**CENTRE OF DEVELOPMENT OF ADVANCED**

**COMPUTING, MOHALI**



**Advanced course on Data Science and Analytics**

**(HP-KVN)**

**Project Report**

**On**

**Predicting Diabetes Mellitus With Machine Learning Techniques**

**Submitted By: Submitted To:**

Diksha

Akshant Chandel Mr. Sanjay Madan

Vivek

# DECLARATION

We hereby declare that we have completed three months industrial training at C-DAC from 24/03/2022 to 22/06/2022 under the guidance of Mr. Sanjay Madan. We hereby declare that the project undertaken by me is the genuine work of mine.

**Name: -**

**Diksha**

**Akshant Chandel**

**Vivek**

# ACKNOWLEDGEMENT

It is our proud privilege and duty to acknowledge the kind of help and guidance received from several people in preparation of this report. It would not have been possible to prepare this report in this form without their valuable help, cooperation and guidance.

First, and foremost, we wish to record our sincere gratitude to C-DAC Coordinators for their constant support and encouragement in preparation of this report and for making available videos and interface facilities needed to prepare this report.

Their contributions and technical support in preparing this report are greatly acknowledged.

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| S.NO | TOPICS | PAGE NO. |
| 1. | 1.Introduction  2. Machine Learning  2.1.Supervised Machine Learning  2.2. Unsupervised Machine Learning | 7-9 |
| 2. | Dataset description | 10 |
| 3. | Proposed Methdology | 11 |
| 4. | 1.1 Importing required Libraries  1.2 Seaborn librabry  1.3 Matplot Library  1.4 Pandas Library  1.5 Numpy Library | 11-13 |
| 5. | Data visualization | 14-15 |
| 6. | 6.1Preprocessing of data  6.2Feature scaling  6.3Train-Test split | 15-16 |
| 7. | 7.1 Build the classifier model using machine  7.2. SVM algorithm  7.3. Random Forest Classifier  7.4. Decision Tree | 16-20 |
| 8. | 1. Building the model and evaluation and finding accuracy   8.1. Confusion matrix | 20 |
| 9 | 9.1.Result  9.2 Conclusion | 21-25 |

**About C-DAC**

Center for Electronics Design & Technology of India (CEDTI), Mohali was setup in May 1989. Primarily with the mission to train manpower in electronic design & technology by offering a variety of training programs in diverse aspects of electronics design, product development, production technology, maintenance engineering, information technology and quality control, etc. In December 2002, CEDTI Mohali merged with C-DAC with a primary mandate to promote high end R&D along with education and training.

The center is engaged in design and deployment of world class IT and electronics solutions in the following domains:

* Health Informatics
* Multilingual Technologies
* Professional Electronics
* Software Technologies
* Cyber Forensics and Security
* Multimedia Technologies

## **ABSTRACT**

Diabetes is a life-long condition that is well-known in the 21st century. Once known as a disease of the West, the rise of diabetes has been fed by a nutrition shift, rapid urbanization and increasingly sedentary lifestyles. In late 2019, a new public health concern was emerging (COVID19), with a particular hazard concerning people living with diabetes. So there is a need to take serious precautions to eradicate this. Medical institutes have been collecting data for years. To eradicate, we should know the level of occurrence. In this project we predict the level of occurrence of diabetes. We predict the level of occurrence of diabetes using Random Forest, a Machine Learning Algorithm. Using the patient’s Electronic Health Records (EHR) we can build accurate models that predict the presence of diabetes,which hopefully will prevent the loss of lives and improve the quality of life using data mining processes. The robust and accurate prediction of diabetes is highly challenging due to the limited number of labeled data and also the presence of outliers (or missing values) in the diabetes datasets. This work proposes a comparative study of data mining techniques for early diagnosis of diabetes. We use a publicly accessible data set containing .The dataset has 8 columns and 769 rows in the

dataset where each row corresponds to a new article. For the classification problem underconsideration we have used ‘title’ and ‘text’ as input columns and column ‘type’ as output

variable. The title, text and type are of string data type. Naive Bayes, Neural Network, AdaBoost, k-Nearest Neighbors, Random Forest and Support Vector Machine methods have been tested. The results suggest that Neural Networks should be used for diabetes prediction. The proposed model presents an AUC of 98.3% and 98.1% accuracy, an F1-Score, Precision and Sensitivity of 98.4% and a Specificity of 97.5%.

INTRODUCTION

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030.However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

The project is sub-divided following section. These are:

1.Loading necessary libraries

2.Loading Dataset from a CSV file or from a Table.

3.Summarization of Data to understand Dataset (Descriptive Statistics)

4.Visualization of Data to understand Dataset (Plots, Graphs etc.)

5.Data pre-processing and Data transformation

6.Hyper parameter tuning to find the optimal parameters for the classification models.

