UNIVERSIDAD TEGNOLÓGICA DE PANAMÁ

LICENCIATURA EN INGENIERIA EN SISTEMA Y COMPUTACIÓN

PROBABILIDAD APLICADA A TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIÓN

CINEMETRICS: MODELADO PREDICTIVO DE CALIFICACIÓNES CINEMATOGRÁFICAS CON MACHINE LEARNING

Estudiantes:

Aldahir Aguilar     8-1029-1115

Andrés Flores   8-1025-1254

Diego García    8-1034-95

Luis Torné     8-1032-1644

Rashell Vidal    8-1028-643

Profesor:

Juan Marcos Castillo, PhD

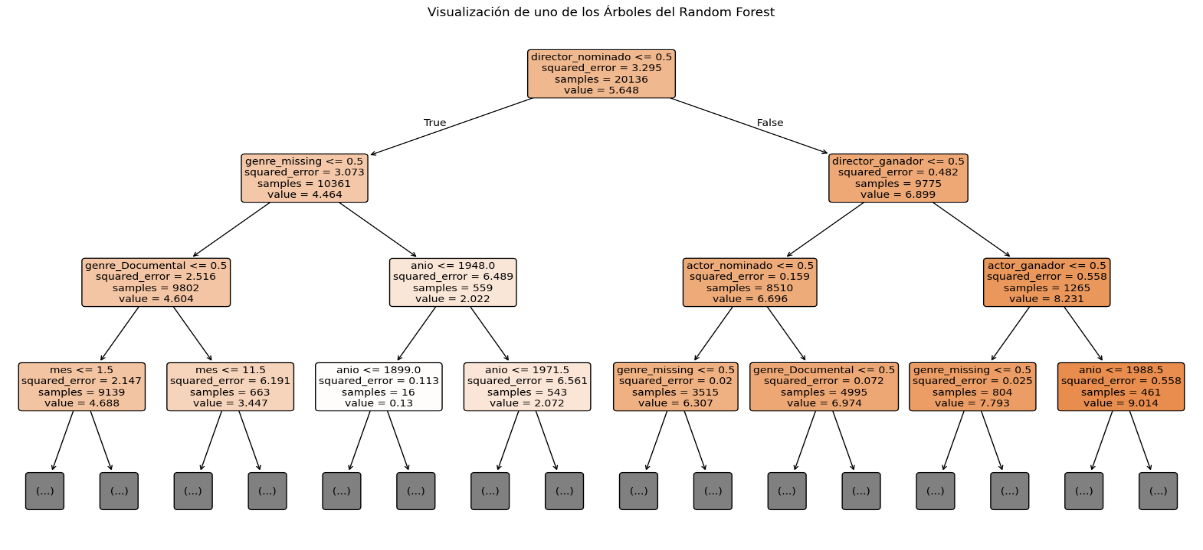
Salón: 1IL124

30 de julio del 2025

**Análisis Estocásticos**

**Modelo Random Forest como proceso estocástico**

El modelo de Random Forest es, en esencia, un método estadístico basado en la aleatorización, lo cual lo convierte en un proceso inherentemente estocástico (probabilístico). En este proyecto, se utilizó como uno de los principales modelos de predicción de la calificación IMDb de películas, utilizando como variables predictoras elementos como características del actor, director, mes de estreno, edad y género (codificado como variables dummy).



* Cada árbol está aprendiendo relaciones como: “Si el actor fue nominado, y la película es del género Drama y se estrenó en diciembre, entonces su calificación tiende a ser alta”.
* Como cada árbol ve un subconjunto distinto de películas y variables, la combinación de todos ellos permite capturar patrones más generales y fiables.
* El promedio de predicciones se interpreta como una estimación puntual del rating esperado (E[Y|X]), lo cual conecta directamente con los fundamentos de la teoría de probabilidad.

