

*CST8390 BI & Data Analytics*

**Assignment II: Decision Tress with Titanic Dataset**

**Elaiza Rivera**

**(040839516)**

**&**

**Abdullah Zeki Ilgun**

**(040991363)**

## Abstract

This famous data set focuses on the passengers of the largest passenger line ever made which sunk over 100 years ago, Titanic. Even though the incident happened a lot of years ago, the data set is still quite useful to develop machine learning algorithms. Most people could not survive from the incident since there were not enough boats.

The likelihood of a passenger's survival depended on certain elements such as, passenger's class, gender, number of relatives, and so on. There were 2224 passengers on Titanic, but in this data set, 887 passengers' information were used. The information includes passenger's name, number of siblings/spouses, number of parents/children, and port of embarkation alongside with the other elements. The aim of the data set is to develop an algorithm that can be used to predict whether the passenger survived or not by the given training and test data sets.

## Attributes & Description

| Attribute   | Description   |
|-------------|---|
| PassengerId | Primary key given to the passenger to differentiate all the passengers                    |
| Pclass      | Class of the passenger (1 = 1 <sup>st</sup> , 2 = 2 <sup>nd</sup> , 3 = 3 <sup>rd</sup> ) |
| Name        | Name of the passenger.  |
| Sex         | Gender of the passenger (Male, Female)  |
| Age         | Age of the passenger.   |
| SibSp       | Number of siblings/spouses does the passenger have  |
| Parch       | Number of parents/children does the passenger have  |
| Ticket      | Ticket number of the passenger  |
| Fare        | Amount of fare that the passenger paid  |
| Cabin       | Cabin of the passenger (If having one)  |
| Embarked    | Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)                      |
| Survival    | Whether the passenger survived (0 = No, 1 = Yes) Not included in the test file            |

## What attributes are further needed?

From the 11 attributes presented above we have picked the most relevant attributes. These are Pclass, Sex, Age, SibSp, Parch, embarked and survival. These are the factors in which we think mainly contribute to the survival rate of each passengers. To classify these factors, we created new attribute Age Group which was derived from Age and Relatives sums up SibSp and Parch. Then the Age, SibSp and Parch was remove from the datasets to not create a biased decision. The image below shows the csv file processed from the original datasets.

|    | A        | B      | C      | D          | E         | F        | G | H | I |
|----|----------|--------|--------|------------|-----------|----------|---|---|---|
| 1  | Survived | Pclass | Sex    | Age Group  | Relatives | Embarked |   |   |   |
| 2  | 0        | 3      | male   | Adult      | 1         | S        |   |   |   |
| 3  | 1        | 1      | female | Adult      | 1         | C        |   |   |   |
| 4  | 1        | 3      | female | Adult      | 0         | S        |   |   |   |
| 5  | 1        | 1      | female | Adult      | 1         | S        |   |   |   |
| 6  | 0        | 3      | male   | Adult      | 0         | S        |   |   |   |
| 7  | 0        | 3      | male   | NK         | 0         | Q        |   |   |   |
| 8  | 0        | 1      | male   | Adult      | 0         | S        |   |   |   |
| 9  | 0        | 3      | male   | Child      | 4         | S        |   |   |   |
| 10 | 1        | 3      | female | Adult      | 2         | S        |   |   |   |
| 11 | 1        | 2      | female | Adolescent | 1         | C        |   |   |   |
| 12 | 1        | 3      | female | Child      | 2         | S        |   |   |   |
| 13 | 1        | 1      | female | Adult      | 0         | S        |   |   |   |
| 14 | 0        | 3      | male   | Adolescent | 0         | S        |   |   |   |
| 15 | 0        | 3      | male   | Adult      | 6         | S        |   |   |   |
| 16 | 0        | 3      | female | Adolescent | 0         | S        |   |   |   |
| 17 | 1        | 2      | female | Adult      | 0         | S        |   |   |   |
| 18 | 0        | 3      | male   | Child      | 5         | Q        |   |   |   |
| 19 | 1        | 2      | male   | NK         | 0         | S        |   |   |   |
| 20 | 0        | 3      | female | Adult      | 1         | S        |   |   |   |
| 21 | 1        | 3      | female | NK         | 0         | C        |   |   |   |
| 22 | 0        | 2      | male   | Adult      | 0         | S        |   |   |   |
| 23 | 1        | 2      | male   | Adult      | 0         | S        |   |   |   |
| 24 | 1        | 3      | female | Adolescent | 0         | Q        |   |   |   |
| 25 | 1        | 1      | male   | Adult      | 0         | S        |   |   |   |
| 26 | 0        | 3      | female | Child      | 4         | S        |   |   |   |
| 27 | 1        | 3      | female | Adult      | 6         | S        |   |   |   |
| 28 | 0        | 3      | male   | NK         | 0         | C        |   |   |   |
| 29 | 0        | 1      | male   | Adolescent | 5         | S        |   |   |   |
| 30 | 1        | 3      | female | NK         | 0         | Q        |   |   |   |
| 31 | 0        | 3      | male   | NK         | 0         | C        |   |   |   |

Titanic\_train\_processed

## Loading the File To Weka

Upon loading to Weka, each attribute was checked if they are set into their expected type. Survived, Pclass and Relatives were in numeric form, so they were converted to be nominal. All attributes are all in nominal form.

