Prediction of Diabetes and Reduction of Type II Error using Ensemble Learning

Thesis submitted in partial fulfilment of requirements for the degree of

Master of Technology in Computer Technology of
Computer Science and Engineering Department of
Jadavpur University

by

Medha Bhowmik
Registration No. 154184 - of 2020-2021
Exam Roll No. - M6TCT23011

under the supervision of

Dr. Anasua sarkar
Assistant Professor

Department of Computer Science and Engineering Jadavpur
University
Kolkata, West Bengal, India 2023

Certificate from the Supervisor

This is to certify that the work embodied in this thesis entitled **” Prediction of Diabetes and Reduction of Type II Error using Ensemble Learning”** has been satisfactorily completed by **Medha Bhowmik** (Registration Number **154184** of **2020−21**; Class Roll No. **002010504018**; Examination Roll No. **M6TCT23011** . It is a bona-fide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata for partial fulfilment of the requirements for the awarding of the **Master of Technology in Computer Technology** degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2020−23.

Dr. Anasua Sarkar
Assistant Professor,
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

Prof. Nandini Mukhopadhyay
Head,
Department of Computer Science and Engineering, Jadavpur
University.

Prof. Ardhendu Ghoshal
DEAN,
Faculty of Engineering Technology Jadavpur University

Certificate of Approval

This is to certify that the thesis entitled **Prediction of Diabetes and Reduction of Type II Error using Ensemble Learning** is a bona-fide record of work carried out by **Medha Bhowmik** (Registration Number 154184 of 2020–21; Class Roll No. 002010504018; Examination Roll No. **M6TCT23011** in partial fulfilment of the requirements for the award of the degree of **Master of Technology in Computer Technology** in the **Department of Computer Science and Engineering, Jadavpur University**, during the 2020 to 2023. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose of which it has been submitted.

Examiners:

(Signature of The Examiner)

(Signature of The Supervisor)

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that the thesis entitled **Prediction of Diabetes and Reduction of Type II Error using Ensemble Learning** contains literature survey and original research work by the undersigned candidate, as a part of his degree of **Master of Technology in Computer Technology** in the **Department of Computer Science and Engineering, Jadavpur University**. All information has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Medha Bhowmik

Examination Roll No.: M6TCT23011

Registration No.: 154184 of 2020−21

Thesis Title: Prediction of Diabetes and Reduction of Type II Error using Ensemble Learning

Signature of the Candidate:

ACKNOWLEDGEMENT

I am pleased to express my gratitude and regards towards my Project Guide **Dr. Anasua Sarkar**, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University, without whose valuable guidance, inspiration and attention towards me, pursuing my project would have been impossible.

Last but not the least, I express my regards towards my friends and family for bearing with me and for being a source of constant motivation during the entire term of the work.

Medha Bhowmik
MTCT Final Year
Exam Roll No. - M6TCT23011
Registration. No. - 154184 of 2020−21
Department of Computer Science and Engineering, Jadavpur University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

Table 1: Abbreviations

LR	Logistic Regresssion
MLP	Multi-layer Perceptron
NB	Naive Bayes
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

NOTATION

English Symbols

Fm	Loss Function
μ	mean

ABSTRACT

Machine learning techniques are used in applications as a routine strategy for analysing the vast amount of available data and extracting relevant knowledge and information to support the main decision-making processes. Diabetes is a common, and fatal syndrome, that affects people all over the world. It is characterised by hyperglycemia brought on by irregularities in insulin secretion, which in turn would cause the glucose level to rise irregularly. Diabetes has significantly worsened in recent years, particularly in emerging nations like India. Machine Learning Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using machine learning analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high.This is primarily caused by the inconsistencies in people’s eating and living routines. As a result, research into the early detection and classification of this fatal disease has increased during the past ten years.There are many clustering and classification approaches that can be used to visualise temporal data in order to spot trends and manage diabetes. This proposed model is based on classifier comparison of various machine learning approaches and reducing the Type II error.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes mellitus is a category of metabolic illnesses in which a person has excessive blood sugar levels due to either insufficient insulin production by the body or improper cell response to the body's production of insulin. Polyuria (frequent urine), polydipsia (increased thirst), and polyphagia (increased hunger) are common in diabetic patients [1, 2]. The 3 Types of Diabetes:

Type I Diabetics

This kind of diabetes results from insufficient insulin production by the body. The terms insulin-dependent diabetes, juvenile diabetes, and early-onset diabetes are also used to describe this kind of diabetes. Typically, type I[3] diabetes strikes before the age of 40, that is, in adolescence or early adulthood. For the remainder of their lives, people with type 1 diabetes will require insulin injections. Additionally, they must maintain correct blood glucose levels by performing routine blood tests and adhering to a certain diet.

Type II Diabetics

In Type II Diabetes, the body does not produce enough insulin or the cells in the body display insulin resistance. Some people may be able to control their type II[4] diabetes symptoms by losing weight, following a healthy diet, doing plenty of exercise, and monitoring their blood glucose levels. However, type II diabetes is typically a progressive disease – it gradually gets worse – and the patient will probably end up having to take insulin, usually in tablet form. Being overweight, physically inactive and eating the wrong foods all contribute to our risk of developing type II diabetes. The risk of developing Type II diabetes also increases with age [5], [6].

Gestational Diabetes

This particular type of diabetic harms women while pregnancy. Some women have very high blood glucose levels, and because their bodies can't create enough insulin to get all of the glucose into their cells, their blood glucose levels continue to rise over time. The majority of women with gestational diabetes may manage their condition with diet and exercise. 10 to 20 percent of them will require the use of a blood glucose-regulating drug. Obstetric problems are more likely if gestational diabetes is un-diagnosed or not under control[7].

1.2 Objective of Project

In this work, we have tried to compare various machine learning classifier based on Confusion Matrix[5] and we have reduced the Type II error using Ensemble Classifier.

- Confusion Matrix : The performance of the classification models for a certain set of test data is evaluated using a matrix called the confusion matrix. Only after the true values of the test data are known can it be determined. Although the matrix itself is simple to understand, some of the terminology used in connection with it might be. It is also referred to as an error matrix since it displays the errors in the model performance as a matrix.

  The above table has the following cases:

  True Negative(TN): Model has given prediction No, and the real or actual value was also No.

