Project 2 Paper

In our Project Raj and I chose to do the Titanic because we found it has meaningful data with categorical variables that would be great for exploratory analysis. The Titanic dataset was looking for analyzing the survival of the people on board the Titanic by their Age,Sex, Embarked, and Pclass. To demonstrate who had the best survival. We found the full dataset of the Titanic and then we split the full Titanic dataset into train and test. We made train to only have columns Survived, Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked. Test only has Survived. Then we cleaned the train dataset by dropping all na values so that there are no values then we analyze. We analyzed the Titanic dataset by showing a graph of Sex and Survived to see which sex had the higher survival rate and which had the higher death rate. Male had the higher death rate than Females and Females had higher survival rate than Male but Female had close values of Survived and Dead while Male had like 90% dead and 10% survived. Then, we did a graph of Pclass compared to Survived to see how each Pclass looked like with their survival. The Pclass 3 had the highest death rate and Pclass 1 had the highest survival rate. We made a graph to show where they Embarked from to show their Survival and we found Cherbourg, France to show the highest Survival rate of where the Titanic was boarding. We made a boxplot to show the age of the pclass to represent what the ages were in each pclass. I wanted to show the total numbers of survived and dead and 424 dead and 288 survived. Then, we made a correlation matrix of the train dataset to show the correlation of all the variables in the train dataset. Then, we made a correlation heatmap of the correlation matrix to give a visual representation of the correlation matrix to better demonstrate how each variable is correlated. Then, we split the train and test and made the test size to 30% and the train size to 70%. Then we dropped the Embarked column in the train so that there are no categorical values when we do our machine learning. Then we cleaned the new train dataset by dropping all na values so that there are no na values when we do our machine learning.

We split the sex column to where male equals 0 and female equals 1 to help us distinguish which is male and female in our new train dataset. The machine learning algorithms we chose to do are Logistic Regression and Decision Tree classifier. The accuracy score of Logistic Regression is .77 if you round it up it is 77%. Then we made a classification report for Logistic Regression and Decision Tree classifiers. For the decision tree classifiers and the accuracy score of that was 73%. The first accuracy score of the decision tree was 76% but we wanted to optimize the decision tree performance and the accuracy score of the new decision tree is .775% if you round it up would be 78% so the new decision tree has the highest accuracy score.

The new decision tree would be the best model to use to evaluate our prediction compared to Logistic Regression and the old decision tree.