IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

A Machine Learning Approach for Diabetes Prediction in Women

Afshan Hashmi[1], Md Tabrez Nafis[2,*], Sameena Naaz[3], Imran Hussain[4]

[1,2,3,4] Department of Computer Science & Engineering, Jamia Hamdard(Deemed University), New Delhi

*Corresponding email: **tabrez.nafis@gmail.com**

Abstract. Diabetes is one of the diseases that are chronic and has seen exponential growth in the recent past. Trends suggest that the number of patients suffering from this disease is going to be doubled very soon which is a cause of serious concern and it needs to be tackled at the earliest. The reason why it is considered a chronic disease is that it is the cause of several other serious diseases such as hypertension, kidney failure, blindness, limb amputation, etc.
So, it is highly required to predict diabetes as early as possible to protect the patient from further damage. Machine learning can be proven as a beneficial tool for the prediction of diabetes. In this study, we have taken the PIMA India dataset, dropped the highly correlated feature, and filled the missing value by KNN imputation. Inter Quartile range was used to get rid of the outliers and Adaptive synthetic sampling was used for class balancing and min-max scaler for normalizing the dataset. Eight machine learning algorithms were used named Support vector classifier, Logistic regression, Naïve Bayes, Decision Tree, Xtreme gradient boosting,K-nearest neighbor, Linear discriminant analysis, and Random Forest.These algorithms were compared based on various performance metrics such as Accuracy, Precision, Recall, F1-score, and Auc-Roc curve. It was found that the linear discriminant analysis and Xtreme gradient boosting was the best performer in terms of accuracy followed by Random Forest, Logistic regression, K nearest neighbor, support vector classifier, and naïve Bayes. The decision tree however showed poor performance. The effect of oversampling on the result was also analyzed and it was found that oversampling enhances the precision and F1 score of all the algorithms but decision tree. Performance can be further improved by using a larger dataset with no or negligible missing values or with a dataset with some additional features such as lifestyle, calorie intake, etc.

Keywords: Support vector machine, diabetes prediction, XGB, logistic regression, random forest,machine learning, accuracy, recall, precision, f1 score, linear discriminant analysis, Adasyn.

INTRODUCTION

Diabetes Mellitus is one of the common chronic diseases that has affected a major part of the population and it is increasing at an alarming rate. Out of all the diabetes cases, 90-95% of cases are of Type 2 diabetes. It is not communicable still it is turning into a silent killer[17]. It is one of the dominant reasons behind heart attack, kidney failure,amputation, blindness, stroke, and nerve damage[1]. Normally when we eat something it breaks down into glucose at that time our pancreas secretes insulin and it is due to the insulin that our cells open up and use that glucose for energy but this mechanism didn't work in the case of diabetes[1]. Diabetes is of 3Types, Type1diabetes-where the body is not able to produce the required amount of insulin, Type 2 diabetes - where the human body is not able to utilize the insulin properly [2]Gestational Diabetes – which affects the pregnant woman during the third trimester of pregnancy mainly because the hormones produced by the placenta cause insulin resistance[3]. Out of the 3 types of Diabetes Type-2, diabetes is very commonly found among people that's why our study mainly focuses on type 2 diabetes. Various factors that play a key role in this disease are aging, a sedentary lifestyle being overweight, an unhealthy diet, and genetic factors. If diabetes is not diagnosed at an early stage it can lead to various health issues and complications that might be dangerous to life[3].To tackle this problem data are collected by the healthcare industry and with the help of data mining techniques and Machine learning algorithms an early prediction of diabetes can be done which might play a vital role in the cure and treatment of this disease. Electronic Health Records are not provided with all requisite information in all conditions and scenarios. Due to these irregularities prediction has become highly challenging and there is an increase in the misclassification rate [14]. The goal of this research.This research aims to explore the PIMA dataset and study and compare the performance of eight machine learning classifiers for the prediction of diabetes inwomen with and without oversampling technique. The literature review part is in section 2. Datasets, machine learning algorithms used, and evaluation metrics are discussed in section 3 Methodology. The result of this study is discussed in section 4and finally,the conclusion of this research is mentioned in section 5 conclusion, and future work.

LITERATURE REVIEW

In 2018 authors used three Machine learning algorithms Neural networks, Random Forest, and Decision Trees, on the dataset from Luzohu China, containing 14 features to predict diabetes mellitus. Fivefold cross-validation was used for model examination, minimum redundancy maximum relevance (mRMR),and principal component analysis (PCA)was used to reduce the dimensionality. Prediction with Random Forest using all the features provides an accuracy of 0.8084. The diagnosis was done based onglucose tolerance, fasting blood glucose, and random blood sugar levels. It was concluded that Random Forest performed slightly better as compared to the

other two Decision tree and Neural Networks. Using PCA was not so good in terms of accuracy. Using all features and RmR provides better results of 80% accuracy on the Luzohu, China data set and 77 % accuracy on the PIMA India dataset. Stress is need to be given on the selection of correct classifiers and valid features. This proves that Machine learning models can be a good predictor of diabetes but Stress is need to be given to the selection of correct classifiers and valid features. In the future predicting the types of diabetes and determining the contribution of each feature can be taken into account [4].

Predictive analysis is being done using many machine learning algorithms and techniques but it is quite a difficult process however it can utilize big data and derive valuable insights about the health and treatment of patients. PIMA India Dataset was used with 768 rows and 8 columns. Six Machine Learning algorithms were applied and the result is shown in **Table 1**

Table 1-Accuracy of six ML algorithms

Algorithm	**Accuracy achieved**
Logistic Regression	74%
Support Vector Machine	77%
Naïve Bayes	74%
Decision Tree	71%
Random Forest	71%
KNN	77%

It was found that the K Nearest Neighbor and Support Vector Machine were found to be the most accurate in terms of accuracy but to get a better result a larger dataset is required with zero missing values. Tuning the parameters in the proposed model with a larger dataset having no missing values can result in better performance in the future [6].

In 2019 The authors found that the existing models are not so accurate in prediction so to improve the accuracy they suggested, including some external features, also as information about Job type is also included as a feature like whether a person is doing fieldwork, machine work, or office work along with the general features like (BMI, age, Insulin, Glucose level). In their study, the dataset was normalized and then K means clustering was applied to classify the patient into diabetic or non-diabetic after this various machine learning algorithms, Logistic Regression, Naïve Bayes, Bagging algorithm, Gradient Boosting algorithm, Random Forest Classifier, Decision Tree Classifier,Support Vector Classifier,Linear Discriminant Analysis algorithm, Extra Tree Classifier, K-Nearest Neighbor, Gaussian Ada Boost algorithm, Perceptron was applied on the data set. This paper also proposes to create a pipeline of the algorithms giving the highest accuracy [2].

