IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 12, No. 3, September 2023, pp. 1302~1311
ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i3.pp1302-1311

Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage

Mohammad Atif[1], Faisal Anwer[1], Faisal Talib[2], Rizwan Alam[3], Faraz Masood[1]
[1]Department of Computer Science, Aligarh Muslim University, Aligarh, India
[2]Department of Mechanical Engineering, Aligarh Muslim University, Aligarh, India
[3]United World School of Computational Intelligence, Karnavati University, Gandhinagar, India

Article Info

Article history:

Received May 31, 2022
Revised Aug 29, 2022
Accepted Sep 20, 2022

Keywords:

AdaBoost
Diabetes prediction
Machine learning classifiers
Gradient boosting
Random forest

ABSTRACT

Diabetes is the most common disease all over the world and it must be detected early to receive proper treatment, which can prevent the condition from becoming more severe. Automated detection plays an essential role in diabetes early diagnosis. Over the last few decades, many complicated machine learning algorithms and data analysis approaches have been applied for diabetes prediction. To determine the best model for early-stage diabetes prediction, ten different machine learning classifiers have been used in this study. These models were evaluated in terms of accuracy, precision, specificity, recall, F1-score, negative predictive value (NPV), false positive rate (FPR), rate of misclassification, and receiver operating characteristics (ROC) curve. The experimental findings indicated that all of the models performed well. Gradient boosting (GB), with 97.2% accuracy, is observed to show the best performance on the early-stage diabetes risk prediction dataset. Random forest (RF) and Adaboost performed similarly to the GB; however, RF and Adaboost's precision was not as good as the GB precision (GB's).

Corresponding Author:

Mohammad Atif
Department of Computer Science, Aligarh Muslim University
Aligarh, India
Email: atif.sidau@gmail.com

1. INTRODUCTION

Diabetes is a metabolic disease that affects millions of individuals throughout the world. Every year, the rate of occurrence rises drastically. In 2014, there were around 387 million diabetics worldwide. These figures will have more than doubled by 2030, according to the World Health Organization (WHO) [1]. Diabetes-related problems in several vital organs of the body can be lethal if left untreated. Diabetes must be detected early to receive proper treatment, which can prevent the condition from escalating to severe problems [2]. As a result, better-accurate automated detection plays an essential role in diabetes early diagnosis [3]. Over the last few decades, many complicated machine learning algorithms and data analysis approaches have been developed in the medical industry, among other fields [4]. For applications like disease diagnosis, brain tumor detection, breast cancer detection, and therapy, machine learning technique has become indispensable tool in the medical profession [5].

Diabetes mellitus (DM), widely known as diabetes, is a collection of metabolic diseases caused primarily by aberrant insulin production [6]. Blood sugar levels rise when cells and/or the pancreas fail to produce enough insulin, damaging multiple organs, including the eyes, kidneys, and nerves. Because of this reason, diabetes is also known as the "silent killer". Diabetes is classified into type I diabetes, type II diabetes, and gestational diabetes [7]. The pancreas secretes very little or no insulin in type I diabetes. Type I diabetes attacks the pancreatic cells, causing them to stop operating. Type I diabetes affects 5% to 10% of the population and can appear in any age group, including childhood and adolescence [8]. Type II diabetes accounts for over 90% of all diabetes cases globally. It develops when the body's insulin production is insufficient [9]. Both adults and children can develop type II diabetes. Gestational diabetes mellitus (GDM) is a third kind of diabetes similar to type II diabetes in that it is caused by an insufficient balance of insulin secretion and responsiveness. This medical issue develops over time as a result of excessive blood pressure and hypertension. Gestational diabetes affects about 2–10% of all pregnant women, and it can progress or disappear after delivery [10]. Diabetes can signal the onset of other diseases. Researchers around the globe are working tirelessly to tackle the illness by creating effective prediction and detection tools and viable therapies [11]. Machine learning approaches play an essential part in predicting this disease in this context. The best-preferred methodology for the categorization of labeled data is classification, which is a supervised machine learning technique [12].

The literature study in this section reviews some of the more well-known early efforts on diabetes prediction using variously supervised and unsupervised machine learning algorithms and also comparisons between different classification methods. A comparison of random forest (RF), K-means clustering, and artificial neural network (ANN) approach for diabetes prediction had been done in [13]. The ANN method had the highest accuracy of 75.7%. On "Pima Indian diabetes dataset" (PIDD), [14], [15] discovered that the Naïve Bayes (NB) classifier outperforms the support vector machine (SVM), NB, and decision tree (DT) machine learning algorithms, with an accuracy of 76.30%. Sadeghi *et al.* [16], utilized the methods of deep neural network (DNN), extreme gradient boosting (XGBoost), and RF for predicting diabetes in Tehran Lipid and Glucose Study (TLGS) cohort data in which DNN outperformed with the highest accuracy. On PIDD, [17] applied machine learning techniques in which generalized boosted regression modeling showed the highest accuracy of 90.91%. Zecchin *et al.* [18] and Reddy *et al.* [19] employed a neural network (NN) and a polynomial model to predict short-term blood glucose. This plan requires continual monitoring and sample-giving, which is time-consuming.