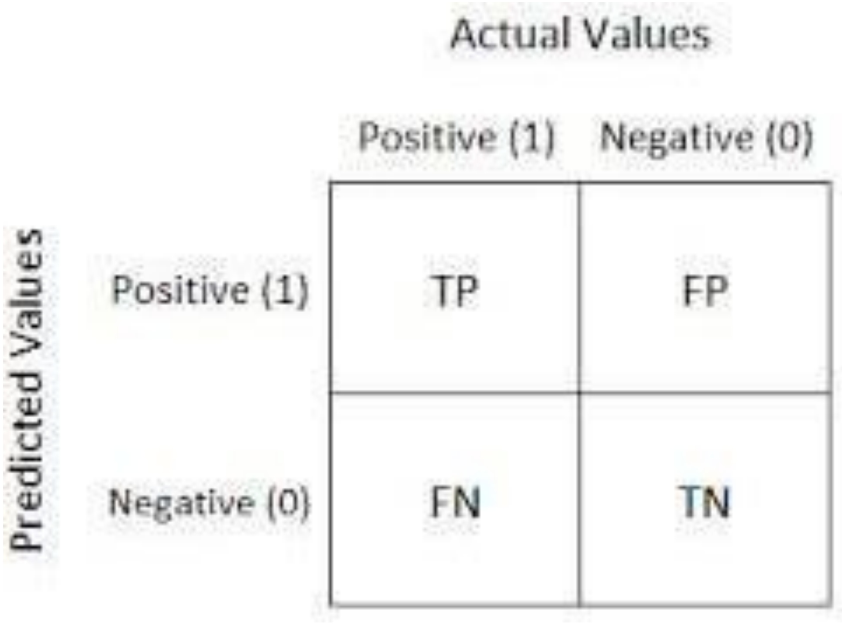
  True Positive(TP): The model has predicted yes, and the actual value was also true.

  False Negative(FN): The model has predicted no, but the actual value was Yes, it is also called as Type-II error.

  False Positive(FP): The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

- Calculations using Confusion Matrix:

  Using confusion matrix, we may calculate the model's accuracy as well as other properties.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig – 1 (Confusion Matrix)

Accuracy: It defines how often the model predicts the correct output. Accuracy can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers.

Accuracy = $\frac{(TP+TN)}{(TP+FP+FN+TN)}$

Error rate: It specifics how frequently the model makes incorrect predictions. The value of error rate can be calculated as the number of incorrect predictions to all number of the predictions made by the classifier.

Error Rate = $\frac{(FP+FN)}{(TP+FP+FN+TN)}$

Precision: It can be determined as the number of accurate outputs produced by the model or as the proportion of correctly expected positive classes that actually occurred.

Error Rate = $\frac{TP}{(TP+FP)}$

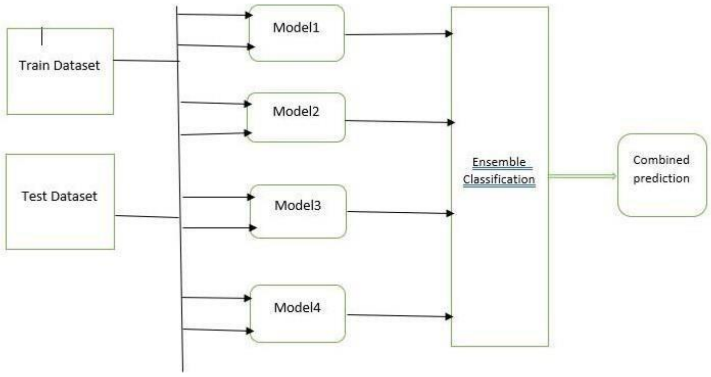
Recall: It is referred to as the percentage of total positive classes that a model accurately predicted.

Error Rate = $\frac{TP}{(TP+FN)}$

F-measure: It is hard to compare two models that have low precision but good recall, or vice versa. F-score can hence be used for this purpose. This score enables us to simultaneously evaluate recall and precision. If the recall and precision are equal, the F-score is at its highest.

F-measure = $\frac{2*recall*precision}{(recall+precision)}$

- Ensemble Learning Method: By combining many models, ensemble learning[7] enhances machine learning outcomes. In comparison to using a single model, this strategy enables the generation of greater prediction performance. It can improve the predictive accuracy. The key objective of the ensemble methods is to reduce bias and variance.

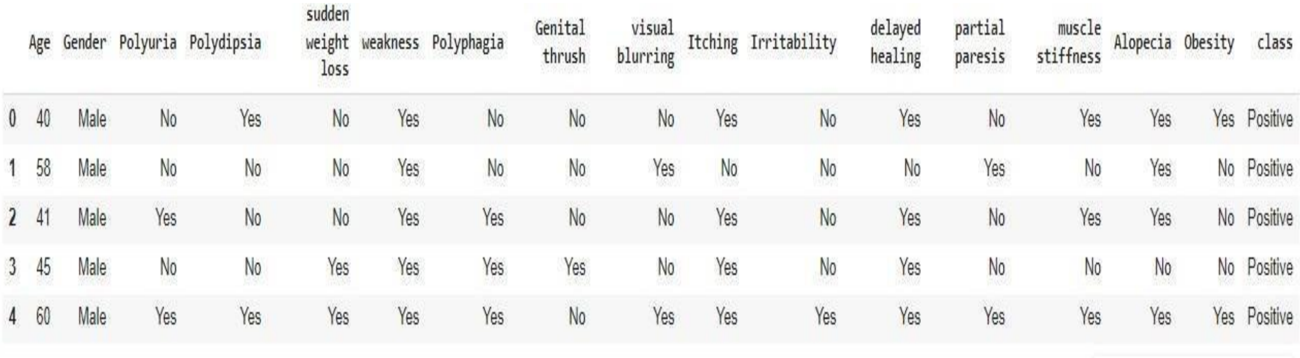


Fig—2(Ensemble Learning Model)

In this model, we have used Weighted Voting classifier to get the best solution. Voting ensembles are an ensemble technique that is used to train many machine learning models before combining the output of each model’s predictions.

Dataset Description :

There are a total of 520 records in the sample, and each of them are characterized by 17 parameters.



	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Table 1 (Dataset Description)

In that records which were identified as a diabetic patient if they met with at least one of the following criteria: plasma fasting glucose 126 mg/dL, serum glucose 200 mg/dL, glycohemoglobin 6.5The dataset is pre-cleaned and the columns have Boolean values except ’Age’ and ’Gender’ columns. So for that we have encoded the datas by using LabelEncoder. As these are boolean categorical values we cannot use pandas.DataFrame.describe to infer about the descriptive statistics including those that summarize the central tendency, dispersion and shape of a dataset’s distribution, excluding NaN values.

The analysis of diabetes data is a challenging issue because most of the medical data are nonlinear, non-normal, correlation structured, and complex in nature[8]. Medical imaging, such as that used to diagnose cancer, coronary artery disease, and stroke, has been dominated by ML-based systems [9–12]. Furthermore, feature selection techniques and classifiers can also be employed with ML-based systems. Additionally, it helps in the proper diagnosis of diabetes, with the best classifier serving as the key to determining an individual’s risk for developing the disease. There were various ML-based classifier are used to classify and predict of diabetic disease like Random Forest, MLP , Decision Tree, Logistic Regression and so on [13-17]. Based on accuracy we have combined the best classifiers to get the best result and we have reduced the type II error.

