2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-9648-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/InCACCT57535.2023.10141734

Diabetes Mellitus Prediction using Supervised Machine Learning Techniques

Srishti Mahajan
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
srishti1480.cse19@chitkara.edu.in

Pradeepta Kumar Sarangi
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
pradeepta.sarangi@chitkara.edu.in

Ashok Kumar Sahoo
Graphic Era Hill University
Dehradun, India
ashoksahoo2000@yahoo.com

Mukesh Rohra
Congnizant Technology Solutions
India
mukesh.rohra@gmail.com

***Abstract*—Diabetes is a long-term condition that occurs when either the body cannot use insulin properly or the pancreas does not produce sufficient amounts of hormone to control blood glucose levels. High blood sugar levels are a hallmark of diabetes, which belongs to a group of metabolic diseases. The two most prevalent varieties of diabetes are type 1 and type 2, but there are other types as well, such as gestational diabetes, which develops during pregnancy. The number of people with type 1 diabetes has significantly increased. The genetic condition known as type 1 diabetes has a long incubation period and frequently manifests early in life. Cells in people with type 2 diabetes do not properly respond to insulin. It changes over time and mostly depends on how people live their lives. According to a 2022 report by the International Diabetes Federation, currently around 382 million people worldwide have diabetes . By 2035, the figure is expected to increase to 592 million. One of the most common causes of tissue and organ damage and dysfunction, including blindness, kidney failure, heart failure, and stroke, is diabetes. As a result, early detection of diabetes is critical. This work aims at implementing two machine learning methods like Logistic Regression and Random Forest for diabetes prediction. Each algorithm is calculated to determine the model's accuracy. Furthermore, the highest accuracy of 99.03% is received by Random Forest.**

Keywords—Diabetes prediction, Random Forest, Logistic Regression, Accuracy, Hybrid model, Machine learning

I. INTRODUCTION

One of the diseases that is constantly spreading and targeting even young people is diabetes and is reported to have increased to 592 million [1]. Diabetes is a metabolic illness that causes the body to behave abnormally, with fluctuating blood glucose levels brought on by pancreatic failure that results in little to no insulin production in the patient's body [2] The root cause of diabetes remains unknown; However, the environment and lifestyle play a significant role in disease development. Despite the fact that it is a fatal disease, treatment and medication are available to treat it [3]. In order to understand diabetes, we need to understand how the body normally uses glucose. Our bodies break down the food we eat, especially the carbs, and convert them to sugar or glucose. Now, the pancreas is supposed to release insulin, which unlocks the cells in the body. Consequently, glucose is able to enter cells and supply the body with energy. However, this approach does not function for diabetic patients. Nowadays, machine learning algorithms are widely used in many sectors and also have shown promising results in the field of medical applications and disease detection [4], [17].

A. Different Types of Diabetes

Diabetes is defined by the presence of high levels of glucose in blood streams and in the urine.

Diabetes has a very complicated etiology and development, and each variety of the disease is caused by a unique set of factors. Type 1 diabetes is of an immune-mediated nature and rises in the body when cells fail to produce insulin in sufficient amounts. All the majority of people living with type 1 diabetes are now adults, also known as "juvenile diabetes". There are no compelling studies that demonstrate the origins of type 1 diabetes or how to prevent it. Type 2 diabetes is primarily due to genetic factors and manner of living. It is a chronic condition which reduces responsiveness to insulin, and can lead to dangerous complications in the body. One of the most prevalent types of diabetes is type 2, thus affecting 95% of the world population. Gestational diabetes appears in pregnant women [18]. Future type 2 diabetes is quite likely to affect these women, as well as perhaps their children. Gestational diabetes is completely treatable, but it requires cautious medication throughout the pregnancy. Maintenance strategies may include blood

ISBN: 979-8-3503-9648-5/23/$31.00 ©2023 IEEE

glucose monitoring, dietary changes and in some cases, insulin may be required.

