Predictive Supervised Learning Model for Classification of Type 2 Diabetes Mellitus

M.S Roobini (roobinims@gmail.com)
Sathyabama Institute of Science and Technology https://orcid.org/0000-0003-1790-5132
M Lakshmi
SRM Institute of Science and Technology

Research Article

Keywords: Supervised Learning, Diabetes Mellitus, Machine Leaning, Ensemble Approach.

Posted Date: October 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1009663/v1

Predictive Supervised Learning Model for Classification of Type 2 Diabetes Mellitus

M.S.Roobini,
Assistant Professor, Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology
Chennai-600127, India
roobinims@gmail.com

Dr.M.Lakshmi
Professor, Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
SRM Nagar,Chennai-603203,India
laks@icadsindia.com

***Abstract*—There is a tremendous increase in severe cases of type 2 diabetes in the day today's life. Therefore, proper assessment of the disease is critical to saving society. Many prediction models help identify type 2 diabetes. At the same time, every model varies based on the performance measures. Various kinds of algorithms such as Decision Tree, Logistic Regression, KNN, Random Forest algorithm are applied to identify type 2 diabetes. At this juncture, used the implementation of type 2 Classification by AdaBoost algorithms, an ensemble approach. Here, the proposed methodology of the paper is to implement an ensemble approach of machine learning to receive a better efficiency compared to other existing algorithms for the classification of type 2 diabetes. When compared to all different algorithms, this ensemble approach shows an efficiency of 83%. The accuracy is calculated based on various performance measures.**

Keywords—Supervised Learning,Diabetes Mellitus,Machine Leaning,Ensemble Approach.

Introduction

1.1. Machine Learning approach in classification and diagnosis of Diabetes Mellitus:

In recent days, the number of people who got the risk of Type2 Diabetes is increasing rapidly day by day. Many different studies are there in recent days. There are many medical issues connected with this disease which leads literally to death. Different kinds of classification algorithms propose to diagnose the disease—ML helps in improving the algorithms and the methods that allow the machine to learn from the past. Machine Learning analyses the data and its uses at various learning processes by collecting different kinds of data. [1]. First, data pre-processing helps in fixing some data-related issues. And ML can solve many complex problems in the real world [2, 3]. Since there are many data to be processed, algorithms are applied according to the needs to analyse the data. Different steps apply to improve the efficiency of the model for overcoming the problem. Many medical-related issues are solved using various classification techniques.

1.2 About Diabetes:

Diabetes keeps your body from fittingly holding essentialness from the food you eat because you can't make insulin or can't utilize it accurately. There is no fix, yet medications grant you to manage your condition—an excessive amount of sugar in the blood classifies as "hyperglycemia" (high glucose). Diabetes has the possibility of leading to severe issues in tissues in the human body. Diabetes categorizations are Type1 and Type2 DM. Patients with type 1 diabetes are consistently more energetic, for the most part under 30 years old. Typical clinical indications lead to various health problems, and the severity will extend to the level of insulin

medications. Type 2 DM regularly affects middle-aged and older people, frequently having a connection with adiposity, dyslipidemia, etc.

1.3. Pre-Diabetes:

Diabetes grows progressively, so when you're in the pre-diabetes stage when your blood glucose level is higher than it ought to be, you might not have any side effects whatsoever. However, if an individual's glucose level stays high, they may start to build up specific side effects of type 2 diabetes, indicating continuous pee and expanded thirst. Regardless of whether you aren't overweight and don't have any of the danger factors, your PCP (primary care providers) must test for the blood glucose level that has to start when you're 45. That is a savvy activity because the danger of creating pre-diabetes (and along these lines, type 2 diabetes) increments with age. Since there are numerous potential confusions of diabetes (e.g., heart issues and nerve issues), it's a smart thought to be careful about identifying blood glucose anomalies early [5].

1.3.1 Diabetes Mellitus with category Type1:

Type1 of Diabetes disease depicts the pancreas producing enough insulin, which is needs for daily processes, a condition likewise as "insulin-subordinate diabetes mellitus." For this situation, the body produces next to no or no insulin by any means. Subsequently, day by day, insulin infusions are expected to keep glucose levels levelled out. Successive pee, unexpected weight reduction, strange thirst, steady craving, obscured vision, and sluggishness is normal side effects of this disease.

1.3.2 Type 2 kinds of Diabetes Mellitus:

Type-2-Diabetes Mellitus is set apart for opposing insulin response. For this situation, the body doesn't wholly react to insulin bringing about higher glucose levels. Therefore, stoutness, unfortunate eating routine, hypertension, and physical idleness are significant danger factors that lead to type 2 diabetes; this can occur as a developing disease of Mellitus with non-insulin deficiency.

2.Literature Survey:

In their work, had used five various predictive modeling techniques such as SVM-linear, radial basis function. K-NN algorithm, ANN, and dimensionality reduction for classification of type 1 and type2 diabetes. Pre-processing steps help to reduce the dimensions to make the process efficient. The accuracy result showed an efficiency of 89% based on the SVM application [1]. Machine Learning helps in the medical field. Sarwar et al. Application of Machine learning in the medical field plays a critical role. Various Machine Learning algorithms apply for medical-related records in the case of prediction. In their paper, the author had worked with methodologies to find out which gives the better efficiency in prediction. Among all the six algorithms applied, SVM showed better efficiency of 77% [2]. Other studies also proposed to find out the better efficiency of the algorithms for prediction. Muhammad et al., in their work, had worked with machine learning models such as regression, SVM, k-NN, random decision forest, boosting algorithm, etc. Among all the algorithms, random forest showed an accuracy of 89%. Gradient boosting algorithm showed an accuracy of 86% [3].

Probst et al., in their study, had proved that the result, which showed the expected error rate of the function, which is non-monotonous, as well as the problem of having a large number of datasets, is computationally feasible with a large number of error measures [4]. Insulin plays a very vital role in the cause of disease. Kanatand, his co-authors, described the sensitivity of insulin and the beta-cell function of Type2 patients. Two treatment options like lifestyle and pharmacologic treatments, are analyzed, leading to a better understanding of Type2 diabetes Mellitus [5]. In this paper, five different machine learning algorithms were applied, conducting experiments on 952 instances with questionnaire sections [6]. In their proposed method, Georga et al. mentioned glucose impact in diabetes patients and described the changes in the glucose level due to food habits and physical activities.

A suggestion is that the glucose level can either prevent or show a difference in the long-term complications in diabetes [7]. Several studies were done before in the insulin impact, which leads to Type2 Diabetes Mellitus. Basu et al. discussed various proportions combined with the insulin level, which showed that it increases diabetes mellitus. A survey for grown-up adults with type 2 diabetes people diagnosed as diabetic need insulin. Cohort studies have proved the need for HBA1c treatment, which needs more medical attention

[8]. Xia et al., in their research, clearly said about the glucose fluctuations in the brain functionality based on the glucose level. The proposed method compared the Type 2 patients with stable glucose range and the people with fluctuations in the glucose level who are Type 2 patients. It showed that the controls in the glucose level were leading to cognitive impairment [9]. Xia et al., in their study, described the connectivity of the interregional cooperation with the MRI in finding out the oxygen level-dependent functionality. The impact of the poor level of cholesterol is one of the main reasons for such health problems. Correlation analysis between the cholesterol level and type 2 patients' rates were examined [10]. Kengne et al., in their suggested work, said about the performances of various models for disease predictions. Clinical-based models for the predictive versions. They implemented the method with a follow-up of ten years, and they validate 12 prediction models; this helped identify the people affected by type 2 diabetes mellitus [11]. In their paper estimated a prediction model which can predict Type2 Diabetes based on the glucose level. The result of the predictions compared with the existing models. Many examinations find the error rate, which was then analyzed with improved grey GM efficiently [12].

