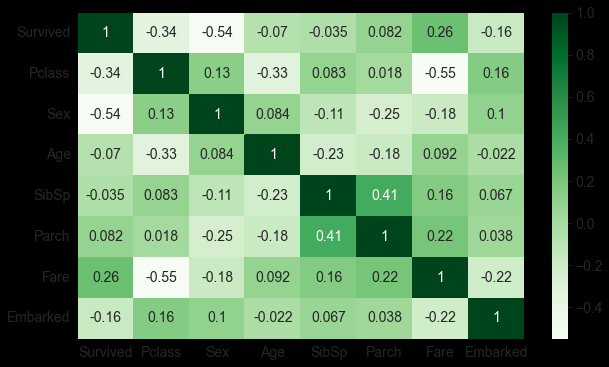
**KNN description**

The K-Nearest Neighbor is a regression, clustering, and classification algorithm. Depending on the need and intended use, it can be used to find clustered values in dataset and it can as well be used to classify data into two or more classes. The algorithm uses the entire dataset during the training, and when a new datapoint is availed or a prediction is required, the algorithm finds and returns the most similar value from the dataset. The algorithm used different techniques to measure the similarity between different variables. The most applied methods are the Euclidean distance, Manhattan distance and Minkowski distance. The evaluated distances are used to evaluate the nearest neighbors to each datapoint, which are then used to predict the classes of unknown data. KNN algorithms use the lazy learning method since an evaluation is only made when a prediction is required. The algorithm does not have the autonomy the learn the data as it only copies the entire dataset into memory and calculates.

**Feature Selection**

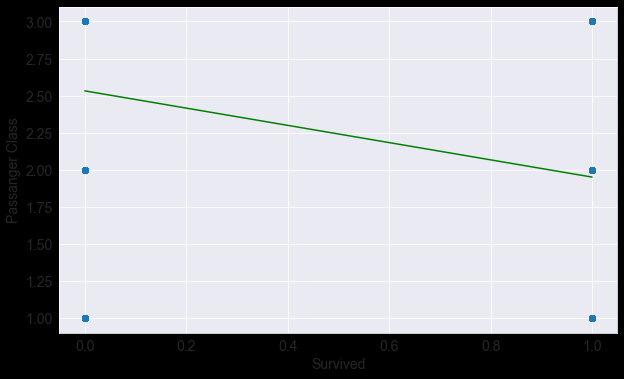
For the classification, three features were selected; Pclass, Sex and Fare. Feature selection is depended upon the understanding of the problem statement. From the dataset, it can be deduced that surviving (Survived column) was based upon the other variables. It is prudent, by elimination, to delete any features that are not likely to aid in the survival of an individual. For instance, their names and ticket numbers.

For the remainder of the features, we deploy a correlation analysis to determine which features were highly correlated to the Survived column. The passenger class, their sex, and the amount they spent showed reasonable correlations to the target variable. 

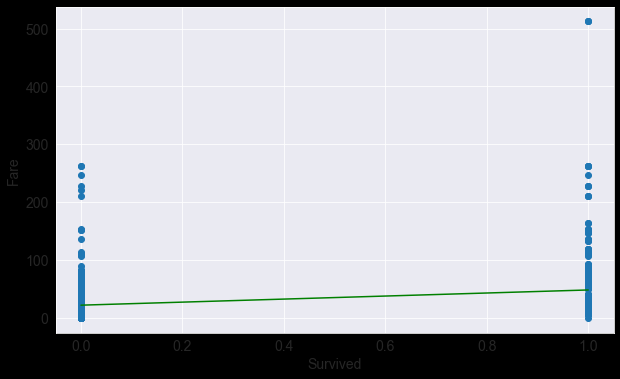
Passenger class had a correlation value of -0.34, sex had a correlation value of -0.54 and finally fare had a correlation value of 0.26. There is also a high correlation between passenger class and fare – implying that the more one spend, the more they were likely to be in a different class, the more they were likely to survive. Other feature that would be considered most relevant would-be age and parch which recorded correlation values of -0.07 and 0.082 respectively.

