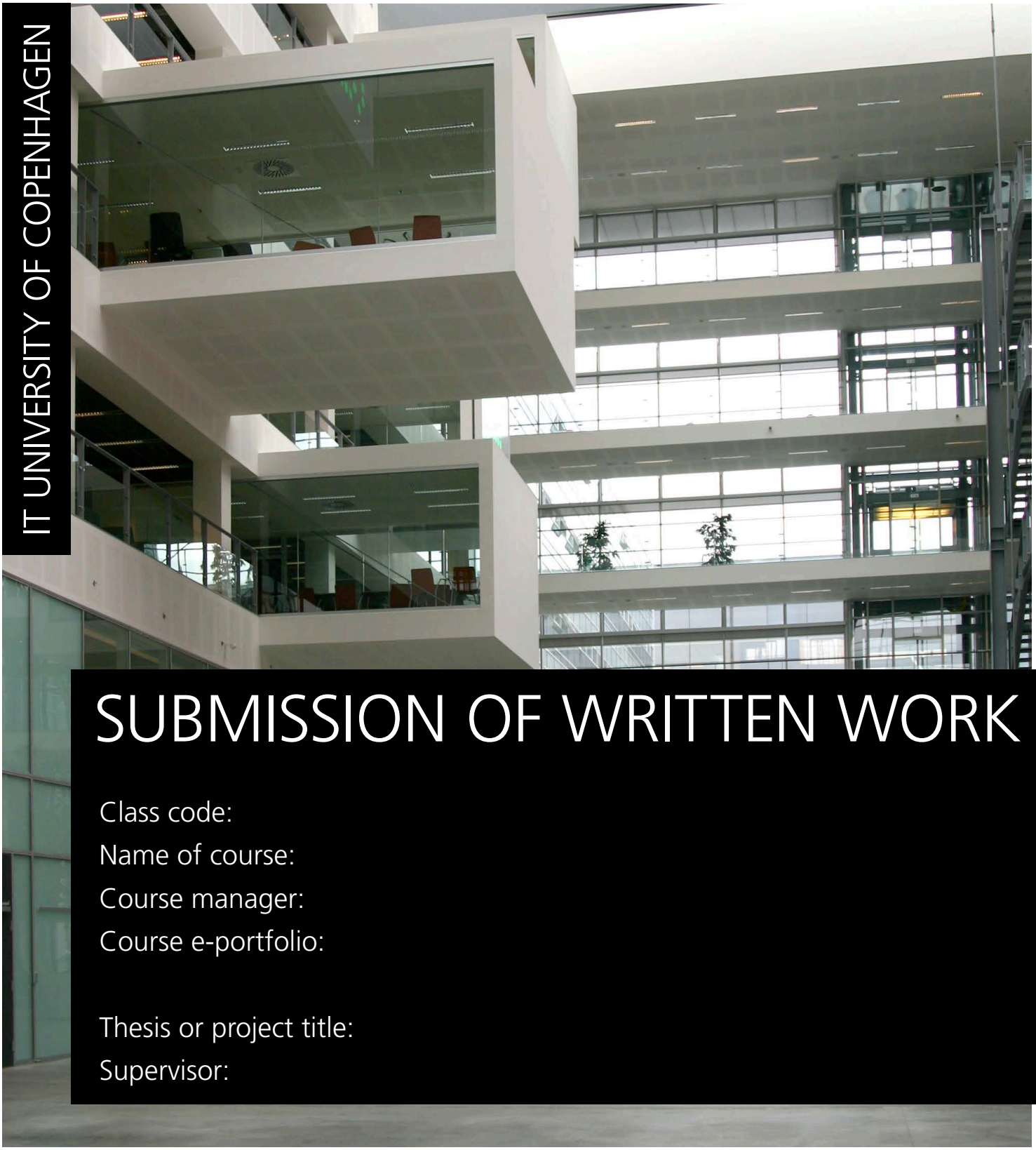
IT UNIVERSITY OF COPENHAGEN



SUBMISSION OF WRITTEN WORK

Class code:

Name of course:

Course manager:

Course e-portfolio:

Thesis or project title:

Supervisor:

Full Name:	Birthdate (dd/mm-yyyy):	E-mail:
1. _____	_____	_____@itu.dk
2. _____	_____	_____@itu.dk
3. _____	_____	_____@itu.dk
4. _____	_____	_____@itu.dk
5. _____	_____	_____@itu.dk
6. _____	_____	_____@itu.dk
7. _____	_____	_____@itu.dk

Data Mining Report

by
Group 21:
Mikkel Stolborg
Hlynur Örn Haraldsson

IT University of Copenhagen
MDMI, S2015
Anders Hartvig Hartzen
Hajira Jabeen
Héctor Pérez Martínez
Sebastian Risi
Noor Shaker
May 15, 2015

Contents

1 Data selection

We worked on a data set regarding the passengers on the Titanic. The data structure is presented in Table 1 with a short name and the value type associated with the variable.

Variable name	Short name	Value Type
survival	Survived	binary
pclass	Passenger Class	numeric
name	Name	string
sex	Sex	binary string
age	Age	numeric
sibsp	Number of Siblings/Spouses Aboard	numeric
parch	Number of Parents/Children Aboard	numeric
ticket	Ticket Number	string
fare	Passenger Fare	numeric
cabin	Cabin	string
embarked	Port of Embarkation	attribute table

Table 1: Titanic data set variables with short description and classification.

The type binary and binary string, means there are only two options, the first of the types is based on numeric binary data, whilst the second is based on string data. The value type attribute table covers three options. The type designation in the table are based on how they are treated in the data mining code. This means that whilst some of the other variables might be a three option choice, it will not be treated as such. The reason for this division is mostly due to the nature of string variables and numeric variables.

1.1 Data description

The variables in the table are specified as follows.

Survival is simply a 1 for survived and 0 for not.

Passenger Class is a numeric value which can take on either 1, 2, or 3. The lower the value, the higher the class. The class is a proxy for the passengers socio-economic status.

Name is simply the name of the passenger, starting with surname.

Sex is the gender of the passenger.

Age is the age in years of the passenger. If the age is less than one it is fractional, and if it is estimated the age is in the form xx.5, meaning it has a decimal value as well.

Number of Siblings/Spouses Aboard, is the number of relatives, yet some of the relations are ignored. The meaning of sibling and spouse is summarized below.

Number of Parents/Children Aboard, is similar to the Number of Siblings/Spouses Aboard. The meaning is as well summarized below.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic

Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)

Parent: Mother or Father of Passenger Aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbours in a village, however, the definitions do not support such relations.

Ticket Number is the number printed on the ticket. It is basically an identification string with characters and numbers. The tickets have no real pattern.

Passenger Fare is how much the passenger paid for its ticket. The price probably is listed in dollars.

Cabin is the cabin number. Again this is an identification of the rooms, yet there is not really any good patterns within the identification string.

Port of Embarkation is the place of embarking for the passenger. The port is abbreviated as follows. C is for Cherbourg, Q is for Queenstown, S is for Southampton.

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 chose to use the free tool called Orange[2]. This tool allowed us to quickly manipulate the data, so that we could extract the interesting elements. In the tool you can manipulate the data by clicking and drawing connections between data elements and processing methods. All manipulations of our data is done in this tool, and the diagram we have drawn has 3 areas of focus. There is an area focused on the preprocessing of the data, deleting and updating data as described later in section 2.1. This area sends the processed data to a classification tree map and onto the K-means Clustering. Each of these constitute the remaining areas. In each of the areas the data is processed using the appropriate algorithm, and visualized.

We have included the map of the process used for our data set in the orange framework in appendix A, figure 6.

2 Data mining

This section goes through the data mining process, detailing the methods used and analysing and explaining the output.

2.1 Preprocessing

The data input we got from Kaggle[3], was quite neat and easy to work with. This lead to very little preprocessing and general expectation of functioning data. However some of the data, even though extensive, had no sensible patterns from which information could be unlocked, and some data was discarded as they were not directly related to our research question. We chose to remove the Ticket Number and Cabin data from our data set, as there were no easily identifiable pattern and there were many missing values. The Name of the passengers were discarded as well, as it was not of any use in the methods we chose to apply to the data.

