Métricas de Error

Objetivo

Usar el dataset 'SBUX Historical Data' para implementar modelos de regresión y analizar los resultados usando las métricas: MSE, RMSE, MAE, RAE, RSE, R2, CC.

Introducción

En el ámbito de la ciencia de datos e inteligencia artificial, hay dos principales tareas que se realizan, la clasificación y la predicción. Para poder evaluar la calidad de los modelos creados para tales fines, existen diversas métricas que permitirán analizar la eficacia y precisión de estos. Este documento se enfoca en presentar las métricas de error usadas para tareas de predicción, específicamente se presentan las métricas: Error Cuadrático Medio (MSE), Raíz del Error Cuadrático Medio (RMSE), Error Absoluto Medio (MAE), Error Absoluto Relativo (RAE), Error Cuadrático Relativo (RSE), Coeficiente de Determinación (R2) y el Coeficiente de Correlación (CC).

A continuación, en las siguientes secciones se dará un breve repaso de las propiedades y peculiaridades de cada métrica y se hará la comparación de cada métrica en la evaluación de regresiones lineales y polinomiales.

Marco Teórico

Se han mencionado ya, las métricas de error que se usarán en el presente documento. A continuación, se analizará cada una de ellas indicando su ecuación y algunas de sus propiedades.

a) Error Cuadrático Medio (MSE): Es uno de los métodos más comunes. Mide el promedio de los errores al cuadrado entre las predicciones realizadas por el modelo y los valores reales observados. Penaliza más severamente los grandes errores debido al uso del cuadrado de las diferencias, esto lo hace también muy sensible a outliers. Se define por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

b) Raíz del Error Cuadrático Medio (RMSE): Representa la magnitud promedio de los errores entre las predicciones y los valores reales. Al calcular la raíz cuadrada, el RMSE devuelve un valor en las mismas unidades que los datos originales, lo que facilita su interpretación y comparación. Es sensible a outliers. Su ecuación es la siguiente:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2)$$

c) Error Absoluto Medio (MAE): Calcula la diferencia absoluta entre cada valor predicho y su correspondiente valor real y luego promedia todos estos valores absolutos. Posee dos ventajas, que las medidas de los errores son fácilmente interpretables pues están en las unidades originales de los datos, y que es menos sensible a outliers. Se calcula de la siguiente forma:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \qquad (3)$$

d) Error Absoluto Relativo (RAE): Permite comparar el desempeño del modelo con respecto a una línea de base que usualmente es el promedio de los valores observados. Proporciona una medida de error relativa en lugar de absoluta, lo que facilita la comparación del rendimiento del modelo en diferentes contextos o entre diferentes conjuntos de datos con escalas distintas. No penaliza fuertemente los errores grandes y puede ser menos sensibles a outliers. Su fórmula es:

$$RAE = \frac{\sum_{i=1}^{n} |y_i - y_i^{pred}|}{\sum_{i=1}^{n} |y_i - \bar{y}|} \qquad (4)$$

e) Error Cuadrático Relativo (RSE): Expresa el error cuadrático del modelo como una expresión del error cuadrático de la línea de base. A diferencia del RAE, penaliza más severamente los errores grandes debido al uso del cuadrado de las diferencias, lo que puede ser útil en situaciones donde se desea identificar y minimizar errores atípicos.

$$RSE = \frac{\sum_{i=1}^{n} (y_i - y_i^{pred})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (5)$$

f) Coeficiente de Determinación (R^2): Indica qué porcentaje de la variación en los valores reales observados es capturado por el modelo predictivo. Varía entre 0 y 1. A diferencia de todas las métricas anteriores, mientras más alto sea el valor, significa que el modelo se ajusta mejor a los datos. Se calcula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^{pred})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (6)$$

g) Coeficiente de Correlación de Pearson (CC): Evalúa la relación lineal entre dos variables. Se utiliza para evaluar la fuerza y la dirección de la relación lineal entre los valores reales observados y los valores predichos por el modelo. Los valores van de -1.0 a +1.0, indicando -1.0 una perfecta correlación negativa y +1.0 una perfecta correlación positiva. Su fórmula es la siguiente:

$$CC = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(y_i^{pred} - \bar{y}^{pred})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 * \sum_{i=1}^{n} (y_i^{pred} - \bar{y}^{pred})^2}} \qquad (7)$$