B. *Symptoms*

The prime symptoms of incurable diabetes are blurry vision, weight loss, increased thirst, urinating frequently, slow healing sores, tiredness/weakness, mood swings/feeling irritable, and getting a lot of infections. Many other symptoms and signs can mark the one set of diabetes.

C. *Causes of Diabetes*

The cause of diabetes depends on the genetics and lifestyle factors. Being overweight or obese may increase the risk, too. Additionally, it has been discovered that viral infections, such as enteroviruses, may be a factor in the death of beta cells that produce insulin. Infection and illness, as well as other forms of stress, can raise the level of blood sugar in the body. 20% of patients who are not overweight, other factors such as specific medication and liver disease contribute to development of diabetes.

II. Literature Review

In [5], the author has used three machine learning techniques for predicting diabetes. Using the Pima Indians Diabetes data set, the authors have reported 76.30% accuracy using Naive Bayes classifier. In [6], the authors like Rastogi et al. have used four machine learning models for diabetes prediction. Using the Kaggle data set, the authors found that the logistic regression gives more accurate results i.e., 82.46% as compared to other machine learning methods. In [7], Patel et al. have proved that Logistic regression gives the highest accuracy of 78% in comparison to other models. In [8], Daanouni et al. predicted diabetes using machine learning, using the Pima Indian Dataset. They tested the data using different neural networks, for instance ANN, KNN, DNN and Decision Tree. They found out that DNN proves to be the optimal algorithm with an accuracy of 90%.

In [9], authors Xue et al. have implemented SVM, Naive Bayes and Light GBM machine learning techniques. Though the authors suggest that it is extremely important to study and compare various predictive methods, to arrive at the most effective one. These techniques have been tested on an open-source dataset from UCI website. The dataset was obtained from Sylhet Diabetes Hospital, Bangladesh. The results depicted that the proposed model beats other methods which includes ANN, DNN, KNN, Decision Tree and many more. Overall accuracy obtained through SVM was 96.54%. According to the authors in [10], Dutta et al. have used Naive Bayes, Random Forest, Decision Tree, XGBoost, and LightGBM machine learning classifiers. Using the newly labeled dataset from Bangladesh, their results are best shown by Logistic regression giving the highest accuracy of 96%.

This paper focuses on and helps in the prediction of disease much before the symptoms start to occur. In [11], Zou et al. made predictions of diabetes using Random Forest (RF), Decision Tree and Neural Networks. The dataset is collected from the China hospital, physical examination data, Luzhou. The authors collectively found out that Random Forest gives the accuracy of 80.84%. In [12], Soni et al. predicted diabetes using six machine learning models, i.e., KNN, SVM, Decision Tree, Logistic Regression, Random Forest and Gradient Boosting. Using the dataset from Pima Indian Diabetes to test the machine learning models, the authors have concluded that Random Forest has maximum performance with 77%. Paper published in 2020 by author Muhammad et al. [13], using SVM, KNN, Logistic Regression, Naive Bayes, Gradient Boosting and Random Forest. The authors have gathered the dataset from a Nigerian hospital named Mohammed Specialist Hospital in Kano state. In reference to other algorithms, Random Forest still holds the highest accuracy of 88.76%.

In [14], Bhat et al. predicted Diabetes at its early stage using supervised algorithms of machine learning like Random Forest, Multilayer perceptron, Logistic Regression, Decision Tree, SVM and Gradient Boosting. According to their observations, a random forest works best with the highest accuracy of 98% among the others. The dataset used in their research was a clinical dataset collected from clinical diabetic professionals. In [15], using KNN and Naive Bayes machine learning classifiers, authors Febrian et al. have predicted diabetes using the data from Pima Indians Diabetes Database. They found out that Naive Bayes has better accuracy than KNN holding up the percentage of 71.37%. The algorithm that is most effective for diabetic vaticination can be determined by comparing the various machine learning techniques used in this study.

