Alumno: Janon Santiago – 106079

Parte I - Análisis Exploratorio (5 puntos):

Deberán realizar 5 visualizaciones **interesantes** que **ayuden a explicar el target** haciendo almenos un plot de cada uno de los siguientes tipos:

- Bar plot (o stacked bar plot o variaciones)
- Violin plot
- Box plot
- Heatmap

Link:

https://colab.research.google.com/drive/1mspHepx-PtumupSrKxiX3gjAl_nN6TCE#scrollTo=zfg0_QlEdxGY

Alumno: Janon Santiago - 106079

Parte II - Machine Learning Baseline (5 puntos):

Utilice **todas las columnas del dataset** (exceptuando columnas que no tenga sentido usar para predecir) con algún encoding donde sea necesario para entrenar una regresión logística, utilizando búsqueda de hiperparametros y garantizando la reproducibilidad de los resultados cuando el notebook corriera varias veces. Conteste las preguntas:

- ¿Cuál es el mejor score de validación obtenido? (¿Cómo conviene obtener el dataset para validar?)
- Al predecir con este modelo para test, ¿Cúal es el score obtenido? (guardar el csv con predicciones para entregarlo después)
- ¿Qué features son los más importantes para predecir con el mejor modelo? Graficar

Link:

https://colab.research.google.com/drive/1BdU6uGunp_3dmsUDptZmrhDIH5ILTm3D

Alumno: Janon Santiago - 106079

Parte III - Random Forest (5 puntos):

Segun el paper con un clasificador basado en Random Forest deberiamos lograr un AUC de 1. Entrenar un Random Forest con búsqueda de hiperparametros que logre un AUC de 1 (¿cómo conviene elegir los datos de validación respecto de los de train?). El modelo debe cumplir las siguientes condiciones:

- Deben utilizar AUC-ROC como métrica de validación.
- Deben medirse solo en validación, **no contra test!!!**
- Deben ser reproducibles (correr el notebook varias veces no afecta al resultado).
- Deben tener un score en validación igual a 1.

Link:

https://colab.research.google.com/drive/1Ysfcm8b9vDLwr5yjWZevzxfnEiK6UgLK

Alumno: Janon Santiago - 106079

Parte IV - Machine Learning (5 puntos):

Entrenar un nuevo modelo (que no sea Random Forest ni el utilizado para el baseline) con búsqueda de hiperparametros (¿cómo conviene elegir los datos de validación respecto de los de train?). El modelo debe cumplir las siguientes condiciones:

- Deben utilizar AUC-ROC como métrica de validación.
- Deben medirse solo en validación, **no contra test!!!**
- Deben ser reproducibles (correr el notebook varias veces no afecta al resultado).
- Deben tener un score en validación superior a 0,9.
- Para el feature engineering debe utilizarse imputación de nulos, mean encoding y one hot encoding al menos una vez cada uno.
- Deben utilizar al menos 40 features (contando cómo features columnas con números, pueden venir varios de la misma variable).
- Deberán contestar la siguientes preguntas:
    - ¿Cuál es el score en test? (guardar el csv con predicciones para entregarlo después)
    - ¿Por qué cree que logro/no logro el mismo valor de AUC que con Random Forest?

Link:

https://colab.research.google.com/drive/1yPBj6IjsPEk4QKEv3n2f96ZdrXGYI13m

Alumno: Janon Santiago - 106079

Punto Extra 4:

- Utilizando los árboles creados por el Random Forest y la importancia de los features, cree un árbol de decisión simple para que una persona normal pueda identificar si un hongo es comestible o no. Qué nivel de error posee este árbol al intentar clasificar un set de datos de testing?

Link:

https://colab.research.google.com/drive/1CmS7PmMAwfGcGttGTZ3X1WVeAiT8vSAh