Summary Report For Titanic DataSet

I covered the following steps for solving the problem.

Data Preprocessing Stage

Step 1: Import libraries

- Imported the numpy and csv libraries. For using Random Forest Classifier imported it from sklearn.

Step 2: CSV, train & test Data (loading & reading data)

- Created the bin for both train and test data.
- Load in the test csv file
- Skip the first line because it is a header
- Skip through each row in the csv file
- Add each row to the data variable
- Then convert from a list to a Numpy array

Step3: Run Statistical summaries to perform computational analysis

- Convert strings to number for gender and embark

Step4: Missing Values Imputation

- All the ages with no data make the median of the data
- All missing embarks just make them embark from most common place
- remove the name data, cabin and ticket
- Did above steps for both training and test data

Step5: Implemented ML Algorithm to train and test

- I tried to implement the Rainforest classification algorithm
- predicted the survival rate by the features gender, the port they embarked, age thinking women and children first.
- A random forest model is a collection of decision tree models that are combined together to make predictions. When you make a random forest, you have to specify the number of decision trees you want to use to make the model. The random forest algorithm then takes random samples of observations from your training data and builds a decision tree model for each sample. The random samples are typically drawn with replacement, meaning the same observation can be drawn multiple times. The end result is a bunch of decision trees that are created with different groups of data records drawn from the original training data.