**1. PREDICTING DIABETICS MELLITUS**

**1.1 Introduction:**

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world’s diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. In order to verity the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.7584) when all the attributes were used.

**1.2 Objectives of Research:**

The main objective of the project is to inform a person about his diabetic condition based the glucose test values. The research is mainly aimed at predicting whether a person is a diabetic or not. it is mainly useful to larger organizations like hospitals for predicting the diabetic condition of patients. When they want to know the diabetic condition of thousands of patients it would be a time taking process for the organizations to study and analyze all the reports.

So in this situations the data model comes into picture.

**1.3 Problem statement:**

Predicting whether a person is having diabetics or not based on the different values like BMI, insulin, glucose levels, etc that are evaluated from the blood sample taken from the person.

**2. REVIEW OF LITERATURE:**

Literature related to life style modification in management of diabetes mellitus . A study to assess type 2 diabetic patients with inadequate glycaemic control on oral hypoglycemic agents (OHA), or a lifestyle intervention programme based on exercise and diet counseling (a) was as effective as insulin treatment in controlling blood glucose, and (b) could prevent the weight gain usually accompanying the introduction of insulin treatment. Thirty eight type 2 diabetic subjects were treated with Oral hypoglycemic agents. Study results showed that, there was no significant difference between the groups in the change observed between start and 12 months of treatment (P = 0.74). There was a significant difference in weight changes between groups (P<0.01).

**3. DATA COLLECTION:**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database.

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Specifically, the training dataset has columns Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. The testing dataset has these same columns but does not contain the column 'Outcome'. This is the column that you predict yourself!

**4. METHODOLOGY:**

Brief Description of Algorithms Used:

Support Vector Machine (SVM):

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyper plane between the two classes[26]. For better generalization hyper plane should not lies closer to the data points belong to the other class. Hyper plane should be selected which is far from the data points from each category.

Decision Tree Classifier:

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes.

Random Forest:

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

K-Nearest Neighbors:

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine learning for regression and classification problem. KNN algorithms use a data and classify new data points based on a similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the most nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

Logistic Regression:

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic regression, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1.

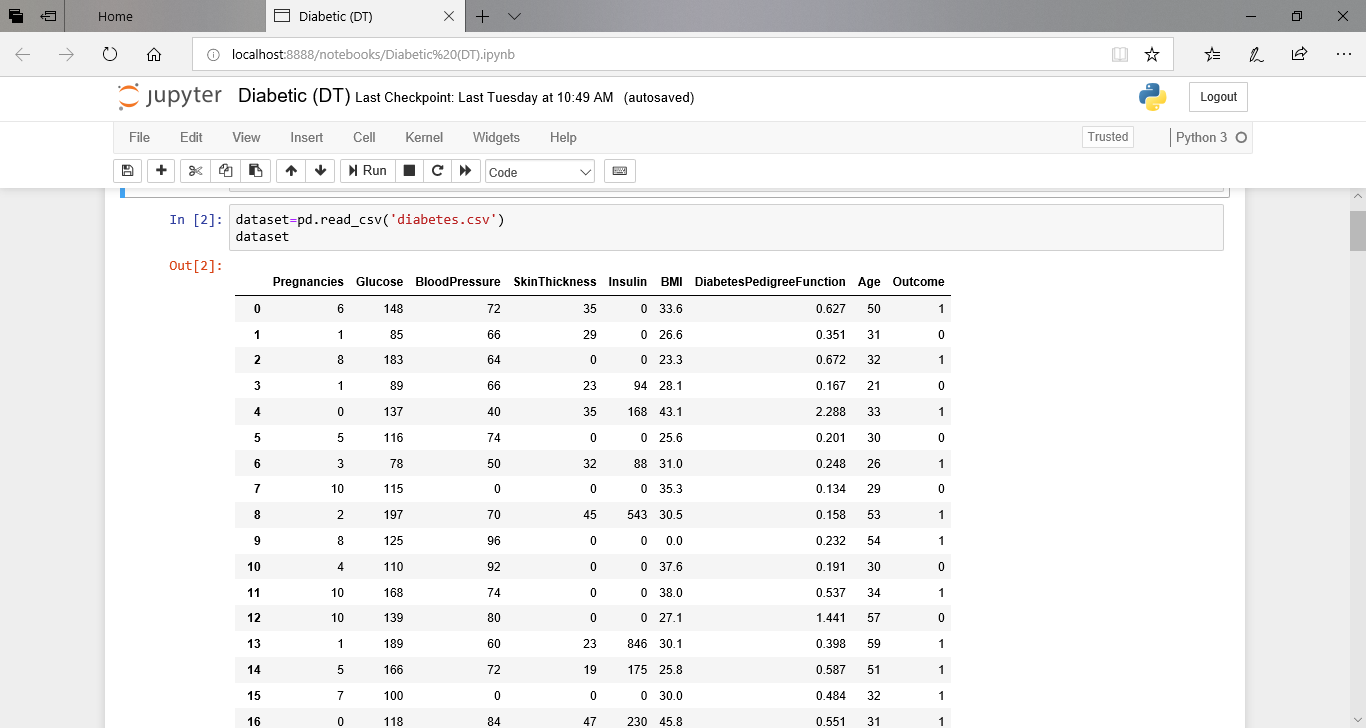
Accuracy Measures:

