Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach

Prajyot Palimkar, Rabindra Nath Shaw, and Ankush Ghosh

Abstract Diabetes is one among many chronic diseases. It is the most common disease and lots of peoples are affected by this. There are many things that are liable for diabetes, mainly age, obesity, weakness, sudden weight loss, and many more. Diabetes patients have high risk of diseases like cardiopathy, renal disorder, stroke, nerve damage, eye damage, etc. Detection of the disease isn't very easy and prediction is additionally costlier. In today's situation, hospitals are extremely busy due to COVID-19 pandemic, and it might be revolutionary if one could know if they're at risk of being diabetic without visiting a doctor. But the rise in Artificial Intelligence techniques can be used for disease prognosis. The objective of this study is to develop a model with significant accuracy to diagnose diabetes in patients. Moreover, this paper also presents an effective diabetes prediction model for better classification of diabetes and to enhance the accuracy in diabetes prediction using several machine learning algorithms. Different machine learning algorithms are utilized for early stage diabetes prediction, namely, Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve Bayes. The performances of these models are measured on respective criteria like Accuracy, Precision, Recall, F-Measure, and Error. For this research work, latest available dataset dated 22nd July, 2020, is being utilized. Latest updated dataset will show comparatively better result.

Keywords Diabetes · Machine learning algorithm · Random forest · Decision tree · Predictive analysis technique · Classification

P. Palimkar · A. Ghosh (✉)
School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

R. N. Shaw
Department of Electrical, Electronics & Communication Engineering, Galgotias University, Noida, India
e-mail: r.n.s@ieee.org

M. Bianchini et al. (eds.), *Advanced Computing and Intelligent Technologies*,
Lecture Notes in Networks and Systems 218,
https://doi.org/10.1007/978-981-16-2164-2_19

1 Introduction

Diagnosis is the most vital part of the medical science. They use different strategies for that. Under this strategy, classification of given data is done in different classes based upon some constraints. The existence of diabetes depends upon different types of factors. Diabetes is an illness due to inability of human body, to secrete harmon insulin. Imbalance insulation show symptoms like intensified thirst and hunger, high blood glucose, frequent urination. If diabetes is untreated, many complications could also be occurring, which end up in serious health problems. So, early prediction of diabetes is critical to cut back the consequences of it.

Different predictive analyses [1] include machine learning algorithms and statistical methods. Under this technique, the classification of past data is done to get knowledge to predict future events. Machine learning and regression technique can be effectively used for predictive analysis. Among different artificial intelligence techniques, machine learning is taken into account as the most vital feature. It supports computing system by acquiring knowledge from the past experience without any programming. So machine learning is an ultimate solution to reduce human effort because it supports automation with negligible error.

For proper prediction of diabetes, machine learning gives preferable good results. Various machine learning techniques are Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve Bayes. The aim of this paper presentation is to predict possibility of being diabetic by using machine learning model.

A brief introduction of this research paper is discussed as follows: Sect. 2 discusses the brief related work of other researchers. In Sect. 3, description of dataset is mentioned. Section 4 proposes applied methodology. Explanation of the proposed model is done in Sect. 5. Final results outcome is produced in Sect. 6. Section 7 is the concluding part of this research paper.

2 Related Works

Komi et al. [2] had evaluated five types of classifier techniques for diabetic candidate classification. The researcher had tested these methods, namely, Logistic Regression, Expectation Maximization, Support Vector Machine, Artificial Neural

Networks (Anns), and Gaussian Mixture Modelling. To improve the accuracy perfection of the model researcher had tuned some hyperparameters. ANNs outperforms comparatively to other tested models.

Perveen et al. [3] had discussed Bagging and AdaBoost Classifier machine technique models for classification [4]. After considering the various diabetes risk parameters for diabetes classification, the authors had used J48 decision tree. The conclusion of the research was that AdaBoost Classifier, an ensemble machine learning algorithm, proved to be better as compared to Bagging and J48 Decision Tree.

Orabi et al. [5] had developed a prediction model for diabetes patients. The objective of the study was to predict whether a patient is non-diabetic at a specific age . Decision Tree gave an optimum accuracy as it had properly predicted patient's diabetes risk factor at specific age [6, 7].

Pradhan et al. [8] had obtained the diabetes dataset from UCI Repository for classifying and analyzing through Genetic Programming (GP) [9, 10]. GP gives the prediction values with maximum accuracy in minimum cost.

Kumar et al. [11] had tested mainly two models for diabetes prediction. The researcher had used ANN as well as Fasting Blood Sugar (FBS). For detecting the existence of diabetes among patients, Decision Tree [12] was used.

