**Project Report**

**Exploratory Data Analysis and Logistic Regression and K-nearest Neighbour**

1. **Introduction:-**

Machine learning playing a critical role in classifying, grouping and predicting the results. Machine learning algorithms come in different types based on their mathematical combination and working principles. Machine Learning has three types:-

1. Supervised Machine Learning

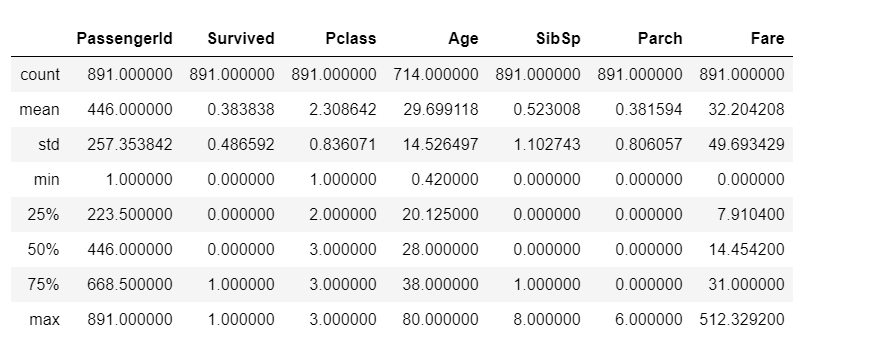
2. Unsupervised Machine Learning

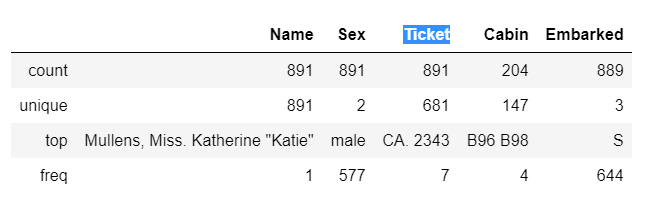
3. Reinforcement Machine Learning

We used in our project two supervised machine learning algorithms. One is Logistic Regression and other is K-nearest Neighbour. We used the famous titanic dataset in our project. In this project, our research question is that how many people were survived and how many people died. We also perform exploratory analysis to find their ages, sex, and passenger class type in project. The age, sex and passenger assist us in prediction of survived target variable.

1. **Methodology:-**
   1. **Dataset Description:-**

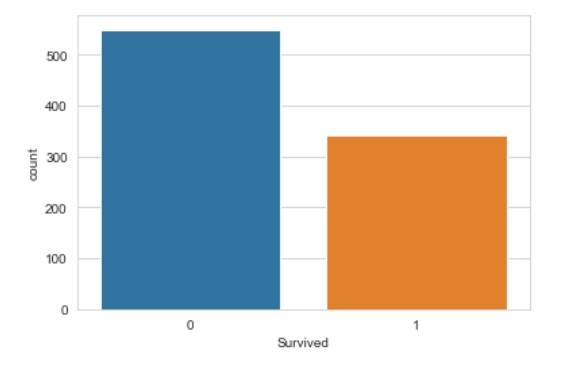
The dataset used in this project is titanic that is publicly available on kaggle platform (<https://www.kaggle.com/c/titanic>). The dataset is sample dataset not population dataset. The dataset consist of 12 attributes and 891 records. The dataset contain the missing values. The dataset contain five categorical features and 7 numeric features. We performed the statistical analysis of dataset w. r.t categorical features and numerical features. The statistical features results are given below:-



