Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-2 Issue-5, August 2022

A Comparative Analysis of Diabetes Prediction using Different Machine Learning Algorithms

Srinivas Mishra, Aruna Tripathy

Abstract: ***The endocrine disorder diabetes is a condition where the body's glucose levels are abnormally high. Diabetes type II is highly prevalent among elderly people. Worldwide, this number is rising quickly. Furthermore, diabetes creates major health issues that might result in organ failure and paralysis in addition to lowering the blood glucose content. Additionally, it shortens the patients' lives [1]. Early diabetes classification involves seeing a patient at a diagnostic facility and consulting doctors, which is a very time-consuming process. A mechanism has been created to deal with these significant problems. A classification of the patient's level of diabetes using machine learning (ML) algorithms has been addressed in this paper. Previous works considered only five different ML algorithms. We have extended and compared the classification of diabetes prediction using eight different ML algorithms. The database used to train the models is taken from the Pima Indian Diabetes datasets as available from the UCI ML repository [2]. Accuracy, Precision, recall, and F1 score are the four metrics that have been used to analyze and compare the performances of prediction. In comparison to other methods, simulation results indicate that the Neural Network model has the highest accuracy, at 93%. Another performance metric has been the receiver operating characteristics (RoC) that also shows that NN has the maximum area among all the eight algorithms. Simulation results show this area as 0.740.***

Keywords: ***Machine learning, Logistic regression, Gaussian Naive Bayes, Decision Tree, Neural Network, Support Vector Classifier, Artificial Neural Network. K-means, Gradient boosting classifier, K-Nearest Neighbor.***

I. INTRODUCTION

A key goal of safeguarding and preventing the community from diseases that pose a health risk is public health. Governments spend a significant portion of their. But over the course of several years, there has a significant increase in the prevalence of hereditary and chronic disorders impacting general health One of the more serious conditions that diseases that are potentially fatal because they trigger other deadly diseases ailments, including as damage to the heart, kidneys, and nerves [3]. Diabetes is a metabolic condition that makes it difficult for a person's body to process blood glucose, also referred to as blood sugar. - Hyperglycemia caused by deficiencies in insulin secretion, insulin action, or both is the disease's hallmark [3]. Type 1 diabetes is brought on by a complete lack of insulin secretion (T1D). Due to the patient's inability to utilize the produced insulin, diabetes rapidly spreads. Type 2 diabetes (T2D) is the name for it [4]. Both types are growing quickly, although T2D is growing at a faster rate than T1D. T2D accounts for 90 to 95 percent of diabetes cases.In this research, we have developed a method for the categorization, early-stage detection, and prediction of diabetes by utilizing the benefits of the development of machine learning techniques.We used eight commonly used classifiers, including Logistic Regression, Gaussian Naive Bayes, Decision Tree, Neural Network, Support Vector Classifier, and Neural Network, to categorize diabetes into preset groups. K-Nearest Neighbor, Gradient Boosting Classifier, and K-means. Second,for predicting the onset of diabetes The PIMA Indian Diabetes data collection is used for experimental evaluation to demonstrate the effectiveness of the indicated strategy. We came to the conclusion that Neural Network outperformed the other classifiers in experimental evaluation, achieving an accuracy of 93.5 percent in the categorization of diabetes.As shown by the fact that machine-learning algorithms are effective in analyzing a variety of diseases, many researchers are conducting studies to identify diseases using machine learning techniques. These techniques include Support Vector Classifier, Support Vector Machine,Logistic Regression, Gaussian Naive Bayes, Decision Tree, , and Neural Network, . K-NearestNeighbor, Gradient Boosting Classifier, and K-means. Algorithms for machine learning are capable of managing vast volumes of data, combining data from many sources, and including narrative exposition into the research. This well-known book concentrates on diabetes in people of all ages. This study uses decision tree machine learning, support vector classifier, and support vector machine approaches on the Pima Indian diabetes dataset to predict a patient's likelihood of developing diabetes Higher accuracy is attained as a result of the three algorithms' divergent investigation capabilities on various computations. Diabetes has been identified from input features using a better model. Various exercise stances have been categorized and counted using a pose classification system. The health of diabetic people improves more quickly with the right type of exercise.