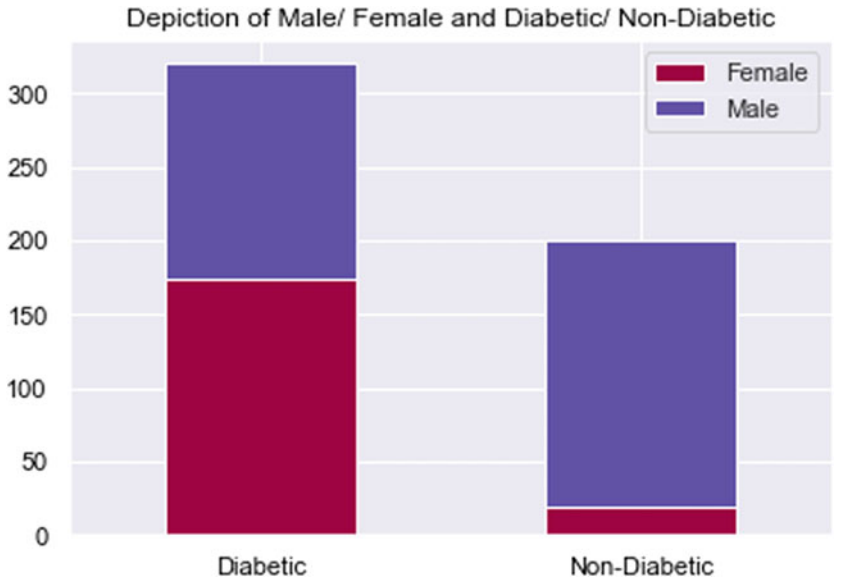
Mandal et al. [13] had described four different types of algorithm for diabetic patient classification, namely, Naïve Bayes, Logistic Regression, ANNs, and Decision Tree. Some hyperparameters were tuned to enhance the accuracy of the model. Random Forest algorithm gave better results.

3 Dataset Description

The patient dataset is collected from Sylhet Diabetes Hospital, Bangladesh. The creation of dataset is based on a direct questionnaire to diabetic patients who have recently been diabetic, and few non-diabetic people who have some symptoms approved by a doctor. The dataset utilized in this paper is present in the UCI Repository [14]. This dataset contains patient's crucial features which are useful for the prediction of diabetes. This diabetes dataset contains 17 attributes of 520 patients out of which 328 are male patients, 192 are female patients as well as 320 are positive and 200 are negative which is represented in the following graph.

Table 1 Dataset description

	Attributes (nos.)	Instances (nos.)
Diabetes symptom dataset	17	520



Graphical Distribution of Male and Female Patients w.r.t. Diabetic/Non-Diabetic

The detailed description of dataset and attributes are as below, respectively (Tables 1 and 2).

Class variables are accustomed to find whether the patient is diabetic (Positive) or not (Negative).

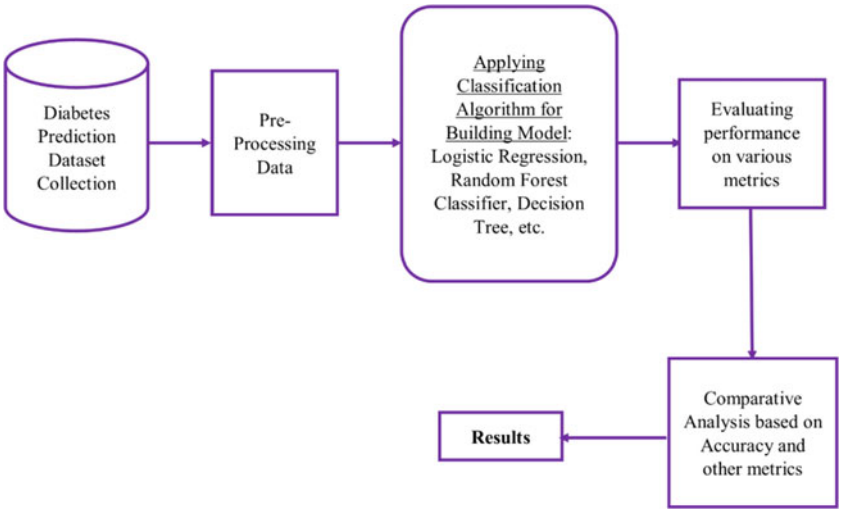
4 Methodology

This research focuses to minimize the complications of diabetes through early prediction so on to enhance the lives of the patients. The person has diabetes because of some considerable features, depending upon their age, gender, weight, and some other factors. The main purpose of research is to find whether the particular patient is diabetic or not by using the Classification technique.

For this, dataset is divided into training and testing sets. For model construction, training dataset is utilized. Testing dataset is employed just for accessing the performance of the model. Therefore, testing dataset is 30% of the entire dataset and rest 70% is utilized as training dataset.

Table 2 Attribute description

Attributes name	Description/Data dictionary
Age	16–90 (years)
Gender	0: Female; 1: Male
Polyuria	0: No; 1: Yes
Polydipsia	0: No; 1: Yes
Sudden weight loss	0: No; 1: Yes
Weakness	0: No; 1: Yes
Polyphagia	0: No; 1: Yes
Genital thrush	0: No; 1: Yes
Visual blurring	0: No; 1: Yes
Itching	0: No; 1: Yes
Irritability	0: No; 1: Yes
Delayed healing	0: No; 1: Yes
Partial paresis	0: No; 1: Yes
Muscle stiffness	0: No; 1: Yes
Alopecia	0: No; 1: Yes
Obesity	0: No; 1: Yes
Class	0: Negative; 1: Positive



Proposed Model Flow Chart

This model has the following different modules:

i. Data Collection
ii. Pre-Processing Data
iii. Building Model

iv. Evaluation
v. Comparative Analysis
vi. Results

Let's have a glance at each model briefly.

4.1 Data Collection

For this research, latest available dataset dated 22 July, 2020, is being utilized. Diabetes dataset is available from the UCI Repository [14]. Latest updated dataset will show comparatively better results. Then information present in this is thoroughly understood by studying its pattern and trends. This diabetes dataset contains information about the symptoms of the patients.

4.2 Pre-processing Data

The data is pre-processed by one hot encoding as the data present is in kind of "Yes" and "No," "Male" and "Female," "Positive" and "Negative," which are converted into 1 and 0, respectively.

