International Journal of Modern Developments in Engineering and Science
Volume 2, Issue 4, April 2023
https://www.ijmdes.com | ISSN (Online): 2583-3138


# Diabetes Prediction using Machine Learning

Aaditi Ranganath Satam[1], Tanmay Dilip Dhumale[2*], Pratik Rajesh Hare[3], Hritika Dinesh Ghosalkar[4], Aarti Bakshi[5]

[1,2,3,4]*Student, Department of Electronics and Telecommunication Engineering, K. C. College of Engineering and Management Studies and Research, Thane, India*

[5]*Professor, Department of Electronics and Telecommunication Engineering, K. C. College of Engineering and Management Studies and Research, Thane, India*

***Abstract*: Making accurate medical diagnoses requires the discovery of knowledge from medical datasets. Diabetes is frequently referred to as diabetes mellitus (DM) by medical professionals and describes a group of metabolic diseases in which a person has high blood sugar due to insufficient insulin production, improper insulin cell response, or a combination of both. Now is the time to start preventing and early-stage diabetes diagnosis. It is not only a disease but also a cause of many other diseases, including kidney disease, blindness, and heart attacks. The standard diagnostic procedure requires patients to visit a diagnostic facility, consult their doctor, and wait for a day or more to receive their results. Making accurate medical diagnoses requires the discovery of knowledge from medical datasets. Here in our project with the help of classifiers we are predicting if the person is diabetic or not. Each classifier has different accuracy. We came to know by calculating the confusion matrix. By which we found out that the highest accuracy is provided by Logistic Regression classifier i.e.79%.**

***Keywords*: Random forest, Decision tree, prediction, classifiers, machine learning.**

## 1. Introduction

After heart and cancer diseases, diabetes is the third leading cause of death. But thanks to the development of machine learning methodologies, we can now resolve this problem. The goal of machine learning is to generate concise and understandable descriptions of patterns by extracting knowledge from datasets of data. We are going to use machine learning to create a diabetes diagnosis system that can determine whether a patient has diabetes or not. Additionally, early disease prediction enables treating patients before their condition deteriorates. From a vast amount of diabetes-related data, machine learning has the capacity to extract hidden knowledge. This project examined and analysed the most recent research on diabetes classification. Additionally, the study used Decision Tree, Support Vector Machine, and K Nearest Neighbour Algorithm to create a classification model for diabetes. A dataset of 768 cases gathered from various National Institutes of Diabetes and Digestive and Kidney Diseases forms the basis of the classification model. There are 406 men and 362 women among these cases. Medical experts can categorize and diagnose diabetic patients using the findings. These findings support the classification of diabetes by medical professionals.

The goal of this study is to predict diabetes disease at an early stage using various machine learning algorithms. KNN, Decision Trees, and Support Vector Machines, for example, can be used to predict this chronic disease early on in order to protect human life. The range of sugar levels before meals is between 80 and 130 milligrams per decilitre (mg/dl), or 4.4 to 7.2 millimoles per litre (mmo/L), and for two hours after meals, less than 180 mg/dl (10.0 mmol/L). If these conditions are met, the result will indicate that the person does not have diabetes. But if exceeds this range then the result will represent that the person is having diabetes. Some people will have slightly higher blood sugar goals, including people who are of age beyond 60years, or have other medical conditions such as heart, lung or kidney diseases.

## 2. Literature Survey

We have referred different international papers and journals in which we came to know that various methods were used by different authors. With the help of the previous done works from the above papers we were able to build this project. Below are some authors and their proposed methodology and their limitations.

In order to assess whether or not a person has diabetes, Yasodha et al. [1] used categorization on a number of datasets. The data set for the diabetic patient was produced using information from the hospital's data warehouse, which comprises 200 instances and nine attributes. These instances of the dataset refer to the blood tests and urine tests groups. WEKA can be used in this study's implementation to categorize the data, and the data is evaluated using a 10-fold cross validation approach because it compares the results and excels on tiny datasets. We employ the naive Bayes, J48, REP Tree, and Random Tree. The results showed that among the others, J48 works best, with an accuracy of 60.2%.

Aishwarya et al. [2] By researching and examining the patterns that emerge in the data via classification analysis utilising Decision Tree and Nave Bayes algorithms, the goal is to find ways to diagnose diabetes. The goal of the study is to suggest a quicker and more effective technique of diagnosing the illness, which will aid in the timely treatment of the patients. The study found that the J48 method has an accuracy rate of

*Corresponding author: tdhumale45@gmail.com

74.8% while the naive Bayes algorithm has an accuracy rate of 79.5% when utilizing a 70:30 split.