## Class and Age Group Attributes

### a. Survived

The screenshot shows the Weka Explorer window with the 'Survived' attribute selected. The interface includes a menu bar (Preprocess, Classify, Cluster, Associate, Select attributes, Visualize), a toolbar (Open file..., Open URL..., Open DB..., Generate..., Undo, Edit..., Save...), and a Filter section (Choose, None, Apply, Stop). The 'Current relation' section shows 'Relation: Titanic\_train\_processed-weka.filters.unsuper...' and 'Instances: 889'. The 'Attributes' section lists attributes: 1. Survived, 2. Pclass, 3. Sex, 4. Age Group, 5. Relatives, 6. Embarked. The 'Selected attribute' section shows 'Name: i>Survived', 'Missing: 0 (0%)', 'Distinct: 2', and 'Type: Nominal'. A table displays the attribute's distribution:

| No. | Label | Count | Weight |
|-----|-------|-------|--------|
| 1   | 0     | 549   | 549.0  |
| 2   | 1     | 340   | 340.0  |

The 'Class' is set to 'i>Survived (Nom)'. A bar chart visualizes the distribution, with a blue bar for '0' (549) and a red bar for '1' (340). The 'Status' section shows 'OK' and a 'Log' button.

## b. Age Group

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** [Apply] [Stop]

**Current relation**  
Relation: Titanic\_train\_processed-weka.filters.unsuper...  
Instances: 889  
Attributes: 6  
Sum of weights: 889

**Attributes**  
[All] [None] [Invert] [Pattern]

| No. | Name  |
|-----|---|
| 1   | <input type="checkbox"/> Survived             |
| 2   | <input type="checkbox"/> Pclass               |
| 3   | <input type="checkbox"/> Sex                  |
| 4   | <input checked="" type="checkbox"/> Age Group |
| 5   | <input type="checkbox"/> Relatives            |
| 6   | <input type="checkbox"/> Embarked             |

[Remove]

**Selected attribute**  
Name: Age Group  
Missing: 0 (0%)  
Distinct: 5  
Type: Nominal  
Unique: 0 (0%)

| No. | Label      | Count | Weight |
|-----|------------|-------|--------|
| 1   | Adult      | 524   | 524.0  |
| 2   | NK         | 177   | 177.0  |
| 3   | Child      | 69    | 69.0   |
| 4   | Adolescent | 111   | 111.0  |
| 5   | Senior     | 8     | 8.0    |

Class: Age Group (Nom) [Visualize All]