CHAPTER 2

Literature Review

2.1 Literature Description

One of the most challenging tasks for medical practitioners is determining the kind of diabetes a patient has. However, examining numerous variables at the time of diagnosis can occasionally produce unreliable results. As a result, interpreting and categorising diabetes is a very difficult undertaking. The healthcare sector has benefited greatly from recent technological developments, particularly those involving machine learning techniques. The literature has offered a variety of methods for classifying diabetes.

Zou et al. [18] studied on the diagnosis of diabetes dataset. The dataset was taken from the hospital physical examination data in Luzhou, China. The dataset contained 14 attributes and consisted of 220,680 patients. Among them, 151,598 patients were diabetic and 69,082 were control. They applied PCA and minimum redundancy maximum relevance to reduce the dimensional and also K5 cross-validation protocol adopted to examine the data. [19] Ahuja et al. used PID dataset in his study. The dataset consisted of 768 patients and 10 attributes. The dataset had some missing values and they were replaced the missing values by median. LDA was used to extract the feature selection. They applied five classification algorithms as: SVM, multi-layer perceptron (MLP), LR, RF, and DT. They showed that LDA with MLP based classifier gave the highest classification accuracy of 78.70 percent. [20] A proposed model by Singh and Singh is a stacking-based ensemble method for predicting type 2 diabetes mellitus. They used a publicly available PIMA dataset from the UCI Machine Learning Repository. The stacking ensemble used four base learners, i.e., SVM, decision tree, RBF SVM, and poly SVM, and trained them with the bootstrap method through cross-validation. However, variable selection is not explicitly mentioned and state-of-the-art comparison is missing. [21] Kumari et al. presented a soft computing-based diabetes prediction system that uses three widely used supervised machine learning algorithms in an ensemble manner. They used PIMA and breast cancer datasets for evaluation purposes. They used random forest, logistic regression, and naïve Bayes and compared their performance with stateof-the-art individual and ensemble approaches, and their system outperforms with

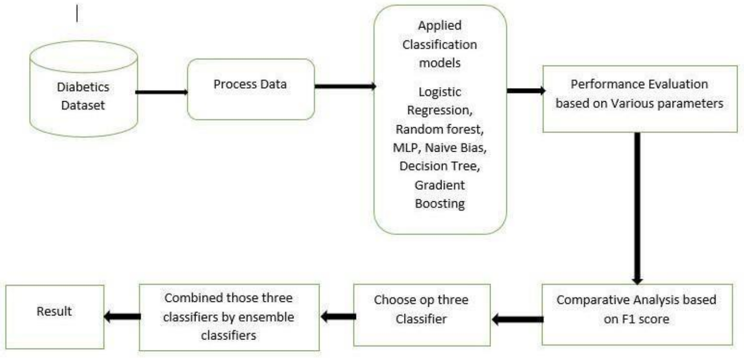
79 percent accuracy. [22] A proposed model by Mohapatra et al.applied MLP and found that MLP gave the classification accuracy of 77.50 percent. [23] Pei et al. also applied DT and gave 94.20 percent classification accuracy. [24] A research by Ratna Patil,

Sharavari Tamane presents an experimentalstudy of several algorithms which classifies Diabetes Mellitus data effectively. The existing algorithms are analyzed thoroughly to identify their advantages and limitations. [25] Hussain and Naaz presented a thorough review of machine learning models presented during 2010–2019 for diabetes prediction. They compared traditional supervised machine learning models with neural networkbased algorithms in terms of accuracy and efficiency. [26] Shruti Iyar analysed step by step Diabetics classification b y KNN classifier. [27] Yang et al. focused on exercise therapy which plays a significant role in treating diabetes and its associated side effects. Specifically, they discovered cytokines which gives a novel insight into diabetes control, but the sequence is still under study. [28] Olexandr Shmatko, Olha Korol, Andrey Tkachov, Vasyl Otenko used the Recursive Feature Elimination method to improve the prediction rate. Their research work is to select the bestclassifier for the diabetes prediction information system. Various machine learning classification algorithms are used to predict diabetes in a patient, such as Linear Regression(LR), K-Nearest Neighbor (KNN), Decision Tree (DT). We have proposed a classification model by comparing different classifiers by Weighted Voting Ensemble Learning. The main goal is to reduce Type II Error we found that with a trade off of accuracy we are able to achieve it.

CHAPTER 3

Methodology

We have Used Weighted Voting Ensemble method. We have analyzed several classifiers like Random Forest, Gradient Boosting, Decision Tree, MLP, Logistic Regression and Gaussian Naive Bias. We have classifies based on confusion matrix. and ensambled top three ones with soft voting for getting a robust decision and lower bias.



Diabetics Dataset
Process Data
Applied Classification models
Logistic Regression, Random forest, MLP, Naive Bias, Decision Tree, Gradient Boosting
Performance Evaluation based on Various parameters
Comparative Analysis based on F1 score
Choose op three Classifier
Combined those three classifiers by ensemble classifiers
Result

Fig3: Proposed Model Diagram

3.1 Brief Description of Algorithm Used:

Below Classifiers we have used to analyse:

Logistic Regression:

Logistic regression is a linear model for classification as opposed to regression. It is sometimes referred to as the logit regression, the log-linear classifier, or maximumentropy classification (MaxEnt). In this model, logistic regression is used to simulate outcomes of a single trial that are mathematically described. It is a fundamental model that explains output variables with two options, and it may be expanded to predict disease categorization[30]. Suppose there are N input variables where their values are indicated by m1,m2,m3,...,mN. Let us assume that the P probability of that an event will occur and

1- P be a probability that event will not occur. Logistic regression model is given by

$$log(\frac{p}{1-p}) = logit(P) = \beta0 + \beta1m1 + \ldots\ldots + \beta NmN$$

Gradient Boosting:

A class of algorithms known as boosting are capable of transforming weak learners into strong learners. The primary idea behind boosting is to fit a series of weak learners—models, like small decision trees, that are just marginally better than random guessing—to weighted versions of the data. The final prediction is then created by combining the guesses using a weighted majority vote (for classification) or a weighted sum (for regression). A generalisation of boosting to any differentiable loss function is gradient tree boosting. It can be applied to classification and regression issues. Gradient Boosting creates the model step-by-step.

$$Fm(x) = Fm - 1(x) = \gamma mhm(x)$$

Decision Tree: A decision tree is a hierarchical model used in decision support that depicts decisions and their potential outcomes, incorporating chance events, resource expenses, and utility. This algorithmic model utilizes conditional control statements and is non-parametric, supervised learning, useful for both classification and regression tasks. The tree structure is comprised of a root node, branches, internal nodes, and leaf nodes, forming a hierarchical, tree-like structure. decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves. Below are the terminologies of Decision Tree -

Root Nodes: It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.