TABLE I. Summary Of Recent Works And Their Findings

S.No.	Author (year)	Method	Accuracy (in %)
1	Sisodia et al. [5] (2018)	SVM, Naive Bayes, Decision Tree	Naive Bayes gives more accuracy than Decision Tree and SVM
2	Rastogi et al. [6] (2023)	SVM, Naive Bayes Classifier, Logistic Regression and Random Forest	Logistic Regression performs better as compared to other 3 techniques
3	Patel et al. [7] (2021)	Logistic Regression, Random Forest, KNN, Decision tree	LR has highest accuracy of 78%
4	Daanouni et al. [8]	ANN, DT, DNN, KNN	DNN - 90%
5	Xue et al. [9] (2020)	SVM, Naive Bayes, Light GBM	SVM works best with accuracy of 96.54
6	Dutta et al. [10] (2019)	KNN, LR, XGB, SVM, RF	LR gives highest accuracy of 96%
7	Zou et al. [11] (2018)	Random Forest	80.84
8	Soni et al. [12] (2020)	SVM, RF, KNN, LR, DT, GB	Random Forest - 77%

9	Muhammad et al. [13] (2020)	LR, SVM, KNN, RF, Naive Bayes, GB	Random Forest - 88.76%
10	Bhat et al. [14] (2022)	LR, DT, GB, SVM, RF and Multilayer Perceptron	Random Forest works best with the accuracy of 98%
11	Febrian et al. [15] (2023)	KNN and Naive Bayes	KNN - 69.37% NB - 71.37%

III. OBJECTIVE

This work implements and compares two machine learning models (Logistic Regression and Random Forest) to predict the diabetic condition using the features like personal data and the health parameters. In particular, this work compares the effectiveness of Logistic Regression and Random Forest in predicting the diabetic condition.

IV. METHODOLOGY

In this section we shall explain about the proposed methodology to improve the accuracy of Diabetes prediction. We shall also learn about different classifiers used in machine learning to predict the output. Two different methods used are defined below with their accuracy metrics used in prediction. Fig 1 shows the summarized procedure of conducting research by composed diagram.

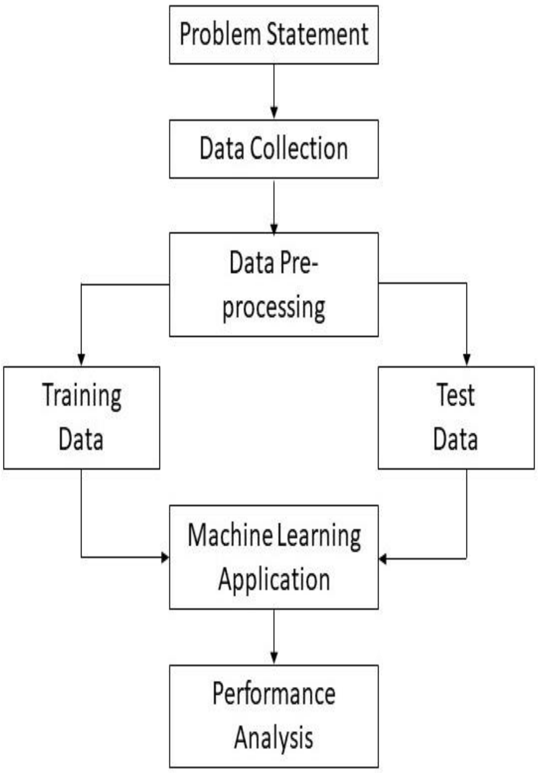
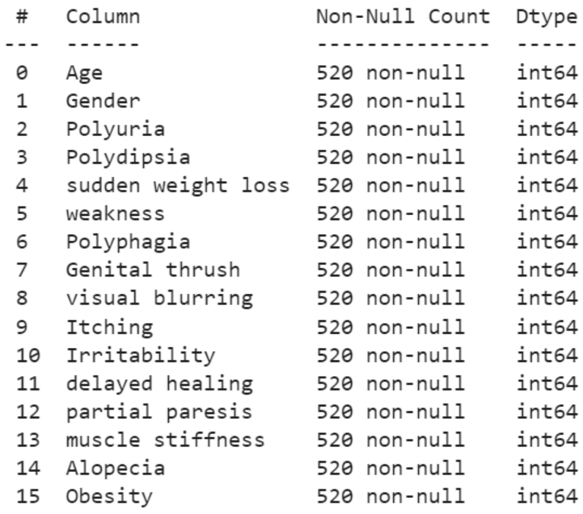