In another study, the authors propose various machine learning algorithms that will help automate the model which can predict diabetes at an early stage with greater accuracy. Distributed computing framework based on a Hadoop cluster is useful for processing and storing large data sets in the cloud environment. Dataset: Data from 75664 patients were collected by the National institute of diabetes. The data set consisted of 13 features out of which Age, Diabetes Pedigree function, BMI, Plasma glucose concentration, Diastolic Blood pressure, Serum Insulin, and Triceps skinfold thickness were proved to be the important features. Differential statistical techniques and Information gain were used for feature selection out of which the former was not so useful and the latter proved useful in increasing the accuracy. It was concluded that the Random Forest algorithm based on the Hadoop cluster performs much better in terms of all the different performance measures (94% precision, 93% accuracy, 90% F measure, and 87% recall) as compared to Naïve Bayes and Decision Tree. In future work, the Meta-Heuristic algorithm can be used as part of Machine learning algorithms by adding more nodes in the Hadoop cluster[5].

A huge data is getting generated by the healthcare industry which is sensitive as well as tedious to handle. This data can be helpful to design a prediction system for diabetes but the system should be reliable enough for the health care professionals. In this paper WEKA software was used for data mining and the dataset used was the PIMA India dataset.

This paper aims to apply bootstrapping resampling techniques for better accuracy and then apply KNN, decision tree, and Naïve Bayes algorithm and then analyze and compare their results. It was found that the ensemble method provides better accuracy of 90.36% as compared to the single one which was only 83.7. This study was only focused on diabetes which can also be used for other diseases using other datasets. In this research, only limited classifiers were used but in the future other classifiers can also be used for further improvements [8].

Diabetes occurs either when the pancreas is not able to produce enough insulin or when the body cannot use it efficiently. It can cause major problems like damage to the blood vessels, heart, eyes, kidneys, and nerves. The study was done on the PIMA India dataset. Mean imputation was done for missing values and Prioritizing the attribute was done to get more accuracy. Receiver operating characteristics and Root mean square error were considered as performance metrics for the ANN model and thus RMSE 0.39 and ROC 0.88 were achieved. A predicted value between 0.5 and 1 was considered diabetic and a value between 0 and < 0.5 was considered non-

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

diabetic. This model is capable of getting 92% accuracy which can be further enhanced by using a large sample [9].

On the PIMA dataset, various methods of preprocessing were used stage-wise. In the first stage outliers removal and missing value imputation were done, in stage 2 dataset was normalized, and in stage 3 balancing of the dataset was done, and after each stage, three machine learning algorithmsSupport vector machine, K Nearest Neighbour, and Random Forest was applied. And it was found that after each stage of preprocessing, accuracy increased and the best performance was of random forest in terms of accuracy of 82%. 5 and 95 percentile values were used for outliers removal, median values were used for missing values imputation, Z score was used for normalization, and SMOTE was used for balancing the dataset. The same method can be used for different datasets and different machine learning algorithms [12].

In 2020 Authors developed a Machine learning-based system that can predict whether a patient is diabetic or not. Four Machine learning algorithms Naïve Bayes, AdaBoost, Decision tree, and Random Forest were applied to the data set received from the survey of National Health and nutritional examination with 6561 records. To identify the risk factors Logistic regression was used based on P-value and odd ratio. For partitioning K2, K5, and K10 were used, and the Area under the curve in addition to accuracy was taken as evaluation metrics.

with the help of logistic regression, it was found that 7 out of 14 features are the risk factors of diabetes which includes education, age, cholesterol, and blood pressure. The overall accuracy of 90.6% was attained on a machine learning-based system and it further increased to 94.25 % when the feature selection algorithm was Logistic Regressor and the classifier was Random Forest. It was concluded that Logistic Regression and Random Forest together Provide better predictions. This framework could be used with different medical datasets in the future[7].

In 2021 PIMA India Dataset was used and then Machine learning algorithms, Data Mining, and Neural networks were applied to it. WEKA software was used and missing values were dealt with mean imputation and outliers were removed so the dataset was left with 699 records. Feature selection was done by using Pearson's correlation method and feature scaling was done by normalizing the data from 0 to 1. K-fold cross-validation was done. 85 % data was used for training and 15% for testing.3 features were dropped using the feature reduction method and 5 features were used in this research. Seven Machine learning classifiers used in this research were Decision Tree, KNN, Random Forest, Naïve Bayes, support vector machine, Ada Boost, and logistic regression. It was found that Logistic regression was 78.8% accurate, Naïve Bayes was 78.2% accurate, Random Forest was 77.34% accurate, and Artificial Neural Network with 2 hidden layers and 400 epochs was the most Accurate with 88.57 % [10].

In this research, two approaches were adopted, a Classification based algorithm (Random Forest) and a Hybrid algorithm (XGBoost). PIMA dataset was used with 768 rows and 9 columns. Missing values were replaced with mean. The correlation of features was derived using the correlation function and the Randomized Search CV was used to optimize the hyper-parameter. Two ML algorithms were applied and the result was Random Forest was 71.9% accurate and XG Boost was 74.10% accurate. It was found that XG Boost gives better accuracy and faster result because it optimizes the hardware and software. Performance can be improved further by optimizing the hyperparameters [11].

In 2022 authors discuss the risk factors of diabetes and analyze 35 machine learning algorithms using or without using the feature selection. 3 different data sets were used and 9 feature selection algorithms. The performance of these algorithms was compared in terms of execution time, F1 score, and accuracy.

It was found that metro culture, unhealthy lifestyle, and genetic factors are also a reason for developing diabetes. It was found that bagging LR was most efficient for a balanced data set and Random Forest was most efficient for an unbalanced dataset[3].

To study the effect ofdata reduction and pre-processing for classification problems in the diagnosis of diabetes a new model was introduced. The model consists of 4 stages Pre-processing data, selection of features, Classifying, and Performance evaluation. For classification boosting, bagging, voting, andstacking was used. The dataset was used with and without preprocessing and it was found that preprocessing certainly improve the F1 score and accuracy to 97.4% and 97.12% respectively[13].

