PROYECTO 7 BOOTCAMP UDD DATA SCIENCE

SAMUEL MONTERO

OBJETIVOS DEL PROYECTO

- 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.

Proyecto elegido

Estadísticas demográficas de los ganadores del premio Oscar de la Academia:

https://www.kaggle.com/datasets/fmejia21/demographics-of-academy-awards-oscars-winners

Explicación dataset

Un conjunto de datos sobre la raza, religión, edad y otros detalles demográficos de todos los ganadores del Oscar desde 1928 en distintas categorías

Elección del modelo

Decidí trabajar con esta base de datos porque me pareció interesante tanto desde una perspectiva de análisis de datos como desde un interés en la industria del cine y la cultura popular. Creo que puede ser útil ver las tendencias el perfil de las personas que ganan los premios con el paso del tiempo, también opino que es un conjunto de datos engañoso debido a que a simple vista parece sencillo de trabajar, pero es mas complicado de lo que parece debido a que hay una raza sobre-expuesta en los ganadores (gente blanca)

Proceso utilizado para clasificar

Elegi el método de RandomForestClassifier debido a su habilidad para manejar variables categóricas y capturar interacciones complejas entre características. Permiten una fácil interpretación a través de la importancia de las características, ofrecen flexibilidad en el manejo de datos faltantes y desbalance de clases, y generalmente proporcionan un alto rendimiento sin la necesidad de preprocesamiento extenso de los datos. Estas cualidades los hacen adecuados para analizar y predecir resultados en datos con múltiples factores influyentes, como es el caso de los premios Oscar.

Esquema de trabajo

El análisis se realizó en la plataforma de Google colab, fue un paso a paso de acuerdo a lo aprendido durante el bootcamp. Se desarrolló de la siguiente manera:

1. Importación de librerías que se usaron

2. EDA
3.Seleccionamos nuestra variable objetivo y decidimos cual método de predicción o clasificación usaremos

4. Codificamos las variables categóricas con valores dummies para poder ingresarlas a nuestro modelo

5. Imputamos valores nulos

6. Empezamos a crear y entrenar nuestro modelo separando en grupos X y Y

Esquema de trabajo

7. Ejecutamos nuestro modelo
8. Realizamos una matriz de confusión para evaluar eficacia
9. Calculamos el Accuracy y el F1-Score
10. Hacemos una segunda grafica, esta vez de importancia de las características
11. Empezamos con el API REST
12. Importamos Flask y ngrok
13. Registré un token único en el portal de ngrok para poder validar el túnel que crearemos
14. Definimos ruta y método para el api
15. creamos la URL, convertimos los datos a JSON y enviamos la solicitud POST a la API