Danta
analytics

Caso de Análisis

Por Ronaldo Vindas Barboza

Agenda

Introducción

Metodología

Análisis Exploratorio Inicial de Características

Análisis Estadístico de Características

Análisis Visual de Características

Comportamiento de Datos

Modelo de Predicción

Hallazgos Importantes

Introducción

OBJETIVO:

Analizar el desempeño de los estudiantes en exámenes de matemáticas, lectura y escritura, utilizando el dataset "Students Performance in Exams" de Kaggle. A través de este análisis, se busca identificar aquellos patrones y factores que influyen en los resultados de los estudiantes.

kaggle

El Conjunto de Datos

Incluye las características:

- Género
- Grupo Étnico
- Nivel de Educación de Padres
- Tipo de Almuerzo
- Curso de Preparación
- Calificaciones

Importancia

Comprender qué factores afectan el rendimiento de las pruebas que puedan ayudar a desarrollar estrategias que mejoren los resultados de los estudiantes.

Metodología

LECTURA DE DATOS

PREPROCESAMIENTO DE DATOS

ANÁLISIS EXPLORATORIO
VISUAL Y ESTADÍSTICO

ANÁLISIS DE RESULTADOS

Lectura de Datos

Los datos fueron leídos por medio de la biblioteca Pandas de Python y se les almacenó en un Dataframe.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Preprocesamiento de Datos

Se aplicó One-Hot-Encoding para transformar features categóricos a numéricos y Normalización para establecer un intervalo de valores [0, 1].

Análisis Exploratorio y Comportamiento de Datos

Análisis Exploratorio

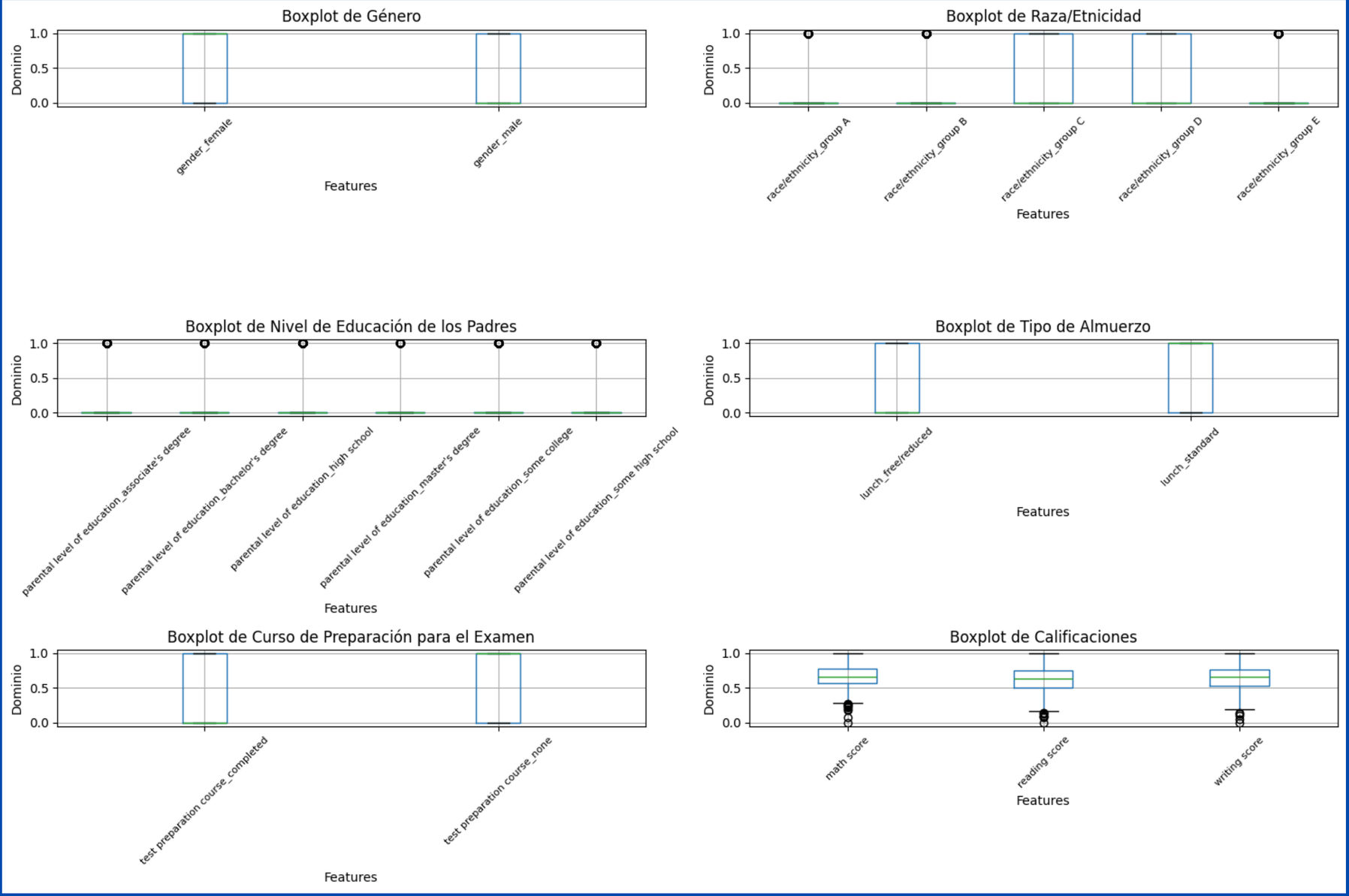
VISUAL

- Gráficos de Caja (BoxPlot)
- Gráfico de Dispersión
- Gráficos de Histogramas
- Matriz de Correlación