Manuscript received on 23 July 2022 | Revised Manuscript received on 04 August 2022 | Manuscript Accepted on 15 August 2022 | Manuscript published on 30 August 2022.

* Correspondence Author
Srinivas Mishra*, Research Scholar, Department of Electronics and Instrumentation Engineering, Odisha University of Technology and Research, Bhubaneswar (Odisha), India. E-mail: mishrasrinivas89@gmail.com
Prof. (Dr). Aruna Tripathy, Professor, Department of Electronics and Instrumentation Engineering, Odisha University of Technology and Research, Bhubaneswar (Odisha), India. E-mail: atripathy@outr.ac.in

II. LITERATURE REVIEW

Umair Muneer Butt;et al in 26 March 2021 have been developed Diabetes classification and prediction for healthcare application,

key points in coverage is Machine learning multilayer perceptron (MLP), and logistic regression (LR). Technique is used Classification of different algorithm for prediction diabetes. Parameter used in this paper is Accuracy, Precision, Recall. Researching in Diabetic classification and prediction using different algorithms. Young Ji Woo.et al have developed Diabetes classification and prediction personal home training system using machine learning. Appl. Sci. 2021, 11, 9029. https://doi.org/10.3390/app11199029 keypoints in coverage is human pose estimation, machine-learning-based personal home training system, random forest (RF), multilayer perceptron (MLP), and logistic regression (LR).Technique is used in machine learning based. An approach has been proposed for the classification, early-stage identification, and prediction of diabetes. Parameter used in this paper is Accuracy, Precision, Recall. Researching in Diabetic detection and developing machine learning based personal home training systems using high possibility of correcting the exercise posture. Many systems have been created utilizing different machine learning methods: A machine learning system that might predict diabetes using large data from the healthcare industry was created by Vijayakumar et al. in 2019. The goal was to create a system that can more accurately do early diabetes prediction for a patient. SVM (Support Vector Machine) classification techniques were employed for the categorization of diabetic and non-diabetic data for the earlier diagnosis of the diabetes disease after pre-processing, noise removal, and clustering. Aminul et al. (2017) created a system employing machine learning methods (SVM, Naive Bayes, Logistic Regression) that may forecast the onset of diabetes. Their goal was to use the results of machine learning classification algorithms to identify the start of diabetes in patients.