The previous brief literature study and the summarised literature analysis are presented in Table 1. Analysis of the literature demonstrates that there is no single algorithm that is best for all issues, several factors, such as the organization and size of the dataset, are critical. Some of the classifiers used include DT, RF, GB, principal component analysis (PCA), k-nearest neighbor (KNN), expectation maximization (EM), NN, logistic regression (LR), radial basis function (RBF), multifactor dimensionality reduction (MDR).

Table 1. List of relevant works from various literature

References	Classifiers **a**pplied	Claimed **r**esult	Constraints	Highest classification accuracy (%)	Dataset **u**sed
Zou *et al.*[20]	DT, RF, and NN	RF predicts diabetes with 80% sensitivity, 89% specificity, and 85% accuracy.	Only the glucose index did well. For effective results, more indexes are required.	80.84 (RF)	Luzhou and Pima Indian diabetes dataset
Heydari *et al.* [21]	DT, SVM, ANN, and Bayesian networks 5NN	ANN outperforms with an accuracy of 97.44%.	Doctors classify data mining. Experts should review disease data mining.	97.44 (ANN)	Dataset from Tabriz University of Medical Sciences, Iran
Wu *et al.* [22]	K-means clustering and LR	Improved K-means algorithm performs well	Data preparation is time-consuming.	95.42 (K-means)	Pima Indian diabetes dataset
Nilashi *et al.*[23]	PCA-KNN, PCA-SVM, EM, PCA-fuzzy rule-based	EM-PCA-fuzzy rule-based performs well	Compared to enormous healthcare data, this dataset's data is simple.	92.9 (EM-PCA)	Pima Indian diabetes dataset
Husain and Khan [24]	LR, KNN, RF, GB, ensemble method	Ensemble method effectively predict diabetes with 75% accuracy.	There are concerns with under-fitting, over-fitting, and high training time overhead,	75 (ensemble method)	National Health and Nutrition Examination Survey (NHANES) 2013-14 dataset
Sarwar and Sharma [25]	ANN, KNN, and NB	ANN predicted 96% accurately, followed by KNN (91%) and NB (95%).	This study's dataset could be expanded to include clinical factors to compare them.	96 (ANN)	500 participants were picked at random from various social groups.
Kaur and Kumari[26]	ANN, k-NN, SVM-linear, SVM-RBF and MDR	k-NN and SVM linear identify diabetes best.	a limited number of parameters	89 (Linear kernel SVM)	Pima Indian diabetes dataset

Diabetes prediction is difficult. This study uses the early-stage diabetes dataset (ESDD) to determine the best classifier model among ten for predicting outcomes. For a thorough evaluation, nine measures were used: confusion matrix, classification accuracy, precision, recall/sensitivity/ true positive rate (TPR), false positive rate (FPR), negative predictive value (NPV), rate of misclassification, F1-score, and receiver operating characteristics (ROC) curve. Contributions: i) to determine the best model for early-stage diabetes prediction, ten different machine learning classifiers, including KNN, ANN, DT, stochastic gradient descent (SGD), RF, SVM, GB, NB, AdaBoost, and LR, have been used; and ii) these models were evaluated in terms of accuracy, precision, specificity, recall, F1-score, NPV, FPR, rate of misclassification, and ROC curve.

The following is how the paper is organized: section 1 discusses the introduction and several connected literature surveys. Section 2 contains a description of various classification methods. The numerous performance evaluation measures of classifiers are depicted in detail in section 3. The experimental approach, its setup, and the description of the dataset have been detailed in section 4. In section 5, obtained results are discussed. Finally, in section 6, this study comes to a conclusion with implications for the future.

2. DESCRIPTION OF CLASSIFICATION METHODS

The purpose of this section is to provide some background information on ten different classifiers used in this study. This information helps to provide some context for the classification methods. In this section, a concise explanation of each of these ten different classifications is discussed.

2.1. Decision tree classifier

DT are supervised learning systems that can address classification and regression problems, but in general, they are used to resolve classification problems. It is a tree-like layout in which each node represents a feature value check, each branch represents a test activity result, and the tree's leaf nodes represent classifications. It may quickly produce intelligible criteria and classify data with minimal processing.

2.2. k-nearest neighbor classifier

This predictive algorithm is suitable for a lazy learning technique prediction mechanism that generates predictions based on the KNN input. When the predictions of any occurrence are requested, the full process of prediction is completed. The Euclidian distance method is frequently used to determine proximity [27].

2.3. Support vector machine classifier

SVM is based on the approach of finding a hyper-plane to separate binary classes of the dataset. Both linear and nonlinear datasets function well with this approach. When the dataset has a large number of attributes, the SVM performs much better.

2.4. Naive Bayes classifier

For high-dimensional inputs, this classifier prefers to employ the Bayesian theorem. The Bayes theorem is used with strong and independent hypotheses in a NB model. The essential assumption of the NB technique is that a particular characteristic of a class is independent of every other property of that class. This technique yields incredible precision when the underlying premise is false [28].