7.Applying different learning algorithms on the training dataset.

8.Evaluating the performance of the fitted model using evaluation metrics like confusionmatrix, precision recall curves.

**MACHINE LEARNING**

Machine learning is a branch of [artificial intelligence(AI)](https://www.ibm.com/in-en/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

**Supervised machine learning**

[Supervised learning](https://www.ibm.com/cloud/learn/supervised-learning), also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the crossvalidation process to ensure that the model avoids [overfitting](https://www.ibm.com/cloud/learn/overfitting) or [underfitting](https://www.ibm.com/cloud/learn/underfitting). Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more.

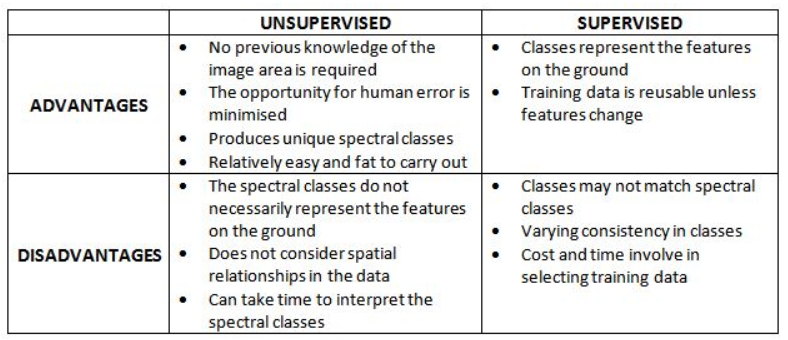


### **Unsupervised machine learning**

[Unsupervised learning](https://www.ibm.com/cloud/learn/unsupervised-learning), also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. It’s also used to reduce the number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more.



### **Supervised machine learning vs Unsupervised machine learning**



1. **Dataset Description-**

The source of this dataset is Kaggle. The dataset has 9 columns and 769 rows in the

dataset where each row corresponds to a new article. For the classification problem under

consideration we have used ‘title’ and ‘text’ as input columns and column ‘type’ as output

variable. The title, text and type are of string data type. The data is classified into following types

- bias, conspiracy, fake, satire, hate, junksci and state. However, in some of the rows some values

were missing for some of the features and value for some of the features are ‘NaN’. The data was

cleaned in the Data pre-processing steps mentioned below, before using it for training different

models.

Dataset URL: <https://www.kaggle.com/diabetes>

Table 1: Dataset Description

|  |  |
| --- | --- |
| **S No.** | **Attributes** |
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

# I. PROPOSED METHODOLOGY

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

## **Procedure of Proposed Methodology-**

**Step1:** Import required libraries, Import diabetes dataset.

**Step2**: Pre-process data to remove missing data.

**Step3:** Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

**Step4:** Select the machine learning algorithm i.e.KNearest Neighbor**,** Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

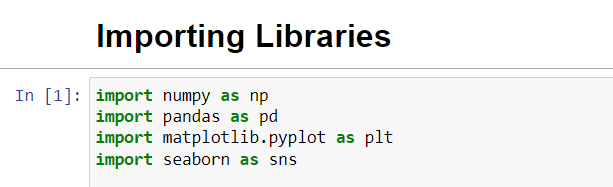
**Step5:** Build the classifier model for the mentioned machine learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned machine learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

**Step8:** After analyzing based on various measures conclude the best performing algorithm.

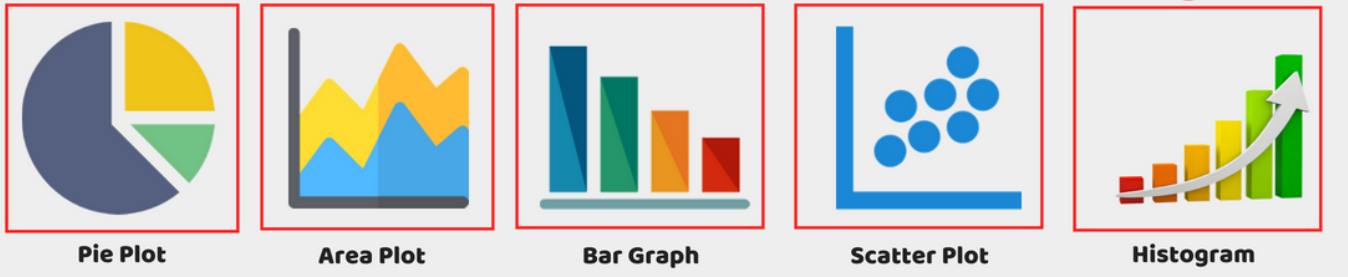
**Step1:**First we Import required libraries, Import diabetes dataset.