Fig. 1. Research Methodology

A. Dataset used

This work uses a public dataset from Kaggle with the name Early Stage Diabetes Risk Prediction Dataset [16]. This dataset comprises signs and symptoms of would be diabetic or newly diabetic patients. This contains numeric-valued 16 features and one target variable named class treated as either tested negative or positive for diabetes. Dataset description and Attributes description are defined below in table-2.

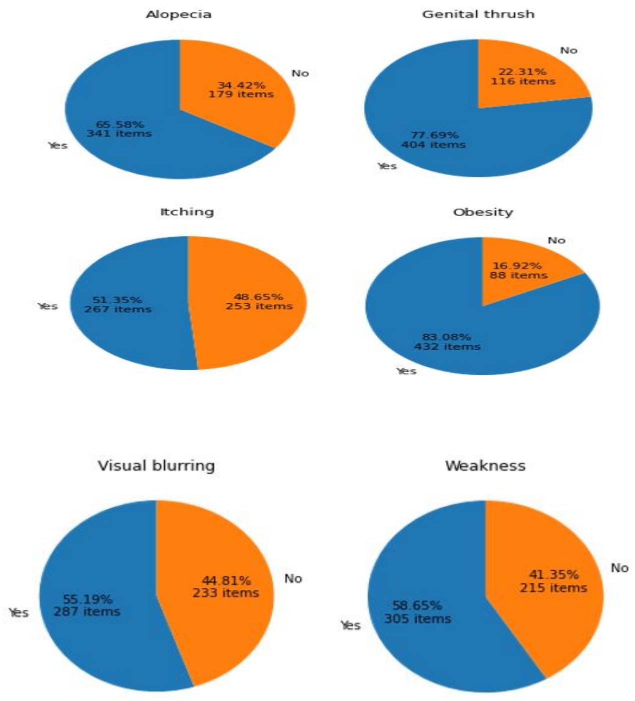
TABLE II. ATTRIBUTES DESCRIPTION



```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    int64
 2   Polyuria            520 non-null    int64
 3   Polydipsia          520 non-null    int64
 4   sudden weight loss  520 non-null    int64
 5   weakness            520 non-null    int64
 6   Polyphagia          520 non-null    int64
 7   Genital thrush      520 non-null    int64
 8   visual blurring     520 non-null    int64
 9   Itching             520 non-null    int64
 10  Irritability        520 non-null    int64
 11  delayed healing     520 non-null    int64
 12  partial paresis     520 non-null    int64
 13  muscle stiffness    520 non-null    int64
 14  Alopecia            520 non-null    int64
 15  Obesity             520 non-null    int64
```

There are no null values in the dataset.