III. METHODOLOGY

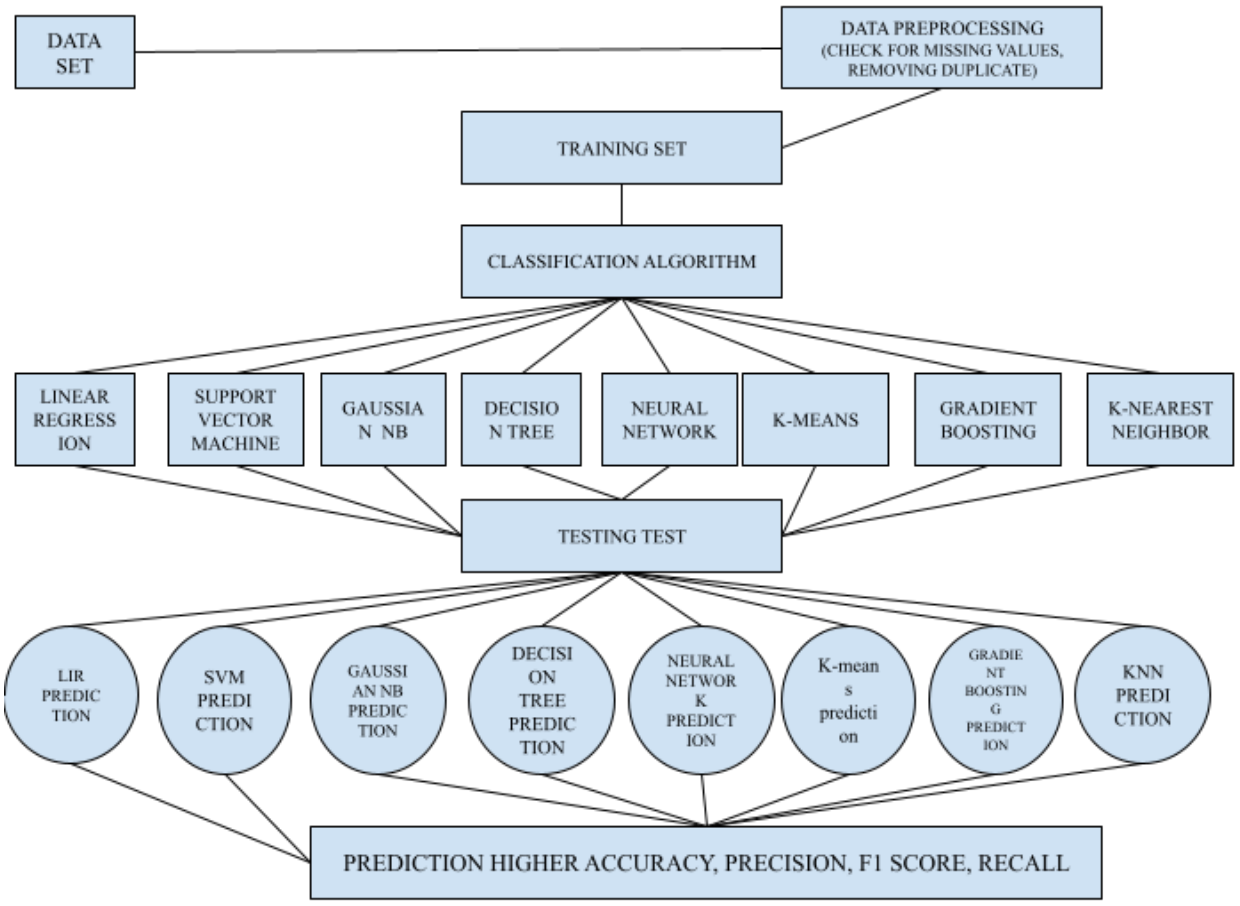

Fig .1 (Model Architecture)

Datasets are screened or verified for missing values at the data preprocessing step; all missing values are then located and replaced with the word "Null." The system was trained using a variety of techniques, including Decision Tree, Linear Regression, Support Vector Machine, K-Nearest Neighbor, Neural Network, Gaussian NB, Gradient Boosting, and K-Means. The training set is made up of a portion of the data that were initially collected. The system was tested using the same classification algorithms that were used to train it on the testing set, which also includes a portion of the information that was initially collected. Individual prediction accuracy is shown, and the most accurate algorithm is chosen as the best model. At the implementation stage, various assessment system components were combined. This involved getting the necessary staff and equipment ready as well as testing the system. Microsoft Excel spreadsheets and Python were used in the system architecture.

The entire system's code was created in Python. It stands for Integrated Development Environment and was utilized to create the machine learning routines.

IV. ALGORITHMS USED

A. Artificial Neural Network

Classification and prediction of patients based on risk factors are one of the applications of artificial neural networks [9]. The human brain consists of billions of nerve cells (neurons). Each neuron is designed to communicate with each other to form a diverse structure of the neural network. It is dedicated to performing different human activities like walking, speaking, breathing, movement, and understanding as well as problem-solving. Artificial neural networks are inspired by biological neural networks and try to mimic human behavior.

Generally, ANN consists of three types of layers.

1. Input Layer: Preprocessed data gets fed into the network.
2. Hidden Layers: These layers consist of inputs, weights, and hidden units. Their relationships can be activated when weights between hidden and input units can be deterministic.
3. Output Layers: These layers give the output depending upon the weights and function of the hidden layer.