The main reduction in entries in our data, rather than attributes, were to remove any entry with missing attributes. 182 entries were removed in total, mainly due to lack of age, whilst a few, 2, were removed because of lacking Embarked. A small reduction in data was made to remove outliers in respect to fare, simply because their presence simply deteriorate the remaining data. This can be supported by looking at the Fare distribution before any preprocessing had been done, see figure 1.

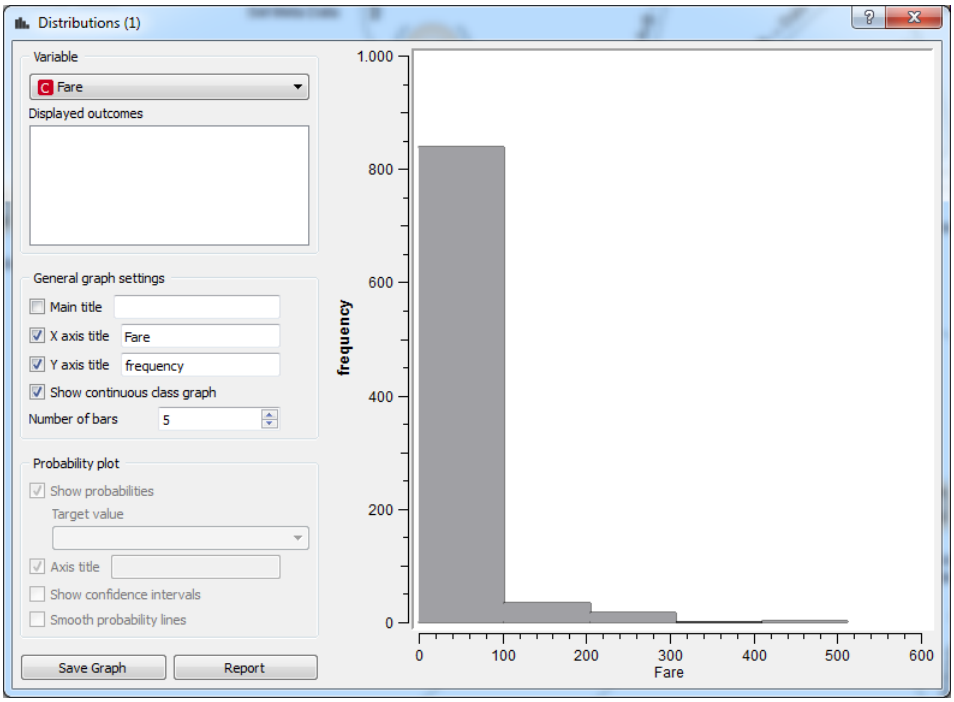


Figure 1: Fare distribution on initial dataset. The black area between 300 and 400, indicates no persons.

From the figure it is clear that the outliers of the 3 people who paid more than 400. Removing the outliers gives a more detailed view of the inter relations of the other values.

The total process in the orange framework can be seen in figure 2. The discretize function is there to tell the orange frame work that the survival data is indeed discrete

data. The meta data in this case is the passenger id, which is used for merging throughout the framework. Finally the discrete values of the survival variable is being renamed to "yes" and "no". The data is merged and sent out for further use.