4.3 Building Model

This dataset is going to be fed to the various classification algorithms like Naïve Bayes, Logistic Regression, AdaBoost Classifier, Random Forest Classifier, and many other for training and testing data using the classification algorithm.

Subsequent paragraphs, however, are indented.

4.4 Evaluation

The performance of the algorithms is going to be tested with an appropriate evaluation model. For evaluation, various metrics were considered like Error, Recall, F1-score, Support, Accuracy, Macro Average, Precision, Weighted, Average, Area Under Curve (AUC), and most significantly, Confusion Matrix.

Accuracy: Accuracy is the ratio between correct numbers of prediction to total number of predictions. It is given as follows:

$$\text{Accuracy} = \frac{\text{Correct Prediction (no.)}}{\text{Total Prediction (no.)}}$$

Confusion matrix: It is the matrix/table that is accustomed to depict the performance of the test data for which actual values are known.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Now, the accuracy will be calculated from this confusion matrix as follows:

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

Precision: Precision is the ratio between the number of actual true positive results divided by total positive results predicted by the given model, i.e., false positive and true positive. It is expressed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Mean Square Error: Mean Square Error (MSE) is the average of the squared error. MSE is the summation of square of the difference between the actual and predicted values of data points, divided by the total number. Root Mean Square Error (RMSE) is the square root of MSE.

$$\text{MSE} = \frac{\sum_{i=1}^{N} (\text{Predicted}_i - \text{Actual}_i)^2}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\text{Predicted}_i - \text{Actual}_i)^2}{N}}$$

Recall: It is the ratio of true positive results divided by the number of actual positive results predicted by the model, i.e., false negative and true positive. It is expressed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score: F1-Score is used to measure the accuracy of the testing dataset. Harmonic mean of recall and precision is called F1-Score. The value of F1-Score lies between 0 and 1. Mathematically, it is given as

$$\text{F1 Score} = 2 * \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$\text{F1 Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Support: Support is term used for the actual finding of specified category within given dataset. Imbalanced support within the training data may show structural weaknesses in the reported scores of the classifier and the necessity for proportional sampling or rebalancing.

Sensitivity: Sensitivity is the ratio of true positive results divided by the number of actual positive results predicted by the model, i.e., false negative and true positive. It is the metrics used to check whether model can predict true positive from the given dataset.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity: Specificity is the ratio of true negative results divided by the number of actual negative results predicted by the model, i.e., true negative and false positive. It is the metrics used to check whether model can predict true negative from the given dataset.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

4.5 Comparative Analysis

So, considering the above metrics, comparative analysis has been done on this dataset. The main objective of this paper is to tune several hyperparameters of different machine learning methods for improving the accuracy of the proposed model.

4.6 Results

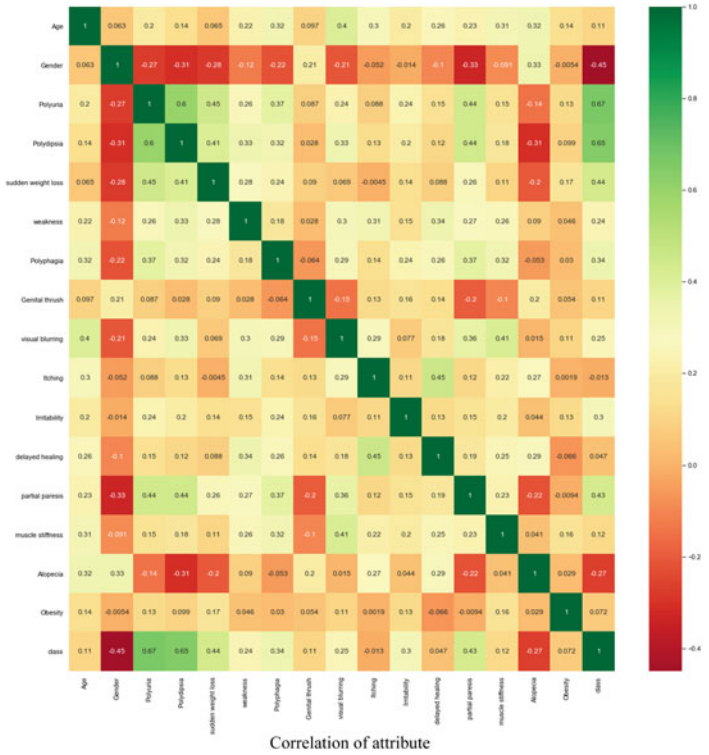
The results are obtained by doing the comparative study and finding the foremost effective algorithm which can be chosen to create the ultimate system for the end users using the dataset as database. So, the final model can be accustomed to predict whether the patient is diabetic or not by taking the symptoms from the user as input and providing it to model.

5 Proposed Models

Modeling

Collinearity, Correlation, and Covariance

When more than two variables are correlated to each other, that will be the case of collinearity. Correlation is a standardized value of strength and direction which shows the connectivity among the given attributes. Covariance denotes non-standardized value of direction and relation among them. The following figure shows the correlation of varied features with each other.



