Andres Siqueiros Cuellar 609026

Introducción

El objetivo de este trabajo es desarrollar un modelo de regresión lineal múltiple para predecir la calificación final (G3) de estudiantes a partir de información académica y demográfica.

El análisis se enfoca en enfrentar los retos asociados al uso de datos reales, particularmente la selección adecuada de variables explicativas, la identificación de relaciones redundantes y la evaluación del desempeño del modelo tanto en datos de entrenamiento como de prueba.

Descripción del conjunto de datos

Para el análisis se utilizó el archivo A1.3 Calificaciones.csv, el cual contiene información de estudiantes de un curso. El conjunto de datos incluye variables demográficas, académicas y de desempeño, entre ellas:

- Escuela
- Sexo
- Edad
- Horas de estudio
- Materias reprobadas
- Acceso a internet
- Faltas
- Calificaciones parciales (G1, G2)
- Calificación final (G3)

La variable G3 se definió como la variable dependiente, mientras que el resto fueron consideradas inicialmente como variables explicativas candidatas.

Preparación y limpieza de los datos

Antes de construir el modelo, fue necesario preparar el conjunto de datos para asegurar su compatibilidad con un modelo de regresión lineal múltiple.

Las variables categóricas binarias (Escuela, Sexo e Internet) fueron transformadas a valores numéricos (0 y 1), permitiendo su inclusión directa en el modelo sin necesidad de variables dummy adicionales.

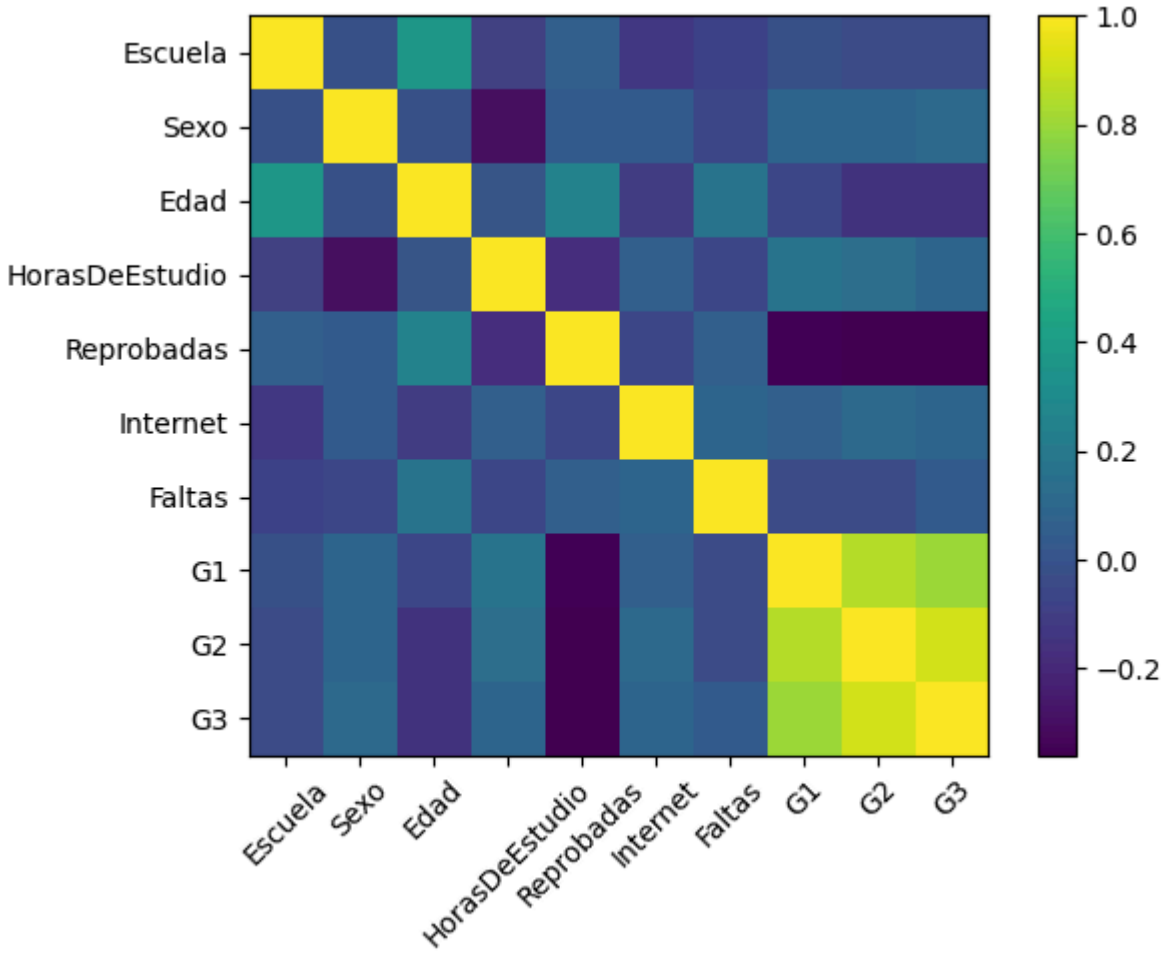
Código utilizado para la transformación:

```
df["Escuela"] = df["Escuela"].map({"GP": 0, "MS": 1})
df["Sexo"] = df["Sexo"].map({"F": 0, "M": 1})
df["Internet"] = df["Internet"].map({"no": 0, "yes": 1})
```