A. Matplotlib

Matplotlib is one of the most popularly used data visualizationlibraries of python. This library was built by a John Hunterwho is along with several contributors, and it had put in agreater amount of time into prompting this software used byevery scientist and philosopher across the globe .Matplotlib is a graphics library for data visualization packagein Python which encompasses as an integral aspect in thepython data science stack and it is easily supported withNumPy, Pandas and other relevant libraries.

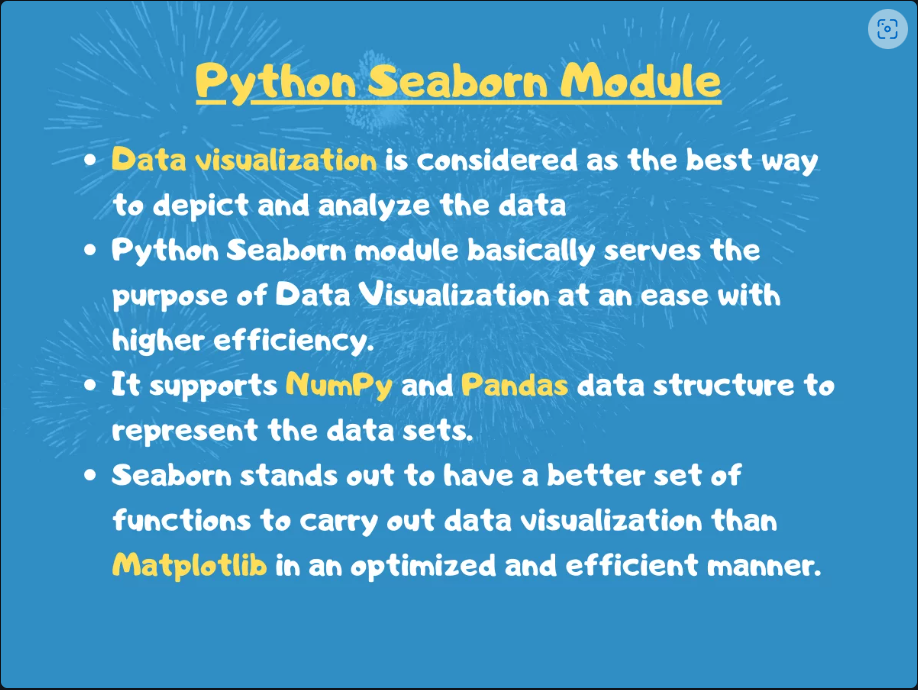
Different plots in matplot library-



B.Seaborn

Seaborn is a graphic visualization library that is built on theprimary configurations of Matplotlib. It provides accessibilityto the users with some of the most commonly provides datavisualizations processes with certain data visualizationsnecessities such as mapping colour to a variable or usingfaceting requirements across the globe. It provides seaborn is

more integrated for working with Pandas DataFrames .



C. Pandas

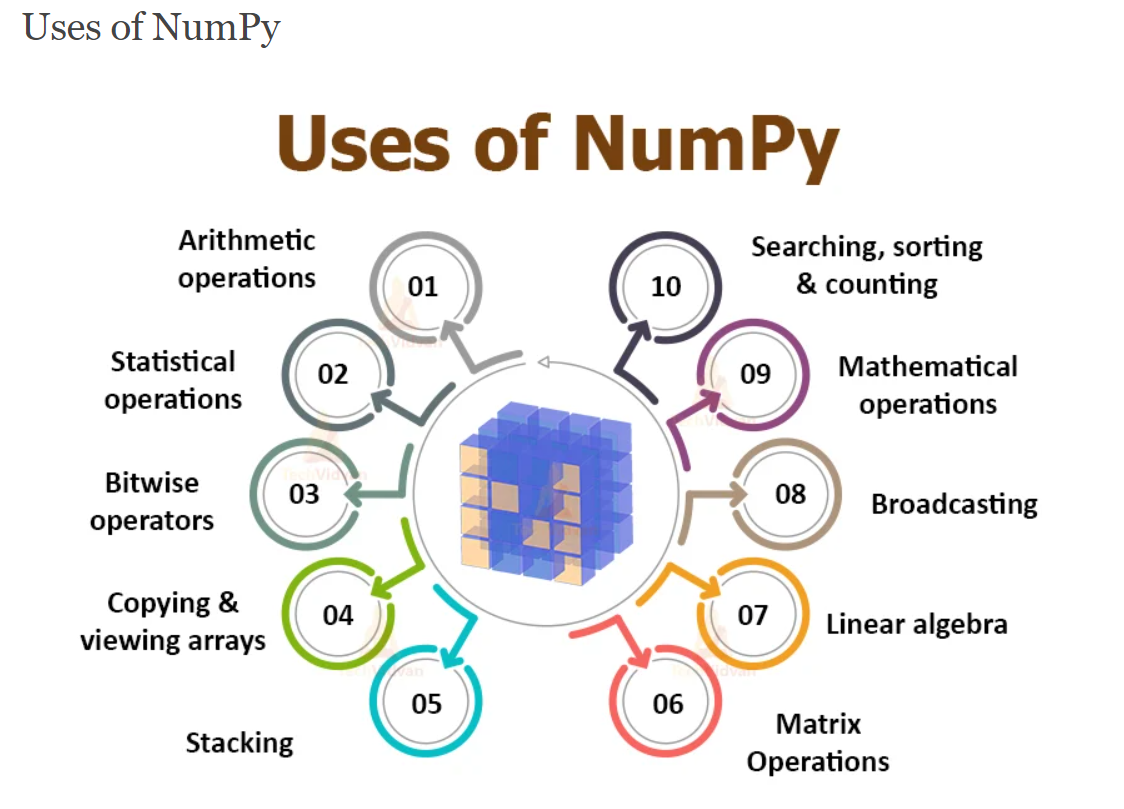
Pandas is an open-source library used in Python that providesenhanced performance metrics, easy to use data structures anddata analysis packages, tools and librariesforPythonProgramming Language .The use of pandas withpython encompassesvarious fields of expertise including datasciences, computational modelling, finance, economics,statistical analysis, machine learning, etc.