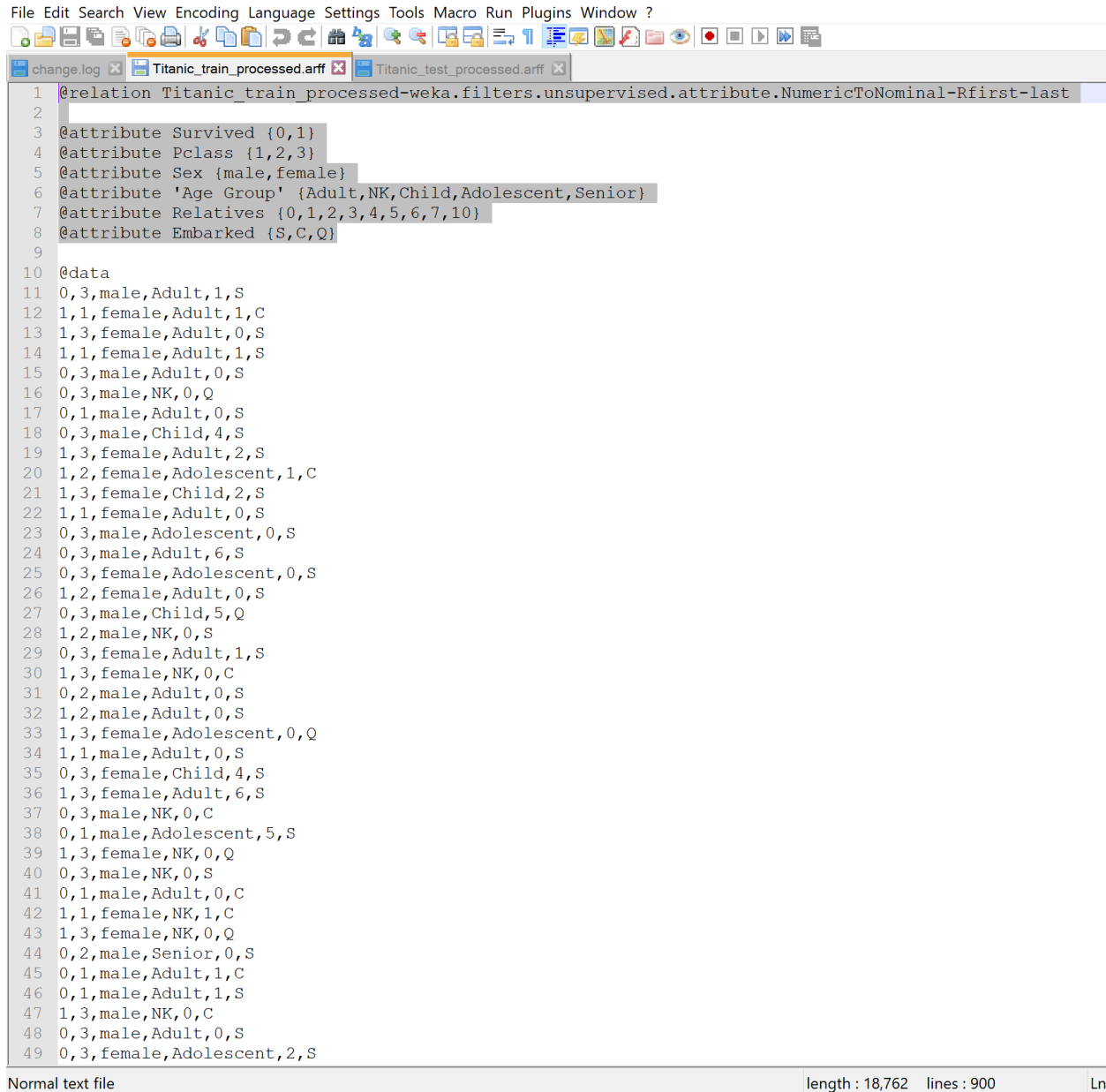
| Age Group  | Count |
|------------|-------|
| Adult      | 524   |
| NK         | 177   |
| Child      | 69    |
| Adolescent | 111   |
| Senior     | 8     |

**Status**  
OK [Log] x 0

## Saving the File in “.arff” format

After converting the attributes to their expected type. The file was ready to save as .arff file.

