International Journal of Computer Science and Informatics

Volume 4 | Issue 3 Article 1

February 2022

Diabetes Prediction: A Study of Various Classification based Data Mining Techniques

Sipra Sahoo
Siksha O Anusandhan Deemed to be University, sipraster@gmail.com

Tushar Mitra
Siksha O Anusandhan Deemed to be University, tushar.99.mitra@gmail.com

Arup Kumar Mohanty
Siksha O Anusandhan Deemed to ne University, arupmohanty@soa.ac.in

Bharat Jyoti Ranjan Sahoo
Siksha O Anusandhan Deemed to be University, bharatjyotisahu@soa.ac.in

Smita Rath
Siksha O Anusandhan Deemed to be University, smitarath@soa.ac.in

Recommended Citation

Sahoo, Sipra; Mitra, Tushar; Mohanty, Arup Kumar; Sahoo, Bharat Jyoti Ranjan; and Rath, Smita (2022) "Diabetes Prediction: A Study of Various Classification based Data Mining Techniques," *International Journal of Computer Science and Informatics*: Vol. 4 : Iss. 3 , Article 1.
DOI: 10.47893/IJCSI.2022.1191
Available at: https://www.interscience.in/ijcsi/vol4/iss3/1

Diabetes Prediction: A Study of Various Classification based Data Mining Techniques

Sipra Sahoo[1], Tushar Mitra[2], Arup Kumar Mohanty[3], Bharat Jyoti Ranjan Sahu[4], Smita Rath[5]

[1,2,3,4,5]Department of Computer Science and Education,
Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar,Odisha, India
sipraster@gmail.com, tushar.99.mitra@gmail.com, arupmohanty@soa.ac.in,
bharatjyotisahu@soa.ac.in, smitarath@soa.ac.in }

Abstract. Data Mining is an integral part of KDD (Knowledge Discovery in Databases) process. It deals with discovering unknown patterns and knowledge hidden in data. Classification is a pivotal data mining technique with a very wide range of applications. Now a day's diabetic has become a major disease which has almost crippled people across the globe. It is a medical condition that causes the metabolism to become dysfunctional and increases the blood sugar level in the body and it becomes a major concern for medical practitioner and people at large. An early diagnosis is the starting point for living well with diabetes. Classification Analysis on diabetic dataset is a part of this diagnosis process which can help to detect a diabetic patient from non-diabetic. In this paper classification algorithms are applied on the Pima Indian Diabetic Database which is collected from UCI Machine Learning Laboratory. Various classification algorithms which are Naïve Bayes Classifier, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier and XGBoost Classifier are analyzed and compared based on the accuracy delivered by the models.

Keywords: Classification; Naïve Bayes; Logistic Regression; Decision Tree; Random Forest; Support Vector Machines.

1 Introduction

Diabetes is a disease that occurs when blood glucose which is the main source of energy is too high. Insulin, a hormone made by the pancreas, helps glucose from food get into cells to be used for energy. It is a medical condition that causes the metabolism to become dysfunctional and increases the blood sugar level in the body. It is prevalent in many nations; however, it is rapidly increasing and is a subject of major concern for healthcare specialists and people at large. Each year diabetes is also one of the major reasons for a significant number of heart attacks, permanent loss of vision, Kidney and Brain failure and even death. Diabetes affects approximately 422 million people worldwide, with the majority living in low- and middle-income countries [1, 2].

Diabetes is a disease that occurs when blood glucose which is the main source of energy is too high. Insulin, a hormone made by the pancreas, helps glucose from food get into cells to be used for energy. It is a medical condition that causes the

metabolism to become dysfunctional and increases the blood sugar level in the body. It is prevalent in many nations; however, it is rapidly increasing and is a subject of major concern for healthcare specialists and people at large. Each year diabetes is also one of the major reasons for a significant number of heart attacks, permanent loss of vision, Kidney and Brain failure and even death. Diabetes affects approximately 422 million people worldwide, with the majority living in low- and middle-income countries [1, 2].

Diabetes is classified into three types according to the Diabetes Federation

- Type 1 Diabetes
- Type 2 Diabetes
- Gestational Diabetes.