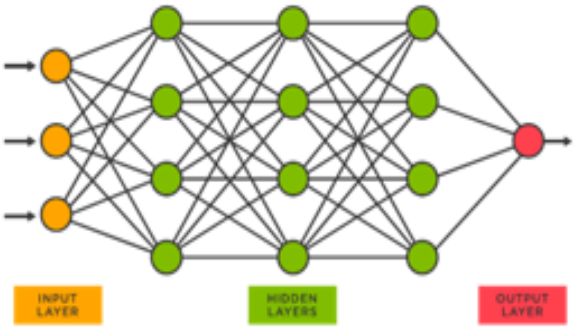

Fig.2. (Graphical Representation of Neural Network)

B. Logistic Regression

It is a supervised learning method that utilizes a predetermined set of independent factors for categorical dependent variables [10]. It explains the relationship between independent and dependent variables and is utilized for predictive analysis. Classifying an input into groups is the outcome of minimizing the cost function. The cost function can be written as:

$$J(\theta)=\frac{1}{m}[\sum_{i=1}^{m} \quad y^i \log h_\theta(x^i)+(1-y^i)\log(1-h_\theta(x^i))] \qquad (1)$$

Where

$$h_\theta(x)=\frac{1}{1+e^{-\theta^T X}},$$

C. Support Vector Machine

Support Vector Machine is working with a supervised technique system. Support vector machines solve two types of problems. One is classification and another is regression. In regression, we are solving continuous type problems. In classification, we are solving to classify the problem. It is best to select a hyperplane that is remote from the data points for each category. The locations that are closest to the margin of the classifier are known as the support vectors. The SVM determines the optimal separation hyperplane by widening the gap between the two decision boundaries. Mathematically, the distance between the hyperplanes $W^T X + b = 1$ and the hyperplane $W^T X + b = -1$is $2/||W||$. [11]The distance has been minimized to $||W||/2$ also SVM should correctly classify x (i), which means

$$y^i(w^T x^i + b) >= 1, \forall i \in \{1, ¢¢, N\}. \qquad (2)$$

D. Gaussian Naive Bayes

The Bayes theorem is used to create a classification approach called Naive Bayes. Under supervised learning techniques, it is a basic but strong approach for predictive modeling. The Naive Bayes approach is simple to comprehend. In the case of missing or imbalanced datasets, it produces better results. Naïve Bayes is a machine learning classifier which employs the Bayes Theorem [12]. The posterior probability P(C|X) can be calculated using the Bayes theorem from P(C), P(X), and
P(X|C).

Therefore,

P(C|X) = (P(X|C) P(C) /P(X) [12]

P (C|X) = posterior probability of target class

P (X|C) = probability of predictor class

P(C) = probability of class C (which is being true)

P(X) = prior probability of predictor class

E. Decision Tree Classifier

A classification-focused supervised machine learning algorithm is a decision tree classifier. Nodes and internodes are used for classification. Instances are categorized by root nodes according to their properties. Additionally, these nodes represent classification while these leaf nodes are made up of two or more branches. [13] Using the most data acquired across all criteria, the decision tree selects each node at each level

F. Gradient Boosting Classifier

Mainy Boosting is divided into 5 types.

1)Light GBM
2)Cat GBM
3)Extreme Gradient Boosting
4)Gradient Boosting
5)Adaptive Boosting

Boosting is work on that machine learning algorithm to increase the accuracy model and combine weak learner to form of strong learner. In Gradient boosting learning happens by optimizing the loss function and use two type of base estimators first is average type model, second is decision tree with full depth. Loss function is calculated by actual values-predicted values. Gradient boosting is method that used for enhance the performance of machine learning model by combining several learners and this algorithm builds models with improved efficiency and accuracy.

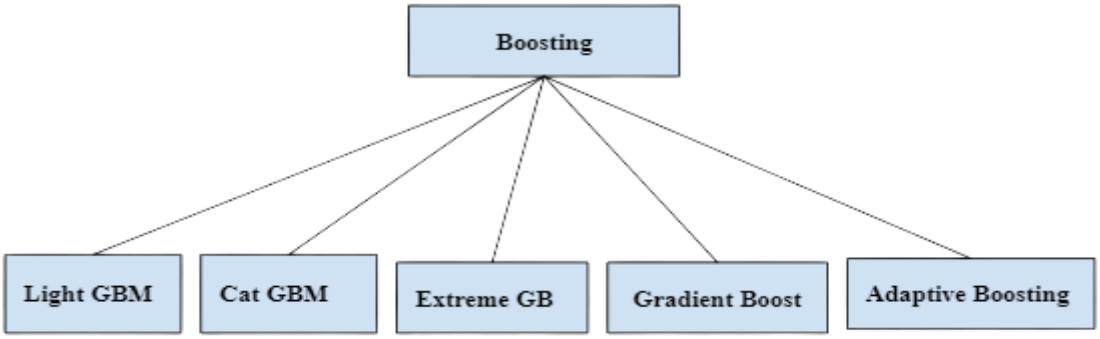

Fig.3. (Boosting classification)