B. Pie charts



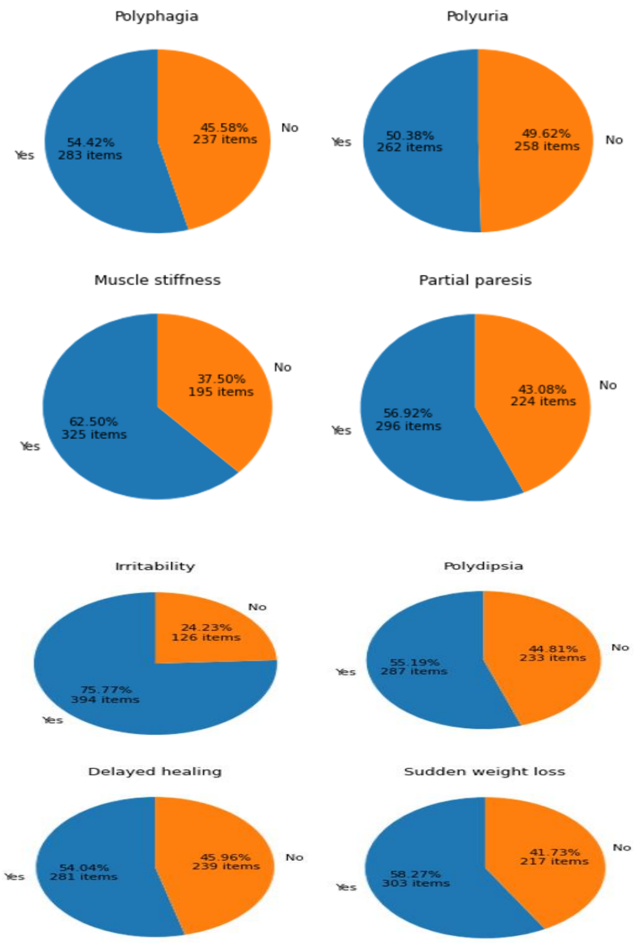


Fig. 2. Occurrence of symptoms

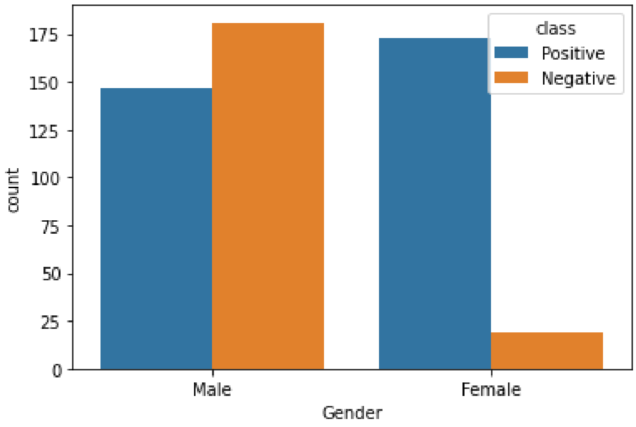


Fig. 3. Gender distribution in the data set

Let's take a look at the above displayed pie charts. Each chart shows the separate features in predicting the diabetes result as Yes - for positive value and No - for negative value. Highest positive results are shown by the factor Obesity and Polyuria almost gives an equal ratio for yes and no.

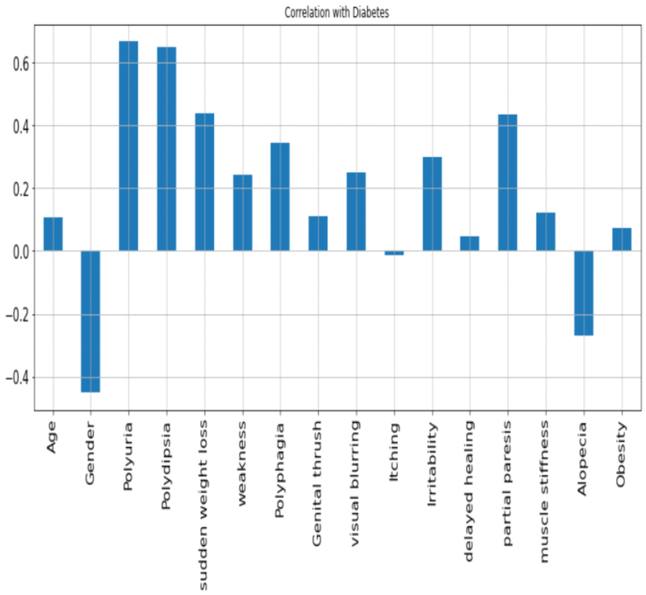


Fig. 4. Correlation with diabetes

C. *Machine learning algorithm used*

1) Logistic Regression : Logistic regression is supervised learning that models the probability of an event. A logistic regression algorithm, sometimes also known as logit model, is a common variable for categorizing the explanatory variables(targets). For instance, to predict whether the patient is diabetic (1) or (0). Logistic regression is used for the disease classification prediction. Let us assume that there are n number of input variables where their values are represented by x1, x2, x3, ………, xn. Let m be the probability of occurring an event and 1-m be the probability of not occurring the event. The following is the Logistic regression equation :

