








Analysis of the Performance of Data Mining Classification Algorithm for Diabetes Prediction

Vijaylakshmi Sajwan¹ , Monisha Awasthi¹ , Prakhar Awasthi², Ankur Goel³ ,
Manisha Khanduja¹ , and Anuj Kumar⁴ 

¹ Uttaranchal School of Computing Sciences, Uttaranchal University, Dehradun, India
uumonishaawasthi@gmail.com

² Department of Computer Science and Engineering, RIT, Bangaluru, Karnataka, India

³ Department of Business Administration, MIET Group, MIT, Meerut, U.P, India

⁴ Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, India

Abstract. The purpose of this paper is to identify solutions for the diagnosis of diabetes disease by analyzing the patterns found in the data using classification algorithms such as Decision Tree, SVM, KNN, Naive Bayes, Random Forest, Neural Network, and Logistic Regression. According to a WHO report, almost 42.2 crores population of the world has diabetes, who are primarily the residents of low and middle income countries, and diabetes is resulting in around 0.15 crores of deaths each year globally [1]. To evaluate and discuss the performance of above-mentioned algorithms, Orange as a data mining tool has been applied. Furthermore, the data set used in this research is the “Pima Indian Diabetic Dataset,” which is obtained from the University of California, Irvine (UCI) Repository of Machine Learning datasets. As this study utilized several classifiers to simulate actual diabetes diagnosis for local and systemic therapy, the results indicated that Logistic Regression outperforms all other classifiers. The experimental data also demonstrated the significance of the suggested model in the study. The disease has been ranked as the fifth-deadliest in the United States, and there is currently no cure in sight. With the advancement of information technology and its continued penetration into the medical and healthcare sectors, diabetes cases and symptoms have become well documented and discussed. The research is original and adds value to the current studies in the same domain as researchers develop a more rapid and efficient method of diagnosing the disease, allowing for more timely treatment of patients.

Keywords: Accuracy · Diabetes · KNN · Logistic regression · Naive bayes · Neural network · Random forest · Support vector machine

1 Introduction

Databases are densely packed with hidden data and are designed to aid in intellectual decision making. Different types of data analysis, such as classification and prediction, are used to make predictions about future data and to describe the data classes. The classification is a process that predicts the labels for categorical classes. The labels for

this class may be discrete or nominal in nature. Classification techniques classify data using a training set and class labels [2]. With the rising prevalence of implementations of various classification and prediction algorithms, there is a need for a central hub that could evaluate the performance of all classification algorithms as well as provide information on which classifier is the best [3].

The objective here is to examine various algorithms of machine learning for classification using the diabetes data set. ORANGE is also used for this purpose. The purpose of this paper is to compare ORANGE classifiers on a diabetes dataset. Such techniques are compared using the results of their ORANGE calculations. We have used the Diabetes dataset because it is a chronic and one of the dramatically increasing metabolic diseases in the world. Diabetes mellitus, more generally referred to as diabetes, is a collection of illnesses (metabolic) characterized by persistently increased levels of sugar in a blood (beyond a certain limit) and caused by lowering the secretion of insulin or biological effects, or both. It is a disorder in which the person's body is not able to metabolize food in an adequate manner. Diabetes can wreak havoc on a variety of tissues, most notably the eyes, kidneys, heart, blood vessels, and nerves, resulting in chronic damage and dysfunction. Diabetes is primarily classified into two segments (types): T1D – Type 1 Diabetes and T2D - Type 2 Diabetes. Type 1 diabetes typically develops in young aged people (below 30 years of age), and the general symptoms include thirst and urination again, as well as elevated levels of sugar in the blood. Only must be treated with insulin as impossible with other oral drugs. Type 2 diabetes is more prevalent in the younger than younger aged and senior population, and is frequently related with obesity, hypertension, dyslipidemia, arteriosclerosis, and other disorders [4]. Numerous data mining classification methods have been developed with the goal of classifying, forecasting, and diagnosing diabetes. However, no meaningful comparison evaluation of the performance of such algorithms has been conducted. There has been no research conducted to determine which of the existing classifier model scans provides the best prediction for diabetes. The decision tree, Naive Bayes, Random Forest, KNN (K-Nearest neighbours) and Support vector machines (SVM) classification methods were utilized in this work to develop classifier models [5].

