RISK ASSESSMENT ON THE TITANIC

To begin with the analysis, the summary statistics of train and test data was prepared. As initial step I checked the relationship between various features and the survival rate using bar graphs and correlation matrix. Upon checking the distribution of survival and sex of the test data, I could find that all females survived while all the males died(fig1). This indicated that there must be a bias in the splitting of data into train and test sets. And sex was highly correlated feature with survival. Therefore, I combined the test and train data. As we can see in fig2 (0-Male, 1-Female) the survival is now distributed unbiased. I filled in the missing values in age and fare with mean values and missing values in embarked with mode. The variables that were just for the individual and that had no predictive power were removed, which are Name, Passenger Id, Cabin, Ticket type. I combined parent child feature and sibling/ spouse feature to generate a new variable to see if travelling alone or not affected the chances of survival (Fig3). The features that I used after cleaning and EDA are in fig4. I ran a decision tree model to check how many survived on the basis of the chosen features. After cost complexity pruning, I chose the alpha that would make optimal amount of splits to predict the model with utmost accuracy which is 0.0025(Fig5 &6). I also ran a probit model to predict the survival, however decision tree(85.9%) was more accurate than probit model(84.7%). Fig 7 & 8 is the confusion matrix presenting the actual and predicted value by decision tree and probit respectively