5.1 Logistic Regression

Statistical model which has binary variable quantity and uses logistic function is called Logistic Regression. Therefore, this can be an appropriate model to predict the category whether the patient is diabetic or not. The general form of logistic regression function is

$$\log \frac{\widehat{\pi}_i}{1 - \widehat{\pi}_i} = x_i^T \beta$$

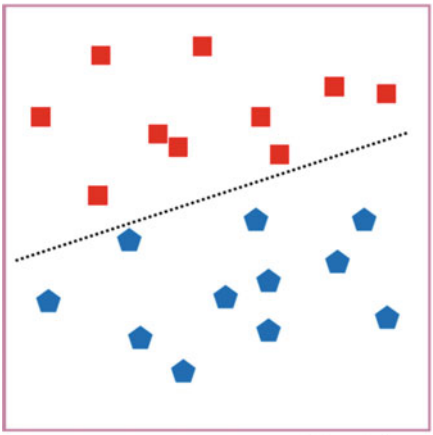
where $\widehat{\pi}_i$ is the estimated probability that observation i is positive, x_i is the ith vector within the design matrix, and β is the vector of coefficients. In this case, the first element of x_i is 1 to activate the intercept in β, the second element of x_i is the age of observation i, and also the remaining elements are 1–0 dummy variables.

5.2 Random Forest

The extension of decision tree is called Random Forest. It builds multiple decision trees and merges them to induce more accurate and stable prediction. Random Forest is used for regression and classification. For an overall better performance, multiple machine learning models are combined which is called ensemble learning. Random forest is one of the ensemble learning models. Ho [15] had created the first algorithm for random forest by using the random subspace method [16].

5.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very efficient, potent machine learning algorithm. It can even achieve far better results than neural networks in some areas. This is often used as a classification model with quite different methodology. Here, the aim is to work out the hyperplane that distinctly classifies the data points and has maximum margin to any or all the other points.



Description of SVM

5.4 K- Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) is the simplest classification algorithm used. It also can be used for regression. K-NN is non-parametric, instance-based (that does not explicitly learn a model). Instead, it chooses to memorize the training instances. The output of K-NN is one of the class member which means result is one of the category from the available categories. There are three key elements in this approach: a collection of labeled object (set of stored record), distance between the objects, and the number of nearest neighbors.

5.5 Decision Tree

Decision Tree (D-Tree) is a supervised learning algorithm used for regression and classification models. D-tree represents the diagrammatic importance of each part. Availability of the number of choice is depicted by branches, Where each leaf shows final decision whether a candidate is diabetic patient or not. D-Tree ishighly interpretable, easy to grasp, and visualize.

5.6 Gaussian Process Classifier (GPC)

It is a classification algorithm based on the Laplace approximation. GPC is a probabilistic classification method which uses Gaussian processes (GP), i.e., output of this is probability of having a particular class.

5.7 AdaBoost Classifier

The AdaBoost Classifier work as a meta-estimator, which applies training and testing methods on the original given dataset. It focuses on improving weights of subsequent identical data. AdaBoost Classifier comes under ensemble boosting classifier technique.

5.8 Gaussian Naïve Bayes (Gaussian NB)

Naive Bayes (NB) could be a powerful classification algorithm used for both multi-class and binary class classification problems. Just by assuming, a Gaussian Distribution Naïve Bayes was further extended to Gaussian Naïve Bayes (Gussian NB).

6 Results and Discussion

See (Table 3).

Table 3 Model accuracy and error

Model name	Training accuracy	Test accuracy	MSE	AUC
Logistic regression	92.8571	93.5897	6.4102	92.7472
Random forest	100.0	99.3589	0.6410	99.2307
Support vector machine	93.6813	94.2307	5.7692	93.9560
K-nearest neighbors	100.0	94.2307	5.7692	94.6153
Decision tree	100.0	98.7179	1.2820	98.9010
Gaussian process classifier	99.7252	98.7179	1.2820	98.6813
AdaBoost classifier	93.1318	94.8717	5.1282	94.5054
Gaussian naïve bayes	89.0109	91.0256	8.9743	90.7692

6.1 Classification Report

Classification report is the report of performance of the machine learning model in the tabular format. It contains the results of metrics on which evaluation of the model had been done. In classification report, it has values of accuracy, precision, and F1-score which gives the detailed idea about the performance of the classification algorithm.Classification reports of various models are shown below:

Logistic Regression

	Precision	Recall	F1-Score	Support
Positive	0.97	0.88	0.92	65
Negative	0.92	0.98	0.95	91
Accuracy			0.94	156
Macro avg	0.94	0.93	0.93	156
Weighted avg	0.94	0.94	0.94	156

Random Forest

	Precision	Recall	F1-Score	Support
Positive	1.00	0.98	0.99	65
Negative	0.99	1.00	0.99	91
Accuracy			0.99	156
Macro avg	0.99	0.99	0.99	156
Weighted avg	0.99	0.99	0.99	156

Support Vector Machine (SVM)

	Precision	Recall	F1-Score	Support
Positive	0.94	0.92	0.93	65
Negative	0.95	0.96	0.95	91
Accuracy			0.94	156
Macro avg	0.94	0.94	0.94	156
Weighted avg	0.94	0.94	0.94	156

K-Nearest Neighbor (K-NN)

	Precision	Recall	F1-Score	Support
Positive	0.90	0.97	0.93	65
Negative	0.98	0.92	0.95	91
Accuracy			0.94	156
Macro avg	0.94	0.95	0.94	156
Weighted avg	0.94	0.94	0.94	156

Decision Tree

	Precision	Recall	F1-Score	Support
Positive	0.97	1.00	0.98	65
Negative	1.00	0.98	0.99	91
Accuracy			0.99	156
Macro avg	0.99	0.99	0.99	156
Weighted avg	0.99	0.99	0.99	156

