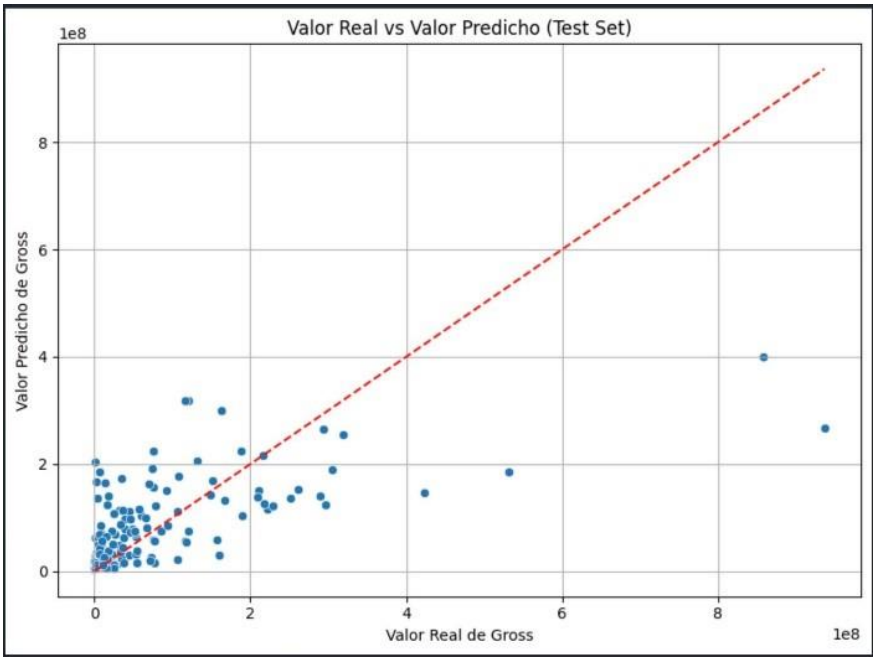
Seleccionamos **Random Forest Regressor** porque, además de ser robusto frente a datos ruidosos y no lineales, y no requerir normalización debido a su capacidad para manejar variables con diferentes escalas, presenta varias ventajas relevantes para el análisis y predicción en el contexto de datos complejos como los de Netflix:

- Capacidad para predecir calificaciones a partir de múltiples atributos
  Las películas en Netflix tienen diversas características (duración, país, año de lanzamiento, género, tipo, etc.). Random Forest puede combinar todas estas variables, incluso si tienen relaciones complejas o no lineales, para estimar con precisión una variable objetivo como la calificación (rating).
- Facilita la segmentación de contenido basada en características importantes
  Gracias a la evaluación de importancia de variables, Random Forest permite identificar qué atributos (como género, año o país) influyen más en el comportamiento del contenido, lo que puede guiar la segmentación de películas en grupos más representativos o estratégicos.
- Robustez ante entradas parcialmente faltantes o codificadas

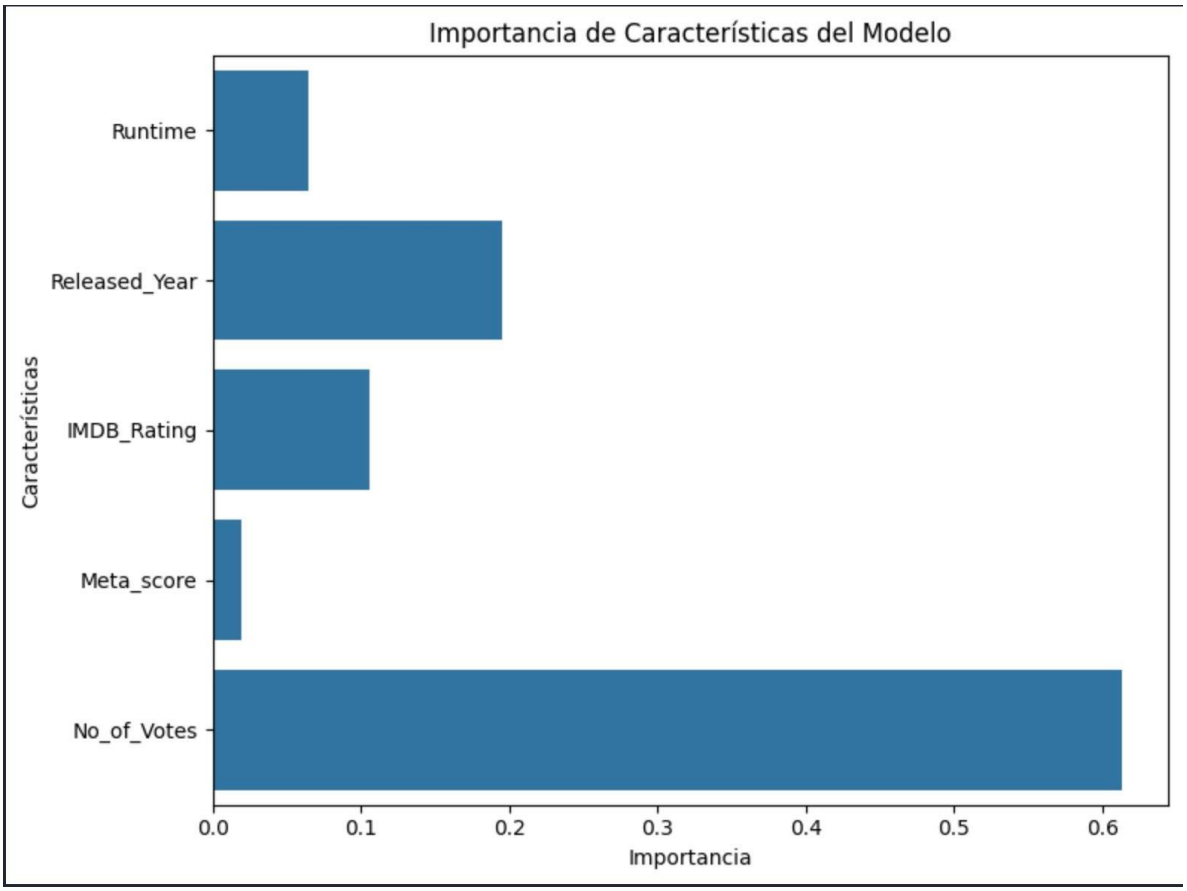
  En muchos casos, algunos atributos pueden no estar presentes en todas las películas (por ejemplo, país o duración). Random Forest puede seguir funcionando sin necesidad de eliminar muchas filas o aplicar imputaciones agresivas, lo cual preserva el valor del dataset original.