2 Related Work

According to Aljumah [6], diabetes is a chronic condition that arises when the body insulin is ineffectively used or when the pancreas produces insufficient insulin. A prominent hormone, Insulin regulates the levels of blood sugar. Unregulated diabetes results in a rise of blood sugar, which leads to serious vandalism to various body parts and systems, like the blood vessels and nerves, over time. According to Health informatics, it is the study of how to collect, retrieve, communicate, store, and utilize health-related data, knowledge, and information to the best of one's ability. Barakat et al. [7] defined how healthcare providers should handle patient information and how citizens should participate in their own health care. It is now widely recognized as a necessary and widespread component of long-term health-care delivery. Machine Learning (ML) is the fastest-growing area in computer science today. When using machine learning in diabetes related data for prediction, it's important to remember that this data isn't being

collected to address specific research questions; instead, learning algorithms are being utilized to analyze biomedical data automatically. Song et al. [8] analyzed multiple categorization algorithms utilizing characteristics such as thickness of skin, pedigree of diabetes, glucose level, Body Mass Index, patient age, insulin and blood pressure. Pradeep and Dr. Naveen compared the machine learning algorithms' performances in [9] and measured the accuracy of each algorithm. There were accuracy variations in terms of techniques utilized, pre-processing and after processing of data. It was noticed that 'Pre-processing of data' had better accuracy and overall performance for prediction of diabetes. In this study, before preprocessing for prediction of diabetes, the Decision tree algorithm provided better accuracy as compared to other techniques like Random forest and Support vector machine. According to Loannis et al. [10], Machine learning techniques, such as the diabetic disorders dataset, have become a significant tool for predicting diabetes using diverse medical data sets (DD). In this work, SVM, Logistic Regression, and Nave Bayes were used. They used 10-fold cross validation for the diabetes dataset (DD). The SVM (Support Vector Machine) strategy outperformed the others in terms of precision and processing, according to the study. For diabetes prediction, Nilashi et al. [11] suggested a CART (classification and Regression Tree) model. Expectation Maximization (EM) and PCA (Principal Component Analysis) were applied to pre-process the data and remove noise before applying the rule. The goal of this study is to design a diabetes decision assistance system. The effect of CART with removal of noise provided efficiency and enhanced prediction, allowing human life to be saved from premature demise. A categorization model was suggested by Kamadi et al. in [12]. One of the most typical problems in categorization, they claim, is reduction of data. PCA (principal component Analysis) was employed in this work for pre-processing of data, as well as for reduction of data to enhance accuracy. The study employed a modified DT (Decision tree) and a fuzzy rule to make predictions. They discovered that reducing the dataset improves the results. Sajida et al. [13] employed the Canadian primary care sentinel surveillance Network(CPCSSN) dataset and three machine learning models to detect diabetes at a primary stage in order to save human lives. To predict diabetes, decision tree (J48), Adaboost, and Bagging were used in this study. Rathore et al. [14] Diabetic disorder can be detected and predicted. The performance measurements were examined using R Studio and the Pima Indians diabetes dataset. SVM and Decision Tree are two machine learning techniques employed. The SVM has an accuracy of 82%.

In [15], S M Hasan Mahmud et al. forecast diabetes. To discover the performance measurements of the classification algorithms, 10-fold cross validation procedures were used. The study found that Naive Bayes outperformed the other classifiers, with an F1 score of 0.74. On the PIMA dataset, Ahuja et al. [16] conducted a comparison examination of various machine learning techniques, including NB, DT, and MLP, for diabetic categorization and found MLP to be superior to other classifiers. Fine-tuning and efficient feature engineering, according to the authors, can improve MLP's performance. Garca-Ordás, M.T. et al. [17] employ min-max normalization and a variant auto encoder sparse auto encoder to solve data standardization, feature augmentation and imbalance. MLP was then used for classification, with an accuracy of 92.31%. Without preprocessing, Bukhari, M.M. et al. [18] state that their ABP-SCGNN (Artificial Back Propagation Scaled Conjugate Gradient Neural Network) obtained 93% accuracy. [19]