Type 1 Diabetes or Insulin-dependent diabetes mellitus (IDDM) or juvenile diabetes manifests as an auto-immune disease that occurs at a very young age, usually before the age of 20. Type 2 Diabetes is a condition in which various organs of the body become insulin resistant, increasing the demand for insulin and, as a result, the pancreas fails to produce the required amount of insulin. This is known as non-insulin-dependent diabetes mellitus "(NIDDM) or diabetes that develops as an adult ". While the cause of Type 1 diabetes is unknown, the cause of Type 2 diabetes is obesity, which can be controlled through exercise and proper diet. If the blood sugar level does not decrease with exercise and diet control, medicine will be prescribed to control the blood sugar level. Gestational diabetes, on the other hand, is more common in pregnant women who do not have a family history of the disease [3].

An early diagnosis is the starting point for living well with diabetes; the longer a person goes undiagnosed and untreated, the worse their health outcomes are likely to be. Classification Analysis on diabetic dataset is a part of diagnosis that can help detect whether a patient is diabetic or not. This would have been otherwise very tough given the multiple symptoms that patients may possess. Data mining techniques that try to discover useful patterns from datasets that are not visible right away to human eyes. Classification is a type of data mining technique that uses classes of output and assigns incoming data to those predefined classes based on the patterns discovered by the model. The primary goal of any Classification algorithm is to correctly assign those classes with the least error that is possible. This article deals with some of the various famous classification algorithms in use today and analyses each based on certain accuracy metrics. Diabetes Prediction is a tough task as classes of attributes are not linearly separable as shown in Fig 1 below.

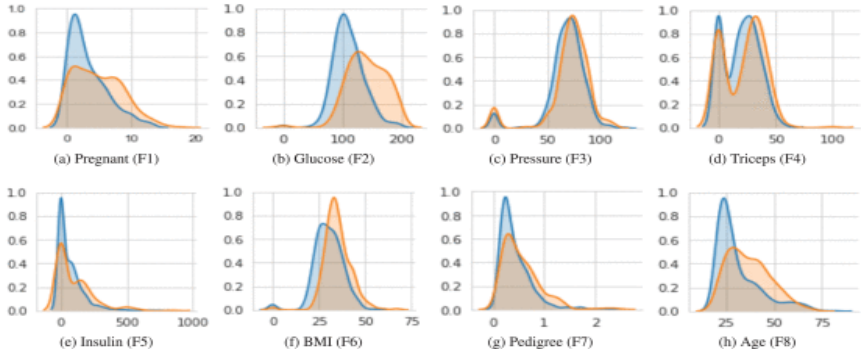

Fig 1: The Population Distribution of all attributes of the Pima Indian Diabetes Dataset [4] where blue and orange color distribution respectively denote non-diabetic and diabetic class.

2 Literature Survey

Siddique et al. discuss the role of Adaboost and Bagging ensemble machine learning methods in classifying Diabetes Mellitus and patients as diabetic or non-diabetic based on diabetes risk factors. The results of the experiment show that the Adaboost machine learning ensemble technique outperforms bagging as well as a JJ48 Decision Tree [5].

Orabi et al. created a diabetes prediction system, the main goal of which is to predict the type of diabetes a candidate will have at a given age. The proposed system is built on the concept of machine learning and employs a decision tree. The obtained results were satisfactory because the designed system performs well in predicting diabetes incidents at a specific age, with greater accuracy using Decision Tree [6].

Pradhan et al. used Genetic programming (GP) for the training and testing of the database for diabetes prediction using the Diabetes data set from the UCI repository. When compared to other implemented techniques, the results obtained using Genetic Programming have the highest accuracy. By reducing the time required for classifier generation, accuracy can be significantly improved [7]. In Zou et al.'s [8] study, they applied Random Forest, Decision Tree, ANN for classification algorithm on PIDD after the feature reduction using Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. They found that Pima Indians' best accuracy is 77.21% obtained from the random forest with the mRMR feature reduction method. The model with Logistic Regression(LR) and Support Vector Machine (SVM) works well on diabetes prediction [9]. The NN model with a different hidden layer with various epochs are implemented and 88.6% accuracy is observed. Kalpana and Kumar [10] proposed fuzzy expert system frameworks for diabetes which has built large scale knowledge based system. The models proposed in [11] is based on the prediction precision of certain powerful machine learning (ML) algorithms based on different measures such as precision, recall, and F1-measure. The Pima Indian Diabetes (PIDD) dataset has been used, that can predict diabetic onset based on diagnostics manner.