Marder et al. in their work, have investigated the middle–age's people having type 2 diabetes and have a link with cognition activities. During this experiment, activation with different correlation levels with HbA1C and insulin levels shows that the people who have type2 diabetes had a problem in their brain activities [13]. Patil et al., in the paper, implemented different types of algorithms in the prediction of disease. The performance analysis identifies all the working algorithms [14,34,35]. Sun et al., in their paper, had concerned about various risk factors which lead to the increase of type2 diabetes severity level to cognitive impairment. It helps identify multiple diagnosis methods and gives the chance to know about the practical applications [15,33]. Baynes et al., in their paper, had discussed numerous pathological issues related to the growth in the severity of diabetes. The insulin variations in the impact of diabetes identify, and the suggestion given to people with T2DM should need insulin treatment to avoid various risk factors [16].

Goldenberg et al., in their paper, had discussed the chronic diabetes condition called hyperglycemia, which was associated with long-term cardiovascular difficulties. Various glucose level-related problems were also are find out, which leads to diabetes mellitus at a profound level [17]. Khawandanah et al., in their paper, discussed the hybrid type of diabetes, which was the combination of type1 and type2 diabetes with many physiological features, which leads to the increase in the risk factors in diabetes. Various levels of insulin resistance problems were also analyzed in their paper [18]. Tigga et al., in their report, discussed the factors which seem to be a risk in the prediction of diabetes two types with different kinds of questionnaire sections online and offline. Among all the classifier algorithms implemented, the Random model given a coherent prediction of type 2 diabetes [19]. Kaur et al. had done five different predictive models with ML models, and the Patterns tracked for various risk factors [20].

Viloria et al., in their paper, predicted diabetes using the algorithm vector support machine with the predisposition of diabetes, and it showed an accuracy of 99.2%. The implementation work examines various ethnic backgrounds [21]. Nai-Arun et al., in their paper, applied ensemble boosting techniques in their work and improved prediction methodologies. They worked with the concept of web applications also for the prediction of diabetes risk [22]. Allen et al., in their work, applied two datasets with the hospital data by implementing various ML algorithms like the random forest, etc. The ensemble approach is involved here, which makes a better improvement in the prediction accuracy. As a result, the accuracy improved up to 94% [23]. Saha et al., in their paper, implemented a new classification model with four phases with knowledge-based data analysis. A hybrid method with a neural network algorithm validated various test sequences compared to existing classifiers [24].

Kamal et al., in their research work, discussed the prediction of severe factors in data. Feature extraction with the FP growth algorithm.ID3 algorithms implementation with a training algorithm with a decision tree [25]. Chaki et al., in their paper, gave an overview of DM detection and various techniques that provide automatic detection systems with several screening procedures for the diagnosis of diabetes disease [26].Wangproposed a novel scheme for calculating indicators for diabetes, which also included feature extraction and extraction of glucose concentration of both types of diabetes. The accuracy of the result was almost 91% [27]. Finally, care et al. suggested various clinical recommendations that provided diabetes components and the goals and guidelines for good quality of diabetes. Different standards mean the care introduction of diabetes [28].

Sarwar et al. discussed the et al. predictive analysis in healthcare, which worked with six various machine learning algorithms. Performance measures for the results obtained showed the efficiency of the implemented algorithm—comparative analysis made to understand the efficiency of the proposed work [29]. Pranto et al. suggested a methodology in analyzing patients who have diabetes with the Pima Indian dataset. Performance investigation with evaluates decision tee algorithm, KNN, etc. Random forest and Nave Bayes showed good performance when compared to other algorithms [30]. Lai et al. had proposed a predictive model with Regression and Boosting algorithms, which showed that AROC evaluated to be 85% for Gradient Boosting with 73% efficiency [31]. Hassali et al. had done a study on twenty-seven papers which showed that the intervention of HbA1C in patients with diabetes mellitus. It also proved that the MBI has an impact on complications of the disease. It also discussed the pharmacist's contribution to the prognosis of the disease [32,33]. Ley et al. had done a study that proved the behavioral risk factor in type 2 diabetes. Identification of metabolic risk factors helps make a delay in the progression of the disease [34,35,36,37].

3.Proposed work

The process of the proposed work in Fig 1 shows the proposed creation of the module.

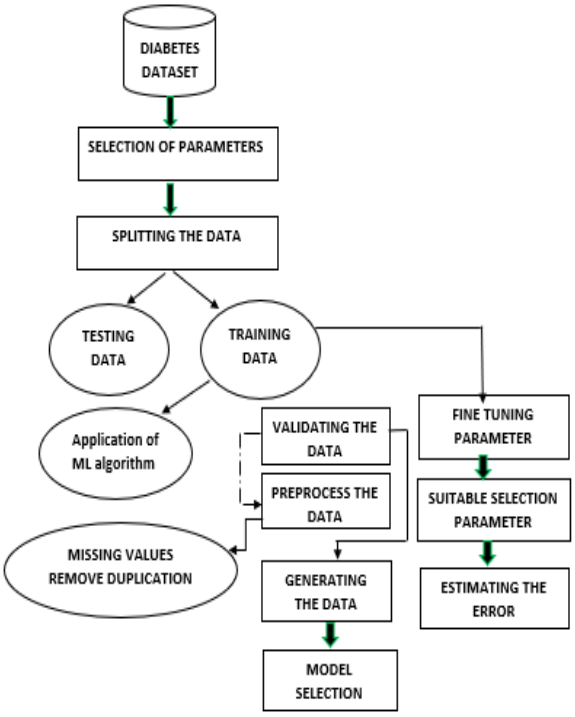

Fig 1: Proposed work Pictorial Representation

Description of the Architecture:

The above architecture diagram description is below. The dataset used is the diabetes dataset for the prediction of type 2 diabetes. Among all the parameters given, the best-suited parameters are selected, making the efficiency better for getting good prediction results. After all the data cleaning process is collected, the information splits for model training. The coherent model is finding out by Various ML models, which shows the better result in the prediction of Type Diabetes Mellitus.

Computational Explanations for Proposed Work of AdaBoost Algorithm:

Initialization:

For every observation assign, w0 as an initial value which equalsw0=1/M, m=1,2…., M, M->no: of words.

For every observation, if the prediction is correct, then w0 is increased, or else if the forecast is incorrect, then w0 is decreased.

Algorithm:

For n=1, 2……N do

Train weaker learners using w0 Distribution. Observations for greater weights given more priority.

a) Fit the classifier Gn(x) to the training data using weights w0.

b) Compute

$$err_n = \frac{\sum_{o=1}^{M} w_o J(y_o \neq G_n(x_o))}{\sum_{o=1}^{M} w_o} \text{ ---------- (1)}$$

c)Compute

$$a_n = \frac{\log(1 - err_n)}{err_n}$$

d)Set

$$w_{o<-} w_o . exp[\propto_n . J(y_o \neq G_n(x_o))] \text{----------(2)}$$

Repeat steps until, observations are perfectly predicted.

Output:

$$G(x) = sign\left[\sum_{n=1}^{N} \propto_n G_n(x)\right] \text{----------(3)}$$

Illustration of Ensemble approach:

An ensemble approach is followed, which improves the efficiency of the prediction of algorithms. The diagram representing the ensemble approach is below. Fig 2 shows the Pictorial representation of the ensemble approach. First, using the Feature Selection method, the most significant parameter which has an impact on Diabetes is identified and classification by the ensemble method. Then, the model implementation and the final ensemble approach are applied, and the category, by other machine learning algorithms, is compared to show the better efficient algorithm for disease prediction.

