PRACTICA 2 TIPOLOGIA DE DATOS.

1. Descripción del dataset. Porqué es importante y que pregunta/problema pretende responder.

El dataset elegido para la practica 2 de tipología y ciclo de vida de datos es el dataset de titanic, perteneciente a Kaggle (<https://www.kaggle.com/c/titanic/data>). Se ha elegido ete dataset, por su contenido; éste tiene suficientes variables para proceder a diferentes estudios de un suceso histórico.

El conjunto de datos contiene 891 registros y 12 variables;

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Tipo | Definición | Descripción |
| PassengerId | Integer | Código de identificación del pasajero. |  |
| Survived | Integer | Supervivencia | 0 = No, 1 = Si |
| Pclass | Integer | Clase de ticket | 1 = Primera, 2 = Segunda, 3 = Tercera |
| Name | Factor | Nombre del pasajero |  |
| Sex | Factor | Género del pasajero |  |
| Age | Double | Edad del pasajero |  |
| Sibsp | Integer | Núm. Hermanos a bordo del barco |  |
| Parch | Integer | Núm. Familiares en el barco |  |
| Ticket | Factor | Número de ticket |  |
| Fare | Double | Tarifa del pasajero |  |
| Cabin | Factor | Número de cabina |  |
| Embarked | Factor | Puerto embarcado | C = Cherbourg, Q = Queenstown, S = Southampton |

Se estudiará diferentes variables fijando siempre como prioridad la variable de supervivencia, nos interesa encontrar que variables son las que tienen más peso a la hora de “predecir” si la persona dependiendo de X variables sobrevivía o moría.

En el cuadro anterior se puede ver perfectamente todas las variables del dataset y su significado, puesto que las clases survived, Pclasss son numéricas pero las podríamos cambiar a factoriales por su significado.

La variable PassengeId no nos aporta nada al estudio, por lo que procederemos a eliminarla.

1. Integración y selección de los datos de interés a analizar.

El proceso de integración y selección de datos se hará al finalizar el proceso de limpieza y análisis de las distintas variables.

Los que creemos que elegiremos serán, la supervivencia, ya que de esta trata el proyecto, en función de otras variables, como la edad, el género de la persona, el puerto de embarque y la clase a la que pertenece la persona. A medida que vayamos avanzando en el estudio se analizará si los otras variables tienen peso en la variable supervivencia.