3 Proposed Model

In Data Mining, the main aim of any classification algorithm is to properly assign classes to the data. This prediction of classes must be done accurately and with the least possible error. We have tried to analyze various Classification algorithms that are widely employed in many Classification type prediction problems. The primary goal of this study is to assess the performance of classification methods for diabetes datasets based on numerical input and imbalance dataset constraints.

The workflow in the article follows two stages:

- Stage 1: This is the data preprocessing step. This step includes primarily outlier rejection (P) and value imputation (Q). An outlier is basically an

observation which is markedly deviated from the other observations. It is necessary to reject such values because the classifiers that would be used are sensitive to data range distribution.
The mathematical formulation used for the detection in the literature can be written as in Equation 1:

$$P(x) = \begin{Bmatrix} x, \ if \ Q1 - 1.5 * IQR < x < Q3 + 1.5 * IQR \\ reject, \ otherwise \end{Bmatrix} \quad (1)$$

Where x is the feature instance, Q1, Q3 and IQR are the First Quartile, Third Quartile, and the Inter Quartile Range respectively.
The attributes after outlier rejection and any null values were imputed to prevent any wrong prediction. The missing values were imputed using mean values in the proposed technique and mathematically it can shown by Equation 2.

$$Q(x) = \begin{Bmatrix} mean(x), \ if \ x \ is \ null \ or \ missing \\ x, \ otherwise \end{Bmatrix} \quad (2)$$

The dataset used is imbalanced. Thus, to deal with it data was sampled randomly, using only 10% of the data at a time [12].

- Stage 2: This is the model training and testing phase. The model is trained upon the data and then predictions are generated. These predictions are further tested against actual values.

Any Classification Algorithm follows certain predefined steps, which have been shown in Fig. 2.

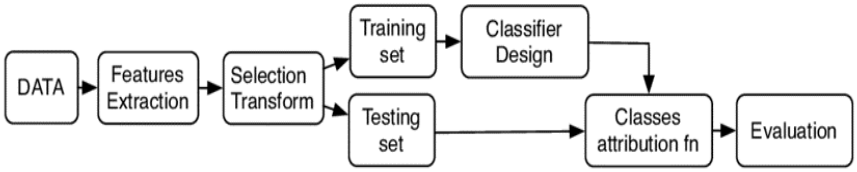

Fig. 2. Stages of a typical Supervised Classification Algorithm

The dataset used in the classification experiment is the Pima Indians Diabetes Database from the National Institute, which has been obtained from the Kaggle Database. There are 768 total instances recorded in the data. This same dataset, however, is imbalanced in target class, with 500 instances of class label for "Negative" or"0" and 268 instances of target class for "Positive" or "1". Thus, the SMOTE oversampling method was used to combat the imbalance dataset, which generated 1036 instances, 500 of which were of the target class "Negative" or"0" and 536 of which were of the target class "Positive" or "1". The dataset was then randomly generated to shuffle the order of newly generated synthetic target classes for "Positive" or"1" in the dataset.

Fig. 3 and Fig. 4 respectively show the share of Negative and Positive classes before and after SMOTE oversampling method. This was done so that the dataset could be balanced.

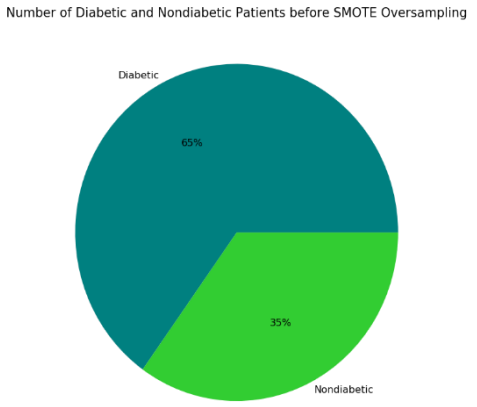

Fig. 3. Total Diabetic and Non-Diabetic Classes before SMOTE Oversampling.

