Club de Datos - Exactas UBA

Consigna para dataset Titanic

August 25, 2021

Acerca del dataset

Aunque puede haber un factor de 'suerte' a la hora de sobrevivir a la tragedia, parece que algunos grupos tuvieron mas chances que otros. En este desafio te invitamos a explorar los datos, y si te animas, te pedimos que construyas un modelo predictivo que nos diga "que tipos de personas tuvieron mas chances de sobrevivir?", basandote en datos de los pasajeros (nombre, edad, genero, clase social, etc).

Objetivo Principiantes

El objetivo es hacer una exploracion de los datos que logre informar, resumir o describir hechos importantes perdidos entre la informacion del dataset.

Primero, vas a necesitar importar (cargar, abrir) los datos. El archivo es titanic_training.csv.

Para entender el dataset, puede ser util averigual cuantas filas tiene, cuantas columnas, que valores toman esas columnas. Puede ser util usar el metodo .describe() de pandas, aprendemos algo util a partir de esta informacion?

Hay valores faltantes? es decir, que esten 'missing', o NaN?

Lo ideal es dar con algunas tablas y/o graficos que nos puedan contar los puntos principales de esta historia, segun la informacion en el dataset. Algunas ideas: que pasa con la supervivencia si agrupamos por genero? y por clase? Por edades?

Objetivo Avanzades

El objetivo es simple: dar con un modelo (machine learning) que prediga que pasajeros sobrevivieron al naufragio del Titanic.

La data se divide en dos grupos:

- training set
- test set

El training set se usa para construir tus modelos de machine learning. Para el training set, tenemos el outcome, es decir la confirmacion de si el pasajero sobrevivio o no. Podes crear

nuevas columnas (features) a partir de transformaciones o combinaciones de las columnas existentes.

El test set se usa para evaluar que tan bien funciona el modelo en datos que no se hayan visto previamente. El test set no tiene la informacion sobre supervivencia.

La idea es, luego de limpiar un poco los datos, probar distintos modelos e ir evaluandolos por su capacidad de predecir correctamente, segun la informacion en training set. Tener en cuenta que es un problema de *clasificacion*.

Pueden evaluar los modelos (en subsets del training set) usando medidas como precision, recall, F1, matrices de confusion, etc. Si estan muy conformes con un modelo pueden probar submitir la respuesta al desafio de kaggle (link).