is another example of good performance utilizing NN-based models. They looked at median value imputation (MVI), KNN and an iterative imputer for imputation of the missing value. Then, to attain an F1-score of 98%, MLP was employed for classification. Khanam and Foo [20] employed MVI and Pearson Correlation for selection of features and missing value imputation. To further standardize the data and eliminate outliers, interquartile ranges were used. The classification model based on DNN achieved an accuracy of 88.6% using several hidden layers. Overall, missing value imputation and feature selection regarding data pretreatment techniques were seen to be highly appropriate for prediction of diabetes classification performance. The majority of data preparation approaches, on the other hand, have been found to perform well when data is normally distributed. Nonlinear approaches will be better adapted to the problem if the data does not conform to normalcy assumptions, and they are likely to add significantly to a classifier’s performance. As a result, this study will look at nonlinear preprocessing approaches and classifiers for data preprocessing.

3 Methodology

This section describes the classification model’s approach as well as its efficacy in DM classification. Figure 1 summarises the process.

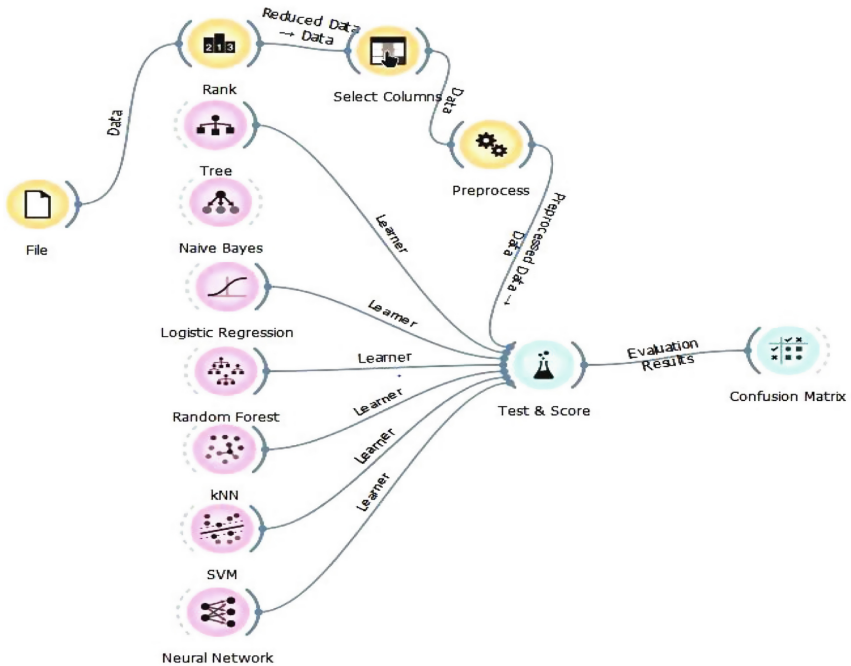


Fig. 1. Methodology of proposed work

For all of the algorithms (Naive Bayes, KNN, ANN, Logistic Regression, Decision Tree, Random Forest, SVM), the ‘Confusion Matrix (CM)’ encapsulates the various steps from raw data to grading, data reduction, pre-processing, scoring, and testing. These steps are described in greater detail in the following subsections as:

- A- It describes the data mining toolkit.
- B- It describes the database and its attributes.
- C- It provides insights into the pre-processing steps.
- D- It discusses the process of classification using the algorithms of seven classifications.

3.1 Data Mining Toolkit

To imitate excellent classification techniques, the Orange Data Mining suite of tools [21] is utilized. Orange was developed as an Open Source Machine Learning (OSML) framework having in-built visualization of data and analytic capabilities at the University of Ljubljana’s Bioinformatics Lab. Orange provides a data preprocessing, classification, regression, clustering, visualization and assessment environment with association rules.

