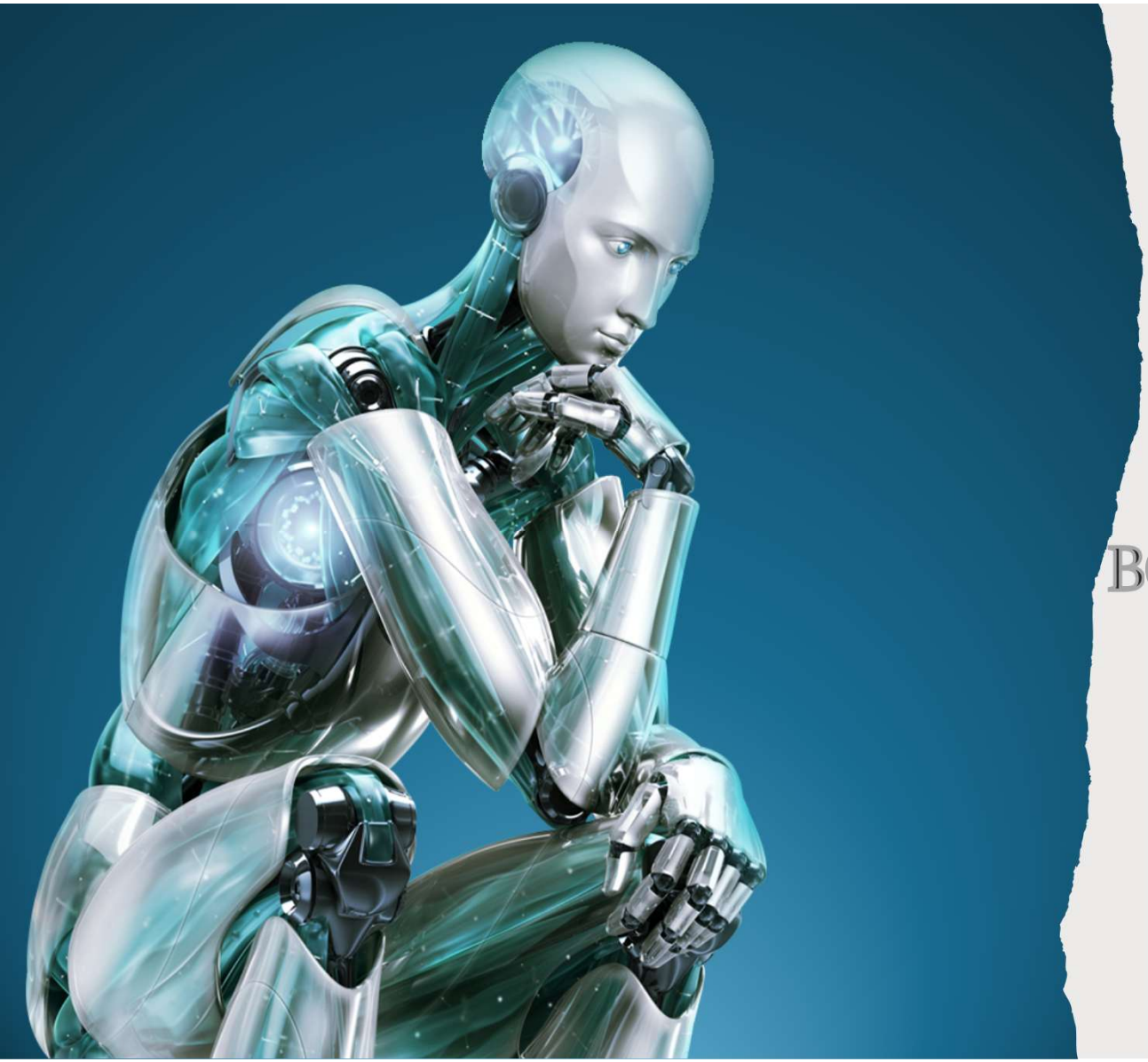

PROYECTO 7
UDD
BOOTCAMP DATA SCIENCE

Juan Carlos Soto Higuera

Objetivos

- Aplicar con éxito todos los conocimientos que has adquirido a lo largo del Bootcamp.
- Consolidar las técnicas de limpieza, entrenamiento, graficación y ajuste a modelos de Machine Learning.
- Generar una API que brinde predicciones como resultado a partir de datos enviados.

Proceso utilizado

Se ha seleccionado el método RandomForestClassifier debido a su capacidad para gestionar variables categóricas y capturar interacciones complejas entre las características. Este modelo facilita la interpretación mediante la evaluación de la importancia relativa de las características, ofrece flexibilidad en el manejo de datos faltantes y en situaciones de desbalance de clases, y suele proporcionar un rendimiento elevado sin requerir un preprocesamiento exhaustivo de los datos. Estas propiedades lo hacen particularmente adecuado para el análisis y la predicción de resultados en conjuntos de datos con múltiples factores interdependientes, como es el caso de los premios Oscar

Marco de trabajo y análisis

El análisis completo se desarrolló en un archivo Colab de Google disponible en la siguiente ruta:

- https://github.com/JuanCarlos-sh/JCSH_UDDCC_Proyecto_M7/tree/main

En donde se puede revisar el detalle de los siguientes pasos:

- Habilitación de Drive donde se encuentra disponible la Base y posterior carga de datos.
- Carga de librerías correspondientes al análisis a realizar
- Limpieza y EDA. Donde se realizaron los siguientes pasos: Completitud, Eliminación de columnas innecesarias, revisar duplicados, cambiar formatos de fechas, Correlación de variables.
- Entrenamiento del modelo, utilizando en este caso RandomForest por las características mencionadas anteriormente.
- Graficación y métricas: se grafica Completitud, Mapa de calor de correlación de variables, Matriz de confusión, importancia de características.
- API REST: se importa Flask y Ngrok, se código en ngrok, ruta y método, se convierten los datos a JSON y se envía la solicitud POST a la API

Hallazgos

Con las variables consideradas (eliminando aquellas innecesarias), se paso de una completitud de 77,9% a 100%.

El modelo obtuvo un buen desempeño y capacidad para realizar predicciones acertadas en el conjunto de datos, logrando una tasa de precisión de 93,2%