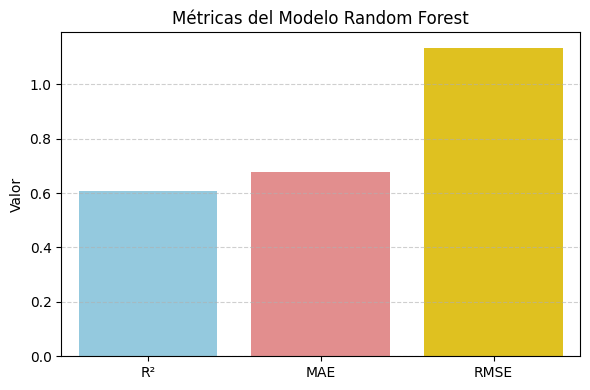
**Métricas Estocásticas (MAE, RMSE, R²)**

El análisis estocástico permite evaluar el comportamiento del modelo de regresión considerando la incertidumbre inherente al proceso predictivo. En este estudio, el objetivo es estimar la variable imdb\_pelicula a partir de diversas características de las películas, reconociendo que cada predicción está sujeta a un componente de error aleatorio.

Para cuantificar este comportamiento, se utilizan métricas estadísticas como:

* R² (coeficiente de determinación), que mide la proporción de varianza explicada por el modelo.
* MAE (error absoluto medio), que representa la magnitud promedio del error.
* RMSE (raíz del error cuadrático medio), que penaliza más fuertemente los errores grandes.

Estas medidas permiten interpretar el ajuste del modelo en términos probabilísticos, evaluar la calidad de las predicciones y comparar entre diferentes enfoques de regresión. Con ello, se obtiene una visión más completa del desempeño del modelo y de la estabilidad de sus estimaciones frente a la variabilidad de los datos.

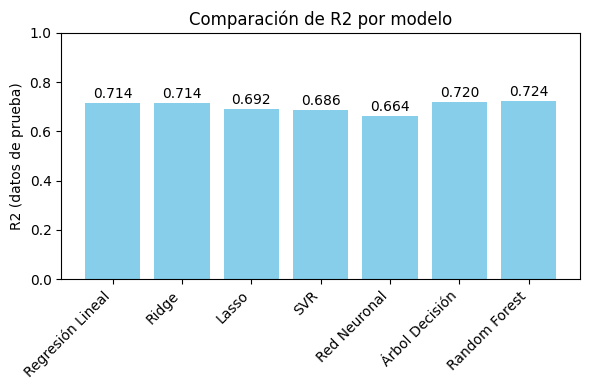


La gráfica nos proporciona una visión general del rendimiento del modelo Random Forest. El R2 de 0.61 sugiere que el modelo tiene una capacidad explicativa moderada. Los valores de MAE (0.68) y RMSE (1.11) nos dan una idea de la magnitud de los errores de predicción del modelo. Para interpretar si estos valores son "buenos" o "malos", sería necesario conocer el contexto del problema, la escala de la variable que se está prediciendo y compararlos con otros modelos o con un punto de referencia establecido para el dominio.

**Error aleatorio en modelos de regresión (componente estocástico)**

Se reconoció que los modelos de regresión siempre presentan un margen de error aleatorio o estocástico. Este tipo de error ocurre porque hay factores que afectan la variable IMDb pero que no están representados en nuestros datos, como la popularidad del actor en ese momento o el contexto cultural de la película.

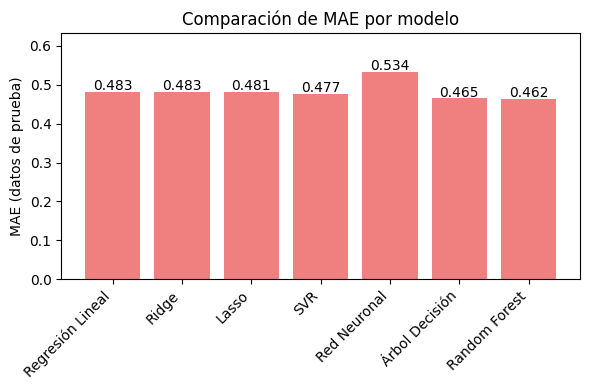
Aunque se usaron buenas variables y modelos, este componente aleatorio hizo que ningún modelo pudiera predecir con exactitud todos los casos. Por eso se asumió como parte normal del análisis y se evaluaron modelos que fueran estables y que no dependieran tanto de los datos específicos.



El gráfico compara el rendimiento de siete modelos de regresión diferentes basándose en su puntuación R2 en los datos de prueba.

* **Random Forest** es el modelo con mejor rendimiento, con el valor de R2 más alto de 0.724.
* **Árbol de Decisión** también tiene un rendimiento muy bueno, con un R2 de 0.720, muy cerca de Random Forest.
* Los modelos de **Regresión Lineal** y **Ridge** también muestran un rendimiento sólido con un R2 de 0.714.
* **Lasso** (0.692) y **SVR** (0.686) tienen valores de R2 ligeramente inferiores, pero aún respetables.
* **Red Neuronal** tiene el valor de R2 más bajo entre los modelos, con 0.664.