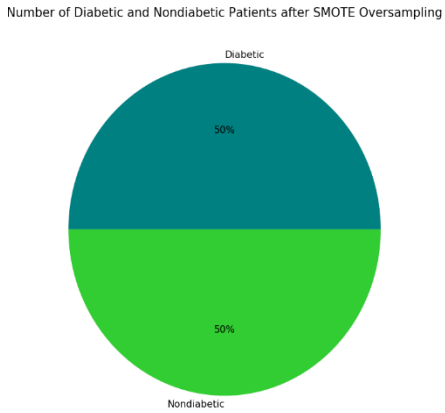

Fig. 4. Total Diabetic and Non –Diabetic Classes after SMOTE Oversampling.

3.1 Models Used

We have used six Supervised Classification Algorithms namely:

- Naïve Bayes Classifier
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Extreme Gradient Boosting (XGBoost Classifier)

- Support Vector Classifier

We shall walk through each one by one:

3.1 Naïve Bayes Classifier: Nave Bayes is a probabilistic machine learning algorithm that is based on the Bayes Theorem and is used for a wide range of classification tasks. The Bayes' Theorem is a straightforward mathematical formula for calculating conditional probabilities. Conditional probability is a measure of the likelihood of one event occurring given that another event has already occurred (via assumption, presumption, assertion, or evidence). By assuming that features are independent of class, the naive Bayes classifier greatly simplifies learning. Although independence is a poor assumption in general, naive Bayes frequently outperforms more sophisticated classifiers in practice. It works well with data that has balancing issues and missing values. The Bayes Theorem is used by Naive Bayes, a machine learning classifier [13]. Using Bayes theorem we can calculate Posterior Probability P(X | C) as shown in Equation 1.

$$P(C|X) = (P(X|C)\, P(C))/P(X) \quad (3)$$

P *C*|*X*) = target class's posterior probability.
P (*X*|*C*) = predictor class's probability.
P (C) = class C's probability being true.
P(X) = predictor's prior probability.

3.2 Logistic Regression: It is much like Linear Regression however the cost function used here is much more complex. A general question arises here that why linear regression cannot be used. The answer is very basic, since the output of linear regression ranges over the entire real plane, it cannot be used for classification type problems. The hypothesis for Logistic regression limits the output variable between 0 and 1. To scale the output within this range a special function is used which is the Sigmoid Function. Thus, the formula for Logistic function is the one as shown in Equation 2.

$$f(x) = \{1\}/\{1 + e^{-x}\} \quad (4)$$

The output of the Sigmoid Function is numerical in nature. Thus, to interpret it as a categorical variable we need a decision boundary. In our model we used that decision boundary of 0.5. It signifies that any data which gave the result greater than equal to 0.5 is labelled as Diabetic and Non-Diabetic otherwise [14].

3.3 Decision Tree Classifier: A decision tree algorithm involves segmenting the predictor space into several simpler regions. Decision trees can be applied to both regression and classification problems.). In a classification type problem, each of these segments is assigned different class labels. For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. A decision tree grows by recursive binary splitting. However, unlike the Regression tree which uses Residual Sum of Squares or RSS as a criterion for binary splitting, here it is not helpful since we have class labels as output variables [15]. A natural alternative to RSS is the classification

error rate which has been mathematically represented in Equation 3. This is simply the fraction of the training observations in that region that do not belong to the most common class:

$$E=1- \max (p_{mk}) \quad (5)$$

Here, p_{mk} represents the proportion of training observations in the m^{th} region that are from the k^{th} class. However, classification error is not sufficiently sensitive for tree-growing, and in practice two other measures are preferable.

Gini Index and Information Gain: The Gini index is defined by a measure of total variance across the K classes, which has been mathematically shown in Equation 4.

$$G=\Sigma \ p_{mk\,(1\,-\,p_{mk})} \quad (6)$$

The Gini index takes on a small value if all the p_{mk}'s are close to zero or one. For this reason, the Gini index is referred to as a measure of node purity - a small value indicates that a node contains predominantly observations from a single class. An alternative to Gini Index is Cross Entropy, and both are almost same.

Information gain uses the concept of entropy which is the degree of randomness or the amount of impurity in the system. Information gain is the decrease in entropy or randomness. The attribute which gives the highest information gain is chosen as the best attribute for split at a particular node[16].