METHODOLOGY

This section discusses the PIMA dataset used in this study, Machine Learning classifiers used for the prediction of type 2 diabetes, and evaluation metrics such as Precision, Accuracy, F1 Score, Recall, and AUC-ROC curve.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

1.1 **Dataset** The Pima India dataset used in this study has the data of women with 768 rows and 9 columns. Features include Pregnancies – which indicates the number of times a woman was pregnant, Glucose –which indicates the concentration of Plasma glucose, blood pressure –which indicates the Diastolic blood pressure of the person (mm Hg), SkinThickness –which indicates skin fold thickness of Triceps(mm), Insulin –which indicates 2-Hour serum insulin (mu U/ml), BMI –which indicates the Body Mass Index of the person, DiabetesPedigreeFunction – scores the probability of occurrence of diabetes based on family history, Age - Age in years and label is Outcome - Whether the person is diabetic or not, 0 represents the person is not diabetic and 1 represents that the person is diabetic. **TABLE 2**. shows the statistics of the dataset.

Table 2- Statistics of the PIMA dataset

Features	**Pregnancies**	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree**	**Age**
Count	768	768	768	768	768	768	768	768
STD	3.37	31.98	19.35	15.95	115.24	7.88	0.33	11.76
Mean	3.84	120.89	69.10	20.53	79.79	31.99	0.47	33.24
Median	3	117	72	23	30.5	32	0.37	29
Minimum	0	0	0	0	0	0	0.078	21
Maximum	17	199	122	99	846	67.1	2.42	81

As shown in TABLE 2 the minimum value of most of the attributes is zero which is practically impossible for some of the attributes like blood pressure, glucose, skin thickness, and BMI this implies that these are missing values. Similarly, maximum values of some attributes are also very high like 17 pregnancies which denotes the presence of outliers. So, outliers and missing values will be dealt with in our data preprocessing section. The correlation between the features can be seen in **FIGURE 1** It is evident that age and pregnancy are positively correlated similarly BMI and skin thickness is very closely related and insulin is also quite dependent on the glucose level. Since it is evident that the correlation between BMI and skin thickness is the highest and among the two of them skin thickness has a significant number of missing values so it's better to drop the feature "skin thickness". So, now we are left with seven dependent features.

3.2 Data Preprocessing

The dataset was checked for any missing value initially and it was found to be none but on carefully examining it was found that some features have 0 values which are not possible so all the 0's were replaced with NaN and then KNN was used for missing value imputation. On exploring further it was found that outliers are there in the dataset and with the help of the box plot, it is shown in **FIGURE 2** which were removed using the IQR Score method.

