# **Predicting the Chronic Kidney Disease in Machine Learning**

## Abstract

This machine learning project aims to predict chronic kidney disease in individuals based on various health parameters. It enables to make predictions of early detection and proactive healthcare intervention. All the algorithms which supports multiple inputs are taken to sort out the best model and the best model will be saved and handed over to the end user.

### Introduction

This project comes under the supervised Machine Learning with the training and test set. The results are in categorical values which falls under the methodology called Classification. A wide array of supervised machine learning algorithms are available for prediction. Here we have chosen Logistic Regression, K-Nearest Neighbors ,Navie Bayes, Support Vector Machine, DT and Random Forest classification algorithms.

##### Dataset Pre-processing Module

The dataset that was used for this project contains Four hundred rows and twenty five columns (399\*25). All the columns holding nominal data (fig.i) are converted into numerical data (fig.ii) using a pre-processing technique called one hot encoding.

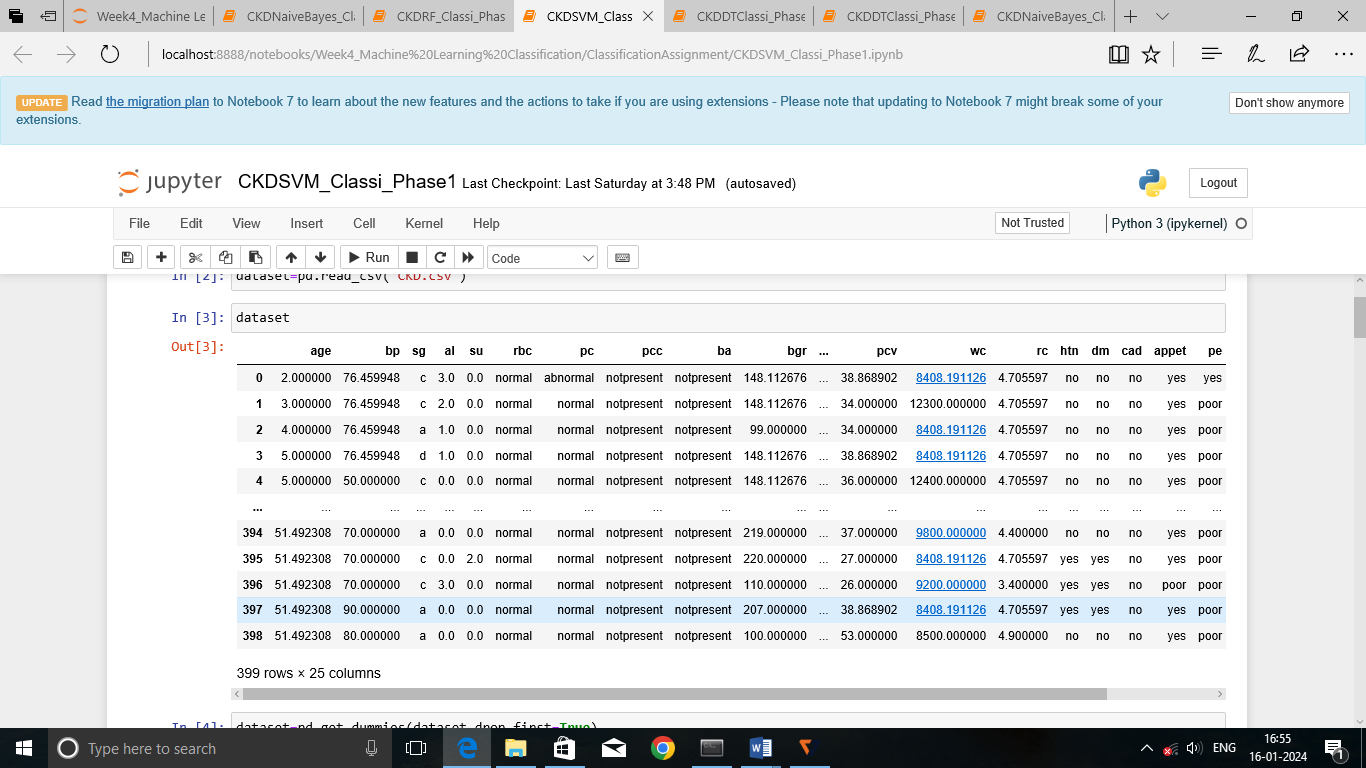


Fig.i) Nominal data

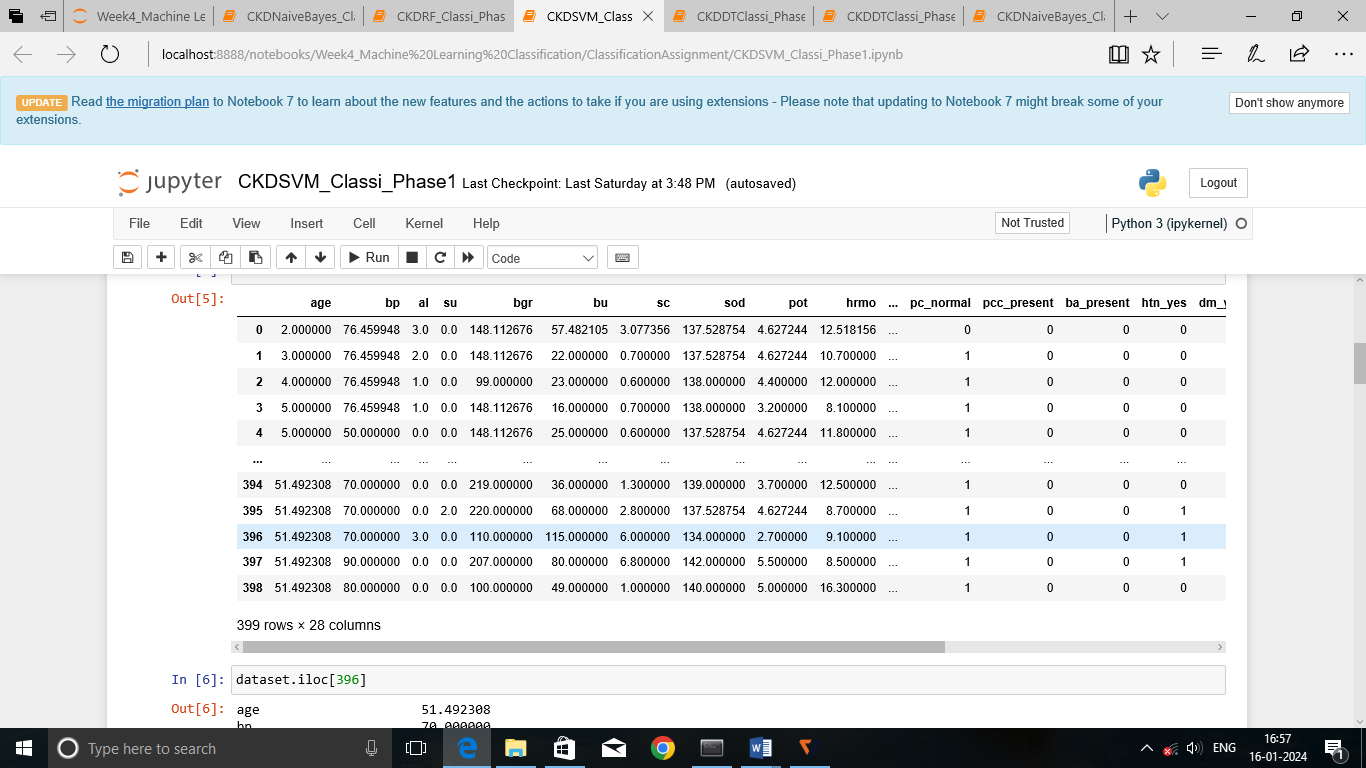


Fig. ii) One Hot Encoding Technique

Algorithms

1. Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear and nonlinear data. The main objective of the SVM algorithm is to find the optimal hyper-plane in an N-dimensional space that can separate data points in different classes as maximum as possible.

