Alex Pardo Ramos

**Tipología y ciclo de vida de los datos**

**Práctica 2**

**1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?**

El dataset utilizado es el adjunto a la práctica “Titanic: Machine Learning from Disaster” (https://www.kaggle.com/c/titanic), este es importante puesto que muestra los pasajeros del titanic y algunos de sus atributos como el sexo, la edad o si sobrevivieron. Se pretende responder a las preguntas de qué atributos tuvieron más supervivencia en el accidente. Estas preguntas son: ¿sobrevivieron más hombres o mujeres?, ¿qué rango de edad tubo más supervivencia?, ¿los pasajeros de que clase sobrevivieron más?, ¿dónde embarcaron influyo en la supervivencia?

**2. Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.**

En nuestro caso concreto utilizaremos una subsección de los datos originales, teniendo en cuenta las variables PassagerId, Pclass, Sex, Age, Embarked y Survived de train.csv. Que serán suficiente para responder a las preguntas formuladas anteriormente. No podremos utilizar test.csv puesto que no dispone de la variable Survived que es la variable objetivo.

**3. Limpieza de los datos.**

**3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.**

Embarked y Age poseen elementos vacíos. Eliminamos todas las filas con valores nulos o vacíos. Y comprobamos que no haya ceros en las columnas de tipo chr. Y así es.

**3.2. Identifica y gestiona los valores extremos.**

Identificamos los valores de Age una variable que puede tener outliers y los gestionamos de forma que no haya. Adjuntamos graficos demostrando la eliminación de estos.

**4. Análisis de los datos.**

**4.1. Selección de los grupos de datos que se quieren analizar/comparar (p. e., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)**

Se van a comparar todas las variables mencionadas anteriormente con la variable Survived para saber si alguna de estas tiene una relación más estrecha con la supervivencia en el accidente.

**4.2. Comprobación de la normalidad y homogeneidad de la varianza.**

Hemos realizado pruebas gráficas sobre la variable Age que muestran que no se puede rechazar la hipótesis de normalidad puesto que es la única variable numérica que no es categórica. Y comprobado la homogeneidad de la varianza respecto Survived y Sex y como p-valor es 0.2 muestra que no existe una diferencia significativa entre las dos variables.

**4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**