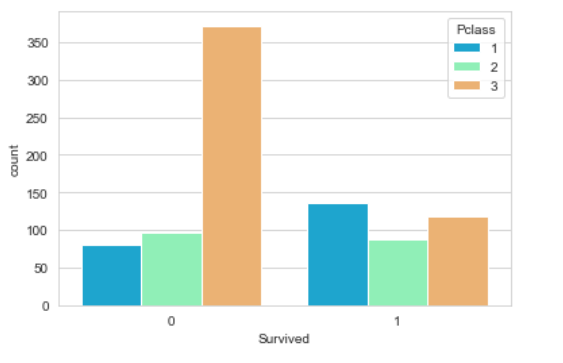


**2.2 Exploratory Data Analysis**

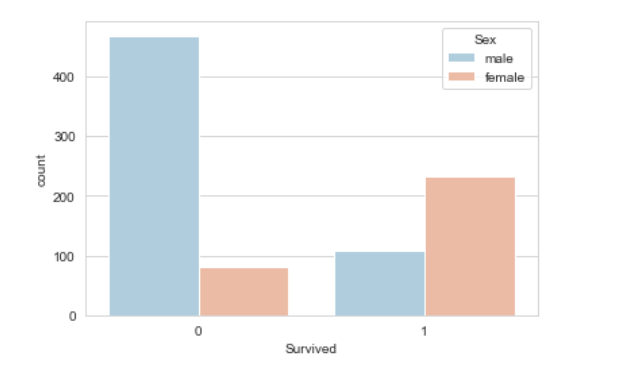
In our dataset, survived attribute is the target variable where as other variable are in depended variable. We performed exploratory analysis with respect survived variable. We find from EDA analysis of survived variable contain that the number of survived and not survived people. People 542 were not survived and 342 survived.



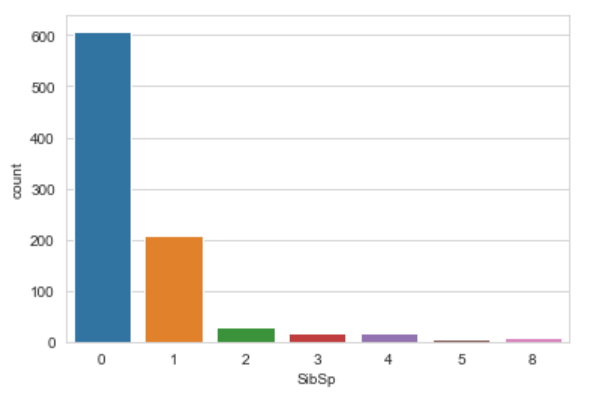
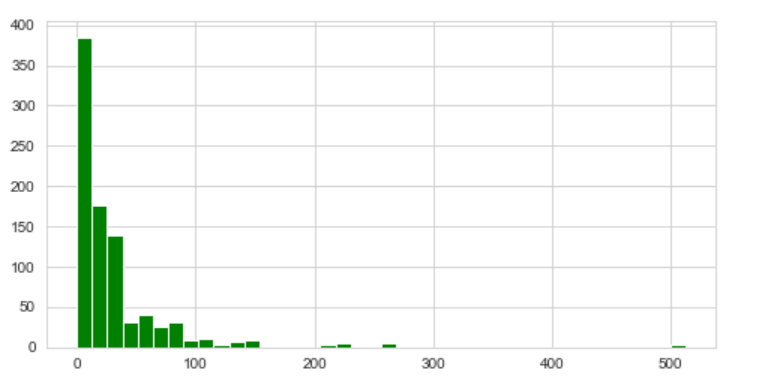
We also do the EDA of Pclass attribute with respect to survived feature to find that in which Pclass category of people survived or not survived. We find that category of 3 people not survived more than the two other categories. W also find the category of 1 people survived more than category 2 and 3.

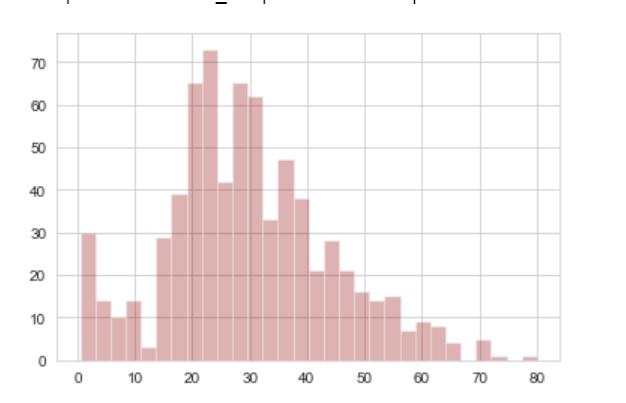


The sex attribute EDA analysis with respect to the target class showed the number of peoples who die or survived were men or women. We found that the number of man were greater than woman who were survived.we also found that number of women were greater than survived than the man.



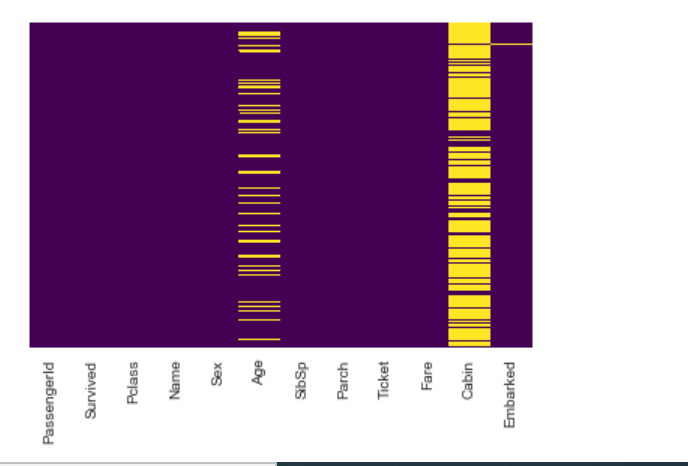
We also performed the Exploratory analysis of Fare group, age and SipSP attributes. The visualization representation of their graph shows the results with respect to the target variable.