SVM, Decision Tree, Random Forest, KNN, Logistic Regression algorithms are used in this research work. Experiments are performed using internal cross-validation 10-folds. Accuracy, confusion matrix and ROC (Receiver Operating Curve)measures are used for the classification of this work.

**4.1 Exploratory data analysis:**

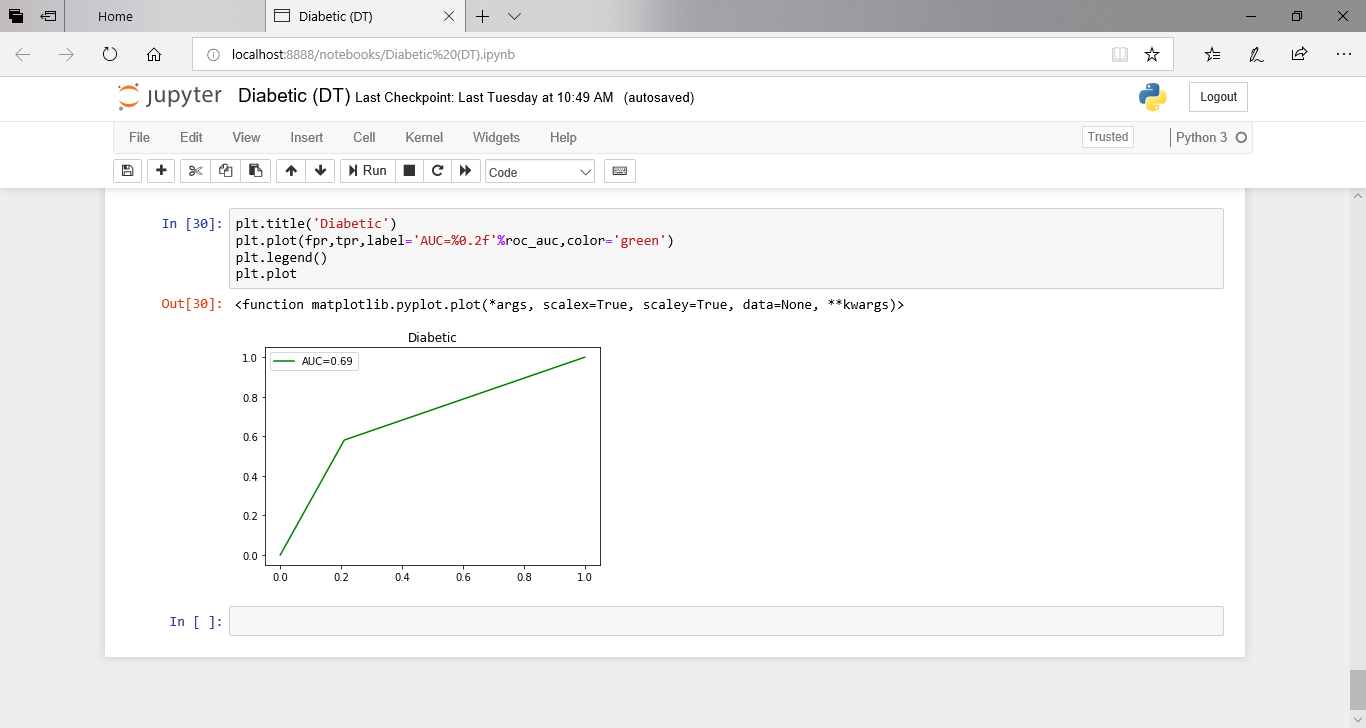
4.1.1 Figures and tables:

Dataset for Diabetic Mellitus:

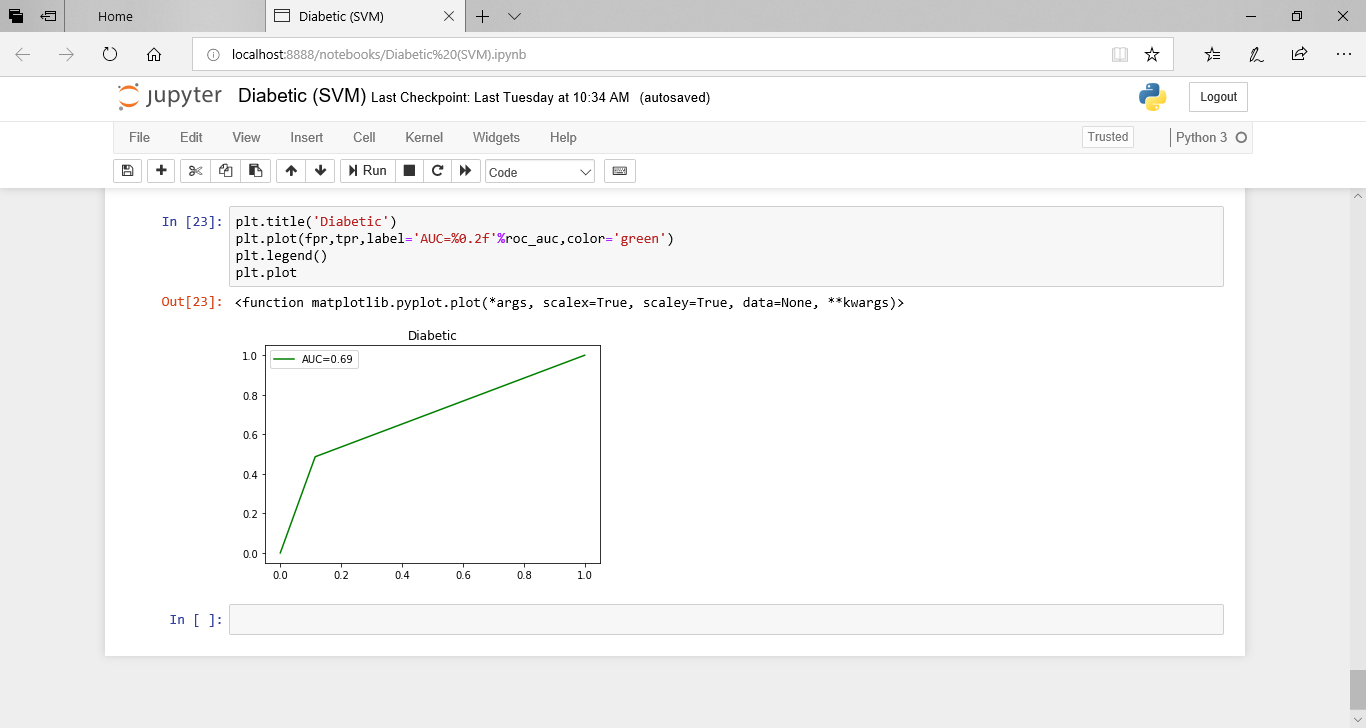


AUC and ROC Curves:

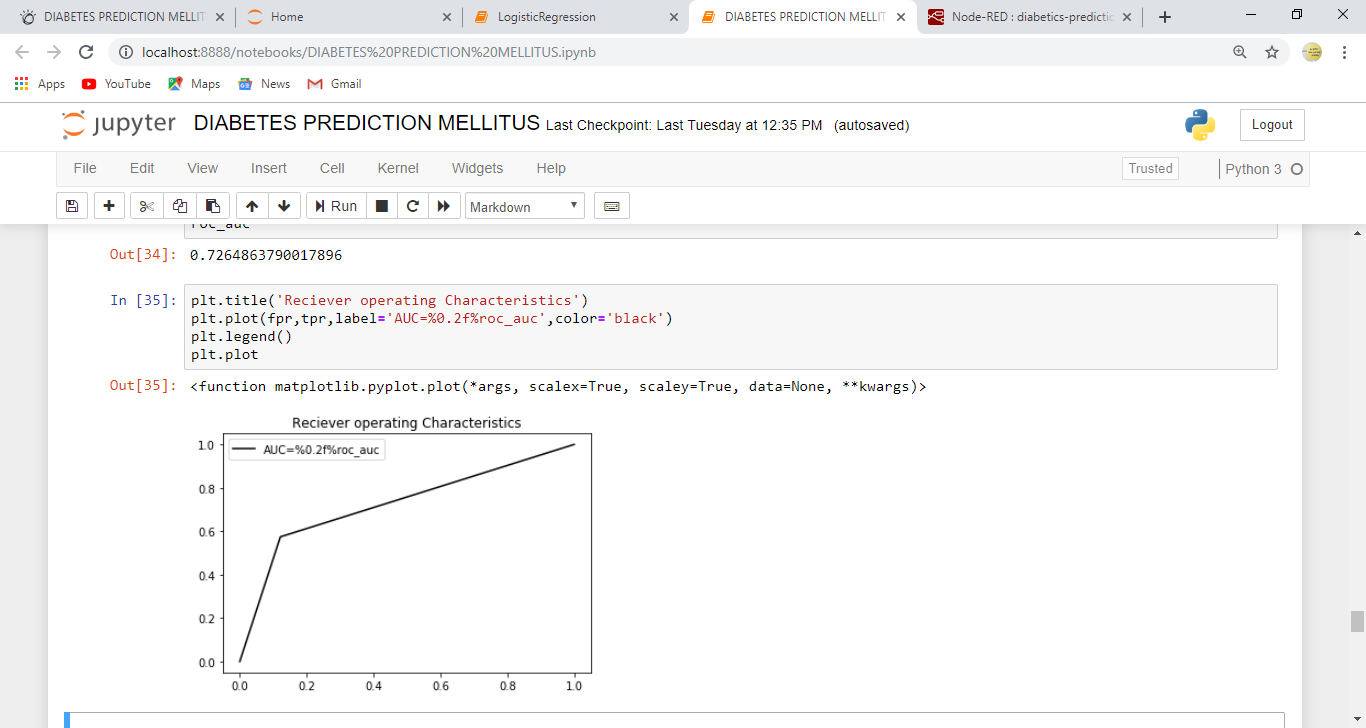
Decision Tree:



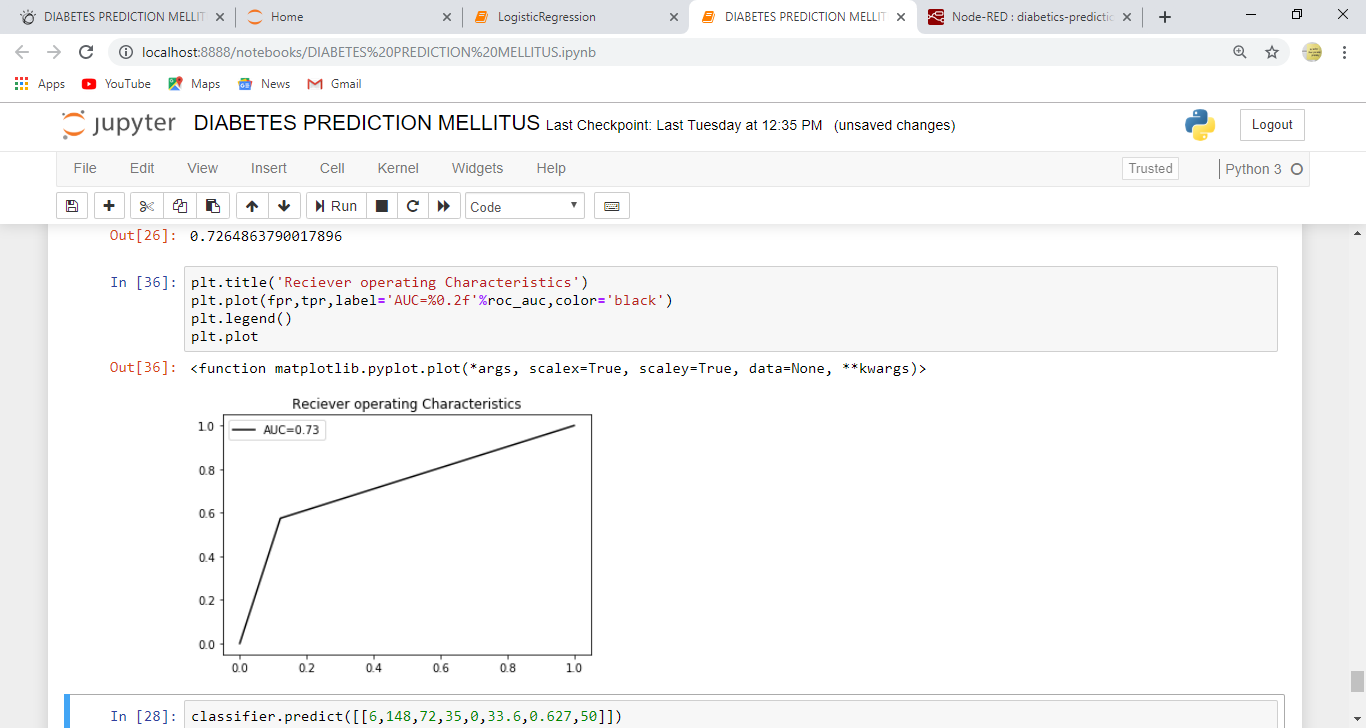
SVM:(Support Vector Machine)