Applications of Pandas –

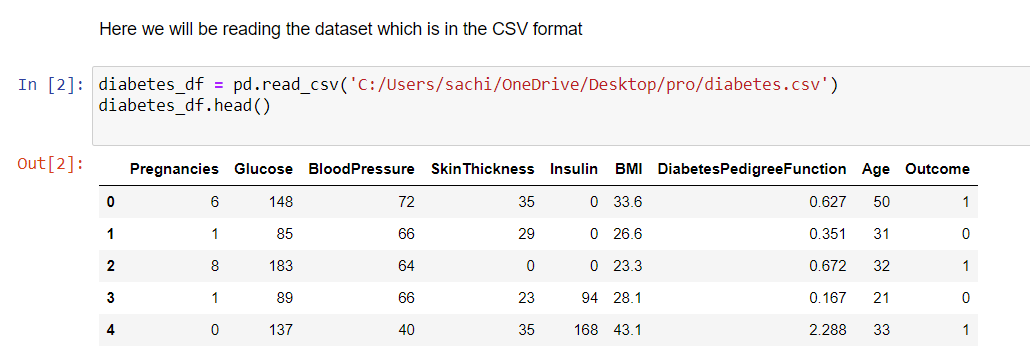


D. NumPy

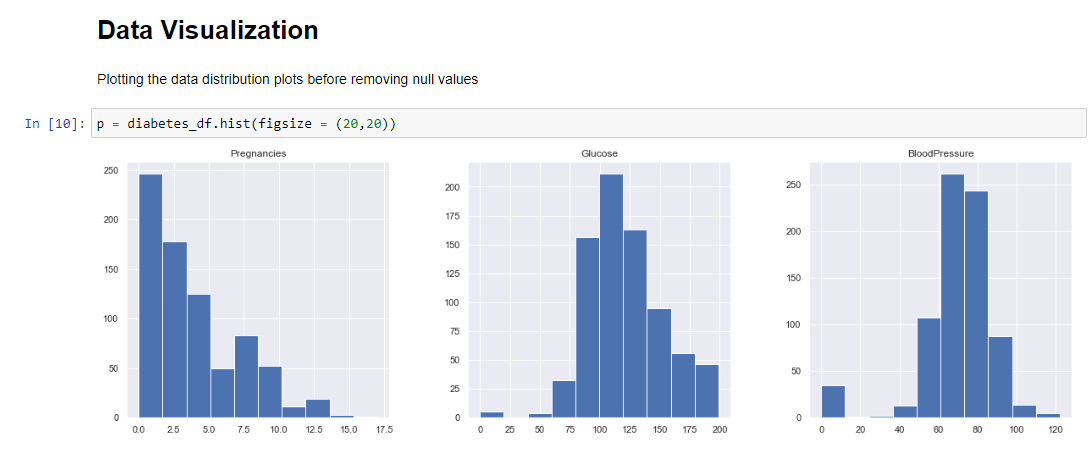
Numerical Python (NumPy) is a data analysis and numericalcomputation library that consists of multidimensional objectsof the array and a collection of routines or procedures that areused for the processing of similar types of arrays. Using theNumPy library, data scientists, system programmers andarchitects can perform mathematical computations, numericaland logical operations on these arrays to build a sustainableand efficient computational model for scientific andnumerical perspective.

