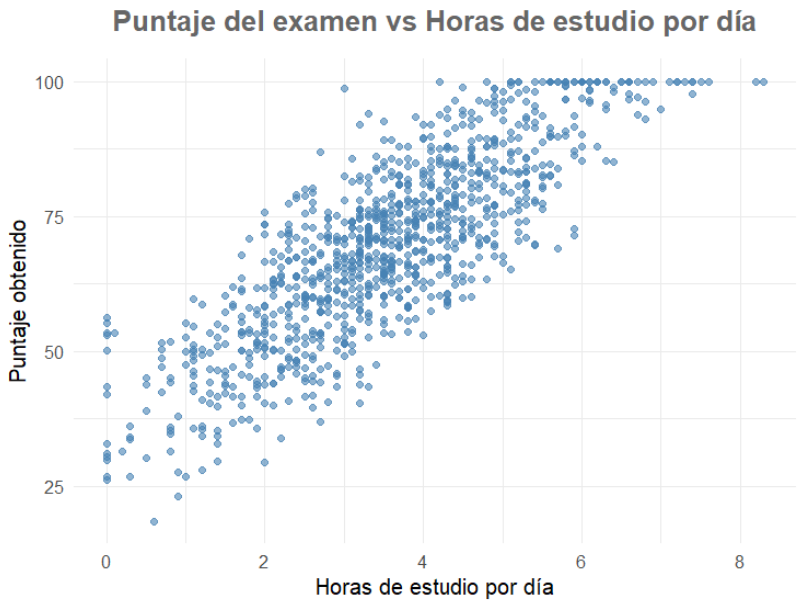

Figura 2: Relación entre `exam_score` y `study_hours_per_day`

1.2.3. exam_score vs. social_media_hours

De la Figura 3 se puede interpretar que.

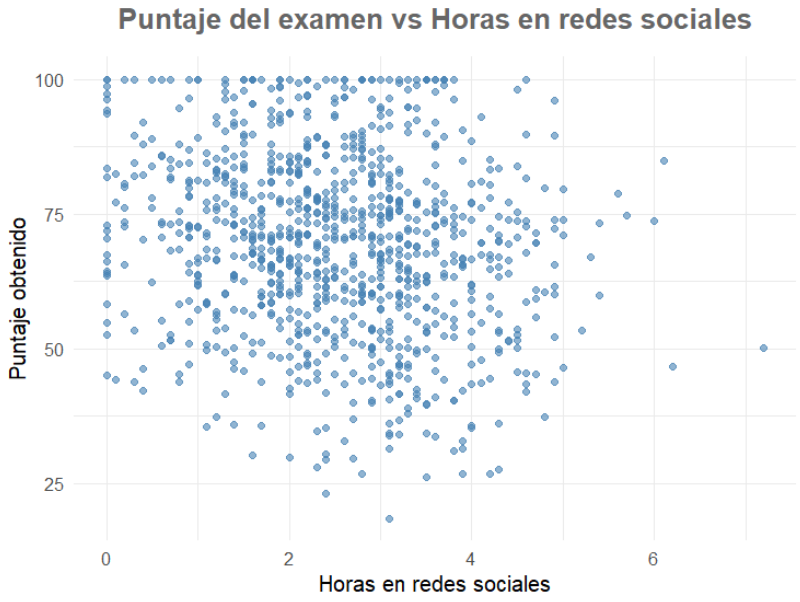

Figura 3: Relación entre `exam_score` y `social_media_hours`

1.2.4. exam_score vs. netflix_hours

De la Figura 4 se puede interpretar que.