ESTADÍSTICOS

- Promedio
- Desv. Estándar
- Rangos de Cuartiles

Gráficos de Caja



Comportamiento

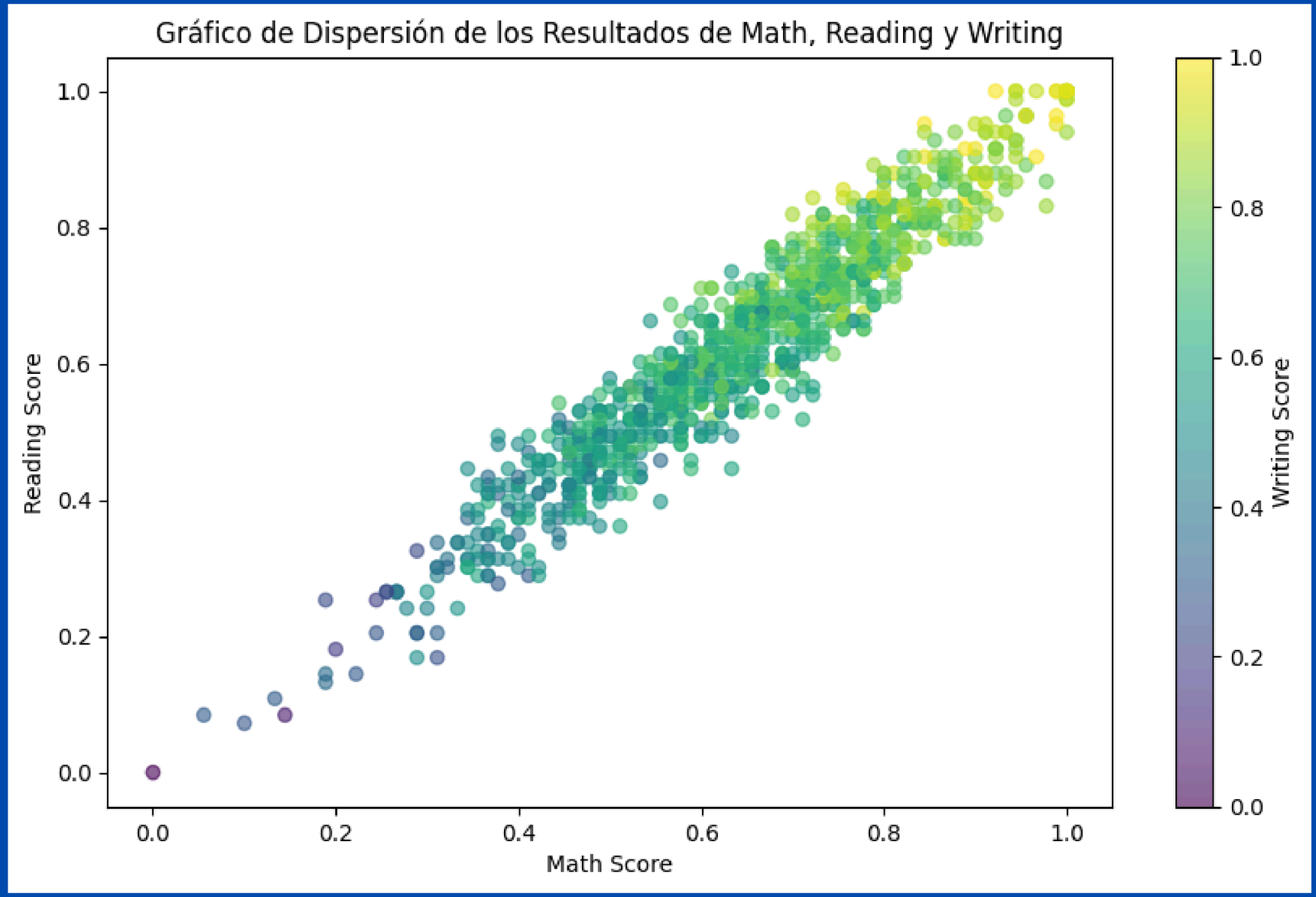
El Análisis Visual muestra:

Según los Gráficos de Cajas:

- Se presentan valores atípicos (**outliers**) principalmente en los grupos étnicos A, B y E así cómo en los niveles de educación de los padres. Se podrían trabajarán métodos de imputación de datos para reducir su impacto en el modelo de predicción.

- Las características de calificaciones presentan también valores atípicos, estos a causa de **variabilidad natural**, y por tanto se tomarán en cuenta en el modelo de predicción.

Gráfico de Dispersión



Comportamiento

Análisis Visual:

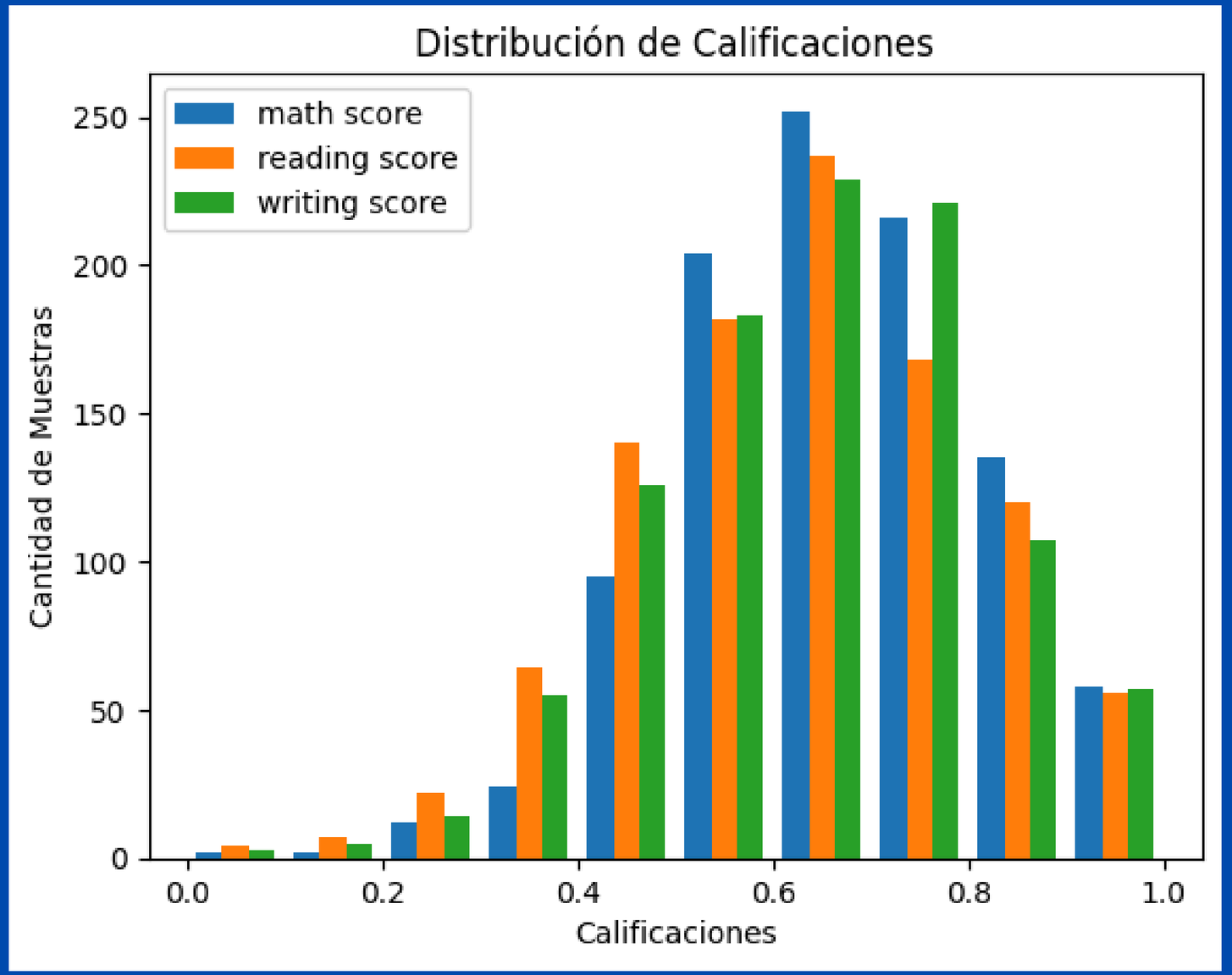
De acuerdo al Gráfico de Dispersión:

- Los estudiantes que suelen **pasar** un examen de una materia suelen hacerlo también con las otras dos.

- Aquellos estudiantes que fallan un examen tienden a **fallar** los demás.

- La mayoría de estudiantes obtienen **notas regulares**.