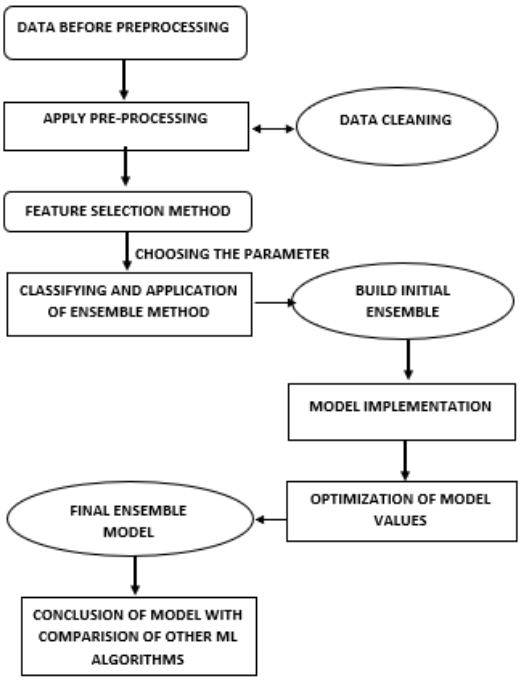

Fig 2: Diagrammatic Representation of Ensemble Approach

4. METHODS AND MATERIALS

A. Dataset Description:

Table 1 shows the illustration of the dataset used. The below given is the description of the dataset.

num_preg	Gluc-level	BP	Thickness	Insulin	BMI	DPG	Age	Outcome
6	148	72	35	0	33.6	0.6275	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Table 1: Dataset Used

Table 2 indicates the brief description of the dataset.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes PedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.6275	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Table 2: Brief Explanation of dataset

B. Pre-processing:

It is necessary to have accurate data when dealing with real-world problems. For good accuracy of the result, various pre-processing steps must clear the data. In this paper, the pre-processing carries three stages, such as checking the null values. If the null value is detected, then it has to be replaced with meaning values. The following process is to find out the duplication of records, which helps identify the repetition of the patients' record details.

The third process is to identify the outliers.

a) Checked null values

b) Checked duplication of data in rows in the dataset

c) Removed outliers

Finding Out the duplication of values:

There is the possibility of having duplication of data in the dataset, which occurs due to merging data or elements from various sources that are heterogeneous. Repetitions may lead to errors in prediction. Duplication of data is finding out in the pre-processing step. Fig 3 shows the duplication process of the values.

```
df1=df.duplicated()
df1=pd.DataFrame(df1,columns=['booli'])
print("Shape",df1.shape)#df1 is a dataframe which shows duplicate values
df1.booli.value_counts()
```

```
Shape (768, 1)

False    768
Name: booli, dtype: int64
```

Fig 3: Finding out the duplication of the values

Null Value detection:

After identification of null values, the places replace with either expected value or meaningful value. Thus, for example, the title of the null values is given below in Table 3.

df.isnull().sum()	
num_preg	0
Gluc-level	0
BP	0
Thickness	0
Insulin	0
BMI	0
DPG	0
Age	0
Outcome	0
dtype:int 64	

Table 3: Identification of Null values

Identification of Outliers:

Detection of Outliers is essential for getting a better outcome. Fig 5 indicates the picture before Outlier Detection.

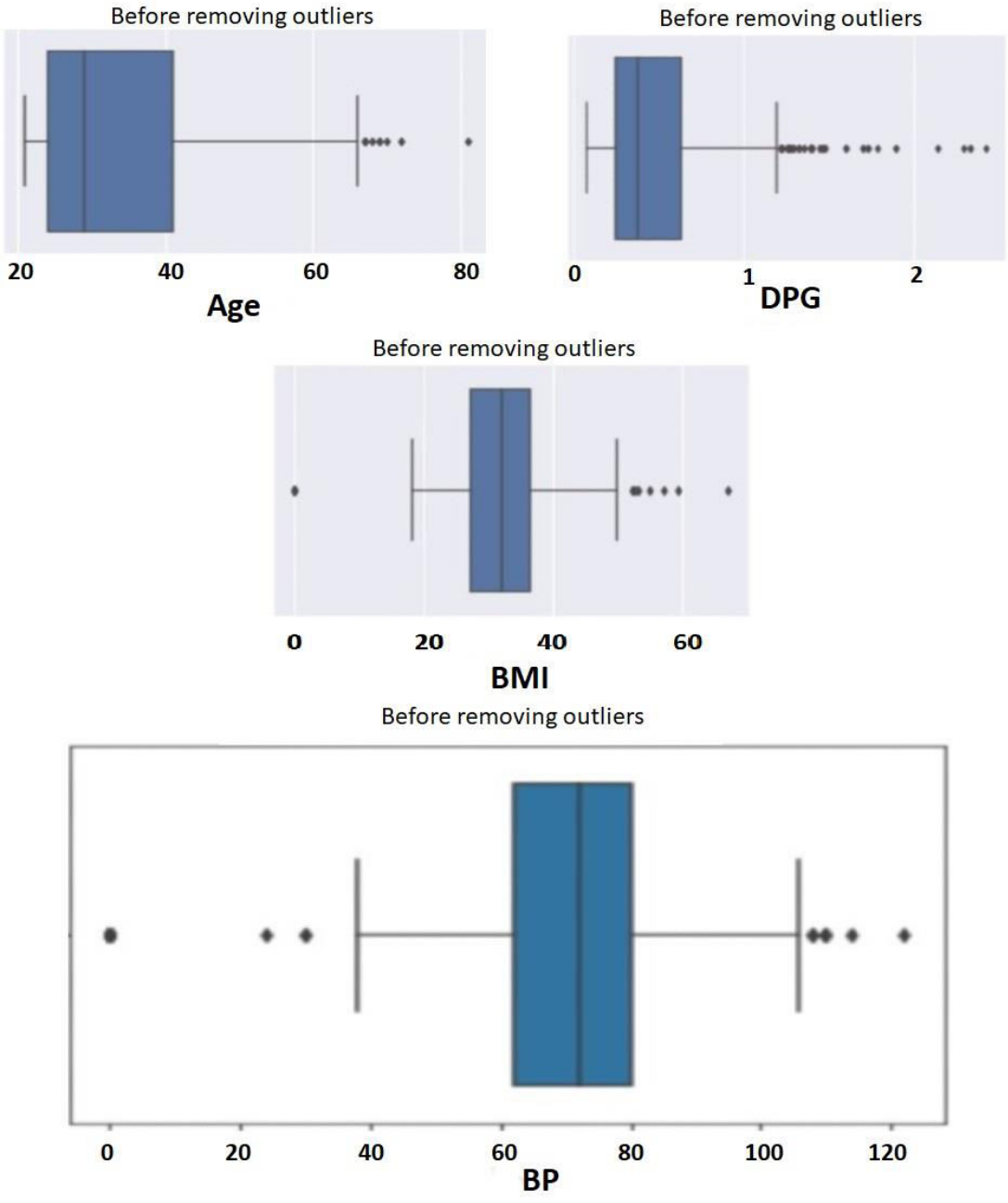

Fig 5: Before Outlier Detection

After removing outliers:

Fig 6 shows the output after Outlier Detection, which is needed to get the efficient outcome.

