REGRESION LINEAL

Maria Roxana Humiri Ramos (227458)

July 1, 2024

1 REGRESION LINEAL

1.1 ¿Qué es la Regresión Lineal?

La regresión lineal es una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal.

1.2 Definición matemática

La relación entre una variable dependiente Y y una o más variables independientes X se define matemáticamente como:

1.2.1 Regresión Lineal Simple

Para el caso de una sola variable independiente:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Donde:

- Y es la variable dependiente.
- X es la variable independiente.
- β_0 es el intercepto.
- β_1 es la pendiente de la línea de regresión.
- ϵ es el término de error.

1.2.2 Regresión Lineal Múltiple

Para múltiples variables independientes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

1.3 Descripción del Dataset

Este CSV se descargó de KAGGLE.

Este conjunto de datos contiene información completa sobre 2392 estudiantes de secundaria, en la que se detallan sus características demográficas, hábitos de estudio, participación de los padres, actividades extracurriculares y rendimiento académico. La variable de destino, GradeClass, clasifica las calificaciones de los estudiantes en categorías distintas, lo que proporciona un conjunto de datos sólido para la investigación educativa, el modelado predictivo y el análisis estadístico.

1.3.1 Número de Datos

Son 2392 datos.

1.3.2 Número de variables

Son 15 variables.

1.3.3 Detalle de las variables

1. StudentID: Identificador único del estudiante (1001 a 3392)
2. Age: Edad del estudiante.
3. Gender: Género de los estudiantes, donde 0 representa masculino y 1 representa femenino.
4. Ethnicity: La etnia de los estudiantes, codificada de la siguiente manera:
   - 0: caucásico
   - 1: afroamericano
   - 2: asiático
   - 3: Otros
5. ParentalEducation: El nivel educativo de los padres, codificado de la siguiente manera:
   - 0: Ninguno
   - 1: Escuela secundaria
   - 2: algo de universidad
   - 3: Licenciatura
   - 4: superior
6. StudyTimeWeekly: Tiempo de estudio semanal en horas, de 0 a 20.
7. Absences: Número de ausencias durante el año escolar, comprendido entre 0 y 30.
8. Tutoring: Estado de la tutoría, donde 0 indica No y 1 indica Sí.
9. ParentalSupport: El nivel de apoyo de los padres, codificado de la siguiente manera:
   - 0: Ninguno
   - 1: Bajo
   - 2: moderado
   - 3: alto
   - 4: Muy alto
10. Extracurricular: Participación en actividades extracurriculares, donde 0 indica No y 1 indica Sí.
11. Sports: Participación en deportes, donde 0 indica No y 1 indica Sí.
12. Music: Participación en actividades musicales, donde 0 indica No y 1 indica Sí.
13. Volunteering: Participación en voluntariado, donde 0 indica No y 1 indica Sí.
14. GPA: Promedio de calificaciones en una escala de 2.0 a 4.0, influenciado por los hábitos de estudio, la participación de los padres y las actividades extracurriculares.
15. GradeClass: Clasificación de las calificaciones de los estudiantes según el GPA:
   - 0: 'A' (GPA ¿= 3,5)
   - 1: 'B' (3,0 ¡= GPA ¡ 3,5)
   - 2: 'C' (2,5 ¡= GPA ¡ 3,0)
   - 3: 'D' (2.0 ¡= GPA ¡ 2.5)
   - 4: 'F' (GPA ¡ 2.0)

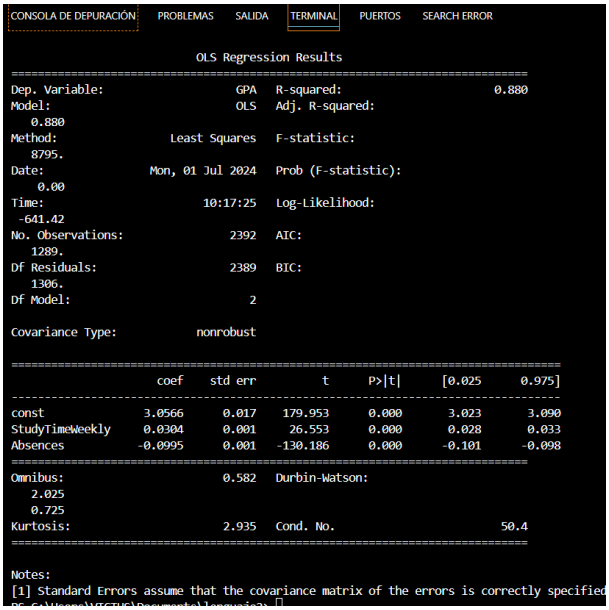
1.4 Código en Python

Como variables independientes se va utilizar: StudyTimeWeekly y Absences
Como variable dependiente: GPA

```
import pandas as pd
import statsmodels.api as sm

# Ruta del archivo CSV con el nombre exacto
file_path = 'C:/Users/VICTUS/Documents/lenguaje2/Student_performance_data _.csv'

try:
    # Cargar el archivo CSV
    data = pd.read_csv(file_path)

    # Seleccionar las variables independientes y dependiente
    X = data[['StudyTimeWeekly', 'Absences']]
    Y = data['GPA']

    # Agregar una constante a las variables independientes
    X = sm.add_constant(X)

    # Ajustar el modelo de regresión lineal
    model = sm.OLS(Y, X).fit()

    # Mostrar el resumen del modelo
    summary = model.summary()
    print(summary)

except FileNotFoundError:
    print(f"El archivo no se encontró en la ruta especificada: {file_path}")
except Exception as e:
    print(f"Ocurrió un error al intentar leer el archivo: {e}")
```

