**DIABETES PREDICTION USING MACHINE**

**LEARNING CLASSIFICATION MODEL**

***A project report submitted by***

# DEVARAPALLI VENUGOPAL REDDY

# (Y21CSE279018)

***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

***in***

# COMPUTER SCIENCE AND ENGINEERING



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**KRISHNA UNIVERSITY COLLEGE OF ENGINEERING AND TECHNOLOGY**

**Rudravaram, Machilipatnam – 521001,Andhra Pradesh**

**INDIA**

**CONTENTS**

**Abstract**  **3**

1. **Introduction 4-6** 
   1. Introduction
   2. Objectives
   3. Motivation
   4. Overview of the Project
   5. Chapter wise Summary

1. **Data Analysis**  **7-13**
   1. Structure of Data
   2. Parameters Implemented
   3. Exploratory Data analysis
   4. Histogram plot of data
   5. Box plot of data
   6. Distribution of classes

1. **Implementation**  **14-20** 
   1. Splitting of Dataset
   2. Feature Scaling
   3. Implementing Machine Learning Algorithms
      1. Correlation Heat Map
      2. Logistic Regression Model
      3. Support Vector Machine Model (Svc)
      4. Decision Tree Model
      5. K-Neighbours Classifier Model
2. **Test results** **21**
3. **Conclusions and Further Scope**  **22**

# ABSTRACT

Diabetes is one of the chronic diseases that causes blood sugar levels to rise. If diabetes is left untreated and undiagnosed, it can lead to complications. The time-consuming identification process leads to a patient's referral to a diagnostic Centre and consultation with a doctor. Predictive analytics in healthcare is a difficult challenge, but it can eventually assist physicians in making timely decisions about a patient's health and condition based on data. The emergence of machine learning methods solves this crucial issue.

The aim of this project is to create a model that can reliably predict the accuracy of diabetes in patients. Dataset splits into three then classification techniques are implemented. Training Dataset, Dataset sample that is used to fit the model. Validation Dataset, Dataset sample that is used for hyper tuning the parameters, and comparing the accuracy and error rates of the model performance between using the training dataset and the validation dataset. Testing Dataset, Dataset sample that is used to test the model performance (predictive power).

To detect diabetes at an early stage, this project employs machine learning classification algorithms: Logistic Regression, Gaussian Naive Bayes, K Neighbours, SVM, Decision tree, Random Forest, Bagging Classifier, Ada Boost Classifier and Gradient Boosting Classifier are implemented. The Pima Indians Diabetes Database (PIDD) is used in the experiments. The National Institute of Diabetes and Digestive and Kidney Diseases provided the results. The dataset's purpose is to diagnose whether a patient has diabetes using diagnostic measures included in the dataset. Various measures like Precision, Accuracy, Specificity, and Recall are measured over classified instances using Confusion Matrix.

The accuracy of the algorithms used are compared and discussed. The study's comparison of the various machine learning techniques shows which algorithm is better suited for diabetes prediction. Using machine learning methods, this project aims to assist doctors and physicians in the early detection of diabetes.

## INTRODUCTION

## 1.1 INTRODUCTION

Various classification strategies are used in the medical field for classifying data into different classes. Diabetes is a condition that affects the body's ability to produce the hormone insulin, which causes carbohydrate metabolism to become irregular and blood glucose levels to increase. High blood sugar is a common symptom of diabetes. If diabetes is not treated, it can lead to a variety of complications. Diabetic ketoacidosis and nonketotic hyperosmolar coma are two significant complications. Diabetes is considered a severe health problem in which the amount of sugar in the blood cannot be regulated. Diabetes is influenced by a variety of factors such as height, weight, genetic factors, and insulin, but the most important factor to remember is sugar concentration. The only way to avoid problems is to identify the problem early. This dataset comes from the ‘National

Institute of Diabetes and Digestive Diseases’ Pima Indians Diabetes Database (PIDD). Several constraints were taken from the massive database.



The dataset is divided into three sections, after which classification techniques are used. The training dataset is a sample of the dataset that is used to match the model. Validation Dataset, a dataset sample used for fine-tuning parameters and comparing model output accuracy and error rates between the training and validation datasets. Testing Dataset is a sample of a dataset that is used to assess the model's output.

Various machine learning techniques are implemented. Confusion matrix is obtained and is compared with all classification algorithms. This comparison of the various machine learning techniques shows which algorithm is better suited for diabetes prediction. Correlation between parameters and the best accuracy score using various supervised machine learning algorithms is obtained.