Logistic Regression:

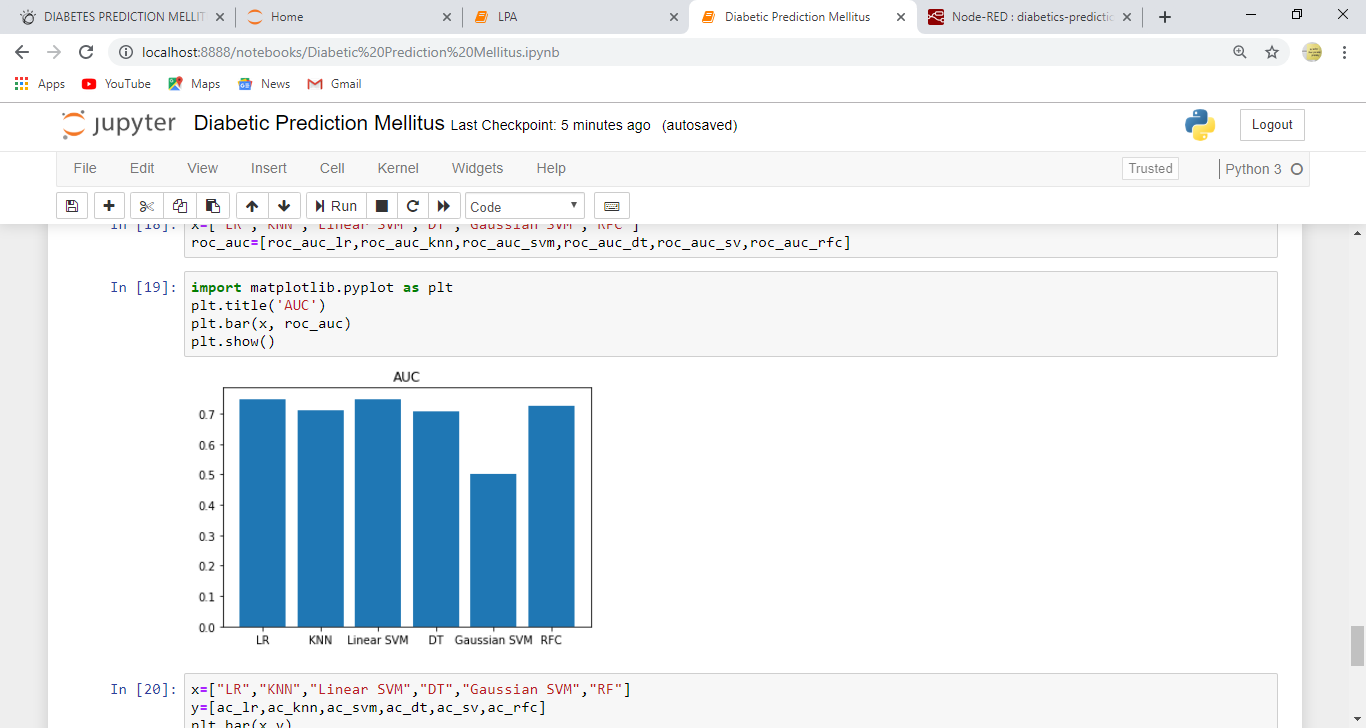


Random Forest:

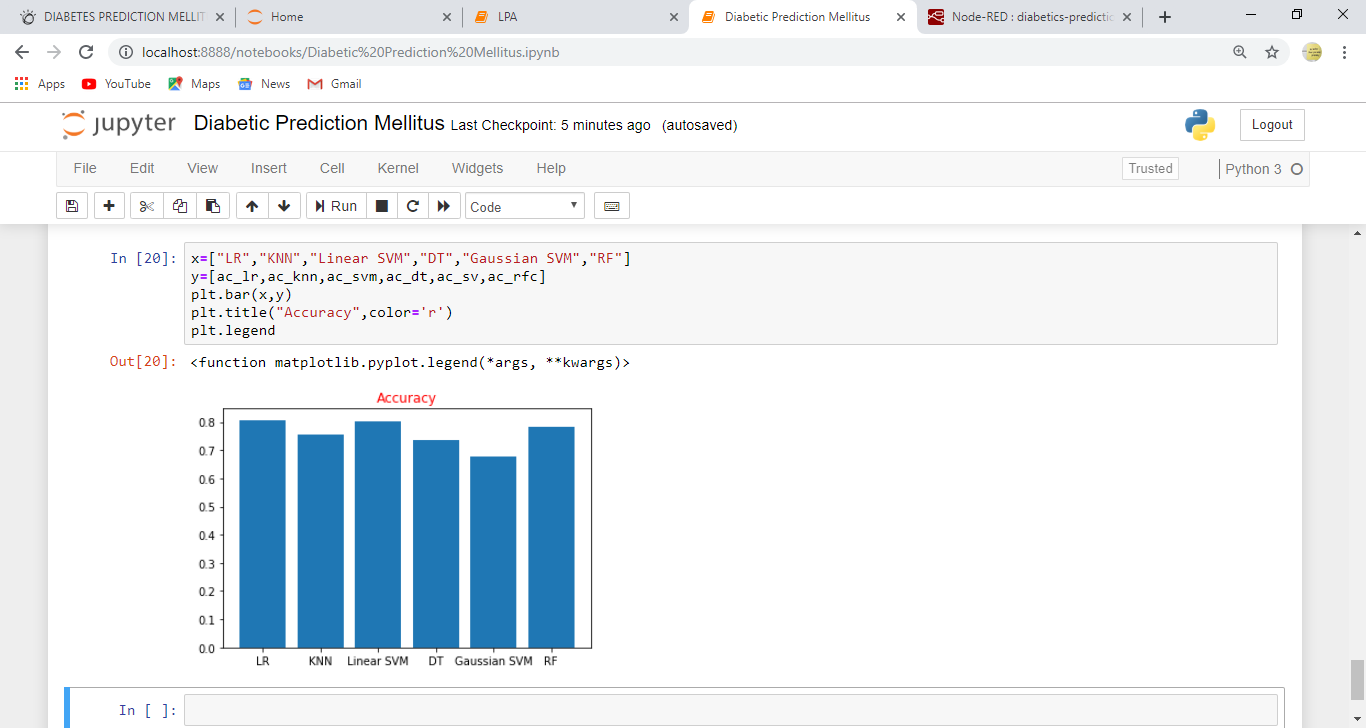


Bar Graphs:

For AUC:



For Accuracy:

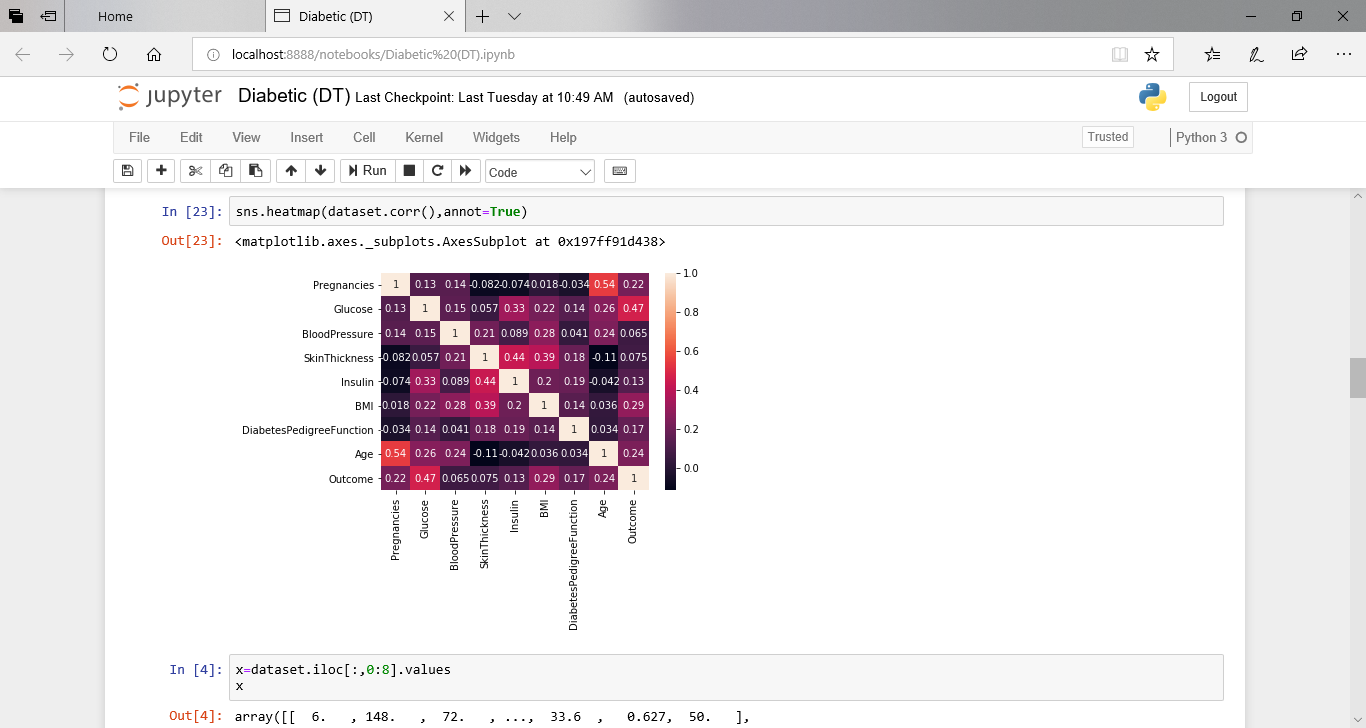


**4.2 Statistical techniques and data visualization:**

By importing matplotlib.pyplot library we have drawn graphs to demonstrate the AUC-ROC curves and by using bar graphs we have visualized the percentage levels of different techniques.

And we have used the co-relation function to demonstrate the impact of every factor on each other.

Correlation graph:



**4.3 Data modeling using supervised learning algorithms:**

In general we have two types of learning algorithms, supervised and unsupervised learning algorithms. and in detail it consists of different techniques like,