$$\log\left(\frac{m}{1-m}\right) = \log it\ (m) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (1)$$

Or can also be written as

$$m = \frac{e^{(\beta 0\,+\,\beta 1 * x1)}}{(1 + e^{(\beta 0\,+\,\beta 1 * x1)})} \qquad (2)$$

where 0 is the intercept and $\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients.

2) Random Forest : Leo Breiman and Adele Cutler created the most widely used machine learning tree algorithm known as Random Forest. This model is made up of multiple decision trees which combine to reach the single output. It is flexible and easy to use as it handles classification as well as regression problems. Random forest also works well with a large dataset containing a large number of input variables. The tree algorithm stated by creating a combination of trees, each will vote for a class as shown in figure (Fig-5) below.

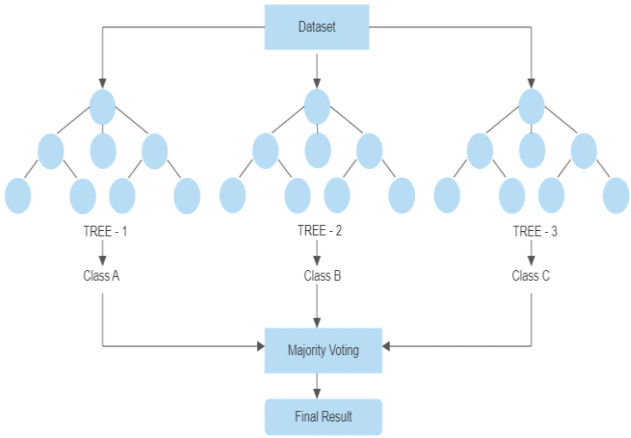


Fig. 5. Random Forest model

The figure represents how the Random Forest Model works. Let us assume that the dataset contains N input variables and K be the number of sampling groups. Each sampling group follows some set of procedure K number of times, these trees become forest. Then comes the classification, that will be selected by majority votes of all the trees in the forest. Therefore, the final result will be evaluated in this process.

D. *Accuracy Measures*

Random forest and logistic regression methods are used in this research work. We use 10-fold internal cross-validation in our experiments. This study is categorized using the following metrics: accuracy, cross-Val accuracy, precision, recall, F1-measure, and ROC (receiver operating characteristic curve) measurements. Table 3 in the text below defines accuracy measures.

TABLE III. PERFORMANCE MEASURE (INDICATORS)

Measures	Definitions	Formula
1. Accuracy (A)	Correct prediction rate of the input patterns	A=(TP+TN)/TP +TN+FP+FN
2. Precision (P)	Measures the classifiers correctness.	P=TP/(TP+FP)
3. Recall (R)	Measure the classifiers recalling capacity	R=TP/(TP+FN)
4. F1-score	To calculate the average precision and recall F1-score is used.	F=2*(P*R)/(P+R)
5. ROC	Is used to compare the functionality of the tests.	

where, TP, TN, FP and FN stand for True Positive, True Negative, False Positive and False Negative respectively. The accuracy of training and testing is also predicted. It is one of the methods to measure the accuracy of a model. From the entire dataset, a testing set and a training set were created.

