*Ahmed et al. LGURJCSIT 2023*

*LGU Research Journal of Computer Science & IT*

*SSN: 2521-0122 (Online)*
*ISSN: 2519-7991 (Print)*

*doi: 10.54692/lgurjcsit.2023.0701411*

*Vol. 7 Issue 1, January – March 2023*

# Diabetes Diagnosis through Machine Learning: An Analysis of Classification Algorithms

Haris Ahmed[1*], Dr.Muhammad Affan Alim[1], Dr.Waleej Haider[2], Muhammad Nadeem[2], Ahsan Masroor[1], Nadeem Qamar[1]

[1]College of Computing and Information Sciences, Karachi Institute of Economics and Technology, Karachi, Pakistan.

[2]Department of Computer Science & Information Technology, Sir Syed University of Engineering and Technology, Karachi, Pakistan

Email: harisahmed19@hotmail.com

**ABSTRACT:**

***Diabetes is a serious and chronic disease characterized by high blood sugar levels. If left untreated, it can lead to numerous complications. In the past, diagnosing diabetes required a visit to a diagnostic center and consultation with a doctor. However, machine learning can help identify the disease earlier and more accurately, providing significant benefits for early intervention and treatment. This study aimed to create a model that can accurately predict the likelihood of diabetes in patients using three machine learning classification algorithms: Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB). The model was tested on the Pima Indians Diabetes Database (PIDD) from the UCI machine learning repository and the performance of the algorithms was evaluated using various metrics such as accuracy, precision, F-measure, and recall. The results showed that Logistic Regression had the highest accuracy at 71.39%, outperforming the other algorithms, demonstrating the potential of machine learning in enhancing diabetes diagnosis and management.***

**KEYWORDS**: Logistic Regression, Naive Bayes, Decision Tree, Machine Learning, Logistic Regression, Diabetes.

## 1. INTRODUCTION

Diabetes is a significant health challenge that affects equally developed and underdeveloped countries. It is a chronic, incurable disease characterized by high blood sugar levels due to complications with insulin production [1]. Numerous factors can cause diabetes, including poor diet, toxic substances in food, environmental pollution, infections, unhealthy eating habits, lifestyle changes, and obesity [2]. If not managed properly, diabetes can result in severe consequences. Such as kidney failure, blindness, coma, damage to the pancreas, peripheral vascular diseases, cardiovascular dysfunction, and weight loss [3]. According to estimates, there were 452 million persons with diabetes worldwide in 2017, and this figure is likely to rise to 700 million by 2045. Some research has indicated that the number of people with diabetes could reach half a billion by 2030 and increase by 25% or 51% by 2045 [4]. While there is no permanent cure for diabetes, early detection and treatment can help to prevent complications. Research and medical professionals agree that the chances of recovery are higher if the disease is detected timely [5].

Machine learning algorithms are beneficial for the timely detection of diseases and analysis of diseases using advanced technology [6]. Machine learning (ML), a subfield of artificial intelligence (AI), encompasses various techniques for discovering patterns in data and achieving

specific outcomes [7]. It is classified into four main categories: supervised learning, semi-supervised learning, reinforcement learning, and unsupervised learning.

In order to demonstrate the potential of machine learning in early diagnosis and guiding healthcare professional's decisions in managing diabetes, this study applies and analyses supervised learning approaches, such as Decision Tree, Logistic Regression, and Naive Bayes, for predicting diabetes. In this study, supervised learning techniques such as Decision Tree (DT), Logistic Regression (LR), and Naive Bayes (NB) were used to predict whether a patient has diabetes. The study is organized as follows: we review earlier related work in unit two. Unit three discusses the methodology and algorithms used to produce the model. Unit four covers the evaluation metrics used. Unit five discourses the experimental method and results, the conclusions are presented in the last section.

## 2. RELATED WORK

Previous research on this topic has been extensively reviewed and taken into consideration.

A study published in the Journal of Medical Internet Research used machine learning to predict diabetes risk in a sample of over 2000 individuals. The study used a Random Forest classifier for classification tasks. The classifier was trained on various features, including body mass index (BMI), age, blood-pressure, and blood-glucose level, and could predict diabetes risk [8].

Another study published in the journal PLOS ONE used machine learning to predict the likelihood of developing diabetes in a sample of over 2000 individuals. The research employed a Decision Tree classifier, a machine-learning method used for classification tasks. The classifier was trained with various features, including body mass index (BMI), age, blood pressure, and blood glucose levels, and it was able to predict diabetes risk [9] accurately.

A review published in a reputed journal examined the use of machine learning for diabetes prediction and management. The review found that machine learning algorithms, for example, Support Vector Machines (SVM) and Neural Networks (NN), have been successfully used for diabetes prediction. The review also highlighted the potential for machine learning to be used in the management of diabetes, including for the prediction of complications and the personalized treatment of individuals [10].