$$Q1 = df.quantile(0.25)$$
$$Q3 = df.quantile(0.75)$$
$$IQR = Q3 - Q1$$

After removing the outliers the dataset was normalized using the min-max scaler

$$x_{std} = (x - x.\min(axis = 0))/(x.\max(axis = 0) - x.\min(axis = 0))$$

$$x_{scaled} = x_{std} * (max - min) + min$$

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

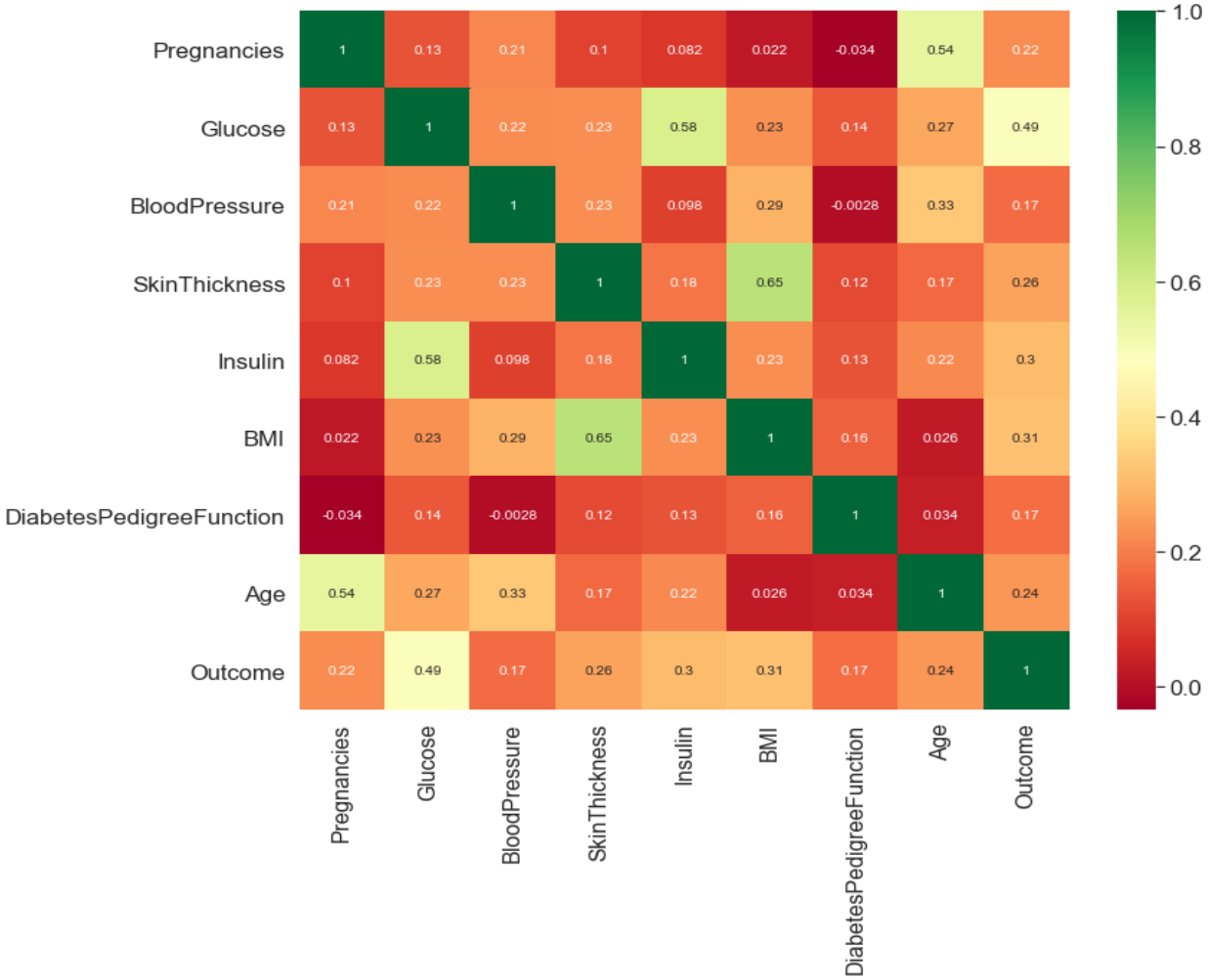


Figure 1 Correlation heat map

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

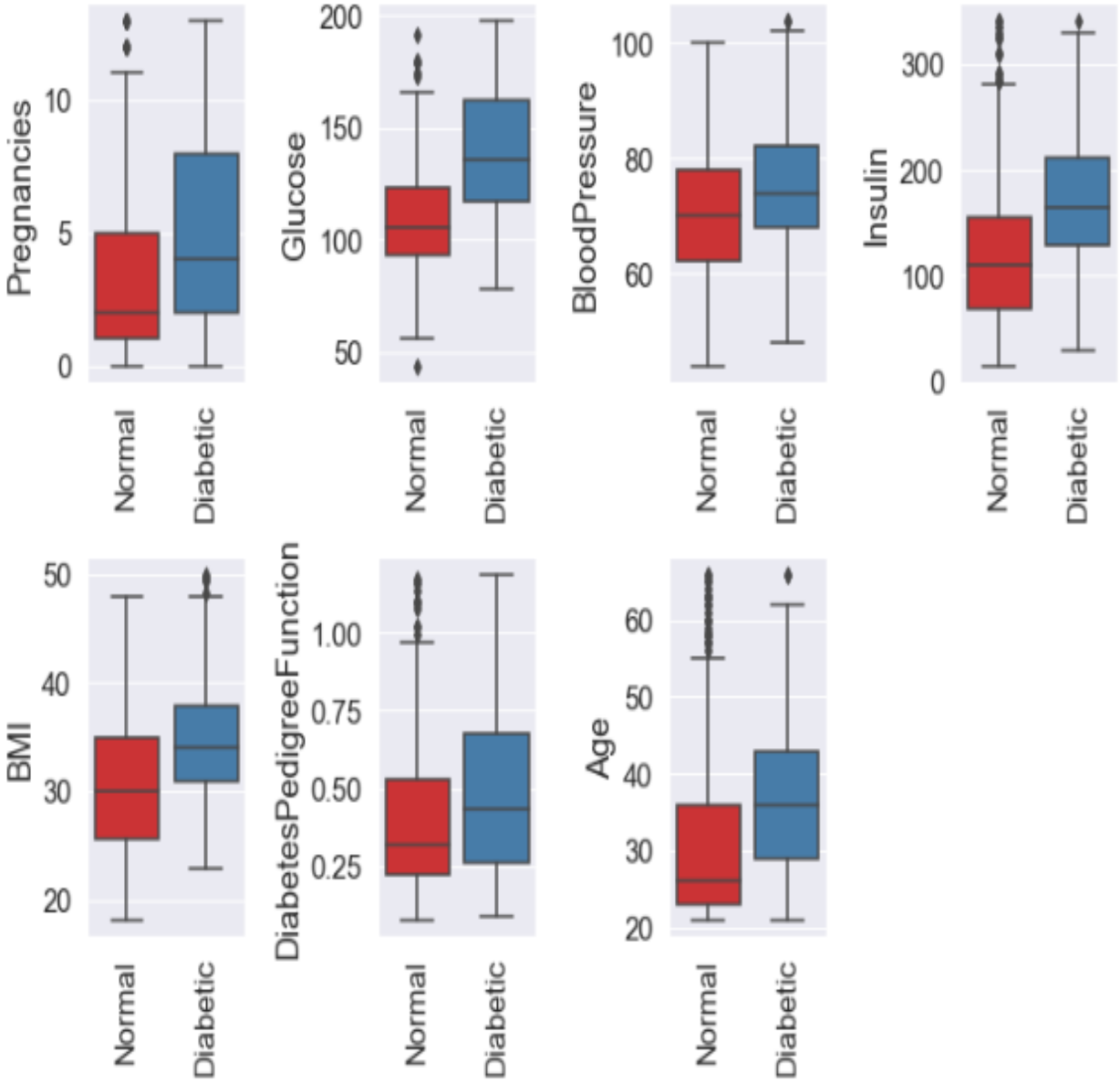


Figure 2 Outliers present in the dataset

On further investigation it was found that the dataset containsa different range of values, to normalize the values between the range of 0 to 1we have used a min-max scaler.

$$x^{norm} = \frac{x_{input} - x_{min}}{x_{max} - x_{min}}$$

Where, x^{norm} denotes the normalized data,
x_{input} denotes input value
x_{min} denotes the minimum value of the feature
x_{max} denotes the maximum value of the feature

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

The dataset is not so balanced as shown with the help of the pie chart in **FIGURE 3**. The percentage of non-diabetic persons is 68.09whereas diabetics are only 31.91%. we will try to balance it during pre-processing.

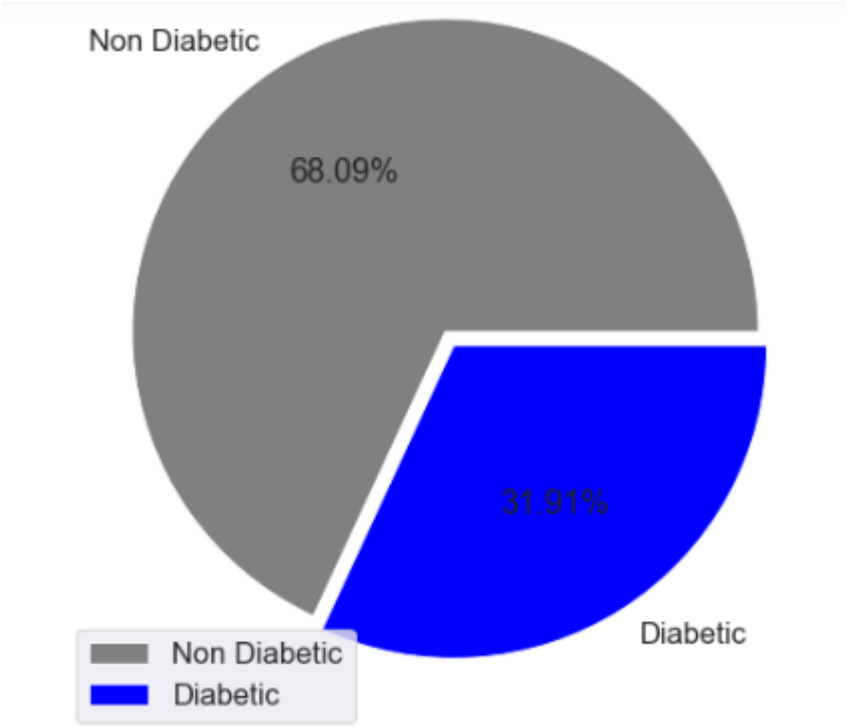


Figure 3 Percentage of Diabetic and non-diabetic in the dataset

Since the dataset was not balanced so ADASYN adaptive synthetic sampling was used. Adaptive synthetic sampling is used to solve the problems of classification with an imbalanced class. It works by synthetically generating data for the minority class. The core concept of ADASYN is to employ a weighted distribution for various minority class examples depending on how challenging it is for them to learn. More synthetic data is generated for difficult-to-learn minority class examples than for easier-to-learn minority examples.The dataset was then split into testing and training data using a train-test split with a75% data as training data and 25% data as the training dataset.