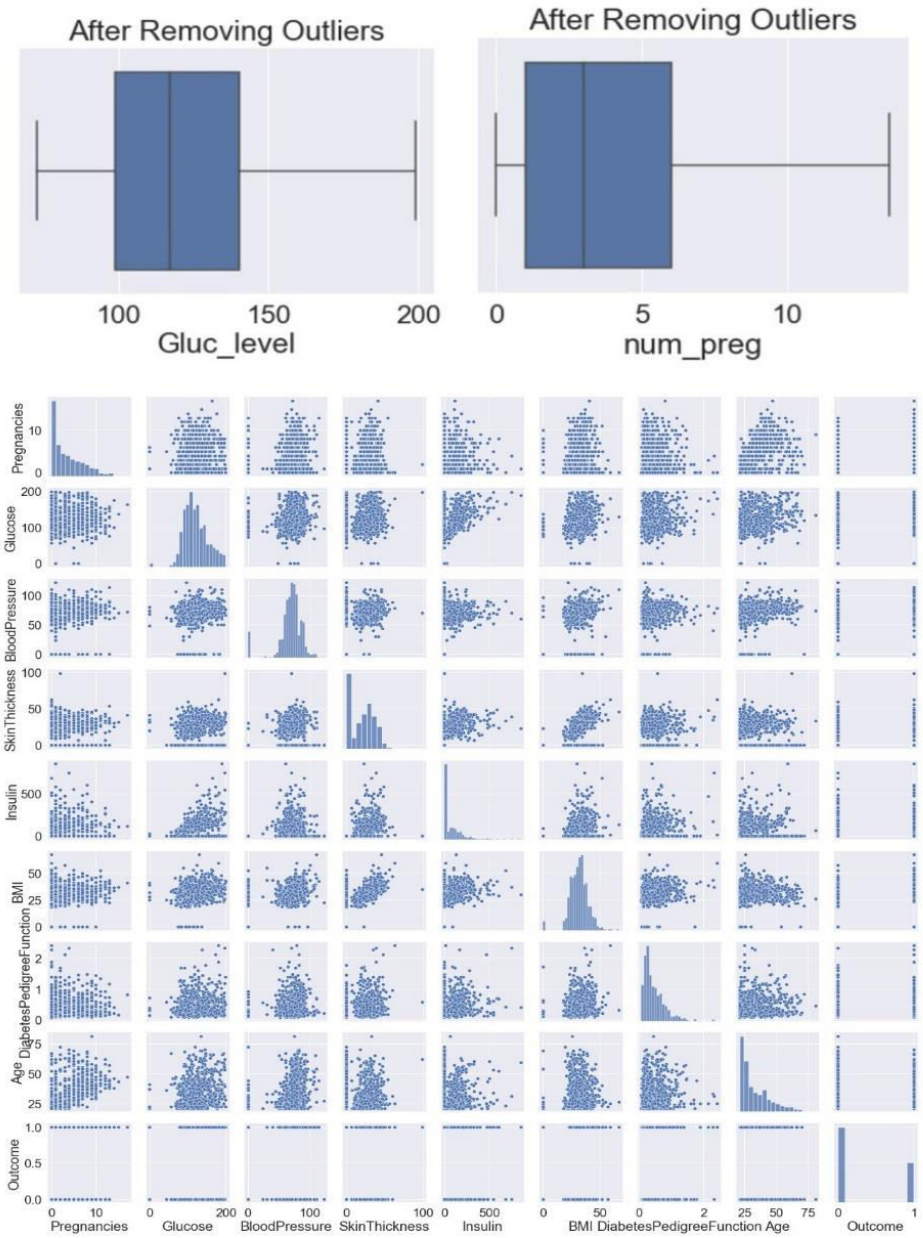

Fig 6: Outlier Detection Outcome

Dividing into training information and testing information:

After pre-processing the dataset, data diverges. For example, 30% of the content is test information from the data, and 70% is considered training information. Fig 7 shows the separation of data.

```
##Dividing the entire into test and train
from sklearn.model_selection import train_test_split
feature_columns = ['num_preg', 'Gluc_level', 'BP', 'Insulin', 'BMI', 'DPG', 'Age']
predicted_class = ['Outcome']
x=df[feature_columns].values
y=df[predicted_class].values
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
```

Fig 7: Separation of Training and Testing data

Various Ages in the dataset:
The below graph shows the observations made from type 2 diabetes of different age groups. Fig 8 shows various ages in the dataset.

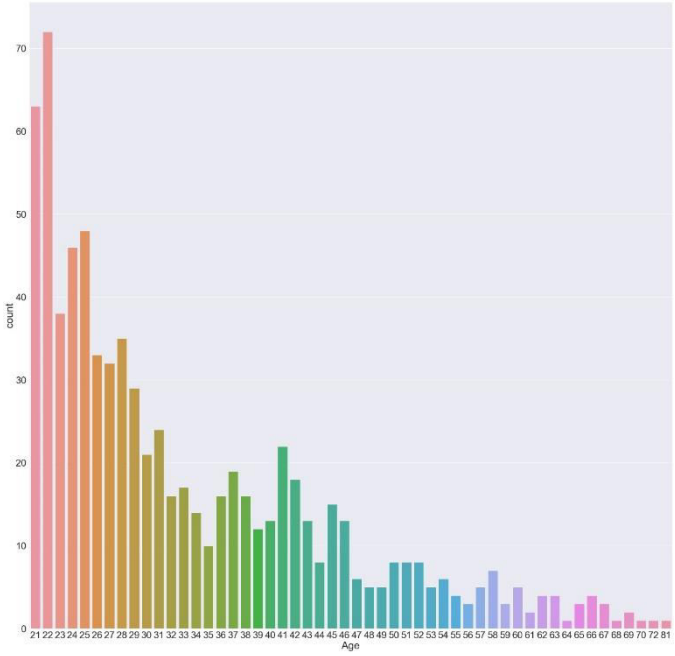

Fig 8: Various ages in the dataset

5. RESULTS AND DISCUSSIONS
Early identification of diabetes may help in improving the lifetime of the patients suffering from this severe illness. Various supervised learning algorithms are there to predict the disease. Some classification algorithms are applied here to make the Diabetes prediction. All ML algorithms with multiple factors like recall, AUC, etc., are below. Accuracy shows the efficient prediction of diabetes by determining the ability of the classifier in the prediction. The projections of the classification algorithms, along with the classification report, are below mentioned, which shows how the algorithms perform better in performance measures.

Confusion Matrix:
Confusion Matrix is an exhibition estimation for AI arrangement, which is incredibly helpful for estimating performance measures. Various factual estimation viewpoints survey execution of all the characterization calculations, such as precision, etc. The details about all the predicted classification and actual classification are in matrix form. It gives the accurate solution of the problem by category.
TP: True, and it is resulted as positive.
TN: True, and it is resulted as negative.
FP: (Type 1 Error): False, and it is predicted as positive.
FN: (Type 2 Error): False, and it is predicted as negative.

Evaluation Measures:
Sensitivity: Certain True rate, affectability, or review is characterized as the calculation that mentions the proportion in positive occurrences that have the disease among genuine actual cases. It is predicted as unfavorable.

$$\text{Recall} = \frac{\text{True Positive (TN)}}{\text{True Positive (TN)} + \text{False Negative (FN)}}$$

Specificity measure: It is a calculation that characterizes proportion where the people do not have diabetes and anticipated as non-diabetes—also, particularity reasonable inverse in review.

$$\text{Specificity Measure} = \frac{\text{True Negative (TN)}}{\text{False Positive (FP)} + \text{True Negative (TN)}}$$

Precision Measure: Apparent multitude of actual classes we have anticipated accurately, the number of really sure. The positive prescient worth or exactness of precise, accurate scores isolated with quantity in positive cases expected in characterization calculation.

$$\text{Precision Measure} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Postitive}}$$

Accuracy: Accuracy to issues is the proportion of correct forecasts made by the model over a wide range of reasonable expectations finished. Precision is the extent of proper outcomes that a classifier accomplishes. That is only, out of the apparent multitude of classes, the amount we anticipated accurately. Be that as it may, it doesn't oblige bogus positives and bogus negatives, where the precision of a classifier may increment while at the same time giving wrong forecast yields.