Initialize model with constant value

$$f_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma)$$

For m= 1 to m

Computes residuals

$$r_{im} = -\left[\frac{\partial L((y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=fm-1(x)} \quad \text{for i=1,....,n} \quad (3)$$

Train regression true with features x against r and create terminal node reasons R_{jm} for j=1,...,j_m

Comput $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \epsilon R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

for j=1,...,j_m

Update model

$$f_m(x) = f_{m-1}(x) + v \sum_{j=1}^{j_m} \gamma_{jm} 1(x \epsilon R_{jm}) \quad (4)$$

G. K-Means Clustering

K-means algorithm is working on an unsupervised technique .Clustering is nothing but creating in a different group .The group consists of similar data elements. The group consists of similar type of points. But points in other groups should be different .Clustering is divided into mainly three types.

1)Exclusive clustering,

2)Overlapping clustering,

3)Hierarchical clustering

Exclusive clustering having hard clustering. It is having data point belongs exclusively to one cluster. Overlapping clustering having soft cluster. It is having data point belongs to multiple cluster. K-means is a clustering algorithm mainly used for grouping similar elements or data points into clusters. K represent number of cluster in K-mean

Initialise cluster centroid $\mu_1, \mu_2, \ldots\ldots, \mu_k \epsilon R^n$ randomly.

Repeat until convergence

$$\text{For every i set } c^{(i)} = \underset{j}{argmin} \left\| x^{(i)} - \mu_j \right\|^2$$

$$\text{For each j set } \mu_j = \frac{\sum_{i=1}^{m} 1\{c^i = j\} x^i}{\sum_{i=1}^{m} 1\{c^i = j\}} \quad (5)$$

H. KNN (K-Nearest Neighbor)

Working on a supervised approach system is K-Nearest Neighbor. The following training phase of the straightforward method K-Nearest Neighbor utilizes the whole data set. Every time an unknown data prediction is needed, the full training data set is searched, and the most comparable instances are returned as the prediction. The K Nearest Neighbor algorithm excels in both regression and classification. In this approach, Neighbor is performed as instances, while k is performed as a number. In order to execute this procedure, we must first obtain the data, define k neighbors, calculate the distance between neighbors, and finally assign a new instance to the majority of neighbors.

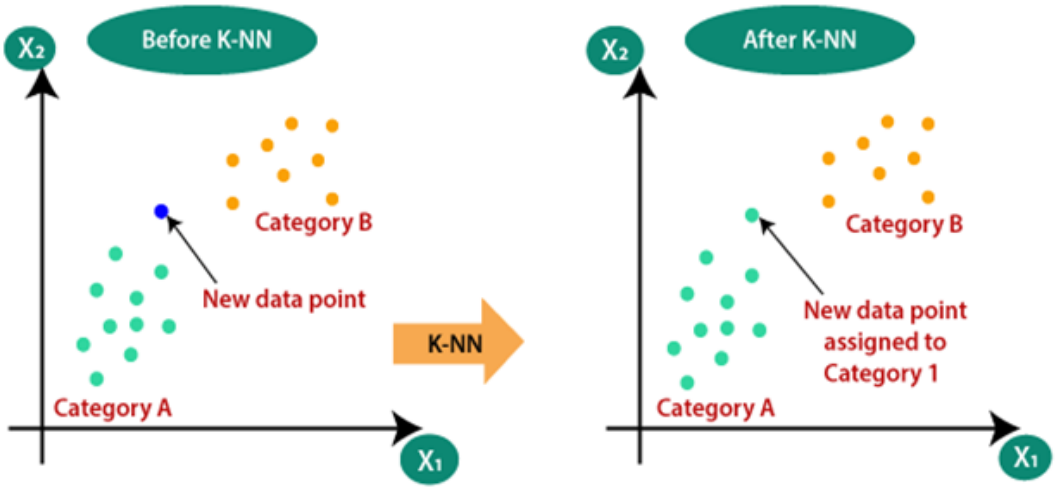

Fig .4. (Figureof K-Nearest Neighbor Algorithm)