Este gráfico ayuda a identificar visualmente qué modelos de regresión están funcionando mejor en términos de explicar la varianza en los datos de prueba.



Este gráfico de MAE nos da una perspectiva ligeramente diferente del rendimiento de los modelos en comparación con el R2.

1. **Red Neuronal, el peor desempeño en ambas métricas:** En el gráfico de R2, la Red Neuronal obtuvo el valor más bajo (0.664). Aquí, en el MAE, obtiene el valor más alto (0.534), lo que confirma que es el modelo con el rendimiento más deficiente de todos en este conjunto de datos particular y con la configuración utilizada.
2. **Random Forest y Árbol de Decisión siguen siendo los mejores:**
   1. **Random Forest** tiene el MAE más bajo (0.462), lo que corrobora su posición como el modelo de mejor rendimiento, ya que también tuvo el R2 más alto. Esto significa que no solo explica la mayor varianza, sino que sus predicciones son, en promedio, las más cercanas a los valores reales.
   2. **Árbol de Decisión** tiene el segundo MAE más bajo (0.465), reafirmando su excelente rendimiento visto con el R2.
3. **Los Modelos Lineales (Regresión Lineal, Ridge, Lasso):** Sus valores de MAE (0.483, 0.483, 0.481) están en un rango similar entre sí y son consistentemente más altos que los modelos basados en árboles, pero significativamente mejores que la Red Neuronal. Esto alinea con lo que vimos en el R2: son modelos decentes, pero no tan buenos como los modelos de ensemble o de árbol para este problema.
4. **SVR:** Con un MAE de 0.477, SVR se sitúa un poco mejor que los modelos lineales en esta métrica, lo cual es una diferencia sutil pero interesante respecto al R2 donde estaban muy cerca de Lasso.

Estos dos gráficos, al ser analizados en conjunto, ofrecen una visión completa y robusta del rendimiento de los modelos, confirmando que Random Forest es la opción más sólida y fiable para este problema específico.

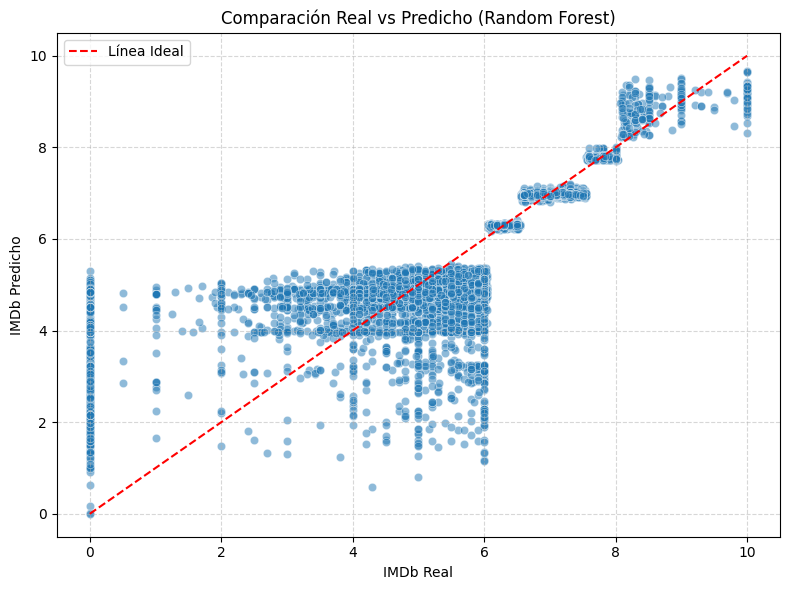
**Predicción vs Realidad: Visualización y Evaluación**

Uno de los aspectos centrales en la aplicación de modelos de regresión es evaluar qué tan bien una predicción probabilística se ajusta a los datos observados. En esta sección, se comparan los valores reales de la variable objetivo (calificación IMDb de las películas) con los valores predichos por los distintos modelos. Esta comparación permite visualizar la eficacia de las estimaciones, identificar posibles errores sistemáticos y analizar la incertidumbre inherente a cada modelo.