Gupta et al. [3] 's study compares the performance of the same classifiers when implemented on some other tools, including Rapid miner and Matlab, using the same parameters in order to find and calculate the accuracy, sensitivity, and specificity percentages of numerous classification methods as well as to compare and analyze the results of several classification methods in WEKA (i.e., accuracy, sensitivity and specificity). The JRIP, Jgraft, and BayesNet algorithms were used. According to the results, Jgraft has the highest accuracy (81.3%), sensitivity (59.7%), and specificity (81.4%). Also, it was found that WEKA performs better than Matlab and RapidMiner.

Lee et al. [4] after applying the resample filter to the data, main focus is on using the decision tree method CART on the diabetes dataset. The author places a strong emphasis on the issue of class imbalance and the need to address it before implementing any method in order to increase accuracy rates. The majority of class imbalances occur in datasets with dichotomous values, which implies that the class variable has two alternative outcomes and can be handled easily if discovered earlier in the data preprocessing stage. This will also assist to increase the predictive model's accuracy.

## 3. Proposed Method/System

In many real-world problems, classification is one of the most crucial decision-making approaches. Classifying the data as diabetic or non-diabetic and increasing classification accuracy are the key goals of this endeavour. The more samples used in a classification challenge doesn't always result in a more accurate categorization. In many instances, an algorithm performs well in terms of speed but poorly in terms of data classification accuracy. Our model's primary goal is to attain great accuracy. If we employ a sizable portion of the data set for training and a small portion for testing, classification accuracy can be increased. In order to classify data as either having diabetes or not, many classification strategies were examined in this survey. As a result, it is noted that the implementation of the diabetes prediction system is best served by techniques like logistic regression. The classification of the diabetes disease is the major goal of the machine learning models. The dataset must first be split into two sets, say 80% for training and 20% for testing. The test and training sets are kept apart. The most important diabetes illness risk indicators have been chosen in the second step using PCA feature selection. We have implemented three classifiers: decision tree, random forest, and logistic regression (RF). The patients have been divided into two groups—diabetic and control—by applying test classifiers after estimating the training classifier coefficients. Five performance parameters—accuracy, F1-score, recall, precision, and AUC—are used to assess the classifiers' performances

## 4. Classifiers Used

*Logistic Regression (LR):* In statistics, a regression model known as logistic regression uses a binary dependent variable, or one that can only have the values "0" or "1," to describe outcomes like pass/fail, win/lose, alive/dead, or healthy/ unhealthy. The majority of medical areas, social sciences, and machine learning all use logistic regression. For instance, logistic regression was initially used to generate the Trauma and Injury Severity Score (TRISS), which is frequently used to predict death in injured patients. Logistic regression has been used to generate numerous more medical measures that are used to evaluate a patient's severity. The method is also applicable to engineering, particularly when determining the likelihood that a certain process, system, or product may fail. The ability to estimate a customer's likelihood to buy a product or cancel a subscription is frequently employed in marketing applications. A business application is about to estimate the likelihood of a homeowner defaulting on a mortgage. It can be used to predict a person's likelihood of choosing to be in the labour force in economics. Natural language processing employs conditional random fields, a logistic regression extension to sequential data. In this study, the presence or absence of diabetes was predicted using logistic regression using seven patient features.

*Random Forest (RF):* One of the well-liked and flexible algorithms employed in ensemble technique is RF. It is the most effective and well-liked machine learning algorithm for the hybrid model concept, which improves performance and predictability. Large data sets and high dimensionality are simple to handle using RF. The samples are chosen at random.

*Decision Tree (DT):* A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. Whereas Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions. It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree.

## 5. Results

A method for summarising a classification algorithm's performance is the confusion matrix. You can acquire a better understanding of the categorization model's successes and failures by calculating a confusion matrix. Each classifier has different accuracy. We came to know by calculating the confusion matrix. By which we found out that the highest accuracy is provided by Logistic Regression classifier ie 79%. Here is the confusion matrix and classification report of each classifier. For every confusion matrix the diagonal values