Materiales y Métodos

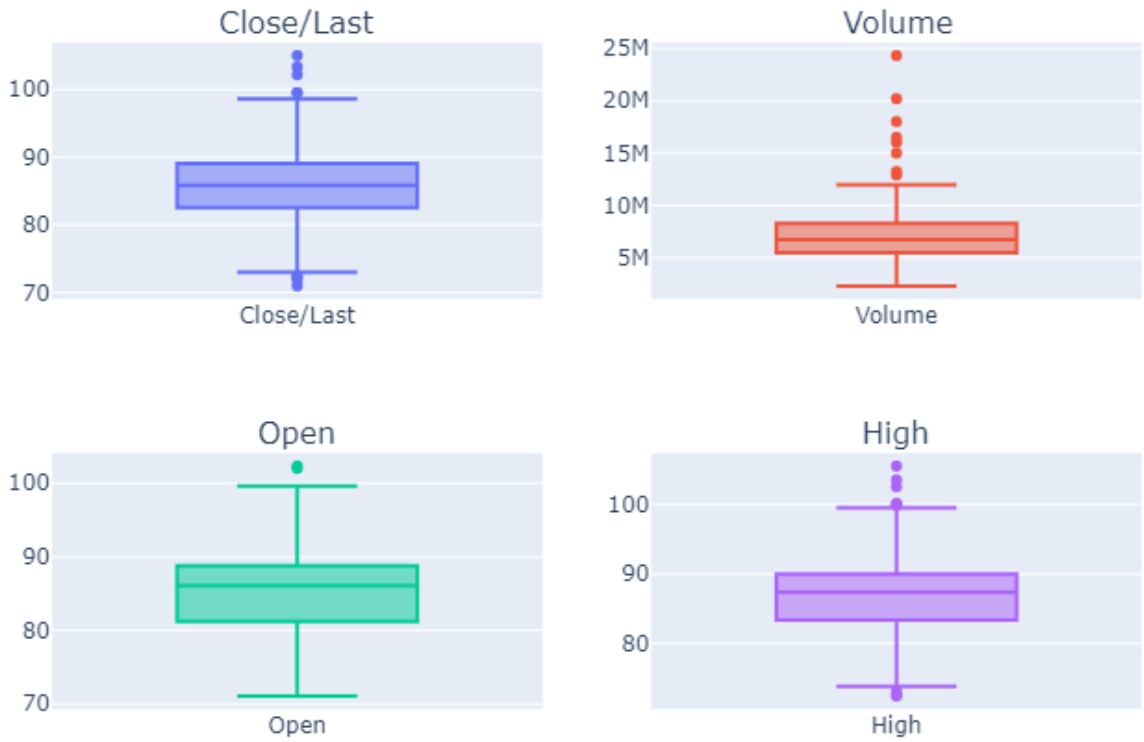
Para el desarrollo de este trabajo se utilizó el dataset ‘SBUX Historical Data’, el cual contiene los diferentes precios y el volumen total negociado en todas las operaciones de la empresa Starbucks. Por su parte, para el manejo de los datos y la creación de las funciones para calcular las métricas, se recurrió a las librerías Pandas y Numpy. Por último, para la creación de las gráficas se usó Matplotlib.

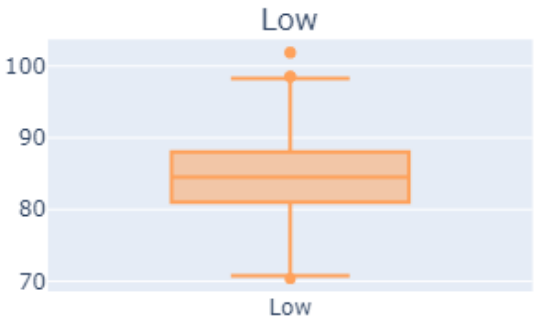
Resultados

Para el desarrollo de esta práctica se siguieron los siguientes pasos: primero se graficaron los boxplot de cada atributo con el fin de visualizar los outliers de cada uno; posteriormente se realizaron regresiones lineales y polinomiales comparando el atributo ‘Close/Last’ con todos los demás, esto debido a que este atributo posee el precio final acordado y se deseaba analizar su relación con las demás variables, a su vez, en este paso, se calcularon y mostraron todas las métricas antes mencionadas para cada regresión; por último, se hace una comparación de cada métrica contra cada una de las diferentes regresiones ejecutadas.

Visualización de Outliers

Se graficaron los boxplot de cada atributo, se puede observar claramente que el atributo ‘Volume’ posee una cantidad significativa de outliers, siendo seguido en la lista por las columnas ‘Close/Last’ y ‘High’.