3.2 Collection of Database

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD) obtained the Pima Indians Diabetes dataset (PIDD) of patients. We would like to express our gratitude to Vincent Sigillito for supplying the data and short detailing is provided in Table 1 which shows the class distribution in PIDD.

Table 1. Distribution of classes in the Pima Indians diabetes dataset

Class value	Number of instances	Relabeled value
0	500	Tested_negative
1	268	Tested_positive

The NIDDKD owns the PIDD downloaded from Kaggle [22]. Diabetes mellitus may be identified with the use of this dataset. It has a total of 2000 records, each with eight characteristics and the class label (outcome). The data set’s description, including its properties, statistical analysis, and values, are included in Table 2.

These eight characteristics are symptoms that people may or may not have that indicate their likelihood of having diabetes mellitus.

3.3 Data Preprocessing

Pre-processing is essential to improving model prediction performance. The Orange toolbox supports a variety of pre-processing techniques [23]. Three different types of pre-processing approaches are used in this article to increase the dataset’s quality and eventually the classification models performance.

Table 2. Data set description, properties, statistical analysis and values of data

S No	Attribute name	Attribute description	Data type of attribute	Range of attribute
1	Preg	Pregnancy frequency	N	0 to 17
2	Plas	Concentration of Plasma Glucose	N	0 to 199
3	Pres	BP (Blood Pressure) (mm, hg)	N	0 to 122
4	Skin	Thickness of skin fold	N	0 to 99
5	Insulin	2 h serum insulin (mm U/ml)	N	0 to 846
6	Mass	BMI – Body Mass Index	N	0 to 67.1
7	Pedi	Function of Diabetes pedigree	N	0.078 to 2.42
8	Age	Age of Person (in yrs.)	N	21 to 81
9	Outcome	Class variable	Tested positive, tested negative

N* - Numeric

- Removal of values which are missing

Due to the fact that the utilized dataset had some missing values, Orange toolkit presents three methods for imputing values which are missing: eliminate such records, change them with values which are random, or lastly, change such values with the mean of other accessible values [24]. As a result, this strategy is selected to be utilized to eliminate missing values from the applied dataset.

- Selection of Relevant feature

It is critical to choose the most relevant elements. This stage assigns a score to each characteristic based on its association with the designated diabetes class. From the dataset, eight characteristics were retrieved. ANOVA is a statistical technique. [25] Once ANOVA was calculated, it was obtained that thickness of skin and BP are the least important characteristics and would play a little role in the process of classification; hence, they were deleted from the features vector, resulting in six rather than eight features. Figure 2. The table below summarizes the results of the ANOVA test on the characteristics.

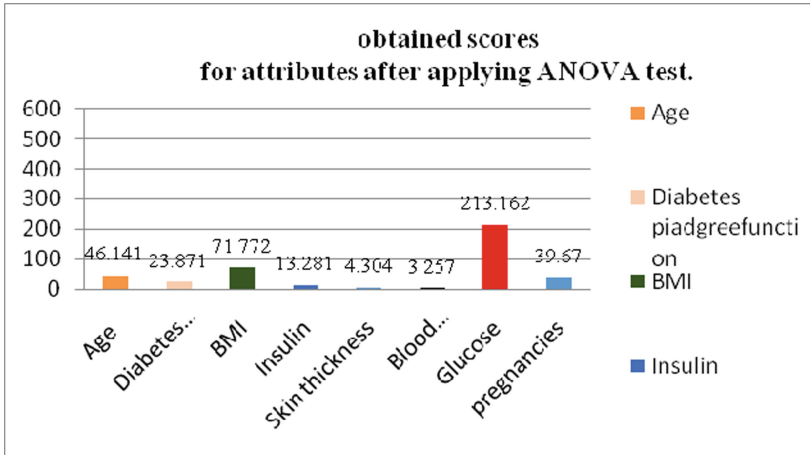


Fig. 2. Result of ANOVA test on the characteristics

- Normalize the Data

Normalization of data can simplify operations and increase computation performance. As a result, the data were normalized to a general scale in a range of zero and one [26]. Scaling by standard deviation (SD) is one of the methods provided in the Orange toolbox.

