Machine Learning based Diabetes Detection

Mrs. L. V. Rajani Kumari,
Dept. of ECE,
VNR Vignana Jyothi Institute of Engineering & Technology,
Hyderabad, India.
rajanikumari_lv@vnrvjiet.in

P. Shreya,
Dept. of ECE,
VNR Vignana Jyothi Institute of Engineering & Technology,
Hyderabad, India.
ponnalashreya36@gmail.com

Mehrunnisa Begum,
Dept. of ECE,
VNR Vignana Jyothi Institute of Engineering & Technology,
Hyderabad, India.
Mehrunnisabegum5132@gmail.com

T. Pavan Krishna,
Dept. of ECE,
VNR Vignana Jyothi Institute of Engineering & Technology,
Hyderabad, India.
pavankrishnatirutullai01@gmail.com

M. Prathibha,
Dept. of ECE,
VNR Vignana Jyothi Institute of Engineering & Technology,
Hyderabad, India.
prathibhamadupu25@gmail.com

2021 6th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-3587-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCES51350.2021.9489058

Abstract— **Diabetes is a serious disease that affects the majority of the population. It is caused due to increased blood sugar level because of imbalance in insulin processing by the body, which leads to varieties of disorders like Coronary failure, blindness, kidney failure, blood pressure, and it can also affects other parts of the body . The subject needs to consult doctor every time which is time taking. The aim of this work is to make an early prediction of diabetes more precisely by using a variety of machine learning algorithms. Machine learning algorithms provide better results in diabetes detection by constructing models from patient datasets. This project has incorporated the algorithms like Naive Bayes, K-Nearest Neighbor (KNN), Logistic Regression and Random Forest. The accuracy is different for both of them. Our result shows that the K- Nearest Neighbor model shows 78.57% accuracy rate, 87% precision, and 72% Specificity. Logistic Regression model shows 72% accuracy rate, 84% precision, and 63% Specificity. Naive Bayes model shows 71% accuracy rate, 81% precision, and 60% Specificity. Random Forest model shows 76% accuracy rate, 84% precision, and 67% Specificity. Among these algorithms KNN algorithm gives high accuracy.**

Keywords— *Diabetes, Machine Learning, Prediction, Precision, Accuracy, specificity,*K-Nearest Neighbor ,Naive Bayes, *Logistic Regression ,Random forest*

I. Introduction

Diabetes is primarily caused by high glucose levels, a lack or increase in insulin in the body, and obesity [1]. Symptoms that Diabetes patients can experience are Extreme hunger, frequent urination, thirst, fatigue and blurred vision. Pancreas releases insulin in the human body. If insulin secretion is not controlled at early stage it results in disorder of all the body parts and damages heart, nerves, kidneys [2] . If diabetes is caused by a failure to produce enough insulin, the daily secretion of insulin is approximately 30-40 units. As stated in World Health Organization (WHO) 422 million people are experienced from diabetes and the count is expected to be 490 billion in the year of 2030.Mainly Diabetes are three types. Type 1 diabetes are caused due to insufficient secretion of insulin by pancreas. Gestational diabetes is mainly caused due to sugar level imbalance in pregnant women. Type 2 are caused due to insufficient secretion of insulin which can breakdown the body cells. According to research, a study depicts that from 2000 to 2010 gestational diabetes are increased to 56%. During pregnancy most of the women are affected by Type3[3].

In India population is now more than 1300 million. According to International diabetes world records total 88 million people are suffering are with diabetes among them 77 million people are from India. Around 1.5 million people dying with diabetes we can say that diabetes is the main reason of death. Early detection of diabetes can cannot the death rate and save human life [4]. In recent years, Machine Learning Technology came into research and developed many algorithms with different accuracy using machine learning the prediction of diabetes became much more earlier due to its simplicity , accuracy and transparency [5]. In medical research machine learning has wide range of application such as drug discovery and manufacturing, medical imaging diagnosis, smart health record . To predict diabetes there are different algorithms such as KNN, Random Forest and ensemble learning [6,7]. To accomplish work we used PIMA diabetes dataset with different attributes and applied various machine algorithms and predicted outcome [8]. KNN algorithm is used to solve the problems of classification and regression by locating the nearest neighbors [9].