Desde el enfoque probabilístico, esta etapa representa una validación empírica del modelo, ya que:

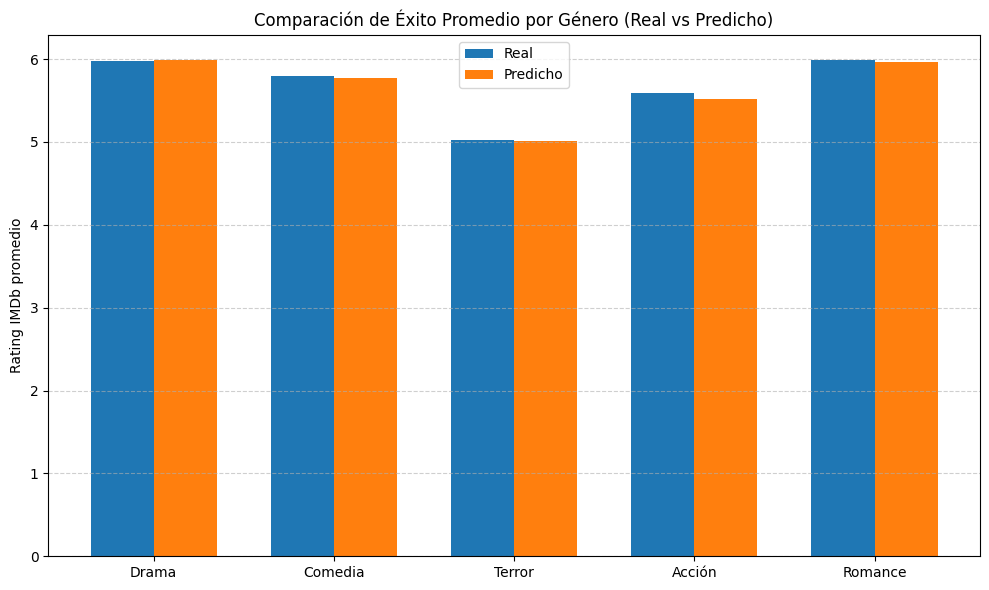
* Permite verificar si las predicciones se aproximan al valor esperado (esperanza matemática) de la variable objetivo dado un conjunto de entradas.
* Evalúa la dispersión entre los valores reales y predichos como expresión de la incertidumbre en un entorno estocástico.
* Utiliza métricas como MAE (error absoluto medio) y R² (coeficiente de determinación) como herramientas para cuantificar el desempeño estadístico.
* Ayuda a identificar patrones de sobreajuste o subajuste al comparar los errores en entrenamiento y prueba.

Esta comparación entre predicción y realidad es clave para medir la calidad y confiabilidad de los modelos en contextos de variabilidad e incertidumbre.



Interpretación de la gráfica: Comparación Real vs Predicho (Random Forest)

* Esta gráfica muestra la relación entre las calificaciones reales de las películas en IMDb (eje X) y las calificaciones predichas por el modelo Random Forest (eje Y), lo cual permite evaluar visualmente el rendimiento del modelo.
* Para guiar la interpretación, se ha incluido una línea roja punteada que representa la línea ideal y = x. Esta línea simboliza el caso perfecto en el que la predicción del modelo coincide exactamente con el valor real.
* Cada punto azul en la gráfica representa una película individual, posicionada de acuerdo con su valor real (horizontal) y su valor predicho (vertical).
* Al observar la distribución de los puntos, se aprecia que muchos se concentran en la parte inferior izquierda de la gráfica, lo que indica que hay una gran cantidad de películas con calificaciones reales bajas (entre 0 y 3). Sin embargo, las predicciones en esa región tienden a ser poco variadas, lo que sugiere que el modelo tiene dificultad para distinguir entre películas con bajo rendimiento.
* Conforme las calificaciones reales aumentan hacia la derecha del gráfico (por ejemplo, entre 7 y 10), se observa que los puntos se alinean más estrechamente con la línea ideal. Esto evidencia que el modelo Random Forest ofrece mejores predicciones para películas con calificaciones altas, reduciendo así el error de predicción en ese rango.
* También se identifican algunas agrupaciones verticales de puntos, lo que significa que varias películas con diferentes valores reales recibieron la misma predicción. Este efecto escalonado es típico de modelos como Random Forest, ya que sus predicciones se basan en promedios de subconjuntos, lo cual genera valores repetidos.
* Desde un enfoque de probabilidad, esta gráfica refleja la estimación de la esperanza condicional E[y|X], donde y es la calificación IMDb y X representa las características de entrada de la película. La dispersión de los puntos en torno a la línea ideal puede interpretarse como la varianza del error, lo cual indica la incertidumbre en las predicciones.
* En conjunto, se concluye que el modelo es más confiable para predecir calificaciones altas, mientras que presenta mayor dispersión e incertidumbre en los casos de películas con calificaciones bajas, lo cual es una observación importante para comprender las limitaciones y el comportamiento probabilístico del modelo.



