Proyecto 2

Descripción del DataFrame

Nombre: IMDB data from 2006 to 2016.

Es sobre películas grabadas entre los años 2006 y 1016.

Tiene 1000 registros y 12 columnas.

Variables

Rank: Posición en el ranking de la película, que indica su popularidad o clasificación en el conjunto de datos.

Title: Título de la película.

Genre: Géneros de la película, listados. Una película puede pertenecer a múltiples géneros.

Description: Descripción breve de la trama de la película.

Director: Nombre del director de la película.

Actors: Nombres de los actores de la película separados por comas.

Year: Año de lanzamiento de la película.

Runtime (Minutes): Duración de la película en minutos.

Rating: Calificación promedio de la película, basada en votaciones de las personas.

Votes: Número total de votos que ha recibido la película.

Revenue (Millions): Ingresos de taquilla de la película en millones de dólares.

Metascore: Puntuación de la película proporcionada por Metacritic, que agrega críticas de fuentes seleccionadas para dar una puntuación ponderada.

Variable respuesta

La variable respuesta que se decidió elegir en este caso es ‘Revenue’ qué son los ingresos totales que obtuvo la película. Es de tipo numérico continuo. De ahí el resto del dataframe se tomaron como variables de X para ayudar a predecir, sin embargo se eliminaron 2 ya que no tenían ninguna relevancia con la predicción de los datos. Se eliminó Title y Description.

Transformaciones

Transformación de Características Numéricas

Imputación

Estandarización

Transformación de Características Categóricas

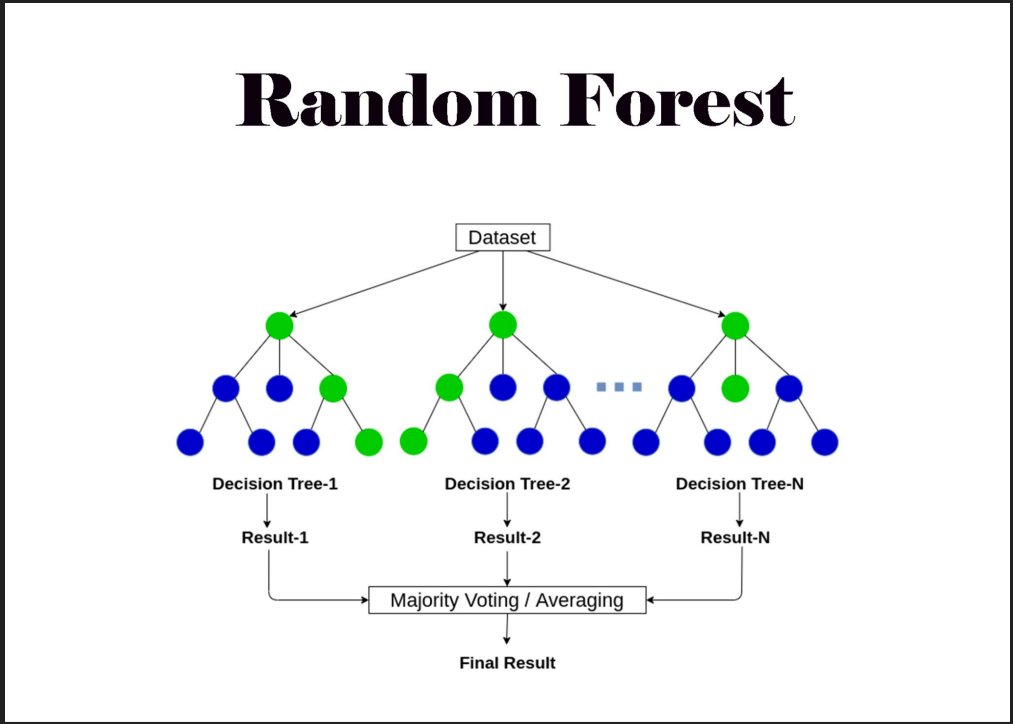
Imputación:

Codificación One-Hot

Modelos Seleccionados

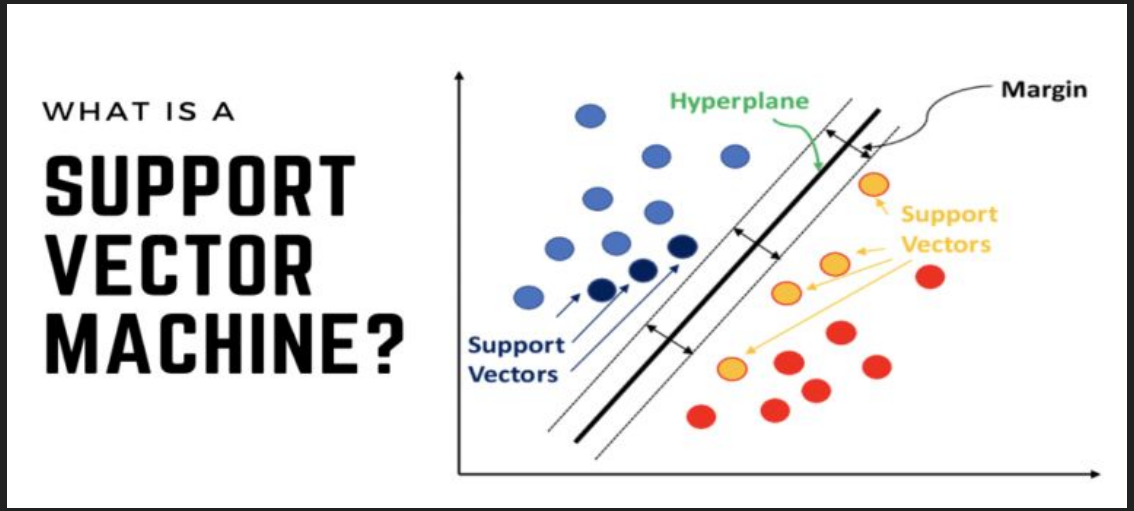
Random Forest:

El modelo es útil para manejar una variedad de tipos de datos y relaciones, proporcionando una buena métrica de la importancia de las características.



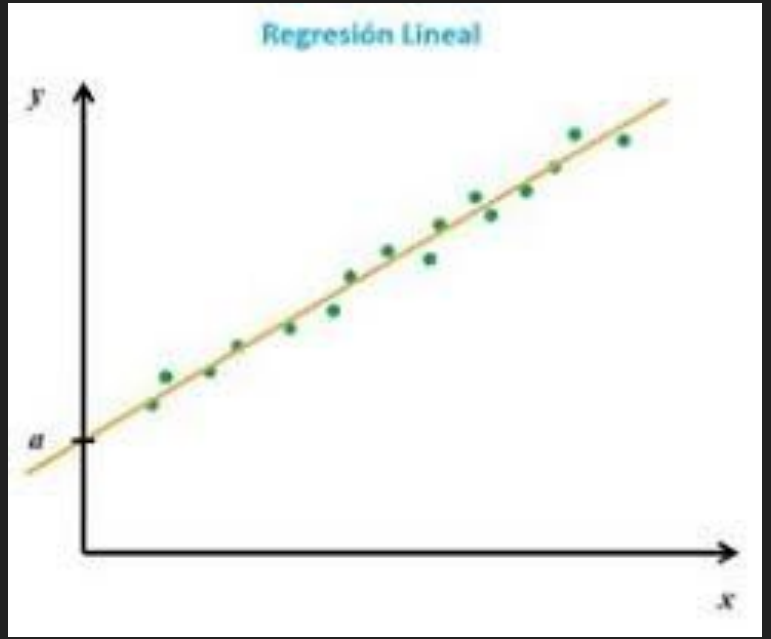
SVM:

Este modelo es efectivo en espacios de alta dimensión y en casos donde el número de dimensiones es mayor que el número de muestras



Regresión lineal:

Un modelo de aprendizaje automático que intenta predecir un valor objetivo mediante la combinación lineal de características de entrada.



Resultados

Random Forest:

