Titanic Survival Prediction

ABSTRACT

*Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg. So more than half people died in this disaster and survival rate is low so in this project we are going to find which factors affects the most on the survival of people by using various Machine Learning algorithms and various techniques of data handling. In this project we are predicting that the person was survived or not depending on his/her age, gender, he/she paid for the ticket etc.*

INTRODUCTION

*Titanic Dataset is one of the most popular datasets used for understanding machine learning basics. It contains information of all the passengers aboard the RMS Titanic, which unfortunately was shipwrecked. This dataset can be used to predict whether a given passenger survived or not. To predict the survival of a particular person based on it’s other features like age, gender, class from which he is travelling and so on we have used various machine learning algorithms like Logistic Regression, K Nearest Neighbours, Decision Tree, Random Forest, Support Vector Machine and Naïve Bayes algorithm. Our dataset is unbalanced dataset and contains null values so before passing this data to our machine learning models we have to clean this dataset and make it balanced dataset. For balancing purpose we have used Over sampling mechanism and made a dataset balanced in terms of survival rate cause the number of not survived people is more than the survived people. And for clean the data set we used to techniques that is where the null values greater than 5% in a particular column we have dropped that column and where the null values are less than 5% we have replace it with mean of that particular column.*

PROBLEM STATEMENT

*Depending on the factors like age, gender, class from which he/she is travelling, is he alone or with his sibling and is he with his parents or not, we will find that the person was survived or not in that disaster.*

DATASET DESCRIPTION

*We have collected the dataset from the Kaggle from Chrome extension for our project. The name of dataset is ‘Titanic – Machine Learning from Disaster’.*

*Our dataset contains 12 columns and 891 rows. The columns are PassengerId, Survived Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked. ‘Survived’ column contains 0 and 1 value, where 0 means person is died and 1 means person is survived. ‘Pclass’ means the class in which passenger was travelling which are denoted by 1, 2 and 3. ‘Sex’ column contains two values Male and Female. ‘Age’ column gives the age of every person. ‘SibSp’ gives that the person’s siblings / spouses aboard the Titanic. ‘Parch’ gives that the person’s parents / children aboard the Titanic. ‘Ticket’ column gives the ticket number. ‘Fare’ gives information about Passenger fare. ‘Cabin’ gives cabin number and ‘Embarked’ gives the station from which the person is boarded on Titanic which contains three station names S, C, Q for Southampton, Cherbourg, Queenstown respectively.*

METHODOLOGY

*As we know our dataset contains null values so we replace it with the mean of respective columns.*

*Our dataset is unbalanced dataset so to make a good prediction and to increase to accuracy we have to make it balanced.*

* Data balancing technique –

*In our project to make a data balanced we used over sampling. Over sampling is used when the amount of data collected is insufficient. A popular over sampling technique is SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples by randomly sampling the characteristics from occurrences in the minority class.*

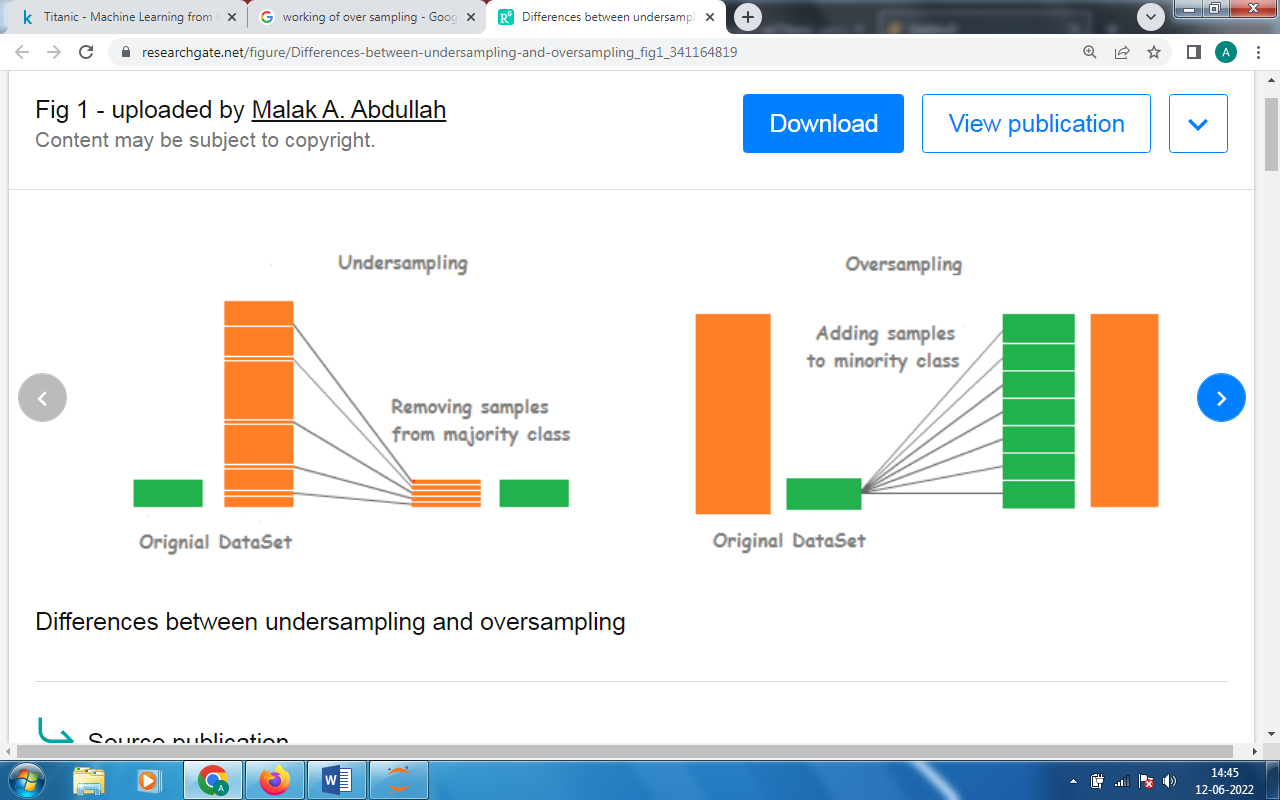


Fig. 1 – Working of Oversampling

* Need of Sampling –

*If we trained our model on imbalanced dataset then the accuracy might be great but it is not correct cause most of the time it predicted only the majority group which is mostly true but when it predict the minority group it may be wrong and that’s why we have to make data balance in order to overcome this problem of wrong prediction.*

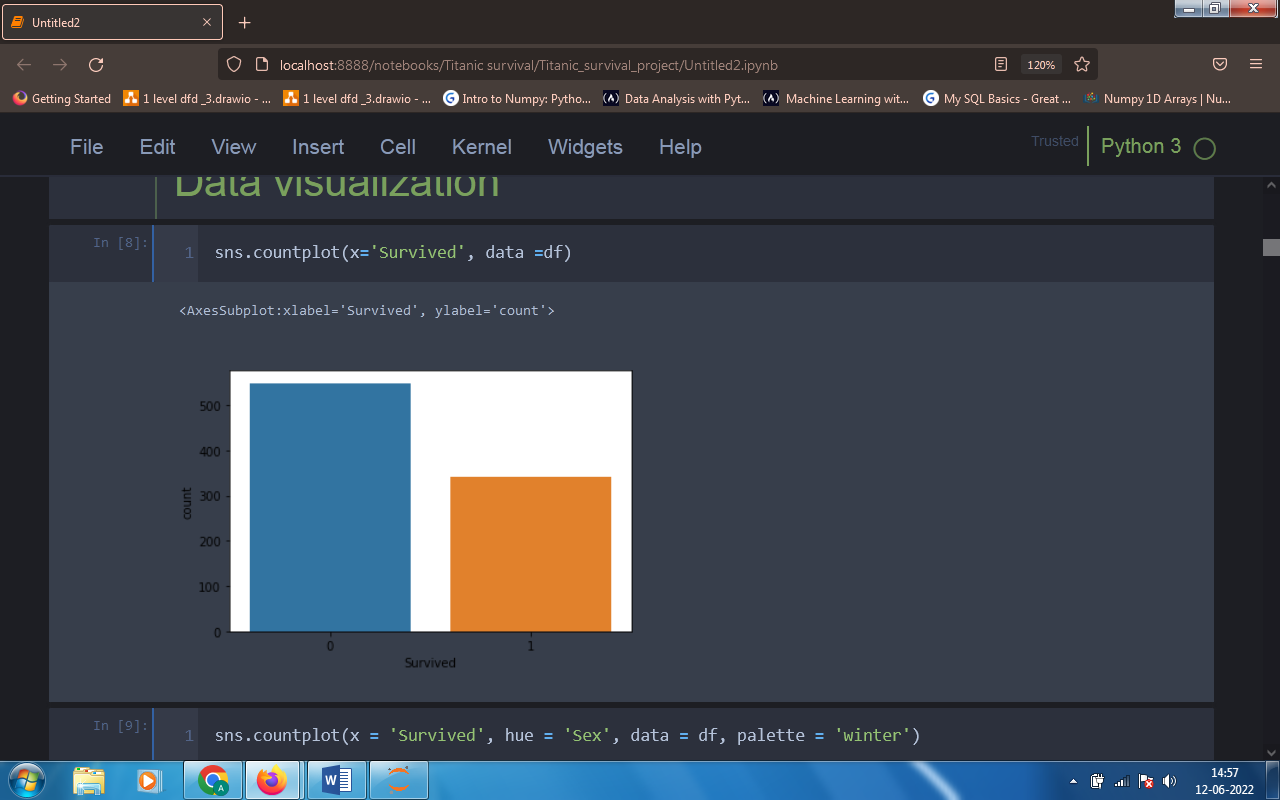


Fig. 2 – Before balancing dataset (original dataset of survived)

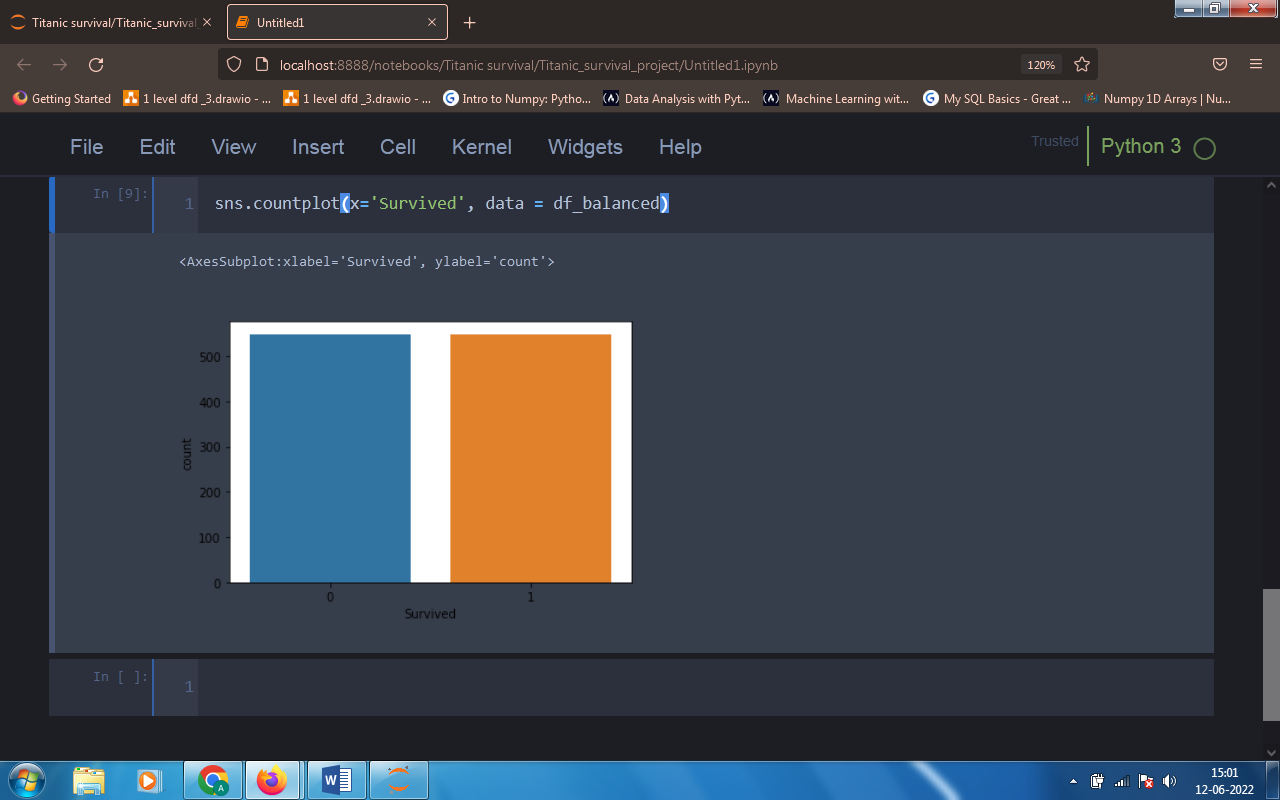


Fig. 3 – After applying Over sampling.