1.5 Resultados de la regresión

1.5.1 Resultado



1.5.2 Detalles

Regresión Lineal usando las variables StudyTimeWeekly y Absences para predecir el GPA:

- R-cuadrado: 0.880
- R-cuadrado ajustado: 0.880
- F-statistic: 8795
- Prob (F-statistic): 0.00

1.5.3 Interpretación de Coeficientes

- Coeficiente de StudyTimeWeekly (Horas de estudio semanal): Por cada hora adicional de estudio semanal, el GPA aumenta en 0.0304, manteniendo constantes las demás variables. Este coeficiente es estadísticamente significativo (p ¡ 0.001).
- Coeficiente de Absences (Número de ausencias): Por cada ausencia adicional, el GPA disminuye en 0.0995, manteniendo constantes las demás variables. Este coeficiente también es estadísticamente significativo (p ¡ 0.001).
- Intercepto: Cuando StudyTimeWeekly y Absences son 0, el GPA es 3.0566.
- R-cuadrado: El 88% de la variabilidad en el GPA puede explicarse por las variables StudyTimeWeekly y Absences.

1.6 Métricas de Evaluación

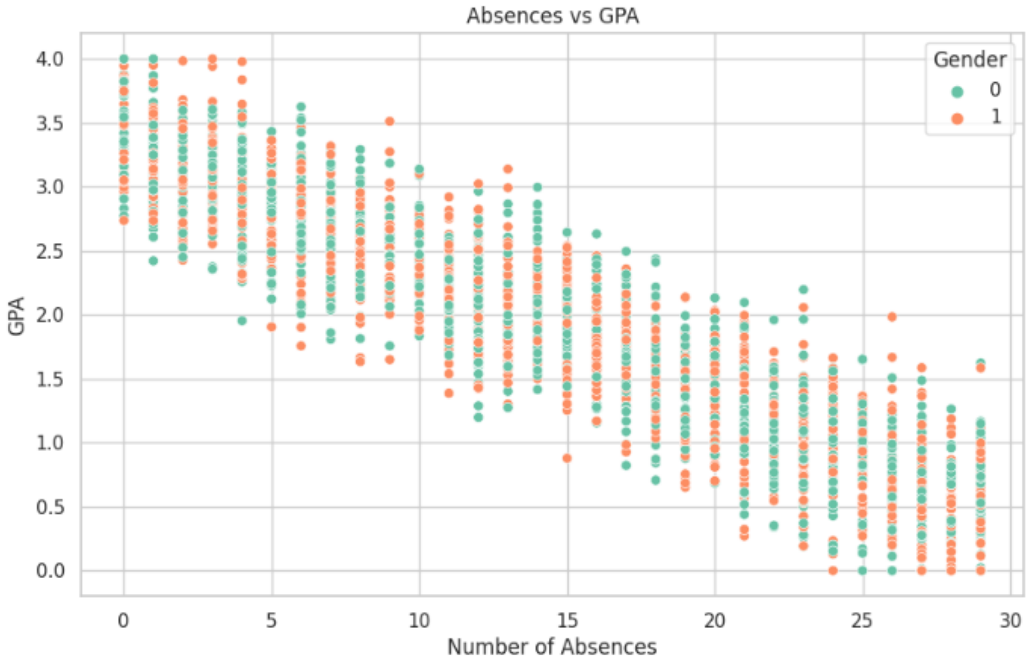


Figure 1: Métricas de evaluación.