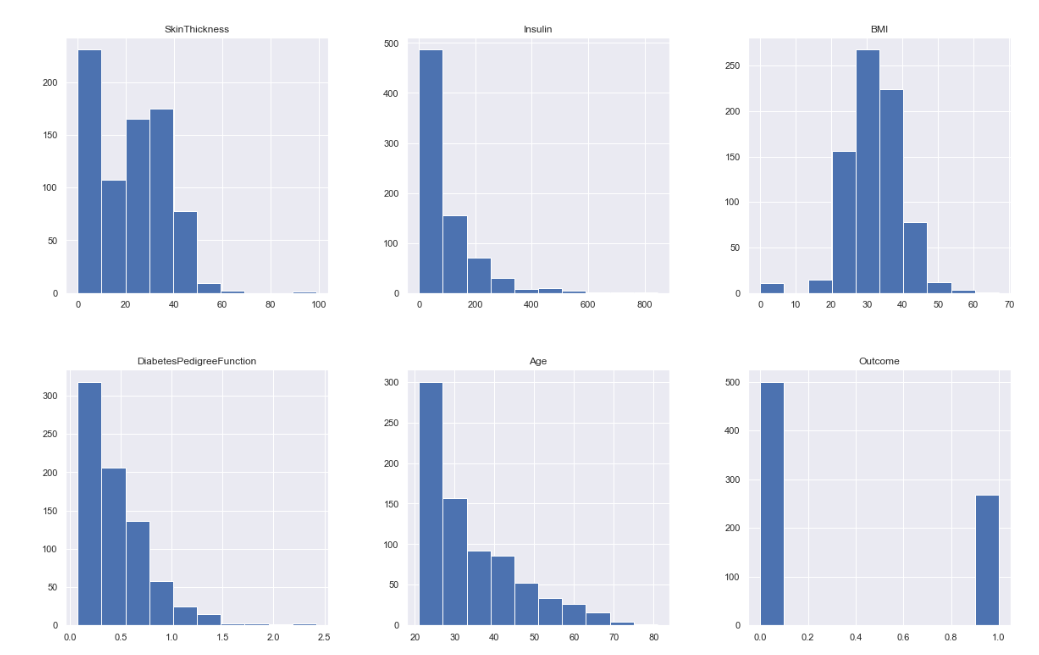


1.2-Reading the dataset



**Visualizing the data:-**

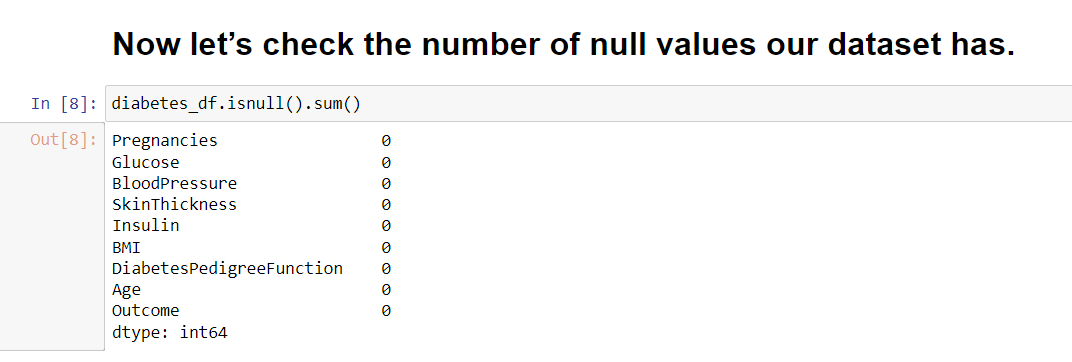




**Step2**: Pre-process data to remove missing data.

**Data Preprocessing-** Data preprocessing is most important process. Mostly healthcare related data contains missing value and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For our dataset we need to perform pre processing in two steps.

**1).Checking for null value/ Missing Values removal-** We check for null values first **.**Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentonality of data and help to work faster.



**2). Splitting of data-** After cleaning the data, data is normalized in training and testing the model. When data is splitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.



**Distribution of Diabetic patient-** We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

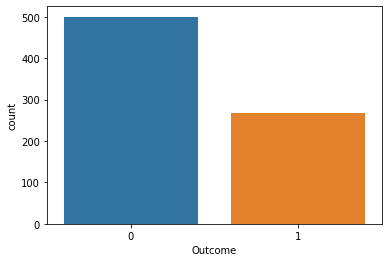


Figure 1: Ratio of Diabetic and Non Diabetic Patient

**Step3: Build the classifier model for the mentioned machine learning algorithm based on training set**:-

When data has been ready we apply Machine Learning Technique. We use differentclassification and ensemble techniques, to predict diabetes. There are 3 methods applied onour dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques are follows-

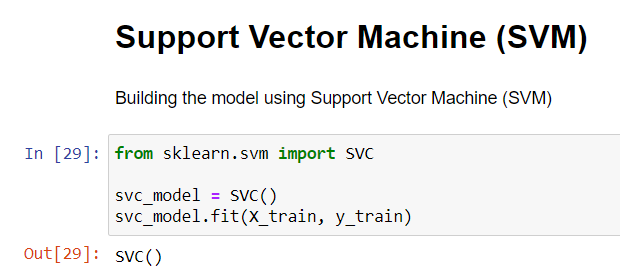
**1) Support Vector Machine-** Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplane in high dimensional space. This hyper plane can be used for classification or regression also. Svmdifferentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

## **Algorithm-**

* Select the hyper plane which divides the class better.
* To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
* If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
* Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.