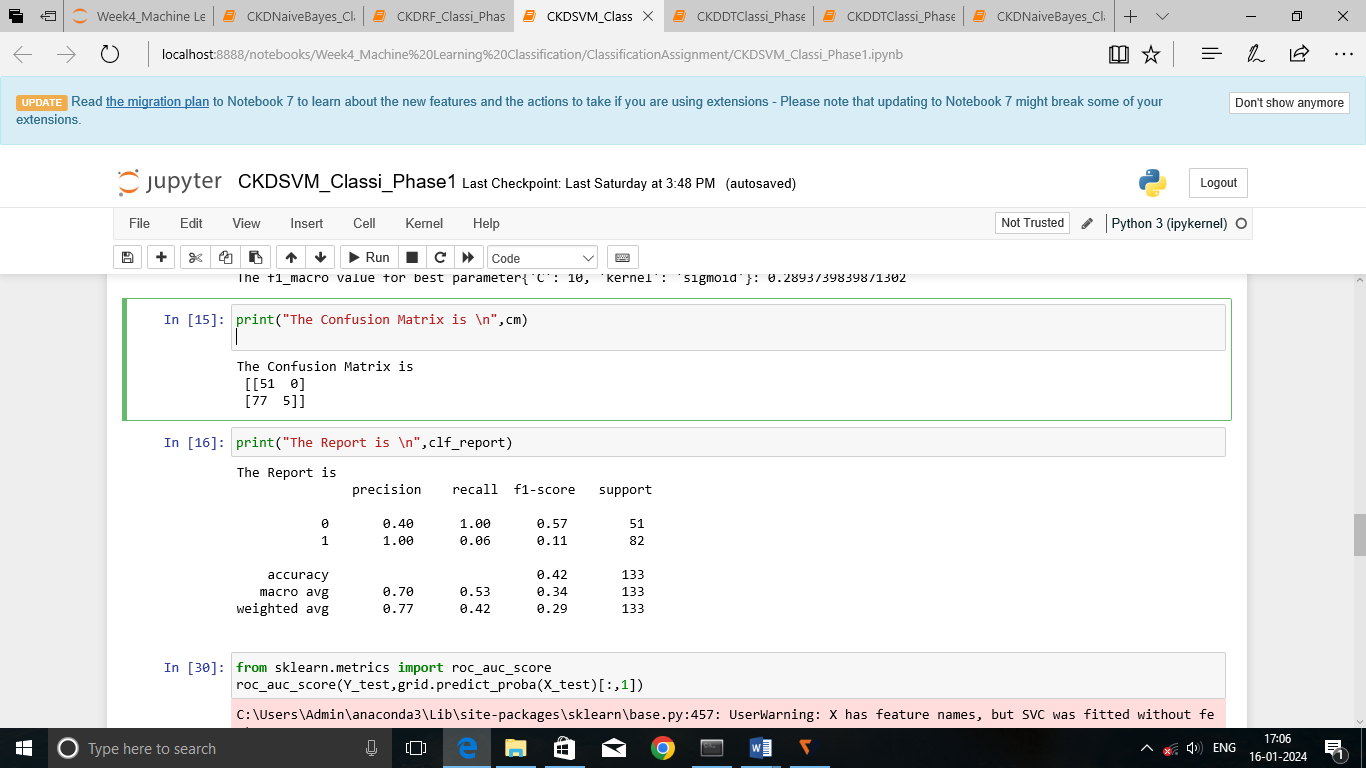


Fig.1.clf\_report(SVM)

#### ROC\_AUC\_SCORE is 0.55

2. Decision Tree Algorithm

This algorithm also supports both classification and regression. It builds a tree like