3.3 Modelling and Evaluation

Seven machine learning classifiers K Nearest Neighbor, Logistic Regression, Random Forest, Naïve Bayes, Support vector machine,Decision Tree, and Xtreme Gradient Boosting were used in this study.

Linear discriminant analysisis also known as normal discriminant analysis. It is a generalized form of fisher's linear discriminant.LDA works by minimizing the variance of each class and maximizing the distance between the mean of two classes. Apart from supervised classification problems, it is also used for preprocessing data by data visualization, and dimensionality reduction.

Logistic Regression is the machine learning algorithm classifier thatuses an analysis of the correlation between one or more pre-existing independent variables to predict a dependent data variable.Based on previous observations of a data set, the statistical analysis technique of logistic regression can be used to forecast a binary outcome, such as yes or no. LR employs the cost function, also referred to as the sigmoid function. Each number between 0 and 1 is transformed by the sigmoid function as shown in **FIGURE 4**

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

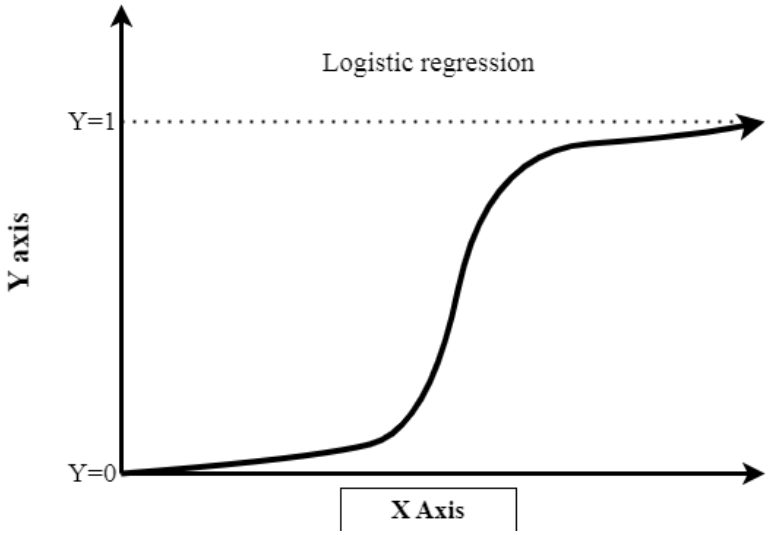


Figure 4 Sigmoid function

Decision Tree is a type of supervised learning algorithm that is non-parametric and utilized for bothregression, and classification. But, it is mostly chosen for classification issues. Its structure is like a tree, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. There are three types of nodes in a decision tree, Root node represents the entire dataset, Decision nodes are used to create decisions and Leaf Nodes are the results and do not have any more branches. We cannot randomly choose a feature as a root node but is decided based on entropy or information gain. The feature with the maximum information gain will be selected as a root node.

Information GainIt measures the reduction of uncertainty and it is also a deciding factor for each attribute to be selected as a root node

$$Information\ gain = Entropy\ (before\ splitting) - Entropy(after\ splitting)$$

Random Forest combines the output of various decision trees to get a single outcome. Given that it can solve classification and regression issues, its popularity has been boosted by its simplicity and adaptability. The bagging approach is extended by the Random Forest algorithm, in this technique, manydecision trees are created using different samples and an average value of all the decision treesis taken in case of a regression problem, and a majority vote of all the decision tree is taken in the classification problem.

K Nearest Neighbor is a supervised learning algorithm used for bothclassification and regression. It calculates the distance between the test data point and the training data points and then predicts the suitable class for the test data. After that, it selects the K spots that are closest to the test data. The KNN method determines which classes of the "K" training data the test data will belong to, and that class is chosen which has the highest probability. The value in a regression situation is the average of all 'K' chosen training data points. Distance between the test data points and training data points are calculated in three ways, Hamming distance is used for classification problem while Manhattan distance and Euclidean distance is used for regression problems.

Naïve Bayes is a machine learning algorithm that, is based on the Bayes theorem. It is the probability of occurrence of event A provided that B has already occurred.

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

It is considered Naïve because it assumes that the probability of occurrence of an event is independent of the occurrence of other events, which is practically not possible.

Support Vector Machine is a type of supervised machine learning algorithm which is used for regression and classification, although classification is where it is most useful. SVR (support vector regressor) is used for regression, and SVC (support vector classifier) is used for classification. In an N-dimensional space, where N stands for the number of features, SVM looks for a hyperplane that uniquely categorizes a data point. If there are only two features, the hyperplane is only a line, and if there are three input features, the hyperplane is a two-

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

dimensional plane. The support vectors are the points that the SVM algorithm finds that are closest to the lines from both classes. The distance between the hyperplane and the vectors is known as a margin and the objective of SVM is to increase the margin, and the ideal hyperplane is the one that does so.

XGBoost Extreme is an ensemble machine learning approach built on decision. This algorithm is an improved or advanced variant of gradient boosting. It is an ensemble method where each tree boosts the attribute that caused the preceding tree to be misclassified. It is a supervised machine learning algorithm that can be applied to both regression and classification problems In a sequential fashion, decision trees are made. All independent variables are given weights, and the decision tree that forecasts outcomes uses these weights to predict the outcomes. The tree increases the weight of variables that it incorrectly predicted. After that, the second decision tree is fed with these variables.A robust and accurate model is then produced by combining these distinct classifiers/predictors.

Using all the eight machine learning algorithms mentioned here we did the modeling on the training data set and the process is shown here in **FIGURE 5**
Four evaluation metrics were considered Precision, Accuracy,F1-Score, Recall, and Auc-Roc curve

Confusion Matrix: It is a tabular representation of the n X n matrix where n is the number of the target class. It is a comparison of predicted values and actual values. False Positive, True Negative, True Positive, and False negative were derived from the confusion matrix for each algorithm separately.*True Positive*when the actual value is positive and the model's predicted value is also positive.*False Positive*when actual is negative but the model's predicted value is positive.*False Negative*when the actual value is positive but the model's predicted value is negative.*True Negative*when the actual value is negative and the model's predicted value is also negative.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ No.of\ prediction}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