3.4 Data Classification

During this step, the diabetes dataset was classified using six different algorithms. The investigated classifiers were Naive Bayes, KNN, ANN, SVM, Random Forest, Decision Tree, Logistic Regression and Adaboost. The data set of features in the data base is separated into two parts as training was 70% and testing was 30% to guarantee that the classification process is exactly fit.

- Naive Bayes

Naive Bayes is a statistical learning technique that uses a condensed version of the Bayes rule to determine the posterior distribution of a category given the input attribute values of an example case. Prior probabilities for groups and attribute values that are conditional on categories are calculated using training data frequency counts. Naive Bayes is a straightforward and fast technique for learning that frequently beats more advanced methods. Bayesian classification is both a supervised learning technique and a statistical classification technique. It is capable of resolving diagnostic and predictive issues [27].

- KNN

The KNN algorithm [28] is a simple classification approach. The detection of the nearest K neighbours during the training phase. The distance between objects and the value of K, the number of closest neighbours, are calculated using a similarity measure.

- ANN

ANN is a supervised learning method [29] that uses a network of layers to represent input data, one or more non-linear layers called hidden layers, and finally an output layer that represents the classification category.

- Random Forest

This classifier creates a collection of decision trees [30], which is a random subset of the training data. The test object's final class is chosen to be one that aggregates votes from the various decision trees.

- SVM

SVM models are a type of supervised learning method that may be used for both classification and regression issues, but is most frequently used for classification problems. This classifier is a widely used statistical model that is built on a logistic function applied to a binary dependent variable in the model [31].

- Decision Tree

A decision tree is a tree structure that resembles a flowchart. It is a method for classification and prediction that uses nodes and inter-nodes to describe the data. The root and internal nodes are test cases that are used to distinguish instances with varying characteristics. Internal nodes are generated as a result of attribute testing. The class variable is denoted by the leaf nodes [32].

- Linear Regression

Logistic regression is a technique for binary classification. The input variables are expected to be numeric and to have a Gaussian distribution. It is not required for the last statement to be true in logistic regression. In other words, the method is capable of producing acceptable results even when the data is not Gaussian. Each input value is assigned a coefficient, which is then linearly merged into a regression function and converted using a logistic function [33].

4 Evaluation & Result

In this part, the results of implemented performance measurements are shown using the Orange toolkit's pleasant graphical interface.

4.1 Setup of Experiments with Results

This subpart explains the procedure of sampling used, the parameters of the classification model, and the CM for every algorithm.

- Method of Sampling

The developed models' performance is evaluated using a K-fold cross-validation sampling approach [27]. The whole datasets are cross-validated tenfold in this article (2000 records). The data were divided into tenfold samples. The classification model is trained on seven folds, with the remaining fold serving as a testing set. As a result, for training the model 70% and for testing the model 30% of data records were utilized.

- Decision Tree

The CM of the Tree classifier is demonstrated in Fig. 3. Out of 500 data points, which are labeled as '0', the correct classification is for 402 records. Out of 268 data points, which are labeled as '1', the correct classification is for 142 records.

The Confusion matrix illustrates four critical metrics for evaluating the Decision Tree Classifier model: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Where $TP = 142$, $TN = 402$, $FP = 126$ and $FN = 98$.

- SVM

To learn the model, the attribute space is transformed into a new feature space using a Radial Basis Function (RBF) kernel. The maximum number of iterations authorized was 100. Figure 4 depicts the SVM classifier's confusion matrix.

Whereas out of 500 data points, which are labeled as '0', the correct classification is for 401 records and out of 268 data points, which are labeled as '1', the correct classification is for 152 records. Again, the values of four critical metrics are $TP = 152$, $TN = 401$, $FP = 116$, and $FN = 99$.

- KNN

Figure 5 illustrates the KNN classifier's confusion matrix. The nearest neighbours' numbers was set to five in the KNN model, and the usage of Euclidean distance was done to calculate the distance between two points, with points weighted according to their distance from the query point.