Gráfico de Histogramas



Comportamiento

Análisis Visual:

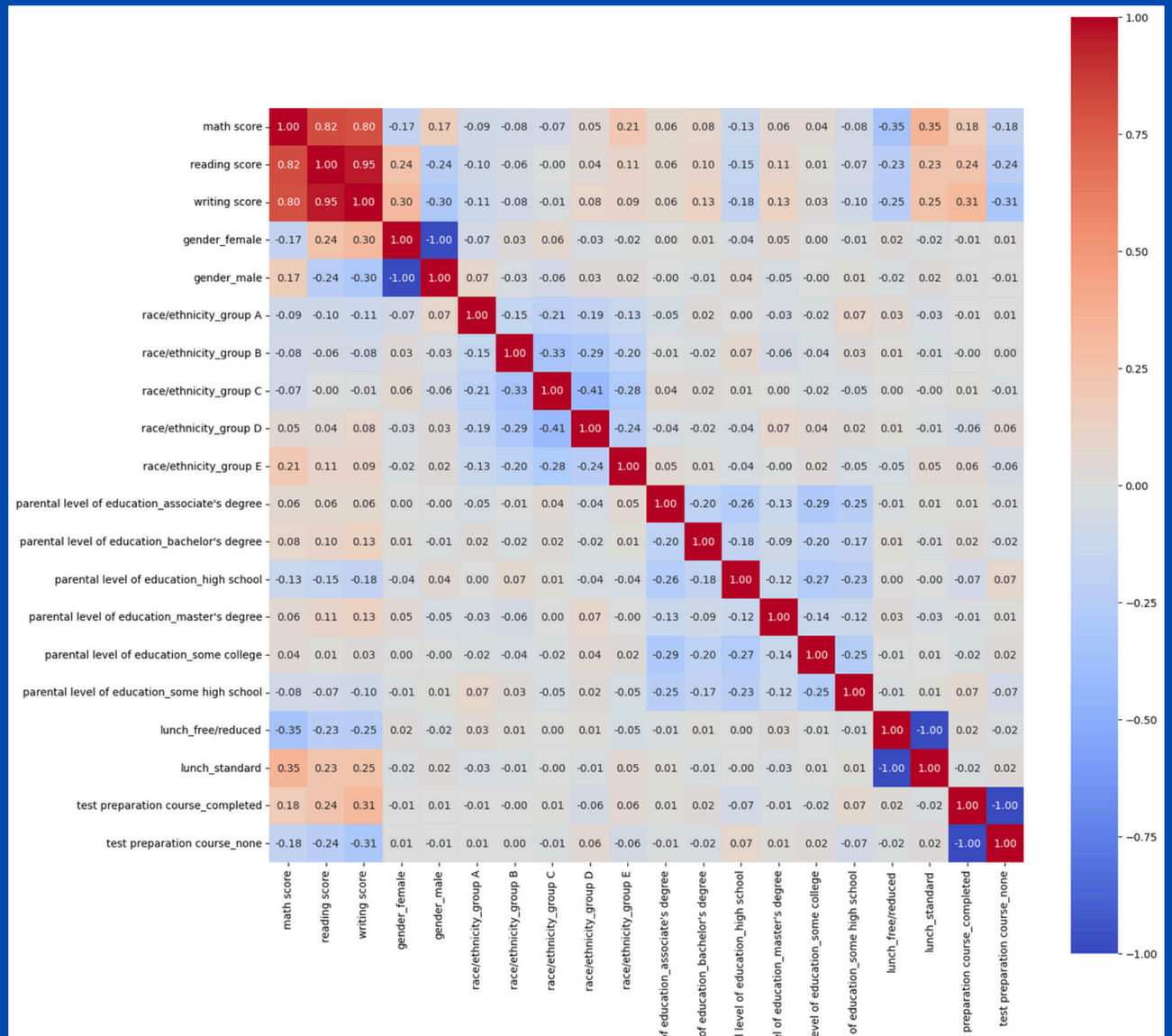
El gráfico de Histogramas nos dice que:

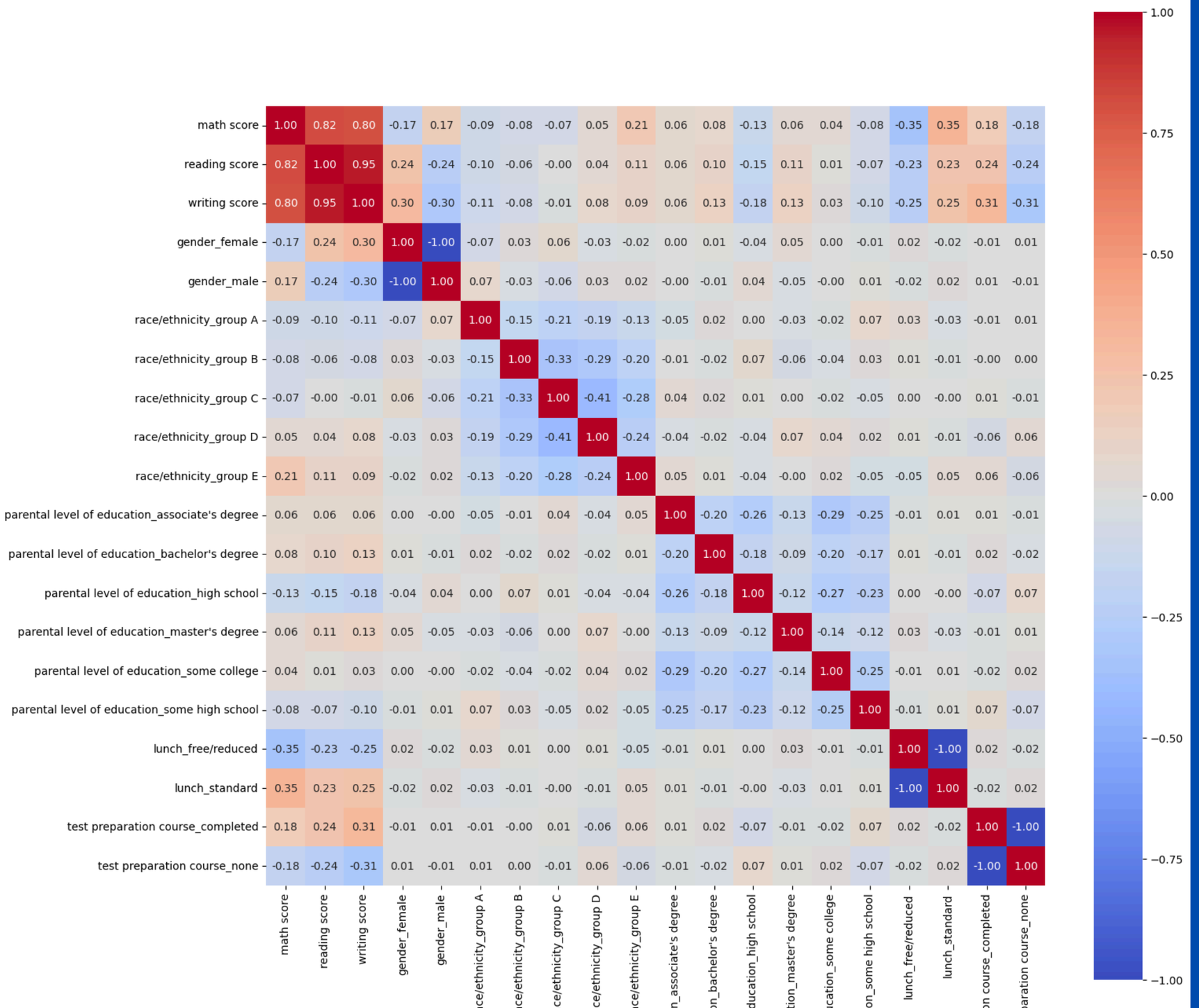
- Las muestras de puntuaciones siguen una **distribución normal**.