2.5. Logistic regression classifier

Instead of forecasts, the logistic regression model produces likelihood approximations. For binary classification, this method is appropriate. The likelihood of every event occurring is handled as a linear transformation of a collection of input characteristics in this.

2.6. Random forest classifier

It is a common machine learning approach that uses the results of several DT constructed on different sets of the dataset to produce forecasts. It is a regression and classification classifier. The mean of all the decision tree outcomes is computed in regression, and the voting from the several DT is pooled to get the final result in classification.

2.7. Artificial neural network classifier

ANN is a supervised classification approach. In this, the neural architecture of the human brain is being implemented as a form of software with the aim to simplify and emulate brain activity. ANN is a group of artificial neurons that absorb data, change their internal state, known as activation, and generate output based on the input provided and used activation function [29].

2.8. Adaboost classifier

Adaboost has the benefits of being easier to develop, having fewer variables to choose from, and having high generality. Even if it only gives sub-optimal results and is vulnerable to extremes and inconsistent data [30]. Gradient boosting (GB) is a method for predicting the residuals of previous models by creating new models, which are then integrated to generate a decision boundary.

2.9. Gradient boosting classifier

Boosting algorithms iteratively aggregate weak learners or those who are slightly better than random into strong learners over time. GB is a regression approach that is similar to boosting. GB aims to obtain an estimate of the function that maps samples to their output values by minimizing the estimated value of an error function given a training sample.

2.10. Stochastic gradient descent classifier

SGD is a straightforward method for locating the local minima of a function whose values are tainted by noisy data. It is a popular optimization method in machine learning. By iteratively computing the gradients of an error function on a single training instance or a batch of a few instances and revising the parameters of the model appropriately, this approach minimizes the error of a model [31].

3. CLASSIFIER PERFORMANCE EVALUATION METRICS

To successfully evaluate any effects of the algorithm, specific performance metrics must be defined that may be used to assess the quality of any classification model under evaluation. In this study, we used nine different metrics to evaluate the performance of classifiers. A brief description of these measures is as:

3.1. Confusion matrix

The confusion matrix (CM) is a table that summarizes an evaluation of the efficacy of a classification model. The learning process yields correct results when it has diagonal entries. True positive (TP): training situations in which we hypothesized that the real class was positive. False positive (FP): this shows that the learning system is wrongly recognizing the instances as positive when they are actually negative. True negative (FN): there are certain training situations where the real class is negative, and we hypothesize that the true class is negative. False negative (FN): this shows that the learning system is wrongly classifying the events as negative when they are actually positive. We may assess the classifier's performance using the confusion matrix.

3.2. Classification accuracy

The overall success rate of the classifiers is displayed here. This success rate is expressed as a percentage of all correct predictions. Accuracy can be expressed mathematically as the following equation:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

3.3. Precision

It is one of the important metrics to evaluate the performance of classifiers. It is defined as the ratio of true positive to the sum of true positive and false positive. The formula for calculating precision is as:

$$Precision = (TP) / (TP + FP)$$

3.4. Recall/Sensitivity/TP rate

True positive recall is often known simply as recall. It is a statistic defined as the ratio of true positive results to the total of true positive and false-negative results. Recall can be expressed mathematically as the following equation:

$$Sensitivity = (TP) / (TP + FN)$$

3.5. False positive rate (FPR)

A FP value is a number that is higher than the sum of FP values and the true negative value. This is called the FP rate. The formula for calculating the FP rate is as:

$$FPrate = (FP) / (FP + TN)$$

3.6. Negative predictive value (NPV)

It is also an important metric to evaluate the performance of classifiers. The NPV is defined as the relationship between the TN value and the sum of the TN and FN values. It can be calculated mathematically as:

$$NPV = (TN) / (TN + FN)$$

3.7. Rate of misclassification

It is calculated as the proportion of incorrectly identified samples to the total number of samples. Erroneous classifications can be divided into two categories. If the presence of diabetes disease is misclassified as the absence of diabetes disease, this is a Type-I Error (E1). If a patient's absence of diabetic illness is interpreted as a diagnosis of diabetes, this is a Type-II Error (E2).

$$RateMisclassification\,(E1 + E2) = (FP + FN) / (TP + TN + FP + FN)$$

3.8. F1-score

The F-measure is another name for the F1-score. This word refers to the value that is calculated by taking the harmonic mean of the accuracy and recall. Its value of 0 indicates the worst performance, whereas F-measure equal to 1 indicates the best performance.

$$F1 - Score = 2 * precision * recall / (precision + recall)$$

3.9. ROC curve

ROC FPR data are plotted against TPR values on a graph with the x-axis representing FPR and the y-axis representing the TPR values. This statistic evaluates a model's ability to discriminate between classes and how effective it is at doing so. The larger the area under the curve (AUC), the better the classifier will be at distinguishing between individuals with and without the condition.