*In our dataset column contains entries in string format that is in object data type but machines only understand the numbers so before passing this data to our machine learning models we have to convert this strings to number and this method is called as Label encoding. In our project we used One Hot Encoding to convert strings into number.*

* One Hot Encoding –

*One hot encoding can be defined as the essential process of converting the categorical data variables to be provided to machine and deep learning algorithms which in turn improve predictions as well as classification accuracy of a model.*



Fig. 4 – Working of One Hot Encoding

* Advantages of Encoding –

*One hot encoding makes our training data more useful and expressive, and it can be rescaled easily. By using numeric values, we more easily determine a probability for our values. In particular, one hot encoding is used for our output values, since it provides more nuanced predictions than single labels.*

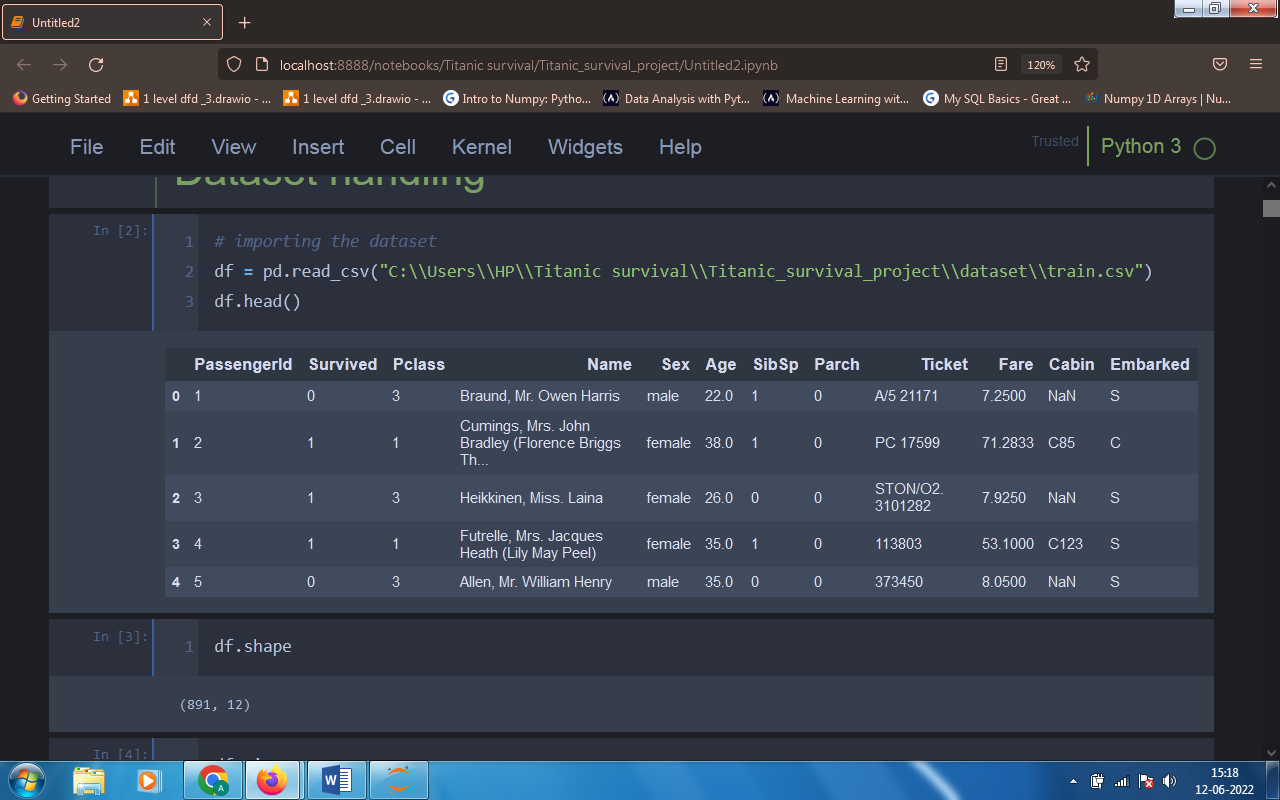
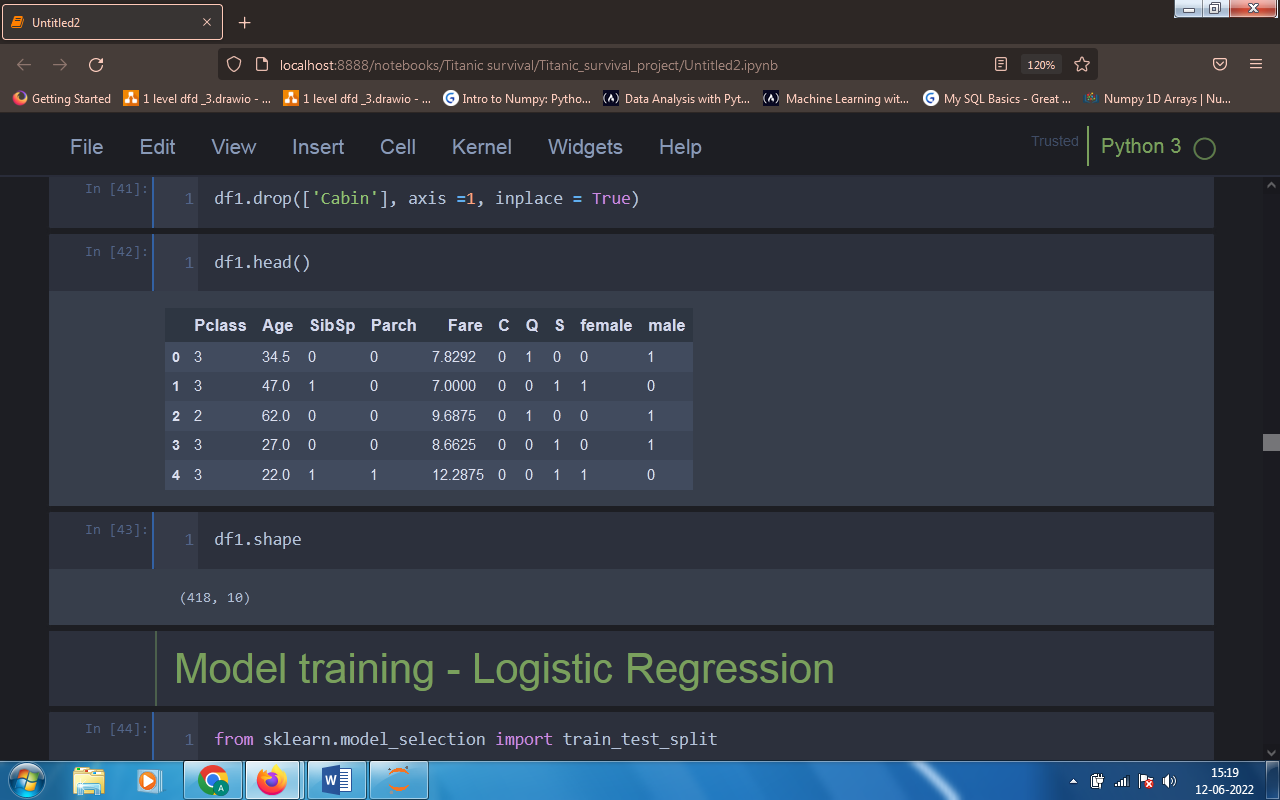
 

Fig. 5 – One Hot Encoding on Embarked column

TECHNIQUES

*In this project we used various machine learning algorithms for prediction purpose.*

* Logistic Regression –

*Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.*

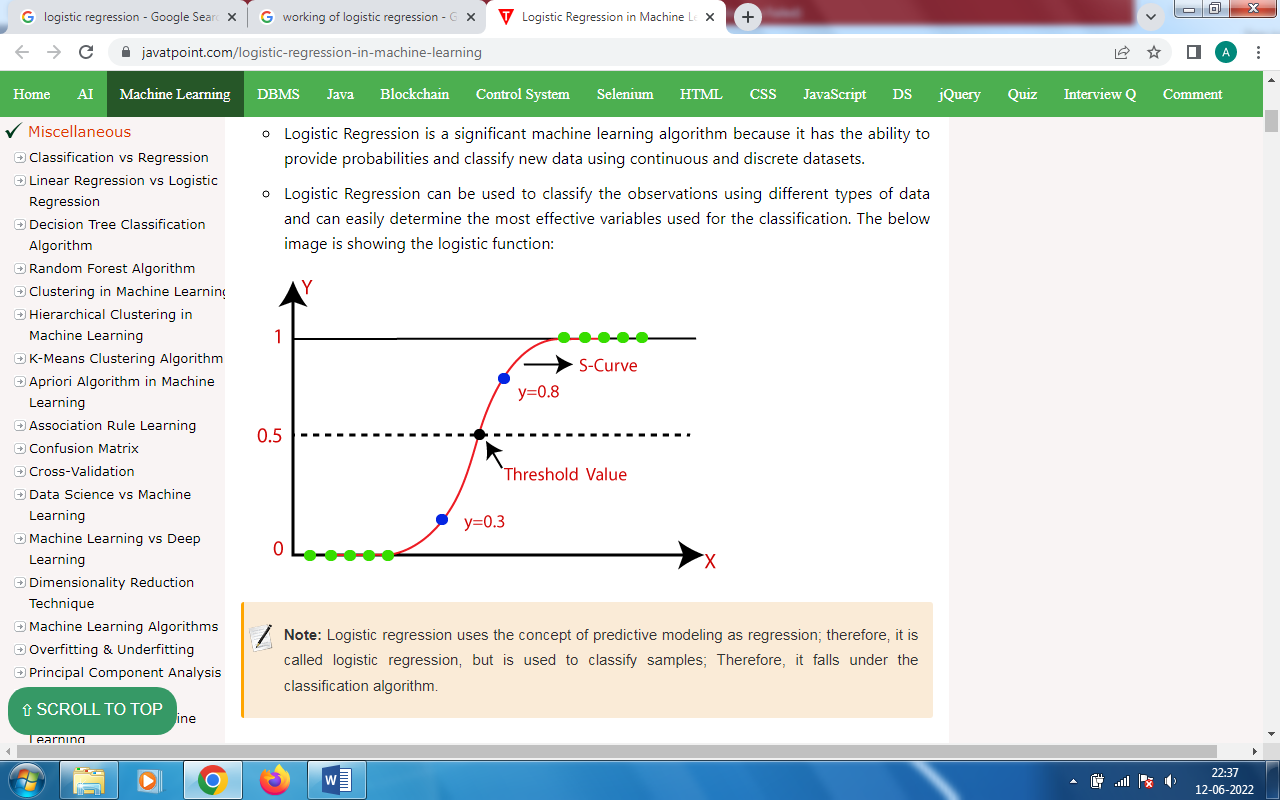


Fig. 6 – Logistic function

* K Nearest Neighbours –

*The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. KNN algorithms decide a number k which is the nearest Neighbor to that data point that is to be classified.*

