Andrea Reyes

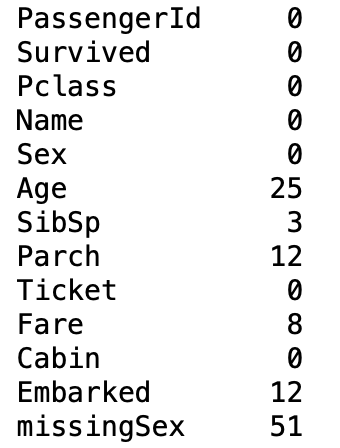
20190265

Laboratorio 8

**Parte 1**

1. Reporte detallado de missing data para todas las columnas.

El dataset tiene 12 variables/columnas de las cuales 6 cuentan con missing data:

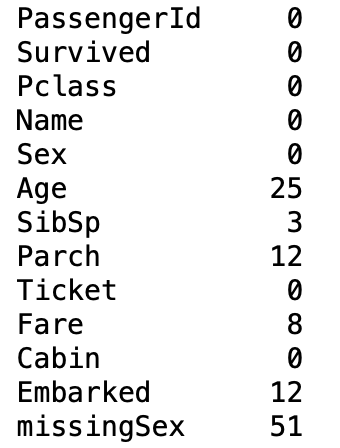


1. Para cada columna especificar qué tipo de modelo se utilizará y qué valores se le darán a todos los missing values.

* Age: imputación de promedio porque así tenemos edades dentro del margen para no sesgar a los datos existentes.
* SibSp: imputación de la moda porque solo hay dos valores posibles entonces no afecta de manera significante usar el que más se repite.
* Parch: imputación de la moda porque solo hay dos valores posibles entonces no afecta de manera significante usar el que más se repite.
* Fare: imputación de promedio porque así tenemos valores dentro del margen y no son muchos los faltantes.
* Embarked: imputación de la moda porque solo hay dos valores posibles entonces no afecta de manera significante usar el que más se repite.

1. Reporte de qué filas están completas

El dataset tiene 12 variables/columnas de las cuales 6 están completas (0’s en cuanto a missing values):



titanicMD["missingSex"] = titanicMD["Sex"].apply(lambda m: None if m == "?" else m)

nll = titanicMD.isna().sum()

nll

titanicMD["missingSex"] = titanicMD["Sex"].apply(lambda m: None if m == "?" else m)

nll = titanicMD.isna().sum()

nll

titanicMD["missingSex"] = titanicMD["Sex"].apply(lambda m: None if m == "?" else m)

nll = titanicMD.isna().sum()

nll

titanicMD["missingSex"] = titanicMD["Sex"].apply(lambda m: None if m == "?" else m)

nll = titanicMD.isna().sum()

nll

titanicMD["missingSex"] = titanicMD["Sex"].apply(lambda m: None if m == "?" else m)

nll = titanicMD.isna().sum()

nll

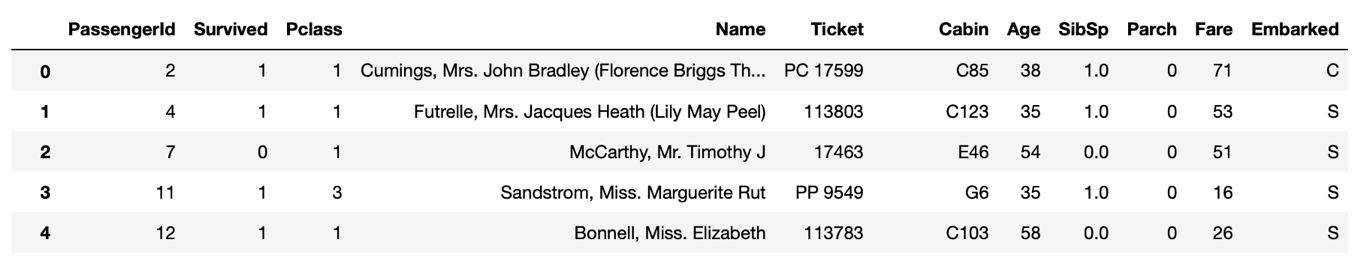
titanicMD["missingSex"] = titanicMD["Sex"].apply(lambda m: None if m == "?" else m)

nll = titanicMD.isna().sum()

nll

1. Utilizar los siguientes métodos para cada columna que contiene missing values

Dataframe sin missing values:



1. Comparar los métodos del inciso 4 contra “titanic.csv”

Age y Fare son los que menos se acercan a la realidad.