Gaussian Process Classifier

	Precision	Recall	F1-Score	Support
Positive	0.98	0.98	0.98	65
Negative	0.99	0.99	0.99	91
Accuracy			0.99	156
Macro avg	0.99	0.99	0.99	156
Weighted avg	0.99	0.99	0.99	156

Adaboost Classifier

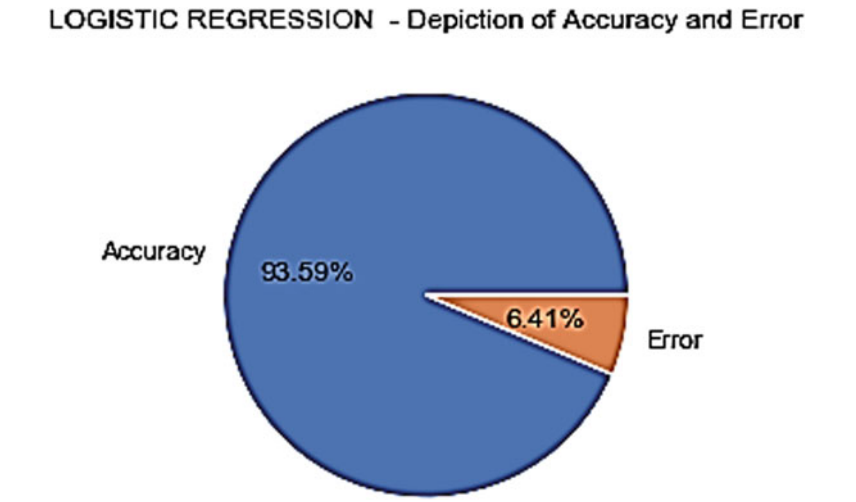
	Precision	Recall	F1-Score	Support
Positive	0.95	0.92	0.94	65
Negative	0.95	0.97	0.96	91
Accuracy			0.95	156
Macro avg	0.95	0.95	0.95	156
Weighted avg	0.95	0.95	0.95	156

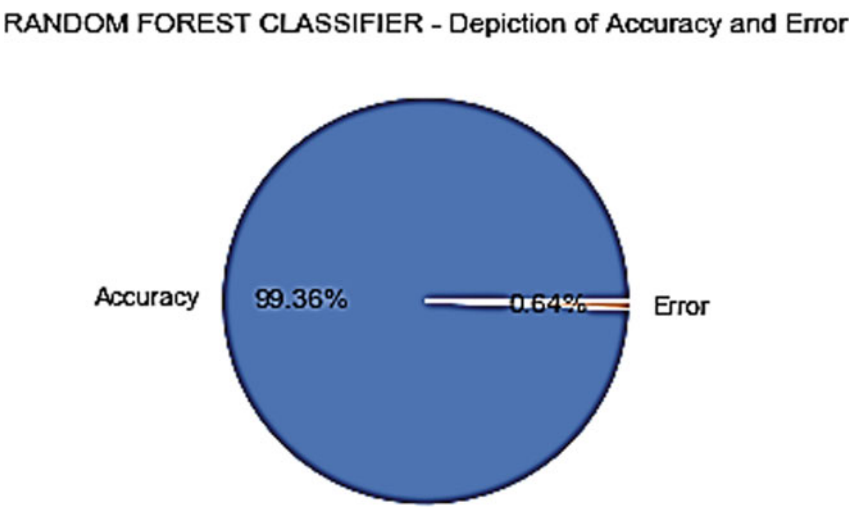
Gaussian Naïve Bayes

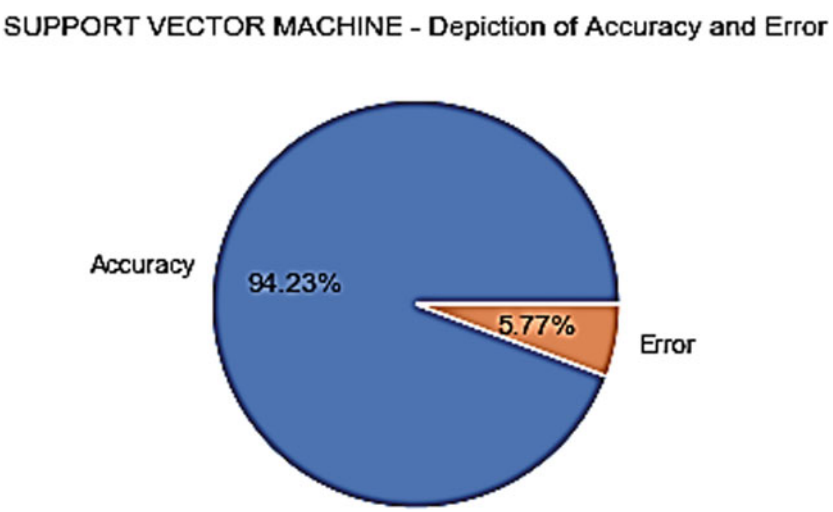
	Precision	Recall	F1-Score	Support
Positive	0.89	0.89	0.89	65
Negative	0.92	0.92	0.92	91
Accuracy			0.91	156
Macro avg	0.91	0.91	0.91	156
Weighted avg	0.91	0.91	0.91	156

