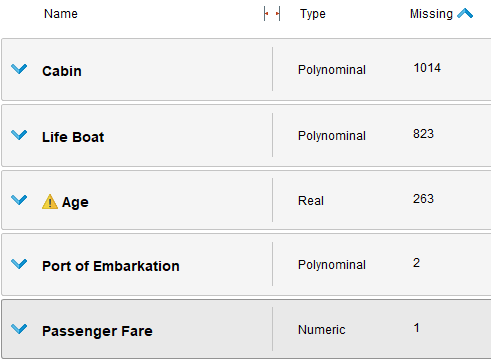
|  |
| --- |
| **Introducción a los Métodos de Aprendizaje Automático** |
| **Ejercicios domiciliarios**  **UT2 – PD1** |
| **Autor:** Gerardo Fernández - CI: 2858230-7  **Docente:** Ernesto Ocampo |
| **04/09/2021** |
|  |

**Ejercicio 1**

**1. Handling Missing Values**

****

Missing values en Titanic data

Con el operador **Select Attribute** se eliminaron los campos con mayor número de missing values, Cabin y Life Boat.

Con **Replace Missing Values** se completaron las edades (campo Age) faltantes con el promedio.

Con el operador **Filter Examples** se filtran también los ejemplos que tienen missing values

|  |
| --- |
| **2. Normalization and Outlier detection** |
|  |
| Another important step of data cleansing is to identify unusual cases and remove them from the data set. In some situations, the outliers themselves might be the most interesting cases (detecting fraudulent credit card transactions, for example), but in most cases outliers are simply the result of an incorrect measurement and should be removed from the data set. This is exactly what we will do in this tutorial. |

Se eliminan los atributos Cabin, Life Boat, Name, and Ticket Number .

The result will be a data set only containing those columns we believe will contribute well to our outlier detection. We will use a distance-based outlier detection algorithm which calculates the Euclidean distance between the data points and marks those points which are farthest away from other data points as outliers. The Euclidean distance uses the distances between two data points for each individual attribute. Think about it: what is the effect on the distance if the attributes have different value ranges (one attribute between 0 and 5 and another attribute between 1 and 1000)? Attributes with larger values will contribute much more than those with smaller values. For this reason we should ensure that all attributes are using similar value ranges. This transformation is called Normalization.

In general, you should always normalize your data before you apply distance-based algorithms like outlier detection or k-Means clustering. Using the default parameters, the Normalize operator will perform a z-Transformation (also known as **Standardization**) **which results in a mean value of 0 and a standard deviation of 1 for each attribute**. In other words, all of the attributes are on the same scale after normalization and can be compared with one another

**Normalize :: Z-transformation** es estandarización

**Normalize :: Range transformation** es normalización

**Ejercicio 2**