**Target variable and select feature’s visualization**

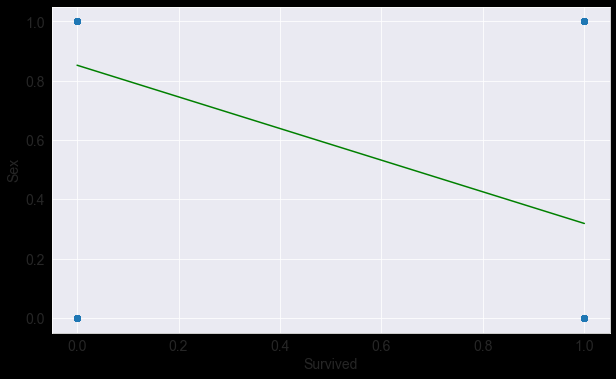
These are correlation plot for each of the three select features and the target variable.



Passenger class correlation score -0.34. The chances of survival decrease as from once passenger class to another, implying that passenger in class 0 had a higher chance for survival compared to their counterparts in class 1.



Fare correlation score 0.26. There is a slight increase in the chances of survival moving from class 0 to 1. People who spent more money had a slightly higher chance of survival compared to their counterparts.

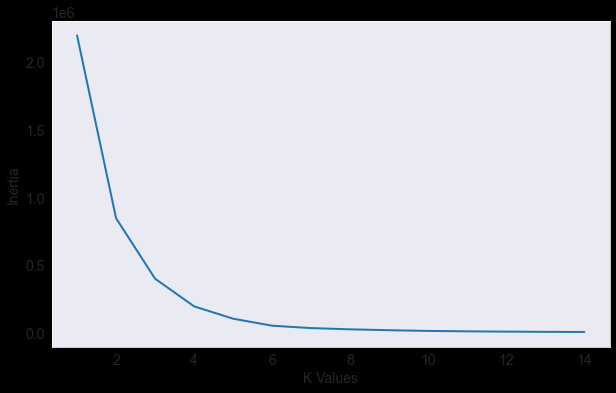


Sex, a correlation score of -0.54. Sex was a key determinant for survival as there is a negative chance for survival moving from subclass 0 (female) for subclass 1 (male).

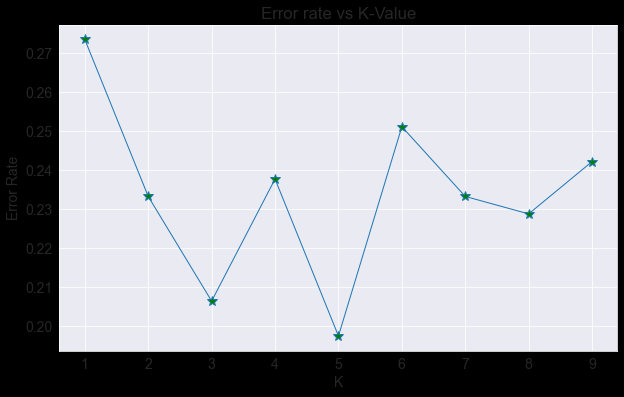
**Results**

Passenger class, Sex and fare were important predictors for survival. This can be seen in the generalization of the KNN classifier which scores an accuracy of 80.269% using these features to predict the chances of survival. A combination of features did not score as high even for different values of k neighbors.

An optimal value for k neighbors is as well important for the optimal performance of the KNN classifier.



Using the elbow method, it was determined that the most optimal range for the value of k-neighbors was between 2 and 5, with 5 n-neighbors scoring the highest.



The error rate plotted against the different values of k, shows an almost trend with elbow curve. The curve bends (create and elbow) at k = 2 and reaches its optimal performance at k=5 (at this point, the classifier has the least error), an indication of a high accuracy. At an error rate of ~0.20, the classifier has an accuracy of ~80% i.e., (1-0.20) \* 100.

**References**