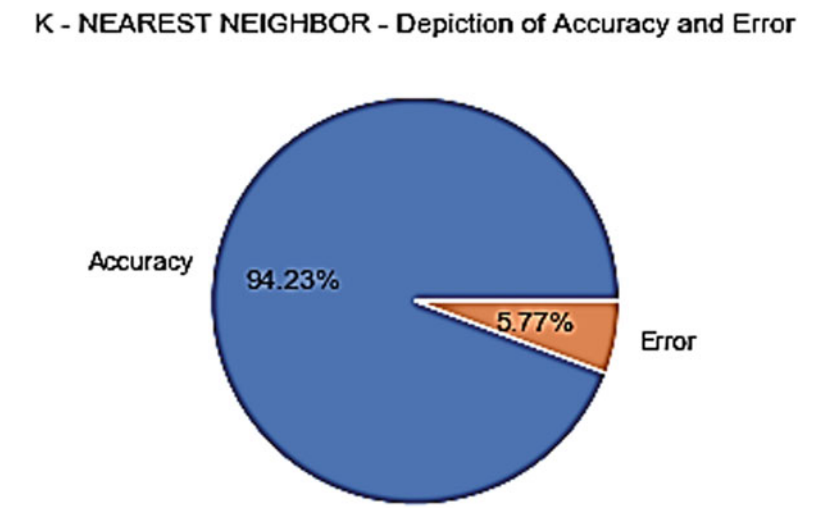
6.2 Graphical Presentation of Accuracy and Error

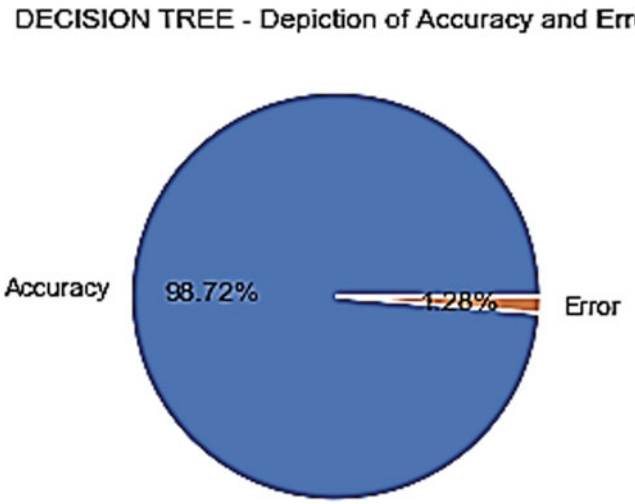
The following pie chart represents a graphical picture of accuracy and error found during the testing of different proposed models. Percentage of accuracy and error denotes a clear concept of productive working of each and every model.

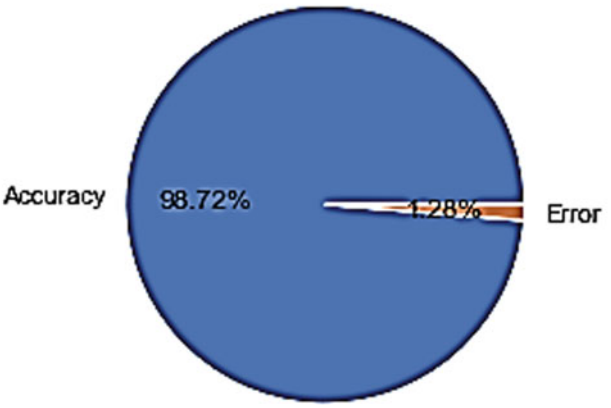

LOGISTIC REGRESSION - Depiction of Accuracy and Error
Accuracy
93.59%
6.41%
Error

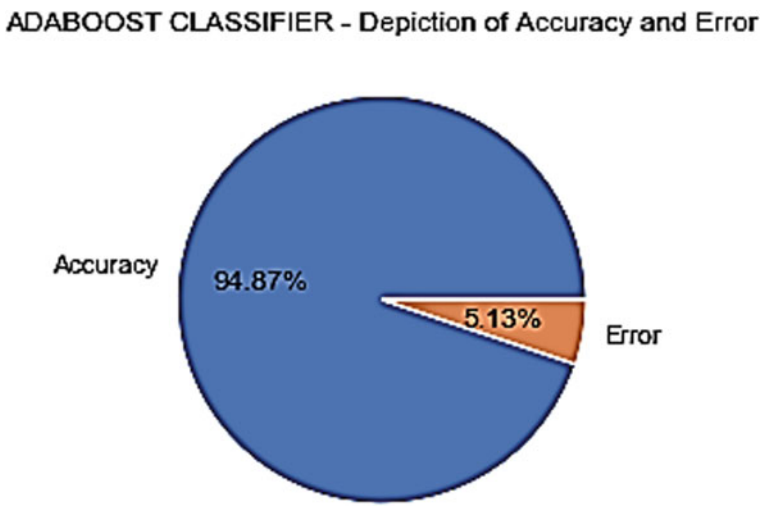

RANDOM FOREST CLASSIFIER - Depiction of Accuracy and Error
Accuracy
99.36%
0.64%
Error