V. DATASET USED

The PIMA Indian diabetes dataset is used, which is taken from the UCI machine learning repositories. [14]The dataset consists of 8 features with 768 instances, 7 out of 8 features can be taken as input. The target class (8th feature) is whether the patient has diabetes (1) and is labeled (0) for negative diabetes. The distribution of the target feature outcome is given in the below figure.

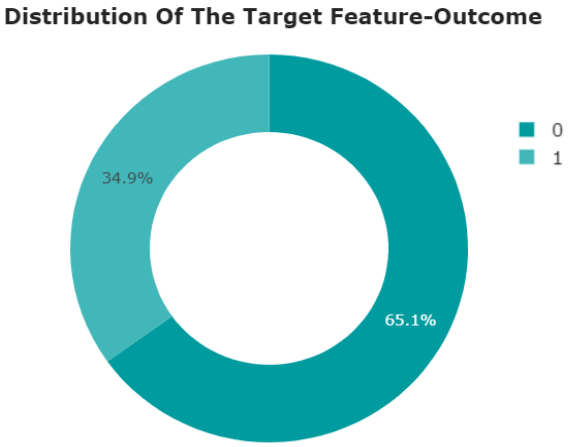

Fig-.5. (Distribution of target feature outcome).

Table -I: Data Available in Data Set

Column	Count
Pregnancies	768
Glucose	768
Blood Pressure	768
Skin Thickness	768
Insulin	768
BMI	768
Diabetes Pedigree Function	768
Age	768
Outcome	768

VI. RESULTS

A. Confusion Matrix

A table called a confusion matrix is used to describe how well a classification system performs. Confusion matrix's major objective is to assist in identifying both the errors and predictions made by a model for various distinct classes. Actual dataset values are shown with columns, whereas anticipated dataset values are shown with rows. True positive is determined by counting the instances in which our real positive values and predicted positive values are equal. Our genuine negative is determined by how frequently our actual negative values match predicted negative values. The frequency with which our model miscalculated a negative value as a positive result gave us false positives. Our approach generates false negatives when it forecasts positive values as negatives.

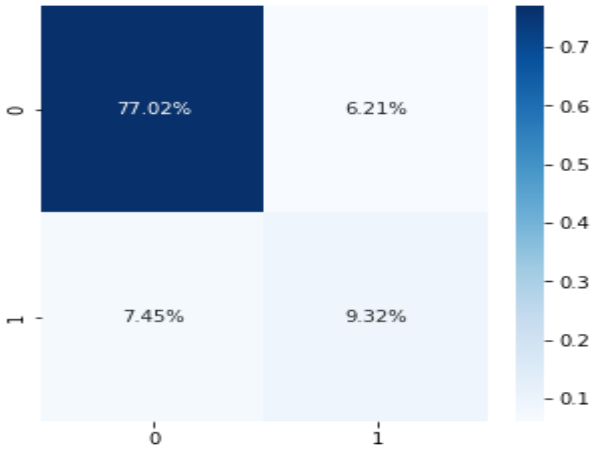

Fig .6. (Confusion Matrix on validation data)

Table. II: Understanding Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

B. ROC

Receiver Operating Characteristics performance of a classification model is represented by a curve at all classification criteria. Between the True Positive Rate and the False Positive Rate is where the Receiver Operating Characteristic plot occurs. A probability curve represents the receiver operating characteristic. The ratio of positive cases in the data set is represented by a horizontal line with a value. The ratio of positive (P) and negative (N) values in the expression Y=P/(P+N) establishes the baseline of the precision-recall curve. The neural network approach performs better in this graph across all categorization thresholds.