structure from the training dataset given. During training it finds the best attribute that maximizes the information gain.

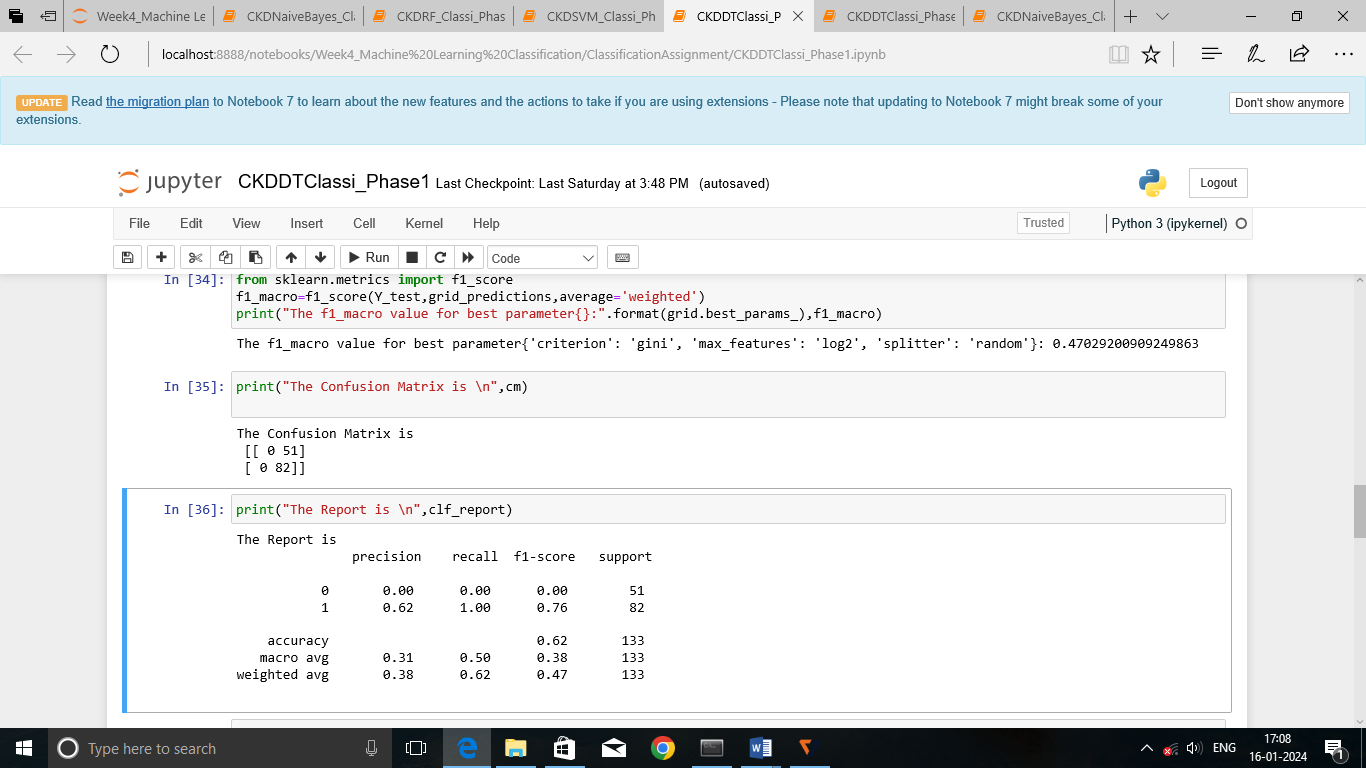
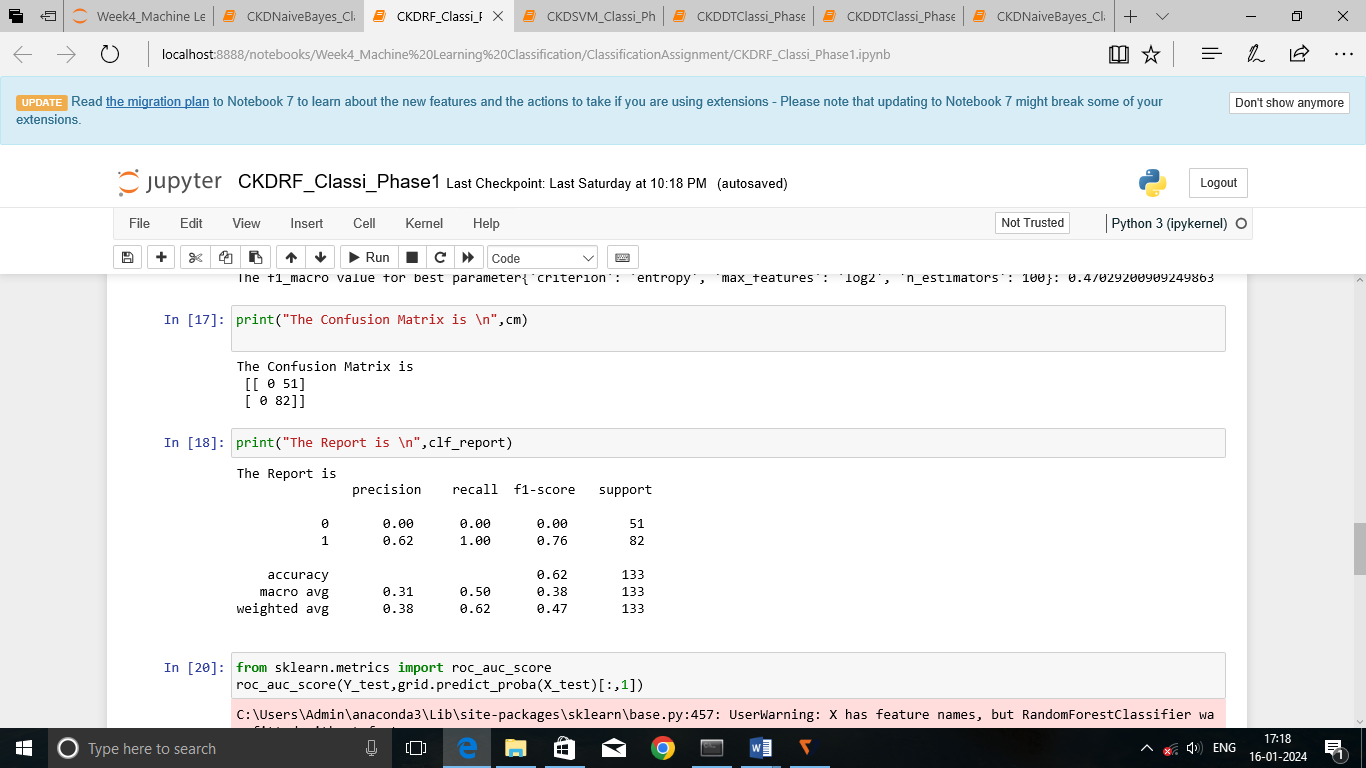