Ensemble methods are divided into 2 types. They are bagging and boosting. In this work, we have used ensemble method gradient boosting algorithm [10]. Machine learning algorithms provide efficient result by classifying dataset and fitting into the model. It is difficult to choose which algorithm gives high accuracy. But it is easy to extract outcome by knowing the dataset. We used popular machine learning techniques that predicts PIMA diabetes dataset are Random Forest, Naïve Bayes and KNN [11-12].

II. MATERIALS AND METHODS

A. *Dataset*

The collected dataset is the PIMA diabetic dataset from Kaggle. The dataset consists of data of 768 subjects Among them 268 subjects were suffering from diabetes. Figure 1 depicts a block diagram of diabetes detection.

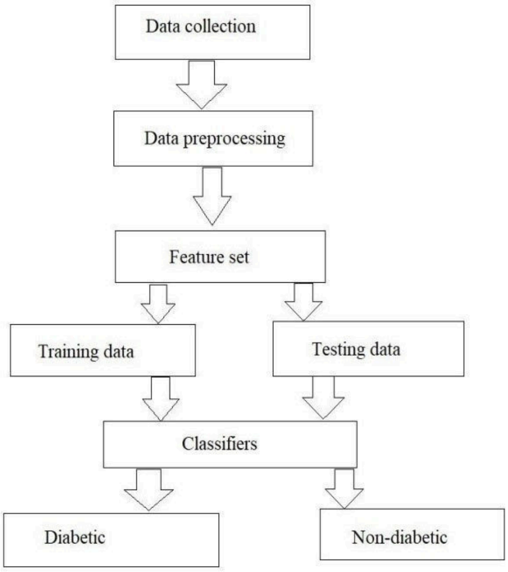


Fig.1. Block diagram of diabetic detection

B. *Data Preprocessing*

Healthcare data may contain impurities and some missing values that may reduce the effectiveness of the data. So, for improving the effectiveness and the quality obtained, data preprocessing is needed. This process is required for successfully getting the result utilizing machine learning methods. We need to perform preprocessing for PIMA diabetes dataset. The data set contains some missing values. So, we removed all the features having with zeros. We created a feature subset by removing irrelevant features. This is known as feature subset selection, and it is useful for reducing data dimensionality and improves the efficiency of the machine learning methods. The reduced feature set is normalized [13]. Then the feature set is decomposed into training set and testing sets.

C. *Feature Set*

The dataset contains 8 features. They are.

I) Total number of times pregnant
II) Glucose/Sugar level
III) Diastolic Blood Pressure
IV) Body Mass Index
V) Skin Fold Thickness in mm
VI) Age of patient in years
VII) Diabetes Pedigree Function
VIII) Outcome

D. *Training and Testing Data*

The total data set is 768*8. 80% i.e., approximately 614*8 (out of 768*8) are used for training and the remaining i.e 154*8 are used for testing data. In Machine Learning, training set is to train the model and the test dataset is used for predictions.

E. *Classifiers*

I. KNN

The KNN classifier is very simple and easy to implement which is used to solve classification problems. This algorithm is a non-parametric method and a lazy learner algorithm since it makes no assumptions about data and does not learn from training datasets during classification. This algorithm assumes the similar ones which are nearby. It helps to group based on the resemblance. Firstly it classify the records on the similarity measure and then finds the distance between points which are closer to the new data, it finds the data points which are closer to the new data point [9]. The K indicates the number of nearest neighbors, this value is chosen. The closeness defined by Euclidean distance.

Steps of KNN algorithm:

Step1: First we need to take a dataset. Here we considered the PIMA dataset which is collected from kaggle.

Step2: After finding training and testing dataset take the testing data attributes.

Step3: Initialize a random value for k which is nearest neighbor.

Step4: Find the distance using the Euclidean Distance formula

Step5: By using the K value and the computed distance, find the nth column for each of them.

Step6: Based on the majority vote, predict the test sample.

II. Random Forest

Random Forest Algorithm is the most popularly known Machine Learning Algorithm, widely used for classification methods. The set of decision tress contributes to form a forest, the random forest algorithm instead of depending on one decision tree, it takes the help of each decision tree and based on the majority, it predicts the final result. To improve the accuracy of the random forest algorithm we need to take more number of decision tree, i.e. The random forest's accuracy improves as the number of features increases [14]. The training time is less when compared with other machine learning algorithms.

Steps of Random Forest algorithm:

Step1: Select the random k features from the set of features.
Step2: Based on the selected features we need to build decision trees.
Step3: Build n number of decision tress by repeating step 1 and 2.
Step 4: Find the prediction of each and every decision tree and based on the majority votes for the category, we can predict the final result/output.