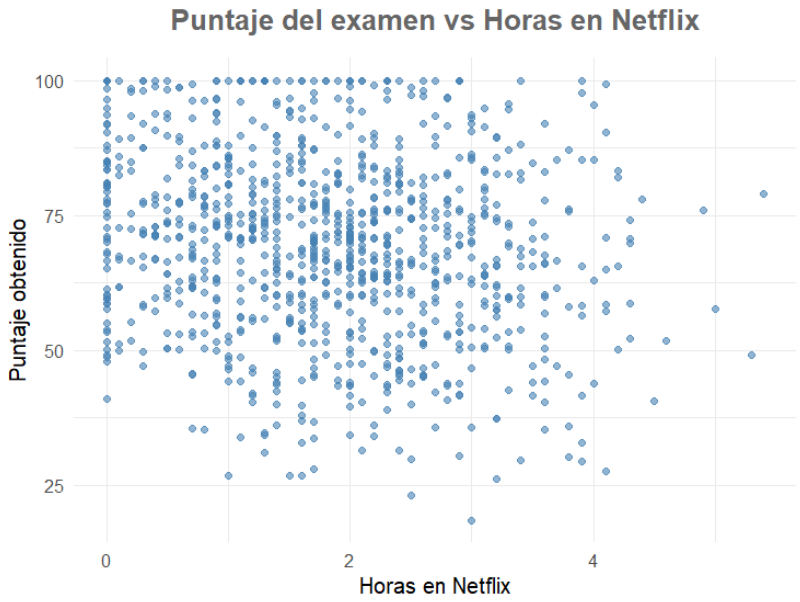

Figura 4: Relación entre `exam_score` y `netflix_hours`

1.2.5. exam_score vs. attendance_percentage

De la Figura 5 se puede interpretar que.

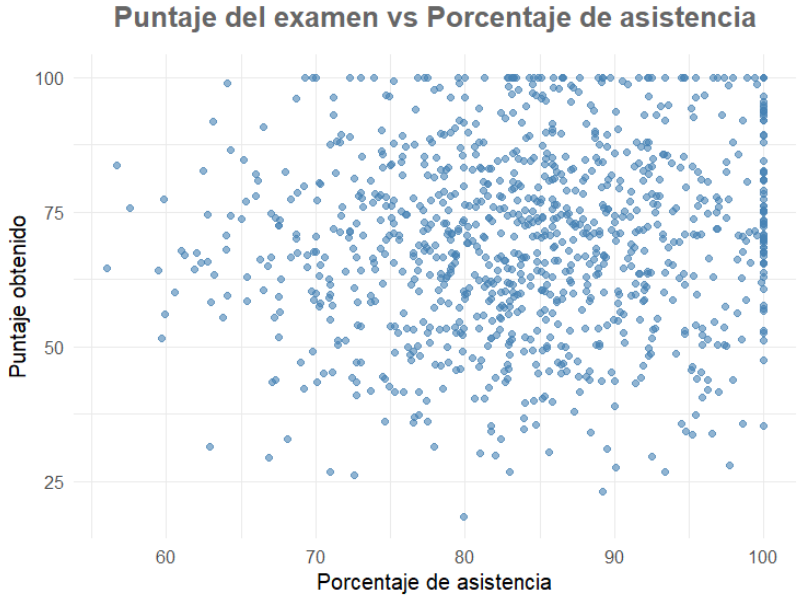

Figura 5: Relación entre `exam_score` y `attendance_percentage`

1.2.6. exam_score vs. sleep_hours

De la Figura 6 se puede interpretar que.

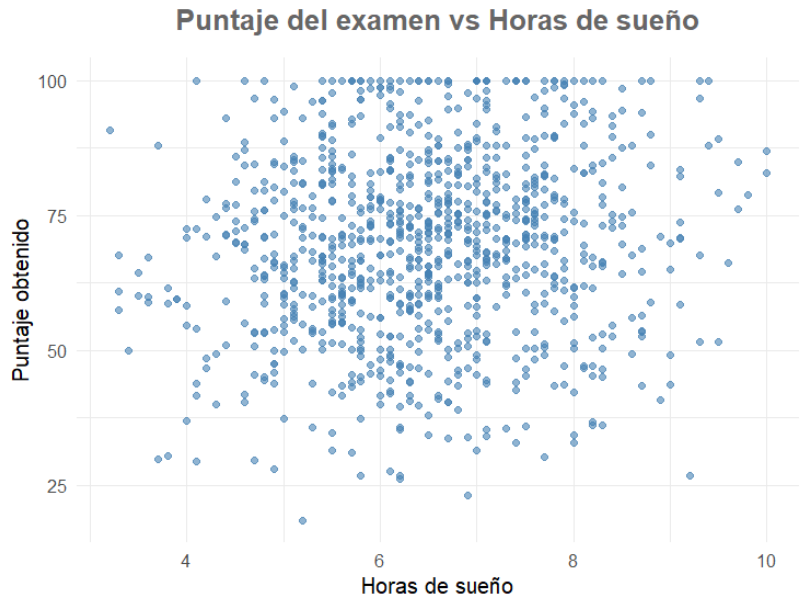

Figura 6: Relación entre `exam_score` y `sleep_hours`

1.2.7. exam_score vs. exercise_frequency

De la Figura 7 se puede interpretar que.

