CSC8111 Machine Learning

Note: If you want to run my code please use the RFandDT.ipynb file using python 3. I have also added an RFandDT.py file just in case the but I haven't executed that one I just copied the code from the .ipynb file.

For this coursework, I analysed the Titanic dataset which consists of approximately 1300 records, divided into training and test subsets for the Titanic passengers, to predict which of the passengers were more likely to survive based on several attributes. From a brief research of the Titanic disaster it can be seen that females passengers and passengers under the age of 12 had a greater probability of surviving the disaster in comparison with the rest of the passengers. Additionally, passengers who belonged to the upper-class had also a higher probability of surviving rather than the passengers that belonged to the lower-class. Because cabins in upper classes were located on the higher desks of the ship, thus the upper-class passengers could access the life boats quicker than the passengers of the lower classes.

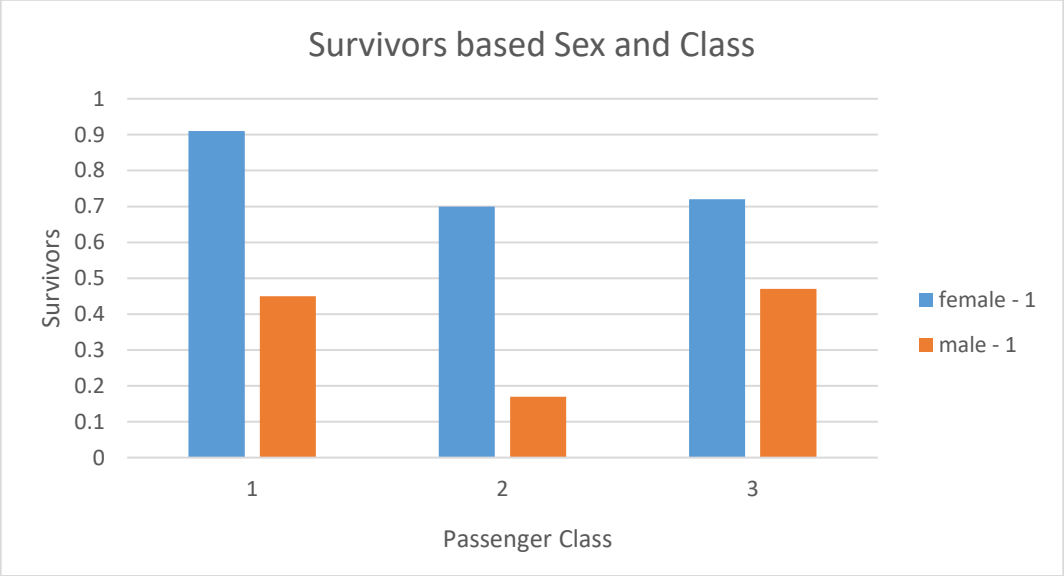

Figure 1. The proportion of survivors based on their class and sex.

I used a Random Forest classifier to predict the survivors of the disaster. The Random Forest classifier works only with integers, thus the strings were converted into integers. Moreover, missing values were filled with the median values of that attribute, and the categories Name, Ticket, Cabin and PassengerId were removed. Figure 1 shows clearly, that regardless of the class in which a passenger belongs to, if they are females the probability of surviving the disaster is greater than for a male passenger. Moreover, a male passenger in first-class has a greater probability to survive than male passengers in lower classes. This means that the Sex and the Class of a passenger are important in the prediction of the survival. However, since Kaggle provides the gendermodel python class we can remove the Sex column as well and work with the rest. Since we are interested to determine which attributes play a role in

whether a passenger on the ship will survive or not the disaster, the response from the classifier would be binary, 0 if the passenger did not survive and 1 if the passenger survived.

I will use two performance evaluation techniques to evaluate the performance of two different classifiers; the 10-fold cross-validation and the Student paired t-test technique. For the 10-fold cross-validation technique, the data is divided into 10 folds and for each fold you partition the data into a test subset and the training subset once you train on the train subset you test on the test subset. This technique is useful when we are dealing with small and imbalanced data sets, because every sample gets to be used in both training and testing. However, for large data sets it can be computationally intensive. Moreover, the Student paired t-test allows us to test if the difference between two classifiers is statistically significant, after the two classifiers are evaluated with the same method, in this case the 10-fold cross validation method.

The classifiers that I used to predict the survival of passengers is the Random Forest classifier and the Decision Tree classifier. The Random Forest Classifier constructs several decision trees while training, each tree gives a classification, and from the classification produced by each of those trees the final output is calculated.

The Decision Tree Classifier relies only on one decision tree to predict whether a passenger survived or not.