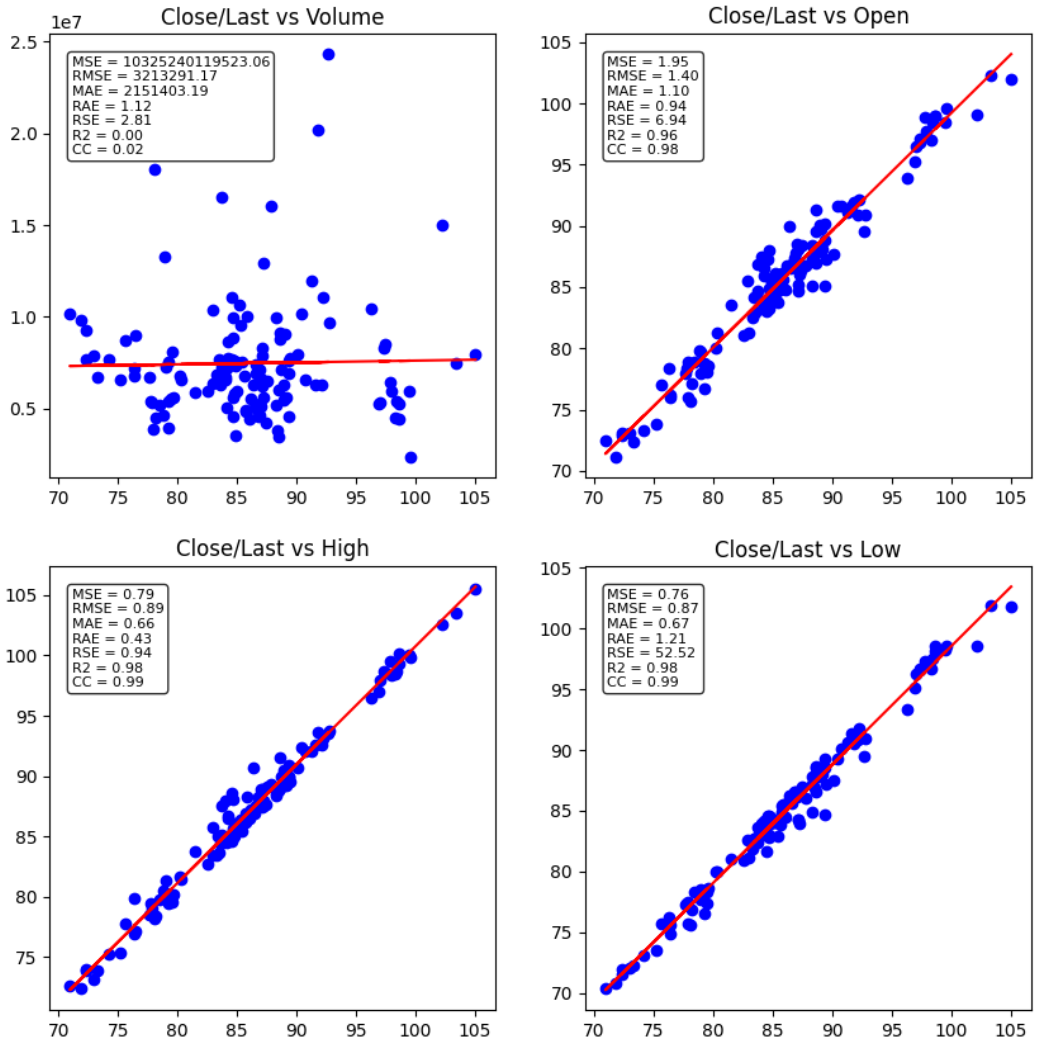




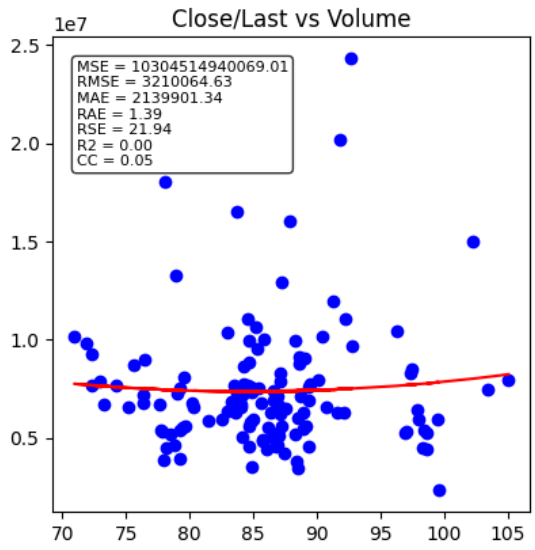
Regresiones

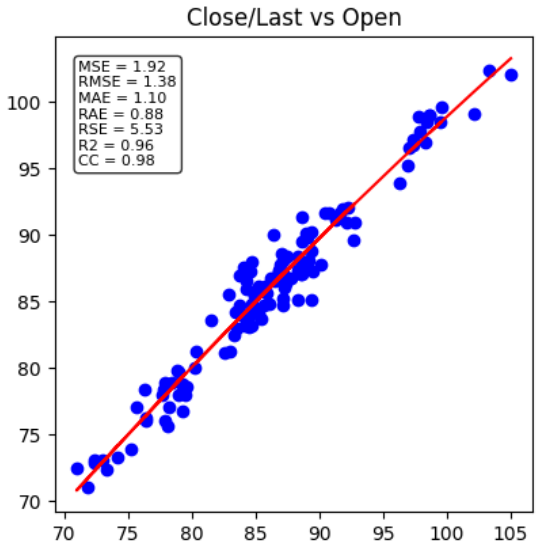
Se efectuaron las regresiones lineales y polinomiales de segundo y tercer grado comparando el atributo ‘Close/Last’ con los demás, a su vez para cada una de estas regresiones se calcularon las métricas mencionadas al principio. Todo esto se puede ver a continuación:

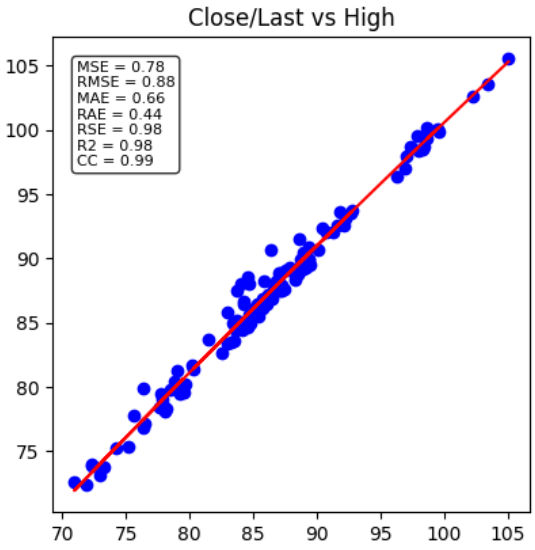
- Regresión Lineal:

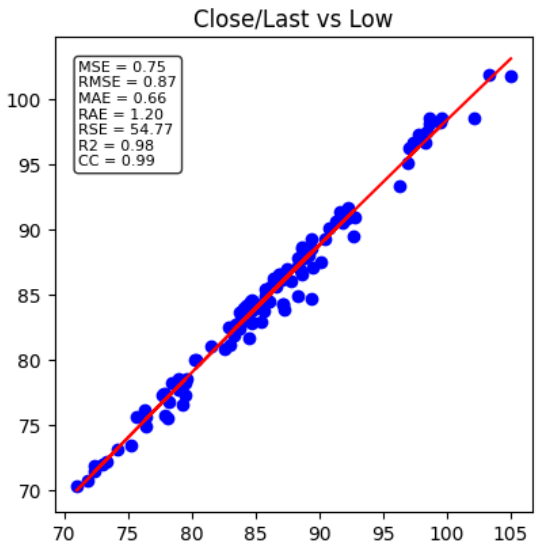


- Regresión Polinomial 2° grado

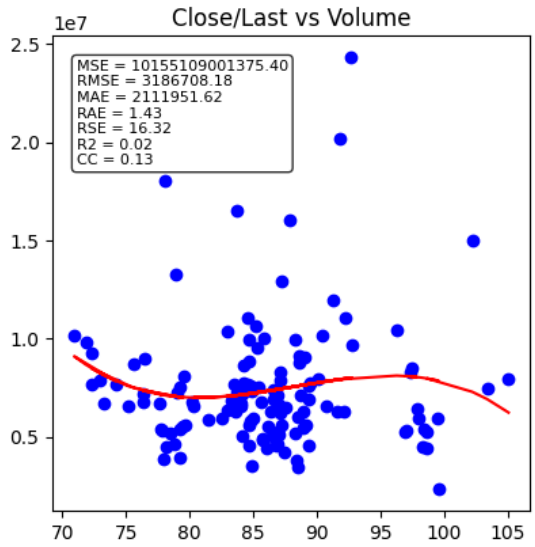


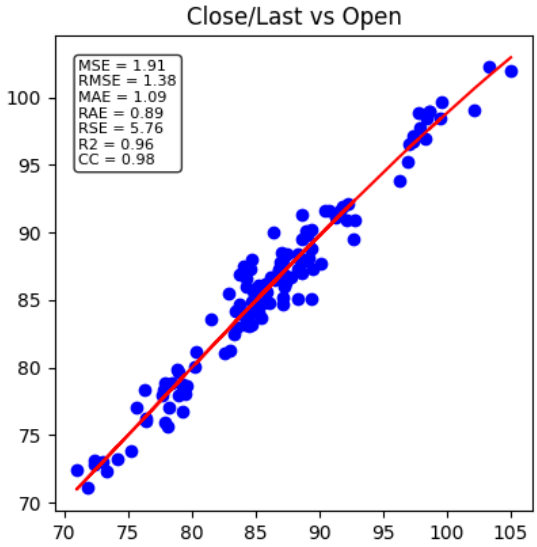


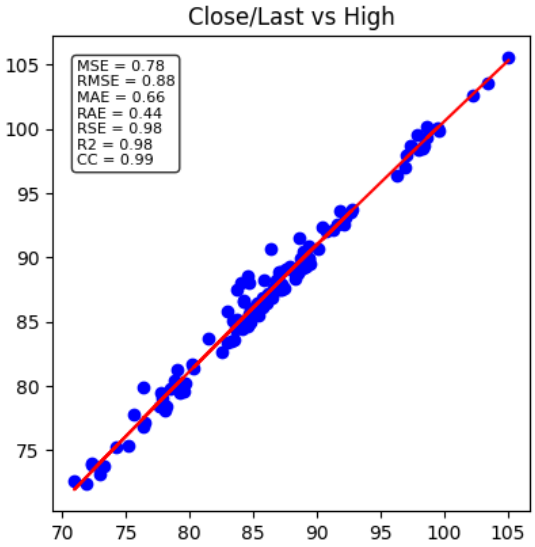


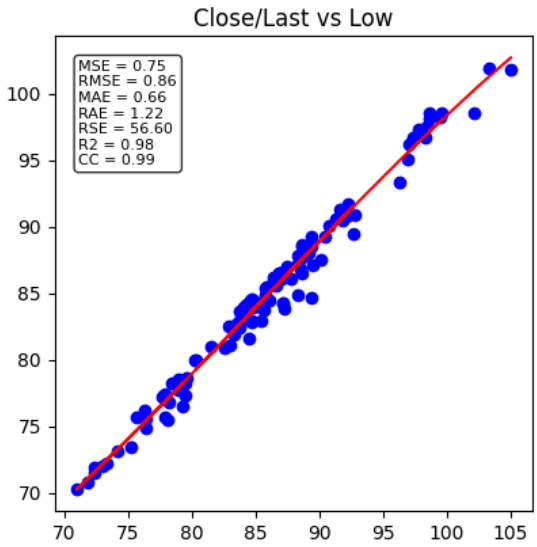


- Regresión Polinomial 3er grado



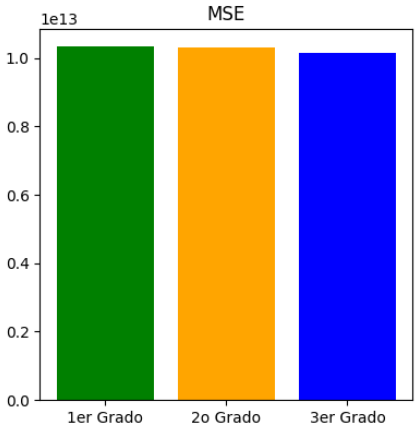


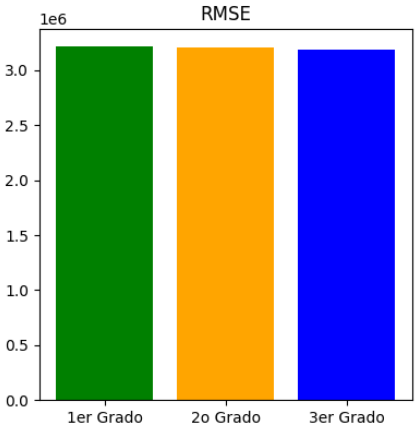


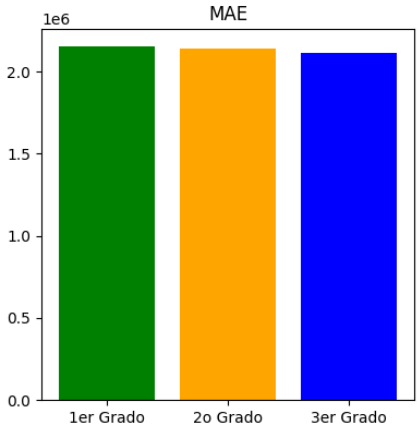


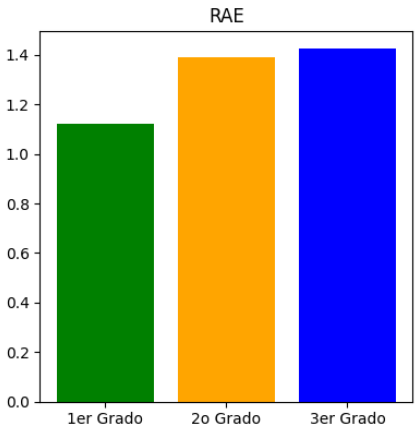
Comparación entre regresiones

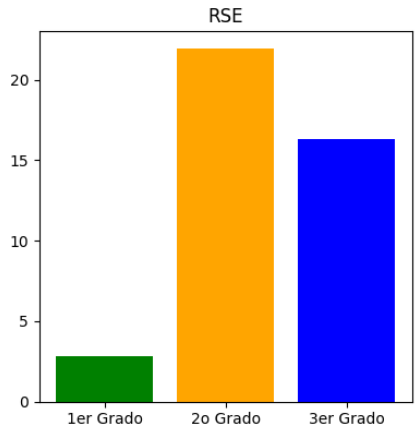
- Close/Last vs Volume

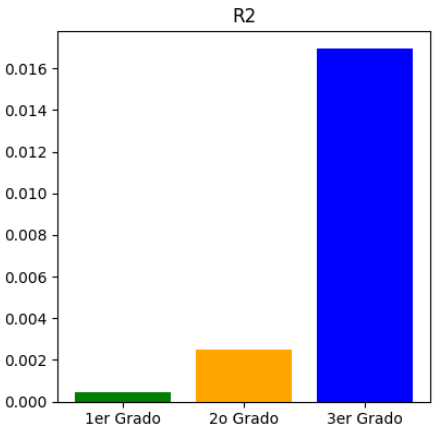


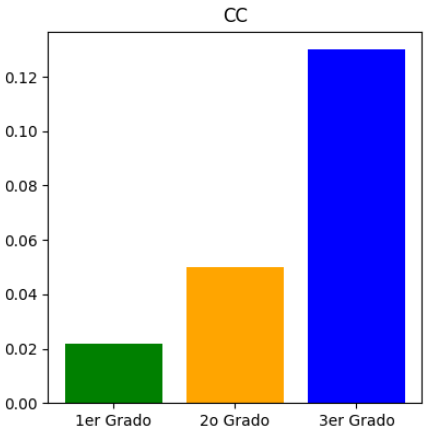




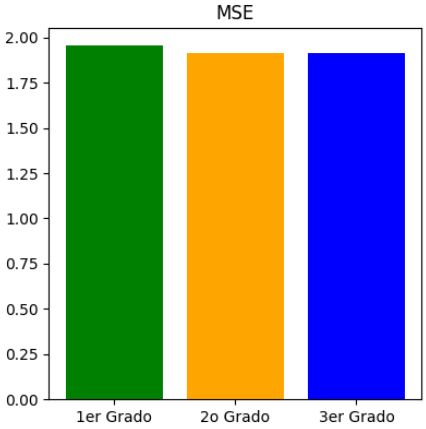


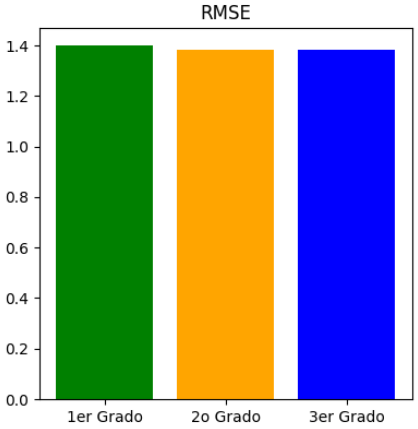


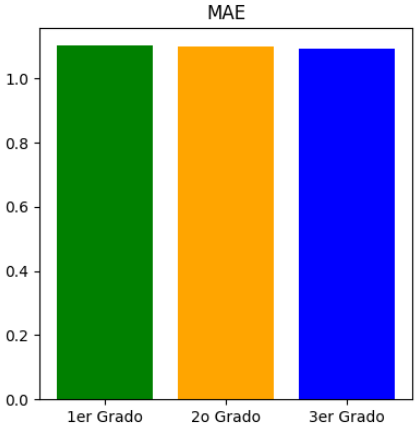