Fig.2.clf\_report(DT)

ROC \_AUC\_Score : 0.50

3. Random Forest

#### This algorithm contains several decision trees derived from each column and takes the average to improve the predictive accuracy of the dataset.



Fig,3.clf\_report(RF)

ROC\_AUC\_Score:0.97

#### 4. Logistic Regression

It is the algorithm to predict the probability of a binary(yes/no) event occurring. Here we use a logistic function, also known as sigmoid function that takes input as independent variables and produces a probability value between 0 and 1.

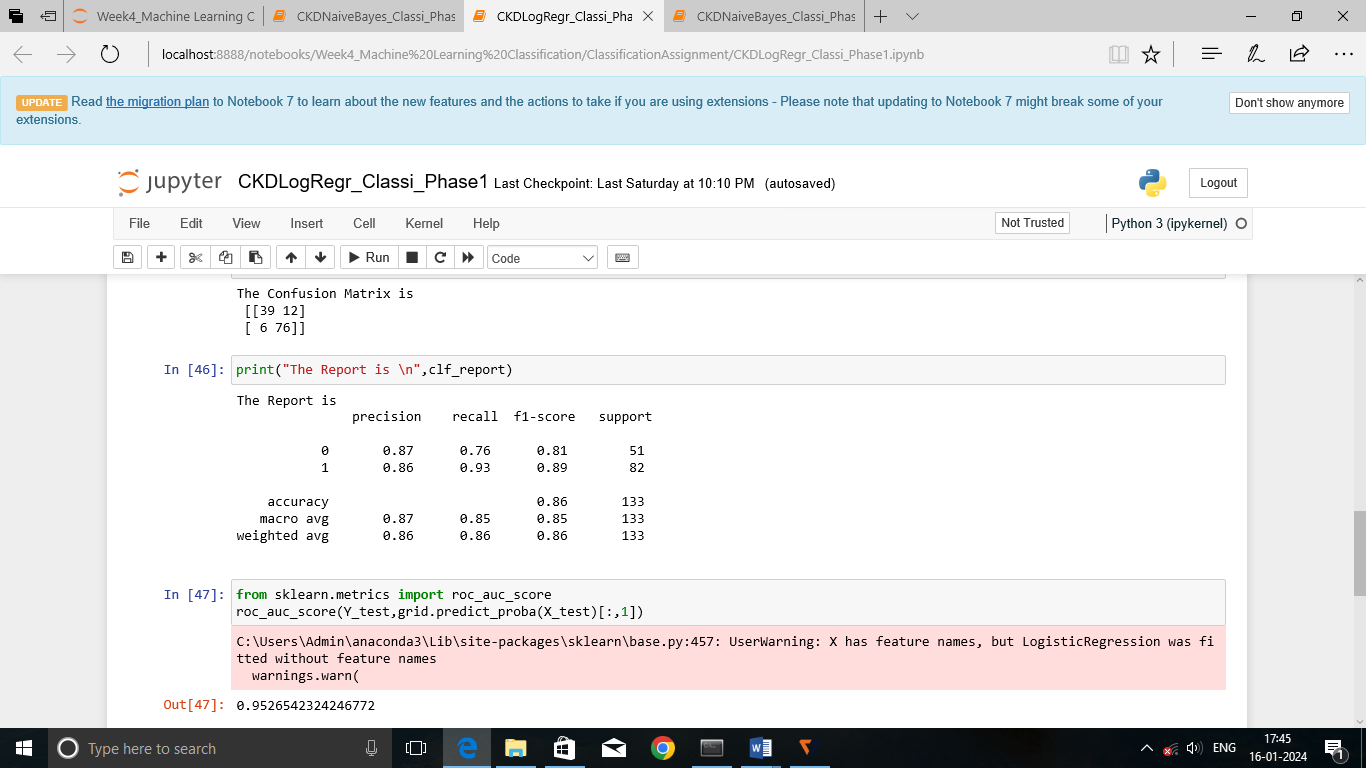


Fig.4.clf\_report(LogisticRegression)

ROC\_AUC\_Score : 0.95

5. K-Nearest Neighbor

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric. The specific method for selecting the nearest neighbors can vary, but a common approach is to sort the distances in ascending order and choose the K data points with the shortest distance.