4. EXPERIMENTAL SETUP

Experiments were used to test the effectiveness and efficiency of various machine-learning algorithms and classifiers. Orange machine learning and data mining toolbox was used to test the classifiers. This toolbox comprises ML algorithms for classification, regression, data preprocessing, association rules, and clustering. It has a graphical data analysis interface. It allows widgets to be used as data processing points on a canvas, related by workflow lines. Figure 1 illustrates the orange data mining toolkit's experimental widget parts and workflow lines shows in appendix.

4.1. Dataset description

This empirical study uses the Early Stage Diabetes Risk Prediction dataset from the UCI machine learning archive. There are 520 instances in this dataset, and each instance has 17 attributes. There are 400 positive samples and 120 negative samples in total. Table 2 summarizes the dataset's features and their Chi2 values. For study reasons, all instances in the dataset without diabetes were assigned to the NEGATIVE (0) class, whereas instances with diabetes were assigned to the POSITIVE (1) class.

Table 2. Dataset features information and Chi2 values for different features

S.No.	Attribute name	Abbreviation	Values	Chi^2	p
1	Age	age	20-65	16.05	0.001
2	Sex	sex	Male/Female	104.94	0
3	Weakness	wk	Yes/No	30.77	0
4	Polydipsia	pd	Yes/No	218.84	0
5	Partial Paresis	pp	Yes/No	97.17	0
6	Polyuria	pu	Yes/No	230.6	0
7	Polyphagia	pph	Yes/No	61	0
8	Muscle Stiffness	ms	Yes/No	7.8	0.005
9	Delayed Healing	dh	Yes/No	1.15	0.284
10	Alopecia	alp	Yes/No	37.21	0
11	Obesity	obs	Yes/No	2.71	0.1
12	Visual Blurring	vb	Yes/No	32.84	0
13	Sudden Weight Loss	swl	Yes/No	99.11	0
14	Genital Thrush	gt	Yes/No	6.32	0.012
15	Irritability	irr	Yes/No	46.63	0
16	Itching	ich	Yes/No	0.09	0.76
17	Class	class	Positive/Negative.	-	-

We selected the Chi-square test for feature selection because the dataset contains mostly categorical attributes, except 'Age,' and the target feature is categorical. We found that Polyuria and Polydipsia had the highest Chi-squared of all 17 attributes. Table 2 also shows poor Chi-squared values for itching and delayed healing. Based on the Chi-squared test, we rejected 'itching' and 'delayed healing.'

5. RESULTS AND DISCUSSION

The experiment used the Early-Stage Diabetes Risk Prediction dataset to detect diabetes disease using random samples, stratified shuffle split, and tenfold cross-validation with an 80% training dataset size. A confusion matrix measures classifier performance. Table 3 summarizes confusion matrices for all ten classifiers studied. According to the results of all the ten potential classifiers used in this study shows in Table 4, RF and Adaboost both achieved a classification accuracy of 97%, whereas ANN, SGD, SVM, Logistic regression, NB, and DT achieved a classification accuracy of 95.6%, 90.2%, 96.3%, 90.2%, 88.4%, and 94.7%, respectively. KNN has the lowest accuracy of 88.2%, while GB excels with a 97.2% accuracy (refer to Table 4).