Figure 2: Overview of the Process



**5) Random Forest –** It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

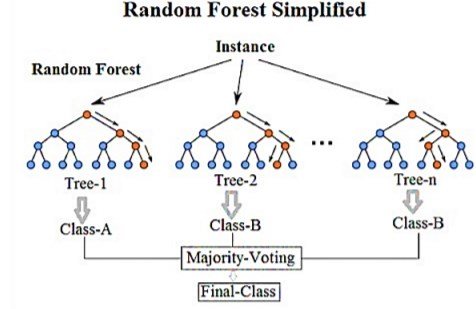


Figure 2: Overview of the Process

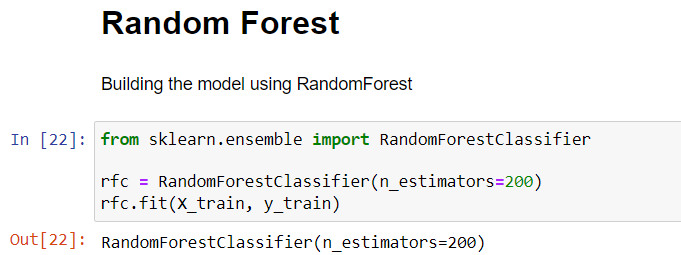
## **Algorithm-**

* The first step is to select the “R” features from the total features “m” where R<<M. •Among the “R” features, the node using the best split point.
* Split the node into sub nodes using the best split.
* Repeat a to c steps until ”l” number of nodes has been reached.
* Built forest by repeating steps a to d for “a” number of times to create “n” number of trees.

The random forest finds the best split using the Gin-Index Cost Function which is given by:



The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest does correct predictions result for a spread of applications are offered.



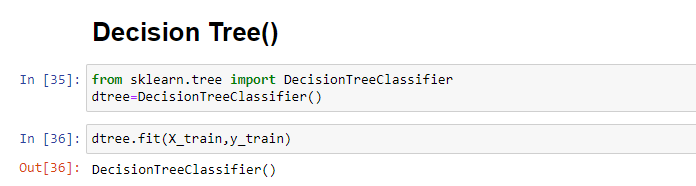
**6Decision Tree-** Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classificationprocess based on input feature. Input variables are any types like graph, text, discrete, continuous etc.

Steps for Decision Tree **Algorithm-**

* + Construct tree with nodes as input feature.
  + Select feature to predict the output from input feature whose information gain is highest.
  + The highest information gain is calculated for each attribute in each node of tree.
  + Repeat step 2 to form a subtree using the feature which is not used in above node.