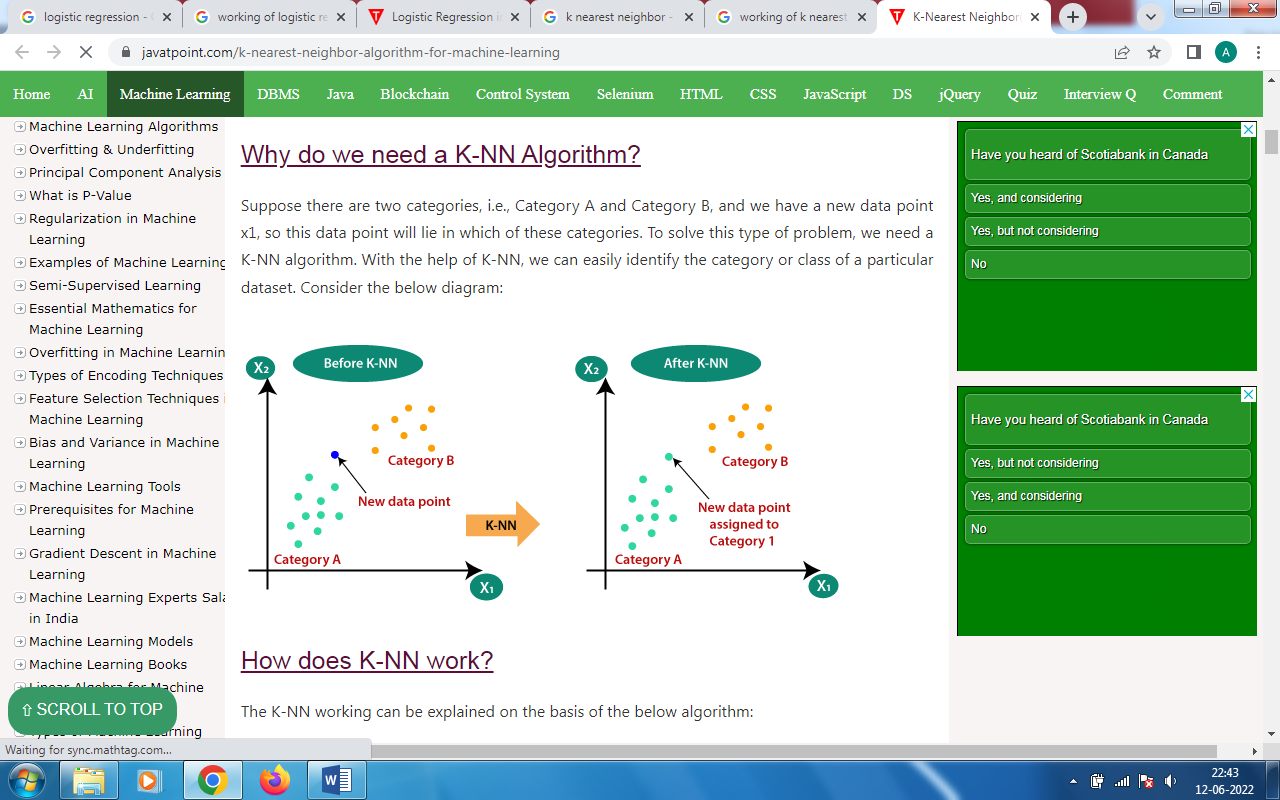


Fig. 7 – Working of KNN

* Decision Tree –

*Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.*

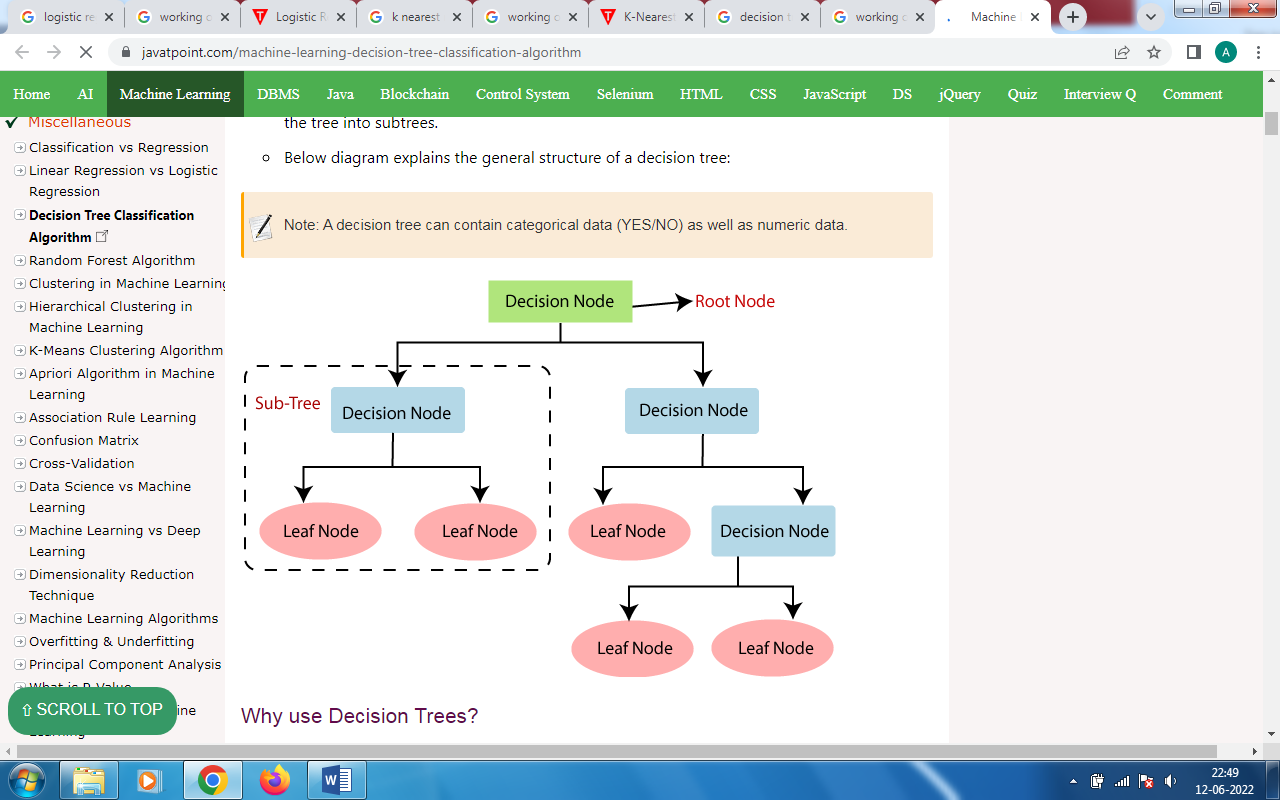


Fig. 8 – General structure of DT

* Random Forest –

*Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.*

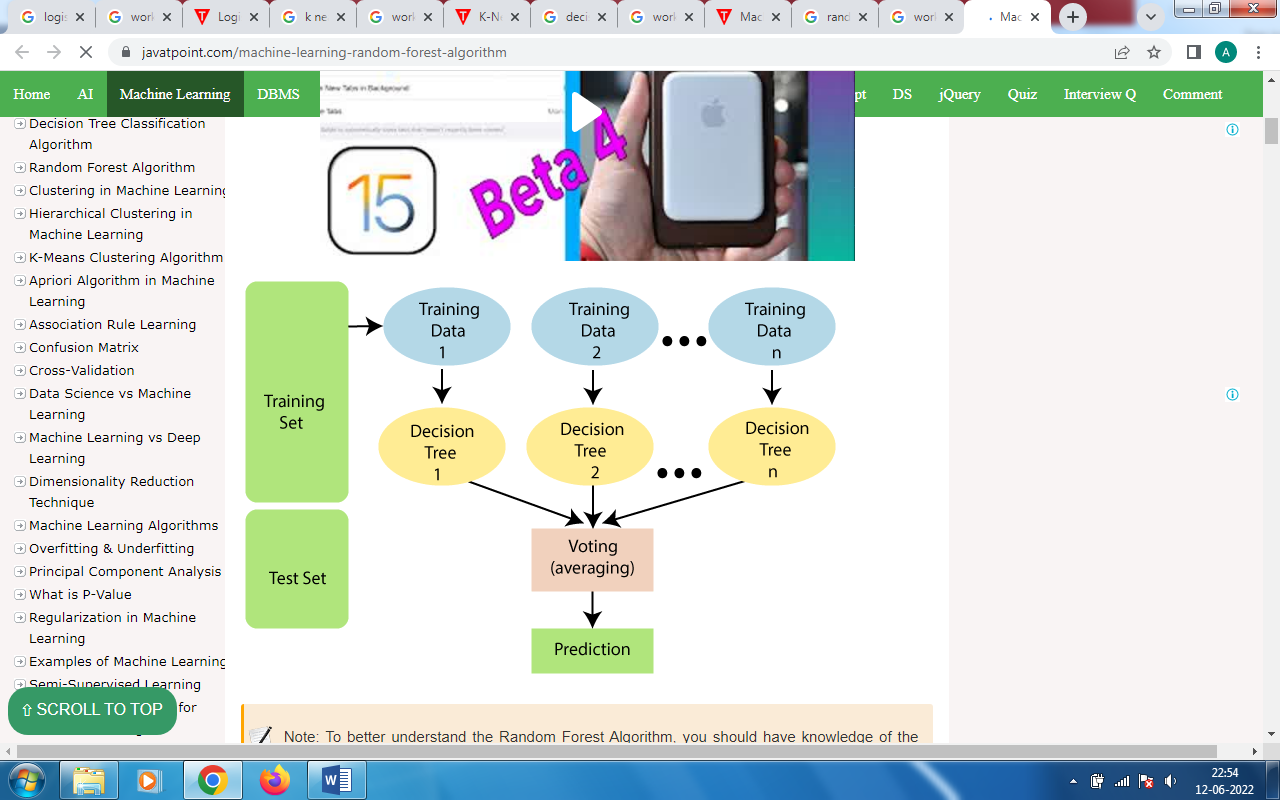


Fig. 9 – General structure of Random Forest

* Support Vector Machine –

*The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.*

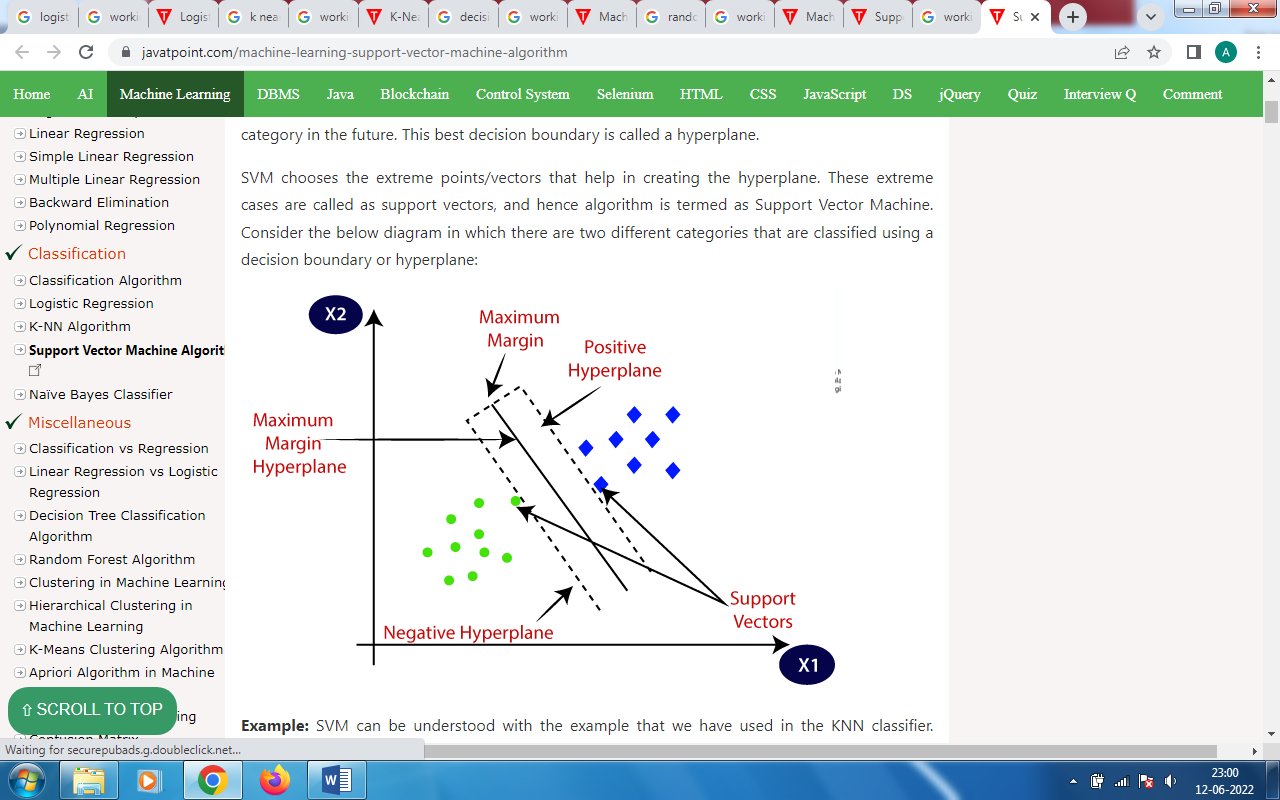


Fig. 10 – SVM algorithm

* Naïve Bayes –

*The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.*

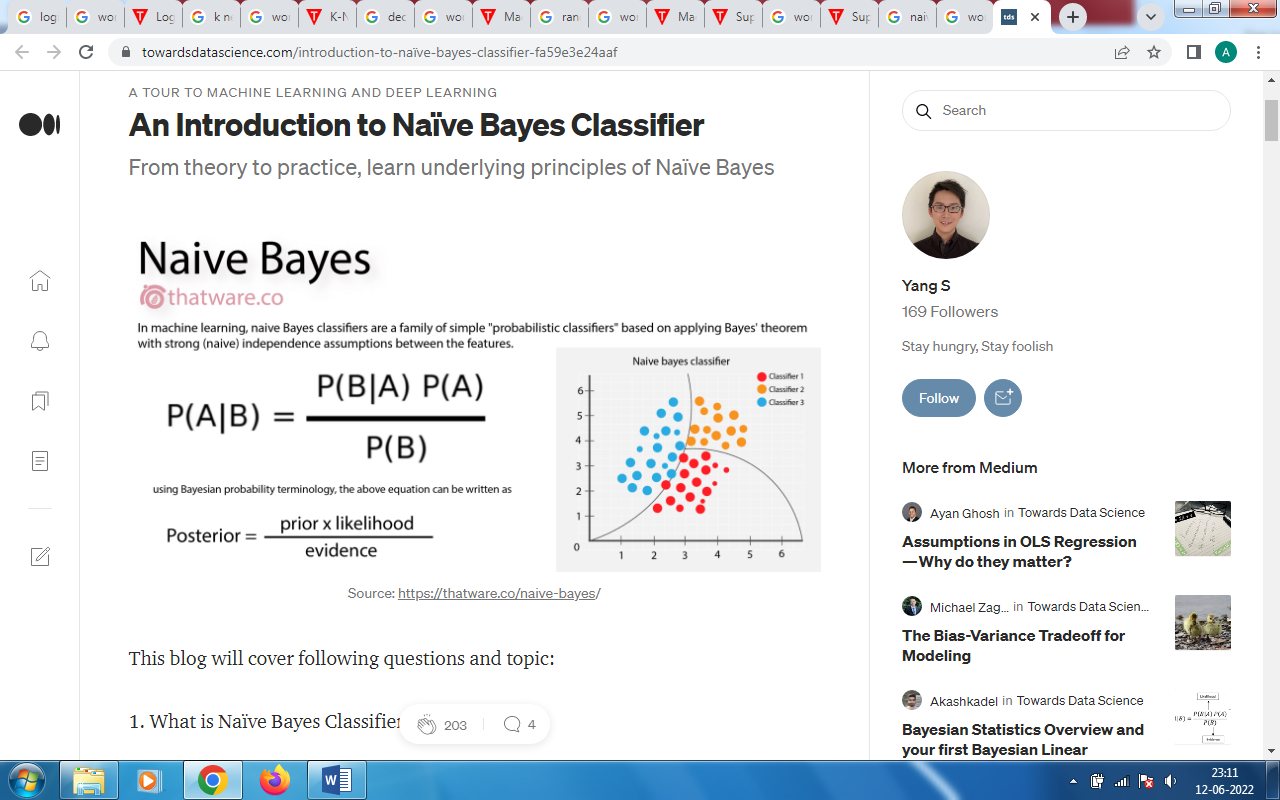


Fig. 11 – Naïve Bayes classifier

SAMPLING

*Our original dataset have 891 rows, after balancing the dataset we get 1098 rows so before giving this dataset to our models we have split it into training and testing dataset. Training is to train our model on, and testing is for checking the how accurate our model is trained. We used 80:20 formula for splitting the dataset i.e. 80% data for training purpose and remaining 20% data is for testing purpose.*