## 1.2 OBJECTIVES

* Since a decade, the number of people diagnosed with diabetes has risen significantly. The current human lifestyle is the primary cause of diabetes rise.
* Main objective of this project is to analyze the data, and see if it is possible to gleam any further information from the data to determine correlation between parameters and diabetes.
* The second is to attempt to get the best accuracy score using various supervised learning machine learning algorithms. To find out which algorithm is able to best predict whether a person has diabetes or not based on this dataset.
* The accuracy of the algorithms used are compared and discussed. The study's comparison of the various machine learning techniques shows which algorithm is better suited for diabetes prediction. Using machine learning methods, this project aims to assist doctors and physicians for predicting whether a person has diabetes or not.

## 1.3 MOTIVATION

The current human lifestyle is the primary cause of increasing diabetes. The three types of errors that may occur in today's medical diagnosis method:

1. The false-negative form, in which a patient is diabetic in fact but test results show that he or she does not have diabetes.

2.The false-positive type. In this type, a patient in reality is not a diabetic patient but test reports say that he/she is a diabetic patient.

3. The third type is an unclassifiable type in which a system cannot diagnose a given case. This happens because of insufficient knowledge extraction from past data, a given patient may get predicted in an unclassified type.

However, in fact, the patient must predict whether he or she will be diabetic or non-diabetic. Such diagnostic errors can result in unnecessary treatments or no treatments at all when they are needed. To prevent or mitigate the magnitude of such an effect, a machine learning algorithm must be used to build a framework that provides reliable results while reducing human effort.

### 1.4 OVERVIEW OF PROJECT

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of a large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes or not. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decisions about disease status.

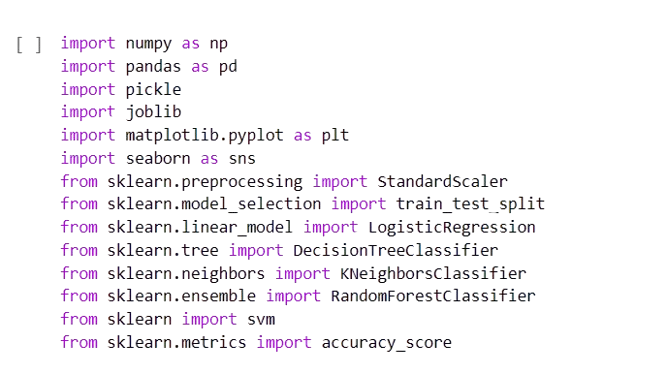
**1.5 CHAPTERWISE SUMMARY**

The first chapter is an introductory chapter, which gives an overview of the project. It includes four divisions - introduction, objectives, motivation, overview and chapter wise summary. The second chapter is data analysis, where the dataset is analyzed and studied for further classifications. Third chapter deals with the different machine learning models used. To detect diabetes at an early stage, this project employs machine learning classification algorithms: Logistic Regression, Gaussian Naive Bayes, K Neighbours, SVM, Decision tree, Random Forest, Bagging Classifier, AdaBoost Classifier and Gradient Boosting Classifier are implemented. The last chapter gives an elaborate idea about the results of different models. Let’s get to know more about the dataset in the upcoming chapter.

# 2.DATA ANALYSIS

## 2.1 STRUCTURE OF DATA

The dataset is originally from the Kaggle data repository. 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. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age etc.



**Fig:- Importing Libraries**

**Fig:-** Importing libraries to implement various machine learning for classification techniques.



**Fig:- Loading Dataset**

**Fig:-**Loading the dataset to understand data structure.



**Fig:- Shape of dataset**

Fig:- represent total number of rows and columns in Dataset

## 2.2 PARAMETERS IMPLEMENTED

Pregnancies: No. of times pregnant

Glucose: Plasma glucose concentration for 2 hours in an oral glucose tolerance test.

Blood Pressure: Diastolic blood pressure (mm Hg). It is the bottom number in blood pressure tests, and is the pressure in the arteries when the heart rests between beats. A normal diastolic blood pressure is < 80 mm HG.

Skin Thickness: Triceps skin fold thickness (mm). Studies have been conducted, with conclusions that there are associations between people with thicker skin and diabetes.

Insulin: 2-Hour serum insulin (mu U/ml). Insulin is a hormone made by the pancreas that allows your body to use sugar (glucose) from carbohydrates in the food that you eat for energy or to store glucose for future use. A high insulin level is associated with diabetes.