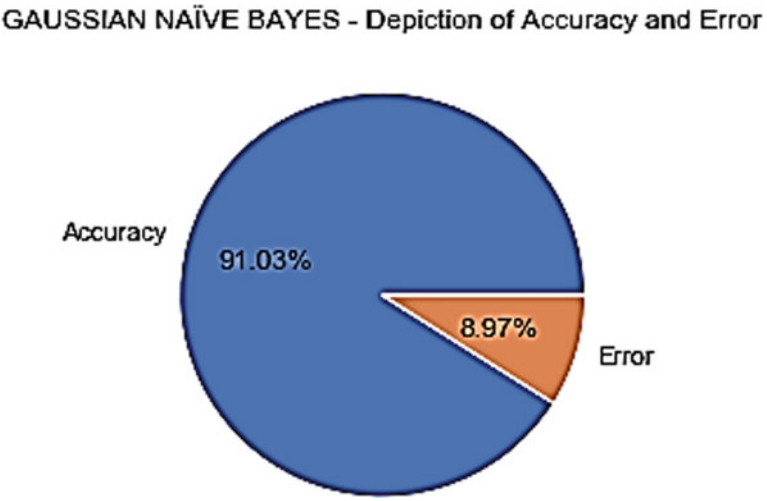

SUPPORT VECTOR MACHINE - Depiction of Accuracy and Error
Accuracy
94.23%
5.77%
Error


K - NEAREST NEIGHBOR - Depiction of Accuracy and Error
Accuracy
94.23%
5.77%
Error


DECISION TREE - Depiction of Accuracy and Error
Accuracy
98.72%
1.28%
Error


GAUSSIAN PROCESS CLASSIFIER - Depiction of Accuracy and Error
Accuracy
98.72%
1.28%
Error





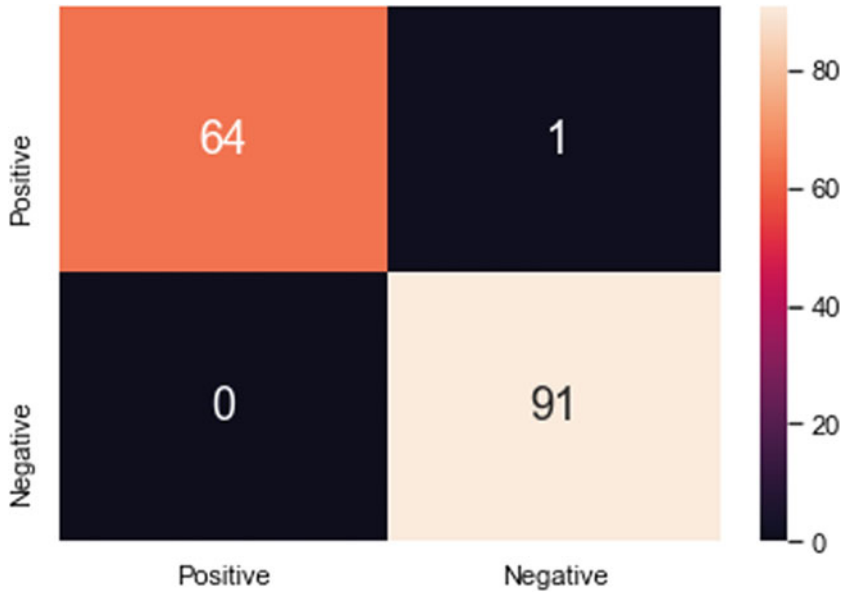
6.3 Confusion Matrix

Confusion Matrix is a 2 × 2 matrix, which used to show the performance of the test dataset for which actual values are known in the schematic format. In this, two types of errors are shown in the diagram. The top right corner in the following confusion matrix depict "Type I" error which means patient was actually non-diabetic but model had predicted it as diabetic patient. And the bottom left corner of confusion matrix depicts "Type II" error which shows diabetic patient was predicted as non-diabetic

by the model. Following are the confusion matrix for different machine learning algorithms.

Logistic Regression

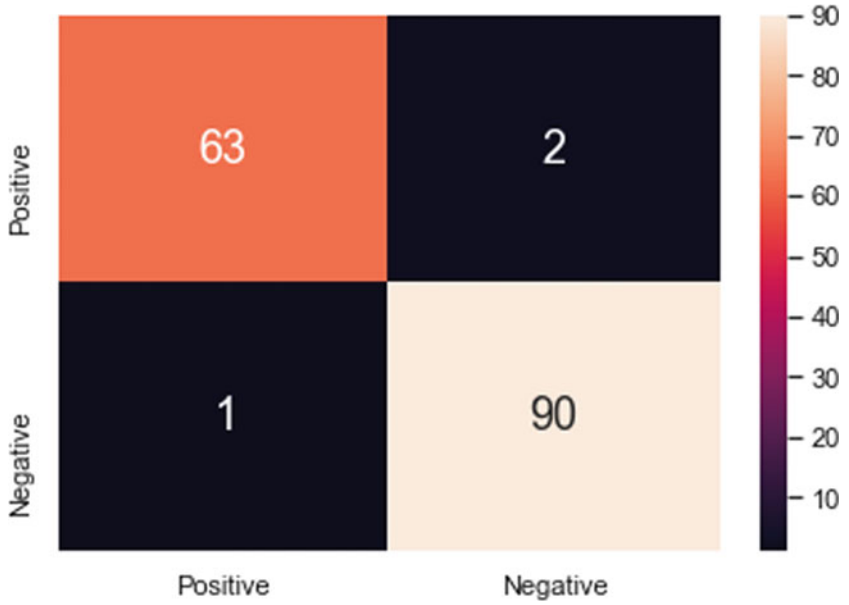
Random Forest



Support Vector Machine (SVM)

K-Nearest Neighbors (K-NN)

Decision Tree



Gaussian Process Classifier

Adaboost Classifier

Gaussian Naïve Bayes

Various classification algorithms were used in the dataset of diabetes. Each and every algorithm was different, because the mathematical logic used for every algorithm was different. Evaluation of model was done based on the subsequent metrics:

i. Accuracy
ii. Precision
iii. Error
iv. Support
v. Area Under Curve (AUC)
vi. Recall
vii. Macro Average
viii. F1-score
ix. Weighted Averages

So, considering these metrics final conclusion has been taken out. Many machine learning algorithms were used and plenty of them gave pretty good accuracy. Logistic Regression gave accuracy of almost 93.59% with error 6.41%. Decision Tree performed well with 98.71% accuracy with error 1.28%. Gaussian Naïve Bayes with accuracy 91.03% and error 8.97%. AdaBoost Classifier with accuracy of 94.87% and error 5.13%. Support Vector Machine (SVM) with 94.23% accuracy and 5.77% error. K-Nearest Neighbors also gave 94.23% with 5.77% error and Gaussian Process Classifier also performed well with 98.71% accuracy and 1.28% error.