Table 3. Summarised depiction of confusion matrices for all ten classifiers

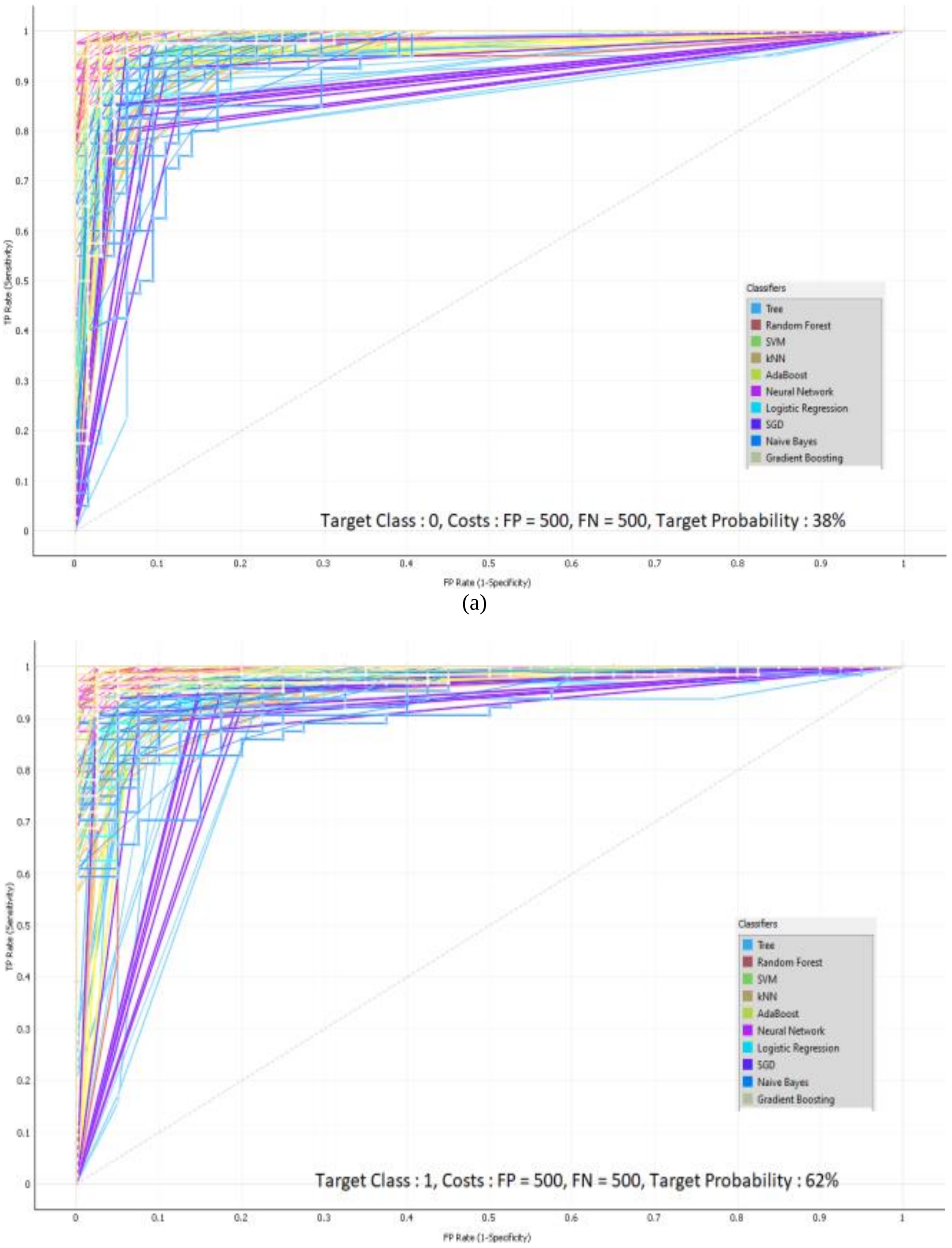
	GB	Tree	RF	SVM	KNN	AdaBoost	ANN	LR	SGD	NB
TN	199	198	199	195	197	200	197	179	180	180
FP	1	2	1	5	3	0	3	21	20	20
FN	1	9	1	6	25	0	3	22	15	45
TP	319	311	319	314	295	320	317	298	305	275

Table 4. Performance statistics of Ten classifiers

Model	AUC	F1	Precision	Recall	Accuracy (%)	NPV (%)	FPR (%)	RMC (%)
k-NN	0.954	0.883	0.894	0.882	88.2	88.74	1.5	5.385
Tree	0.943	0.947	0.947	0.947	94.7	95.65	1	2.115
SVM	0.991	0.962	0.962	0.963	96.3	97.01	2.5	2.115
SGD	0.896	0.902	0.902	0.902	90.2	92.31	10	6.731
RF	0.996	0.966	0.967	0.966	96.6	99.5	0.5	0.385
NN	0.991	0.956	0.956	0.956	95.6	98.5	1.5	1.154
NB	0.953	0.885	0.891	0.884	88.4	80	10	12.5
LR	0.964	0.902	0.902	0.902	90.2	89.05	10.5	8.269
GB	0.988	0.972	0.972	0.972	97.2	99.5	0.5	0.385
AB	0.967	0.966	0.966	0.965	96.5	100	0	0

The F1-score is the sum of the recall and precision ratios; a higher value implies a model's classification capability; the classifiers LR and SGD have the same F1-score value, 0.902. Similarly, the classifiers RF and Adaboost produce equal F1-score value, i.e., 0.966; however, ANN, DT, NB, and Logistic regression demonstrates the F1-score value of 0.956, 0.947, 0.885, and 0.902. KNN offered the lowest F1-SCORE of 0.883 while GB outperformed with an F1-score of 0.972. In terms of the rate of misclassification provided by classifiers, Adaboost surpassed all other classifiers with a 0% rate of misclassification, while NB has the lowest RMC of 12.5%. For other classifiers, RMC ranged from 0.4 % to 8.3 %. In terms of the NPV provided by classifiers, Adaboost surpassed all other classifiers with 100%, while NB has the lowest NPV of 80%. For other classifiers, NPV ranged from 88.7 % to 99.5 %.

Except for SGD (0.896), DT (0.943), NB (0.953), KNN (0.954), LR (0.964), Adaboost (0.967), and GB (0.988), the other two classifiers, SVM and ANN, have an AUC value of 0.99, indicating greater classification performance shows in Table 3. In terms of the patient, the negative class prediction is the most sensitive; if the classification is erroneous, the negative patient condition may become dangerous, as the patient would not be treated for a negative case. AdaBoost has an FPR of 0%, while all other classifiers have FPRs ranging from 1% to 11%, as shown in Table 4. On the early-stage diabetes risk prediction dataset, the AdaBoost classifier is expected to produce the best results. The ROC curves in Figures 2(a) and 2(b) illustrate the primary reason for the enhanced results obtained by using the AdaBoost classifier. The fundamental objective of employing the AdaBoost classifier is to achieve superior results. Adaboost makes it possible to merge several "weak classifiers" into a single "strong classifier." The weak learners in AdaBoost are decision trees with a single split, often known as decision stumps. The AdaBoost classifier works by giving more weight to cases that are hard to categorize and less weight to those that are already well-classified. As a result, it is well suited to the dataset used in this study.