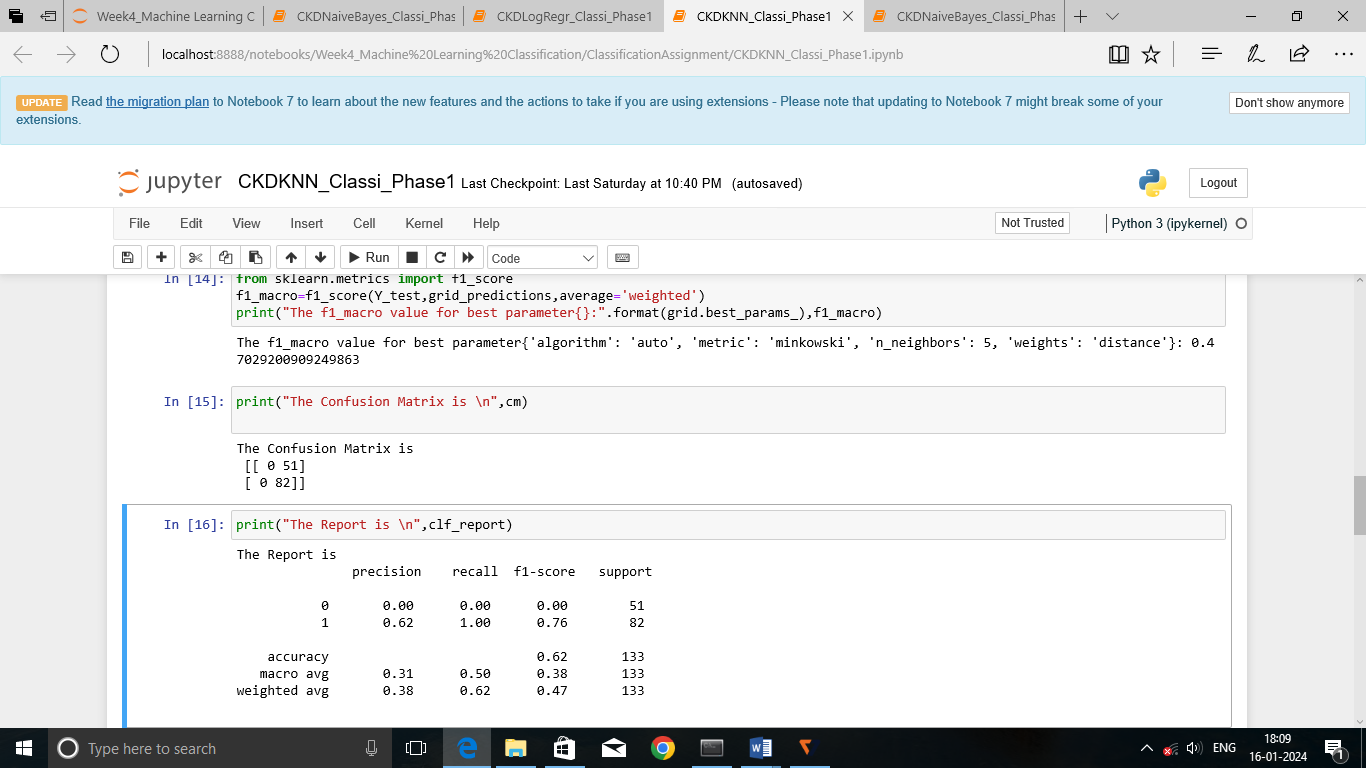


Fig.5.K-Nearest Neighbor

ROC\_AUC\_Score : 0.50

6. Naive Bayes

Naïve Bayes algorithm is a supervised learning algorithm,which is based on Bayes theorem and used for solving classification problems. The Bayes theorem gives us the conditional probability of event A, given that event B has occurred.