$$\text{Accuracy} = \text{TP+FN/TP+FP+TN+FN}$$

It is hard to contrast between models and exactness with another way around. So, to equalize that, F1-Score is used. F1-score assists with estimating Recall and Precision simultaneously. F1 score has an instinctive significance. It reveals how accurate our classifier is (the number of occurrences it arranges accurately), just as how vigorous it will be (it doesn't miss a considerable number of cases).

$$\text{F1 measure} = \frac{2 + \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

ROC Curve: This utilized diagram sums up the presentation of a classifier over every imaginable limit. It develops by d by plotting the (y-hub) against the (x-pivot) as you change the limit for appointing perceptions to a given class. ROC bend. A ROC bend (collector working trademark bend) is a diagram indicating the exhibition of order limits. This plots two boundaries:

True Positive Rate:

True Positive Rate (TPR) gives the meaning of recall, and it can represented as:

$$\text{True Positive Rate} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

- False Positive Rate (FPR):
- FPRis defined as follows:

$$\text{False Positive Rate (FPR)} = \frac{\text{FalsePositive}}{\text{FalsePositive} + \text{TrueNegative}}$$

For ROC bend, there is a need to assess a calculated logistic model commonly in various arrangement edges, yetthis will be wasteful. Luckily, there's effective arranging with the calculation which can give this data AUC.

AUC: ROC Curve Area:

AUC measures the Two-dimensional area inside the ROC Curve location from (0,0) to (1,1).

Decision Trees

Decision trees are dependent on a greedy approach in which the optimal decisions are made. Fig 11 shows the code for Decision Tree Classification. Fig 12, Fig 13, Fig 14 shows the report of classified data and Matrix of Confusion of Decision Tree Classification along with the ROC curve, respectively.

```
from sklearn import tree
t=tree.DecisionTreeClassifier()
t.fit(x_train,y_train)
t_1=t.predict(x_test)
print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, t_1)))
```

```
Accuracy = 0.662
```

Fig 11.Decision Tree Classification

	precision	recall	f1-score	support
0	0.79	0.88	0.83	154
1	0.69	0.55	0.61	77
accuracy			0.77	231
macro avg	0.74	0.71	0.72	231
weighted avg	0.76	0.77	0.76	231

Fig 12: Classified Report of Decision Tree

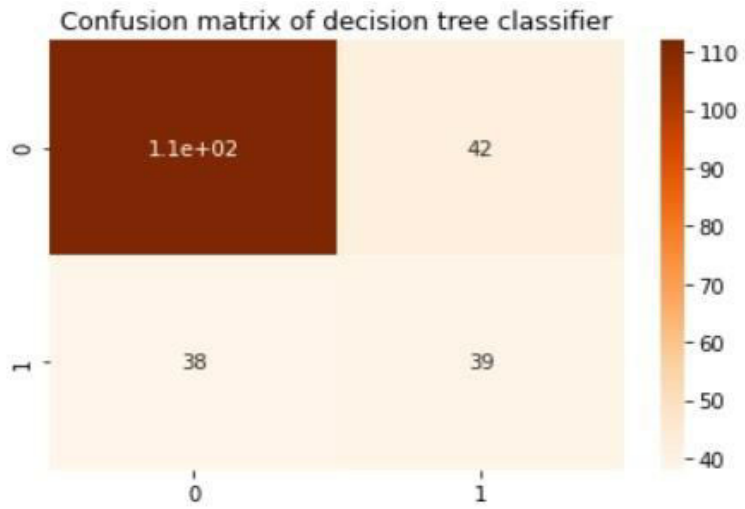

Fig 13: Matrix for Decision Tree

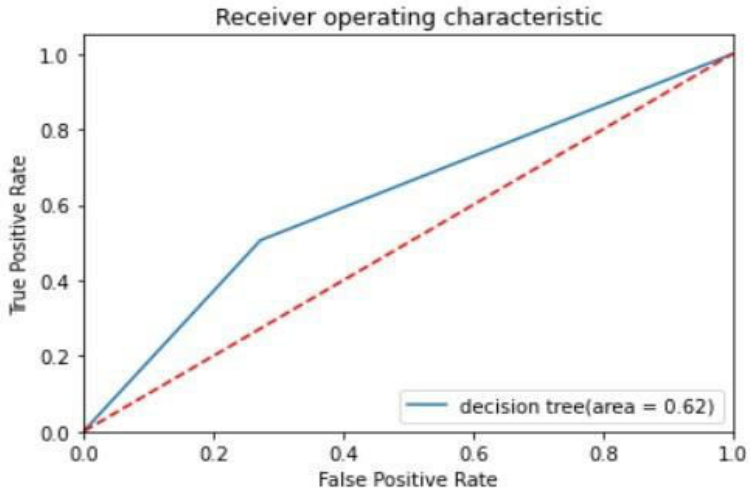

Fig 14: ROC for Decision Tree

Logistic Regression:

There is an opportunity to learn of functions in the interval [0, 1]. Since it is the likelihood, the result lies somewhere in the range of 0 and 1. Hence, to utilize the LR as a paired classifier, an edge should be allocated to separate two classes. Calculated relapse is anything but difficult to execute and clear. It is an LR-based model which can be refresh effectively, and it doesn't make a suspicion concerning the circulation of free factors. Fig :15, Fig:16, Fig: 17, and Fig:18 show the Classified report, matrix of confusion, and ROC in the Logistic model.

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
clf=LogisticRegression()
clf.fit(x_train,y_train.ravel())
a=clf.predict(x_test)
print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, a)))
```

```
Accuracy = 0.766
```

Fig 15: Logistic Regression Classification

	precision	recall	f1-score	support
0	0.84	0.82	0.83	160
1	0.61	0.65	0.63	71
accuracy			0.77	231
macro avg	0.73	0.73	0.73	231
weighted avg	0.77	0.77	0.77	231

Fig 16: Classified Information of Logistic model

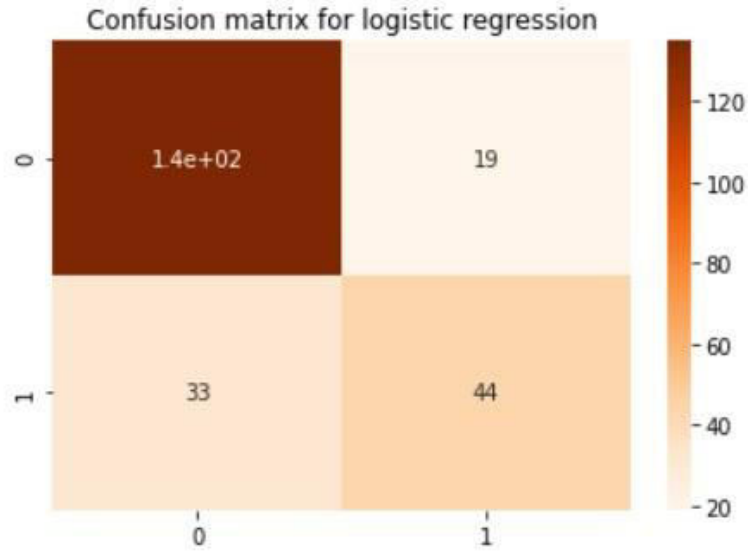

Fig 17: Matrix for Logistic model

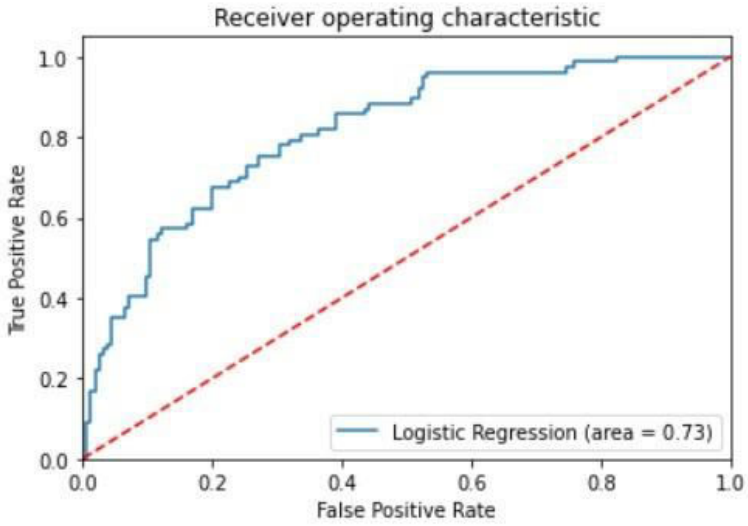

Fig 18: ROC for Logistic model

KNN:

The 'K' is the KNN calculation is the quantity of closest neighbors considered to take 'vote' from. The determination of various qualities for 'K' can produce exceptional arrangement results for a similar example object. It is a simple calculation that can arrange occurrences quickly.

It can deal with noisy occasions for characterization and relapse. However, KNN is computationally costly as the quantity of properties increases. Fig:19,20,21,22 shows the Classified outcome and matrix and ROC curve for KNN, respectively.