III. Naive Bayes

The Naïve Bayes method is depending on Bayes' Theorem. Each and every pair of features that are classified are independent of one another [15]. It helps in making quick predictions by building the fast machine learning models. It is a possible classifier, that predicts based on an objects' probability. One of its advantages is it training data is not much required. It can handle both the continuous as well as discrete data. It is highly measurable with the help of predictors and data points. It is used to make real-time predictions and it is also fast.

Bayes theorem: This theorem is used when we need to find the probability of hypothesis with previous knowledge. It is dependent on the conditional probability. Its formula is given by

$$P\left(\frac{M}{N}\right) = \frac{P\left(\frac{N}{M}\right) P(M)}{P(N)} \qquad (1)$$

P(M/N) - Posterior probability
P(N/M) - Likelihood probability
P(M) - Prior probability
P(N) - Marginal probability

Steps to implement Naive Bayes classifier:

Step 1: Convert the given dataset into tables of frequency.
Step 2 Generate a table called likelihood table. This table finds the probabilities of the features that are given.
Step 3: Bayes theorem is required to calculate the posterior probability.

IV. Logistic Regression

Logistic Regression is a statistical regression model in which the output variable is categorical rather than continuous [16]. The categorical output variable can represent n different groups. The one vs. all algorithms is used for multi-class classification. The dataset has two classes i.e diabetic and non-diabetic. The main goal of logistic regression is to find the best fit, which is in charge of describing the relationship between the target and predictor variables. The sigmoid function has a range of values from 0 to 1.

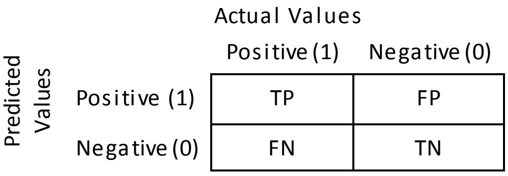
$$s = \frac{1}{1+e^{-z}} \qquad (2)$$

Steps to implement the Logistic Regression :

Step 1:At this stage, we will prepare the data in order to be able to use it effectively in our code.
Step 2:To provide training or adapt the template to the training package, we will import the Logistic Regression class from the sklearn library.
Step 3:The sigmoid function is used in the logistic regression model to predict the probability of positive and negative classes.
Step 4:The confusion matrix will be created here to verify the accuracy of the classification.

F. Performance Matrix

Evaluate the performance of the classifier is done using the confusion matrix.



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

Fig.2. Confusion Matrix

TP is True Positives   TN is True Negatives
FP is False Positive   FN is False Negatives [17].

Accuracy: The fraction of correct predictions to the total predictions [18,19].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (3)$$

Precision: It is the fraction of correct predicted positives to the total predicted positives [17].

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

Specificity: The fraction of no. of true negative predictions to total no. of negatives.

$$Specificity = \frac{TN}{TN + FP} \qquad (5)$$

III. RESULTS

In this paper, diabetic and non diabetic subjects are classified. We took the PIMA diabetic dataset from kaggle which consists of 768 subjects and 8 features among them 268 are suffering from diabetes. This collected dataset is divided into training and testing dataset and given to KNN and gradient boosting classifiers.

A.CONFUSION MATRIX

The KNN classifier confusion matrix is shown in Table 1. The confusion matrix of gradient boosting is shown in table2

Table 1. Confusion matrix for KNN

		True Values	
Predicted Values		positive	negative
	positive	87	13
	negative	20	34

Table 2.Confusion matrix of Logistic Regression

		True Values	
Predicted Values		positive	negative
	positive	84	16
	negative	27	27

Table 3.Confusion matrix of Naive Bayes

		True Values	
Predicted Values		positive	negative
	positive	81	19
	negative	25	29

Table 4.Confusion matrix of Random Forest

		True Values	
Predicted Values		positive	negative
	positive	84	16
	negative	21	33

Table 5.Performance measures of KNN, Logistic Regression, Naive Bayes And Random Forest algorithms

Performance Parameter	KNN	Logistic Regression	Naive Bayes	Random Forest
Accuracy	78.57%	72%	71%	76%
Specificity	72%	63%	60%	67%
Precision	87%	84%	81%	84%

The figure 3 shows the accuracy of the algorithms of KNN Logistic Regression, Naive Bayes and Random Forest. The graph clearly shows that the KNN provides highest accuracy and Naive Bayes gives low accuracy.