We can see in Fig. 5, The CM summarizes four critical metrics for evaluating the KNN Where $TP = 156$, $TN = 413$, $FP = 112$ and $FN = 87$.

- Random Forest

A forest was incorporated here with 10 decision trees. In Fig. 6, the model's confusion matrix is depicted. The CM illustrates four critical metrics for evaluating the Decision Tree Classifier model, where $TP = 161$, $TN = 425$, $FP = 107$, and $FN = 75$.

- Naive Bayes

Whereas out of 500 data points, which are labeled as ‘0’, the correct classification is for 403 records and out of 268 data points, which are labeled as ‘1’, the correct and successful classification is for 182 records.

The CM illustrates four critical metrics for evaluating the Naive Bayes Classifier model, where $TP = 182$ $TN = 403$, $FP = 86$ and $FN = 97$.

- Artificial Neural Network

In this model, back-propagation was applied with a multi-layer perceptron (MLP) approach. Each buried layer had 200 neurons with a Rectified Linear Unit (ReLU) activation function. The Adam technique was then employed to efficiently optimise stochastic weights. In Fig. 8, the con-fusion matrix for the neural network model is shown. Whereas out of 500 data points, which are labeled as ‘0’, the correct classification is for 431 records and out of 268 data points, which are labeled as ‘1’, the correct classification is for 157 records. The Confusion matrix summarizes four critical metrics for evaluating an ANN Classifier model as $TP = 157$, $TN = 431$, $FP = 111$, and $FN = 69$.

- Logistic Regression

This model’s regularization is set to ridges (L2), and the cost strength is set to its default value of one ($C = 1$). The model’s CM is depicted in Fig. 9.

From 500 data points labeled 0, 442 records were successfully identified, while from 268 data points labeled 1, 151 records were correctly classified.

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are four significant metrics used to assess Logistic Regression Classifier model (FN). $TP = 151$, $TN = 442$; $FP = 117$; $FN = 58$ (Fig. 7).

		Predicted		Σ
		0	1	
Actual	0	402	98	500
	1	126	142	268
Σ		528	240	768

Fig. 3. CM of tree classifier

- Comparison of Performance

The classification methods performance on the dataset of diabetes is examined and compared. The following sections contain details on performance measurements and comparisons.

		Predicted		Σ
		0	1	
Actual	0	401	99	500
	1	116	152	268
Σ		517	251	768

Fig. 4. CM of SVM

		Predicted		Σ
		0	1	
Actual	0	413	87	500
	1	112	156	268
Σ		525	243	768

Fig. 5. CM of KNN

		Predicted		Σ
		0	1	
Actual	0	425	75	500
	1	107	161	268
Σ		532	236	768

Fig. 6. CM of random forest

		Predicted		Σ
		0	1	
Actual	0	403	97	500
	1	86	182	268
Σ		489	279	768

Fig. 7. CM of Naïve Bayes

		Predicted		Σ
		0	1	
Actual	0	431	69	500
	1	111	157	268
Σ		542	226	768

Fig. 8. CM of artificial neural network

		Predicted		Σ
		0	1	
Actual	0	442	58	500
	1	117	151	268
Σ		559	209	768

Fig. 9. CM of logistic regression

- Evaluation Measures of Performance

As mentioned before, the CM illustrates four critical metrics for evaluating classification models: true negative (TN), true positive (TP), false negative (FN) and false positive (FP). These metrics are applied to calculate the following measures of performance:

a) Recall b) Precision c) Accuracy D) F1-measure. These performance metrics are derived by the use of (TP, TN, FP, and FN). The following metrics are used in this study to examine and evaluate categorization models:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - measure = \frac{2 \times (precision \times recall)}{precision + recall} \quad (4)$$

- Classification Model Comparison

The performance of the implemented classifiers is assessed in this subsection using the aforementioned metrics. Table 3 summarizes the performance metrics for the classifiers used.