Table 1
Classification report

| MODEL | PRECISION | | RECALL | | F1- SCORE | | ACCURACY |
|---|---|---|---|---|---|---|---|
| | Diabetic | Non-Diabetic | Diabetic | Non-Diabetic | Diabetic | Non- Diabetic | |
| Logistic Regression | 70% | 83% | 66% | 85% | 68% | 84% | 79% |
| Random Forest | 71% | 81% | 60% | 87% | 65% | 84% | 78% |
| Decision Tree | 63% | 82% | 68% | 79% | 65% | 81% | 75% |

indicate the correct prediction and the off-diagonal values indicate the confused predictions. The fig. 1, indicates the confusion matrix for decision tree and for this we get the recall as 66% for diabetic and 85%, precision score as 63% for diabetic and 82% non-diabetic, F1-score as 65% for diabetic and 81% for non-diabetic and an accuracy of 75%.Then for the next classifier which is random forest the fig. 2 shows the classification matrix and for which the recall is 60% for diabetic and for non-diabetic is 87%, precision score for diabetic is 71% and for non-diabetic 81%, F1-score as 65% for diabetic and 84% for non-diabetic and an accuracy of 78%.For the final classifier which is Logistic Regression the classification matrix is shown in fig. 3. This classifier gives us the highest accuracy of 79% for the overall data. The recall for diabetic data is 66% and for non-diabetic is 85%, precision is about 70% for diabetic and 81% for non-diabetic and the F1-score as 68% for diabetic and 84% for non-diabetic.


Fig. 1. Decision Tree confusion matrix

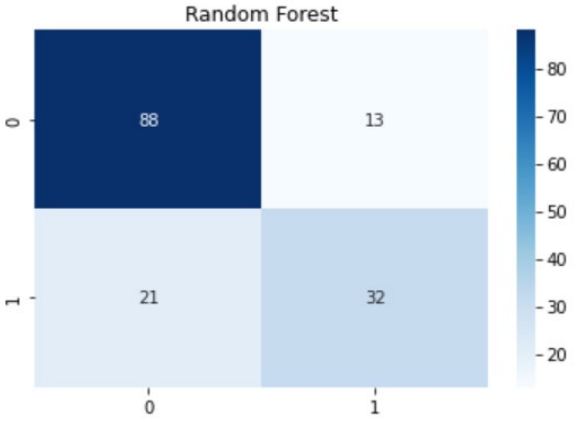


Fig. 2. Random Forest confusion matrix

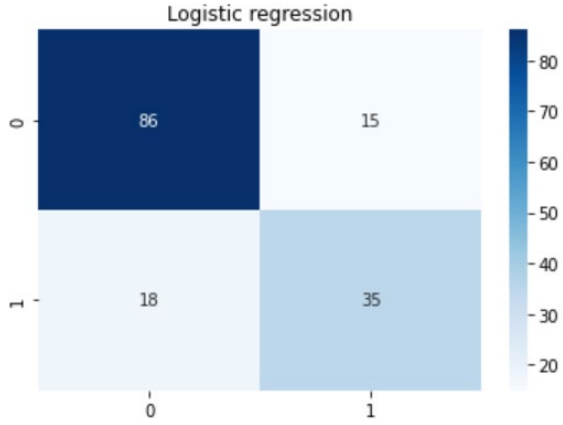


Fig. 3. Logistic Regression confusion matrix

## 6. Conclusion

We demonstrate the effectiveness of machine learning classification algorithms for the classification of diabetes. For this, we examine a number of well-liked classification methods, such as logistic regression, decision trees, and random forests. Three classifiers have also been modified, and their performance has been evaluated in terms of accuracy, F1-score, recall, precision, and AUC.

## References

[1] S. Dewangan et al., "Diabetes diagnosis using machine learning algorithms" Int. Journal of Engineering Research and Application, vol. 8, Issue 1, (Part -II) January 2018, pp. 09-13.

[2] Rahul Joshi et al., "Machine learning and data mining methods in diabetes research" International Research Journal of Engineering and Technology, Volume 4, Issue 10, Oct. 2017.

[3] Desmond Bala Bisandu et. al. "Detection and Prediction using Data mining" International Journal of Research and Innovation in Applied Science, Volume 4, Issue 6, June 2019.

[4] Aditya Saxena et al., "Prediction and Detection using ML," Indian Journal of Artificial Intelligence and Neural Networking, Volume 1, Issue 2, April 2021.

[5] Md. Kamrul Hasan et al., "Diabetes Prediction Using ensembling of Different Machine Learning Classifiers", 2020.

[6] Rahul Joshi et al., "Analysis and prediction of diabetes diseases using machine learning algorithm", Volume 04, Issue 10, Oct. 2017.

[7] M. Young, "The Technical Writer's Handbook," Mill Valley, CA: University Science, 1989.