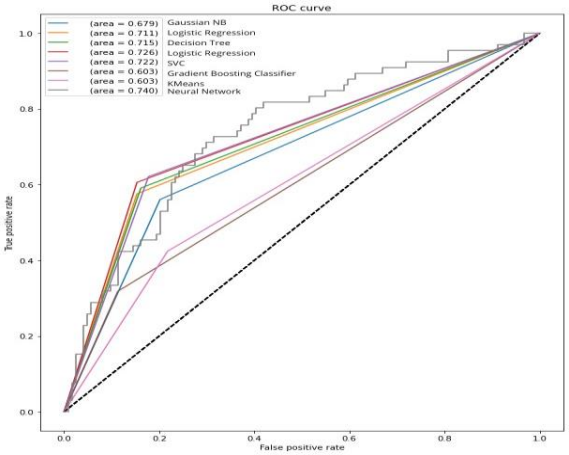

Fig .7. (Confusion Matrix on Validation Data)

C. Evaluation Metrics for Algorithms

As per the methodology described in Fig. 1, the dataset is trained on all the given algorithms and the best performing model has been taken for prediction. Performance of all classifiers based on various measures are plotted by graphs in the given Fig.-8

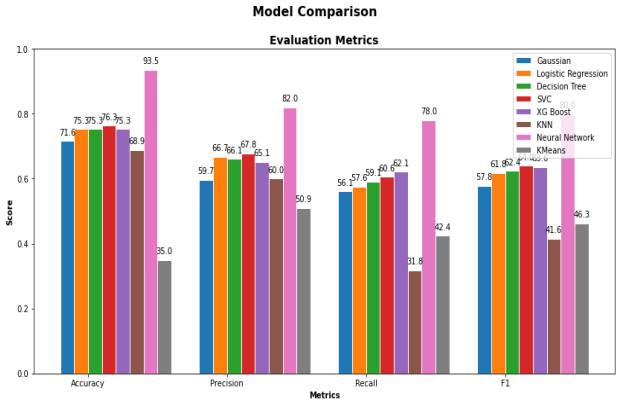

Fig. 8. (Evaluation metrics for different algorithms)

Accuracy: accuracy is the ratio of the addition of true positive and true negative divided by the addition of false-positive, false-negative, true positive, and true negative. The formula is given below.

$$\text{Accuracy} = \frac{tp+tn}{fp+fn+tp+tn}$$

Error rate: It is represented as one minus accuracy, or it is the ratio of the addition of false positive and false negative divided by the addition of false positive, false negative, true positive, and true negative.

$$\text{Error rate} = \frac{(fp+fn)}{(fp+fn+tp+tn)}$$

Precision: Precision is the ratio of true positives divided by the addition of true positives and false positives.

$$\text{Precision} = \frac{tp}{(fp+tp)}$$

Recall is the ratio of true positive divided by the sum of true positive and false negative.

$$\text{Recall} = \frac{tp}{(fn+tp)}$$

F1-Score: It is defined as the harmonic mean of precision and recall. It is the ratio of the product by the sum of precision and recall.

$$\textbf{F1-Score} = \frac{2\ X\ Precision\ X\ Recall}{Precision\ Recall}$$

According to Fig.8 and the analysis, the neural network model has achieved the highest accuracy. The table given below is Parameter's Performance Evaluation of all Algorithms .

Table. III. Parameter's Performance Evaluation of all Algorithms

Model	Acc.	Precision	Recall	F1 Score
Gaussian NB	71.6	59.7	56.1	57.8
Logistic Regression	75.3	66.7	57.6	61.8
Decision Tree	75.3	66.1	59.1	62.4
Support Vector Classifier	76.3	67.8	60.6	64
Neural Network	93.5	82	78	80
K-means	35	50.9	42.4	46.3
Gradient Boosting Classifier	75.3	65.1	62.1	63.6
K-Nearest Neighbor	68.9	60	31.8	41.6

VII. CONCLUSION

This study mainly addressed two problems. The first is diabetes classification, and the second is treating it with a personalized plan. With the aid of the availability of a sizable amount of genetic diabetes dataset, machine learning has the huge potential to transform diabetes prediction. Early diabetes detection is essential for effective treatment. Diabetes cannot be cured, but early detection can save costs and long-term problems. The capacity to forecast diabetes early is crucial for the patient's suitable treatment plan because millions of people worldwide have the disease unaware. However, present machine learning algorithms frequently have correct prediction of accuracy, precision, F1, and recall. The performance of the neural network model was the best. It indicated whether someone would have diabetes or not.