- Aproximadamente solo un 0.5% de los estudiantes sacó una **nota mayor** o igual a 90 en los tres exámenes.

- Menos de 25 estudiantes obtuvieron una **nota menor** a 20 en los tres exámenes.

Matriz de Correlación




math score 1.00 0.82 0.80 -0.17 0.17 -0.09 -0.08 -0.07 0.05 0.21 0.06 0.08 -0.13 0.06 0.04 -0.08 -0.35 0.35 0.18 -0.18
reading score 0.82 1.00 0.95 0.24 -0.24 -0.10 -0.06 -0.00 0.04 0.11 0.06 0.10 -0.15 0.11 0.01 -0.07 -0.23 0.23 0.24 -0.24
writing score 0.80 0.95 1.00 0.30 -0.30 -0.11 -0.08 -0.01 0.08 0.09 0.06 0.13 -0.18 0.13 0.03 -0.10 -0.25 0.25 0.31 -0.31
gender_female -0.17 0.24 0.30 1.00 -1.00 -0.07 0.03 0.06 -0.03 -0.02 0.00 0.01 -0.04 0.05 0.00 -0.01 0.02 -0.02 -0.01 0.01
gender_male 0.17 -0.24 -0.30 -1.00 1.00 0.07 -0.03 -0.06 0.03 0.02 -0.00 -0.01 0.04 -0.05 -0.00 0.01 -0.02 0.02 0.01 -0.01
race/ethnicity_group A -0.09 -0.10 -0.11 -0.07 0.07 1.00 -0.15 -0.21 -0.19 -0.13 -0.05 0.02 0.00 -0.03 -0.02 0.07 0.03 -0.03 -0.01 0.01
race/ethnicity_group B -0.08 -0.06 -0.08 0.03 -0.03 -0.15 1.00 -0.33 -0.29 -0.20 -0.01 -0.02 0.07 -0.06 -0.04 0.03 0.01 -0.01 -0.00 0.00
race/ethnicity_group C -0.07 -0.00 -0.01 0.06 -0.06 -0.21 -0.33 1.00 -0.41 -0.28 0.04 0.02 0.01 0.00 -0.02 -0.05 0.00 -0.00 0.01 -0.01
race/ethnicity_group D 0.05 0.04 0.08 -0.03 0.03 -0.19 -0.29 -0.41 1.00 -0.24 -0.04 -0.02 -0.04 0.07 0.04 0.02 0.01 -0.01 -0.06 0.06
race/ethnicity_group E 0.21 0.11 0.09 -0.02 0.02 -0.13 -0.20 -0.28 -0.24 1.00 0.05 0.01 -0.04 -0.00 0.02 -0.05 -0.05 0.05 0.06 -0.06
parental level of education_associate's degree 0.06 0.06 0.06 0.00 -0.00 -0.05 -0.01 0.04 -0.04 0.05 1.00 -0.20 -0.26 -0.13 -0.29 -0.25 -0.01 0.01 0.01 -0.01
parental level of education_bachelor's degree 0.08 0.10 0.13 0.01 -0.01 0.02 -0.02 0.02 -0.02 0.01 -0.20 1.00 -0.18 -0.09 -0.20 -0.17 0.01 -0.01 0.02 -0.02
parental level of education_high school -0.13 -0.15 -0.18 -0.04 0.04 0.00 0.07 0.01 -0.04 -0.04 -0.26 -0.18 1.00 -0.12 -0.27 -0.23 0.00 -0.00 -0.07 0.07
parental level of education_master's degree 0.06 0.11 0.13 0.05 -0.05 -0.03 -0.06 0.00 0.07 -0.00 -0.13 -0.09 -0.12 1.00 -0.14 -0.12 0.03 -0.03 -0.01 0.01
parental level of education_some college 0.04 0.01 0.03 0.00 -0.00 -0.02 -0.04 -0.02 0.04 0.02 -0.29 -0.20 -0.27 -0.14 1.00 -0.25 -0.01 0.01 -0.02 0.02
parental level of education_some high school -0.08 -0.07 -0.10 -0.01 0.01 0.07 0.03 -0.05 0.02 -0.05 -0.25 -0.17 -0.23 -0.12 -0.25 1.00 -0.01 0.01 0.07 -0.07
lunch_free/reduced -0.35 -0.23 -0.25 0.02 -0.02 0.03 0.01 0.00 0.01 -0.05 -0.01 0.01 0.00 0.03 -0.01 -0.01 1.00 -1.00 0.02 -0.02
lunch_standard 0.35 0.23 0.25 -0.02 0.02 -0.03 -0.01 -0.00 -0.01 0.05 0.01 -0.01 -0.00 -0.03 0.01 0.01 -1.00 1.00 -0.02 0.02
test preparation course_completed 0.18 0.24 0.31 -0.01 0.01 -0.01 -0.00 0.01 -0.06 0.06 0.01 0.02 -0.07 -0.01 -0.02 0.07 0.02 -0.02 1.00 -1.00
test preparation course_none -0.18 -0.24 -0.31 0.01 -0.01 0.01 0.00 -0.01 0.06 -0.06 -0.01 -0.02 0.07 0.01 0.02 -0.07 -0.02 0.02 -1.00 1.00
math score
reading score
writing score
gender_female
gender_male
race/ethnicity_group A
race/ethnicity_group B
race/ethnicity_group C
race/ethnicity_group D
race/ethnicity_group E
parental level of education_associate's degree
parental level of education_bachelor's degree
parental level of education_high school
parental level of education_master's degree
parental level of education_some college
parental level of education_some high school
lunch_free/reduced
lunch_standard
test preparation course_completed
test preparation course_none
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Comportamiento

Análisis Visual:

El Gráfico de Correlación arroja que:

- Sobresale el hecho de que la **calificación** en una materia se relaciona mucho con las demás, esto podría traducirse en que al igual a como se salga en una prueba, saldrá similar en las demás.

- El prepararse para los exámenes parece influir positivamente en el resultado de estos de manera baja/moderada.

Comportamiento

Análisis Visual:

El Gráfico de Correlación arroja que:

- Características como el **tipo de almuerzo** parece tener una correlación positiva y negativa **moderada** con las puntuaciones de los exámenes. Aquellos estudiantes que tienen un almuerzo estándar puede que salgan un poco mejor que aquellos que no almuerzan o que almuerzan poco.

- El **estudio de los padres** no parece influir demasiado con las calificaciones de los exámenes.

Comportamiento

Análisis Visual:

El Gráfico de Correlación arroja que:

- Las **mujeres** tienden a salir un poco mejor en Escritura y Lectura que los hombres.

-Los **hombres** suelen salir un poco mejor en Matemáticas que las mujeres.

- La mayoría de grupos étnicos no influyen mucho en el resultado de los exámenes, aunque el **grupo étnico E** pareciera influir de manera baja/moderada sobre las calificaciones.