FLOWCHART

**Data Collection:-**

***Data is collected from the Kaggle.***

**Data Visualization:-**

***Visualize the data to get more idea about data.***

**Data Encoding:-**

***Used One Hot Encoding for get better results.***

**Data Balancing:-**

***Used Over sampling for balance the imbalanced data according to survival rate.***

**Data Wrangling:-**

***Did the preproccessing on data.***

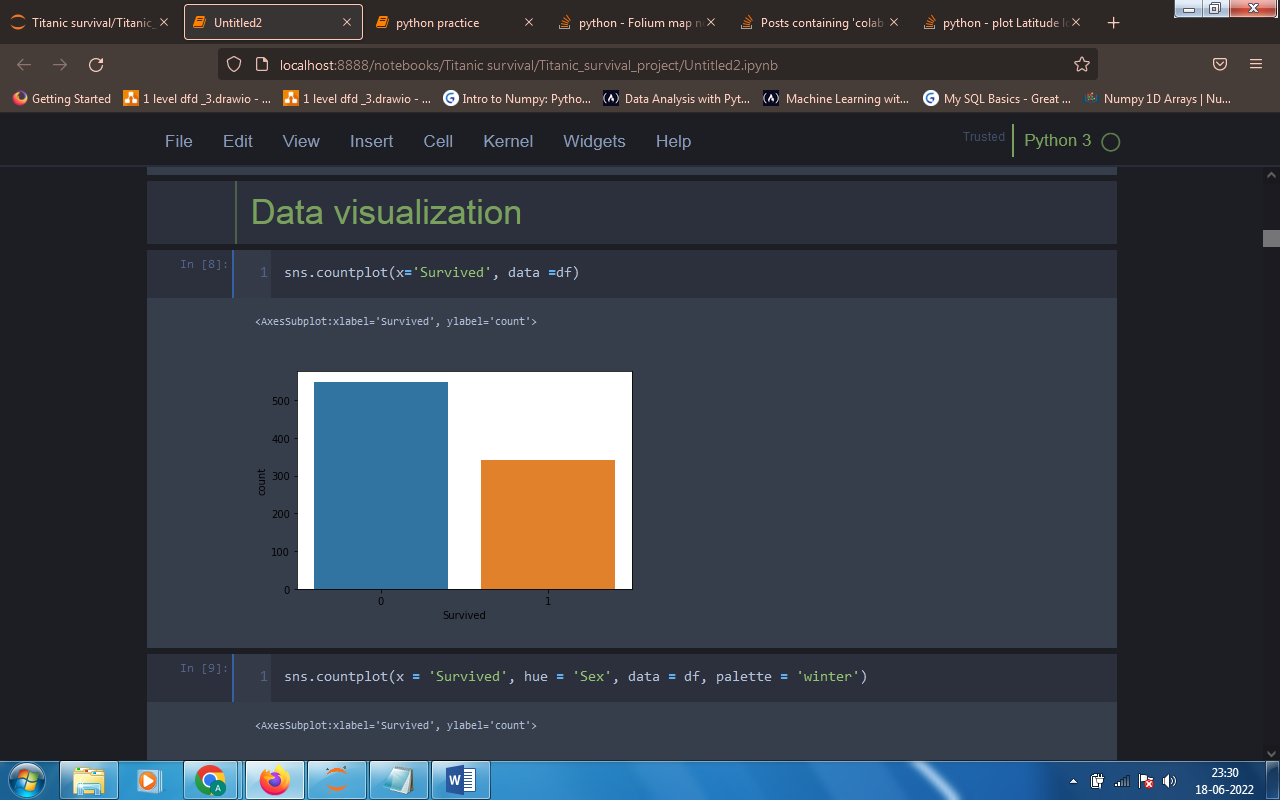
**Model Fitting:-**

***Various Machine Learning models are fitted on data and checked the accuracy of each model.***

EXPLORATORY DATA ANALYSIS

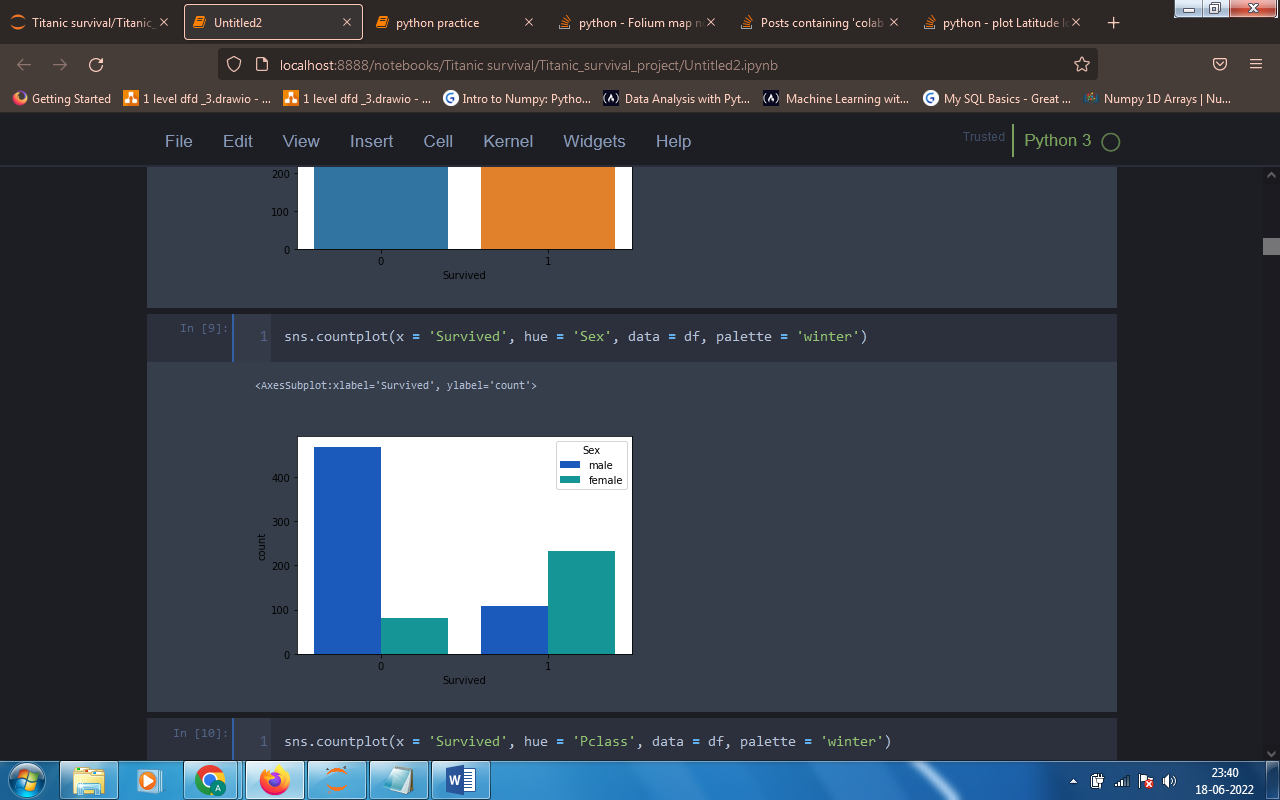
*It always good for use visualize the data rather than using numbers for comparison. By visualizing the data we get more clear idea about the data. So data visualization we used Matplotlib and Seaborn Libraries in Python.*

*1. Countplot for checking survival number.*



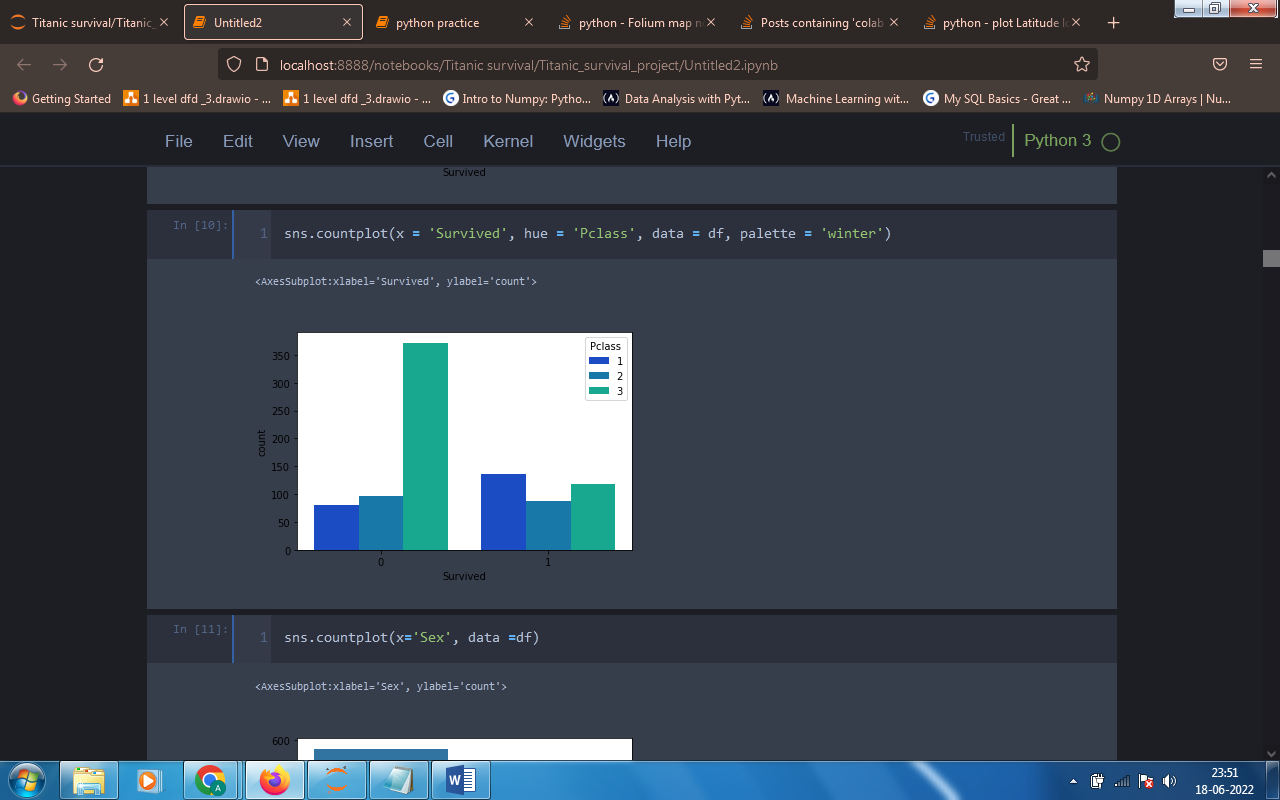
*Here ‘0’ means not-survived and ‘1’ means survived. So from above plot we can easily conclude that the number of survived people are much smaller than not-survived people.*

*2. Counplot to check comparative study of gender and survival rate.*



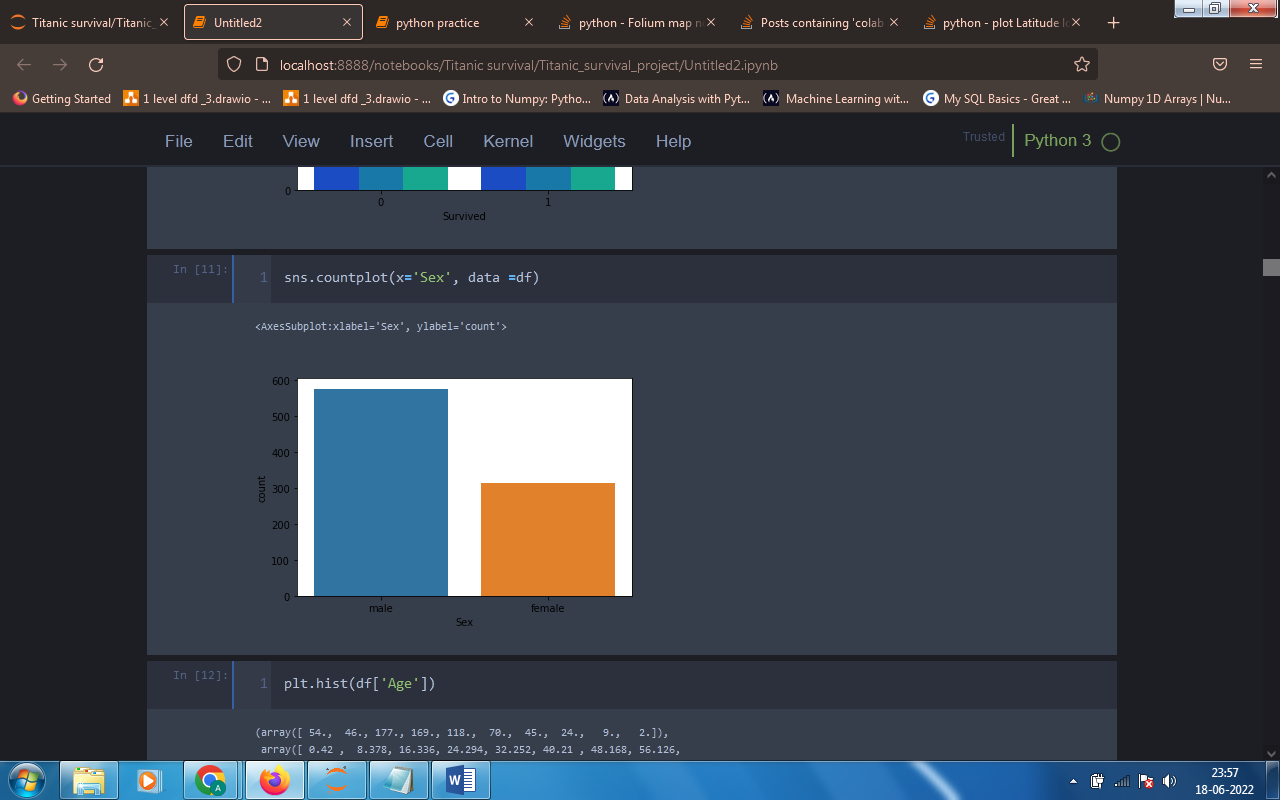
*In this fig. we can see that there are less female compared to male who died. And more females are survived than male.*