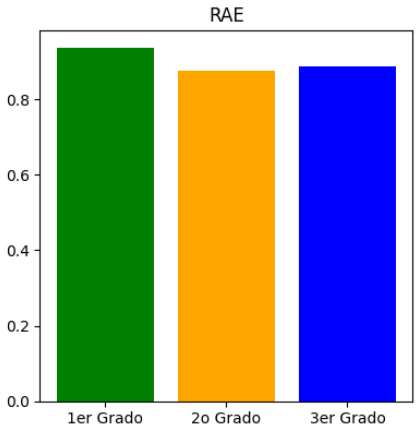


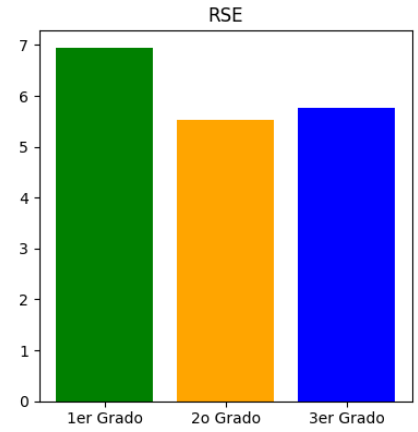
- Close/Last vs Open

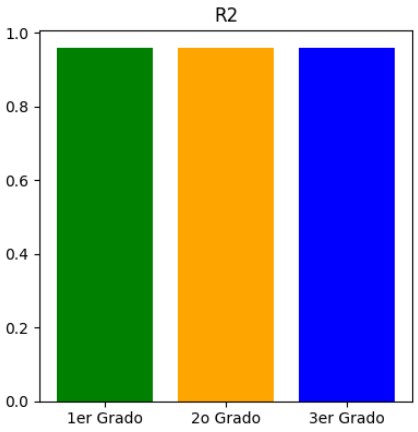


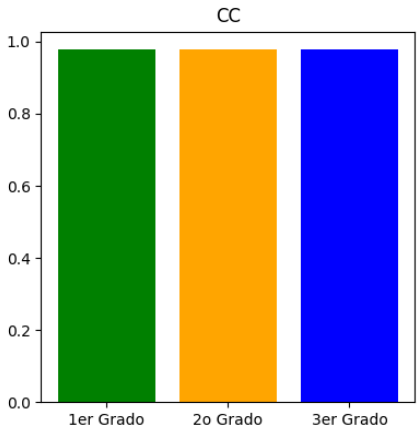




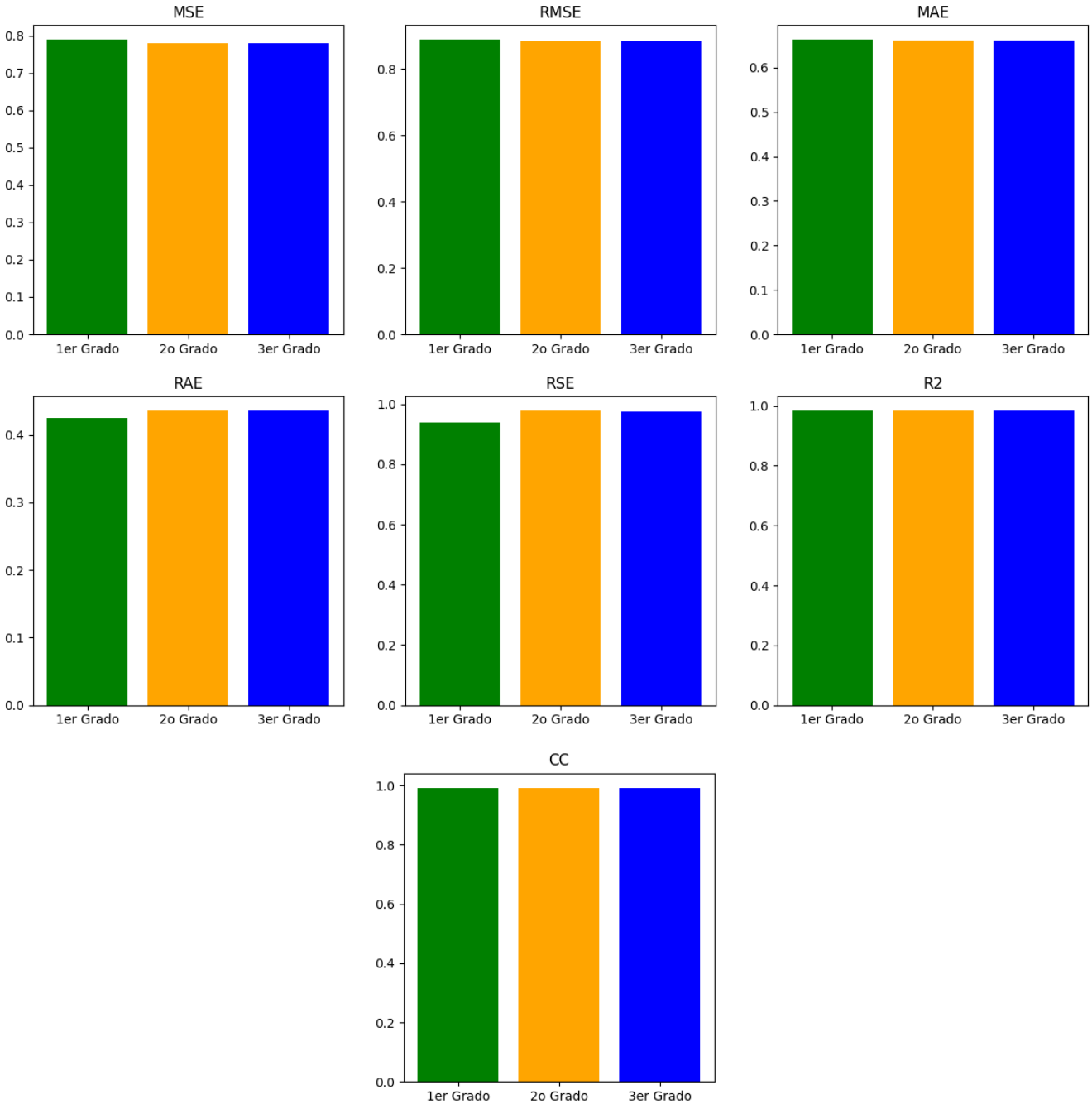




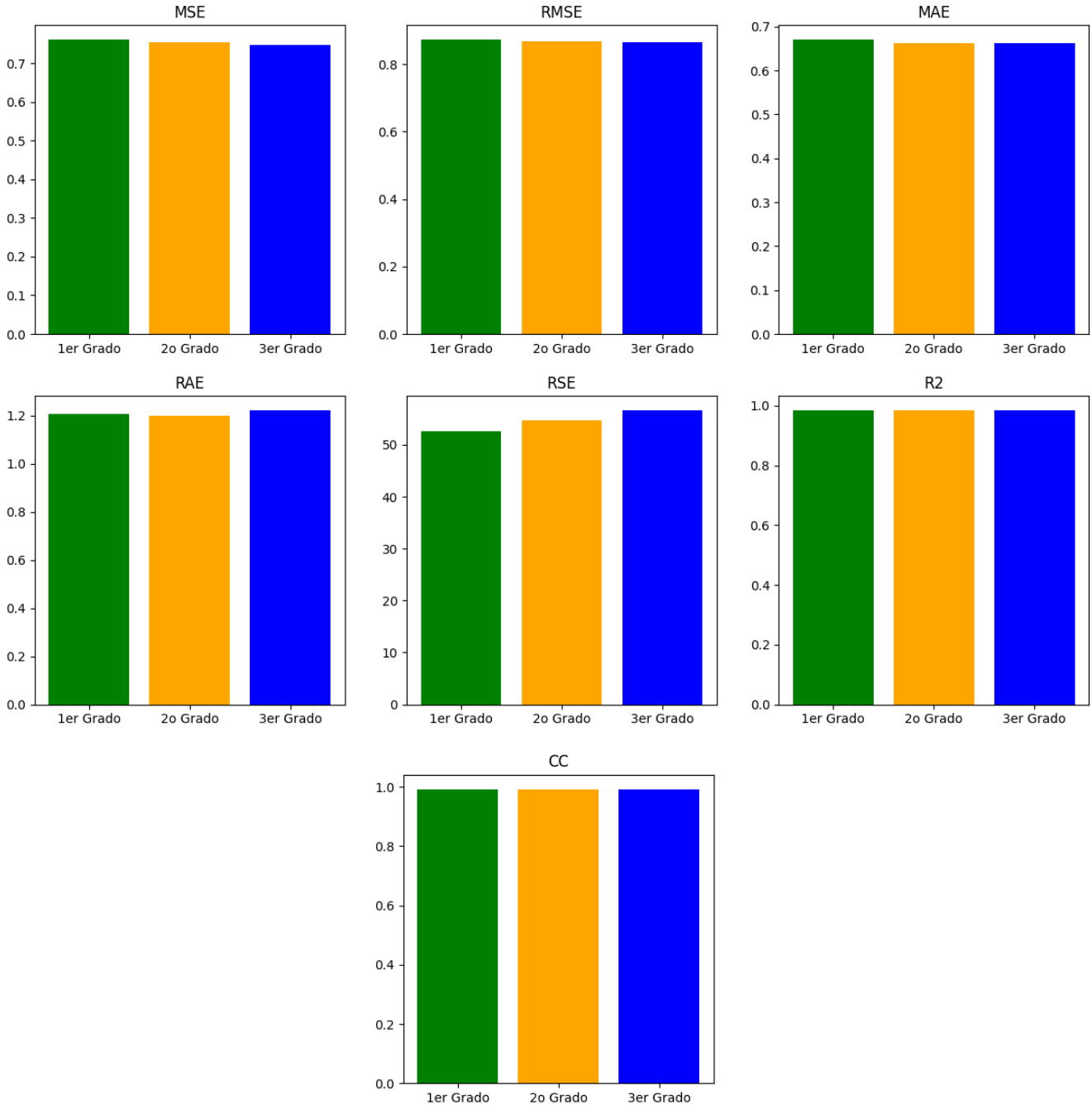




- Close/Last vs High



- Close/Last vs Low



Análisis de Resultados

Para poder analizar las métricas se aplicaron regresiones lineales y polinomiales, el objetivo era observar cómo disminuía o aumentaba el error al cambiar el ajuste de la línea en cada regresión de diferente grado. Se puede observar que, de todas las comparaciones entre atributos, las columnas 'Close/Last', 'High' y 'Low' poseen una alta correlación, algo positivo en muchos contextos, pero no tan bueno para poder comparar propiedades de las diversas métricas, por