The following figure 2 is the result from the 10-fold cross validation and the paired t-test methods I have used.

```
Accuracy of Random Forest: 0.81 (+/- 0.11)
Accuracy of Decision Tree: 0.79 (+/- 0.09)
Paired t-test:  6.32455532034
```

Figure 2. Results from the two classifiers

The Random Forest Classifier has a slighter better accuracy than the Decision Tree Classifier, but we still need to use the paired t-test method to measure the statistical significance of the results from the two classifiers. Since we have 9 degrees of freedom, which corresponds to 3.250 based on the distribution table in the lecture notes and since the **t** value I have calculated is greater than 3.250, we ascertain that the observed difference in performance is significant at a level 1%, which mean that the Random Forest classifier is better than the Decision Tree Classifier.

Following Figure 3, was taken from the Kaggle website it shows the importance of the attributes in making predictions whether a passenger survived or not for the Random Forest classifier. We can clearly see that for the Random Forest classifier the Sex, the Class, the Fare and the Age of the passengers are the most important attributes in predicting whether a passenger has survived the disaster or not. These results come agree with the initial brief

research (Figure 1) which showed that the sex, and Class of a passenger plays a role in predicting if a passenger survived or not.

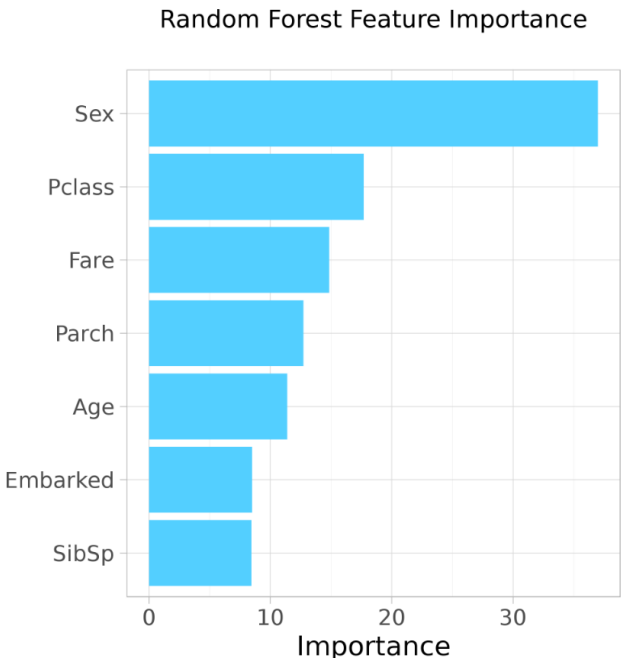

Figure 3. The Importance of the attributes in the Random Forest Classifier [1]

Moreover figure 4 shows an example of tree used by a Decision Tree Classifier for the Titanic disaster. As it can be seen Sex is again the most important attribute, followed by the Class and Age of a passenger. We can also see that for a male passenger his Age is more important rather than his class in predicting if he survived or not. However, for a female passenger her class, and then her family size and her Age are more important attributes to predict if she survived or not in comparison with a male passenger. Thus, we can say for a male passenger they will have to be in the first class to have a higher probability of survival, but for a female the class she belongs in does not matter that much, because her Sex, Family size and Age are as important as her class.

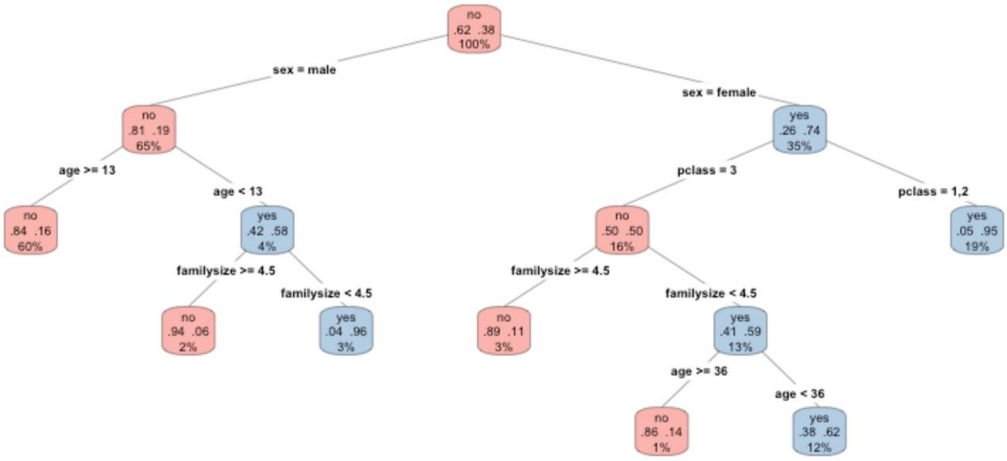

Figure 4: Example Decision Tree of the Titanic Disaster. [2]

References:

[1] Inc, K. (2016) Titanic: Machine learning from disaster. Available at: https://www.kaggle.com/benhamner/titanic/random-forest-benchmark-r/code (Accessed: 14 December 2016).

[2] Buhrmann, T. (2014) Titanic survival prediction. Available at: https://buhrmann.github.io/titanic-survival.html (Accessed: 14 December 2016).

Name: Stella Englezou, 130422246

Date: 14/12/2016