Below is the screenshot of the file opened in Notepad++

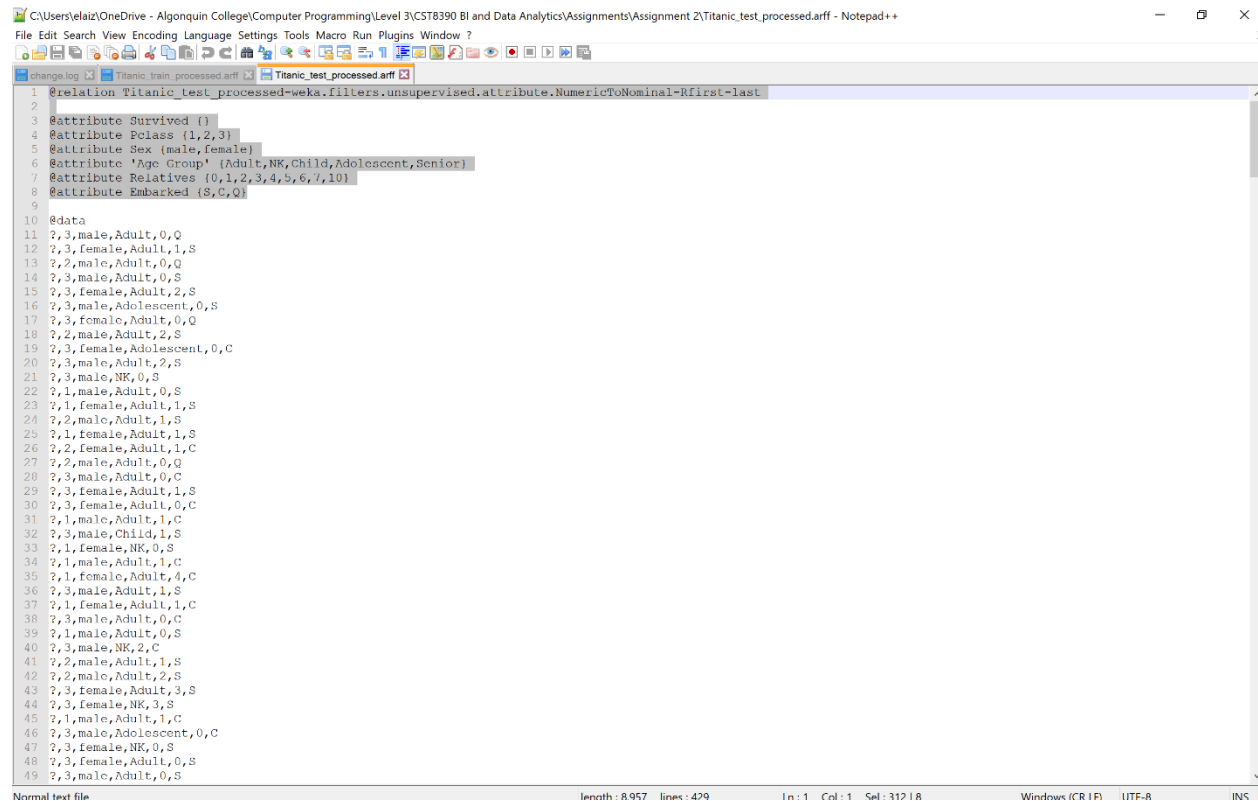


```
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
change.log x Titanic_train_processed.arff x Titanic_test_processed.arff x
1 @relation Titanic_train_processed-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
2
3 @attribute Survived {0,1}
4 @attribute Pclass {1,2,3}
5 @attribute Sex {male,female}
6 @attribute 'Age Group' {Adult,NK,Child,Adolescent,Senior}
7 @attribute Relatives {0,1,2,3,4,5,6,7,10}
8 @attribute Embarked {S,C,Q}
9
10 @data
11 0,3,male,Adult,1,S
12 1,1,female,Adult,1,C
13 1,3,female,Adult,0,S
14 1,1,female,Adult,1,S
15 0,3,male,Adult,0,S
16 0,3,male,NK,0,Q
17 0,1,male,Adult,0,S
18 0,3,male,Child,4,S
19 1,3,female,Adult,2,S
20 1,2,female,Adolescent,1,C
21 1,3,female,Child,2,S
22 1,1,female,Adult,0,S
23 0,3,male,Adolescent,0,S
24 0,3,male,Adult,6,S
25 0,3,female,Adolescent,0,S
26 1,2,female,Adult,0,S
27 0,3,male,Child,5,Q
28 1,2,male,NK,0,S
29 0,3,female,Adult,1,S
30 1,3,female,NK,0,C
31 0,2,male,Adult,0,S
32 1,2,male,Adult,0,S
33 1,3,female,Adolescent,0,Q
34 1,1,male,Adult,0,S
35 0,3,female,Child,4,S
36 1,3,female,Adult,6,S
37 0,3,male,NK,0,C
38 0,1,male,Adolescent,5,S
39 1,3,female,NK,0,Q
40 0,3,male,NK,0,S
41 0,1,male,Adult,0,C
42 1,1,female,NK,1,C
43 1,3,female,NK,0,Q
44 0,2,male,Senior,0,S
45 0,1,male,Adult,1,C
46 0,1,male,Adult,1,S
47 1,3,male,NK,0,C
48 0,3,male,Adult,0,S
49 0,3,female,Adolescent,2,S
```

Normal text file length : 18,762 lines : 900 Ln

## The Test File is Prepared in the Same Format with the Train File

The Titanic test file was then prepared the same way as the training sets. With the same format but with the Survive attribute set to “?” as a value. To test and compare the result with the training sets. Below is the screenshot of the test.arff file opened in notepad++.



```
1 @relation Titanic_test_processed-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
2
3 @attribute Survived {}
4 @attribute Pclass {1,2,3}
5 @attribute Sex {male,female}
6 @attribute 'Age Group' {Adult,NK,Child,Adolescent,Senior}
7 @attribute Relatives {0,1,2,3,4,5,6,7,10}
8 @attribute Embarked {S,C,Q}
9
10 @data
11 ?,3,male,Adult,0,Q
12 ?,3,female,Adult,1,S
13 ?,2,male,Adult,0,Q
14 ?,3,male,Adult,0,S
15 ?,3,female,Adult,2,S
16 ?,3,male,Adolescent,0,S
17 ?,3,female,Adult,0,Q
18 ?,2,male,Adult,2,S
19 ?,3,female,Adolescent,0,C
20 ?,3,male,Adult,2,S
21 ?,3,male,NK,0,S
22 ?,1,male,Adult,0,S
23 ?,1,female,Adult,1,S
24 ?,2,male,Adult,1,S
25 ?,1,female,Adult,1,S
26 ?,2,female,Adult,1,C
27 ?,2,male,Adult,0,Q
28 ?,3,male,Adult,0,C
29 ?,3,female,Adult,1,S
30 ?,3,female,Adult,0,C
31 ?,1,male,Adult,1,C
32 ?,3,male,Child,1,S
33 ?,1,female,NK,0,S
34 ?,1,male,Adult,1,C
35 ?,1,female,Adult,4,C
36 ?,3,male,Adult,1,S
37 ?,1,female,Adult,1,C
38 ?,3,male,Adult,0,C
39 ?,1,male,Adult,0,S
40 ?,3,male,NK,2,C
41 ?,2,male,Adult,1,S
42 ?,2,male,Adult,2,S
43 ?,3,female,Adult,3,S
44 ?,3,female,NK,3,S
45 ?,1,male,Adult,1,C
46 ?,3,male,Adolescent,0,C
47 ?,3,female,NK,0,S
48 ?,3,female,Adult,0,S
49 ?,3,male,Adult,0,S
```