Las demás variables ya se encontraban en formato numérico, por lo que no fue necesario realizar transformaciones adicionales en esta etapa.

Análisis de relaciones entre variables

Con el objetivo de identificar posibles problemas de colinealidad y redundancia de información, se calculó la matriz de correlación entre todas las variables numéricas.



La tabla de correlación muestra que las calificaciones parciales G1 y G2 presentan una alta correlación tanto entre sí como con la calificación final G3. En particular, G2 presenta la correlación más alta con G3, lo que indica un alto poder explicativo.

Por otro lado, variables como Escuela, Sexo, Internet y Faltas presentan correlaciones bajas con la calificación final, lo que sugiere un aporte limitado al modelo.

Este análisis permitió detectar que incluir simultáneamente G1 y G2 podría generar problemas de multicolinealidad, por lo que se decidió conservar únicamente G2, al ser la calificación parcial más cercana temporalmente a la calificación final.

Selección de características

Con base en el análisis de correlación y en criterios académicos, se seleccionó el siguiente subconjunto de variables explicativas:

- G2: principal predictor de la calificación final
- Reprobadas: indicador del historial académico del estudiante
- Edad: variable demográfica que aporta contexto adicional
- HorasDeEstudio: variable conductual relacionada con el desempeño académico

Tabla. Coeficientes del modelo de regresión lineal múltiple

Variable	Coeficiente	Valor p	Interpretación
Intercepto	10.5137	0.000	Nivel base estimado de la calificación final cuando las variables explicativas se encuentran en su media
G2	4.0178	0.000	Incrementos en la segunda calificación parcial se asocian con aumentos significativos en la calificación final
Reprobadas	-0.1557	0.178	Mayor número de materias reprobadas tiende a reducir la calificación final, aunque no es estadísticamente significativo
Edad	-0.1788	0.093	La edad presenta una relación negativa débil con la calificación final
Horas de estudio	-0.0419	0.699	No se observa un efecto significativo de las horas de estudio sobre la calificación final en este modelo

Variables como G1, Escuela, Sexo, Internet y Faltas fueron excluidas del modelo debido a su bajo aporte explicativo o a problemas de redundancia.

Entrenamiento del modelo de regresión lineal múltiple

Se entrenó un modelo de regresión lineal múltiple utilizando un 80% de los datos para entrenamiento y un 20% para prueba, con el objetivo de evaluar la capacidad de generalización del modelo.

El modelo fue ajustado utilizando mínimos cuadrados ordinarios (OLS) sobre el conjunto de entrenamiento, lo que permitió obtener coeficientes, valores p y métricas de ajuste.

Código utilizado para el entrenamiento del modelo:

```
X = df[["G2", "Reprobadas", "Edad", "HorasDeEstudio"]]
y = df["G3"]
```

Para evaluar la capacidad de generalización del modelo, se calcularon los coeficientes de determinación en los conjuntos de entrenamiento y prueba:

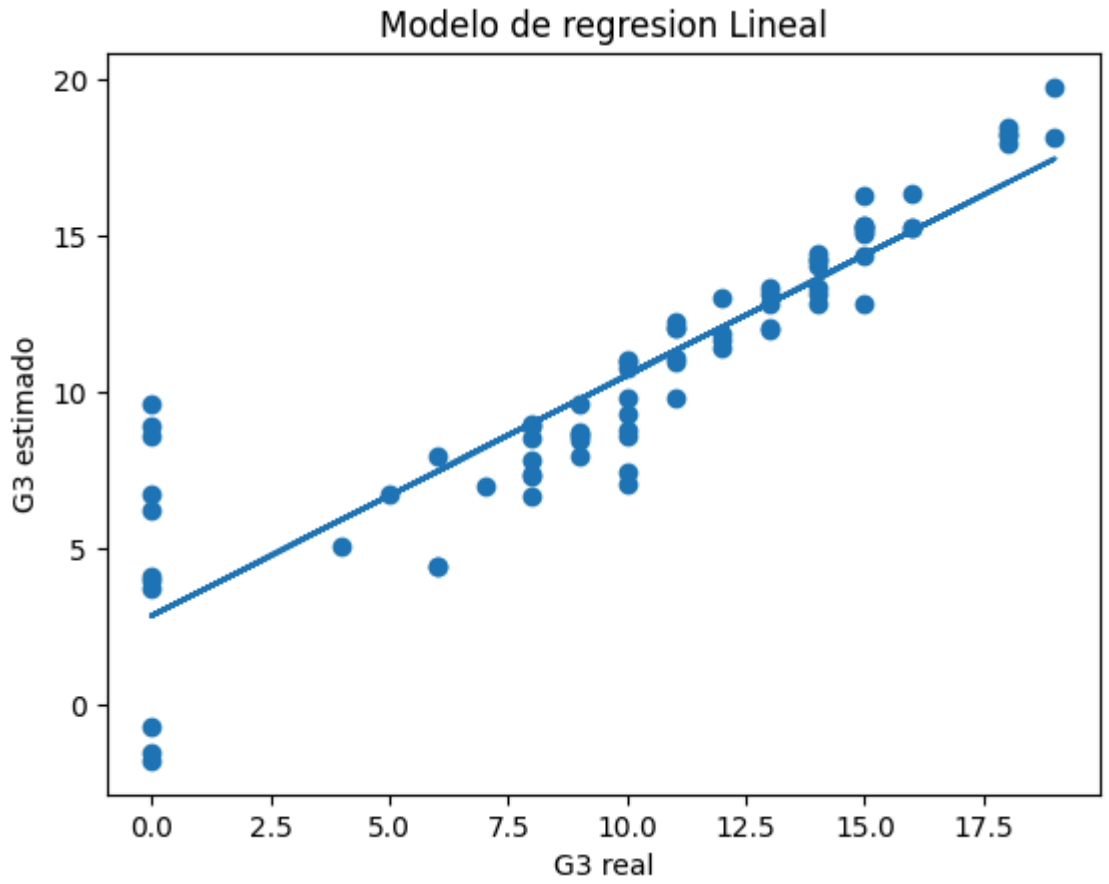
- **R^2 train:** 0.829
- **R^2 test:** 0.796
- **Diferencia:** 0.033

La diferencia entre el R^2 de entrenamiento y prueba es pequeña, lo que indica que el modelo **no presenta sobreajuste significativo** y generaliza adecuadamente a datos no utilizados durante el entrenamiento. Esto refuerza la robustez del modelo y valida el proceso de selección de variables realizado previamente.

Andres Siqueiros Cuellar 609026

Evaluación del ajuste del modelo

Para evaluar visualmente el ajuste del modelo, se compararon los valores reales de la calificación final con los valores estimados por el modelo en el conjunto de prueba.



La Figura muestra la gráfica de dispersión entre G3 real y G3 estimado, junto con la línea de tendencia del modelo.

La concentración de los puntos alrededor de la línea indica un buen ajuste general, confirmando que el modelo captura adecuadamente la relación entre las variables explicativas y la calificación final.

Conclusiones

Andres Siqueiros Cuellar 609026

El análisis realizado demuestra la importancia de una adecuada selección de características en la construcción de modelos de regresión lineal múltiple. A través del análisis de correlación fue posible identificar variables redundantes y descartar aquellas con bajo poder explicativo.

El modelo final, construido a partir de la calificación parcial G2 y variables académicas adicionales, logró explicar una proporción significativa de la variabilidad de la calificación final, manteniendo un buen desempeño tanto en entrenamiento como en prueba. Este trabajo evidencia que el desempeño académico final de un estudiante no depende únicamente de una sola variable, sino de una combinación de factores académicos y personales.