Análisis Estadístico

	math score	reading score	writing score	gender_female	gender_male	race/ethnicity_group A	race/ethnicity_group B	race/ethnicity_group C	race/ethnicity_group D	race/ethnicity_group E	parental level of education_associate's degree	parental level of education_bachelor's degree	parental level of education_high school	parental level of education_master's degree	parental level of education_some college	parental level of education_some high school	lunch_free/reduced	lunch_standard	test preparation course_completed	test preparation course_none
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	0.660890	0.628542	0.645044	0.518000	0.482000	0.089000	0.190000	0.319000	0.262000	0.140000	0.222000	0.11800	0.196000	0.059000	0.226000	0.179000	0.355000	0.645000	0.358000	0.642000
std	0.151631	0.175906	0.168841	0.499926	0.499926	0.284886	0.392497	0.466322	0.439943	0.347161	0.415799	0.32277	0.397167	0.235743	0.418448	0.383544	0.478753	0.478753	0.479652	0.479652
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.570000	0.506024	0.530556	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.660000	0.638554	0.655556	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000
75%	0.770000	0.746988	0.766667	1.000000	1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Comportamiento

Análisis Estadístico:

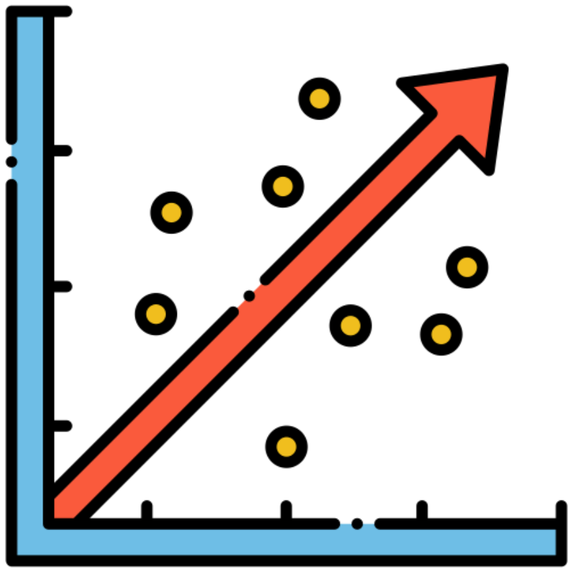
- En su mayoría, los estudiantes tienen calificaciones regulares (alrededor de los 60 pts).

- La media de estudiantes que no se preparan ante un examen es alta, aproximadamente un 64% de ellos.

- La mayoría de estudiantes pertenecen al grupo étnico C.

- Las características de "almuerzo" y "preparación" se encuentran algo desbalanceadas,. lo que podría llegar a sesgar el modelo predictivo.

Modelo de Predicción



Regresión Lineal

Uso de Biblioteca Scikit-Learn
Técnica 80/20.
Predecir “Math Score”.
Evaluado mediante:

- Mean Squared Error
- Coeficiente de Determinación (R^2 Score)
- Coeficientes de Variables
- Gráficos (Evaluación, Error y Coeficientes)

Experimento Adicional

Igual al modelo anterior pero entrenado con datos limpios de valores atípicos que puedan “entorpecer” el modelo.

Evaluación del Modelo

MEAN SQUARED ERROR (MSE)