## Classification

The classification was performed by using decision tree method. First the file is loaded to Weka, and J48 classifier was chosen with 10-fold validation. The confusion matrix produced from the cross validation can be seen below, which has overall 713 correctly classified instances out of 889 instances (80% correct).

```
a    b    c    <-- classified as
498  51    0 |    a = 0
125 215    0 |    b = 1
  0   0    0 |    c = ?
```

### Confusion Matrix

## Decision Tree

Here are the predictions made by the decision tree from the training data set.

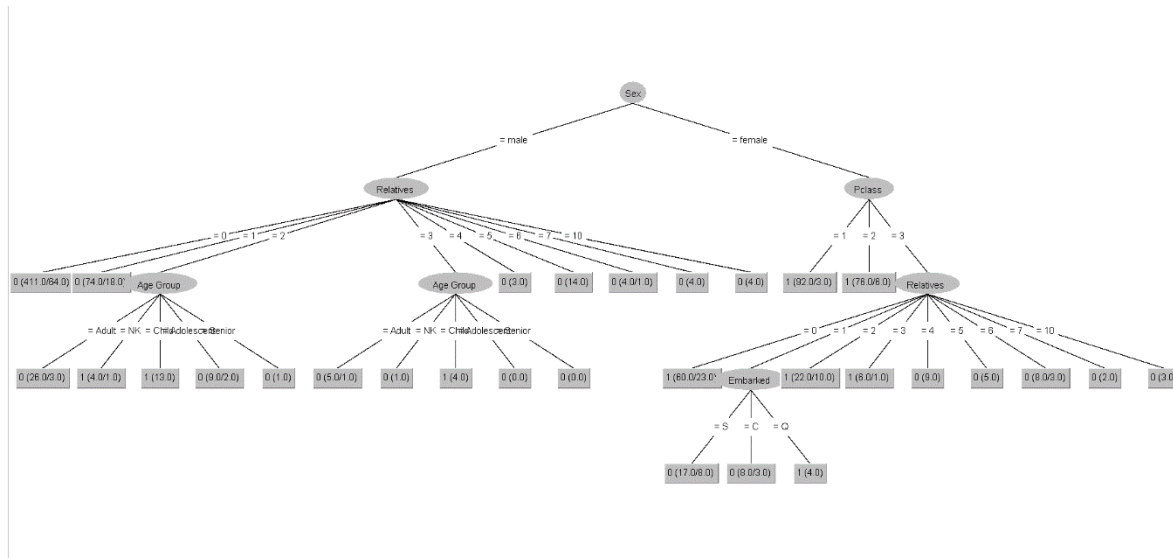
-> The decision tree has the root value of sex, indicating it is the most relevant attribute of a passenger's survival. Females had a better chance of survival.

-> Females in first and second classes survived.

-> Third class females who have more than 3 relatives did not survive. The ones having one relative and embarked from Queenstown survived, the others did not.

-> Regarding the males, only the children having 2 or 3 children had a chance of survival. There are also survivors with the unknown ages.

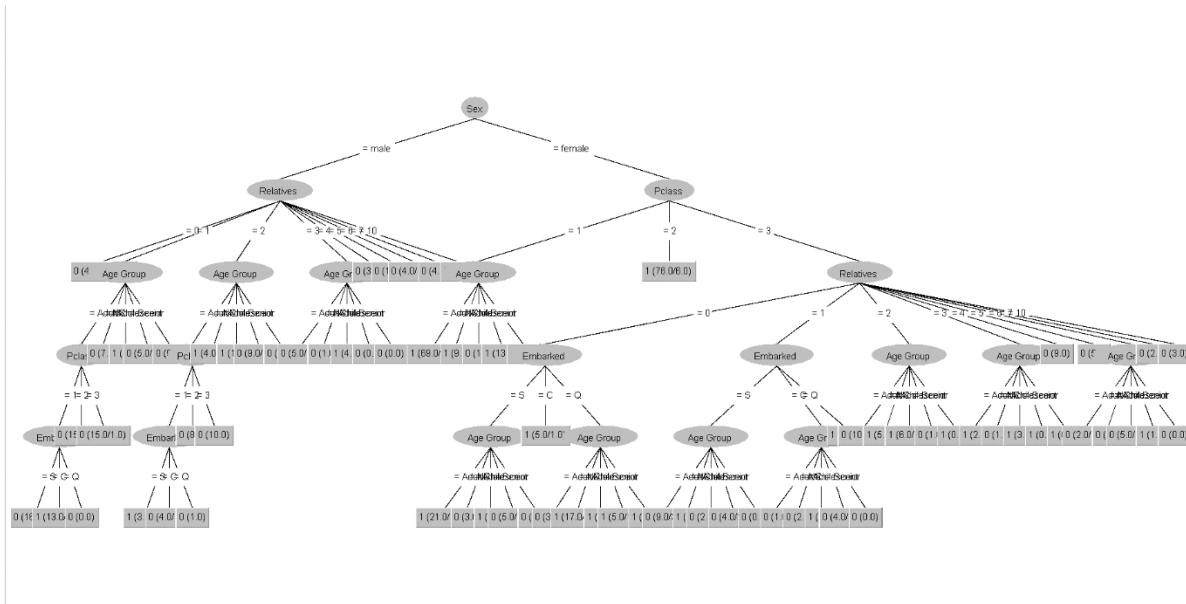
Below is the decision tree for more detailed information.