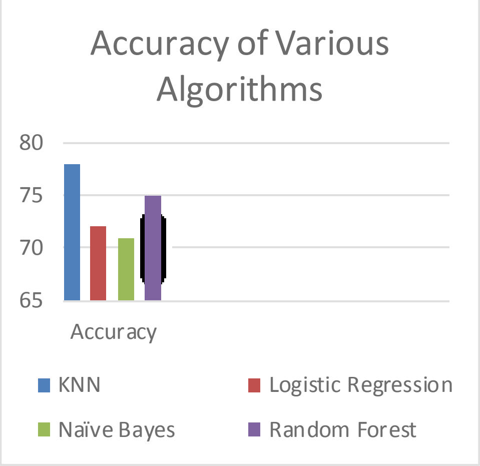


Fig.3. Comparison of an accuracy of all algorithms

IV. CONCLUSION

From the experimental result, we can conclude that the KNN gives more accuracy than the other algorithms. If people want to prevent Diabetes, they should maintain their glucose level constant and should intake proper diet food to keep up their insulin level. People with family that has diabetic history should take care of themselves. KNN algorithm provides an accuracy of 78.57%.

References

[1] Madhusmita Rout, Amandeep Kaur."Prediction of Diabetes Risk based on Machine Learning Techniques", 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020

[2]Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February,2019.

[3] S. Siddiqui, Depression in type 2 diabetes mellitus—a brief review. Diabetes Metab. Synd.Clin. Res. Rev. 8(1), 62–65,2014

[4] K. Rajesh, V. Sangeetha, Application of data mining methods and techniques for diabetes diagnosis. Int. J. Eng. Innov. Technol. 2(3),2012

[5] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928,2018.

[6] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes".Proceeding of International Conference on Systems Computation Automation and Networking,2019.

[7] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and

Application, Vol. 8, Issue 1, (Part -II) January 2018,pp.-09-13

[8] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC),2018.

[9] S. C. Gupta and N. Goel, "Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 980-986, doi: 10.1109/ICSSIT48917.2020.9214129.

[10] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC),2018.

[11] Kemal Polat, Salih Gunes, and Ahmet Arslan, “A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine,” Expert Systems with Applications, vol. 34. 1, January. 2008, pp.482-487.

[12] Jobeda Jamal Khanam, Simon Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express, 2021, ISSN 2405-9595

[13] Aishwarya Mujumdar, V Vaidehi, “Diabetes Prediction using Machine Learning Algorithms”,Procedia Computer Science,Volume 165,2019,Pages 292-299,ISSN 1877-0509

[14] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878802.

[15] K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy and G. R. Mohan Reddy, "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 2020, pp. 603-607, doi: 10.1109/ICOEI48184.2020.9142959.

[16] Karan Bhatia, Shikhar Arora, Ravi Tomar. "Diagnosis of diabetic retinopathy using machine learning classification algorithm", 2016 2nd International Conference on Next Generation Computing Technologies (NGCT),2016

[17] Mafizur Rahman, Linta Islam. "DiabetesRecognition in Pregnant Women by Extracting Features Using PCA and Data Mining Algorithms", 2019 IEEE Pune Section International Conference (PuneCon), 2019

[16] Karan Bhatia, Shikhar Arora, Ravi Tomar. "Diagnosis of diabetic retinopathy using machine learning classification algorithm", 2016 2nd International Conference on Next Generation Computing Technologies (NGCT),2016

[17] L. V. Rajani Kumari, Y. Padma Sai & N. Balaji “ R-Peak Identification in ECG Signals using Pattern-Adapted Wavelet Technique”, IETE Journal of Research, 2021DOI: 10.1080/03772063.2021.1893229

[18] V.Rajani Kumari, L.; Padma Sai, Y.; Balaji, N.. “Performance Evaluation of Neural Networks and Adaptive Neuro Fuzzy Inference System for Classification of Cardiac Arrhythmia”. International Journal of Engineering & Technology, [S.l.], v. 7, n. 4.22, p. 250-253, nov. 2018. ISSN 2227-524X

[19] L.V.R. Kumari, Y. Padma, Sai, N. Balaji and K. Viswada, "FPGA Based Arrhythmia Detection", Procedia Computer Science, vol. 57, pp. 970-979, 2015