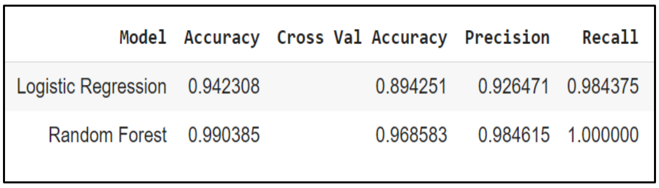
V. RESULTS AND ANALYSIS

Different performance values represented in table 4, classifies algorithms on various measures. Table 4 analysis reveals that Random Forest exhibits the highest degree of accuracy. Therefore, when compared to other classifiers, the Random Forest Machine Learning Algorithm can predict the possibility of diabetes in patients with the highest degree of accuracy.

TABLE IV. PERFORMANCE MEASURE (VALUES)

Model	Precision	Recall	F1-score	ROC
Logistic Regression	0.926471	0.984375	0.954545	0.929688
Random Forest	0.984615	1.000000	0.992248	0.987500

From the table above, the precision values for LR and RF are 0.926471 and 0.984615 respectively. This indicates that the accuracy by LR and RF are 92% and 98% respectively. Similarly, the recall values indicate that the LR model is able to classify 98% cases correctly and that for RF is 100%. The output produced by the execution is given in the figure-6.



```
                Model  Accuracy  Cross Val Accuracy  Precision    Recall
  Logistic Regression  0.942308            0.894251   0.926471  0.984375
        Random Forest  0.990385            0.968583   0.984615  1.000000
```

Fig. 6. Model accuracy

The confusion matrix for the models is shown in the table-5 and table-6 respectively

TABLE V. CONFUSION MATRIX: LOGISTIC REGRESSION

Total: 104	Actual Class: NO	Actual Class: YES
Predicted Class: NO	38	4
Predicted Class: YES	2	60

TABLE VI. CONFUSION MATRIX: RANDOM FOREST

Total: 104	Actual Class: NO	Actual Class: YES
Predicted Class: NO	39	1
Predicted Class: YES	1	63

The performance comparison of the models in terms of accuracy, precision, recall and F1-score is shown in the figure-7.

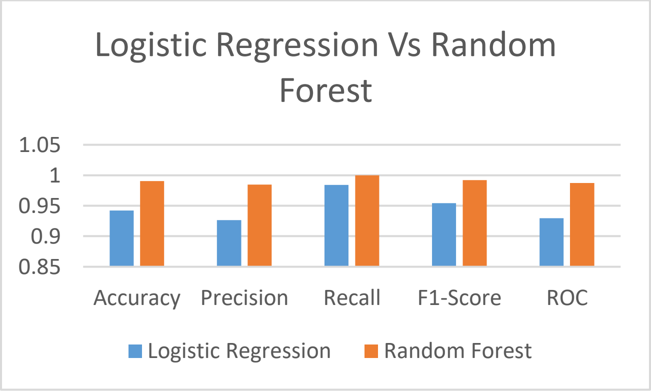


Fig. 7. Performance Comparison

The ROC curve for both the models is given in the figure-7.

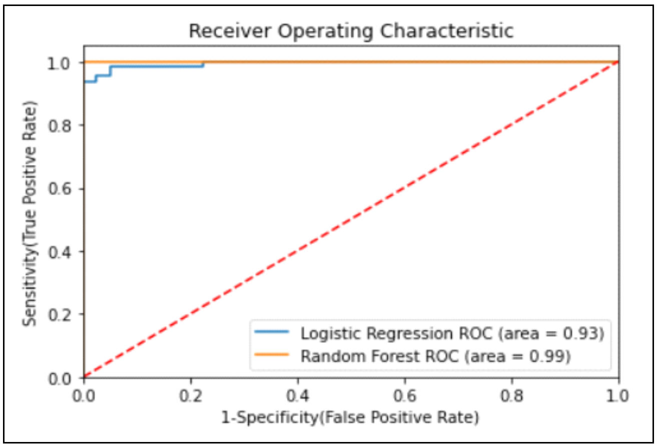


Fig. 8. ROC areas of all classified algorithms

It has been found that the accuracy of the Random Forest model is much higher than that of the Logistic Regression model. The total number of instances is used to evaluate each algorithm's performance. With a greater accuracy of 99.03%, Random Forest is therefore regarded as the best supervised machine learning method.