## Bonus: What if the Tree Was Unpruned?

By default, the tree is pruned, which means the non-critical and redundant classify instances are removed. When the “unPruned” option is set to true, the size of the tree gets drastically bigger as it can be seen below.



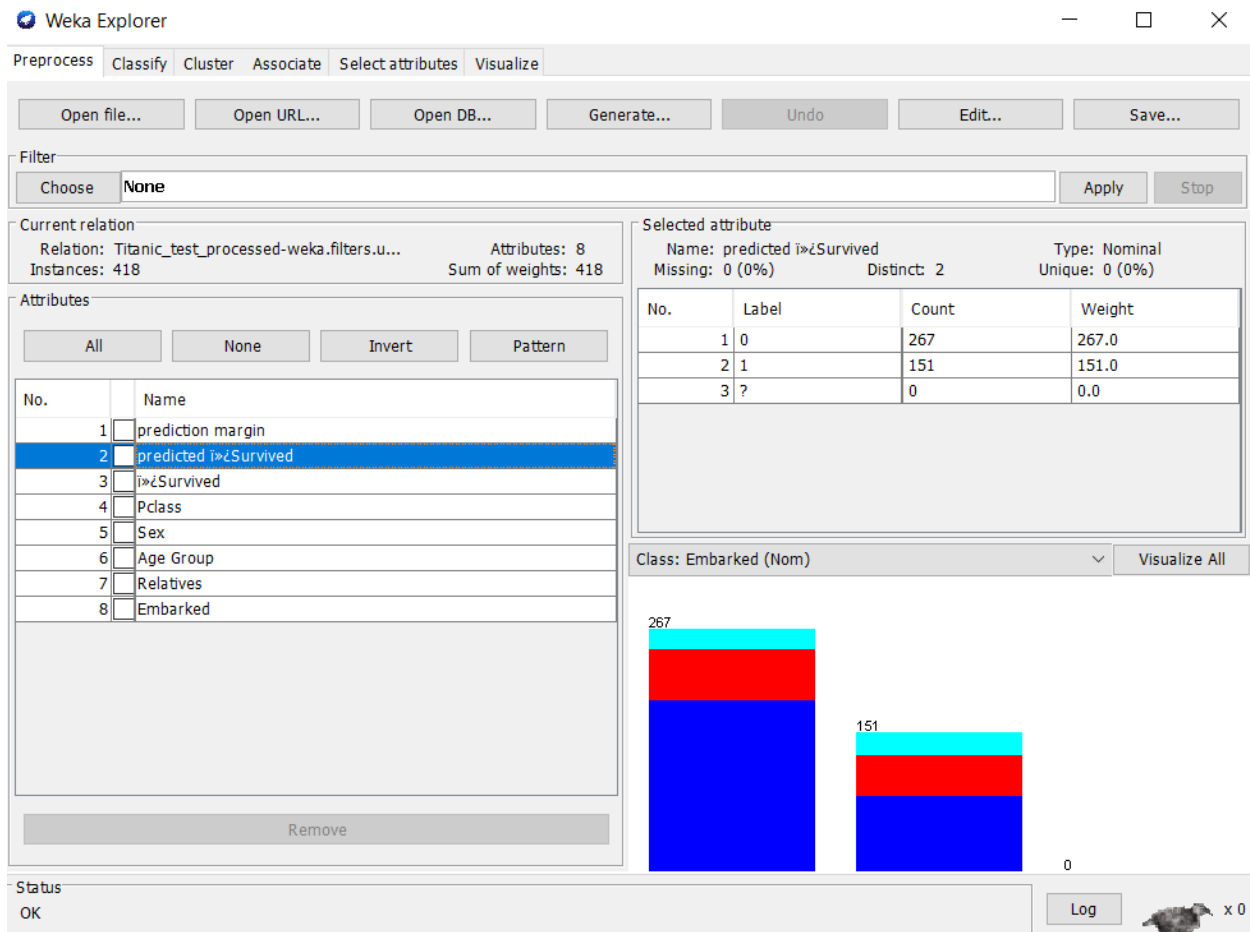
The pruned tree is used for the analysis as it is less complex and it improves predictive accuracy by reduction of overfitting. (Info retrieved from Wikipedia.)