BMI: Body mass index (weight in kg/ (height in m) ^2) Range of BMI:

BMI < 18.5 - underweight

18.5 < BMI < 24.9 - ideal weight

25 < BMI < 29.9 - overweight

29.9 < BMI - obese

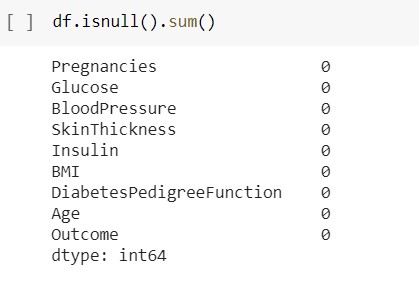
Diabetes Pedigree Function: It is a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject.

Results show that a person with a higher pedigree function tested positive and those who had a lower pedigree function tested negative.

Age: Age of the patient in years

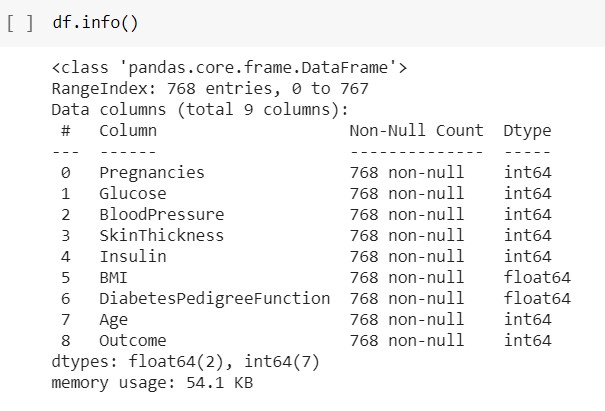
Outcome: The target column which we are interested in finding out. 1 - diabetic, 0 - non-diabetic

## 2.3 EXPLORATORY DATA ANALYSIS



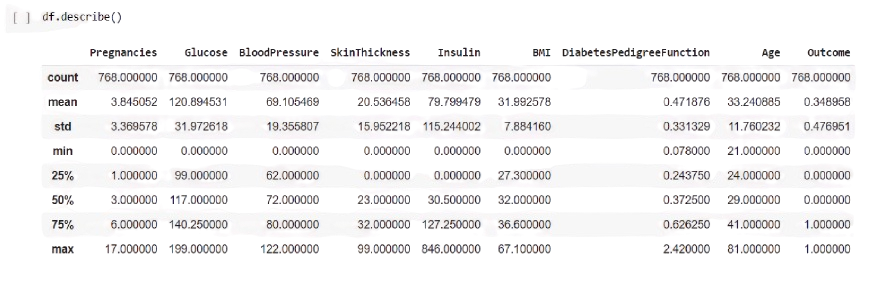
**Fig 2.3.1 Exploratory Data Analysis**

**Fig 2.3.1**, is analyzing the dataset and checking any missing values.



**Fig:- Dataset Information**

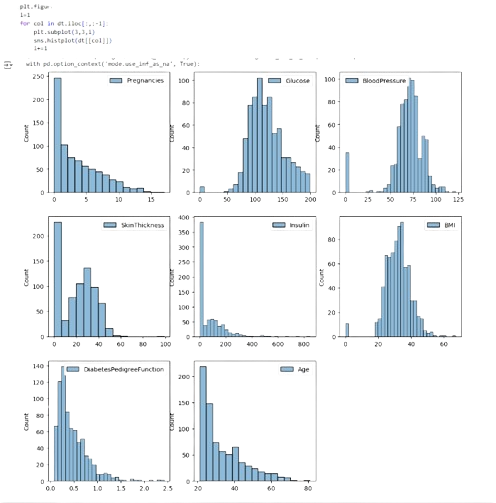
**Fig:-**Dataset information's are checked.



**Fig:- Calculating Mean, Count, Min, Max and Standard Deviation.**

## 2.4 HISTOGRAM PLOT OF DATA

The below histogram plots give a high-level view of the bucket distribution of the dataset parameters. At first glance, most of them appear to be positively skewed, with Glucose and Blood Pressure with the closest distribution to a normal distribution. Outcome is a bimodal distribution which is to be expected.