(a)

(b)

Figure 2. ROC analysis of (a) negative class and (b) positive class

6. CONCLUSION

Diabetes is a disease that affects a large number of people. Diabetes may be detected early, which not only lowers treatment costs but also saves lives. In the prediction of disease, a reliable prediction system is quite helpful. The chi-squared test is used for feature selection. This research also implies that the Chi-squared test can be used for feature selection for small datasets and that it is preferable to select attributes without medical domain knowledge. Thus, with a limited number of parameters, we were able to achieve the higher performance standards of Machine Learning models using this feature selection strategy. To determine the best model for early-stage diabetes prediction, ten different machine learning classifiers, including KNN, ANN, DT, SGD, RF, SVM, GB, NB, AdaBoost, and LR, have been used. These models were evaluated in terms of accuracy, precision, specificity, recall, F1-score, NPV, FPR, rate of misclassification, and ROC curve. The experimental findings indicated that all of the models performed well. GB, with 97.2% accuracy, is the best performance on the Early-Stage Diabetes Risk Prediction Dataset. RF and Adaboost performed similarly to the GB; however, RF and Adaboost's precision was not as good as the GB's. With an accuracy of 88.2%, the KNN classifier was the worst performer in classification. In the center of the performance range of RF and KNN were SVM, RF, ANN, SGD, and LR. When it came to Recall and F1 score, GB topped the pack with both of those numbers at 0.972. Because our dataset is an example of an unbalanced dataset, the F1 score offers a better understanding of our models' performance. The F1 score achieves a good mix of precision and recall. It can also be noted that RF has the greatest AUC value (0.996). AUC of this magnitude implies that RF is a trustworthy model. Finally, we may conclude that, among all the classifiers tested in this work, GB and RF are the best for predicting early-stage diabetes. As this study evaluates the performance of models individually, and the performance of the combination of these models is yet to be explored. Future work will focus on developing prediction models using the ensemble methods like soft or hard voting classifiers or the development of hybrid classifiers, following to enhance those models for better performance.