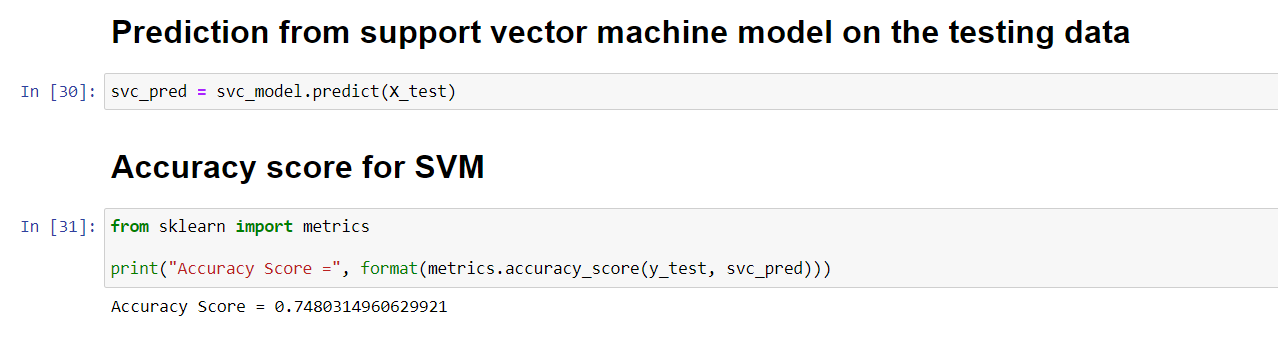


Figure 2: Overview of the Process

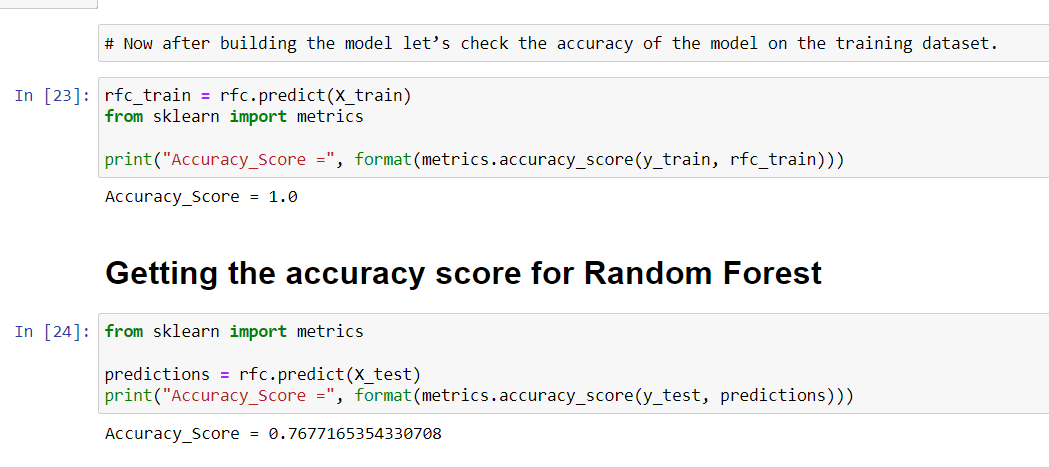


**Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set and find the accuracy :-**

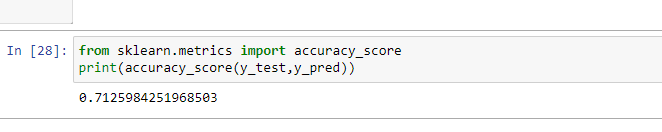
1. **Support Vector Machine-** accuracy for support vector machine is given below



1. **Random forest classifier-** accuracy for Random forest classifieris given below



1. **Decision Tree -** accuracy for Decision tree is given below



# V. EXPERIMENTAL RESULTS

In this work different steps were taken. The proposed approach uses different classification and ensemble methods and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see that random forest classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy.

**RESULTS:**

We use classification reports, confusion matrix for each class and precisionrecall-

f1 curves as the evaluation metrics to analyze the results obtained for our models.

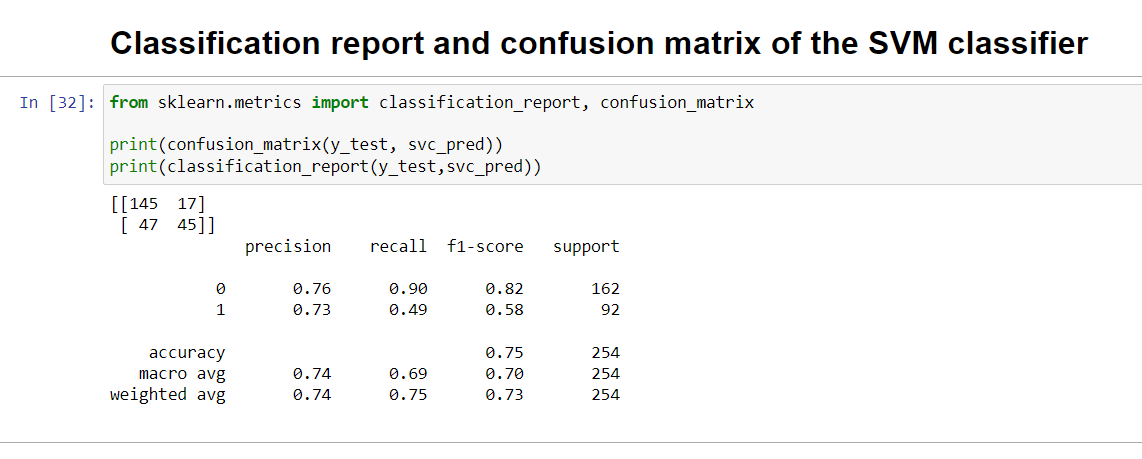
Overall,generalizes well when compared to, Random Forest and Support Vector Machine.