```
from sklearn.neighbors import KNeighborsClassifier as kn
k=kn()
k.fit(x_train,y_train.ravel())
k_1=k.predict(x_test)
print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, k_1)))
```

```
Accuracy = 0.779
```

Fig 19: KNN Classification

	precision	recall	f1-score	support
0	0.82	0.74	0.78	160
1	0.52	0.63	0.57	71
accuracy			0.71	231
macro avg	0.67	0.69	0.67	231
weighted avg	0.73	0.71	0.71	231

Fig 20: Classification Report of KNN

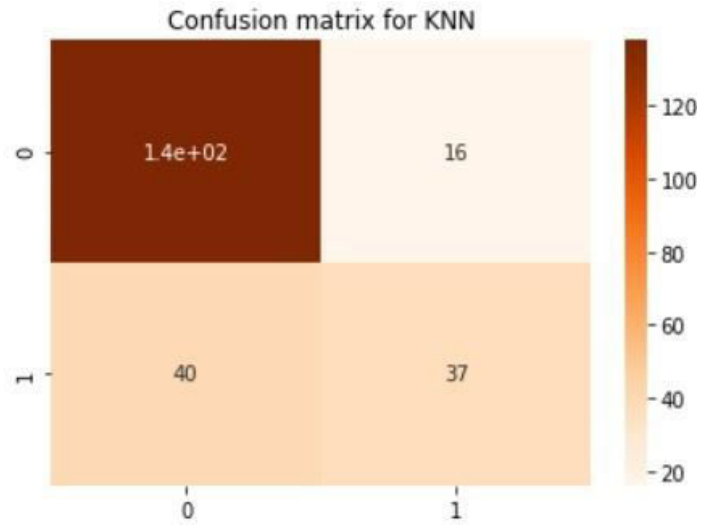

Fig 21: Matrix for KNN(Confusion Matrix)

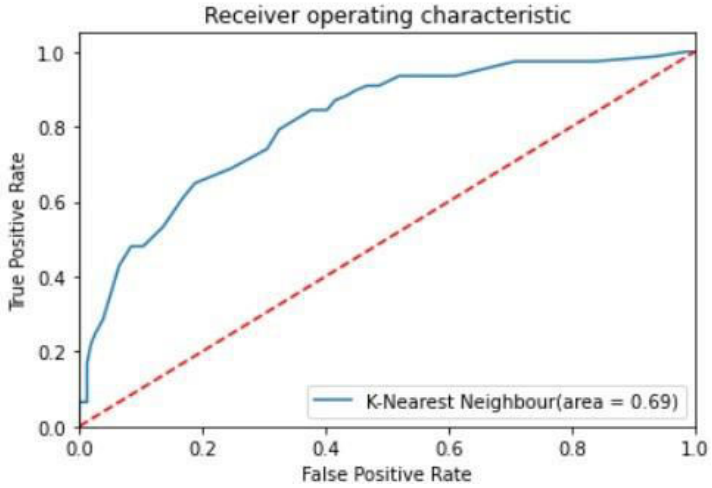

Fig 22: ROC for KNN

Random Forest model:

In RF, T is one of a few significant boundaries which the client must painstakingly pick. A portion of these boundaries are tuning boundaries as both excessively high and too low limit esteems yield problematic exhibitions; see Segal (2004) for an early investigation on the impact of such limitations.

This inquiry is pertinent to any client of RF. It has been the subject of much casual conversation in established researchers. However, as far as anyone is concerned, it has never been tended to methodically from a theoretical and observational perspective. Fig 23, Fig 24, Fig 25and Fig 26 show the Classified report, matrix of confusion, and ROC for Random Forest.

```
from sklearn.ensemble import RandomForestClassifier
random_forest_model = RandomForestClassifier(random_state=10)
random_forest_model.fit(x_train, y_train.ravel())
```

```
RandomForestClassifier(random_state=10)
```

```
b=random_forest_model.predict(x_test)
print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, b)))
```

```
Accuracy = 0.771
```

Fig 23: Random Forest Classification

	precision	recall	f1-score	support
0	0.79	0.88	0.83	154
1	0.69	0.55	0.61	77
accuracy			0.77	231
macro avg	0.74	0.71	0.72	231
weighted avg	0.76	0.77	0.76	231

Fig 24: Classification Report of Random Forest

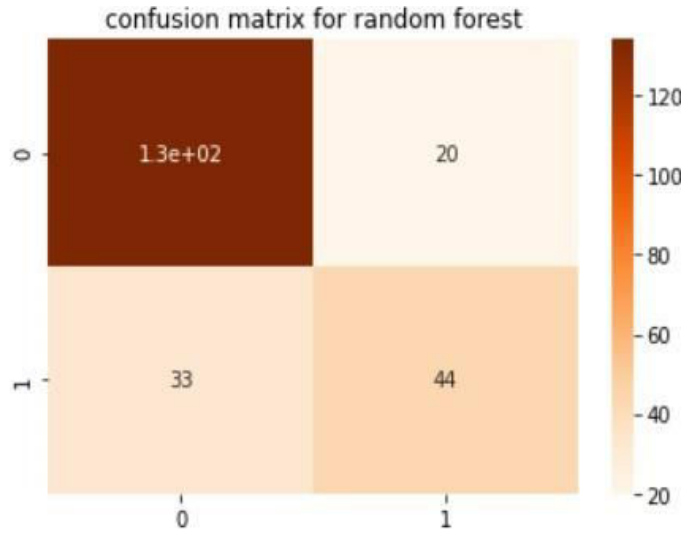

Fig 25: Confusion Matrix for Random Forest

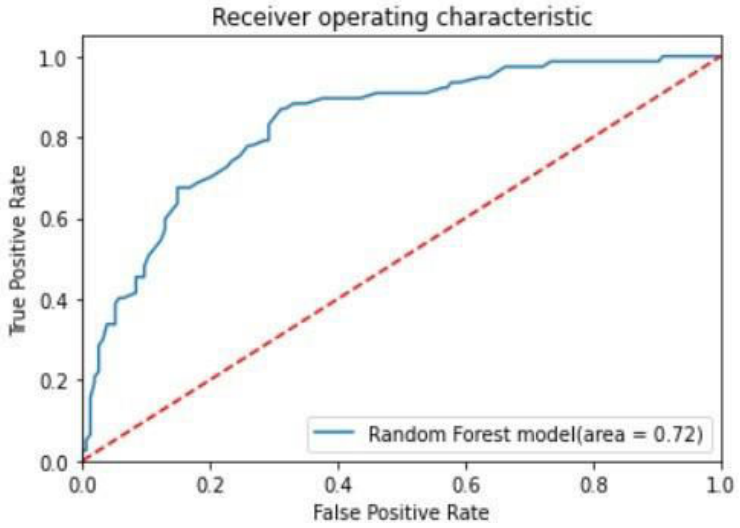

Fig 26: ROC for Random Forest

Support Vector Machines**:**
SVM is a binary classification algorithm. Therefore, separating the data with a more significant margin in Hard-SVM and the soft-SVM does not assume it separately.

AdaBoost:
Fig:27, Fig:28, Fig:29 and Fig: 30 show the Classification report, matrix, and ROC, respectively.