APPENDIX

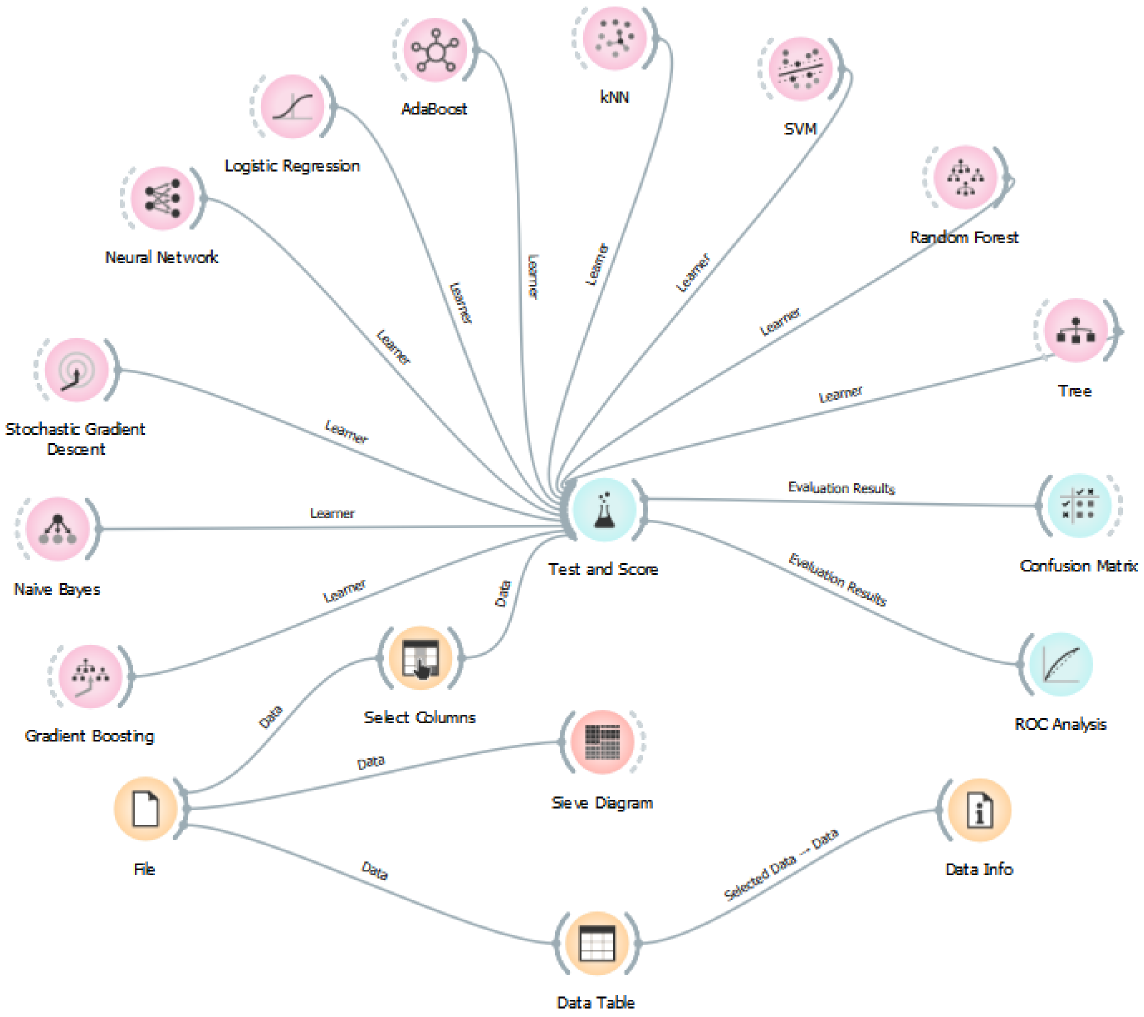


Figure 1. Orange data mining toolkit's experimental setup

REFERENCES

[1] M. Maniruzzaman *et al.*, “Comparative approaches for classification of diabetes mellitus data: machine learning paradigm,” *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 23–34, Dec. 2017, doi: 10.1016/j.cmpb.2017.09.004.

[2] G. Swapna, R. Vinayakumar, and K. P. Soman, “Diabetes detection using deep learning algorithms,” *ICT Express*, vol. 4, no. 4, pp. 243–246, Dec. 2018, doi: 10.1016/j.icte.2018.10.005.

[3] J. Chaki, S. Thillai Ganesh, S. . Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: a systematic review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3204–3225, Jun. 2022, doi: 10.1016/j.jksuci.2020.06.013.

[4] S. S. Reddy, N. Sethi, and R. Rajender, “Mining of multiple ailments correlated to diabetes mellitus,” *Evolutionary Intelligence*, vol. 14, no. 2, pp. 733–740, Jun. 2021, doi: 10.1007/s12065-020-00432-6.

[5] Y. K. Afework and T. G. Debelee, “Detection of bacterial wilt on enset crop using deep learning approach,” *International Journal of Engineering Research in Africa*, vol. 51, pp. 131–146, Nov. 2020, doi: 10.4028/www.scientific.net/JERA.51.131.

[6] A. D. Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 32, no. Supplement_1, pp. S62--S67, Jan. 2009, doi: 10.2337/dc09-S062.

[7] Z. Tao, A. Shi, and J. Zhao, “Epidemiological perspectives of diabetes,” *Cell Biochemistry and Biophysics*, vol. 73, no. 1, pp. 181–185, Feb. 2015, doi: 10.1007/s12013-015-0598-4.

[8] J. L. Chiang, M. S. Kirkman, L. M. B. Laffel, and A. L. Peters, “Type 1 diabetes through the life span: a position statement of the American Diabetes Association,” *Diabetes Care*, vol. 37, no. 7, pp. 2034–2054, Jun. 2014, doi: 10.2337/dc14-1140.

[9] K. Azbeg, M. Boudhane, O. Ouchetto, and S. Jai Andaloussi, “Diabetes emergency cases identification based on a statistical predictive model,” *Journal of Big Data*, vol. 9, no. 1, pp. 1–25, Mar. 2022, doi: 10.1186/s40537-022-00582-7.

[10] S. S. Reddy, R. Rajender, and N. Sethi, “A data mining scheme for detection and classification of diabetes mellitus using voting expert strategy,” *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 23, no. 2, pp. 103–108, Jul. 2019, doi: 10.3233/KES-190403.

[11] V. Kumar, “Evaluation of computationally intelligent techniques for breast cancer diagnosis,” *Neural Computing and Applications*, vol. 33, no. 8, pp. 3195–3208, Jul. 2021, doi: 10.1007/s00521-020-05204-y.

[12] R. Maronna, “Charu C. Aggarwal and Chandan K. Reddy (eds.): data clustering: algorithms and applications,” *Statistical Papers*, vol. 57, no. 2, pp. 565–566, Apr. 2016, doi: 10.1007/s00362-015-0661-7.

[13] T. M. Alam *et al.*, “A model for early prediction of diabetes,” *Informatics in Medicine Unlocked*, vol. 16, no. 1, p. 100204, 2019, doi: 10.1016/j.imu.2019.100204.

[14] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” *Procedia Computer Science*, vol. 132, no. 1, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[15] S. S. Reddy, N. Sethi, and R. Rajender, “A comprehensive analysis of machine learning techniques for incessant prediction of diabetes mellitus,” *International Journal of Grid and Distributed Computing*, vol. 13, no. 1, pp. 1–22, 2020, doi: 10.33832/ijgdc.2020.13.1.01.

[16] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, “Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods,” *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/s12911-022-01775-z.

[17] K. C. Howlader *et al.*, “Machine learning models for classification and identification of significant attributes to detect type 2 diabetes,” *Health Information Science and Systems*, vol. 10, no. 1, pp. 1–13, Feb. 2022, doi: 10.1007/s13755-021-00168-2.