Decision Nodes: the nodes we get after splitting the root nodes are called Decision Node

Leaf Nodes: the nodes where further splitting is not possible are called leaf nodes or terminal nodes

Sub-tree: just like a small portion of a graph is called sub-graph similarly a sub-section of this decision tree is called sub-tree.

Pruning – is nothing but cutting down some nodes to stop over fitting.

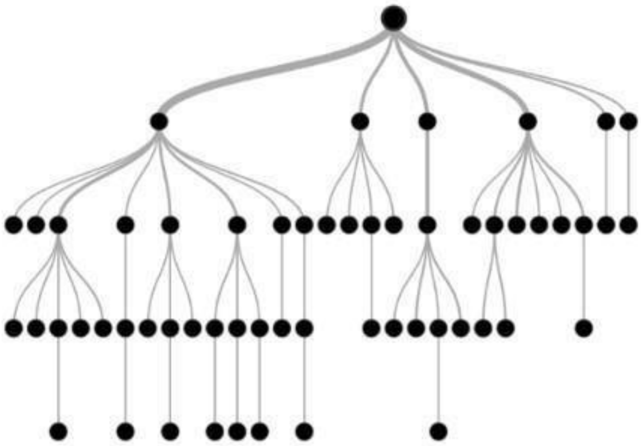


Fig- 4 (Decision Tree)

Multi-Layer Perceptron:

Multi-layer perception is also known as MLP. It is made up of dense, completely connected layers that may change any input dimension into the desired dimension. A neural network with numerous layers is referred to as a multi-layer perception. In order to build a neural network, we combine neurons so that some of their outputs are also their inputs. A multi-layer perceptron contains one input layer with one neuron (or node) for each input, one output layer with one node for each output, and any number of hidden layers with any number of nodes on each hidden layer. Below is a schematic illustration of a Multi-Layer Perceptron (MLP).

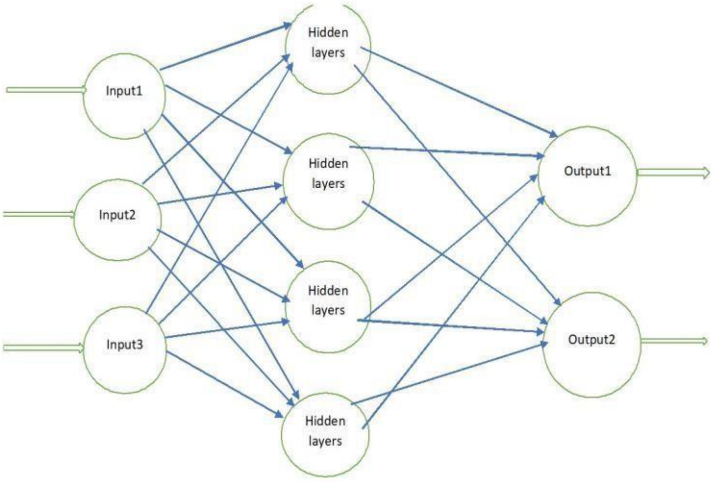


Fig – 5 (Multi-Layer Perceptron)

Random Forest:

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

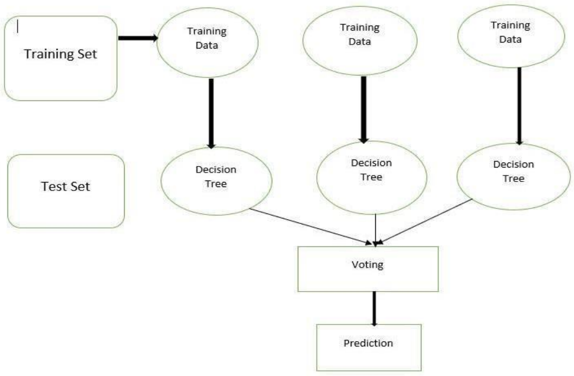


Fig- 6 (Random Forest)

Gaussian Naive Baise:

Gaussian Distribution is also called Normal Distribution. A statistical model known as the normal distribution is used to represent the statistical distributions of continuous random variables seen in nature. Its bell-shaped curve serves as the definition of the normal distribution. A mean and a standard deviation are the two key components of a normal distribution. The standard deviation is the "width" of the distribution around the mean, and the mean is the average value of a distribution.

It is important to know that a variable (X) that is normally distributed, is distributed continuously (continuous variable) from $-\infty < X < +\infty$ and the total area under the model curve is 1.

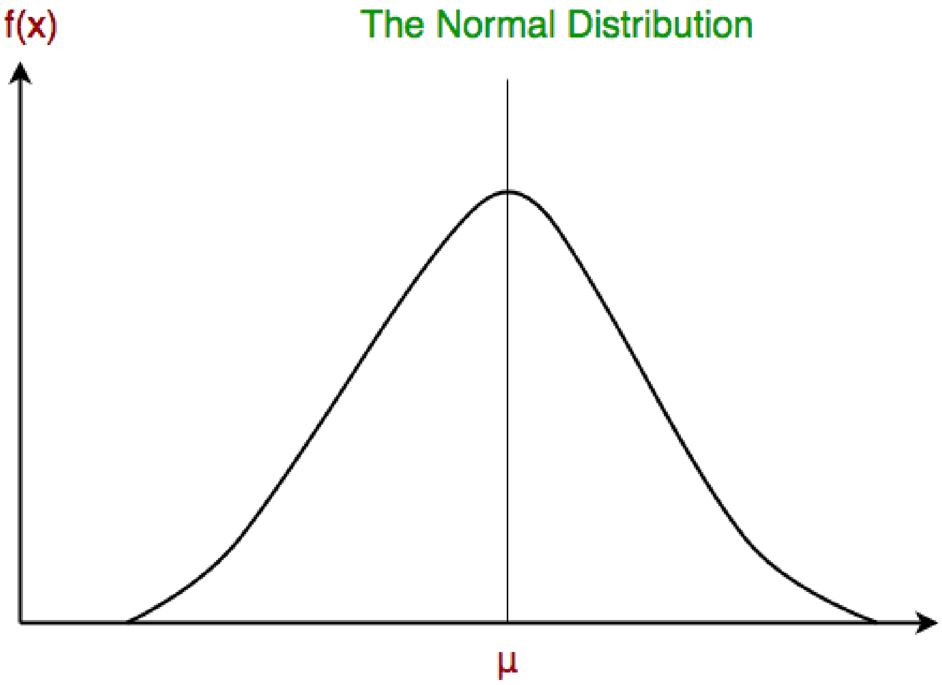
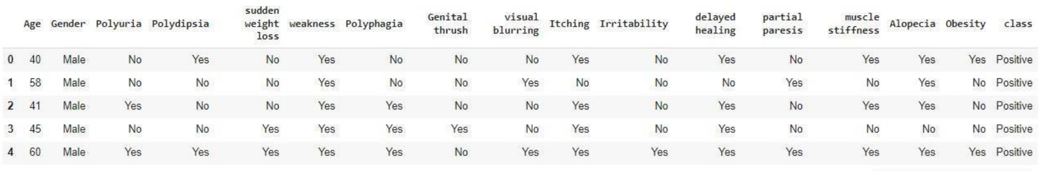


Fig Gaussian Naive Baise:

3.2 Experimental Analysis :

There are a total of 520 records in the sample, and each of them are characterizedby 17 parameters.