1. GaussianNB

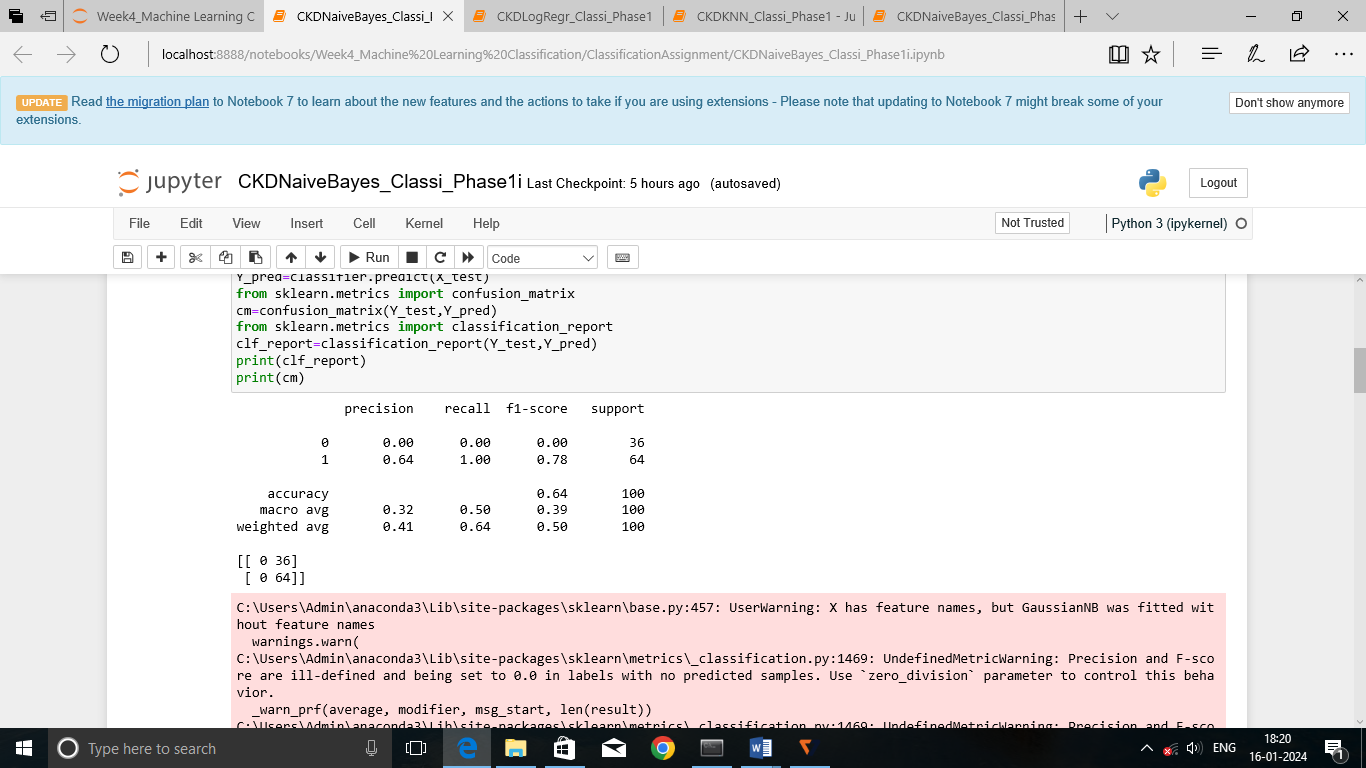


Fig.6.1 clf\_report GaussianNB

ROC\_AUC\_Score : 0.50

1. MultinomialNB

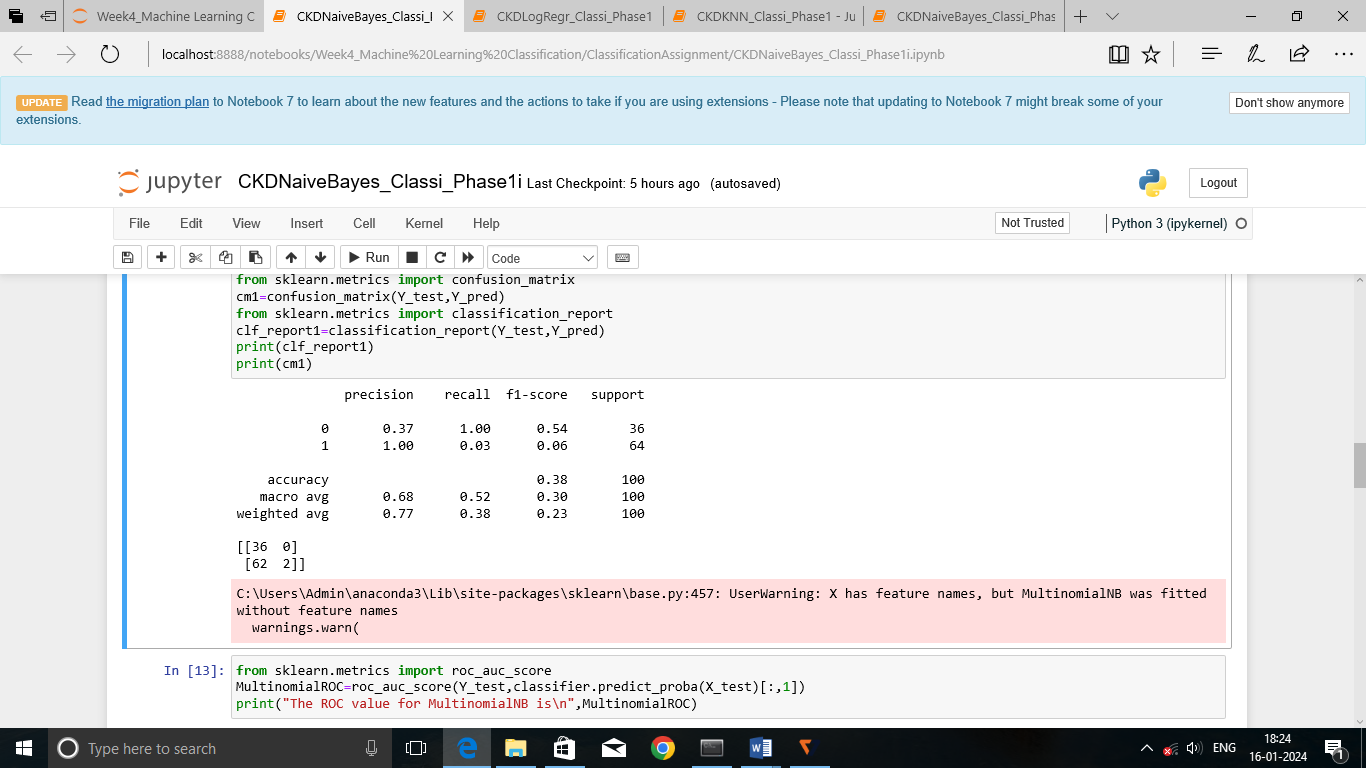


Fig.6.2 MultinomialNB

ROC\_AUC\_Score : 0.65

1. BernoulliNB