## 3. RESEARCH METHODOLOGY

Our approach consists of numerous steps, as shown in Figure 1. Initially, we gather the Pima Indian Diabetes dataset. Then, we performed preprocessing on the PIDD with the aim of developing the predicting model. Next, we use a range of machine learning techniques to train the model. Finally, we use the test database to estimate the performance of these approaches and select the finest classifier for predicting diabetes. We will give additional information about each of these steps in the following sections.

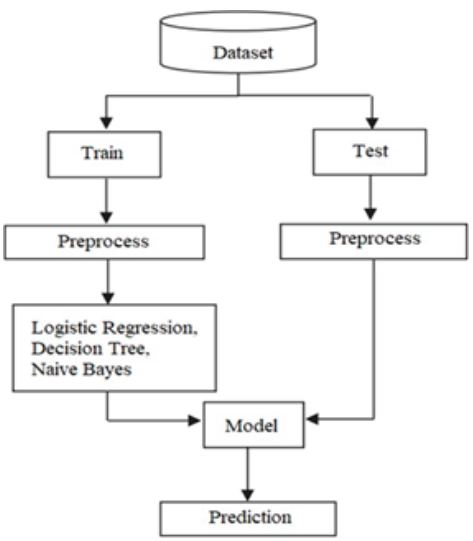


***Figure 1: The diagram of the steps in our model***

### *3.1. Dataset*

Our model aims to predict whether a patient has diabetes or not. We executed it using Python language and several machine learning libraries, including Numpy, Scikitlearn, and pandas. We divided the dataset into a testing set (30%) and a training set (70%) using the PIDD, a medical dataset containing 768 instances and 9 attributes. Of these instances, 268 “class-1” have diabetes, while

500 “class-0” do not [11]. The characteristics include glucose levels, skin thickness, insulin-levels, number of pregnancies, BMI, blood-pressure, diabetes pedigree, age and class (either 0 for not diabetic or 1 for diabetic).

### 3.2. Data Preprocessing

Preprocessing the Data-set can increase the precision of predictions. One way to do this is through data normalization, a great method for enhancing the performance of machine learning models. In our approach, we used Min-Max Scalar as our normalization method. Min-Max Scalar converts the dataset's features by scaling them to a specific range, such as {0 to 1}, using the smallest and largest data values [12]. Furthermore, using Min-Max Scaler normalization facilitates faster convergence of the learning process for our machine learning models. It reduces the possibility of skewed weight updates due to disparate feature scales. The Min-Max Scaler is described by equation (1).

$$y' = \frac{y - \min(y)}{\max(y) - \min(y)} \qquad (1)$$

### 3.3. Machine-Learning Techniques

After arranging the dataset for the model, we employed three popular machine-learning classification methods to predict diabetes mellitus. The following section provides an overview of these algorithms.

#### 3.3.1. Logistic Regression (LR)

It's a statistical model used for binary classification. It is a supervised learning procedure that takes in a set of inputs and forecasts the probability of an event belonging to one of two categories, which are usually represented as 0 and 1. The goal of logistic regression (LR) is to find the finest model that describes the association between the dependent variable (the label or output) and one or more independent variables (the features or inputs). The model is trained using labelled data, and the resulting model is used to make predictions on new, unseen data [13].

In logistic regression, the prediction is made using the following equation:

$$p = \frac{e^{(b_0 + b_1 x)}}{\left(1 + e^{(b_0 + b_1 x)}\right)} \qquad (2)$$

where (p) is the probability of the event occurring, b0 and b1 are coefficients learned during training, and x is the input. The predicted output is then obtained by applying a threshold to the predicted probability. If the probability is greater than the threshold, the predicted output is 1, and if it is less than the threshold, the predicted output is 0.

Logistic regression is a popular technique for binary classification since it is easy to implement and provides fast, reliable results. It is often used in fields such as marketing, finance, and healthcare to predict whether a customer will make a purchase or a patient will have a specific health condition.

#### 3.3.2. Naïve Bayes (NB)

The probabilistic machine learning method uses Baye's theorem to make predictions. It is based on the assumption that the features in the data are not dependent on one anouther. This allows the algorithm to simplify assumptions and improve the efficiency of the learning process [14]. There are several dissimilar types of Naïve Bayes algorithms, including Multinomial Naive Bayes (MNB), Gaussian Naïve Bayes (GNB), and Bernoulli Naive Bayes (BNB). These algorithms differ in the way that they calculate the likelihood of the features given the class, which is a key part of Bayes' theorem

$$P(A \mid B) = \frac{P(A) * P(B \mid A)}{P(B)} \qquad (3)$$

Where (A) is the event (the class), B is the evidence (the features), and P(A), P(B|A), and P(B) are probabilities.