**2.5 BOX PLOT OF DATA**

plt.figure(figsize=(12,12))

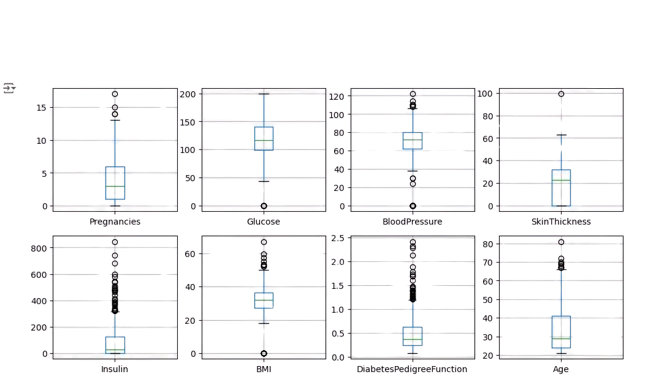
i=1

for col in dt.iloc[:,:-1]:

    plt.subplot(4,4,i)

    dt[[col]].boxplot()

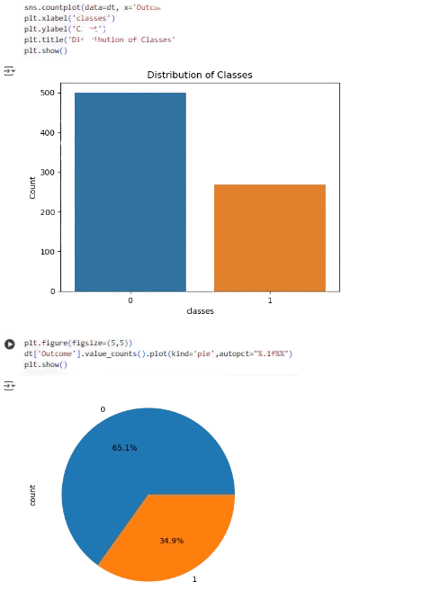
    i+=1



**2.6 DISTRIBUTION OF CLASSES USING PIE CHART & BAR CHART**

**Pie Chart:** A pie chart is ideal for illustrating the relative proportions or percentages of different classes within a dataset. In the case of diabetes detection prediction, a pie chart could be used to show the distribution of outcomes, such as the percentage of patients classified as diabetic versus non-diabetic. Each segment of the pie represents a class (e.g., diabetic or non-diabetic), and the size of each segment corresponds to its proportion of the whole dataset. This visualization method helps stakeholders quickly grasp the balance or imbalance between different predicted outcomes.

**Bar Chart:** A bar chart is useful for comparing quantities across different categories. In the context of diabetes detection prediction, a bar chart could display the absolute counts of each class, such as the number of individuals predicted as diabetic and non-diabetic. Each bar represents a class, and the height of the bar indicates the frequency or count of instances in that class. This allows for a straightforward comparison of class distributions and can highlight any disparities or trends in prediction outcomes.



# 3. IMPLEMENTATION

## 3.1 SPLITTING OF DATASET (TRAINING/VAILDATION/TESTING)

The splitting of the dataset for validation and testing.

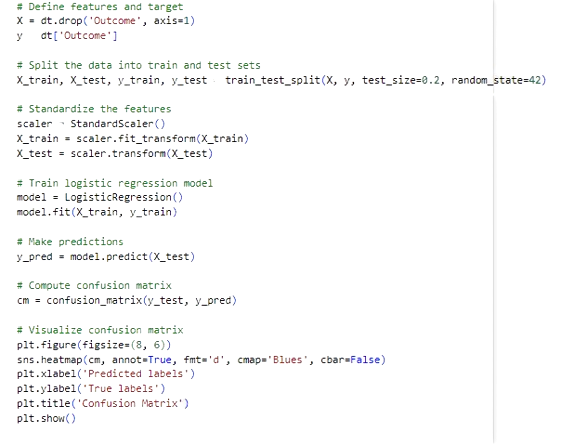
Training Dataset: Dataset sample that is used to fit the model.

Validation Dataset: Dataset sample that is used for hyper tuning the parameters, and comparing the accuracy and error rates of the model performance between using the training dataset and the validation dataset.

Testing Dataset: Dataset sample that is used to test the model performance (predictive power).

## 3.2Feature Scaling

Here StandardScaler() is used to perform feature scaling. This will retain the mean and the standard deviation of the sample distribution of the data set, and reuse it to transform the X\_train and X\_test subsequently. I try to reuse the mean and standard deviation obtained from the training set and apply it to the testing set as well. Standardizing data after data splitting is to prevent data leakage from test dataset into train dataset.