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

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

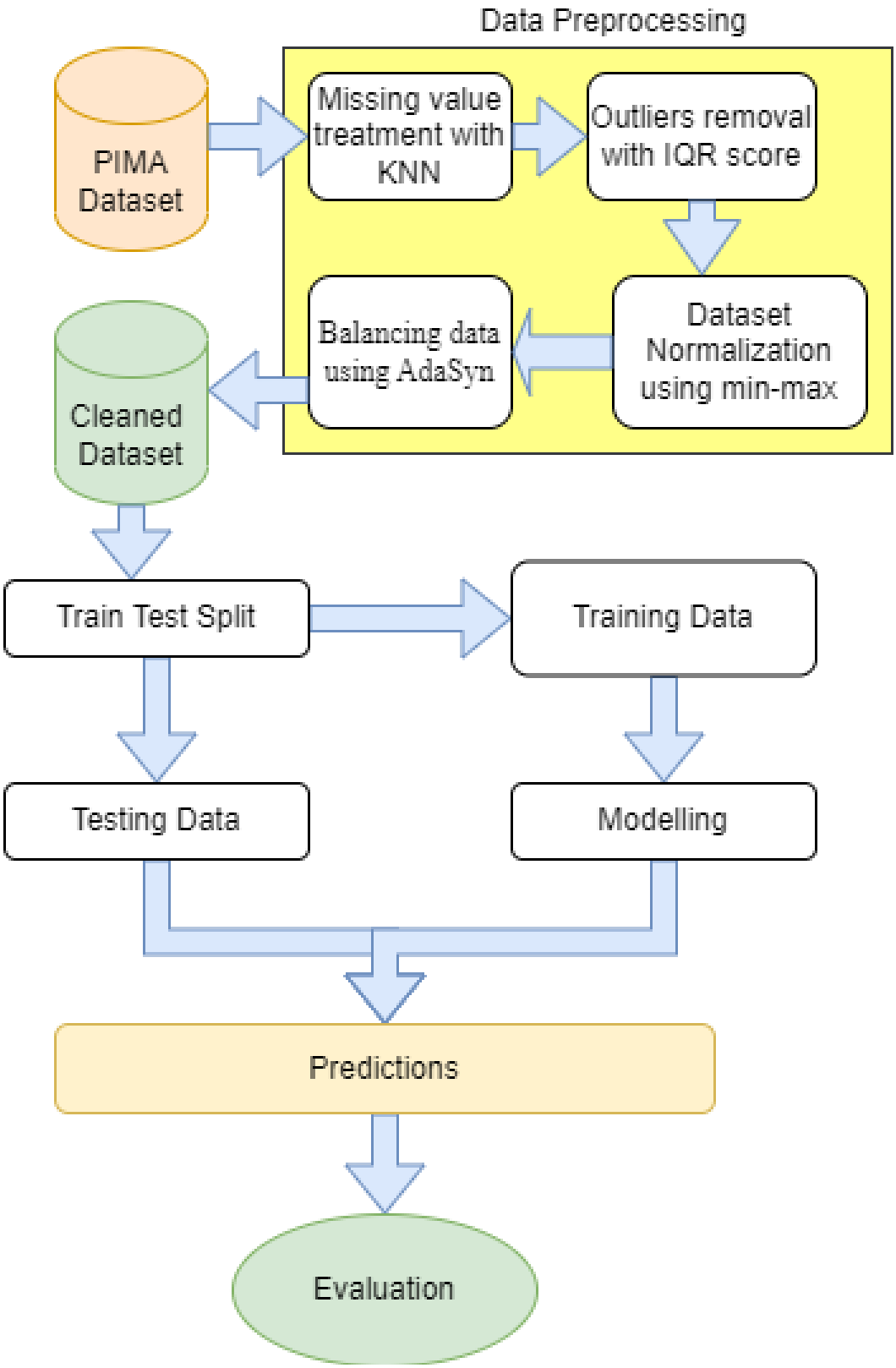


Figure 5 Process used in this study

All eight algorithms were compared for their precision, accuracy, f1 score, and recall, which is shown in **Table 3 and Table 4**

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

Table 3 Performance of eight machine learning algorithms without oversampling

S.No.	ML Algorithm	Accuracy	Precision	Recall	F1 Score
1	Linear Discriminant Analysis	78%	48%	69%	57%
2	XGB	77%	52%	66%	58%
3	Random Forest Classifier	75%	48%	61%	54%
4	Logistic Regression	76%	44%	66%	53%
5	K Nearest Neighbors	79%	58%	70%	63%
6	Gaussian Naïve Bayes	75%	54%	60%	57%
7	Support vector classifier	79%	48%	74%	58%
8	Decision Tree	72%	58%	56%	51%

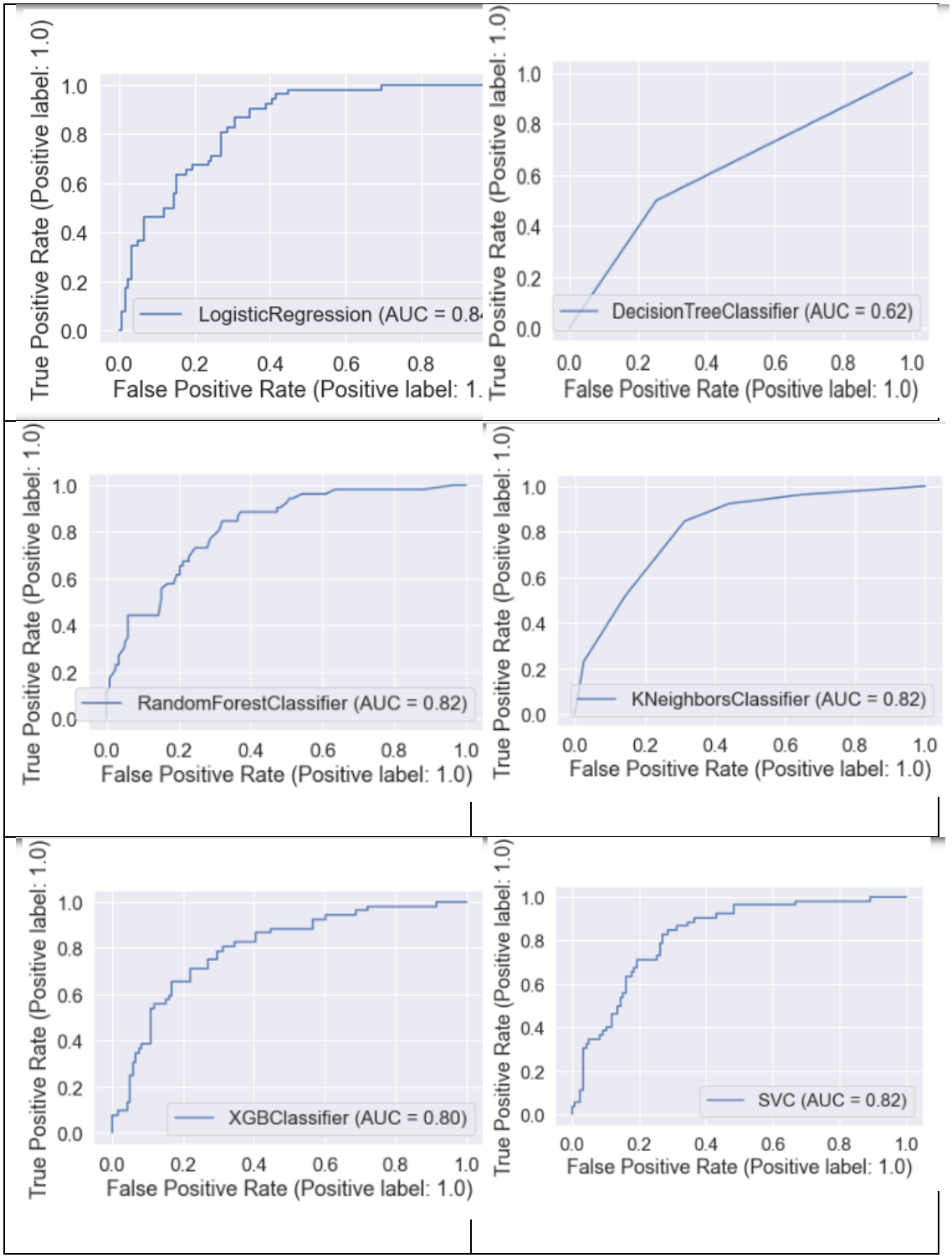
Table 4 Performance of eight machine learning algorithms with oversampling (AdaSyn)