0.0026945732490891833

R^2 SCORE

0.8788700517507204

Gráfico de Evaluación

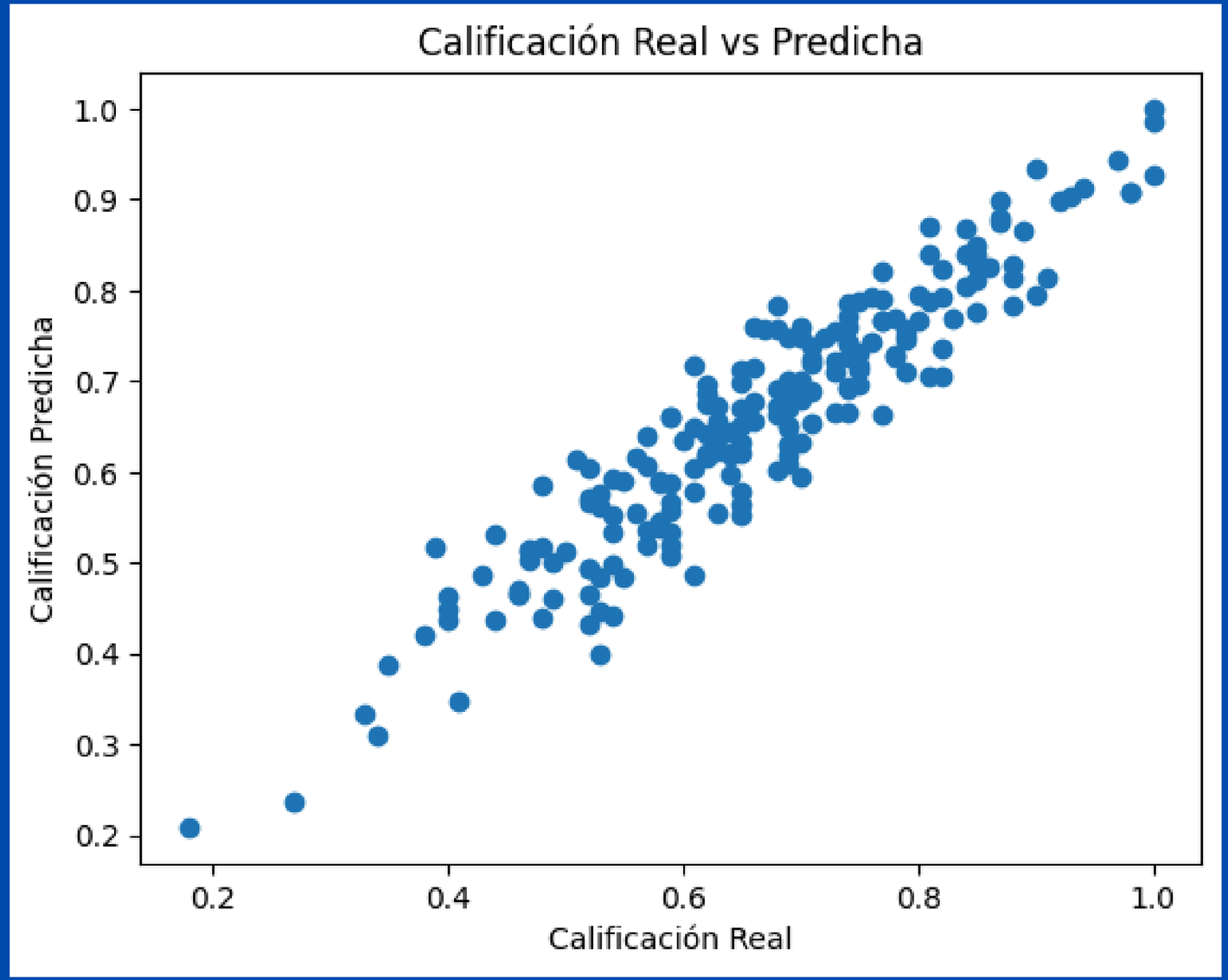


Gráfico de Error

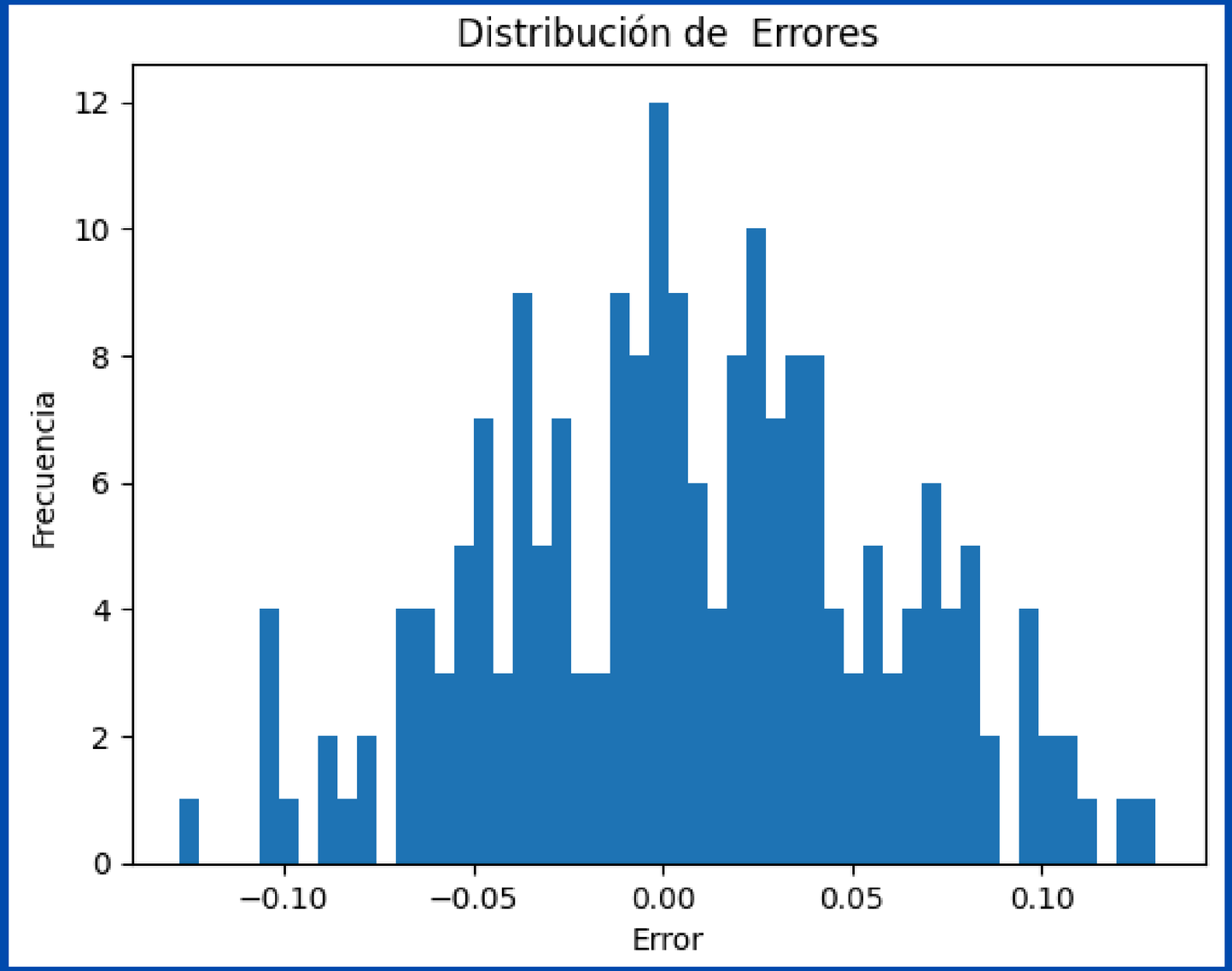
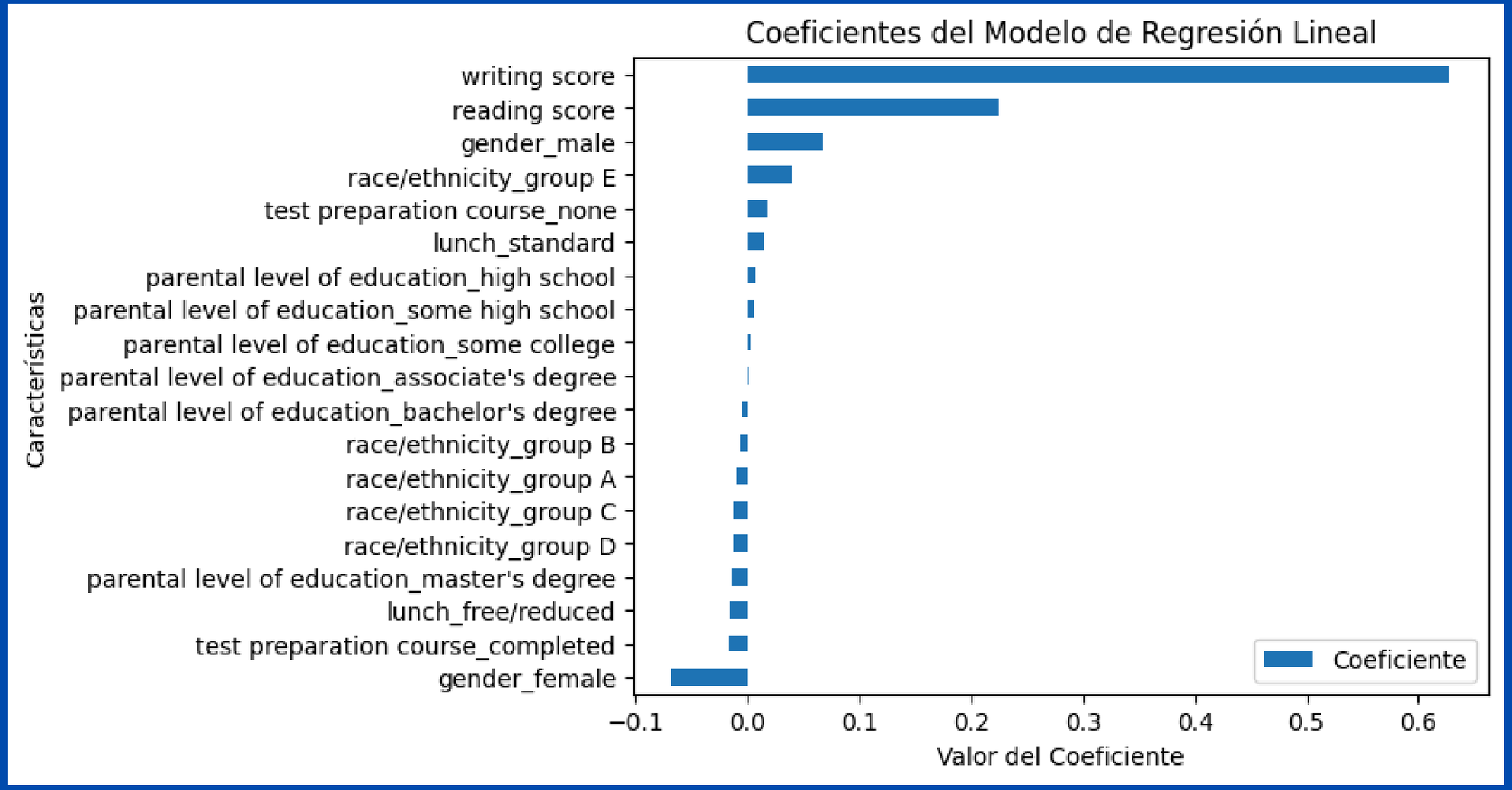