* 1. **Missing Values-Data Cleaning:-**

The two attribute of dataset contain the missing values that are age and cabin. The age contains almost 20% missing values while Cabin contains 90%. So we dropped the cabin attribute and removed the missing values age attribute with respect to the value of Pclass attribute. The results of missing values are show through heatmap chart.



We dropped the attribute that do not contribution in predicting the model. The name of those attributes are ‘Name', and ‘Ticket'

**2.4 Transformation:-**

As the Machine Learning does not accept the categorical features. So that we convent them through One hot encoding method to numeric features. We convert Sex and Embarked attributes by using get dummies method. After that original attributes are dropped.

2.5 **Train/Split** :-

We split the dateset into train and test parts. The test part contains 30% while remaining is the training part. We used random state 101.

2.6 **Machine Learning:-**

a. **Logistic Regression:-**

Logistic Regression is type of supervised machine learning algorithm that deals the regression problems. This is probabilistic model and its value lie between 0 and 1. It threshold values is 0.5. The Logistic regression take the independent variable and make comparison of their with respect to threshold and predict the depended variable. The formula of logistic regression is 1/1+e(-z).

b. **K-nearest Neighbour:-**

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time.

1. **Performance Evaluation:-**

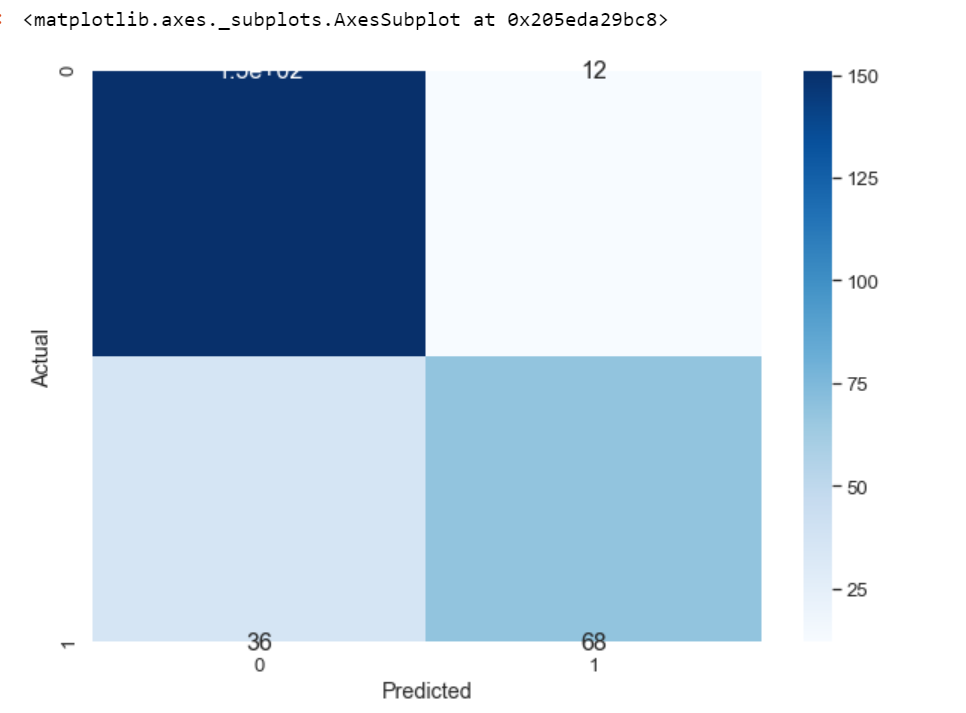
We performed the evaluation of both algorithms by using the cross validation technique. Both algorithm are powerful and outperform on the benchmark dataset. The overall logistic regression performance better than KNN.