S.No.	ML Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
1	Linear Discriminant Analysis	79%	71%	64%	67%	0.85
2	XGB	78%	65%	63%	64%	0.80
3	Random Forest Classifier	75%	63%	58%	61%	0.82
4	Logistic Regression	75%	67%	57%	62%	0.84
5	K Nearest Neighbors	74%	85%	54%	66%	0.82
6	Gaussian Naïve Bayes	74%	67%	56%	61%	0.83
7	Support vector classifier	74%	73%	55%	63%	0.82
8	Decision Tree	68%	50%	47%	49%	0.62

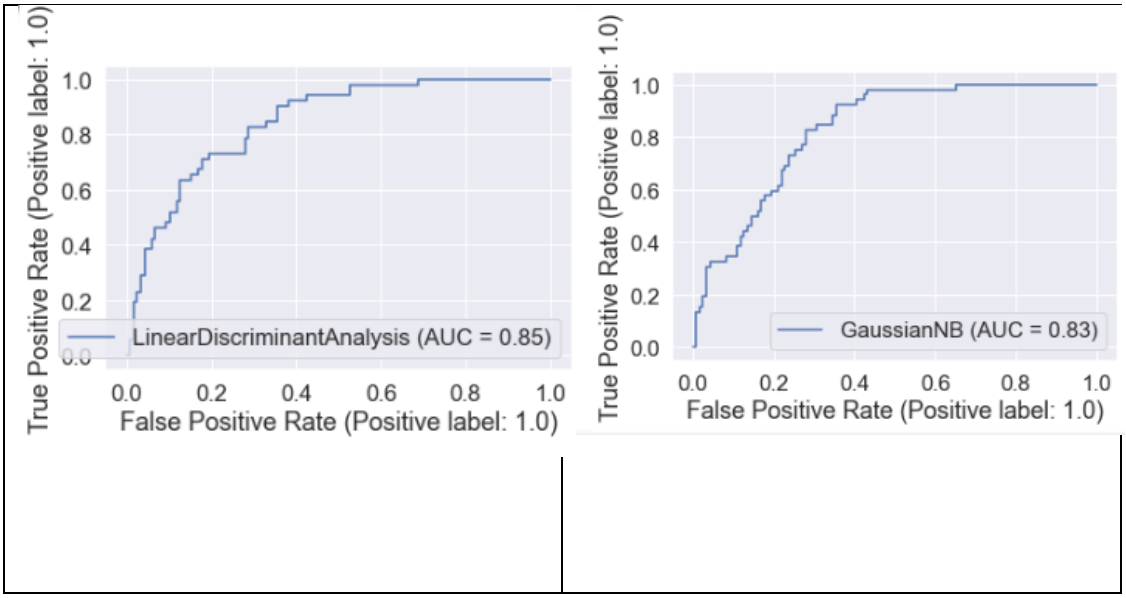
The AUC-ROC curve is sometimes referred to as the Area under curve and the receiver operator characteristics curve. The AUC-ROC curve is used to assess classification algorithm performance at various threshold levels. The AUC is a measure of separability, while the ROC is a probability curve. This indicates the model's ability to differentiate across classes. The AUC of a model reveals how effectively it predicts 0 classes as 0 and 1 class as 1. The greater the AUC value, the better. For example, the AUC value is inversely proportional to a model's capacity to differentiate between individuals who have the illness and those who do not. The ROC curve with TPR vs FPR is depicted on the y-axis and x-axis, respectively. The ROC curve of all the algorithms is shown in tabular form in **Table 5**

ISSN PRINT 2319 1775 Online 2320 7876
Research paper , UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

Table 5- AUC-ROC curve of eight algorithms



ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022



4. RESULTS

Linear Discriminant Analysis was the best-performing algorithm with an accuracy of 79%, Precision of 71%, followed by XGB with 78% accuracy. Logistic regression and Random Forest's performed almost similarly with 75% accuracy and very similar with other metrics like precision, F1 score, and recall.

The accuracy of K nearest neighbors, Gaussian Naïve Bayes, and Support Vector Classifier, were 74%, out of them K nearest neighbor achieved the highest precision of 85%. In terms of AUC score Linear Discriminant Analysis got the highest value of 0.85 followed by Logistic regression at 0.84, Gaussian Naïve Bayes at 0.83 K Nearest Neighbour, Random Forest, and Support vector Classifier at 0.82.XGB was at 0.80 and the decision tree was at 0.62. The decision tree performs poorly with the lowest accuracy, precision, recall, F1 score and AUC.

It was found that upon oversampling the dataset the precision and F1 Score were greatly increased.A detailed comparison of the eight machine-learning algorithms is represented in **Figure 6**.