VI. CONCLUSION

Predicting Diabetes at its early stage may save some one’s life. In this work we have made finest efforts in designing a system and getting the most appropriate results. Two machine learning algorithms are compared across a range of criteria in this study. Publicly available dataset is used from Kaggle for this experimental work. Experimental results of the designed model give the accuracy of 99.03% using the Random Forest classification algorithm. Logistic regression classification can also be implemented for good accuracy of 94.23%. The future work includes prediction or diagnosis of other diseases using other machine learning algorithms. Thus, the work can be refined and expanded in order to automate the analysis of diabetes.

REFERENCES

[1] Diabetes, World Health Organization (WHO): 16 September, 2022.

[2] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin and M. K. Islam, "A Comparative Analysis of Early-Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," 6th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 2021, pp. 654-659.

[3] P. Saini and R. Ahuja, “A Review for Predicting the Diabetes Mellitus Using Different Techniques and Methods”. Proceedings of International Conference on Data Science and Applications. Lecture Notes in Networks and Systems, vol 288. Springer, Singapore. 2022.

[4] S. Bhardwaj, J. Sachin, N. Trivedi, A. Kumar and R. Tiwari. “Intelligent Heart Disease Prediction System Using Data Mining Modeling Techniques”, Soft Computing: Theories and Applications, pp. 881-891, 2022.

[5] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms”, International Conference on Computational Intelligence and Data Science, Procedia Computer Science, pp. 1578–1585, 2018.

[6] R. Rastogi and M. Bansal, “Diabetes prediction model using data mining techniques”, Measurement: Journal of the International Measurement Confederation (IMEKO), Measurement: Sensors Volume 25, February 2023.

[7] K. Patel, M. Nair and S. Phansekar, “Diabetes Prediction using Machine Learning”, International Journal of Scientific & Engineering Research Volume 12, Issue 3, March-2021.

[8] O. Daanouni, B. Cherradi and A. Tmiri, “Diabetes Disease Prediction Using Supervised Machine Learning and Neighborhood Components Analysis”, NISS2020, March 31-April 2, 2020,

[9] J. Xue, F. Min and F. Ma, “Research on Diabetes Prediction Method Based on Machine Learning”, Journal of Physics: Conference Series, 2020.

[10] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud and H. Meshref, “Early Prediction of Diabetes Using an Ensemble of Machine Learning Models”, Int. J. Environ. Res. Public Health, 19, 2022.

[11] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, “Predicting Diabetes Mellitus with Machine Learning Techniques”, Front. Genet. Vol. 9, pp. 1-10, 2018.

[12] M. Soni and S. Varma, “Diabetes Prediction Using Machine Learning Techniques”, International Journal of Engineering Research & Technology, Vol. 9, Issue 09, September-2020.

[13] L. J. Muhammad, E. A. Algehyne and S. S. Usman, “Predictive Supervised Machine Learning Models for Diabetes Mellitus", SN Computer Science, 1, 240, 2020.

[14] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M H. Rahman, “Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora”, Computational Intelligence and Neuroscience, Volume 2022, Article ID 2789760.

[15] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum and R. Yunanda, “Diabetes Prediction using Supervised Machine Learning", Procedia Computer Science, 216 , pp. 21-30, 2023.

[16] Public Dataset used from Kaggle - https://www.kaggle.com

[17] P. Kumar, R. Kumar, and M. Gupta, “Deep learning based analysis of ophthalmology: A systematic review,” *EAI Endorsed Trans. Pervasive Health Technol.*, p. 170950, 2018.

[18] R. Kaur, R. Kumar, and M. Gupta, “Food image-based nutritional management system to overcome polycystic Ovary Syndrome using DeepLearning: A systematic review,” *Int. J. Image Graph.*, 2022.