But, among all of these machine learning algorithms, Random Forest Classifier gave the highest accuracy of 99.4% with precision of 99.4%, recall of 99.23%, and error of just 0.6%. Random Forest algorithm proved to be the best as compared to other because it gave accuracy of 99.4%. As random forest takes the random subset to build each tree, it had not suffered from high number of predictors. The confusion matrix also shows that random forest had achieved great accuracy, as it had accurately identified all the patients in the test dataset set with no Type II error and only one patient in Type I error. It means patient was non-diabetic but model had predicted it as a diabetic patient.

7 Conclusion

For this research, the patient dataset is collected from Sylhet Diabetes Hospital, Bangladesh. During this study, classification has been done by applying various machine learning algorithms, namely, Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve Bayes were used to create the model for carrying out the diagnosis of diabetes. Furthermore, the machine learning algorithm is tested by evaluating the performance in terms of accuracy. Our comparative analysis is done on the features in dataset and it also shows Random Forest Classifier is the best algorithm for the prediction of newly created datasets made for early stage diabetic risk prediction because it gives the most effective fit to the data

with respect to the diabetic and non-diabetic patients. So, conclusion of this paper is that by applying Random Forest Classifier algorithm with some hyper parameter optimization on diabetes dataset we are able to notice performance of the Random Forest Classifier had achieved higher accuracy. The preferred algorithm technique can be incredibly useful for diabetes prediction. Each prediction of diabetes would be of lot helpful for patient's health. Finally, this research will be extended further by optimizing hyperparameters and considering only important attributes of diabetes dataset to reinforce the performance of model by increasing its accuracy and also to predict in next few years possibility of diabetes to non-diabetic patients.

References

1. Kalyankar, G.D., Poojara, S.R., Dharwadkar, N.V.: Predictive analysis of diabetic patient data using machine learning and hadoop. In: International Conference On I-SMAC (2017). ISBN 978-1-5090-3243-3
2. Komi, M., Li, J., Zhai, Y., Zhang, X.: Application of data mining methods in diabetes prediction. In: Image, Vision and Computing (ICIVC), 2017 2nd International Conference on, pp. 1006–1010. IEEE (2017)
3. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance analysis of data mining classification techniques to predict diabetes. Proced. Comput. Sci. **82**, 115–121 (2016). https://doi.org/10.1016/j.procs.2016.04.016
4. Nai-Arun, N., Sittidech, P.: Ensemble learning model for diabetes classification. Adv. Mater. Res. **931–932**, 1427–1431 (2014). https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427
5. Orabi, K.M., Kamal, Y.M., Rabah, T.M.: Early predictive system for diabetes mellitus disease. In: Industrial Conference on Data Mining, pp. 420–427. Springer (2016)
6. Priyam, A., Gupta, R., Rathee, A., Srivastava, S.: Comparative analysis of decision tree classification algorithms. Int. J. Current Eng. Technol. **3**, 334–337, 2277–4106 (2013). arXiv: ISSN
7. Esposito, F., Malerba, D., Semeraro, G., Kay, J.: A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern Anal. Mach. Intell. **19**, 476–491 (1997). https://doi.org/10.1109/34.589207
8. Pradhan, M., Bamnote, G.R.: Design of classifier for detection of diabetes mellitus using genetic programming. Adv. Intell. Syst. Comput. **1**, 7630770 (2014). https://doi.org/10.1007/978-3-319-11933-5
9. Sharief, A.A., Sheta, A.: Developing a mathematical model to detect diabetes using multigene genetic programming. Int. J. Adv. Res. Artif. Intell. (IJARAI) **3**, 54–59 (2014). https://doi.org/10.14569/IJARAI.2014.031007
10. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020). https://doi.org/10.1109/gucon48875.2020.9231239
11. Kumar, M., Shenbagaraman, V.M., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability Book Chapter, Springer. In: Favorskaya, M.N., Mekhilef, S., Pandey, R.K., Singh, N. (eds.) Innovations in Electrical and Electronic Engineering, pp. 765–773 (2020). ISBN 978-981-15-4691-4
12. Han, J., Rodriguez, J.C., Beheshti, M.: Discovering decision tree based diabetes prediction model. In: International Conference on Advanced Software Engineering and Its Applications, pp. 99–109. Springer (2008)

13. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 768–773. https://doi.org/10.1109/iccca49541.2020.9250790
14. UCI−Machine Learning Repository, Early stage diabetes risk prediction dataset. Data Set
15. Ho, T.K.: Random decision forests (PDF). In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, pp. 278–282 (1995). Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016
16. Ho, T.K.: The random subspace method for constructing decision forests (PDF). IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998). https://doi.org/10.1109/34.709601