- Resistencia al “ruido” en datos provenientes de descripciones no estructuradas
  Algunas variables pueden venir de procesamiento de texto (como sinopsis, categorías, etiquetas), que tienden a ser ruidosas. Random Forest tolera este ruido mejor que modelos lineales o de menor complejidad.
- Ideal para prototipos rápidos y producción escalable
  Dado que no requiere normalización ni una ingeniería de características intensiva, es ideal para prototipar rápidamente. Además, su entrenamiento es paralelizable, lo cual permite escalar si se desea aplicar a todo el catálogo de Netflix.



Esta gráfica muestra la comparación entre los valores reales y los valores predichos de la recaudación bruta (“Gross”) de películas utilizando un modelo de árbol de decisión. En el eje horizontal se representan los valores reales y en el eje vertical los valores que el modelo predijo. La línea roja discontinua representa el caso ideal donde las predicciones serían exactamente iguales a los valores reales.

La mayoría de los puntos azules están cercanos a esa línea en valores bajos, lo que indica que el modelo funciona relativamente bien para películas con recaudaciones moderadas o bajas. Sin embargo, a medida que los valores de recaudación aumentan, los puntos se dispersan más, lo que evidencia que el modelo pierde precisión al predecir películas con altos ingresos. Esto es común en los árboles de decisión, ya que tienden a simplificar en exceso las predicciones cuando los datos son muy variables. En general, el modelo acierta en estimaciones promedio, pero presenta errores considerables en casos extremos.



Esta gráfica muestra la importancia relativa de las características utilizadas por el modelo de árbol de decisión para predecir la recaudación bruta de las películas. La variable más influyente con diferencia es *No_of_Votes* (número de votos), con una importancia superior al 60 %, lo que indica que el modelo se basa principalmente en la cantidad de personas que votaron para estimar los ingresos.

La segunda característica más relevante es *Released_Year* (año de lanzamiento), seguida por *IMDB_Rating*, aunque ambas tienen una influencia mucho menor. *Runtime* (duración) también aporta algo de valor, pero en menor medida. Por otro lado, *Meta_score* (puntuación crítica) es la variable menos relevante para el modelo, con una importancia casi nula. En resumen, el modelo considera que la popularidad (medida por el número de votos) es el mejor predictor de la recaudación, más que la calidad crítica o la duración de la película.

```
Fitting 3 folds for each of 36 candidates, totalling 108 fits
Mejores hiperparámetros: {'max_depth': 5, 'min_samples_split': 10, 'n_estimators': 50}
Validación - MSE: 6149126977158980.0, R2: 0.5111358627966409
Prueba - MSE: 1.0045800738067074e+16, R2: 0.41534639520946537
```

Predicción con nuevos datos

```
# Crear un nuevo ejemplo para predecir
nuevo_dato = pd.DataFrame([{
    'Runtime': 120,          # Duración de la película en minutos
    'Released_Year': 2020,   # Año de lanzamiento
    'IMDB_Rating': 8.5,      # Calificación en IMDB
    'Meta_score': 80,        # Puntuación en Metascore
    'No_of_Votes': 500000    # Número de votos
}])
```

```
Ingreso bruto predicho para el nuevo dato: $206,121,391.10
```