*3. Countplot for comparing class wise survival rate.*



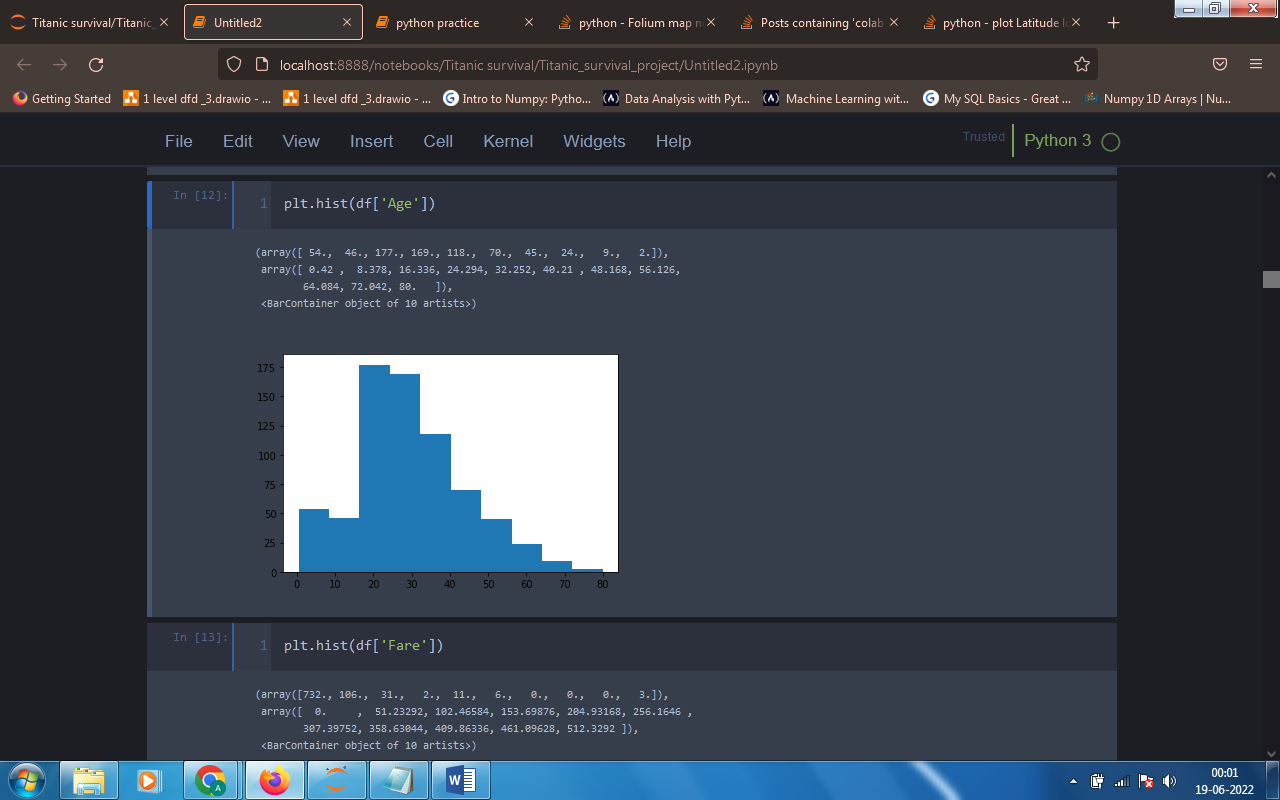
*From the above figure it is see that people belonging to first class are get rescued more than the people who belongs to second and third class.*

*4. Checking proportion of male and female.*



*Number of females are less than the number of male in the Titanic.*

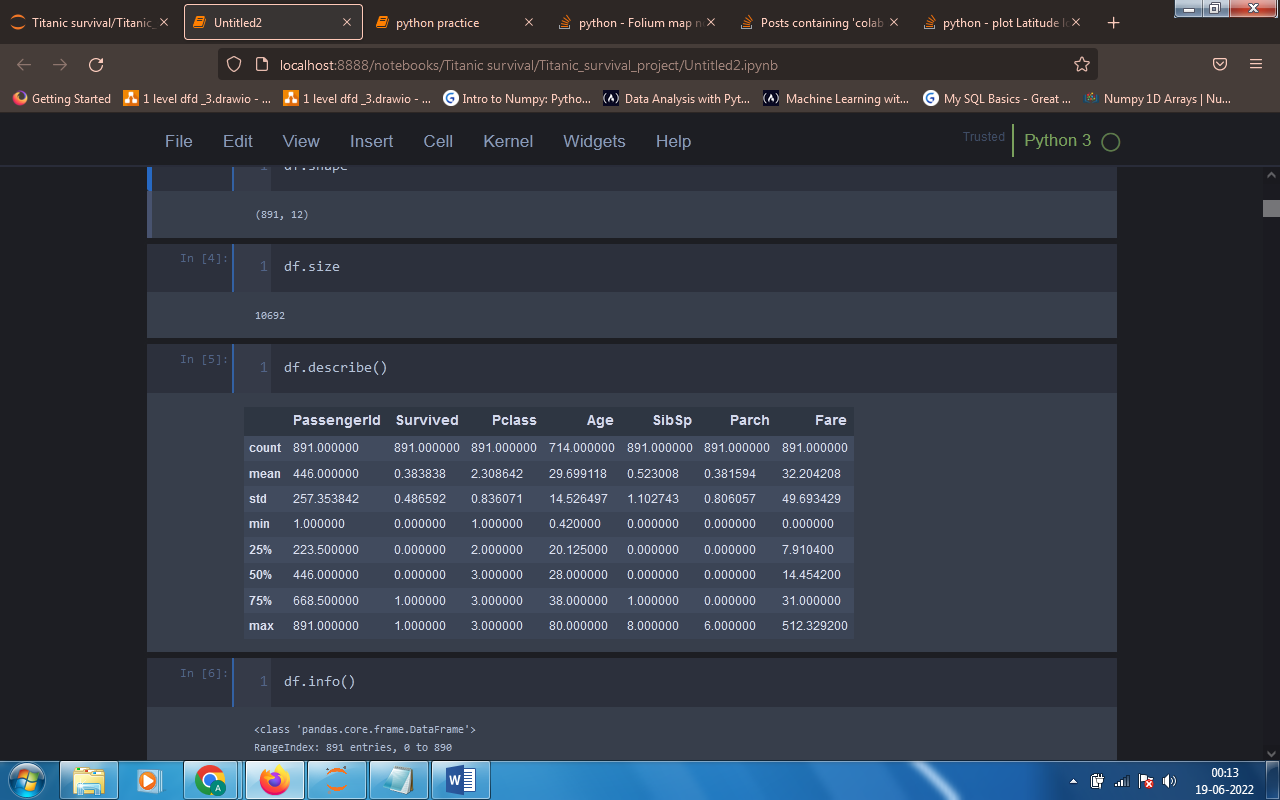
*5. Histogram for checking the age wise distribution on the Titanic.*



*People belonging to age group 17 to 40 are more in the Titanic ship and above age 65 are very less.*

SUMMERY STATISTICS

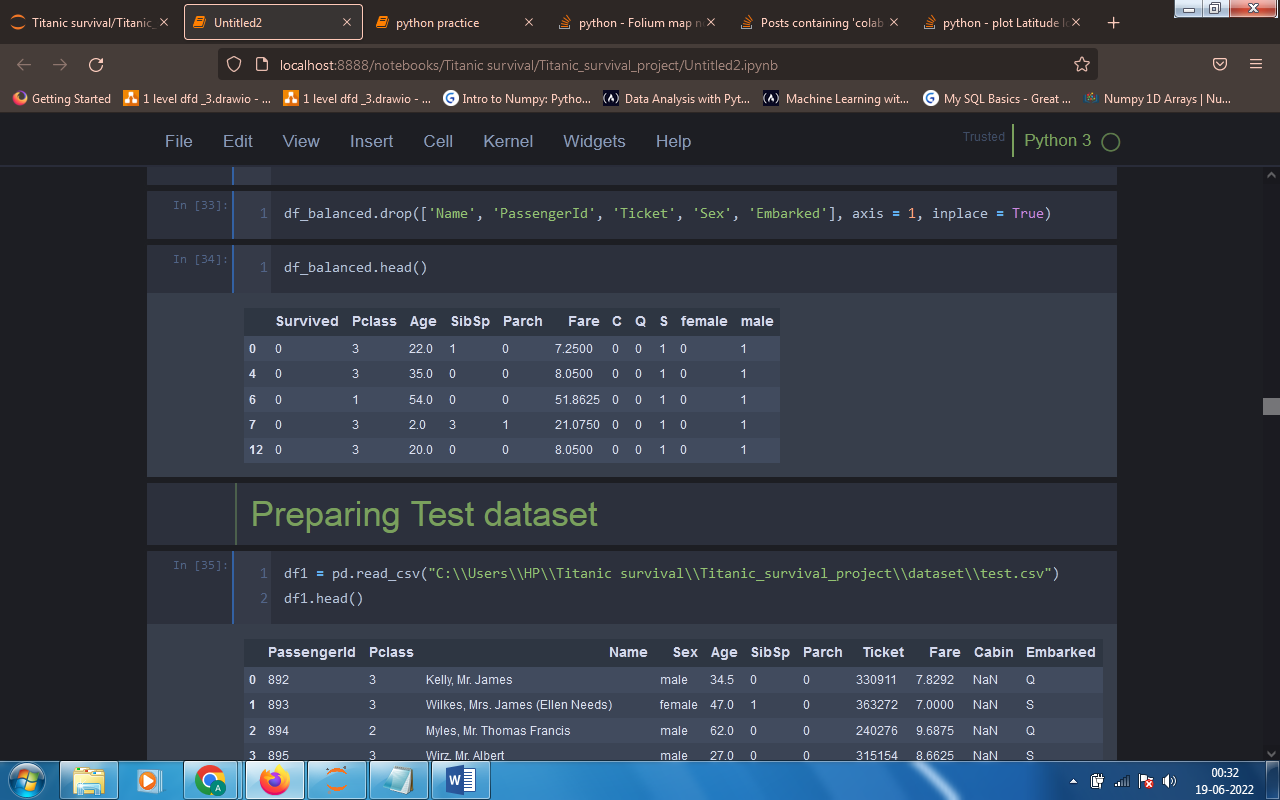
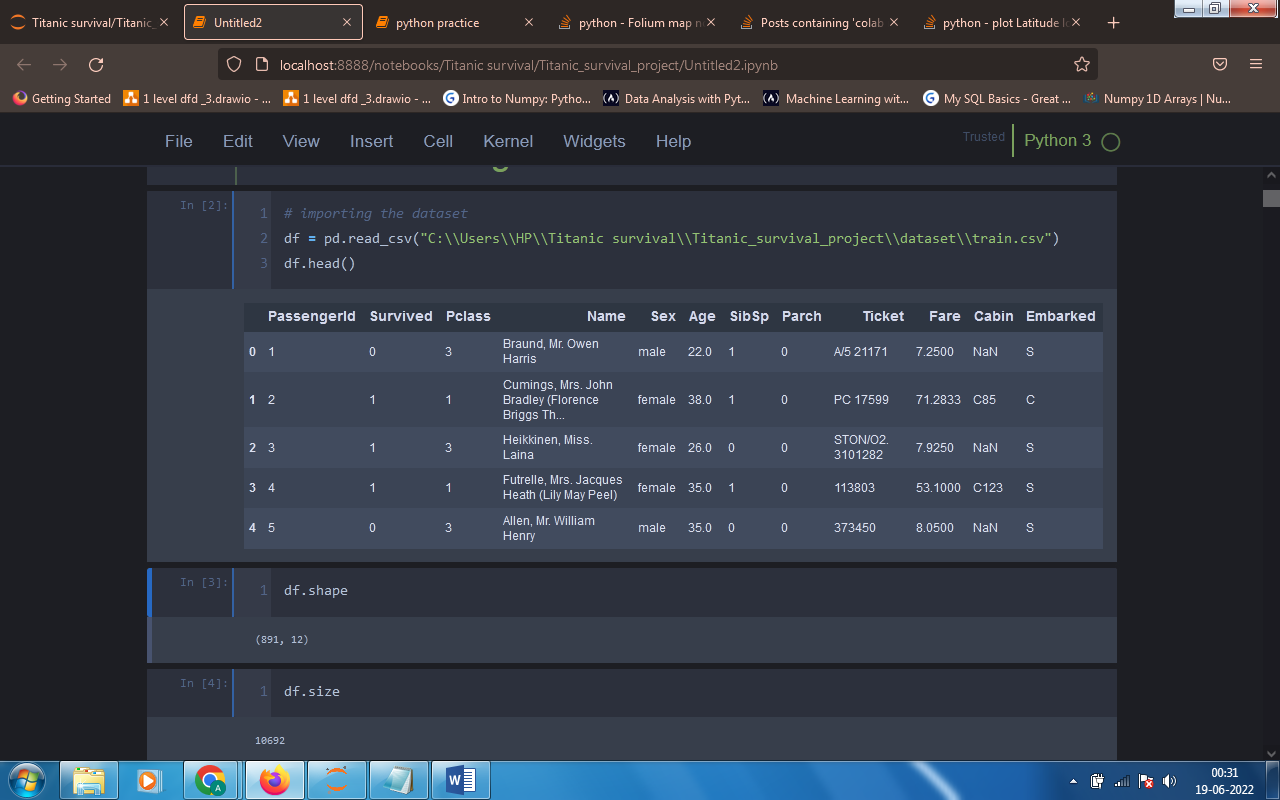
*Summery statistics give the very clear idea about the data of numeric class. Summery statistics gives us the count, mean, standard deviation, minimum, first quartile, second quartile (median), third quartile and maximum value of each numeric data column.*



*To get this summery we have a describe function in python.*

LABEL ENCODING

*We used One Hot Encoding to convert the values of ‘Embarked’ columns into number. The ‘Embarked’ column contains ‘C’, ‘Q’ and ‘S’ so we in one hot encoding the function creates column of each unique value and put 1 if it is present and put the zeros in other columns.*



MACHINE LEARNING ALGORITHMS

*Classification report contains various terms that is measures of accuracy like precision, recall f1-score support, total accuracy.*