## 3.3 Implementing Machine Learning Algorithms

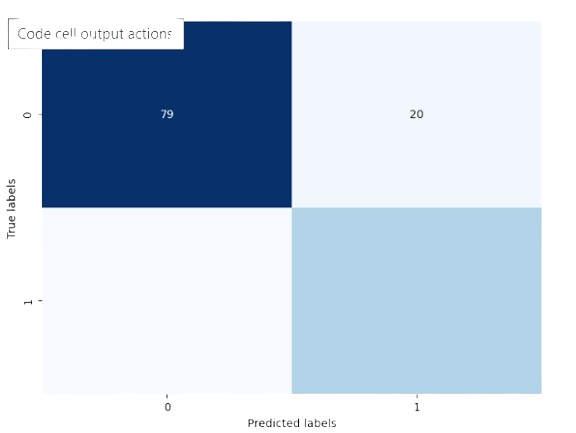
Different machine learning algorithms to try and classify the pima Indian diabetes dataset. First a confusion matrix function is formed.

Accuracy: (TP+TN)/All

Recall: TP/(TP+FN)

Precision: TP/(TP+FP)

Specificity: TN/(TN+FP)



**3.3.1 CORRELATION HEATMAP**

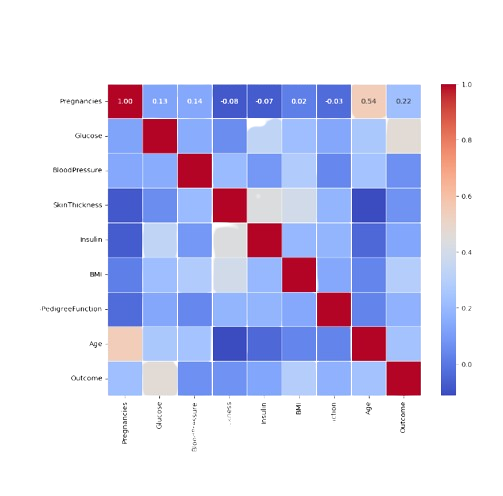
plt.figure(figsize=(10, 8))

correlation\_matrix = dt.corr()

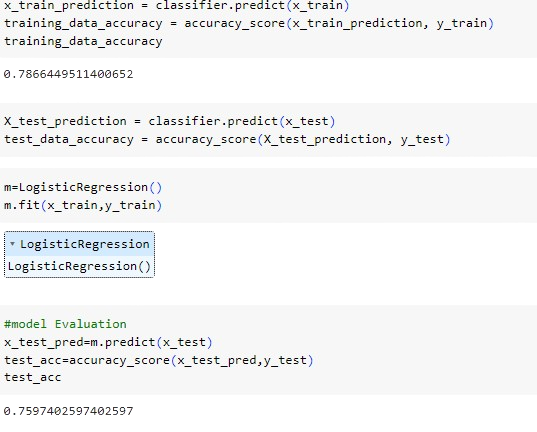
sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)

plt.title('Correlation Heatmap')

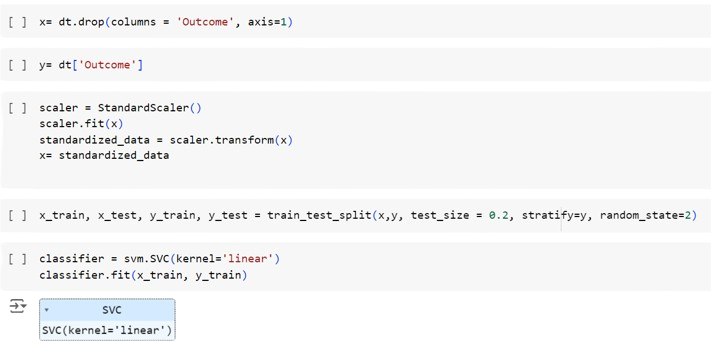
plt.show()