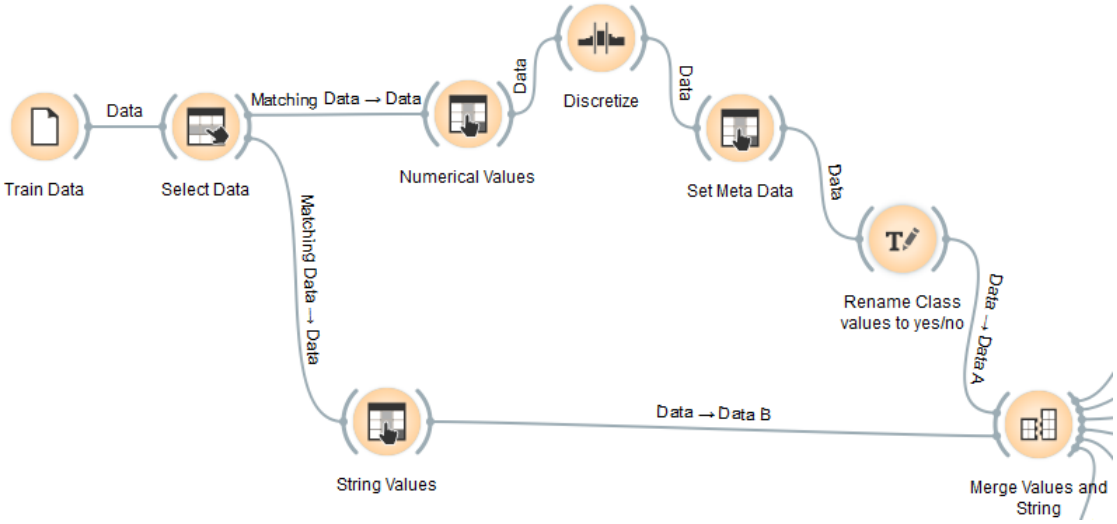


Figure 2: Orange map of preprocessing. The data is prepared for the algorithms and unnecessary data is removed.

2.2 Classification tree

The classification tree[4] is built upon the information gain of the attributes give to it. The information gain lets us known on which attribute the tree should split upon. The tree is created upon the information gain, it will be built from top to bottom and the tree will keep being built until all the the samples of a given tree node belong to the same class, no remaining attributes to split upon or no samples are left. In appendix A, figure 7, we see the structure of the orange map where we create the classification tree.

2.2.1 Implementation and analysis

Our classification tree is built using the standard gain value and splits in regard to whether a passenger survived. This means the information gain is based on whether we learn more about who died after the split. The resulting tree, taken to five nodes depth, is visible in figure 3.

From the figure there quickly arises an interesting distinction, namely that the females have a much greater chance of survival, compared to the males. Even at the first split into genders, the females have an astonishing survival rate of 75.2 percent. From there after a simple split in class, you see the females who are both middle or upper class, have a 94.2 percent chance of survival. From there it splits on the fare they paid, where the ones who paid the most are more likely to not survive. This can be explained in the fact that many of the upper class high paying females already have been sorted out. With this in mind, the fact that there is a difference in fare, though signification in survival rate, tells us little about why it is important. Next split is, for both sub-splits, where the younger age prevails in both cases. From there the splits yield very little information and have been excluded from this report.