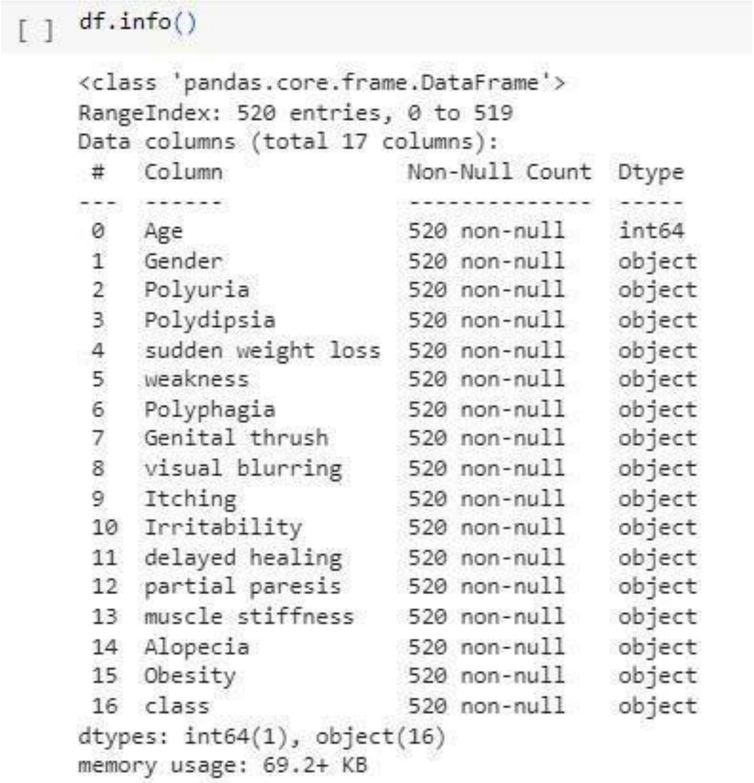


	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Below steps are followed for predicting early stage diabetics:

Data wrangling :

Data Wrangling is a prerequisite step for machine learning and analytic purposes. It involves reorganizing, mapping, and transforming data from its raw, unstructured form into a more usable format. It is also known as data cleaning or munging. According to mythology, this wrangling consumes up to 80 percent of the time spent by analytical specialists, leaving only 20 percent of their time for investigation and modelling. But the dataset is pre-cleaned and there is no missing data. As a result, it is quite an easy step here. Below we have checked the data type of columns and if there is any null value.

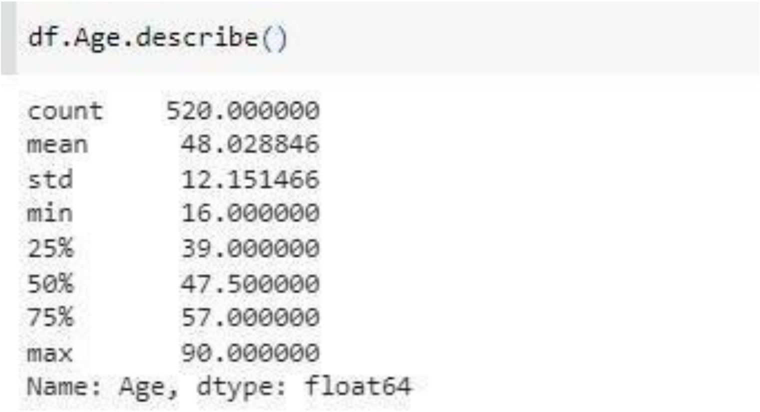


```
[ ]  df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 520 entries, 0 to 519
     Data columns (total 17 columns):
      #   Column              Non-Null Count  Dtype
     ---  ------              --------------  -----
      0   Age                 520 non-null    int64
      1   Gender              520 non-null    object
      2   Polyuria            520 non-null    object
      3   Polydipsia          520 non-null    object
      4   sudden weight loss  520 non-null    object
      5   weakness            520 non-null    object
      6   Polyphagia          520 non-null    object
      7   Genital thrush      520 non-null    object
      8   visual blurring     520 non-null    object
      9   Itching             520 non-null    object
      10  Irritability        520 non-null    object
      11  delayed healing     520 non-null    object
      12  partial paresis     520 non-null    object
      13  muscle stiffness    520 non-null    object
      14  Alopecia            520 non-null    object
      15  Obesity             520 non-null    object
      16  class               520 non-null    object
     dtypes: int64(1), object(16)
     memory usage: 69.2+ KB
```

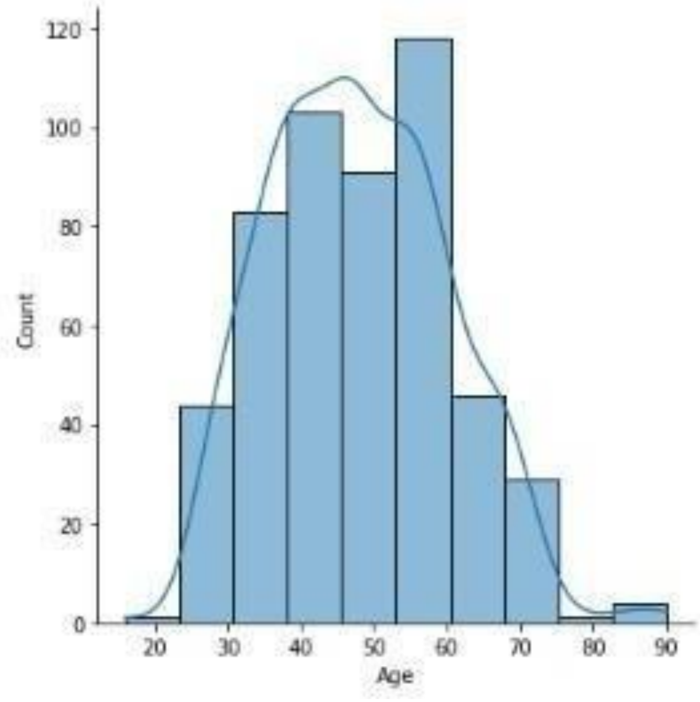
Exploratory Data Analysis :