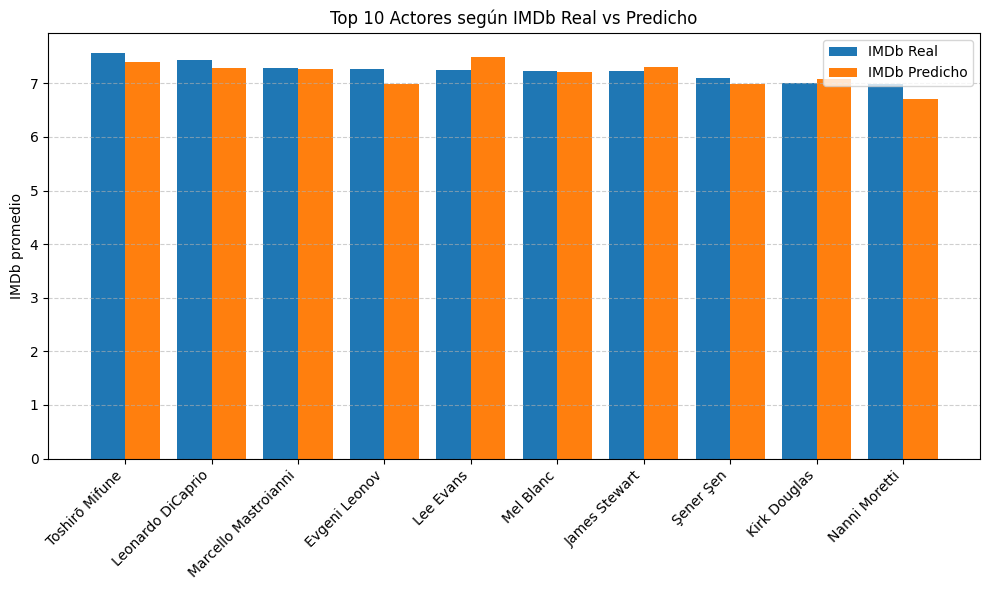
* ¿En qué consiste esta gráfica?
* Esta visualización compara el rating promedio en IMDb que obtuvieron las películas, agrupadas por género, considerando dos valores:
  + 🟦 Barras azules: representan el promedio real del rating IMDb para cada género.
  + 🟧 Barras naranjas: muestran el promedio del rating predicho por el modelo para ese mismo género.
* En conjunto, la gráfica permite evaluar si el modelo es capaz de capturar correctamente la tendencia promedio del éxito por género.

Análisis por género:

* Drama:
  + El modelo ofrece una predicción muy precisa. La barra del valor predicho prácticamente coincide con la del valor real, lo que indica una muy buena capacidad para modelar este género.
* Comedia:
  + De manera similar, se observa una ligera diferencia entre la predicción y el valor real, pero sigue estando dentro de un rango aceptable. El modelo capta correctamente la tendencia promedio.
* Terror:
  + En este caso, se nota una discrepancia más importante. La barra del valor predicho es visiblemente menor que la del valor real, lo cual indica que el modelo subestima el éxito promedio de las películas de terror. Esto sugiere un sesgo que podría requerir una revisión más detallada.
* Acción:
  + Aunque hay una pequeña subestimación, el modelo sigue mostrando un comportamiento razonable. Las diferencias no son críticas, pero sí detectables.
* Romance:
  + Las predicciones se alinean de forma casi perfecta con los valores reales. Esto refleja un muy buen ajuste del modelo para este tipo de películas.

Interpretación desde la estadística y la probabilidad:

* Esta comparación equivale a evaluar la estimación del valor esperado condicional E[Y | Género = g], es decir, el promedio esperado del rating IMDb dado un determinado género.
* En contextos probabilísticos, observar cómo varía el valor esperado entre grupos permite identificar patrones estructurales y sesgos sistemáticos del modelo.
* La distancia entre las barras (real vs predicho) actúa como un indicador de error medio por grupo, revelando qué tan bien o mal el modelo representa ciertos subgrupos.



