



Prueba técnica Data Scientist 0



Prueba - Construcción de un modelo de *machine learning* en un Jupyter Notebook

● Descripción de la prueba

El hundimiento del Titanic es uno de los naufragios más infames de la historia.

El 15 de abril de 1912, durante su viaje inaugural, el RMS Titanic, ampliamente considerado "insumergible", se hundió después de chocar con un iceberg. Desafortunadamente, no había suficientes botes salvavidas para todos a bordo, lo que resultó en la muerte de 1502 de los 2224 pasajeros y la tripulación.

Si bien hubo algún elemento de suerte involucrado en sobrevivir, parece que algunos grupos de personas tenían más probabilidades de sobrevivir que otros.

En este challenge, pedimos que se cree un modelo predictivo que responda a la pregunta: "¿Qué tipo de personas tenían más probabilidades de sobrevivir?" utilizando datos de pasajeros (es decir, nombre, edad, sexo, clase socioeconómica, etc.).

Para esta prueba se usarán los datos adjuntos formato csv y la realización se hará en un Jupyter Notebook.

El lenguaje de programación para la implementación de la prueba debe ser Python3  . Utiliza todos aquellas librerías que te necesites. Además de celdas de código añade celdas de markdown en las que expliques de forma breve los distintos pasos y expongases cualquier *insight* que consideres interesante.



● Objetivos

Desarrollar las siguientes fases en el proceso de *data science*: construcción del *dataset* , analítica descriptiva, limpieza de datos, *feature engineering*, preprocessado, selección del modelo, entrenamiento, validación y métricas de *performance*.

- En la medida de lo posible, emplear conocimientos que se tengan del sector del crédito online, scoring de riesgo, detección de fraude...
- Escribir código elegante y eficiente
- Comunicar los resultados
- Pasar un buen rato y aprender ;)

● Entrega

Para la entrega descárgate los siguientes ficheros:

1. El notebook de trabajo de la prueba en .ipynb
2. El notebook de trabajo de la prueba en .html

Mándanos los archivos por separado o en un fichero zip.

● Documentación de los campos del dataset

Un conjunto de datos se titula `train.csv` y el otro se titula `test.csv` .

- Train.csv contendrá los detalles de un subconjunto de pasajeros a bordo (891 para ser exactos) y, lo que es más importante, revelará si sobrevivieron o no, también conocido como la "verdad terrestre".
- El conjunto de datos `test.csv` contiene información similar pero no revela la "verdad en tierra" para cada pasajero. Es su trabajo predecir estos resultados.
- Utilizando los patrones que encuentre en los datos de train.csv, prediga si los otros 418 pasajeros a bordo (que se encuentran en test.csv) sobrevivieron.