1. **Logistic Regression:-**

**Performance Evaluation:-**

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1-Score |
| 0.80 | 0.81 | 0.93 | 0.86 |

The confusion matrix is shown as



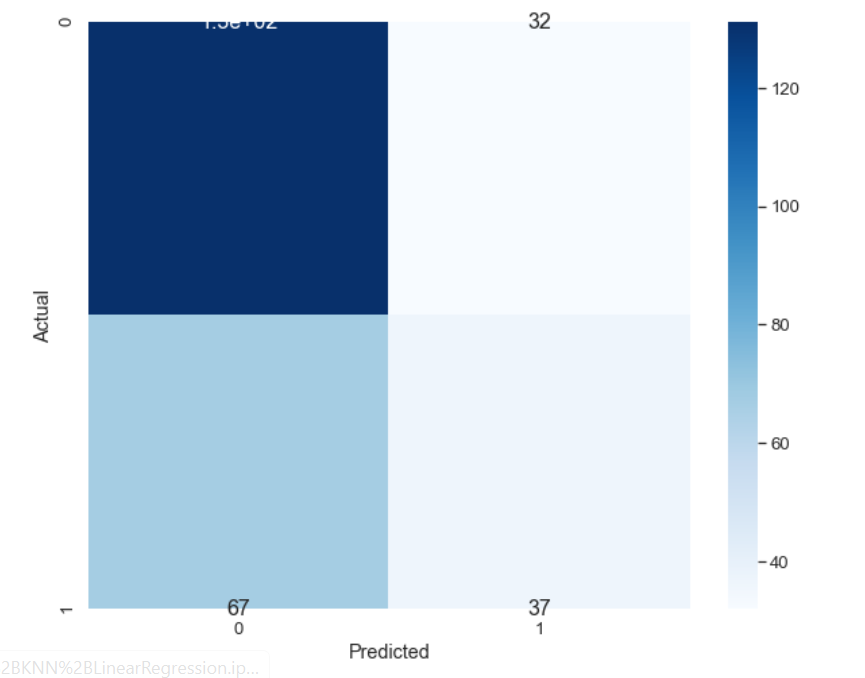
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sensitivity** | **Specificity** | **FPR** | **TP** | **TN** |
| 0.65 | 0.92 | 0.73 | 0.85 | 0.80 |

1. **K-nearest neighbour**

**Performance Evaluation:-**

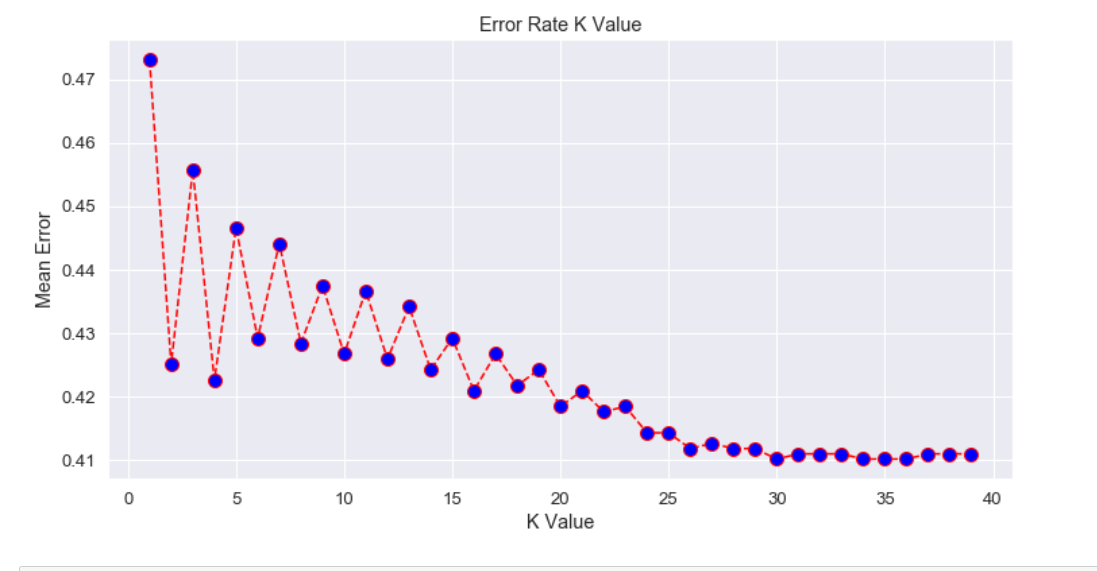
|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| 0.80 | 0.66 | 0.73 | 0.73 |

The confusion matrix is shown as



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sensitivity** | **Specificity** | **FPR** | **TP** | **TN** |
| 0.35 | 0.80 | 0.19 | 0.53 | 0.66 |

**Error / Loss Computation diagram;-**

****