## Adjusting the arff File For the Test Set

When we tried to run the decision trees for the test set, we encountered with the error saying these data sets are not compatible. The reason was that all the attributes should have been identical in which, the Survived attributes had different value categories. So, both attributes were changed in the .arff files to “Survived {?,0,1}” since the training set had 0 and 1, whereas the test set had only “?” As the value.

## Test Set and “res” File

After making the changes, the test set became able to be supplied. The same decision tree algorithm was run on the test set, and got the output of all empty values as there was no actual survival information. In order to see the actual results, the classifier errors were visualized on the output file and saved as another file(res.arff).



The prediction results can be seen on res.arff file. Prediction margin and predicted survived attributes are automatically generated by Weka, and we are looking at the predicted survived attribute for the relevant information. Here are the results of the prediction based on the decision tree.

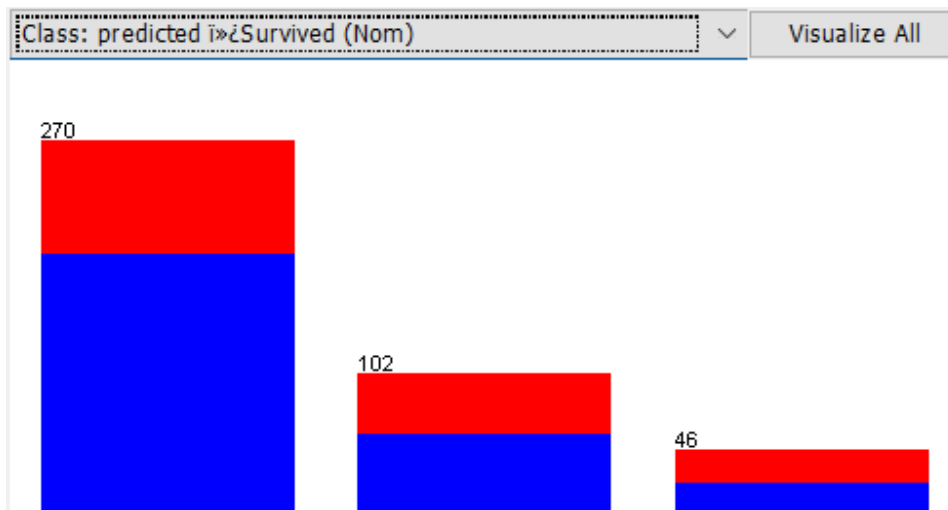
- There are total of 418 instances in the test file.
- 151 persons were predicted to survive.
- 267 persons were predicted not to survive.
- Percentage of predicted survival is 36.12%.

## Comparing to the Actual Results

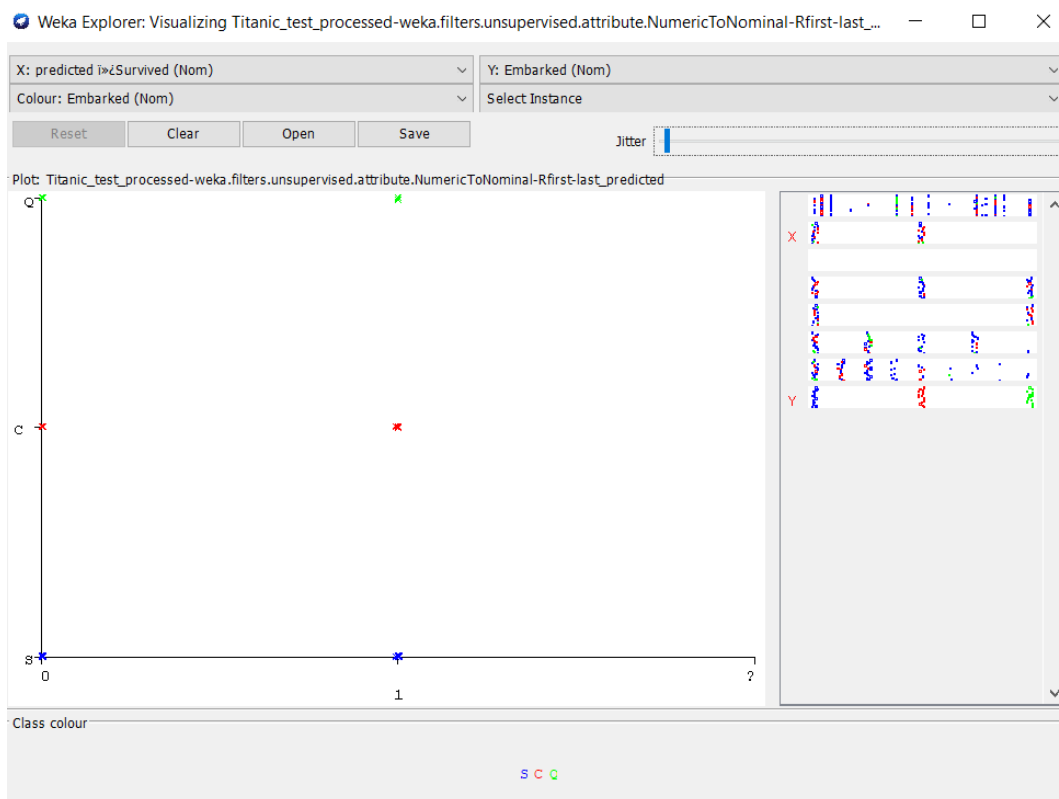
According to the actual results, 492 passengers were survived from the incident, which has a survivor percentage of 37%. The result is fairly close to our algorithm's prediction with roughly 1% deviation.