Before proceeding further let us understand what **Ensemble Learning** is. It is a technique in which multiple weak learners are trained simultaneously to produce a single strong learner to enhance the accuracy of prediction. They are primarily of three types:

- Bagging or Bootstrap Aggregating
- Boosting
- Stacking

3.4 Random Forest Classifier: It is a specific type of Ensemble Learning**.** We shall speak strictly about bagging here because Random Forest Classifier is a type of Bagging Algorithm. Bagging considers many homogeneous weak learners and trains each of them independently in parallel and combines their results in a deterministic averaging technique.

The major priority here is generating a model with lower variance. In a Random Forest type classifier, the single weak learners are Decision Trees. Unlike Decision Trees, since Random Forests do not sample over the same features, rather they split on a small subset of features, the final outcomes have very little correlation with them. Also, it restricts over fitting and can also handle missing values, which is, a major problem in Decision Trees [17].

Firstly, we bootstrapped multiple samples from the dataset which are independent of each other. Then we trained the Decision Tree on each of these independent samples. Then each of these results were combined to find the results.

3.5 XGBoost Classifier: It is otherwise known as Extreme gradient Boosting which is a type of Ensemble technique that is based upon Decision Trees and uses Boosting. It is a method that goes through cycles iteratively to add models into an ensemble. It begins by initializing the ensemble by a weak learner or a base model whose predictions are very naïve. With Subsequent iterations of the algorithm the errors are addressed. Firstly, the current ensemble is used to generate predictions for each observation. To make a prediction, all the predictions from different models are considered, which are then used to calculate the loss function. This loss function is used to fit a new model that gets added to the ensemble. The gradient in XGBoost stands for gradient descent which is used in the loss function to determine the parameters. The loss function that we used was **binary: logistic** since the problem was of a binary classification type. XGBoost has several parameters that can substantially alter the accuracy of prediction [18].

- **n_estimators**: It determines how many times to go about the modelling cycle. It is equal to the number of models we include in the ensemble. Typical values range from 100-1000. **The value we used was the default value 100**.
- **early_stopping_rounds**: It automatically provides an ideal value for n_estimators. The early_stopping_rounds cause the model to stop iterating when the validation score stops improving after number of cycles equal to the value, set for early_stopping_rounds is reached. **The value we used was equal to 5**.
- **learning_rate**: This value is multiplied to the output of each model while calculating the overall result for the ensemble. This ensures that each individual weak learner contributes less and thus prevents over fitting of the model. It is usually suggested to keep the value for n_estimators high and the learning rate low. This ensures the XGBoost model predicts with higher accuracy. **The value we used was equal to 0.05** [19]**.**

3.6 Support Vector Classifier: Support vector machines (SVMs, also known as support vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. A Support Vector Machine (SVM) is a discriminative classifier that is formally defined by a separating hyper plane. Support vector Classifiers are fast and dependable when it comes to limited size of data to analyze. For the case of a binary classification a SVC considers a plane of output variables from the train set. The SVC takes the plane of these points and outputs the hyper plane that best separates these variables. It is worthy to note that Support Vector Classifiers also work well for nonlinear data [20].

4 Results and Discussions

All of the six Machine Learning models that have been specified in the paper were implemented using Python Programming and Python and Keras API's. The machine hardware specification are as follows:

Operating System: Windows 10
RAM: 8GB
Processor: Intel Core i7 vPRO
Python Version: Python 3.6
To study the performance of all the Classification Algorithms we have used many accuracy measures. Let us go through each of these Accuracy measures one by one first.

A **Confusion Matrix** is a measurement for accuracy for Classification Type Algorithms. As the name suggests it is a matrix containing values. For a typical Binary Classification type problem, a confusion matrix contains four values that are: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [16].

- **True Positive**: We predicted positive, and it is true.
- **True Negative**: We predicted negative, and it is true.
- **False Positive**: We predicted positive, and it is false.
- **False Negative**: We predicted negative, and it is true.

The Confusion Matrices for each of the models have been shown below in Table 1, Table 2, Table 3, Table 4, Table 5, and Table 6.

Table 1. Confusion Matrix for Naïve Bayes Classifier.

Actual → Predicted ↓	Positive	Negative
Positive	73	18
Negative	42	67

Table 2. Confusion Matrix for Logistic Regression.

Actual → Predicted ↓	Positive	Negative
Positive	73	18
Negative	43	66

Table 3. Confusion Matrix for Decision Tree Classifier.