Gráfico de Coeficientes



Evaluación del Modelo Experimento

MEAN SQUARED ERROR (MSE)

0.0028874734116149433

R^2 SCORE

0.8439658801376765

Gráfico de Evaluación

Experimento Adicional

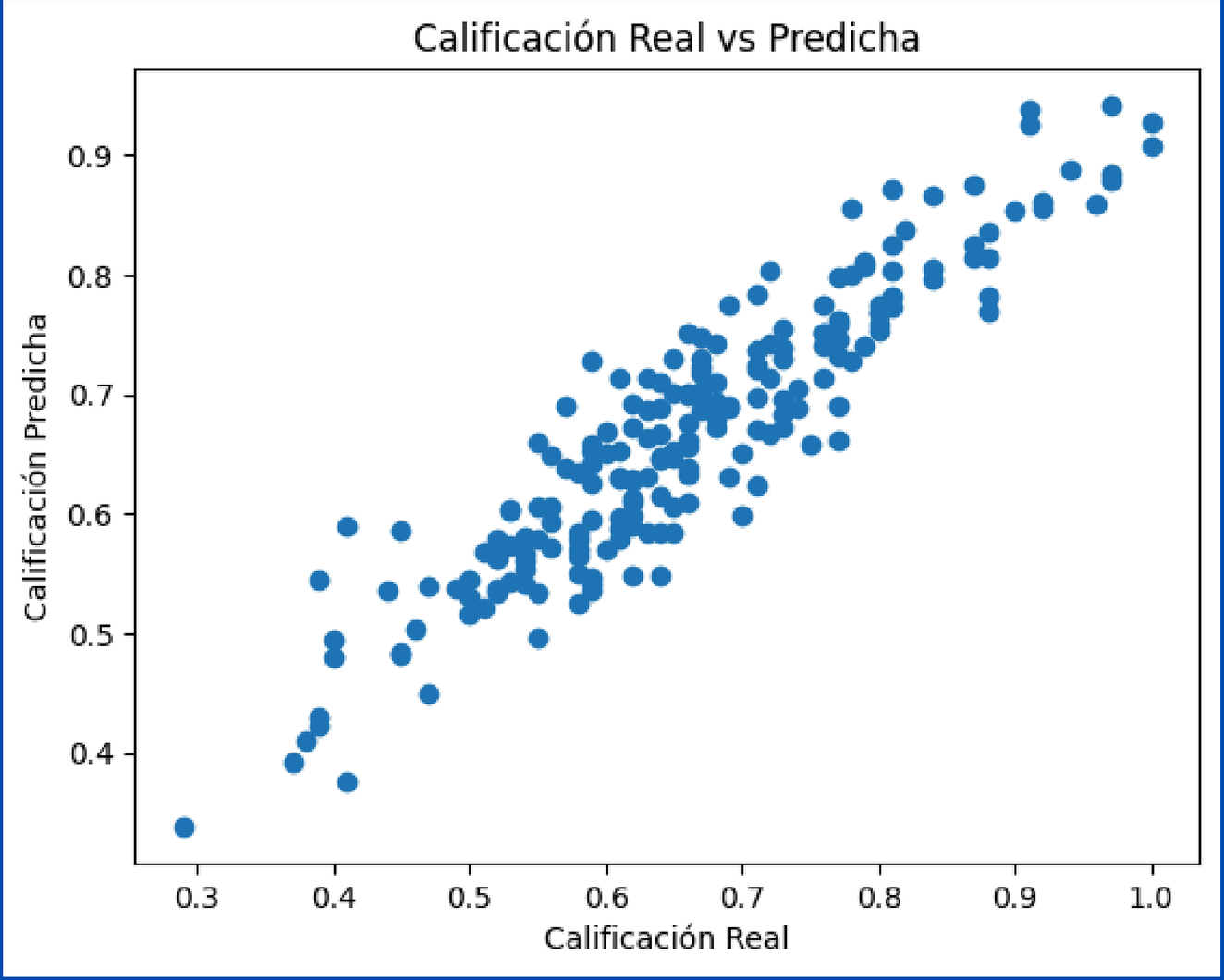


Gráfico de Error

Experimento Adicional

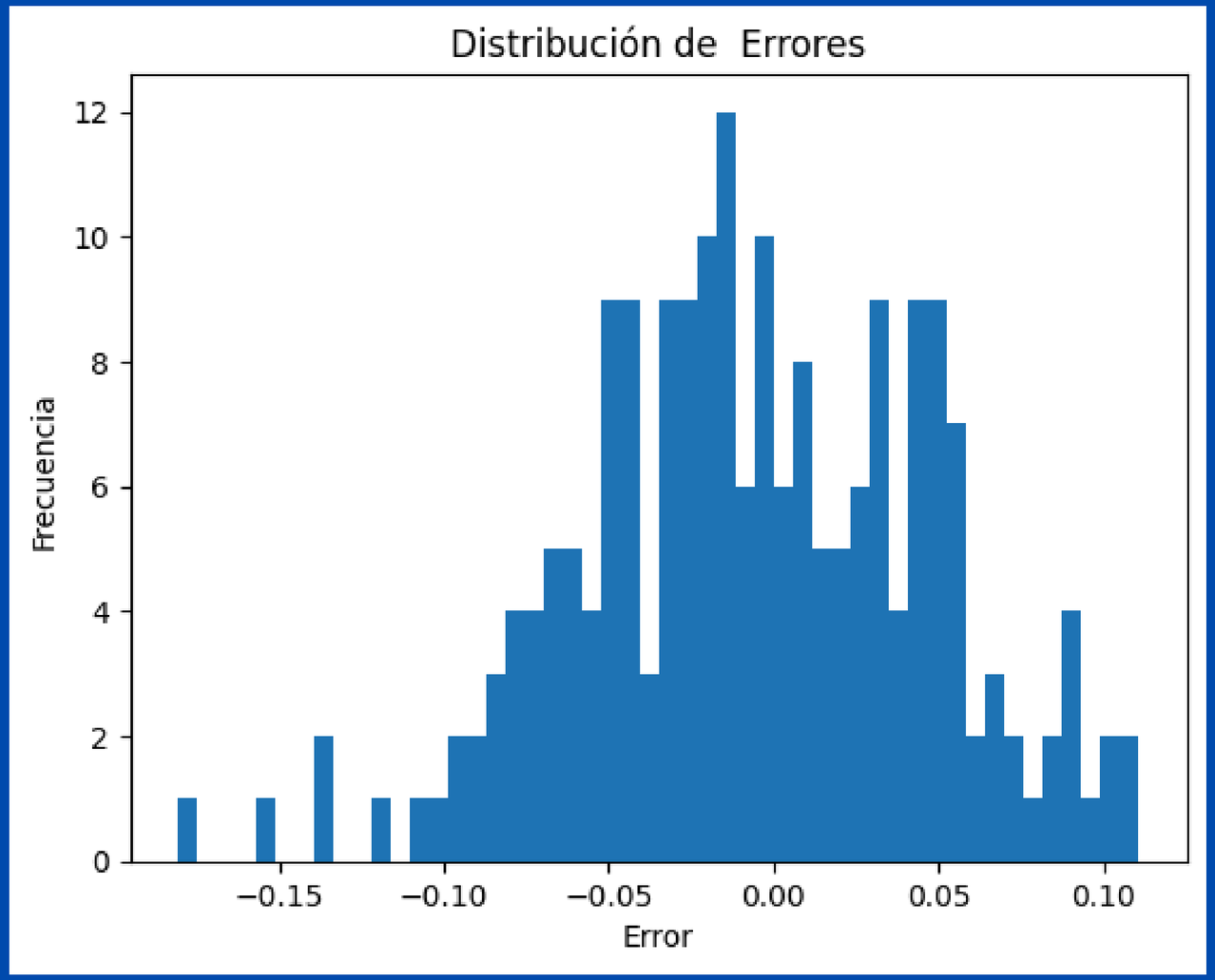
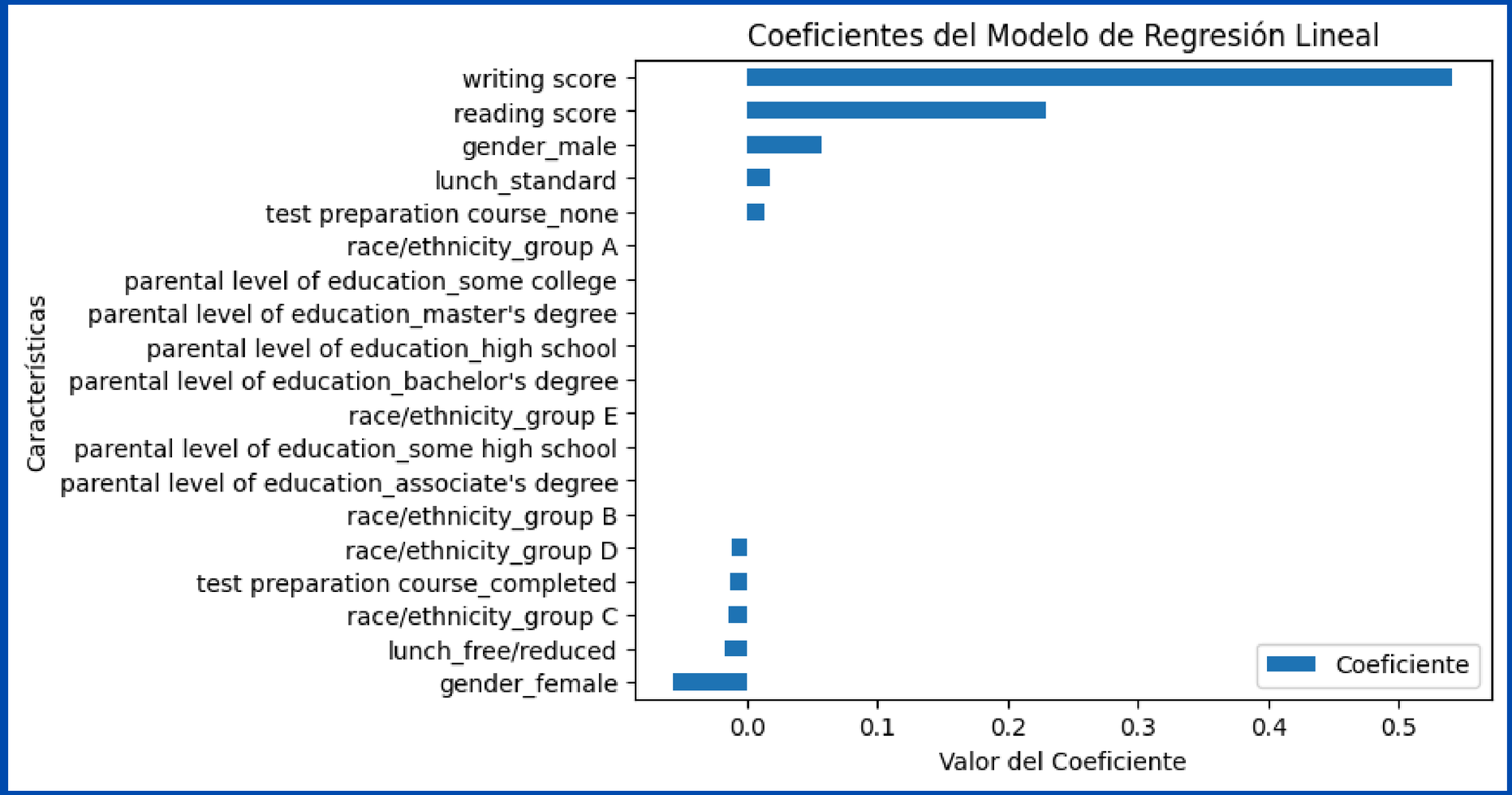


Gráfico de Coeficientes

Experimento Adicional



Hallazgos Clave

Resultados muestran:

De acuerdo a los resultados arrojados por la evaluación:

- La proporción de **error** del MSE en ambos modelos es **muy bajo**, lo que indica que estos modelos pueden llegar a predecir muy bien.

- El Coeficiente de Determinación (R^2 Score) tiene un valor de **ajuste muy bueno** sin llegar a un overfitting, además menciona que el 87% de la varianza en la calificación de matemáticas se puede explicar con las variables independientes.

Hallazgos Clave

Resultados muestran:

Según los gráficos de evaluación:

- El gráfico de dispersión muestra que los valores predichos siguen una distribución normal muy **similar** a la del dataframe **original** concentrando la mayoría de calificaciones alrededor de los 60 puntos. Además, este gráfico muestra la capacidad de aproximarse muy cercanamente al valor real de la calificación.

- El gráfico de distribución de errores muestra una **distribución** muy **simétrica** y con una varianza razonable aunque muestra aún algunos valores atípicos, esto podría deberse al haberse utilizado aún todos los features para el entrenamiento del modelo. Por lo que se pueden realizar más experimentos y hacer un nuevo entrenamiento sin los features menos significativos. Sin embargo, el modelo aún se comporta de muy buena manera.

Hallazgos Clave

Resultados muestran:

Según los gráficos de evaluación:

- Los índices de coeficientes de variables muestran que efectivamente, características como **"Reading Score"**, **"Writing Score"**, **"Gender Male"**, **"Race/Ethnicity Group E"** y **"Lunch Standard"** influyen de manera positiva en el resultado de la calificación de matemáticas.

- Sin embargo, resalta que la característica **"Test Preparation None"** influye aunque de poca manera en el resultado del examen. Y que, por otro lado, las características de **"Gender Female"** al igual que el análisis previo, parece influir de manera un poco negativa sobre el resultado del examen; Aunque curiosamente la característica **"Test Preparation Completed"** también influye negativamente.

¡Muchas Gracias!

¿Alguna consulta?

Enlace a Github

https://github.com/RonaldoVindas/Danta-CasoAnalisis