1.7 Gráficos

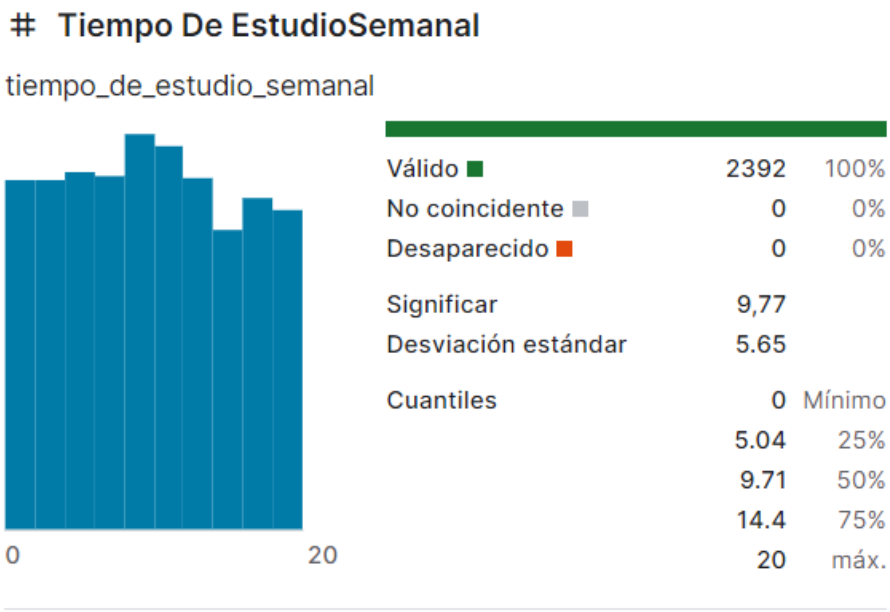
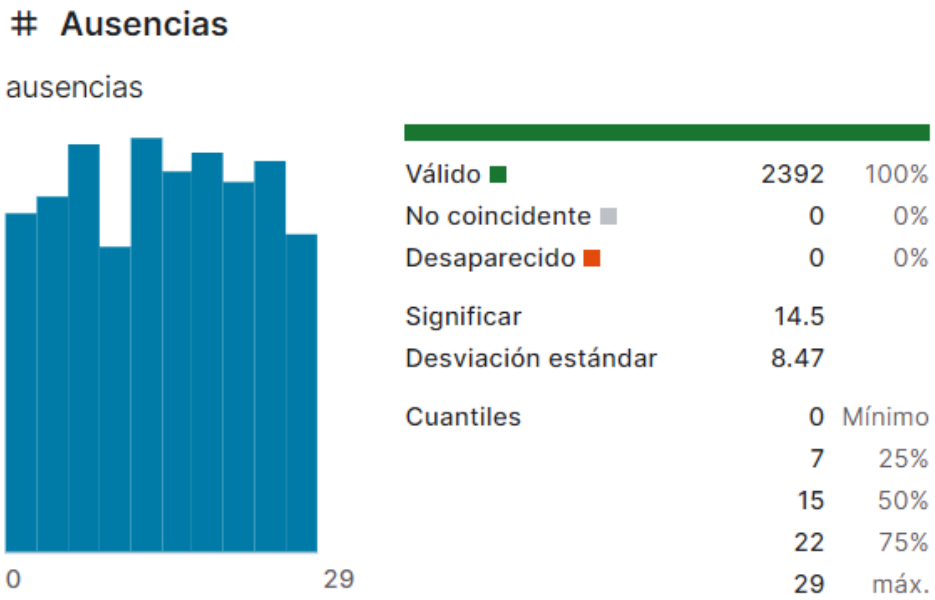
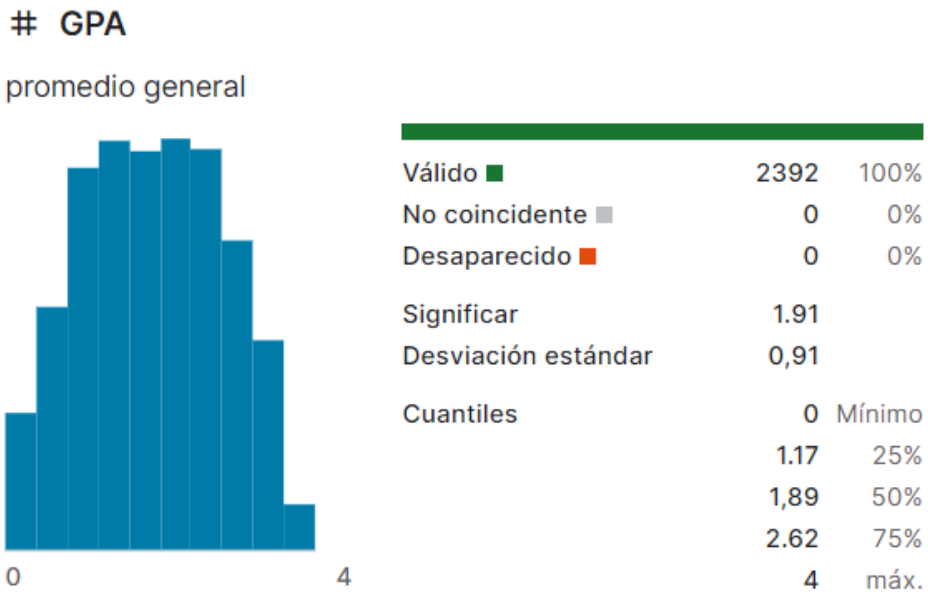


Figure 2: Descripción de la figura 2.





1.8 Bibliografía

References

[1] AWS. (s.f.). *What is Linear Regression?*. Recuperado de `https://aws.amazon.com/es/what-is/linear-regression/#:~:text=La%20regresi%C3%B3n%20lineal%20es%20una,independiente%20como%20una%20ecuaci%C3%B3n%20lineal`.

[2] Rabie El Kharoua. (2024). *Students Performance Dataset*. Kaggle. DOI: `https://doi.org/10.34740/KAGGLE/DS/5195702`.