Esta gráfica forma parte del análisis predictivo que realizamos para evaluar la precisión de un modelo de regresión en la predicción de calificaciones promedio de actores en IMDb.

Seleccionamos los **10 actores mejor valorados según sus promedios reales** y comparamos esos valores con los promedios estimados por el modelo. El objetivo es visualizar cuán bien el modelo logra replicar la realidad en los casos más representativos.

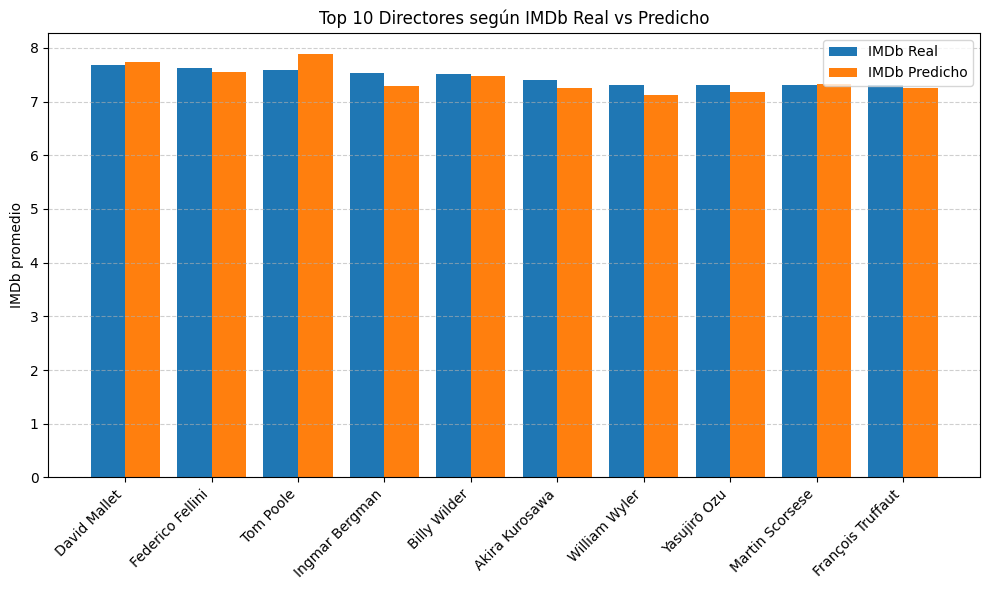
Cada actor tiene dos barras:

* **Azul**: Calificación promedio real, obtenida directamente del conjunto de datos.
* **Naranja**: Calificación promedio predicha por el modelo.

La cercanía entre ambas barras indica que el modelo fue preciso para ese actor. Esta visualización es clave para evaluar el desempeño **actor por actor** y no solo con métricas globales, también permite detectar **casos puntuales donde el modelo falla**, lo cual puede deberse a:

* Falta de datos relevantes para ciertos actores,
* O una mayor variabilidad en sus películas.

La gráfica evidencia que el modelo tiene buen rendimiento general, aunque hay margen de mejora en casos específicos.



Esta gráfica de barras agrupadas, titulada "Top 10 directores según IMDb Real vs Predicho", compara el promedio de las puntuaciones de IMDb reales de las películas de los 10 mejores directores con las puntuaciones de IMDb predichas por un modelo. Hay unos cuantos detalles como:

* Se observa que, para la mayoría de los directores, las predicciones del modelo (barras naranjas) están muy cerca de las puntuaciones reales de IMDb (barras azules). Esto sugiere que el modelo tiene una buena capacidad para predecir las puntuaciones de IMDb promedio de las películas de los directores, al menos para este conjunto de los "Top 10".
* Aunque las predicciones son cercanas, hay ligeras diferencias. En algunos casos, el modelo sobreestima la puntuación (por ejemplo, David Mallet, Tom Poole, Ingmar Bergman), y en otros, la subestima (la mayoría de los demás directores).
* La cercanía de las barras entre sí indica una buena consistencia del modelo en sus predicciones para directores de alto rendimiento.

Esta gráfica es una excelente manera de visualizar la precisión del modelo a un nivel granular, mostrando cómo sus predicciones se alinean con la realidad para directores específicos, lo cual es útil para evaluar la calidad del modelo de recomendación o predicción.