1. Conclusiones

La cantidad de sobrevivientes fue de 123

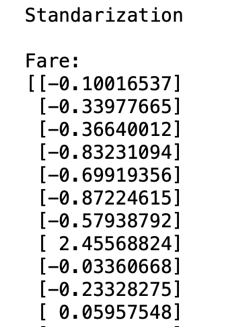
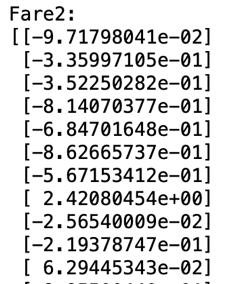
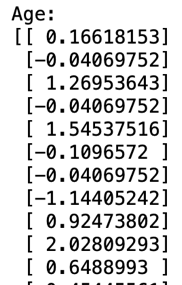
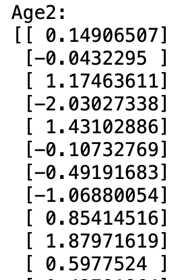
El dataframe tiene variables de distintos tipos y hay modelos que no son muy efectivos dependiendo de a que tipo de variable se aplique.

Es importante tomar en cuenta la cantidad de missing values en cada columna y si hay correlación.

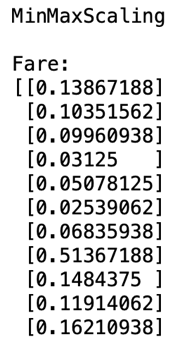
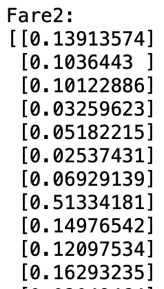
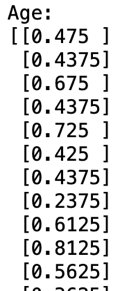
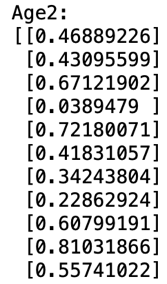
**Parte 2**

1. Luego del pre-procesamiento de la data con Missing Values, normalice las columnas numéricas por los métodos:
   1. Standarization
   2. MinMaxScaling
   3. MaxAbsScaler

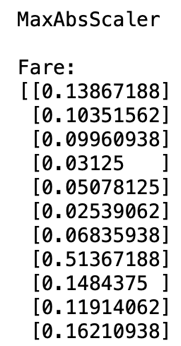
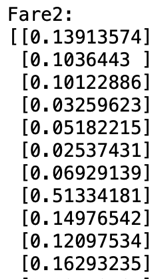
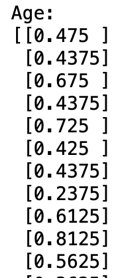
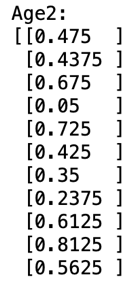
a.

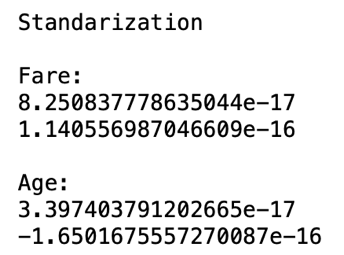
b.

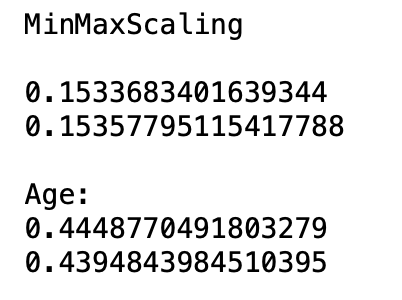
   

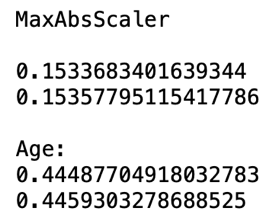
c.

1. Compare los estadísticos que considere más importantes para su conclusión y compare contra la data completa de “titanic.csv” (deberán de normalizar también).







Hay métodos que funcionan mejor dependiendo del tipo de variable. Como se observa en las imágenes de arriba, los mejores modelos son MinMaxScaling y MaxAbsScaler.

Código: <https://github.com/AndreaNathalia/data-wrangling/blob/main/Laboratorio8/Lab8.ipynb>