## **Table 1: Results Summary**

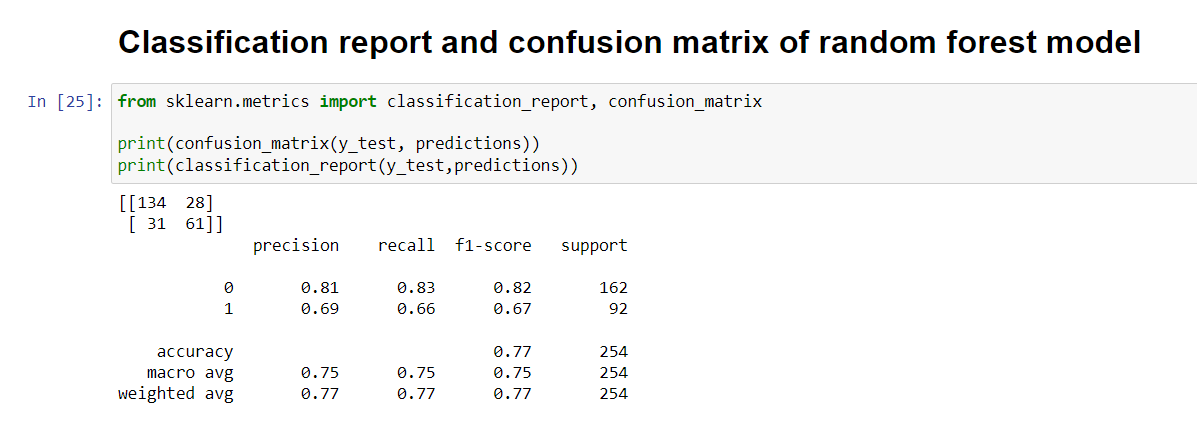
|  |  |  |  |
| --- | --- | --- | --- |
| **MODEL** | **RANDOM FOREST** | **SVM** | **Decision Tree** |
| Precision | 0.81 | 0.76 | 0.77 |
| Recall | 0.83 | 0.90 | 0.77 |
| Accuracy | 0.76 | 0.74 | 0.74 |
| F1-Score | 0.82 | 0.82 | 0.78 |

# Figure shows the Classification report and confusion matrix of these Machine Learning methods:-

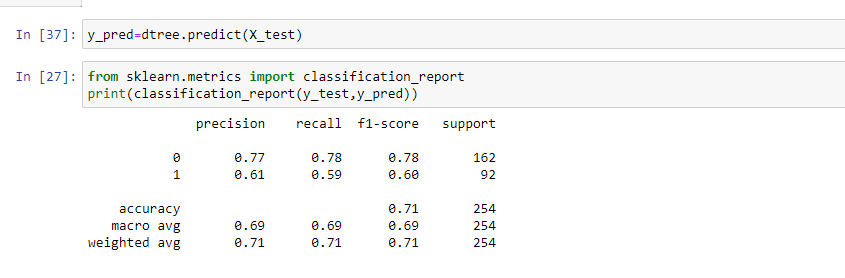
1. **Support Vector Machine-** Classification report and confusion matrix of support vector machine is given below



1. **Random Forest Classifier-** Classification report and confusion matrix ofRandom forest classifier is given below



1. **Decision Tree -** Classification report and confusion matrix ofDecision Treeis given below



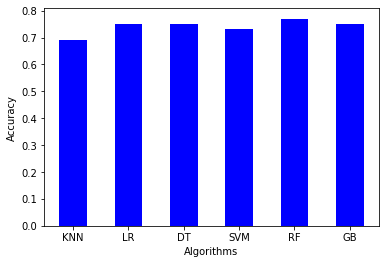


Figure3: Accuracy Result of Machine learning methods

Here feature played important role in prediction is presented for random forest algorithm. The sum of the importance of each feature playing major role for diabetes have been plotted, where X-axis represents the importance of each feature and Y-Axis the names of the features.

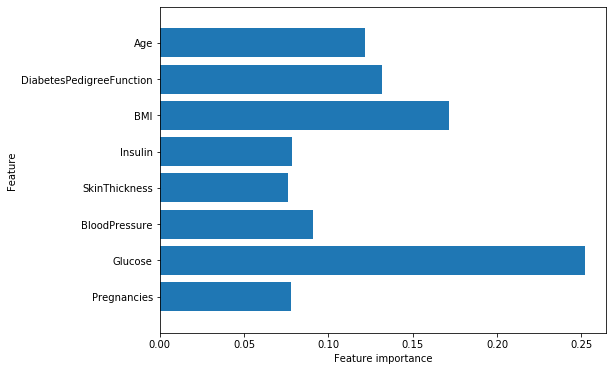


Figure 4: Feature Importance Plot for Random Forest

# VI. CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Random Forest and Decision Tree classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

**DISCUSSION:**

We can improve the model in the following ways:

1. Adding more data

Using huge volumes of data to train the models will help improve performance, leading to more accurate models.

1. Enhancing the quality of data by collecting data from various domains
2. Implementing it with gui to make it easier to read for everyone and improving accessibility to everyone in mobile without any issue.
3. Using domain specific knowledge and human insight can help in choosing better features for the prediction task under consideration.
4. Multiple Approaches:

Using different classification algorithms for the news classification task can help in choosing the best solution. Hence, using multiple approaches and comparing their performance on validation dataset can help in choosing the correct approach to solve the classification problem under consideration