Data Mining Report

by
Mikkel Stolborg
Hlynur Örn Haraldsson

Contents

1 Introduction

1.1 Data selection

1.2 Research question

Our primary question which we want answered was:

"Which attributes contributes mostly to the survival rate of a passenger on the Titanic
Here we wished to figure out what set of parameters would ensure the highest rate of survival on the Titanic. We would use classification through a classification tree to figure out which set gives the highest percentage of survival.

The secondary question arose when looking at clusters of the data.

"Which societal data can be found in clusters of the titanic data"
Looking at the data, we decide to try clustering to see if there was an emergent pattern. It would be interesting to see if the were a relation between wealth and number of children and the like.

1.3 Tools for data mining

We choose to use the free tool called Orange[1]. This tool allowed to quickly manipulate the data, such that we could extract the interesting elements. In the tool you manipulate the data by clicking and drawing connections between data elements and processing methods. Before explaining further, we have included the map of the process used for our data set in the orange framework, see figure 1.
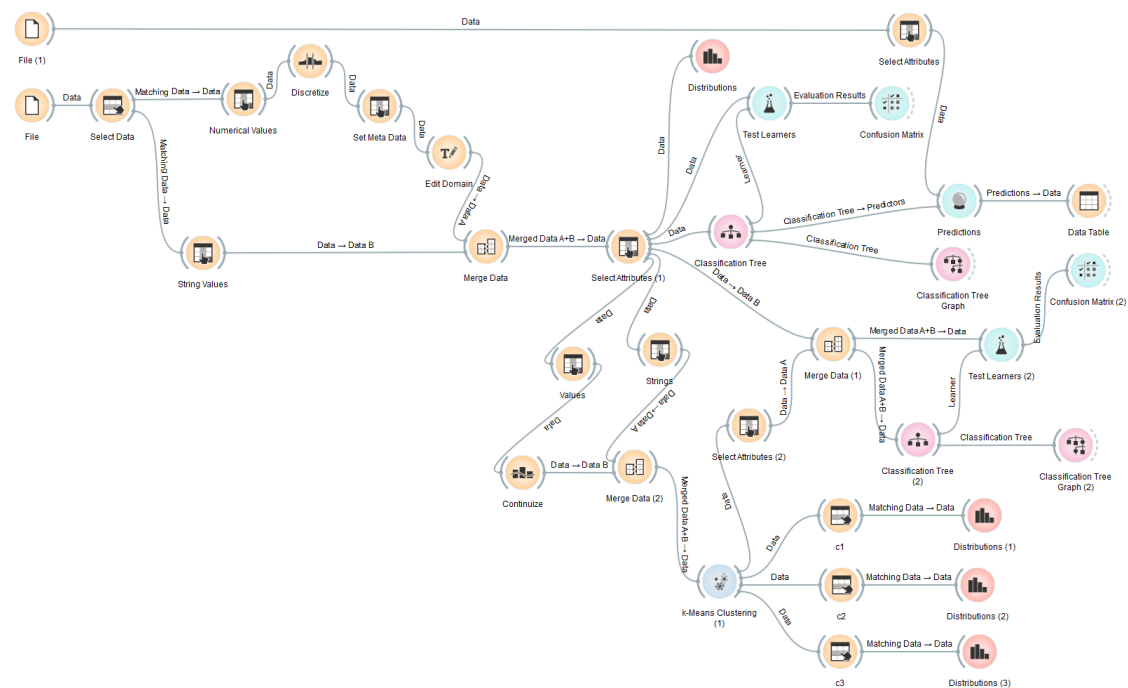

Figure 1: The map of the methods and processes used on the data.

2 Data mining

2.1 Preprocessing

2.2 Classification tree

2.2.1 Cross Validation

2.3 K-means Clustering

2.3.1 Data validation

3 Conclusion

3.1 Societal impact

References

[1] *Orange Data Mining*, http://orange.biolab.si/, 13-05-15

[2] *EVE online*, https://www.eveonline.com/, 12-05-15

[3] *XCOM Enemy Unkown*, http://www.xcom.com/enemyunknown/entry, 12-05-15