Table 3. Measures of performance of applied classifiers

Method	AUC	CA	F1	Precision	Recall
Tree	67.9%	70.8%	70.04%	70.2%	70.8%
SVM	75.9%	72%	71.8%	71.6%	72.0%
KNN	78.8%	74.1%	73.8%	73.6%	74.1%
Naïve Bayes	82.9%	76.2%	76.3%	76.4%	76.2%
Random Forest	81.1%	76.3%	75.9%	75.8%	76.3%
Neural Network	82.6%	76.6%	76%	76.0%	76.6%
Logistic Regression	82.9%	77.2%	76.4%	76.7%	77.2%

It also compares the accurate performance of all applicable models. It is self-evident that Logistic Regression surpasses other classifiers with 77.2% accuracy. Logistic Regression is followed by a Artificial Neural Network model in second place with an accuracy of 76.6% and Random Forest in third place with a accuracy of 76.3%. Random forest is followed by the KNN model in fourth place with accuracy of 74.1%. And SVM got fifth position with accuracy of 72%. Decision tree with the accuracy of 70.8% is the worst case. Logistic regression outperforms in all performance measures like AUC, F1-score, Precision and Recall, which can be shown in Table 3.

5 Conclusion

Automatic diabetes detection is a significant real-world medical issue. Early detection and management of diabetes are critical. This article demonstrates the use of several classifiers, including Decision Trees, SVM, KNN, Naive Bayes, Random Forest, Neural Network, and Logistic Regression, to simulate actual diabetes diagnosis for local and systemic therapy, as well as presenting relevant work in the field and the outcome indicates that Logistic Regression outperforms all other classifiers. The suggested model's usefulness is demonstrated by experimental data. The performance of the strategies was evaluated in relation to the problem of diabetes diagnosis. Experiments validate the given model. In the future, it is planned to compile data from several locations across India and develop a more precise and broad predictive model for diabetes diagnosis. Future research will similarly focus on accumulating data from a later time period and identifying additional possible prognostic factors to integrate. The technique might be expanded and refined to automate the analysis of diabetes.

References

1. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
2. Amin, D.M., Garg, A.: Performance analysis of data mining algorithms. J. Comput. Theor. Nanosci. **16**(9), 3849–3853 (2019). <https://doi.org/10.1166/jctn.2019.8260>