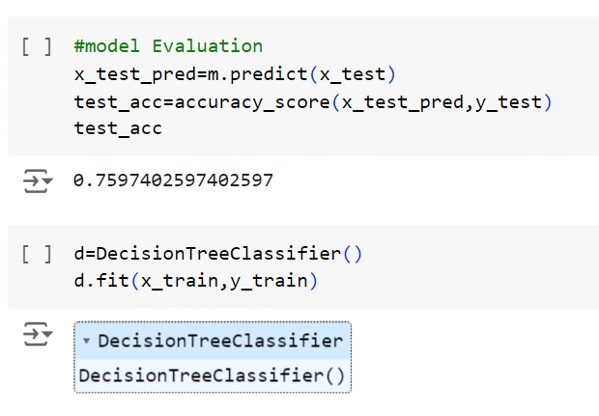
### 3.3.2 Logistic Regression Model



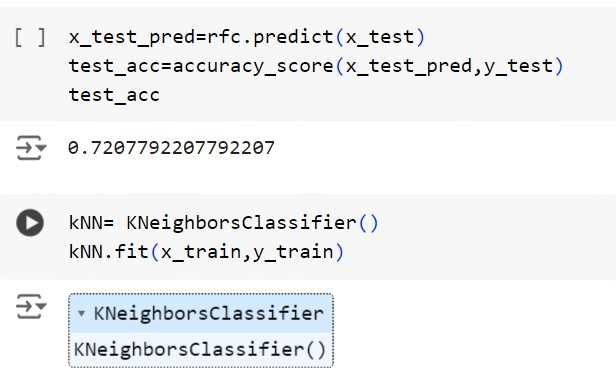
**3.3.3 Support Vector Machine Model**

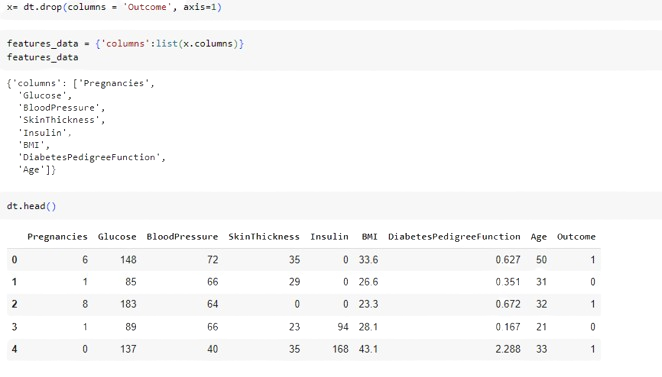


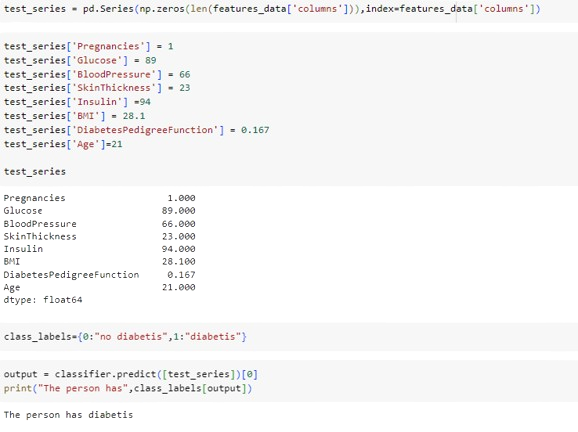
**3.3.4 Decision Tree Model**



**3.3.5 K Neighbours Classifier Model**







## 4.TEST REULTS ANALYSIS

Finally, we have trained our models, and summarized table of the metrics of the various models.

Objectives were,

1. To attempt to see if it is possible to glean any further information from the data to determine correlation between parameters and diabetes.
2. To attempt to get the best accuracy score using various supervised learning machine learning algorithms.

For the first objective, based on the hypothesis test, we can tell that glucose levels are positively correlated to a person having diabetes, but we are not able to confirm if there is causality. For the second objective, based on the comparison between the various algorithms used, Random Forest seems to produce the best results to me.

The aim of this project is to create a model that can reliably predict the accuracy of diabetes in patients. The main aim of this project is 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, KNN, Random Forest, Decision Tree, Logistic regression and Gradient Boosting classifiers are used. A machine learning algorithm must be used to build a framework that provides reliable results while reducing human effort.

The test accuracy of the various models is generally within the same range, from approximately 73% to 81%. Based on Accuracy and Recall score, overly the Random Forest Classifier produced the best results.

## CONCLUSION AND FUTURE SCOPES

* Machine learning has the great ability to revolutionize the diabetes prediction with the help of advanced computational methods.

* Detection of Diabetes in its early stage is the key for treatment.

* The technique may also help researchers to develop an accurate and efficient tool that will reach at the table of clinicians to help them make better decisions about the disease.

* More parameters and factors would be involved in the future scope of this project.

* The accuracy will increase even more when the parameters increase.

* Using traditional techniques and algorithms, we can enhance the accuracy by improving the data.