Figure 1

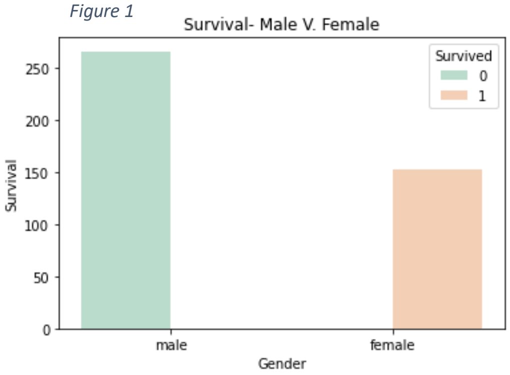


Figure 2

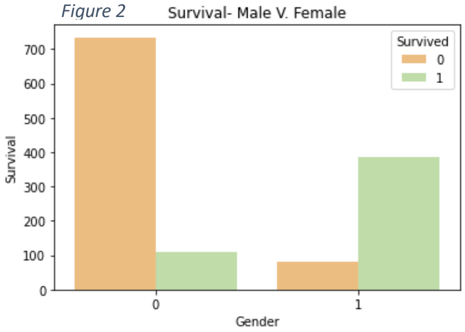


Figure 3

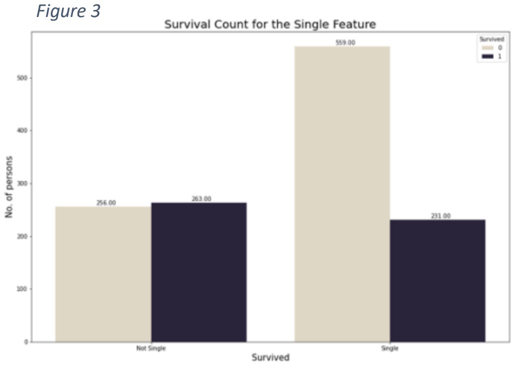


Figure 4

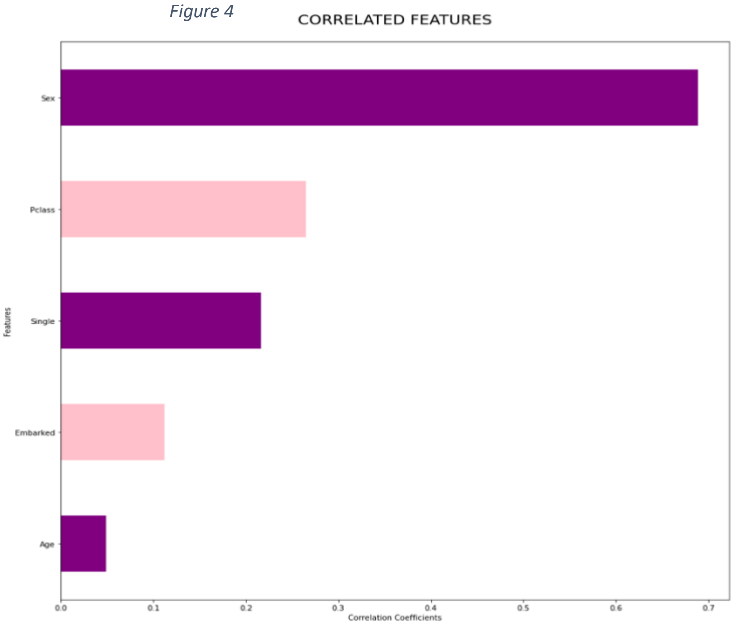


Figure 5

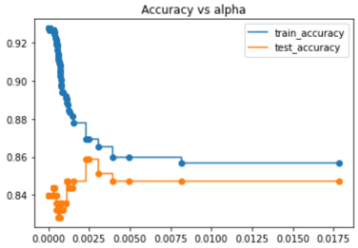
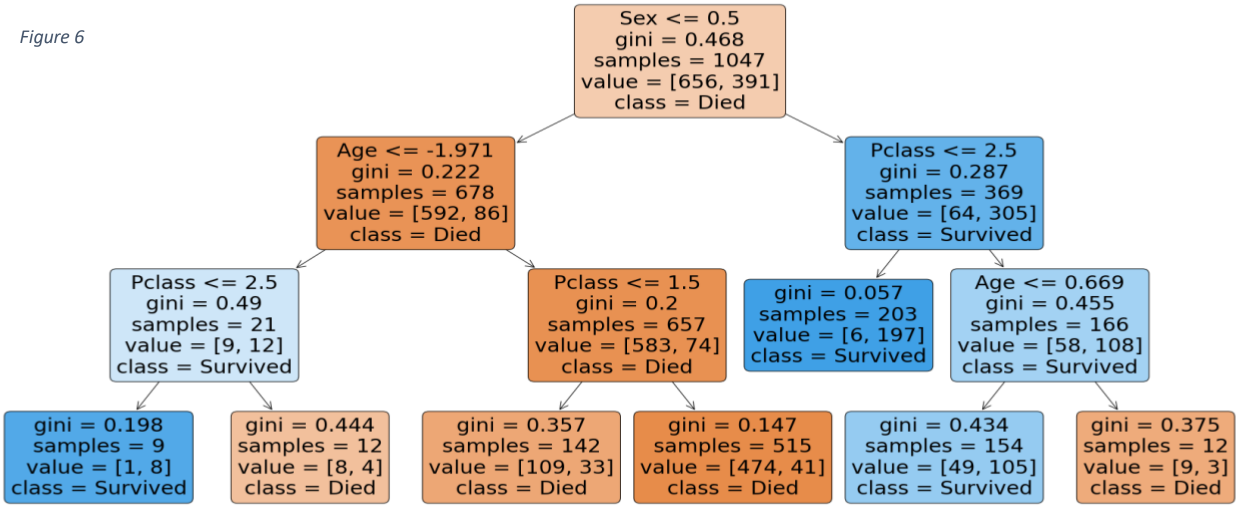
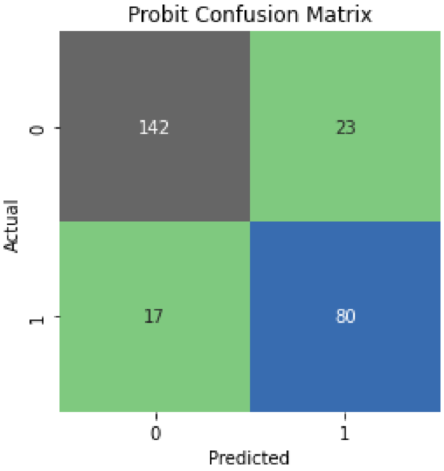
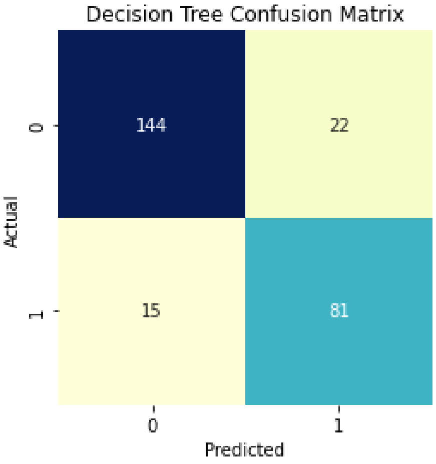


Figure 6


Sex <= 0.5
gini = 0.468
samples = 1047
value = [656, 391]
class = Died
Age <= -1.971
gini = 0.222
samples = 678
value = [592, 86]
class = Died
Pclass <= 2.5
gini = 0.287
samples = 369
value = [64, 305]
class = Survived
Pclass <= 2.5
gini = 0.49
samples = 21
value = [9, 12]
class = Survived
Pclass <= 1.5
gini = 0.2
samples = 657
value = [583, 74]
class = Died
gini = 0.057
samples = 203
value = [6, 197]
class = Survived
Age <= 0.669
gini = 0.455
samples = 166
value = [58, 108]
class = Survived
gini = 0.198
samples = 9
value = [1, 8]
class = Survived
gini = 0.444
samples = 12
value = [8, 4]
class = Died
gini = 0.357
samples = 142
value = [109, 33]
class = Died
gini = 0.147
samples = 515
value = [474, 41]
class = Died
gini = 0.434
samples = 154
value = [49, 105]
class = Survived
gini = 0.375
samples = 12
value = [9, 3]
class = Died


Probit Confusion Matrix
0
1
Actual
142
23
17
80
0
1
Predicted


Decision Tree Confusion Matrix
0
1
Actual
144
22
15
81
0
1
Predicted