**Name: Muhammad Samiullah**

**Roll no : NIM-BSCS-2020-27**

**Machine Learning**

**Assignment no 01**

# **Supervised Learning:**

In Supervised learning, we trained the model on label data where labels help the model to show correct output with given input. The Algorithm is provided with a set of labeled data where each sample data consist of an input vector and a corresponding output label. Then Algorithm use these data to learn mapping between input and output variables, which can then be used to predict the output for new unseen input.

## **Example:**

In titanic dataset, which contain information about the passengers on board the ship including survived or not. In this dataset, our goal is to build the supervised learning model which predict either the passenger survived or not.

In this case, input variables (features) include attributes like Age, gender, tickets class and Embarked port. In output variable (which also knows as target variables) which get binary variable indicating passengers are survived or not.

In this dataset, we split the dataset into trainning and testing set for evaluating some result like accuracy, precision, recall and F1 score. But first we need to select the appropriate model for getting such result. So, we used Decision tree (Which used the concept of if else condition and divide the data into some block as child of the root tree), Confusion matrix (it is a matrix of actual versus predicted class and show numbers of TP, TN, FP, FN) and Also Naïve Bayes (it is use to find the probability of each feature for given classes to predict the probability of new sample belonging to each class. It suppose that each feature are independent to each other.

## **Cleaning the Dataset.**

First we identity the missing values of each feature of dataset and do some mathematical calculations to fill these values. Like Age, Embarked in our dataset. We take median for age because it gives the value as a middle value which are not affected for dataset and help us to find outliers in data.

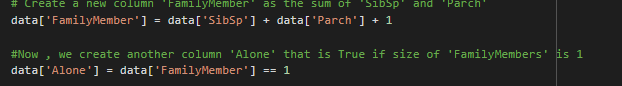
And In Embarked column, we take mode because we fill none value on that port which have largest number of passengers.



If we have some outliers then we remove from our dataset. Then we removing some unnecessary column which are not useful for Analysis and Learning. So, we remove some columns from our dataset. Like Passenger id, Name, Ticket etc.



Now, we reduces the number of column which make our dataset more precise which help the model to get maximum accuracy. In our dataset, we merge to columns like Sibsp and Parch to complete the family members because these are much related to each other. So, we combine them.



Now we divide the Embarked column into their related port, which help the supervised leaning model to learn and show good result and also help us in Analysis. If we have one column for different sites, we are not easily see the which passenger belong to which port.so, we divide it and make it more precise for Exploratory Analysis.

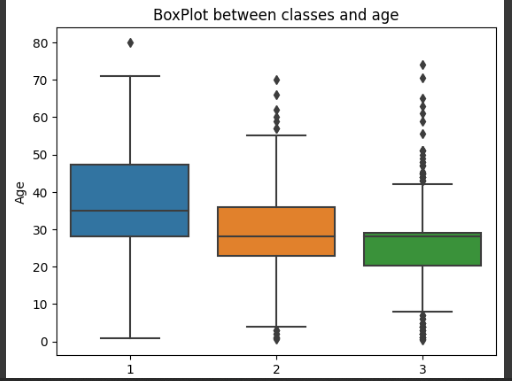
Almost our dataset is clean and ready for Exploratory Analysis and for supervised learning models.

## **Exploratory Analysis.**

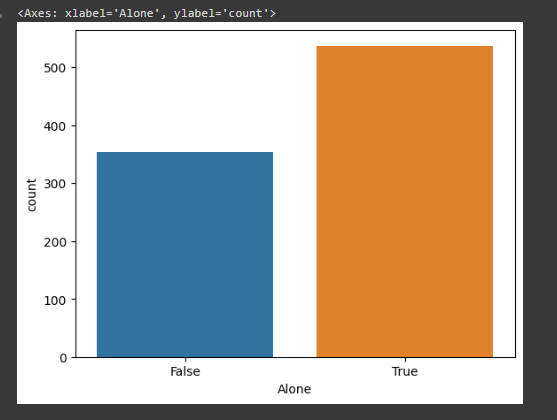
In Exploratory Analysis, we analyze and summarized the dataset to gain insight and identify pattern and their relationship between them for further analysis. It involves the visualizing and exploring the data using statistical and graphical techniques.

So, we used boxplot which help us to find the outliers easily and help us to visualize the distribution of data based upon five number like minimum, maximum, median etc.it is also useful to find the variability , spread and skewness of the dataset.

We make a plot between age and class .if passenger age is very large then it show us that it is outliers and dirty our data. Also it show us some graphical and statistical relation between them.



We also count the number of passenger which are not with their families but travelling in ship . we count and visualize such passenger with help of bar chart.



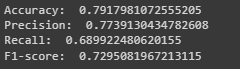
We also check that can every feature of dataset have some relation with another dataset. Then we use heat map which show relation between features. Heat map use some statistical approach to show the relation between them using corelation function. In this plot, we see that relation probability graph between attributes.

## 

## 

## **Models and their Results :**

We used different models for trainning and testing the dataset. Then we get different accuracy results, which means that different models follow different criteria. On the basic of these criteria’s, they show different result. We used only those model which are used for label dataset so, first we used naïve Bayes which assumes that input features follow the Gaussian or normal distribution. Their result is show in figure



But we also performed the 10 fold-cross validation to get the more reliable estimate of the model performance and compute the average accuracy score.



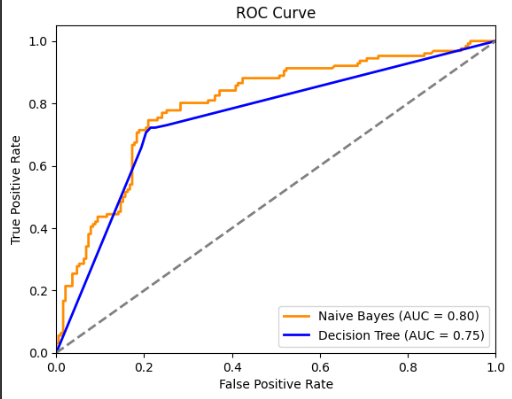
If we talk about the decision tree classifier which is based upon only if else condition on every branch of the tree which solve the problem very easily. Their result is also show in figure.



## **ROC Curve:**

ROC curve is a graphical representation of the performance pf the binary classification models. It is created by plotting the true positive rate with false positive rate at various threshold. This is useful while checking the performance of different models. If we talk about perfect classifier then it will gain the AUC of 1 and a random classifier will gain AUC of 0.5.

So, we draw the curve between naïve Bayes and decision tree classifier. The graphical representation of curve is given below.



In this figure, we see that Naïve Bayes have good result of 0.80 while decision tree will have 0.75 which is not bad for this model.