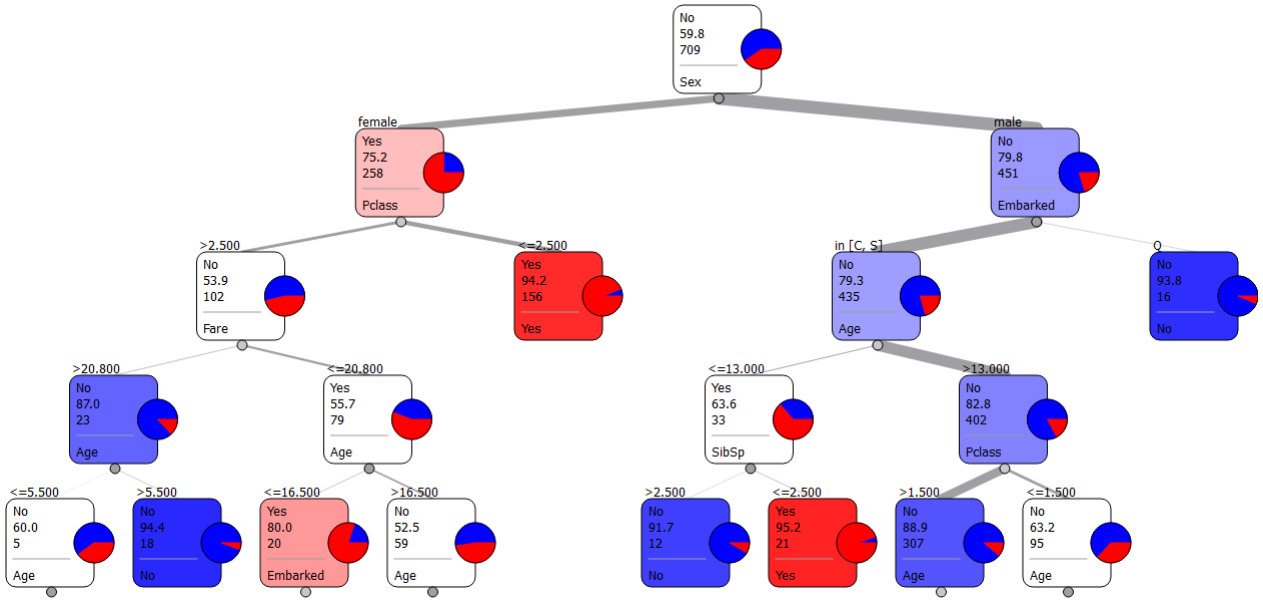


Figure 3: Classification tree on Titanic data set.

Looking at the male side of the tree, just by being male, you have 79.8 percent risk of dying. The first split in city of embarkation, if you were one of the 16 people from Queenstown, your chances of survival is only 6.2 percent. The passengers from the other cities fare a little better, but in order to have a decent chance of survival, you will need to be of age less than or equal to 13, and have less than 2.5 siblings or spouses.

All in all this tells us if you wanted to survive on the Titanic, your best choice would be to be an upper or middle class female. Failing that you should have been one of the lower class females which were younger than 16.5 years. This means, that for surviving the Titanic as a female, your greatest chance occurs when you are young, when not in the upper class.

For males the greatest chance of survival occurs when you are less than or equal to 13 years of age, whilst having fever than 2.5 siblings, and have not embarked at Queenstown.

2.3 K-means Clustering

K-Means clustering[1, p.451] works by first defining K number of clusters, then you feed normalized data to the clustering algorithm, the algorithm chooses K-initial centroids and starts calculating the distance to each data point and then clusters the data around those centroids. After each loop the centroid is recalculated in accordance to all data points in cluster and then the algorithm recalculated the distance to every point. This loop continues until the algorithm runs its loop once without changing anything or is within the minimal change value.

2.3.1 Implementation and analysis

We decided to use K-Means Clustering in order to increase the accuracy of our Classification Tree by adding the clusters to the attribute of the original data.

However after reclassifying with the clusters as an extra attribute we saw that it did

not increase accuracy as expected. After looking into the clustered data we saw that the clustering and the classification focused on the same attributes so that did not help us in increasing accuracy. The final centroids of our clusters are visible in Table 2.

Sex	Embarked	Class	Age	Sibling/Spouse	Parent/Children	Fare
Male	S	2.428	30.28	0	0	21.5697
Female	S	2.258	26.32	1	1	35.2690
Male	C	1.552	34.14	0	0	66.8577

Table 2: Cluster Centroids after clustering.

However while looking over the clusters we could see interesting facts about the clusters and were able to find connections regarding clusters.