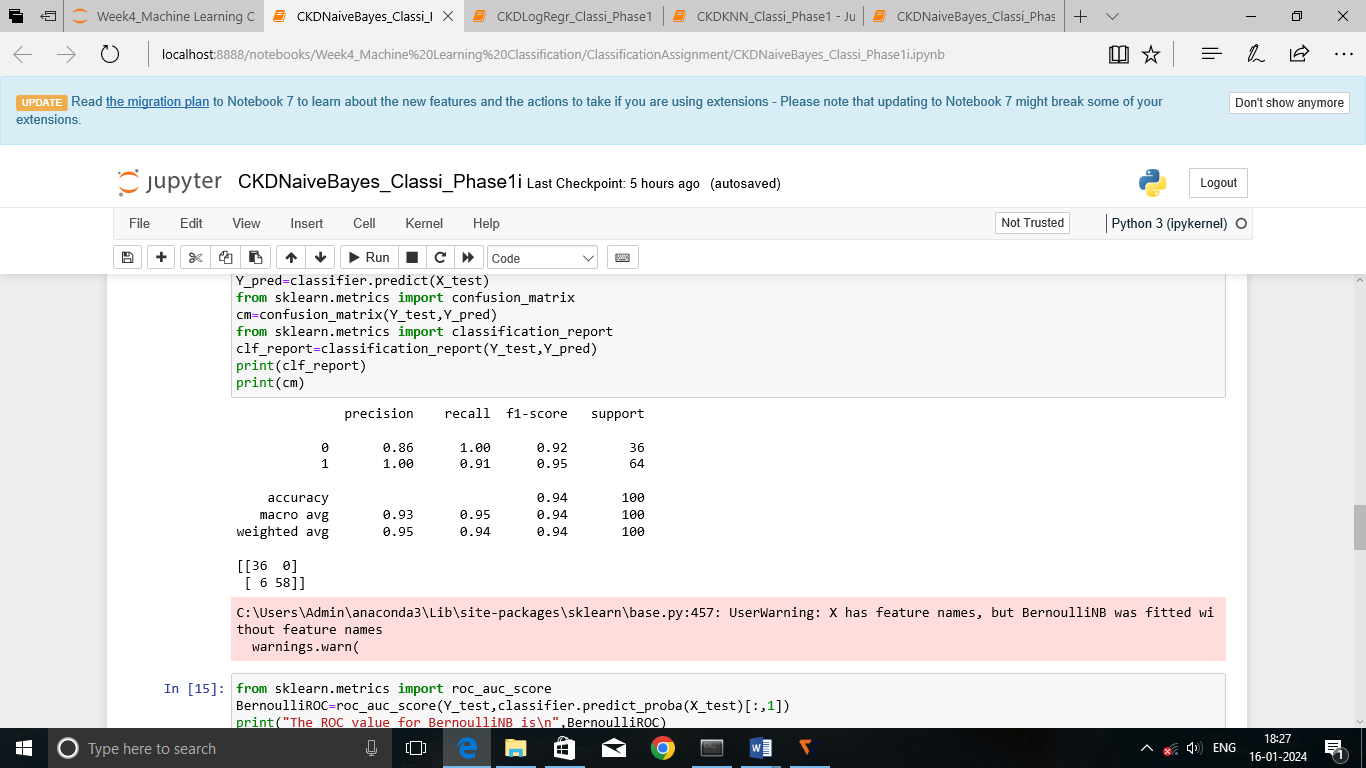


Fig.6.3 BernoulliNB

ROC\_AUC\_Score :0.99

ComplementNB

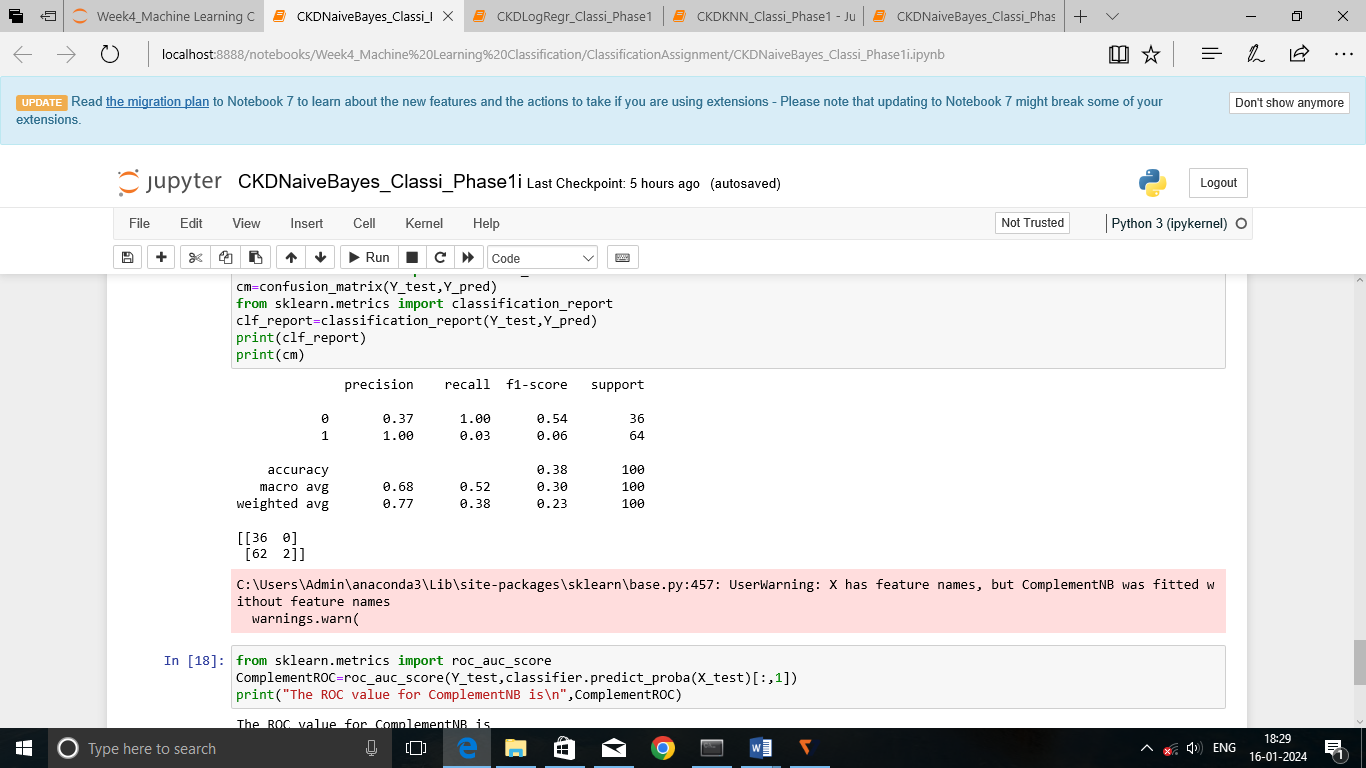


Fig.6.4 ComplementNB

ROC\_AUC\_Score: 0.65

###### Conclusion

As a result the accuracy of the model created by Bernoulli Naïve Bayes algorithm is good compared to the other algorithms. Hence the model created from learning the training set through Bernoulli Naïve Bayes is sorted out and saved to be deployed to the end users.