In this section, we will try to infer about the trends in the dataset using data visualization and statistics. We can see that every columns without age consists of Boolean values. So at first we need to encode them by using LabelEncoder. As these are boolean categorical values we cannot use pandas.DataFrame.describe to infer about the descriptive statistics including those that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values. Rather we can analyze the frequency of attributes over different age groups and gender. below shows the descriptive statistics of the Age Column and Distribution of patient Age.



```
df.Age.describe()
```

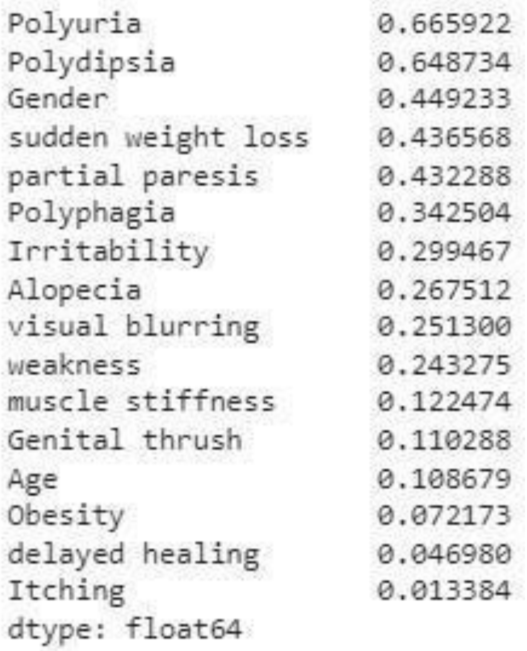
```
count    520.000000
mean      48.028846
std       12.151466
min       16.000000
25%       39.000000
50%       47.500000
75%       57.000000
max       90.000000
Name: Age, dtype: float64
```


120
100
80
60
40
20
0
Count
20
30
40
50
60
70
80
90
Age

Data Pre processing and Feature Important analysis :

In this process we will find he most important features of this data set by: Finding out the pearson correlation between features and Class Feature importance techniques to reduce computational cost.

Correlation analysis : we are able to determine which characteristics attributes are most important in determining the class. So that we can lighten the strain on machine learning models, we can also eliminate relatively trivial characteristics. We have taken the Pearson correlation coefficient. We took the absolute values of the correlating features to find out top 5 features.



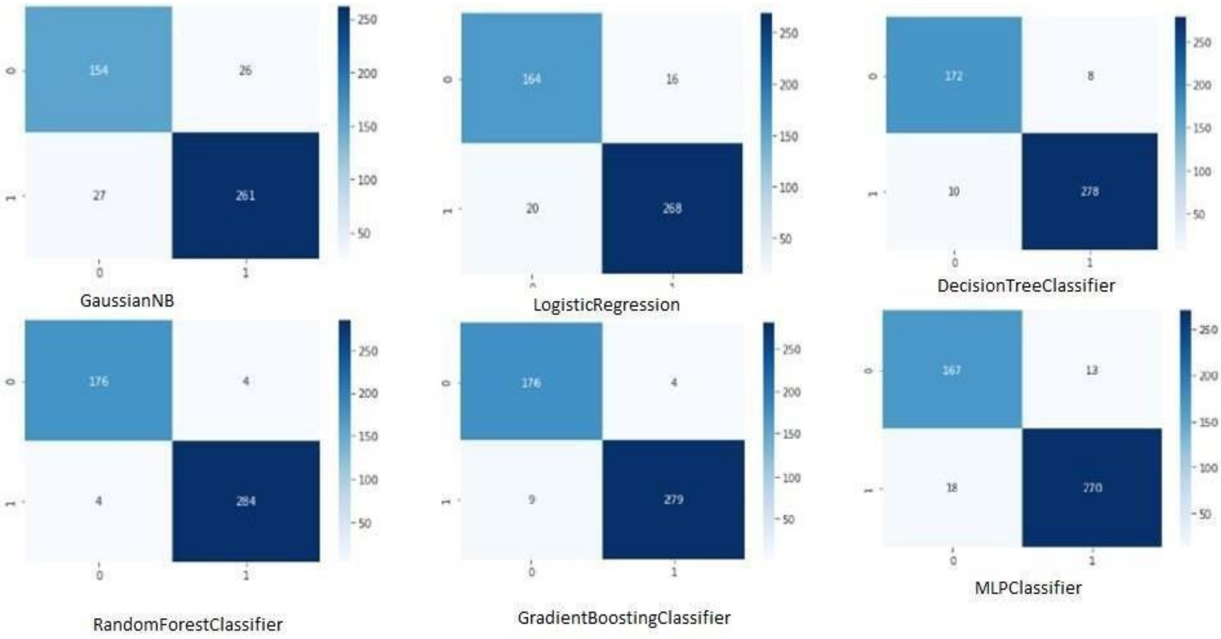
```
Polyuria              0.665922
Polydipsia            0.648734
Gender                0.449233
sudden weight loss    0.436568
partial paresis       0.432288
Polyphagia            0.342504
Irritability          0.299467
Alopecia              0.267512
visual blurring       0.251300
weakness              0.243275
muscle stiffness      0.122474
Genital thrush        0.110288
Age                   0.108679
Obesity               0.072173
delayed healing       0.046980
Itching               0.013384
dtype: float64
```

From that we can see that Age is one of the most important factors but it is not quite correlated with the target variable Class. So, we should not rely on only one method in determine important features for predicting the target.

Predictive Analysis : In predictive analysis, we need to come to an strategy for making a robust classifier to classify the likelihood of a person having early stage diabetics by using new features. In this project, our goal should be minimizing false positives even if it reduces over all accuracy.
We have created a DummyRegression model that allows to create a very simple model that we can use as a baseline to compare against our actual model. We got results 0.61 so we have developed a classifier for more than 62 percent accuracy and with lowest false negative rate.

Developing classifiers:

We have analyzed the accuracy, F1 score and other parameters based on Confusion Matrix fordifferent classifiers. Below diagram represent the confusion matrix for all the classifiers.