lo que se procederán a analizar específicamente las regresiones de la columna 'Close/Last' vs 'Volume' que son las que más información nos otorgan para realizar análisis de cada métrica debido a su diferencia de unidades y al gran número de outliers en el atributo 'Volume'.

La primera observación importante, es ver cómo disminuyen los dígitos de error primero entre el MSE, RMSE y MAE con siete dígitos de diferencia entre el primero (MSE) y los dos últimos (RMSE y MAE) y después igualmente ver cómo disminuyen considerablemente las demás métricas siendo todas ellas, de uno o dos dígitos con algunos decimales. Esto es importante, debido a que se pueden visualizar las propiedades mencionadas en el marco teórico de cada una, especialmente del MSE en donde se observa cómo penaliza muy fuertemente los errores grandes debido a la alta cantidad de datos atípicos en esta comparación. Por su parte también se puede ver que entre RMSE y MAE hay una diferencia de alrededor de 1 millón en cada regresión, esto ya que MAE no penaliza tan fuertemente los outliers. Destaca igualmente el valor R^2, el cual tiene un valor de 0 en la regresión lineal y polinomial de segundo grado, indicando que el modelo se ajusta nada a los datos, y un valor de 0.02 en la regresión de tercer grado en el que el modelo tiene al menos un ajuste mínimo; también algo parecido pasa con CC aumentando ligeramente el valor de la primera a la tercera regresión, lo cual tiene sentido. Un aspecto curioso es lo que sucede con las métricas RAE y RSE, en donde en lugar de disminuir el error, parece aumentar en cada regresión de mayor grado (RAE) o aumenta y disminuye después (RSE). En todo caso se puede observar que siempre es mayor el RSE, debido a que penaliza los outliers.

Conclusiones

En esta práctica se realizó un análisis de algunas de las diferentes métricas que existen para tareas de predicción. A pesar de que se hicieron varias regresiones comparando algunos atributos, se decidió analizar específicamente las regresiones de 'Close/Last' vs 'Volume' debido a que estas arrojaban ciertos datos útiles que permitían destacar algunas características de cada métrica. Finalmente se observó que, si bien el resultado de cada métrica puede variar mucho, en realidad cada una tiene su aplicación, por lo que no existe algún tipo de métrica que sirva en todos los casos y dependerá del contexto.

Bibliografía

[1] M. A. Fernandez, «Inteligencia Artificial para Programadores con Prisa,» 2021.