REFERENCES

1. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. *Phys Ther*. 2008;88(11):1254-1264. doi:10.2522/ptj.20080020 [CrossRef]
2. Bădescu, S V et al. "The association between Diabetes mellitus and Depression." *Journal of medicine and life* vol. 9,2 (2016): 120-5.
3. "Facts & figures." 2021. International Diabetes Federation. https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.
4. United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality, "National Survey on Drug Use and Health, 2012." Inter-university Consortium for Political and Social Research [distributor], 2015.
5. M.-A. Moreno-Ibarra, Y. Villuendas-Rey, M. D. Lytras, C. Yáñez-Márquez, and J.-C. Salgado-Ramírez, "Classification of Diseases Using Machine Learning Algorithms: A Comparative Study," *Mathematics*, vol. 9, no. 15, p. 1817, Jul. 2021, doi: 10.3390/math9151817. [CrossRef]
6. S. H *et al.*, "Leisure-time physical activity is a significant predictor of stroke and total mortality in Japanese patients with type 2 diabetes: analysis from the Japan Diabetes Complications Study (JDCS)," *Diabetologia*, vol. 56, no. 5, pp. 1021–1030, 2013, doi: 10.1007/s00125-012-2810-z. [CrossRef]
7. S. A. Kaveeshwar και J. Cornwall, 'The current state of diabetes mellitus in India', *Australas Med J*, τ. 7, τχ. 1, pp. 45–48, Ιανουαρίου 2014. [CrossRef]
8. S. R. Colberg et al., "Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association," Diabetes Care, vol. 39, no. 11, pp. 2065–2079, Oct. 2016. [CrossRef]
9. H. I. Fahmy, G. Develekos and C. Douligeris, "Application of neural networks and machine learning in network design," in *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 2, pp. 226-237, Feb. 1997, doi: 10.1109/49.552072. [CrossRef]
10. Schober, Patrick MD, PhD, MMedStat*; Vetter, Thomas R. MD, MPH† Logistic Regression in Medical Research, Anesthesia & Analgesia: February 2021 - Volume 132 - Issue 2 - p 365-366 doi: 10.1213/ANE.0000000000005247 [CrossRef]
11. D. Sisodia και D. S. Sisodia, 'Prediction of Diabetes using Classification Algorithms', *Procedia Computer Science*, τ. 132, pp. 1578–1585, 2018. [CrossRef]
12. Z.-J. Bi, Y.-Q. Han, C.-Q. Huang, and M. Wang, "Gaussian Naive Bayesian Data Classification Model Based on Clustering Algorithm," in *Proceedings of the 2019 International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)*, 2019, pp. 396–400.
13. R. nbspPatelBrijain and N. K. Rana, "A Survey on Decision Tree Algorithm for Classification," International Journal of Engineering Development and Research, vol. 2, pp. 1–5, 2014.
14. D. Dua and C. Graff, "UCI Machine Learning Repository." 2017.

AUTHORS PROFILE

Srinivas Mishra, M. Tech. in Electronic Communication Engineering Department of Electronics and Instrumentation Engineering, Student of Odisha University Technology and Research (OUTR), Bhubaneswar, Odisha, India. Has strong interest in prediction of Diabetes using Machine Learning with different Algorithms. Has worked upon many projects in Machine Learning and Artificial Intelligence. E-mail: mishrasrinivas89@gmail.com

Prof. (Dr). Aruna Tripathy, Phd in Electronics communication Engineering. Motivating and Talented Professor in Department of Electronics and Instrumentation Engineering, Odisha University of Technology and Research, Bhubaneswar, Odisha, India. My domain includes Machine Learning, communication and signal processing. I always try to improve the performance of students by motivating them and thinking out of the box. With more than 15 years of experience E-mail: atripathy@outr.ac.in