```
from sklearn.ensemble import AdaBoostClassifier
from sklearn.metrics import accuracy_score
ada=AdaBoostClassifier()
ada.fit(x_train,y_train.ravel())
x=ada.predict(x_test)
accuracy_score(y_test,x)
```

```
0.7835497835497836
```

Fig 27: AdaBoost Classification

	precision	recall	f1-score	support
0	0.82	0.87	0.84	156
1	0.69	0.61	0.65	75
accuracy			0.78	231
macro avg	0.75	0.74	0.75	231
weighted avg	0.78	0.78	0.78	231

Fig 28: Classification Report of AdaBoost

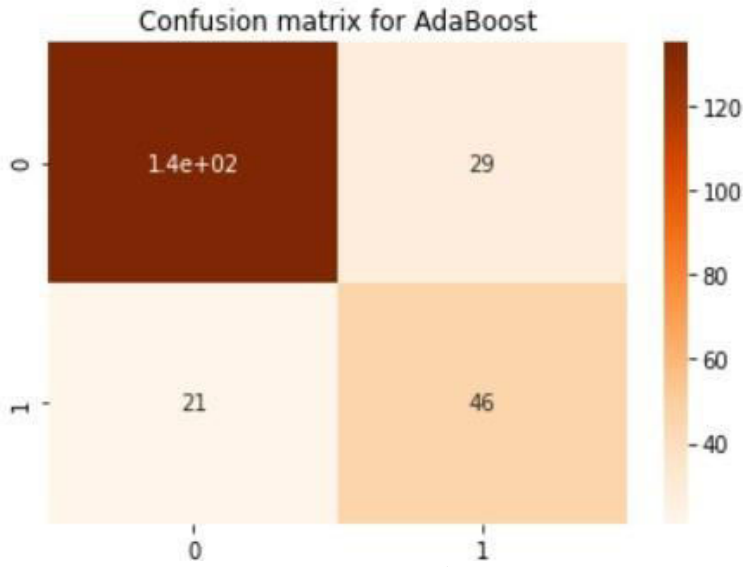

Fig 29: Matrix of AdaBoost

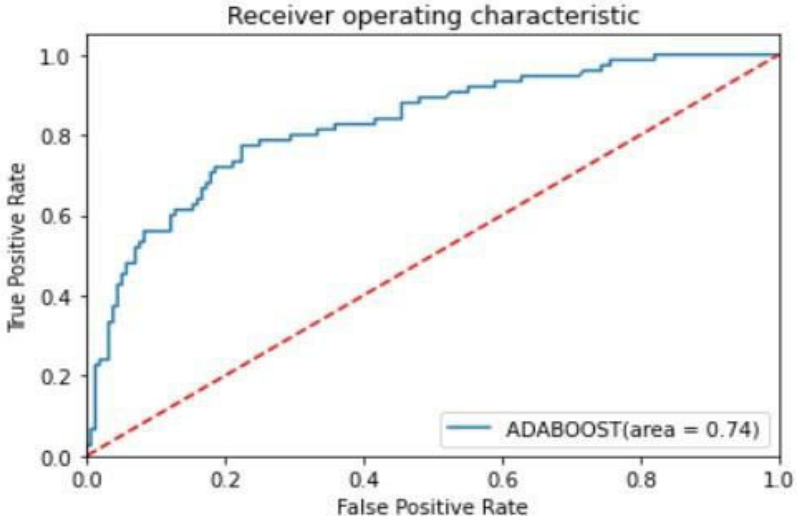

Fig 30: ROC of AdaBoost

Correlation Matrices:

The correlation matrix helps identify the most relevant features which lead to the disease. It helps in proving the prediction efficiently. Memory taking time reduced when the model trained. By this method, the noise is earnt by the model, which is the irrelevant features in the dataset. By implementing this, the accuracy of the predictive model improved. Fig:31 shows the correlations among parameters in the dataset.

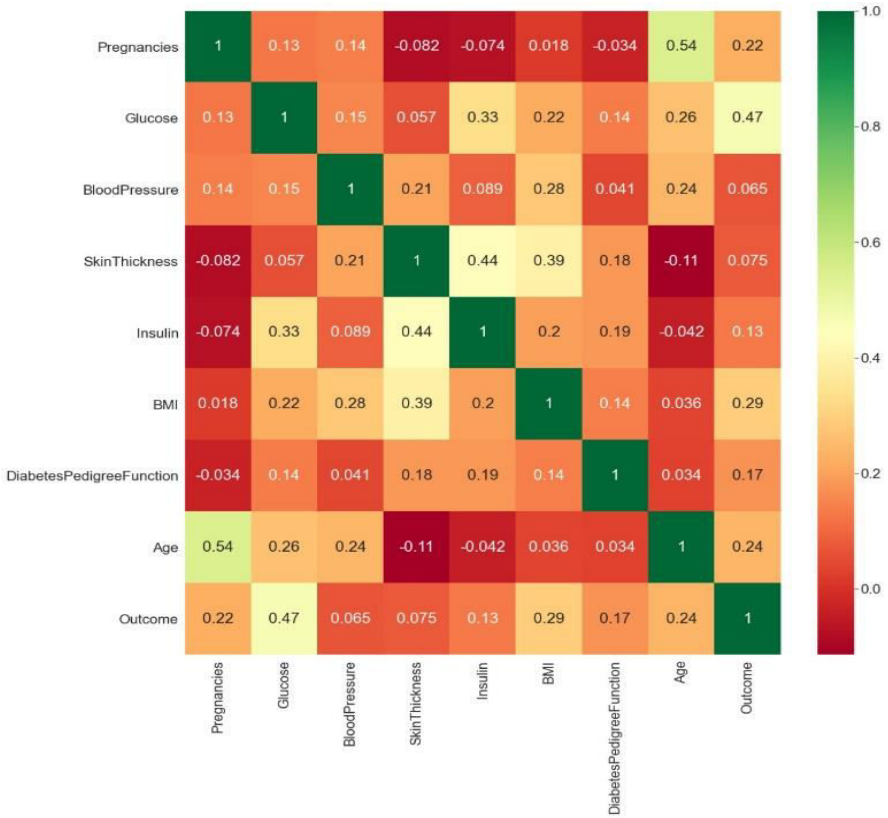

Fig 31: Correlations among parameters in the dataset

CONCLUSION:

Diabetes seems to be a rapidly developing issue for many kinds of related cognitive problems. Prediction of Type2 Diabetes helps a lot in predicting the problems before the problem's severity increases. From the above analysis done, the ensemble approach here shows an efficiency of 83%, which is better when compared to all other algorithms implemented. AdaBoost algorithms offer better accuracy and, based on various performance measures done, the Ensemble approach gives betterment for the prediction of type 2 Diabetes Mellitus. Future enhancement is in tracing out the accuracy by improvising the algorithm.

DECLARATIONS

Funding:

The authors did not receive financial support from any organization for the submitted work.

Conflicts of interest/Competing interests:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and material : 'Not applicable'

Code availability : 'Not applicable'

Authors' contributions : 'Not applicable'

Ethics approval : Compliance with Ethical Standards

Consent to participate ' : Not applicable'

Consent for publication:
Authors give consent to Soft Computing to publish their article.

REFERENCES:

[1] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and informatics.

[2] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

[3] Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive Supervised Machine Learning Models for Diabetes Mellitus. SN Computer Science, 1(5), 1-10.

[4] Probst, P., & Boulesteix, A. L. (2017). To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research, 18(1), 6673-6690.

[5] Kanat, M., DeFronzo, R. A., & Abdul-Ghani, M. A. (2015). Treatment of prediabetes. World journal of diabetes, 6(12), 1207.W. Zaremba, I. Sutskever, and O. Vinyals, ``Recurrent neural network regularization,'' 2014, arXiv:1409.2329. [Online]. Available: https:// rxiv.org/abs/1409.2329.

[6] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716.

[7] Georga, E. I., Protopappas, V. C., & Fotiadis, D. I. (2011). Glucose prediction in type 1 and type 2 diabetic patients using data driven techniques. Knowledge-oriented applications in data mining, 277-296.

[8] Basu, S., Yudkin, J. S., Kehlenbrink, S., Davies, J. I., Wild, S. H., Lipska, K. J., ... & Beran, D. (2019). Estimation of global insulin use for type 2 diabetes, 2018–30: a microsimulation analysis. The Lancet Diabetes & Endocrinology, 7(1), 25-33.

[9] Xia, W., Luo, Y., Chen, Y. C., Chen, H., Ma, J., & Yin, X. (2020). Glucose Fluctuations Are Linked to Disrupted Brain Functional Architecture and Cognitive Impairment. Journal of Alzheimer's Disease, (Preprint), 1-11.

[10] Xia, W., Zhang, B., Yang, Y., Wang, P., Yang, Y., & Wang, S. (2015). Poorly controlled cholesterol is associated with cognitive impairment in T2DM: a resting-state fMRI study. Lipids in health and disease, 14(1), 47.

[11] Kengne, A. P., Beulens, J. W., Peelen, L. M., Moons, K. G., van der Schouw, Y. T., Schulze, M. B., ... & Tormo, M. J. (2014). Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. The lancet Diabetes & endocrinology, 2(1), 19-29.

[12] Wang, Y., Wei, F., Sun, C., & Li, Q. (2016). The Research of Improved Grey GM (1, 1) model to predict the postprandial glucose in Type 2 diabetes. BioMed research international, 2016.

[13] Marder, T. J., Flores, V. L., Bolo, N. R., Hoogenboom, W. S., Simonson, D. C., Jacobson, A. M., ... & Musen, G. (2014). Task-induced brain activity patterns in type 2 diabetes: a potential biomarker for cognitive decline. Diabetes, 63(9), 3112-3119.

[14] Patil, R., & Tamane, S. (2018). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes. International Journal of Electrical & Computer Engineering (2088-8708), 8.