GaussianNB

LogisticRegression

DecisionTreeClassifier

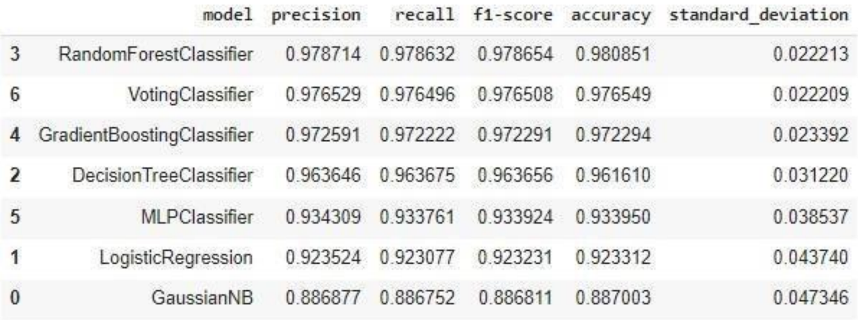
RandomForestClassifier

GradientBoostingClassifier

MLPClassifier

Fig Confusion matrix for different classifiers

We have combined the classifiers using Voting classifier and from the below table we can determined that Random Forest, Gradient Boosting and Decision tree are the top three.

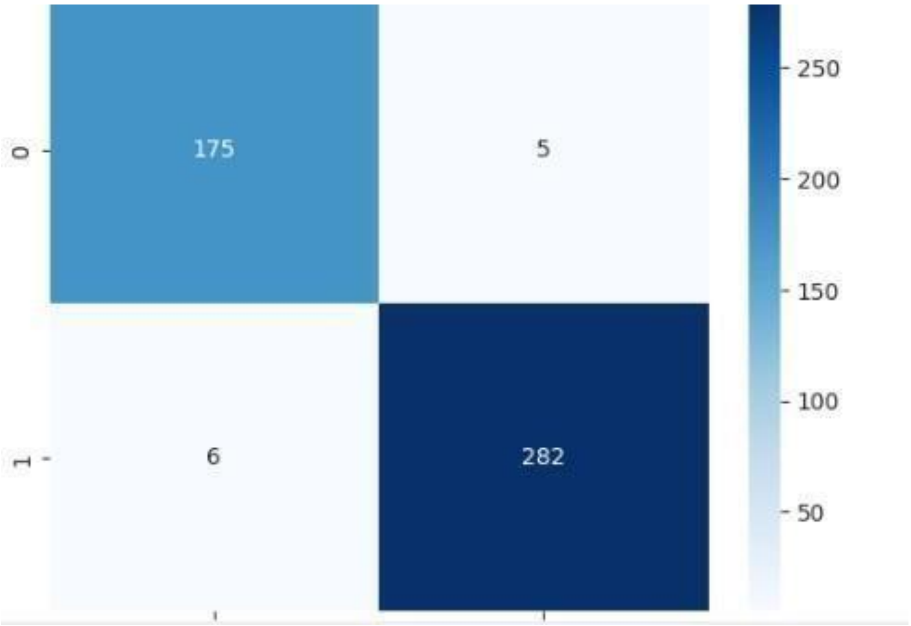


	model	precision	recall	f1-score	accuracy	standard_deviation
3	RandomForestClassifier	0.978714	0.978632	0.978654	0.980851	0.022213
6	VotingClassifier	0.976529	0.976496	0.976508	0.976549	0.022209
4	GradientBoostingClassifier	0.972591	0.972222	0.972291	0.972294	0.023392
2	DecisionTreeClassifier	0.963646	0.963675	0.963656	0.961610	0.031220
5	MLPClassifier	0.934309	0.933761	0.933924	0.933950	0.038537
1	LogisticRegression	0.923524	0.923077	0.923231	0.923312	0.043740
0	GaussianNB	0.886877	0.886752	0.886811	0.887003	0.047346

3.1 Developing a Voting Classifier by ensembling top three models :

Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.Although there are a seemingly unlimited number of ensembles that you can develop for your predictive modeling problem, there are three methods that dominate the field of ensemble learning. So much so, that rather than algorithms per se, each is a field of study that has spawned many more specialized methods.

From the above dataframe and confusion matrix we can see that RandomForest, GradientBoosting, and Decision Tree classifiers gives better F1-score and lower False Negatives. So, we can ensemble them for developing a robust classifier usign soft voting.Ensemble methods are techniques that create multiple models and then combine themto produce improved results. Below shows the confusion matrix for Ensembled Classifier.

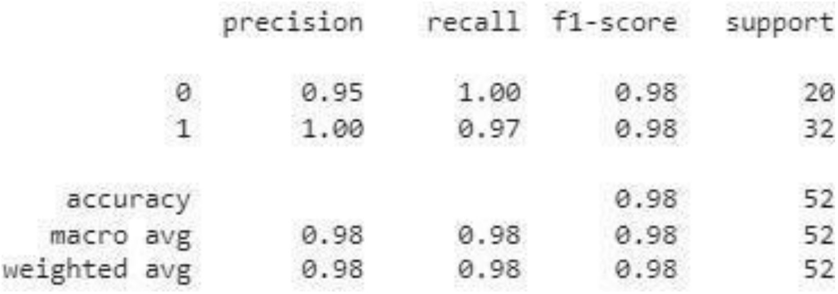


-

3.2 Discussions :

We can see that we can achieve a significant accuracy using Random Forest Classifier but with the help of other two classifiers we are able to reduce type II error. Our main goal was to reduce Type II error and we can see that with a combination of accuracy we are able to achieve it.

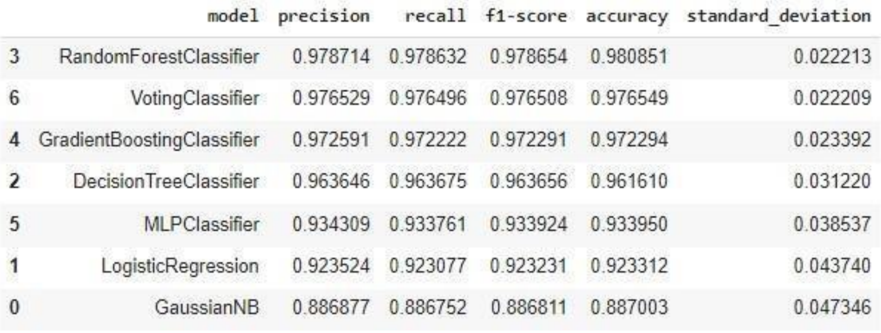
From the confusion matrix, we can see that we have reduced the type II error, by using weighted voting classifier. and we have also performed the model in validation dataset.



```
              precision    recall  f1-score   support

           0       0.95      1.00      0.98        20
           1       1.00      0.97      0.98        32

    accuracy                           0.98        52
   macro avg       0.98      0.98      0.98        52
weighted avg       0.98      0.98      0.98        52
```



	model	precision	recall	f1-score	accuracy	standard_deviation
3	RandomForestClassifier	0.978714	0.978632	0.978654	0.980851	0.022213
6	VotingClassifier	0.976529	0.976496	0.976508	0.976549	0.022209
4	GradientBoostingClassifier	0.972591	0.972222	0.972291	0.972294	0.023392
2	DecisionTreeClassifier	0.963646	0.963675	0.963656	0.961610	0.031220
5	MLPClassifier	0.934309	0.933761	0.933924	0.933950	0.038537
1	LogisticRegression	0.923524	0.923077	0.923231	0.923312	0.043740
0	GaussianNB	0.886877	0.886752	0.886811	0.887003	0.047346