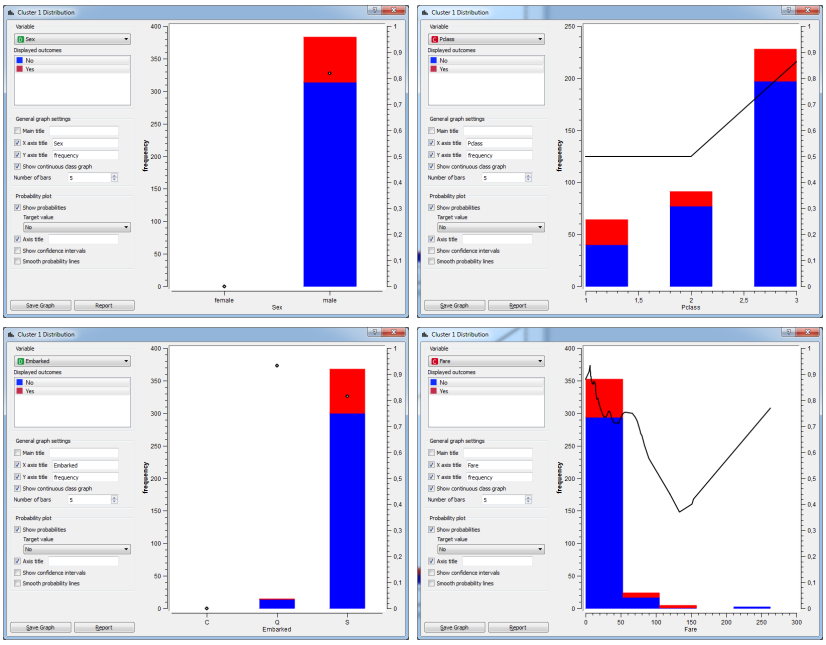


Figure 4: Cluster 1 data. From top left to bottom right the distributions are as follows: Sex, Pclass, Embarked, Fare.

For example we saw that in cluster 1, see figure 4, every person there was male, mostly lower class, most of them paid under 20 dollars for their Fare and almost all of them embarked from Southampton(S).

However if we look at Cluster 3, see figure 5, that cluster is mostly male, mostly upper class and most of them embarked from Cherbourg(C).

Cluster 2, see appendix B, figure 8, is only females, but with no other interesting correlation.

What we can extract from this data is, that if you were upper class you would most likely be living in Cherbourg. We can also see that people living in Cherbourg and Queenstown(Q) generally had more money than people living in Southampton. If we look at the fare graph difference between Cluster 1 and 3 we can see a big difference, people living in Southampton would generally pay less than 20 dollars for the ticket and almost no one paid more than that. How ever in Cluster 3 we can see that people still mostly paid under 20 dollars however we can also see that there are a lot more people willing to pay more than the 20 dollars.

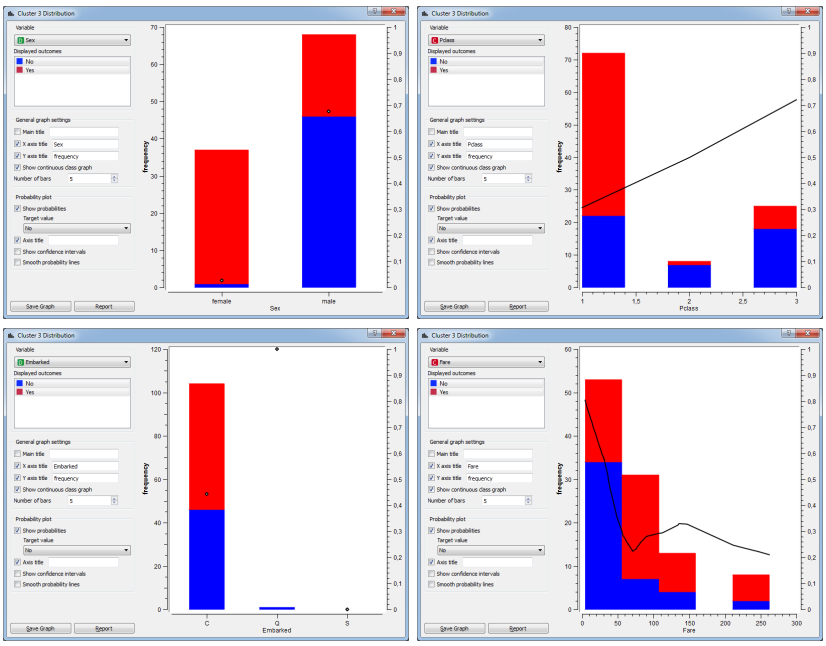


Figure 5: Cluster 3 data. From top left to bottom right the distributions are as follows: Sex, Pclass, Embarked, Fare.

2.4 Data validation

K-fold Cross validation [1, p.370] works by randomly partitioning the data into K exclusive subsets. The subsets are then run K times, in each loop a different subset will be used as the test data while the other k-1 subsets will be used as a training set.

2.4.1 Results

We decided to use 10-fold cross validation to validate our classification from the classification tree.