Sun, L., Diao, X., Gang, X., Lv, Y., Zhao, X., Yang, S., ... & Wang, G. (2020). Risk Factors for Cognitive

[15] Impairment in Patients with Type 2 Diabetes. Journal of Diabetes Research, 2020.

[16] Baynes, H. W. (2015). Classification, pathophysiology, diagnosis and management of diabetes mellitus. J diabetes metab, 6(5), 1-9.

[17] Goldenberg, R., & Punthakee, Z. (2013). Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. Canadian journal of diabetes, 37, S8-S11..

[18] Khawandanah, J. (2019). Double or hybrid diabetes: A systematic review on disease prevalence, characteristics and risk factors. Nutrition & diabetes, 9(1), 1-9.

[19] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716.

[20] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and informatics.

[21] Viloria, A., Herazo-Beltran, Y., Cabrera, D., & Pineda, O. B. (2020). Diabetes Diagnostic Prediction Using Vector Support Machines. Procedia Computer Science, 170, 376-381.
[22] Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142.
[23] Alehegn, M., Joshi, R. R., & Mulay, P. Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach.
[24] Saha, S., & Bhattacharya, T. (2020). An Approach to Enhance the Design of Protein Sequence Classifier Using Data Mining. Procedia Computer Science, 167, 717-726.
[25]Kamal, J., Tanveer, S., & Nafis, M. T. (2017). Disease symptoms analysis using data mining techniques to predict diabetes risk. International Journal of Advanced Research in Computer Science, 8(3).
[26] Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2020). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University-Computer and Information Sciences.
[27]Wang, Y., Liu, S., Chen, R., Chen, Z., Yuan, J., & Li, Q. (2017). A novel classification indicator of Type 1 and Type 2 diabetes in China. Scientific Reports, 7(1), 1-7.
[28]Care, D. (2020). Medical Care in Diabetes22020. Diabetes Care, 43, S135.
[29]Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.
[30]Pranto, B., Mehnaz, S., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. (2020). Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. Information, 11(8), 374.
[31] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders, 19(1), 1-9.
[32] Hassali, M. A., Nazir, S. U., Saleem, F., & Masood, I. (2015). Literature review: pharmacists' interventions to improve control and management in type 2 diabetes mellitus. Altern Ther Health Med, 21(1), 28-35
[33]. Nirmalraj, S., and G. Nagarajan. "Biomedical image compression using fuzzy transform and deterministic binary compressive sensing matrix." Journal of Ambient Intelligence and Humanized Computing 12, no. 6 (2021): 5733-5741..
[34] Nagarajan, G., Minu, R. I., & Devi, A. J. (2020). Optimal nonparametric bayesian model-based multimodal BoVW creation using multilayer pLSA. Circuits, Systems, and Signal Processing, 39(2), 1123-1132.
[35] Nagarajan, G., and K. K. Thyagharajan. "A machine learning technique for semantic search engine." Procedia engineering 38 (2012): 2164-2171.
[36] Nagarajan, G., Minu, R. I., & Jayanthiladevi, A. (2019). Brain computer interface for smart hardware device. International Journal of RF Technologies, 10(3-4),131-139.
[37] Minu, R., G. Nagarajan, A. Suresh, and Jayanthila A. Devi. "Cognitive computational semantic for high resolution image interpretation using artificial neural network." BIOMEDICAL RESEARCH-INDIA 27 (2016): S306-S309.

Author Details:

M.S.Roobini received B.E degree in Computer Science and Engineering from Cape Institiute of Technology and the M.E degree in Software Engineering from Noorul Islam University in 2009 and 2011,respectively.she is a Faculty Member of Department of Computer Sience and Engineering,School of Computing,Sathyabama Institute of Science and Technology, Chennai,India.Her current research interests include Artificial Intelligence, Machine Learning,and Computer Vision.

Dr.M Lakshmi presently working as Professor in Department of CSE,SRM Institute of Science and Technology, Kattankulathur.She previously worked as Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences from 2019 to 2018 and she worked as Principal in Sri Krishna College of Technology, Coimbatore from Sep 2018 to April 2019. She also worked as Professor and Dean, School of Computing, Sathyabama University, Chennai, till June 2018 from Nov 1995. She has an experience of more than 24 years in Teaching. She has completed her B.E. (Computer Science and Engineering) from Bharathidasan University and M.E. (Computer Science and Engineering) from Madras University and Ph.D. from Sathyabama University in the area of Wireless Ad Hoc networks.Her current research interests include Computer Vision, Artificial Intelligence, Machine Learning, and Wireless Ad Hoc networks. She has published more than 100 research papers in peer-reviewed journals.