Actual → Predicted ↓	Positive	Negative
Positive	51	40
Negative	13	96

Table 4. Confusion Matrix for Random Forest Classifier.

Actual → Predicted ↓	Positive	Negative
Positive	73	18

Negative	32	77

Table 5. Confusion Matrix for XGBoost Classifier.

Actual → Predicted ↓	Positive	Negative
Positive	72	19
Negative	41	68

Table 6. Confusion Matrix for Support Vector Classifier.

Actual → Predicted ↓	Positive	Negative
Positive	72	19
Negative	41	68

From here come the concept of Precision, Recall, Accuracy and F1-Score. Let us understand each of these next.

Precision: Out of all the positive classes we have predicted correctly, how many are positive. It has been mathematically shown through Equation 5.

$$\text{Precision} = TP/(TP+FP) \quad (7)$$

Recall: Out of all the positive classes how any did we correctly classify. It has been mathematically shown through Equation 6.

$$\text{Recall} = TP/(TP+FN) \quad (8)$$

Accuracy: Out of all the classes how many did we classify properly. It has been mathematically shown through Equation 7.

$$A = (\text{Correctly classified data})/(\text{Total number of data}) \quad (9)$$

F1- Measure: The F1-Measure is the Harmonic mean of the precision and recall. It gives a better measure of incorrectly classified data. It has been mathematically shown through Equation 8.

$$\text{F1- Score} = 2*(\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \quad (10)$$

Out of Accuracy and F1-Score, F1-Score is more helpful in real life problems because there are imbalanced classes in real life. We have calculated all the values for precision, recall, accuracy and F1- score for all the algorithms and displayed in Table 7(up to 5 decimal points of accuracy).

Table 7. Table containing values for Precision, Recall, Accuracy and F1-Score for all the classification algorithms.

	Precision	Recall	Accuracy	F1-Score
Naïve Bayes	0.62222	0.50909	0.71428	0.56000
Logistic Regression	0.69230	0.49090	0.74026	0.57446
Decision Tree	0.63265	0.56363	0.72727	0.59615
Random Forest Classifier	0.64102	0.45454	0.71428	0.53191
XGBoost Classifier	0.67500	0.49090	0.73376	0.56842
Support Vector Classifier	0.67500	0.49090	0.73376	0.56852

The metrics stated above for all the models have also been shown in a graphical manner as in Fig. 5.

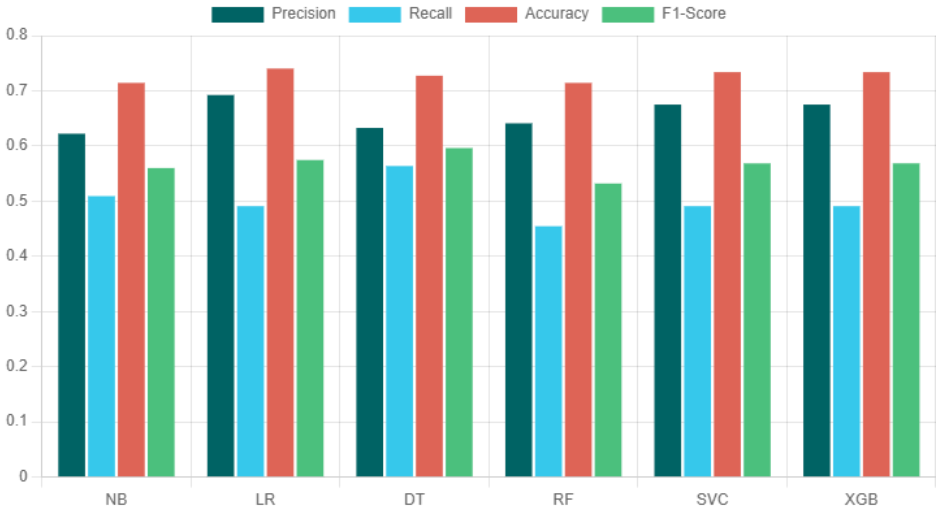

Fig. 5. Bar graph representing the Precision, Recall, Accuracy and F1-Score of all the models.