	No	Yes	
No	367	57	**424**
Yes	75	210	**285**
	442	**267**	**709**

	No	Yes	
No	369	55	**424**
Yes	81	204	**285**
	450	**259**	**709**

Table 3: Left: Non clustered Confusion Matrix. Right: Clustered Confusion Matrix. Grey colored cells are misclassified

By looking at Table 3, we can see that the clustering did not increase the accuracy of the classification. But the accuracy is pretty good none the less, we managed to get around 80% accuracy when predicting if people from test data survived or not.

3 Conclusion

To finish off this report, even though we failed our initial goal of using the clustering to increase the accuracy of our classification we still learned a lot from the data. We learned that majority of survivors were female of every class but for males the largest group of survivors were boys under the age of 13.

But what we also learned is a bit of class geography, we saw that lower class people mostly came from Southampton(S) whilst the upper class people mostly came from Cherbourg. What we could also assume from the data is that Queenstown might be more poverty stricken compared to the other ports due to amount of passengers coming from that town.

3.1 Societal impact

There isn't a big societal impact from this data mining but what could be done with this is feed information about passenger whose survived status is unknown and accurately predict their survival.

In theory, time travellers could use this information to go on the Titanic with the best amount of odds of surviving.

A Orange Process map

In this appendix are the maps of the processes being used on the data.

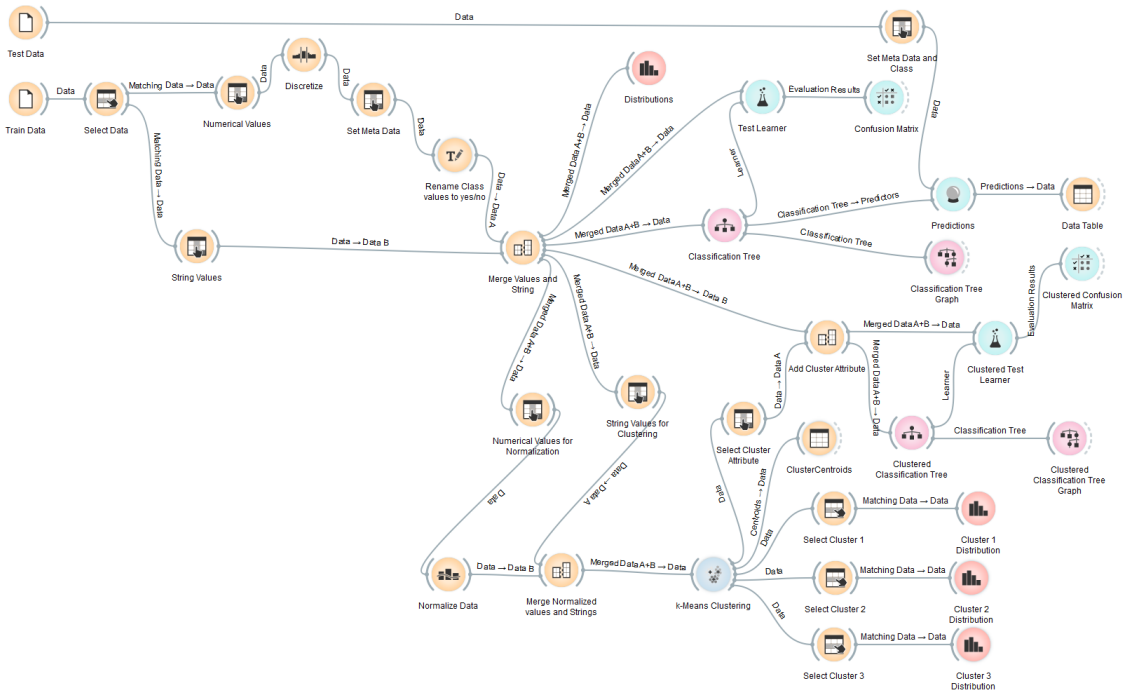


Figure 6: The main map of the methods and processes used on the data.