3. Saichanma, S., Chulsomlee, S., Thangrua, N., Pongsuchart, P., Sanmun, D.: The observation report of red blood cell morphology in Thailand teenager by using data mining technique. *Adv. Hematol.* **2014**, 1–5 (2014). <https://doi.org/10.1155/2014/493706>
4. Canlas, R.D. (2009). *Data Mining in Healthcare: Current applications & Issues*, Unpublished Master Thesis, 1–10
5. Iyer, A., Jeyalatha, S., Sumbaly, R.: Diagnosis of diabetes using classification mining techniques. *Int. J. Data Min. Knowl. Manag. Process* **5**(1), 01–14 (2015). <https://doi.org/10.5121/ijdkp.2015.5101>
6. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K.: Application of data mining: diabetes health care in young and old patients. *Journal of King Saud University - Computer and Information Sciences* **25**(2), 127–136 (2013). <https://doi.org/10.1016/j.jksuci.2012.10.003>
7. Barakat, N., Bradley, A.P., Barakat, M.N.H.: Intelligent support vector machines for diagnosis of diabetes mellitus. *IEEE Trans. Inf. Technol. Biomed.* **14**(4), 1114–1120 (2010). <https://doi.org/10.1109/titb.2009.2039485>
8. Komi, M., Li, J., Zhai, Y., Zhang, X.: Application of data mining methods in diabetes prediction. In: *2nd International Conference on Image, Vision and Computing (ICIVC)*, pp. 1006–1010 (2017)
9. Pradeep, K.R., Naveen, N.C.: Predictive analysis of diabetes using J48 algorithm of classification techniques. In: *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 347–352 (2016)
10. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **15**, 104–116 (2017). <https://doi.org/10.1016/j.csbj.2016.12.005>
11. Nilashi, M., Ibrahim, O.B., Ahmadi, H., Shahmoradi, L.: An analytical method for diseases prediction using machine learning techniques. *Comput. Chem. Eng.* **106**, 212–223 (2017). <https://doi.org/10.1016/j.compchemeng.2017.06.011>
12. Kamadi, V.V., Allam, A.R., Thummala, S.M.: A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. *Appl. Soft Comput.* **49**, 137–145 (2016). <https://doi.org/10.1016/j.asoc.2016.05.010>
13. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance analysis of data mining classification techniques to predict diabetes. *Procedia Comput. Sci.* **82**, 115–121 (2016). <https://doi.org/10.1016/j.procs.2016.04.016>
14. Rathore, A., Chauhan, S., Gujral, S.: Detecting and predicting diabetes using supervised learning: an approach towards better healthcare for women. *Int. J. Adv. Res. Comput. Sci.* **8**(5), 1192–1195 (2017)
15. Mahmud, S.M.H., et al.: *Machine Learning Based Unified Framework for Diabetes Prediction*. Association for Computing Machinery. China (2018). <https://doi.org/10.1145/3297730.3297737>
16. Ahuja, R., Sharma, S.C., Ali, M.: A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing* **3**(3), 44–52 (2019). <https://doi.org/10.33166/aetic.2019.03.005>
17. García-Ordás, M.T., Benavides, C., Benítez-Andrades, J.A., Alaiz-Moretón, H., García-Rodríguez, I.: Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Comput. Methods Programs Biomed.* **202**, 105968 (2021). <https://doi.org/10.1016/j.cmpb.2021.105968>
18. Bukhari, M.M., Alkhamees, B.F., Hussain, S., Gumaei, A., Assiri, A., Ullah, S.S.: An improved artificial neural network model for effective diabetes prediction. *Complexity* **2021**, 1–10 (2021). <https://doi.org/10.1155/2021/5525271>

19. Roy, K., et al.: An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. *Complexity* **2021**, 1–21 (2021). <https://doi.org/10.1155/2021/9953314>
20. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. *ICT Express* **7**(4), 432–439 (2021). <https://doi.org/10.1016/j.ict.2021.02.004>
21. Orange – Data Mining Fruitful & Fun. <https://orange.biolab.si/>
22. Diabetes –dataset. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/>. Accessed 01 April 2022
23. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J.: Progress in computing, analytics and networking. In: *Proceedings of ICCAN 2017* (2018)
24. Garcia, S., Luengo, J., Herra, F.: *Data Preprocessing in Data Mining*. Springer (2015). <https://doi.org/10.1007/978-3-319-10247-4>
25. Alsalamah, M., Amin, S., Palade, V.: Clinical practice for diagnostic causes for obstructive sleep apnea using artificial intelligent neural networks. In: Miraz, M.H., Excell, P., Ware, A., Soomro, S., Ali, M. (eds.) *iCETiC 2018*. LNICSSITE, vol. 200, pp. 259–272. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95450-9_22
26. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann (2016)
27. Rennie, J.D., et al.: Tackling the poor assumptions of naive bayes text classifiers. In: *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, pp. 616–623 (2003)
28. Chen, G.H., Shah, D.: Explaining the success of nearest neighbor methods in prediction. *Found. Trends® Mach. Learn.* **10**(5–6), 337–588 (2018). <https://doi.org/10.1561/22000000064>
29. van Gerwen, M., Bohte, S.: Artificial neural networks as models of neural information processing. *Front. Comput. Neurosci.* **11**, 114 (2017). <https://doi.org/10.3389/fncom.2017.00114>
30. Davies, A., Ghahramani, Z.: The random forest kernel and other kernels for big data from random partitions (2014). arXiv.1402.4293
31. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004). <https://doi.org/10.1023/b:stco.0000035301.49549.88>
32. Rokach, L.: *Data Mining with Decision Trees: Theory and Application*, vol. 81. World Scientific (2014)
33. Weisberg, S.: *Applied Linear Regression*, 4th ed. Wiley (2013)