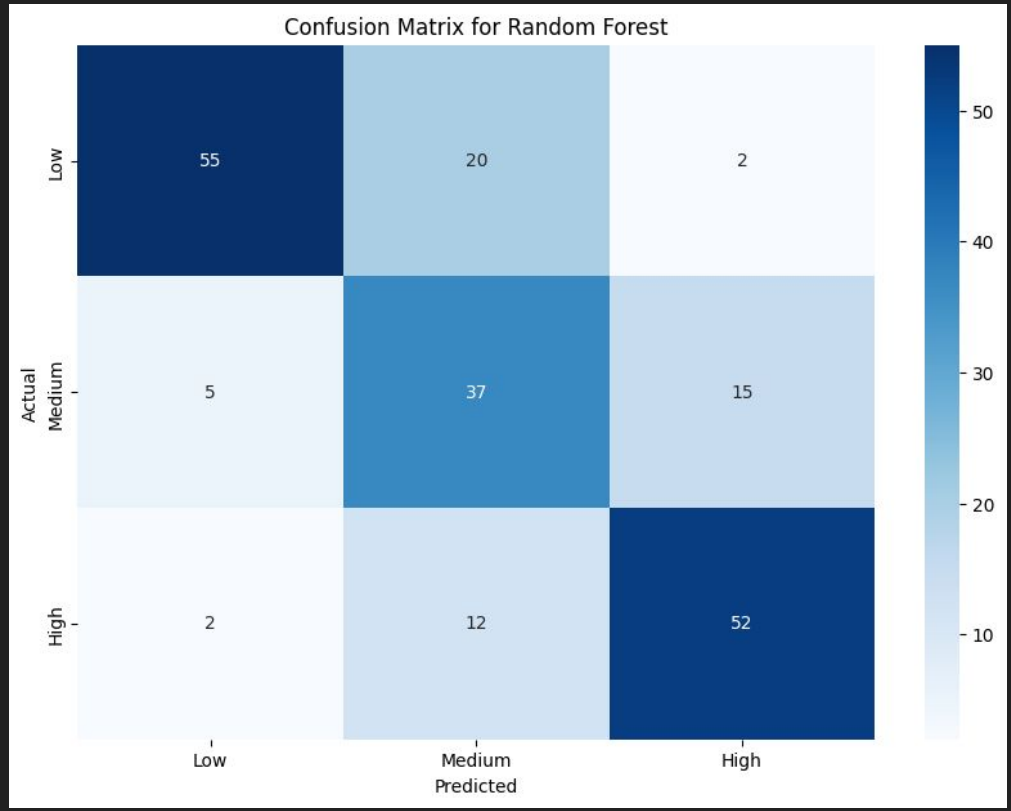
MAE: 38.744454701898576,

MSE: 4534.268309468408,

RMSE: 67.33697579687113,

MAPE: 9204.947929421971,

R2: 0.5568600932099148



Resultados

SVM:

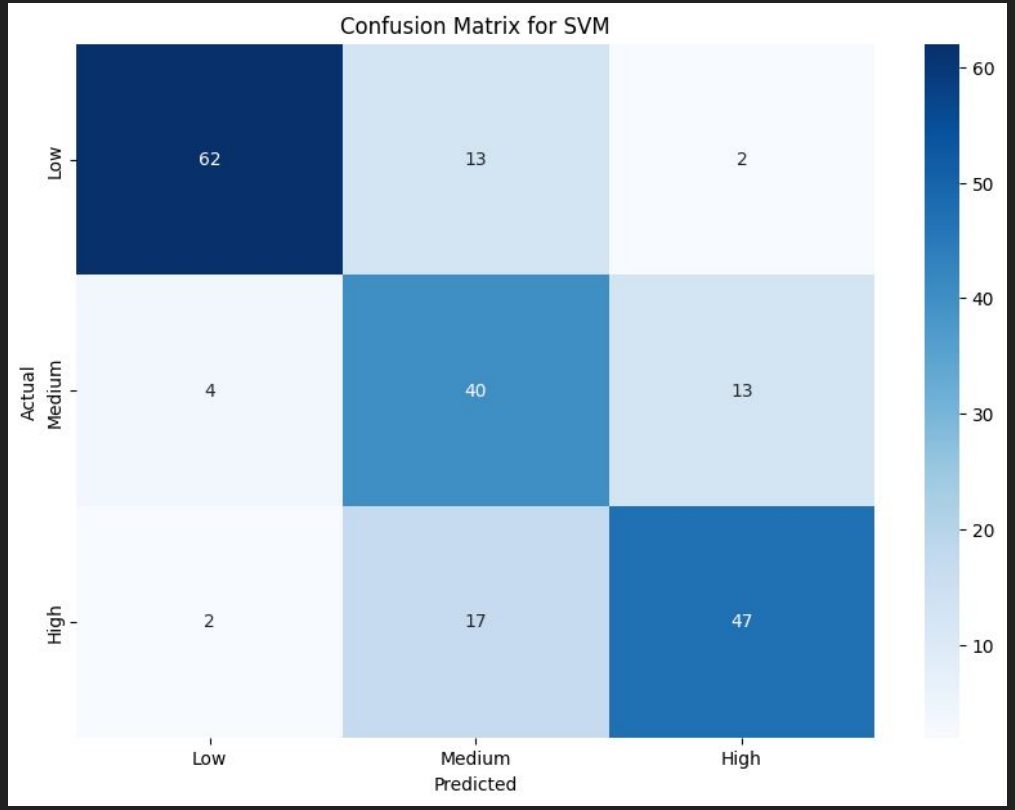
MAE: 39.43147760439113,

MSE: 4677.005936350044,

RMSE: 68.388638942079,

MAPE: 8770.10747322437,

R2: 0.5429101603090138



Resultados

Linear Regression:

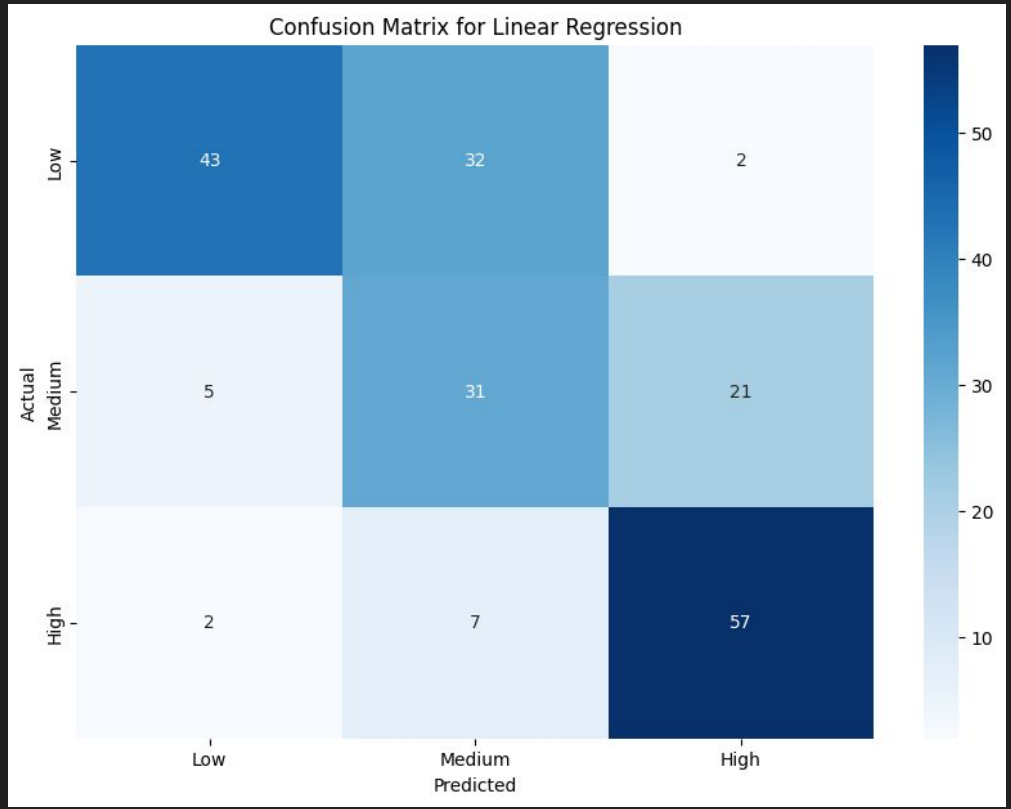
MAE: 38.73809139744471,

MSE: 3442.974023914162,

RMSE: 58.67686106050802,

MAPE: 14966.43849183058,

R2: 0.6635136952852974



Conclusiones Generales

Después de varios intentos de mejorar las predicciones de los modelos, se concluyó que los resultados obtenidos son los mejores posibles, aunque no perfectos debido a factores inherentes a los datos. Se aplicaron diversos hiperparámetros y modificaciones en las columnas, logrando el mejor resultado con el proceso final.

La Regresión Lineal tiene la mejor capacidad explicativa, pero su alto MAPE indica problemas con valores atípicos. Random Forest ofrece un equilibrio razonable en la clasificación de los extremos, aunque necesita mejoras en la categoría media. SVM tiene un rendimiento comparable a Random Forest, pero requiere reducir las grandes desviaciones en sus predicciones. La elección del modelo depende del contexto específico y del balance deseado entre precisión general y manejo de valores atípicos.