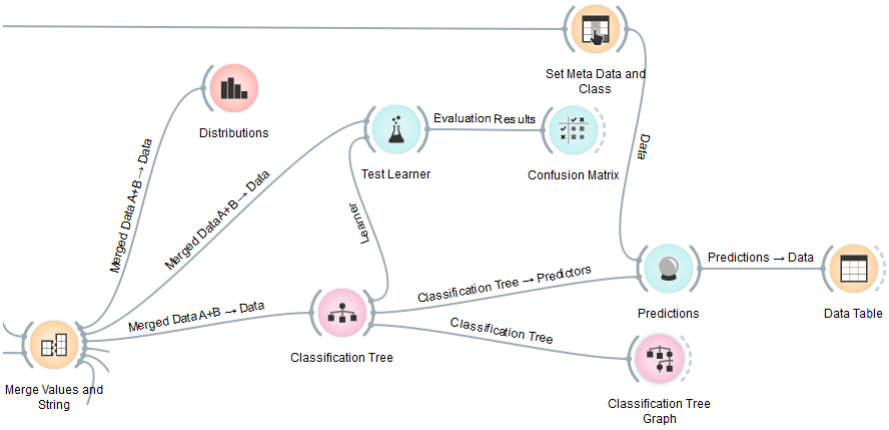


Figure 7: Orange map of the classification tree process.

B K-Mean Clustering

The distributions from the clustered data.

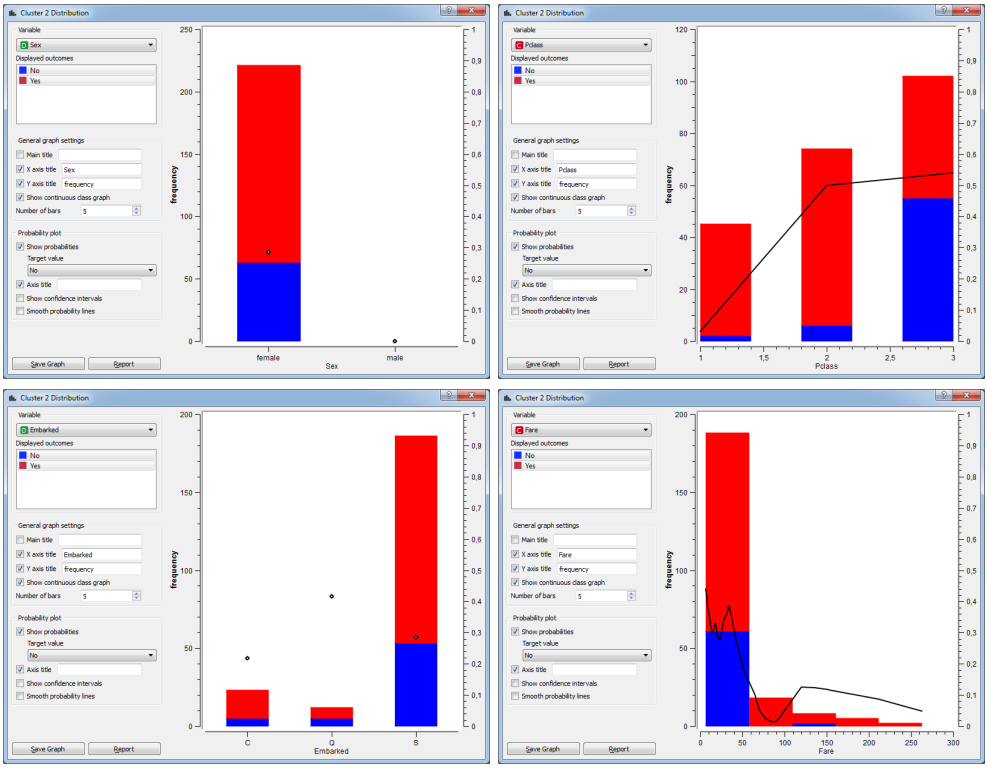


Figure 8: Cluster 2 data. From top left to bottom right the distributions are as follows: Sex, Pclass, Embarked, Fare.

References

[1] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, USA Third edition, 2012.

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

[3] *Kaggle Titanic data*, https://www.kaggle.com/c/titanic, 14-05-15

[4] *Classification Tree*, http://en.wikipedia.org/wiki/Decision_tree_learning, 14-05-15