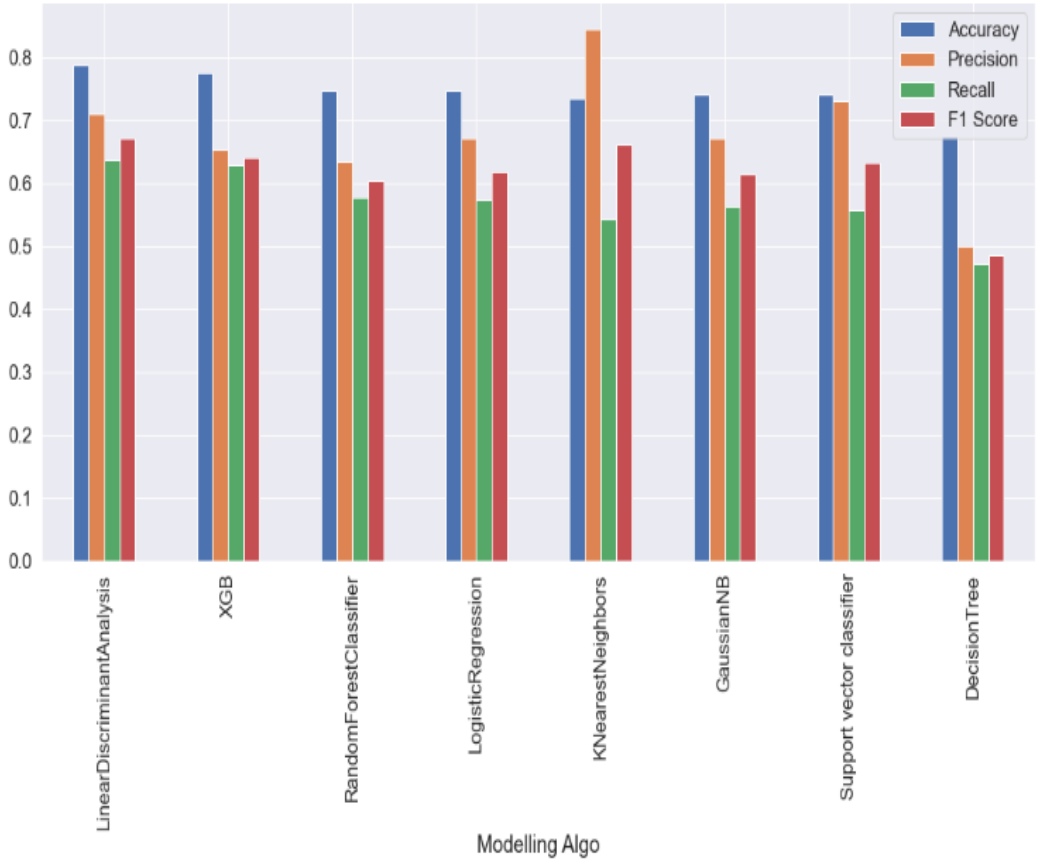


Figure 6 Performance analysis of eight algorithms

5. CONCLUSION AND FUTURE WORK

On analyzing and comparing the performance of eight machine learning classifiers Random Forest, Logistic Regression, Linear Discriminant Analysis, Decision Tree, Naïve Bayes, K nearest neighbor, Support vector machine, and Xtreme Gradient Boosting it was found that Linear Discriminant Analysis achieved the highest accuracy of 79% followed by XGB with 78% accuracy. The performance of Random Forest and Logistic Regression was almost the same. In terms of precision K Nearest Neighbor was the best performer with 85% followed by the support vector classifier. Decision Tree achieves a pretty low score in all the performance metrics.The comparison and analysis of eight machine learning algorithms were done both with oversampling and without oversampling and it was found that oversampling increases the precision and F1 score of all the algorithms but this was not the case with the decision tree as it performs well without oversampling.The AUC score of Linear Discriminant Analysis was the maximum among the rest at 0.85 followed by Logistic regression and Gaussian Naïve Bayes. It was also evident that dropping the feature did not reduce the performance.
Future work: This methodology can also be implemented with another dataset of different diseases. The performance can be further improved by using a larger dataset, or a dataset, with no missing values or outliers. The dataset with additional informative features like lifestyle, calorie intake, etc. also has the scope for further improvement in predictions.

References

1. D. K. Choubey, M. Tech, and Y. K. Rathore, "IJERT-A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques [Online]. Available: www.ijert.org.
2. A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in Procedia Computer Science, 2019, vol. 165, pp. 292–299, doi: 10.1016/j.procs.2020.01.047.
3. L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck, "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation," Arch. Comput. Methods Eng., vol. 29, no. 1, pp. 313–333, Jan. 2022, doi: 10.1007/s11831-021-09582-x.
4. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, Nov. 2018, doi: 10.3389/fgene.2018.00515.
5. N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," Cluster Comput., vol. 22, pp. 1–9, Jan. 2019, doi: 10.1007/s10586-017-1532-x.
6. M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare 1.
7. M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," Heal. Inf. Sci. Syst., vol. 8, no. 1, Dec. 2020, doi: 10.1007/s13755-019- 0095-z.
8. S. Saru, "ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING," 2019. [Online]. Available: https://ssrn.com/abstract=3368308.
9. S. Srivastava, L. Sharma, V. Sharma, A. Kumar, and H. Darbari, "Prediction of diabetes using artificial neural network approach," in Lecture Notes in Electrical
10. J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.
11. S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," in Smart Innovation, Systems and Technologies, 2021, vol. 153, pp. 399–409, doi: 10.1007/978-981-15-6202-0_41.
12. M. S. Padmavathi* and C. P. Sumathi " A New Method of Data Preparation for Classifying Diabetes Dataset" , Indian Journal of Science and Technology, Vol 12(22) DOI: 10.17485/ijst/2019/v12i22/144929, June 2019
13. Razieh Asgarnezhad1*, Karrar Ali MohsinAlhameedawi,"Improving of Diabetes Diagnosis using Ensembles and Machine Learning Methods,"Majlesi Journal of Telecommunication Devices Vol. 11, No. 1, March 2022
14. J. Omana1 · M. Moorthi2," Predictive Analysis and Prognostic Approach of Diabetes Prediction with Machine Learning Techniques",Wireless Personal Communications ,Feb 2021
15. Ahsan, M.M.; Luna, S.A.;Siddique, Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare 2022, https://doi.org/10.3390/healthcare10030541
16. Ahmad Shaker Abdalrada1,2 · Jemal Abawajy2 · Tahsien Al- Quraishi1,2 · Sheikh Mohammed Shariful Islam3,"Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study", Journal of Diabetes & Metabolic Disorders 21:251–261,January 2022 https://doi.org/10.1007/s40200-021-00968-z
17. Ashwini Tuppad1 · Shantala Devi Patil1," Machine learning for diabetes clinical decision support: a review" Advances in Computational Intelligence (2022) 2:22 April,2022. https://doi.org/10.1007/s43674-022-00034-y