CHAPTER 4

Conclusion and Future Work:

One of the most serious health issues is the early detection of diabetes. Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. The system architecture and classifier for an information system are proposed in this work, system with a high degree of accuracy for predicting diabetes. We have applied many machine learning algorithms for get the best result. We decided the result based on confusion matrix and we obtained the precision, recall,f1-score,support from the matrix. Mostly all classifiers gave more than 70 percent accuracy. NB(88%),LR(92%), Decision Tree(96%),GB(98%),MLP(93%), Random Forest(98%). So we choose Random Forest as best classifier. But we found only Random forest was not able to reduce the Type II error, so for that using other two (Gradient Boosting and Decision Tree classifiers) we have reduced the Type II error Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes.

Reference

[1] K. Selvakuberan, et al., “An Efficient Feature Selection Method for Classification in Health Care Systems Using Machine Learning Techniques”, IEEE, pp. 8610-8615, 2011.

[2] M. Seera, et al., “A Hybrid Intelligent System for Medical Data Classification”, Expert Elsevier: Systems with Applications, vol. 41, pp. 2239-2249, 2014.

[3] T. Karthikeyan, et al., “An Intelligent Type-II Diabetes Mellitus Diagnosis Approach using Improved FP-growth with Hybrid Classifier Based Arm Research”, Journal of Applied Sciences, Engineering and Technology, vol. 11, no. 5, pp. 549-558, 2015.

[4] D. K. Karumanchi, et al., “Early diagnosis of Diabetes mellitus through the eye”, 2nd International Conference on Endocrinology, 2014.

[5] Sofia Visa, Brian Ramsay Confusion Matrix-based Feature Selection.

[6] Luis Fregoso-Aparicio, Julieta Noguez, Luis Montesinos & José A. García-García Diabetology & Metabolic Syndrome volume 13, Article number: 148 (2021), “Machine learning and deep learning predictive models for type 2 diabetes-“

[7] Daniela Mennickent 1, Andrés Rodríguez 2, Marcelo Farías-Jofré 3, Juan Araya 4, Enrique Guzmán-Gutiérrezm - Machine learning-based models for gestational diabetes mellitus prediction before 24-28 weeks of pregnancy.

[8] Maniruzzaman M, Kumar N, Abedin MM, Islam MS, Suri HS, El-Baz AS, Suri JS. Comparative approaches for classification of diabetes mellitus data: machine learning paradigm. Comput Methods Programs Biomed. 2017;152:23–34. [PubMed] [Google Scholar]

[9] Maniruzzaman M, Rahman MJ, Al-MehediHasan M, Suri HS, Abedin MM, El-Baz A, Suri JS. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst. 2018;42(5):92. [PMC free article] [PubMed] [Google Scholar]

[10] Srivastava SK, Singh SK, Suri JS. Healthcare text classification system and its performance evaluation: a source of better intelligence by characterizing healthcare text. J Med Syst. 2018;42(5):97. [PubMed] [Google Scholar]

[11] Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst. 2016;4(1):2. [PMC free article] [PubMed] [Google Scholar]

[12] Shakeel PM, Baskar S, Dhulipala VS, Jaber MM. Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. Health Inf Sci Syst. 2018;6(1):16. [PMC free article] [PubMed] [Google Scholar]

[13] Banchhor SK, Londhe ND, Araki T, Saba L, Radeva P, Khanna N, Suri JS. Calcium detection, its quantification, and grayscale morphology-based risk stratification using

machine learning in multimodality big data coronary and carotid scans: a review. Comput Biol Med. 2018;101:184–198. [PubMed] [Google Scholar]

[14] Bashir S, Qamar U, Khan FH. IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. J Biomed Inform. 2016;59:185–200. [PubMed] [Google Scholar]

[15] Zhao X, Zou Q, Liu B, Liu X. Exploratory predicting protein folding model with random forest and hybrid features. Curr Proteomics. 2014;11:289–299. [Google Scholar]

[16] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018;132:1578–1585. [Google Scholar]

[17] Ahuja R, Vivek V, Chandna M, Virmani S, Banga A. Comparative study of various machine learning algorithms for prediction of Insomnia. In: Advanced classification techniques for healthcare analysis; 2019. p. 234–257.

[18] N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster," Cluster Computing, vol. 22, no. 1, pp. 1–9, 2019.

[19] R. Ahuja, S. C. Sharma, and M. Ali, "A diabetic disease prediction model based on classification algorithms," Annals of Emerging Technologies in Computing, vol. 3, no. 3, pp. 44–52, 2019.

[20] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 1–22, 2020.

[21] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, 2021.

[22] S. K. Mohapatra, J. K. Swain, and M. N. Mohanty, "Detection of diabetes using multilayer perceptron," in Proceeding of the International Conference on Intelligent Computing and Applications, pp. 109–116, Springer, Ghaziabad, India, December 2019.

[23] A. B. Bazila, R. K. Priyadarshini, and P. Thirumalaikolundusubramanian, "Prediction of children diabetes by autoregressive integrated moving averages model using big data and not only sql," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 8, pp. 3510–3513, 2019.

[24] N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia, "Diabetes prediction using artificial neural network," Deep Learning Techniques for Biomedical and Health Informatics, Springer, Singapore, 2020.

[25] A. Hussain and S. Naaz, "Prediction of diabetes mellitus: comparative study of various machine learning models," in Proceeding of the International Conference on Innovative Computing and Communications, pp. 103–115, Springer, Delhi, India, January 2021.

[26] H. Temurtas, et al., “A Comparative Study on Diabetes disease Diagnosis using Neural Networks”, Elsevier: Expert Systems with Applications, vol. 36, 2009.

[27] Wang S, Li D, Song X, Wei Y, Li H. A feature selection method based on improved fisher’s discriminant ratio for text sentiment classification. Expert Syst Appl. 2011;38(7):8696–8702.

[28] M. N. Devi, et al., “An Amalgam KNN to Predict Diabetes mellitus”, IEEE, 2013

[29] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010;10(1):16–23.

[30] Diwani, Salim Amour, and Anael Sam."Diabetes forecasting using supervised learning techniques." Adv Comput Sci an Int J 3 (2014): 10-18.