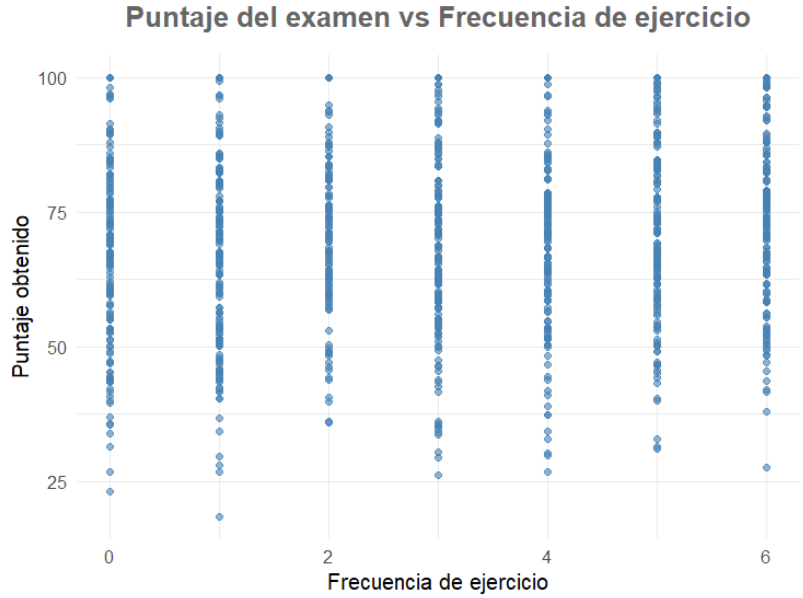

Figura 7: Relación entre `exam_score` y `exercise_frequency`

1.2.8. exam_score vs. mental_health_rating

De la Figura 8 se puede interpretar que.

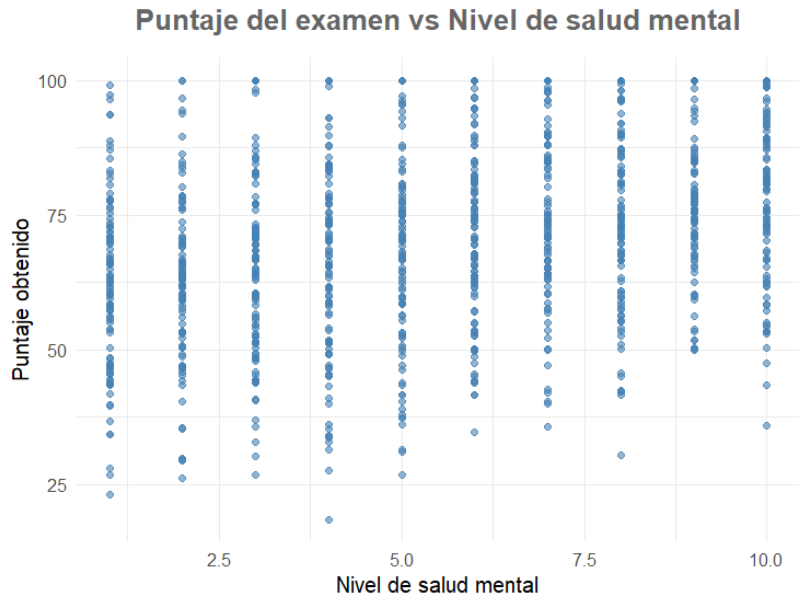

Figura 8: Relación entre `exam_score` y `mental_health_rating`

1.3. Variables Explicativas vs. Explicativas

1.4. Relaciones Lineales y No Lineales

1.5. Transformación de Variables

2. Estimación de modelos, ajuste y validación

2.1. Estimación del Modelo Completo

2.1.1. Significancia Global del Modelo

2.1.2. Significancia Individual de Parámetros

2.2. Validación de Supuestos del Modelo

2.2.1. supuesto 1

2.2.2. supuesto 2

2.2.3. supuesto 3

2.3. Eliminación de Variables No Significativas

2.4. Bondad de Ajuste entre Modelos

2.5. Elección del Modelo final

2.5.1. Interpretación de los Parámetros

2.5.2. Optimización de Parámetros (opcional)

2.6. Origen del Dataset

2.6.1. Kaggle y Diccionario de Datos

2.6.2. Código Utilizado - Repositorio