Support Vector Machines

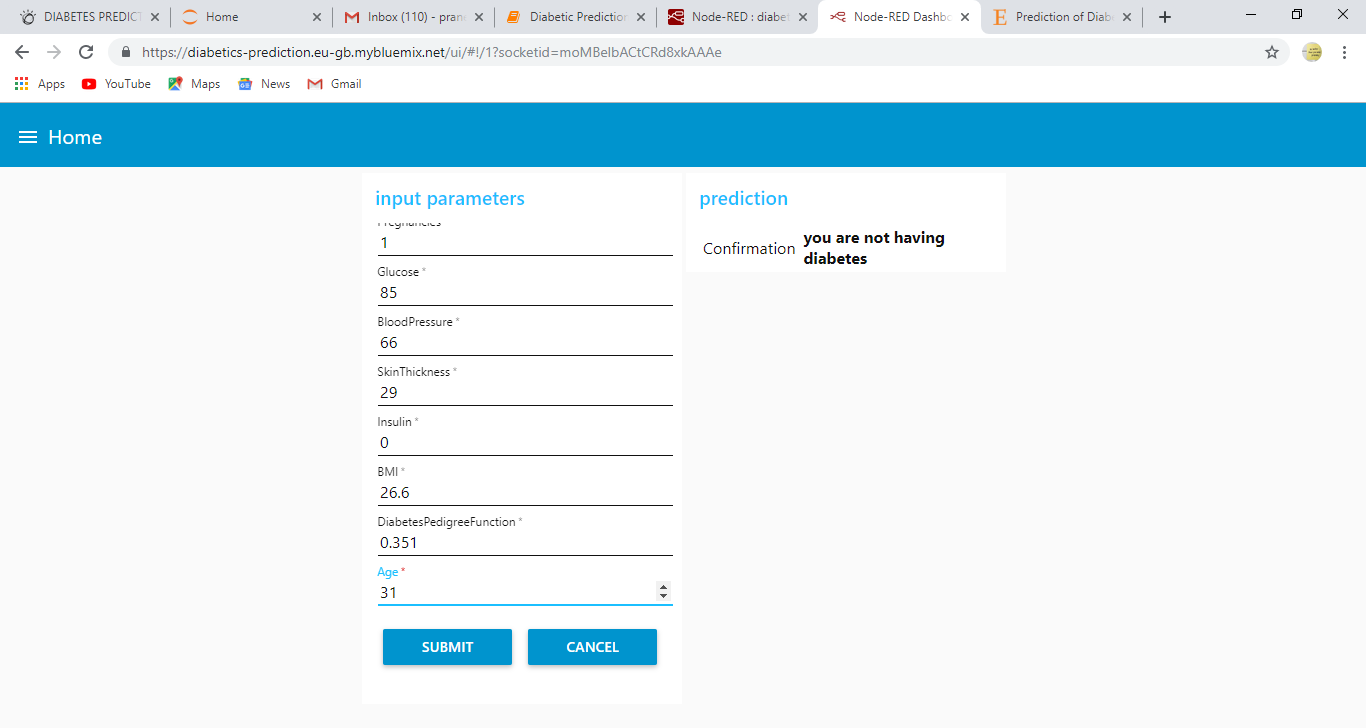
logistic regression

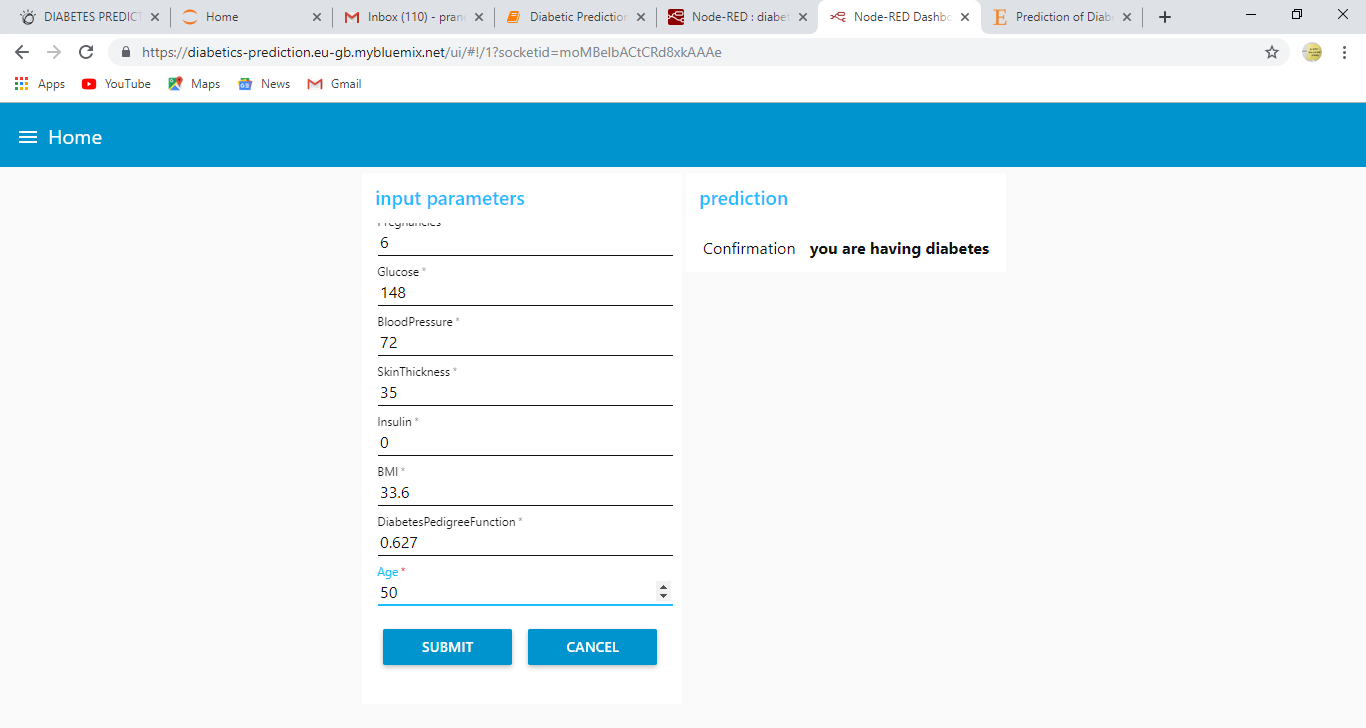
decision tree algorithm

k-nearest neighbor algorithm

since our model comes under supervised learning algorithm we applied every technique of the algorithm and based on the accuracy values we obtained we chosen random forest technique. and our model has  dependent variable which is dichotomous(binary) means the output can be either a person is having diabetes or not having diabetes.

**5.FINDINGS AND SUGGESTIONS:**





**6.CONCLUSION:**

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 0.75 using the Random Forest classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.