*The Accuracy is calculated by how many predictions are right divided by total number of predictions. Precision is the prediction of positive class and it is calculated by TP/ (TP + FP) this is the formula for calculating precision. Recall is used when we want out of all positive value how many we got right and it is calculated by formula TP / (TP + FN). Precision and Recall are calculated for individual classes. F1-score is simply a harmonic mean of precision and recall.*

*Confusion Matrix gives the number of correct and wrong prediction it gives that True Positive, False Positive, False Negative and True Negative.*

*True Positive (TP) – Positive class that is predicted correctly.*

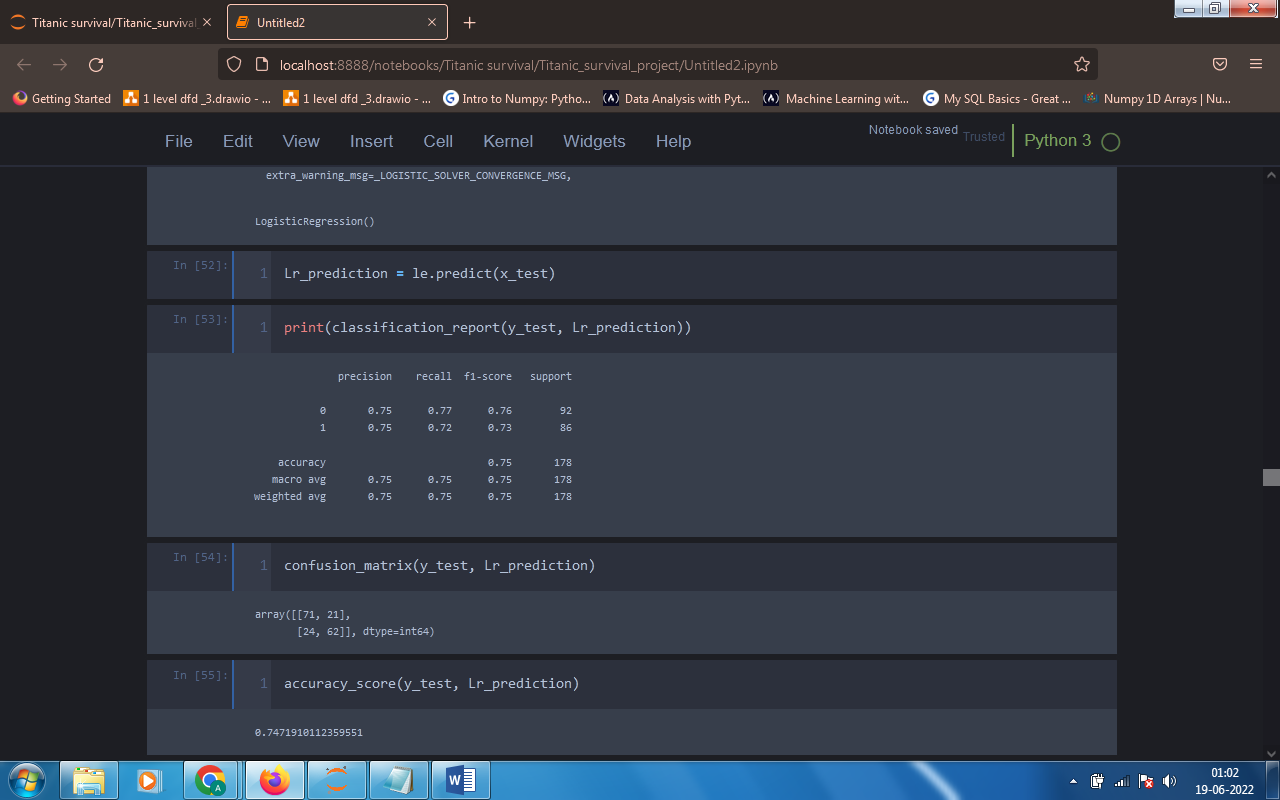
*False Positive (FP) – Predicted positive but it is wrong.*

*False Negative (FN) – Predicted negative but it is positive.*

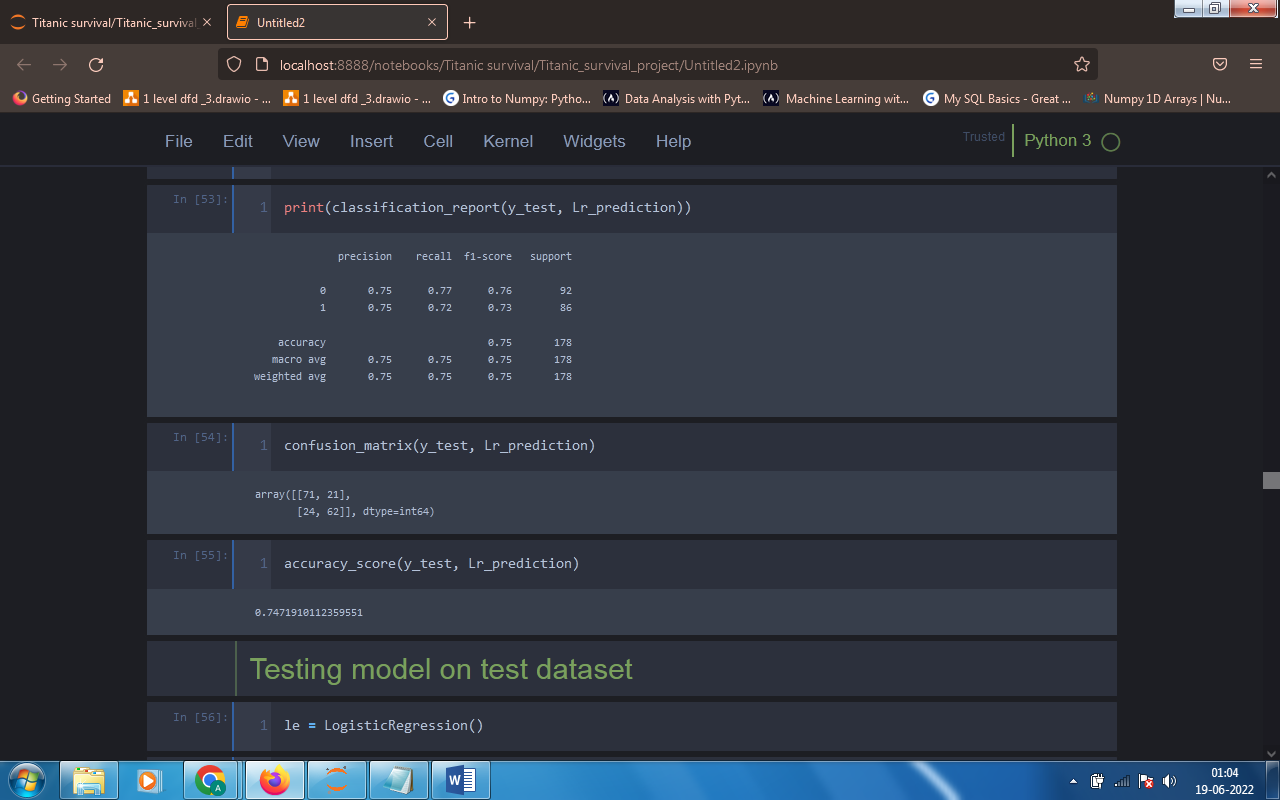
*True Negative (TN) – Predicted negative and it is correctly.*