There are certain reasons influencing the deviation difference.

As it can be seen from the chart of port of embarkation and survival prediction, the port has no significant effect on a passenger's survival more than a coincidence. However, in the decision tree generated, it was one of the factors deciding whether the person will survive or not.

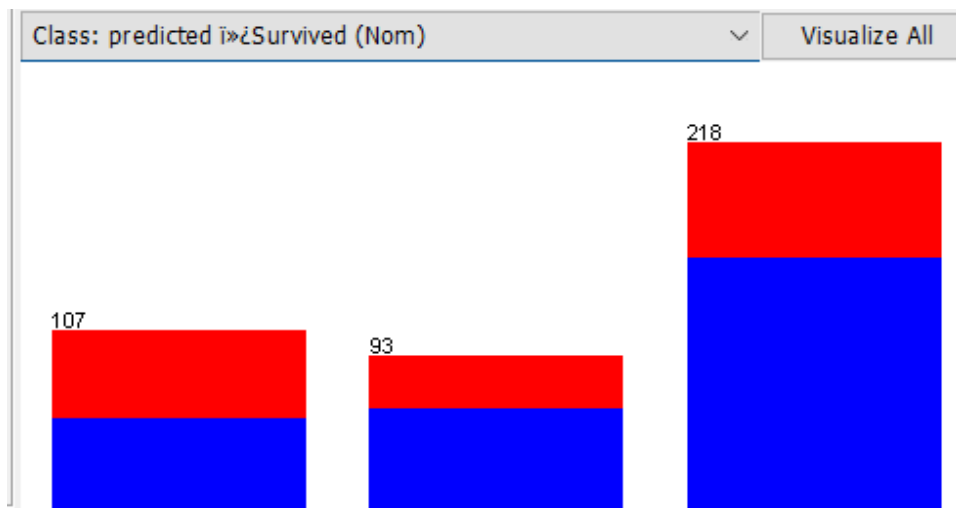


Here is the comparison of embarkment ports and survived data. There is no real-life information on if the embarkment port has an effect on survival.

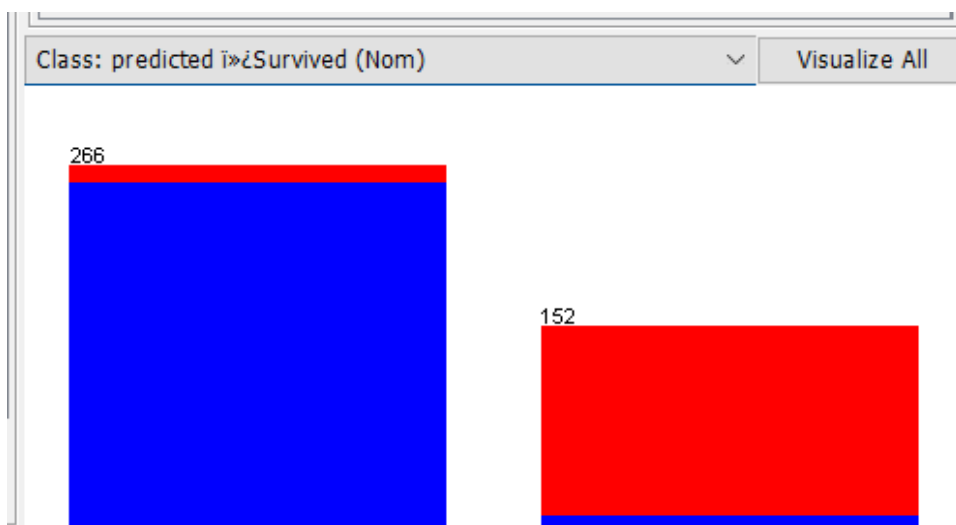


Also, the cabin might have been a factor for prediction, but since there were too many missing data on cabinet, we preferred to not include it on the data sets.

The success of the prediction is because for the decision tree, the most significant factors were sex and class, so was for the actual statistics.



*Respectively first, second, and third class (Victims are blue, survivors are red)*



*Respectively female and male (Victims are blue, survivors are red)*