5 Conclusion and Future Scope

This Literature uses the Pima Indian Dataset to study and analyze various Classification Algorithms. It has been established, how preprocessing can improve the precision of Classification. With Outlier Rejection and Missing value imputation being the core concern, they were dealt followed by SMOTE sampling technique.
This work is based on comparing various models for prediction from the Diabetes Dataset. We used several State-of-the-art Supervised Classification Algorithms which are namely Naïve Bayes, Logistic Regression, Decision Tree, random forest, XGBoost and Support Vector Classifier. From the above generated outputs, it is quite

evident that Decision Tree Classifier outperforms any other model when it comes to F1- Score. But when we consider Precision and Accuracy as measures for the best model Logistic Regression outperforms any other model. Thus, we can say that for the Pima Indian Dataset Logistic Regression and Decision Tree are best suited models.

However, over time the focus has shifted from a highly accurate system from diabetes prediction to a system that is highly accurate, for the greater population.

It has become evident that preprocessing improves Classification outcomes. Furthermore, different attribute subset selection techniques could be employed in the preprocessing step to improve. This may enhance the outcome. Along with these multiple pipelines could be created for best performing algorithms. However, these are beyond the scope of this paper. Apart from using hybrid models, that is a combination of different best performing models, the algorithms could be trained on various datasets to compare and find the most reliable algorithm for diabetes prediction. Additionally, the proposed framework could be used in the branch of medicine to detect chances of diabetes and prevention of diabetes.

References

1. Rashid, T.A., Abdullah, S.M., & Abdullah, R.M.: An intelligent approach for diabetes classification, prediction and description. In Innovations in Bio-Inspired Computing and Applications. Springer, Cham (2016) 323-335

2. Tervaert, T. W. C., Mooyaart, A. L., Amann, K., Cohen, A. H., Cook, H. T., Drachenberg, C. B., & Bruijn, J. A.: Pathologic classification of diabetic nephropathy. Journal of the American Society of Nephrology 21.4 (2010) 556-563

3. Haneda, M., Utsunomiya, K., Koya, D., Babazono, T., Moriya, T., Makino, H., Kimura, K., Suzuki, Y., Wada, T., Ogawa, S. and Inaba, M.: A new classification of diabetic nephropathy 2014: a report from joint committee on diabetic nephropathy. Journal of diabetes investigation 6.2 (2015) 242-246

4. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 261-265, Nov. 1988.

5. 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 (2013) 127-136

6. Orabi, K. M., Kamal, Y. M., & Rabah, T. M.: Early predictive system for diabetes mellitus disease. Industrial Conference on Data Mining. Springer, Cham, (2016) 420-427

7. Pradhan, M.A., Bamnote, G.R., Tribhuvan, V., Jadhav, K., Chabukswar, V. and Dhobale, V., 2012.: A genetic programming approach for detection of diabetes." International Journal of Computational Engineering Research 2.6 (2012) 91-94

8. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.

9. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express.

10. Kalpana, M., & Kumar, A. S. (2011). Fuzzy expert system for diabetes using fuzzy verdict mechanism. International Journal of Advanced Networking and Applications, 3(2), 1128.

11. Khaleel, F. A., & Al-Bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings.

12. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
13. Choubey, D.K., Paul, S., Kumar, S. and Kumar, S.: Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. Communication and computing systems: proceedings of the international conference on communication and computing system (ICCCS 2016). (2017) 451-455
14. Zhu, C., Idemudia, C. U., & Feng, W.: Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked 17 (2019) 100179
15. Mirza, S., Mittal, S., & Zaman, M.: Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree." International Journal of Applied Engineering Research 13.11 (2018) 9277-9282
16. Nurjahan, Mohammad Abu Tareq Rony, Md. Shahriare Satu, Md Whaiduzzaman, "Mining Significant Features of Diabetes through Employing Various Classification Methods", *Information and Communication Technology for Sustainable Development (ICICT4SD) 2021 International Conference on*, pp. 240-244, 2021.
17. Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., Chen, L. and Qiu, L., 2021: Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. BMC medical informatics and decision making 21.1 (2021). 1-14
18. Chen, P., & Pan, C.: Diabetes classification model based on boosting algorithms. BMC bioinformatics 19.1 (2018) 1-9
19. Kumari, V. Anuja, and R. Chitra.: Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications 3.2 (2013) 1797-1801
20. P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.