John Robinson

Lab 4 Report

In finding a k-NN classifier for the titanic survivor dataset I first examined the data for missing entries. After finding that Age, Cabin and Embarked had missing values in the training set, along with Age Cabin and Fare missing values in the testing data I knew that there was missing data to be compensated for. To fill in the missing data I got the idea to fill in the remainder of missing values with the median or mean depending on which seemed more relevant from another Kaggle user. Also the idea to drop some of the columns as they technically were not relevant to survival rate such as name or passenger ID. Next I changed non numeric columns to numeric values such as sex or embarked. Next was finding kNN classifier using the sklearn import. Getting an averaged knn accuracy of about 76.47. Along with that using the sklearn naïve bayes import I got an accuracy of about 76.59. Lastly I used another sklearn function to complete a tenfold cross validation for both the kNN and NB classifiers. I did not get to the analysis with the ROC curves and confusion matrixes as I did not completely understand them at the time. I left citations of the examples I used to assist me in the Jupyter worksheet, but I will add them here too. I am not exactly sure what you wanted me to display in the report, but I explained to the best of my knowledge