[18] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, “Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1550–1560, Jun. 2012, doi: 10.1109/TBME.2012.2188893.

[19] S. S. Reddy, N. Sethi, and R. Rajender, “A review of data mining schemes for prediction of diabetes mellitus and correlated ailments,” in *Proceedings - 2019 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019*, Sep. 2019, pp. 1–5, doi: 10.1109/ICCUBEA47591.2019.9128880.

[20] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” *Frontiers in Genetics*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[21] M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, “Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran,” *International Journal of Diabetes in Developing Countries*, vol. 36, no. 2, pp. 167–173, Apr. 2016, doi: 10.1007/s13410-015-0374-4.

[22] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” *Informatics in Medicine Unlocked*, vol. 10, no. 1, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.

[23] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, “An analytical method for diseases prediction using machine learning techniques,” *Computers and Chemical Engineering*, vol. 106, pp. 212–223, Nov. 2017, doi: 10.1016/j.compchemeng.2017.06.011.

[24] A. Husain and M. H. Khan, “Early diabetes prediction using voting based ensemble learning,” in *Communications in Computer and Information Science*, vol. 905, Springer Singapore, 2018, pp. 95–103.

[25] A. Sarwar and V. Sharma, “Comparative analysis of machine learning techniques in prognosis of type II diabetes,” *AI and Society*, vol. 29, no. 1, pp. 123–129, Apr. 2014, doi: 10.1007/s00146-013-0456-0.

[26] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” *Applied Computing and Informatics*, vol. 18, no. 1–2, pp. 90–100, Jul. 2020, doi: 10.1016/j.aci.2018.12.004.

[27] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2888, Springer Berlin Heidelberg, 2003, pp. 986–996.

[28] M. Di Marzio and C. C. Taylor, “Kernel density classification and boosting: an L2 analysis,” *Statistics and Computing*, vol. 15, no. 2, pp. 113–123, Apr. 2005, doi: 10.1007/s11222-005-6203-8.

[29] J. J. Hopfield, “Artificial neural networks,” *IEEE Circuits and Devices Magazine*, vol. 4, no. 5, pp. 3–10, Sep. 1988, doi: 10.1109/101.8118.

[30] A. Kadiyala and A. Kumar, “Applications of Python to evaluate the performance of decision tree-based boosting algorithms,” *Environmental Progress and Sustainable Energy*, vol. 37, no. 2, pp. 618–623, Mar. 2018, doi: 10.1002/ep.12888.

[31] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: stability of stochastic gradient descent,” in *International conference on machine learning*, Jun. 2016, pp. 1225–1234.

BIOGRAPHIES OF AUTHORS

Mohammad Atif received a master's degree in computer application from the Department of Computer Science, Aligarh Muslim University. He is currently pursuing his Ph.D. at the Department of Computer Science, Aligarh Muslim University (AMU). Before joining AMU, he worked as a Senior Engineer at Tata Consultancy Services (TCS) from 2012 to 2017. He has published several research papers in international/national conferences and journals. His research interests include machine learning, artificial intelligence, bioinformatics, and data mining. He can be contacted at email: atif.sidau@gmail.com.

Faisal Anwer received a master's degree in computer application and a Ph.D. degree in information security from Jamia Millia Islamia, New Delhi. He is currently employed as an Assistant Professor at Aligarh Muslim University's Department of Computer Science. Before joining AMU, he worked at Computer Science Corporation (CSC) in Noida as a Senior Software Engineer. From 2009 to 2010, he worked for CSC in the United Kingdom. He has a number of research articles that have been presented at international and national conferences and journals. Software security, cryptography, program robustness, and machine learning are among his research interests. He can be contacted at email: faisalanwer.cs@amu.ac.in.

Faisal Talib is working as a Professor in the Department of Mechanical Engineering at Aligarh Muslim University. He received a master's degree in Industrial Engineering from AMU and a Ph.D. degree from IIT Roorkee. He has over 100 publications in national and international journals and conferences to his credit, as well as 24 years of teaching and research experience. Total quality management (TQM), sustainability, operations management, multi-criteria decision making (MCDM) methods, and quantitative research are some of his areas of expertise. He can be contacted at email: ftalib77@gmail.com.

Rizwan Alam received his B.Sc., MCA and Ph.D. (Computer Science) from Aligarh Muslim University, Aligarh, UP, India. He has worked with NIELIT, IIT Goa, and NCERT Bhopal. Currently, he is working as Assistant Professor at USCI, Karnavati University, Gandhinagar, Gujarat, India. His research interests include learning analytics and healthcare analytics. He can be contacted at email: riz.alig@gmail.com.

Faraz Masood received his Master of Computer Application degree in 2015 from the Department of Computer Science, Aligarh Muslim University. He is currently pursuing the Ph.D. degree in Computer Science at Aligarh Muslim University, as well. From 2015 to 2017, he was working as a System Engineer in Tata Consultancy Services in New Delhi, India. He has published several research papers in international/national conferences and journals. His research interests include blockchain technology, distributed ledger technology, machine learning and data mining. He can be contacted at email: fmasood@myamu.ac.in.