*1.* *Logistic Regression –*

*a. Classification Report :*

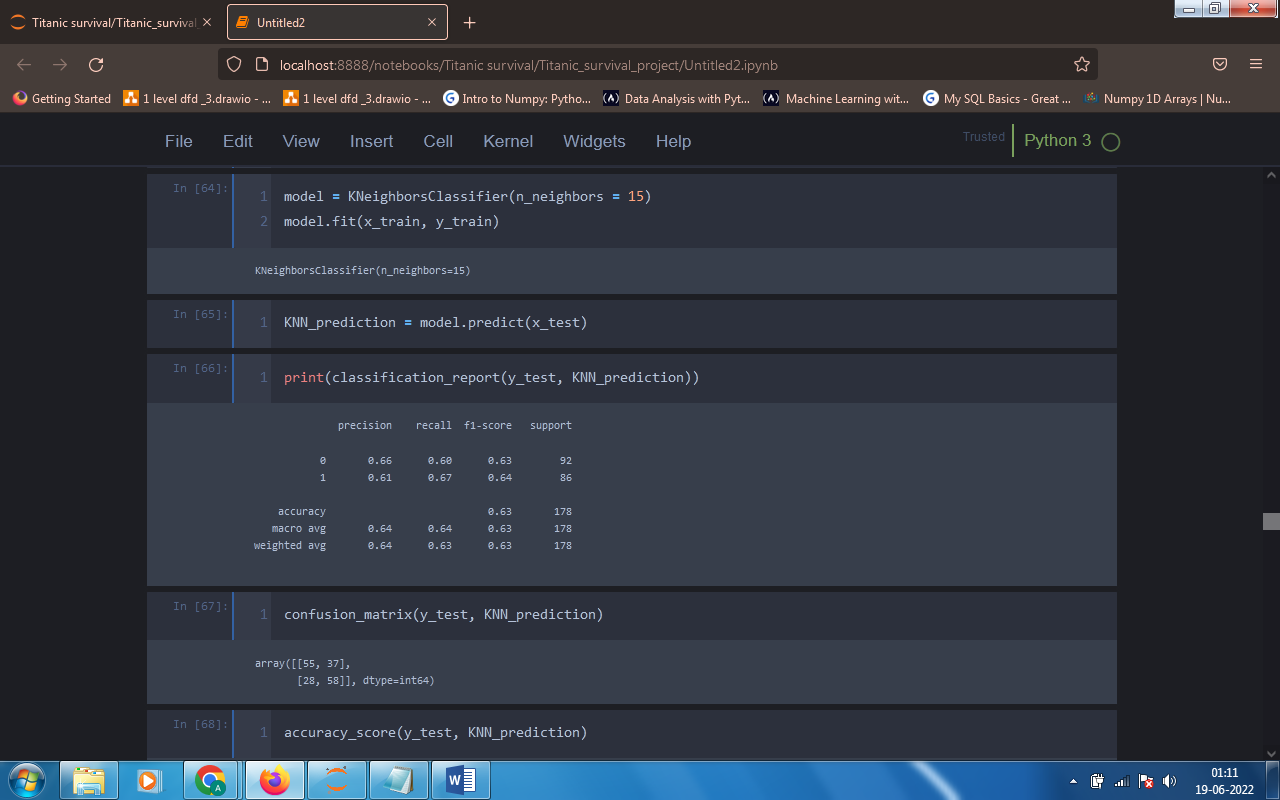


*b. Confusion Matrix :*

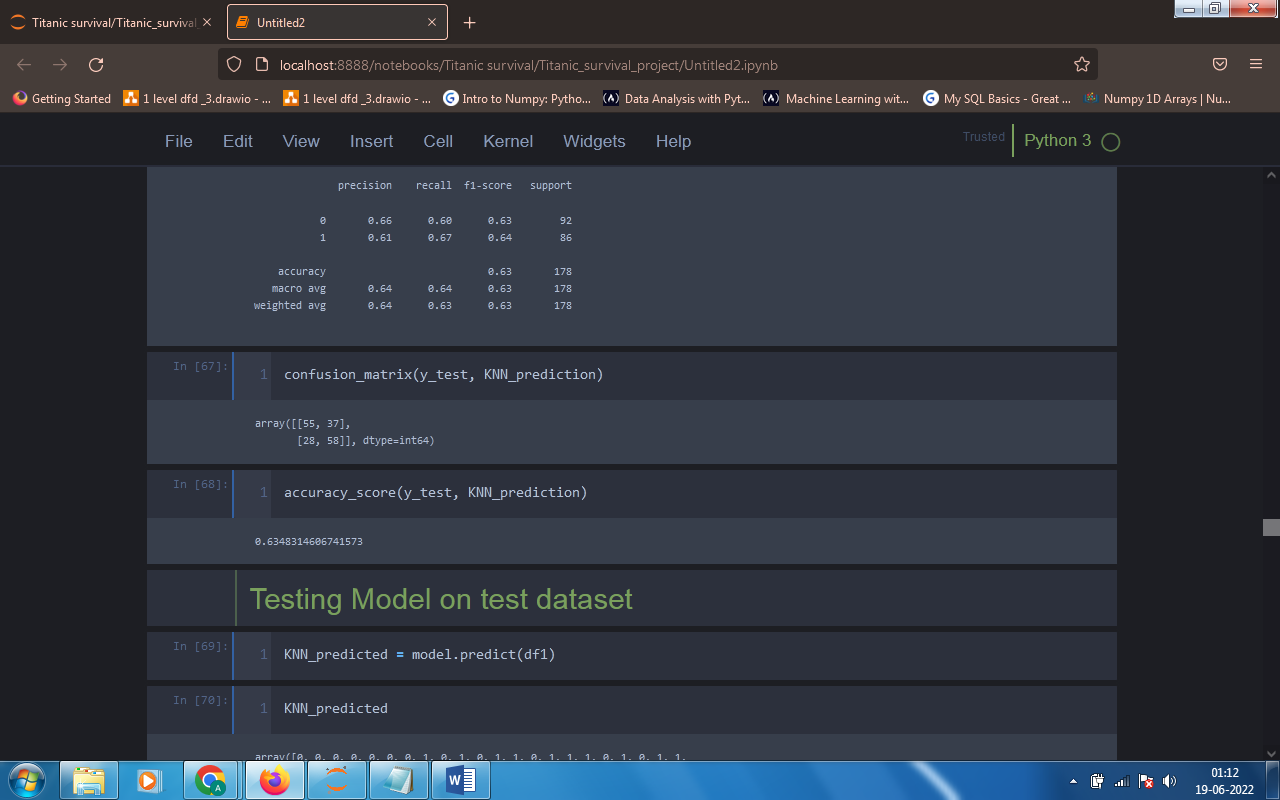


*2.* *K Nearest Neighbours –*

*a. Classification Report :*

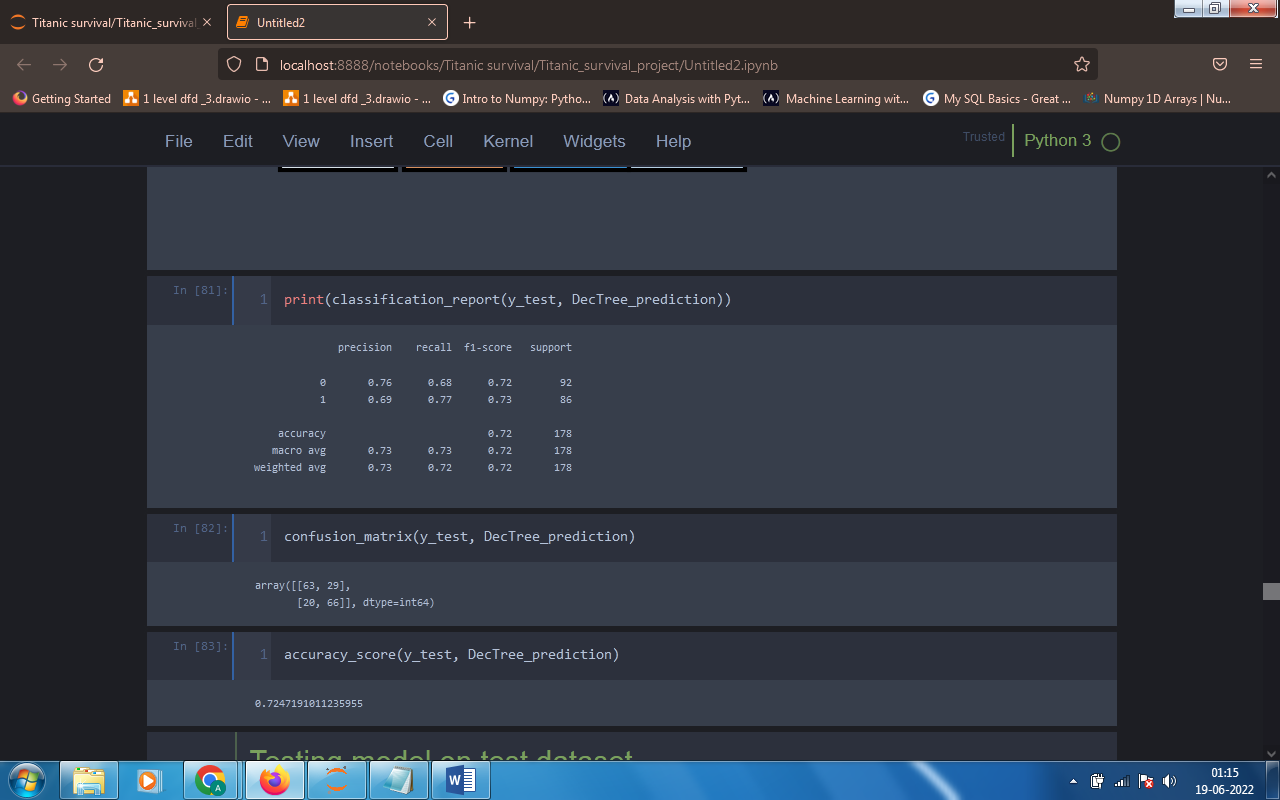


*b. Confusion Matrix :*

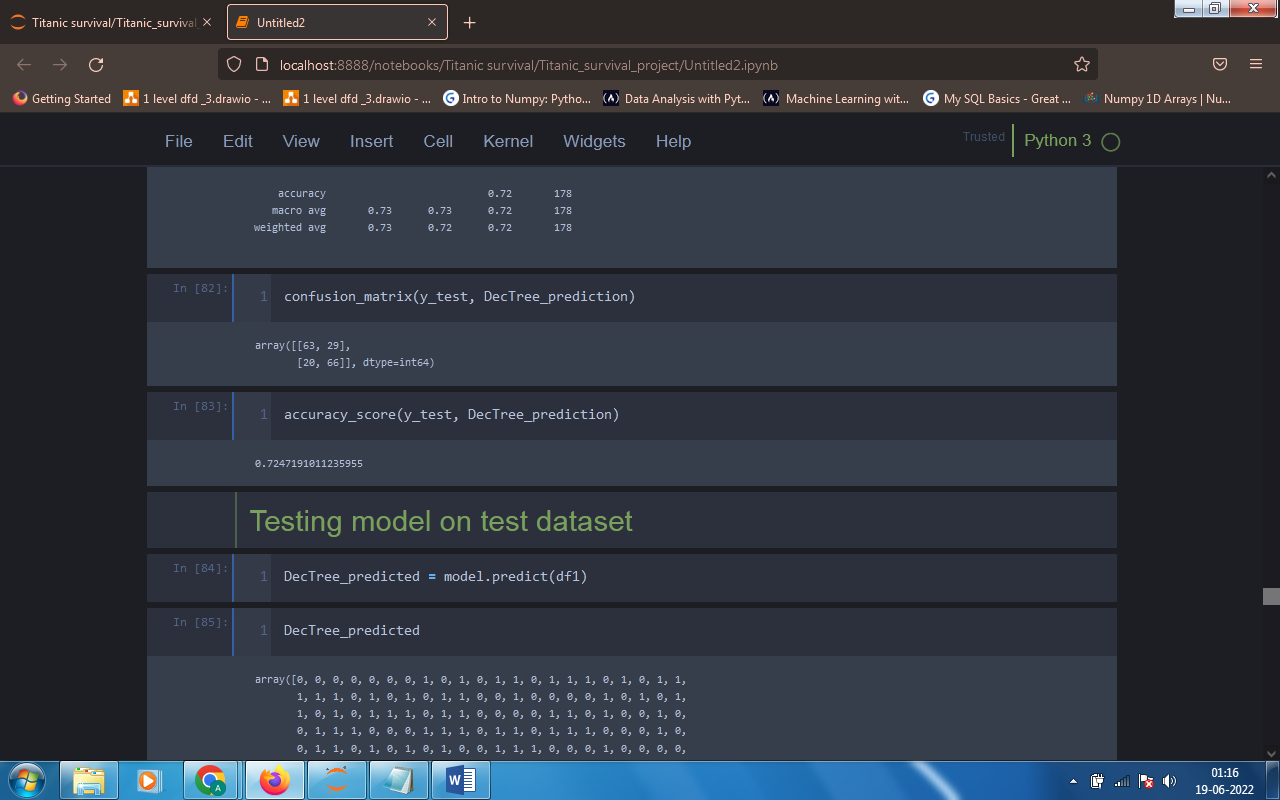


*3.* *Decision Tree –*

*a. Classification Report :*

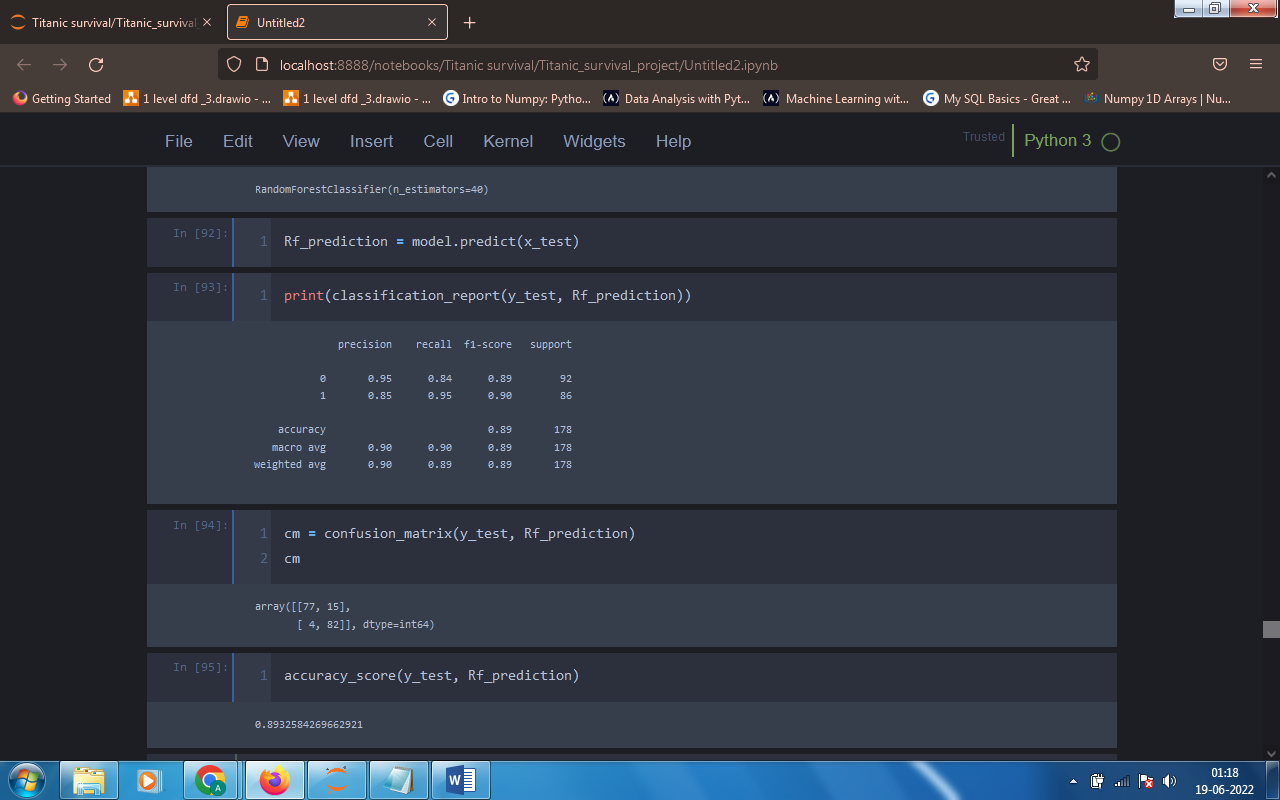


*b. Confusion Matrix :*

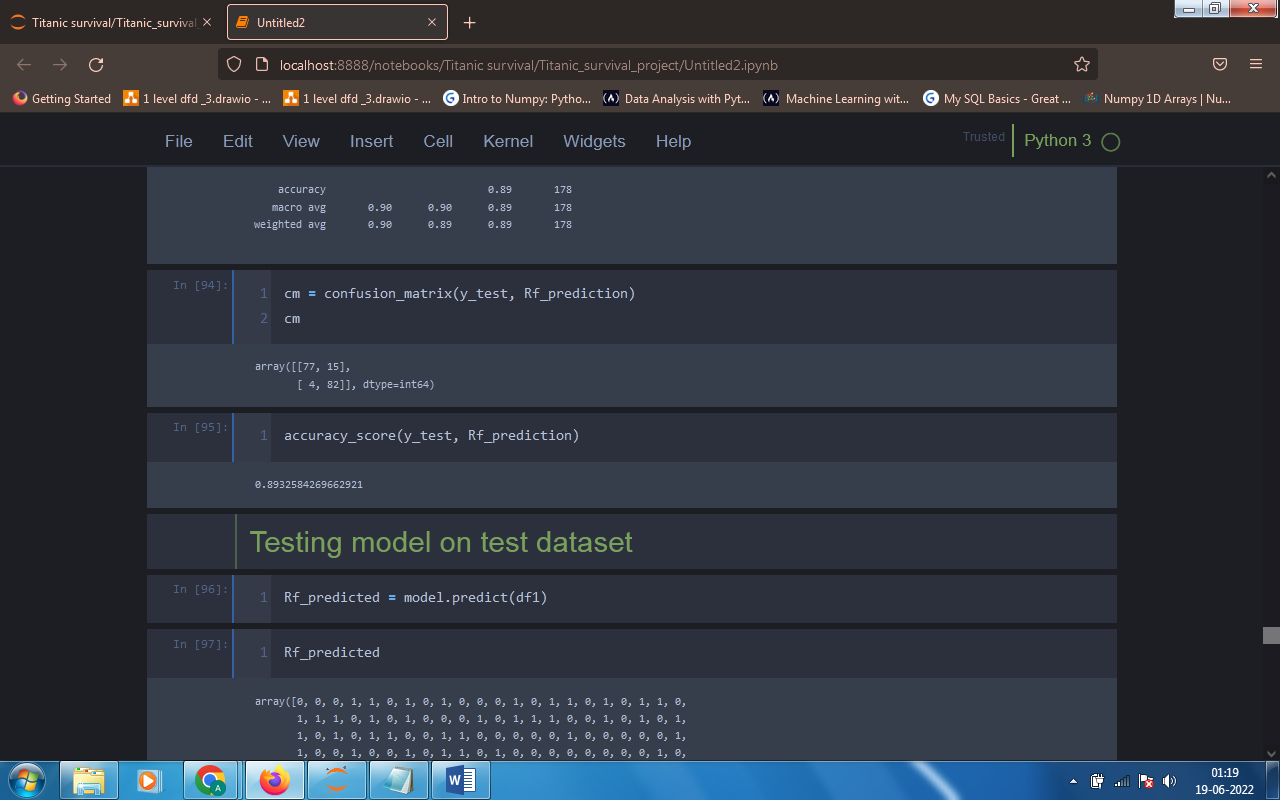


*4.* *Random Forest –*

*a. Classification Report :*

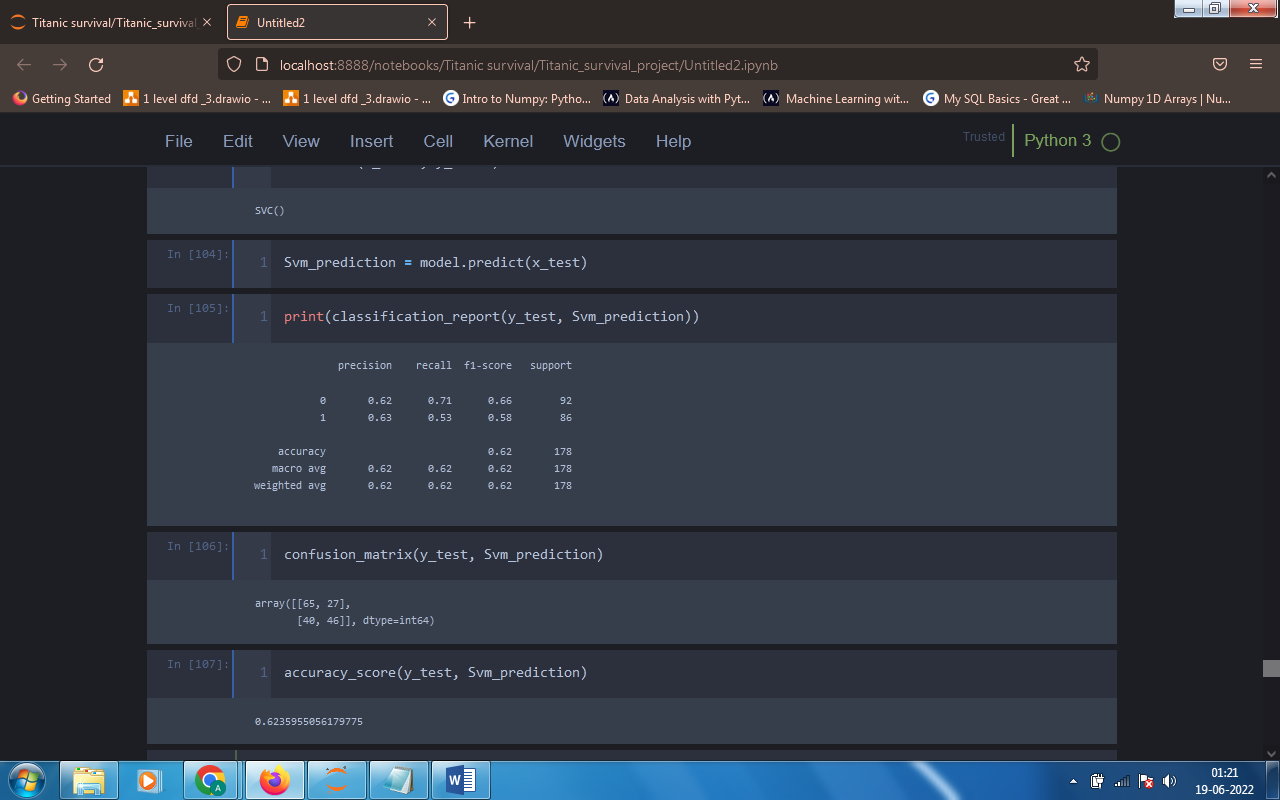


*b. Confusion Matrix :*

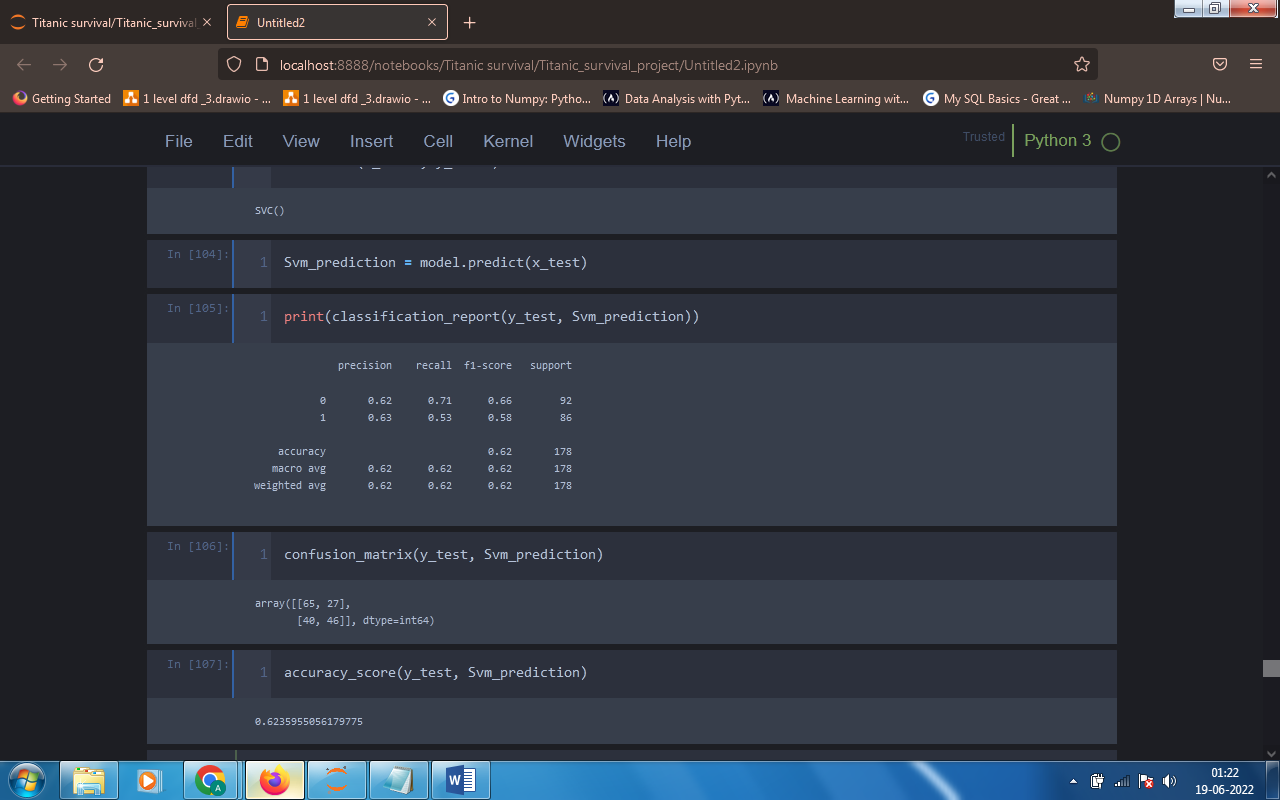


*5.* *Support Vector Machine –*

*a. Classification Report :*

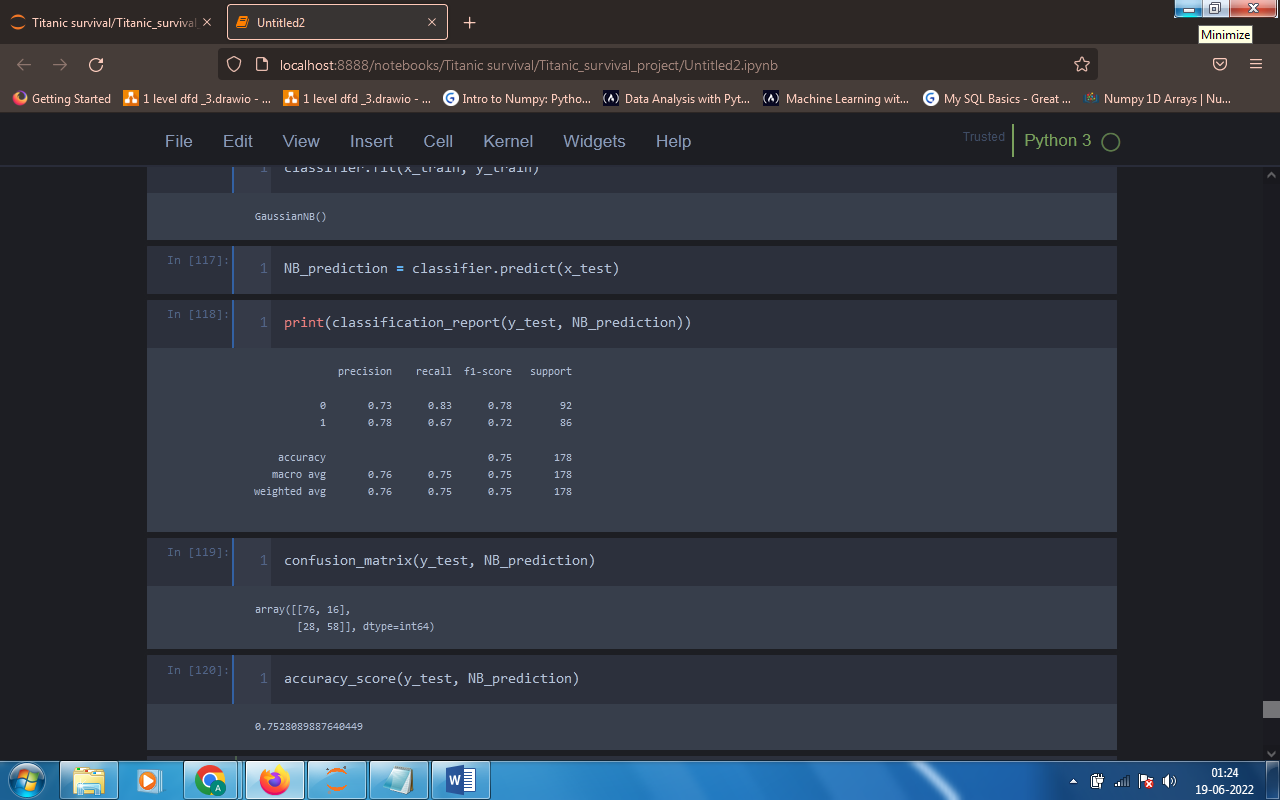


*b. Confusion Matrix :*

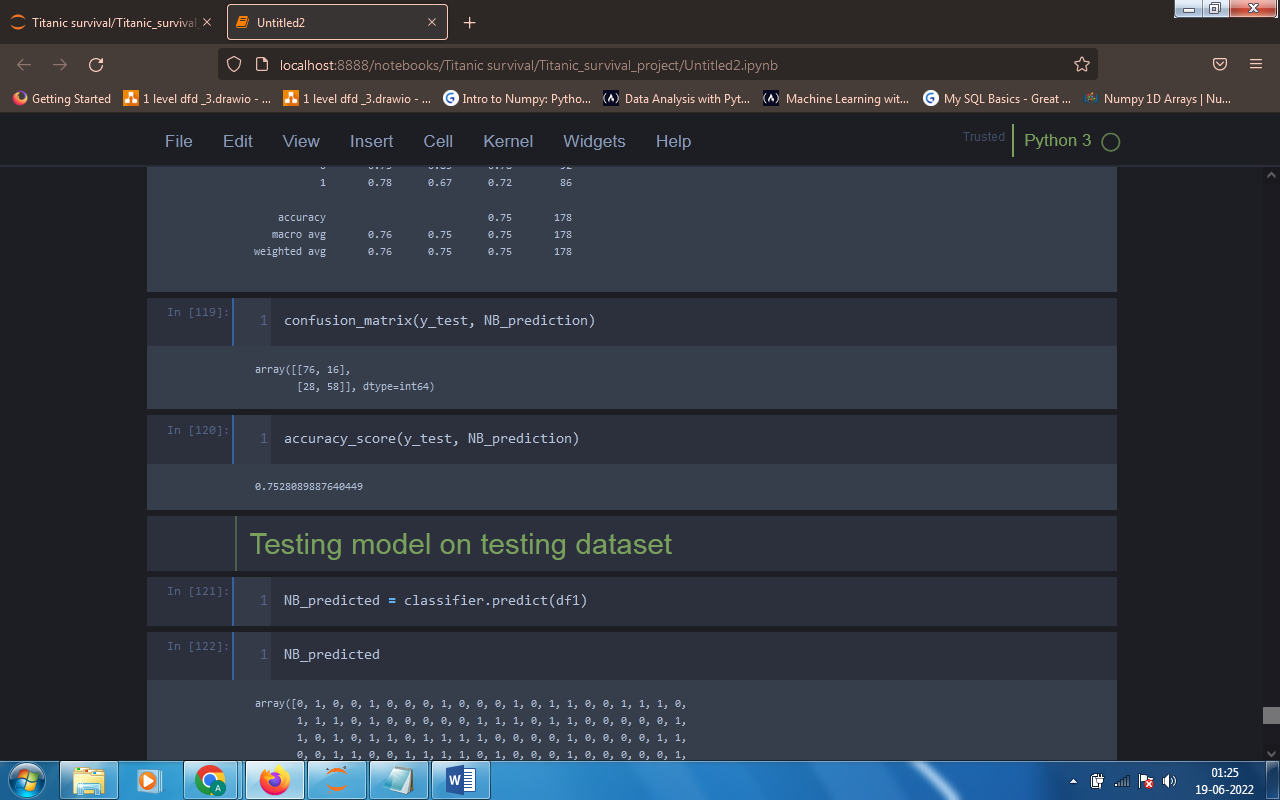


*6.* *Naïve Bayes –*

*a. Classification Report :*



*b. Confusion Matrix :*



*Comparative study –*

|  |  |  |  |
| --- | --- | --- | --- |
| ***ML Model*** | ***Not survived*** | ***Survived*** | ***Total accuracy*** |
| *Logistic Regression* | *0.75* | *0.75* | *0.75* |
| *KNN* | *0.66* | *0.61* | *0.63* |
| *Decision Tree* | *0.76* | *0.69* | *0.72* |
| *Random Forest* | *0.95* | *0.85* | *0.89* |
| *SVM* | *0.62* | *0.63* | *0.62* |
| *Naïve Bayes* | *0.73* | *0.78* | *0.75* |

CONCLUSION

*Among all the Machine learning algorithms the Random Forest algorithm gives the high accuracy so random forest is the best algorithm for our dataset for prediction.*

*From the visualization part we can conclude that the survival and not survival mainly depends on the columns like gender, fare, age and pclass.*