#### 3.3.3. Decision Tree (DT)

A decision tree is a tree-like model used for classification and regression tasks. It makes predictions by learning a hierarchy of if/else questions and subsequent decisions based on the answers to those questions.

A root node at the top of the tree represents the entire dataset. The root node is divided into two or more child nodes based on a decision rule, a threshold value that divides the data into two or more groups. The child nodes are then split into additional child nodes based on additional decision rules, and this process continues until the leaf nodes are reached. The leaf nodes represent the final prediction made by the decision tree [15]. The decision rules used to split the nodes are chosen based on the feature resulting in the most significant impurity decrease. Impurity measures

how mixed the data is at a given node. The most common measure of impurity used in decision trees is the Gini index, which is calculated using the following equation (4):

$$Gini(p) = 1 - sum(p\_i^2) \quad (4)$$

Where p is the probability of a sample belonging to a particular class, and p_i is the probability of a sample belonging to the ith class.
A decision tree is a fast and easy-to-interpret machine learning method that is often used for classification tasks. However, it can be prone to overfitting, especially when the tree grows too deep, and it is not well-suited for datasets with many features or continuous features.

## 4. EVALUATION METRIC

Numerous assessment metrics can be used to calculate the performance of a machine learning model. Here are some popular evaluation metrics, along with the corresponding equations:

### *4.1. Accuracy*

The essential evaluation metric is simply the ratio of correct predictions to total predictions. The equation for accuracy is:

$$Accuracy = \frac{TruePositive + TrueNegatives}{Total\ \Pr edictions} \quad (5)$$

### *4.2. Precision*

Precision is a measure of the accuracy of the model when it predicts a positive outcome [16]. It is the ratio of true- positive forecasts to the total number of positive-predictions made. The equation for precision is:

$$\Pr ecision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (6)$$

### *4.3. Recall*

It measures the model's capability to identify all positive instances in the dataset properly. It is the ratio of true-positive forecasts to the total number of actual positive instances [17]. The equation for the recall is:

$$\Re call = \frac{TruePositives}{TruePositives + FasleNegatives} \quad (7)$$

### *4.4. F1 Score*

this is a mixture of recall and precision and is a single-metric that combines both measures. It is calculated using the harmonic average of precision and recall [16]. The equation for the F1 score is:

$$F1 - Score = \frac{2 * (\Pr ecision * \Re call)}{\Pr ecision + \Re call} \quad (8)$$

### *4.5. Confusion-Matrix:*

The confusion-matrix is a table used to estimate a classification model's performance. It shows the amount of true-negative, true-positive, false-negative and false-positive predictions prepared by the model.

## 5. RESULTS AND DISCUSSION

This study divided the data-set into two subsections: a test set (30%) and a train set (70%). We then used these subsets to evaluate the overall prediction performance of various algorithms, including Logistic Regression. Our results, shown in Table 1, indicate that Logistic Regression has a high precision of 93.79%.
Additionally, among the algorithms tested, Logistic Regression had the highest accuracy. These results were obtained using the Pima Indian Diabetes dataset. The performance of each classification was further visualized using various evaluation measures, as shown in Figures 2-4.

***Table 1: Outcomes for the algorithm on the Diabetes data set***

| Classification Algorithms | Precision | Recall | F1-measure |
|---|---|---|---|
| Logistic Regression | 93.79% | 71.39% | 80.20% |
| Naive Bayes | 80.12% | 66.93% | 71.09% |
| Decision Tree | 69.49% | 70.37% | 69.33% |

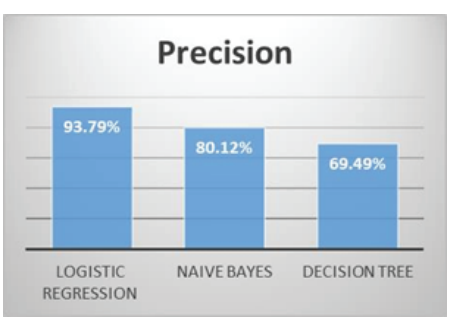


***Figure 2: Precision-measurement***

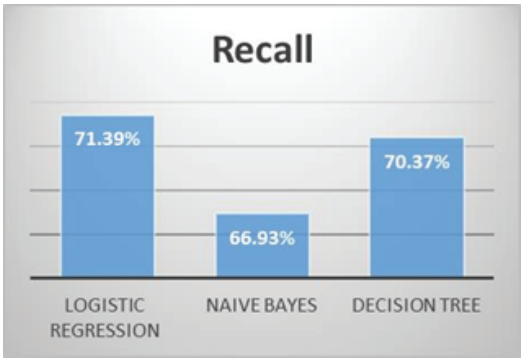


*Figure 3: Recall-measurement*

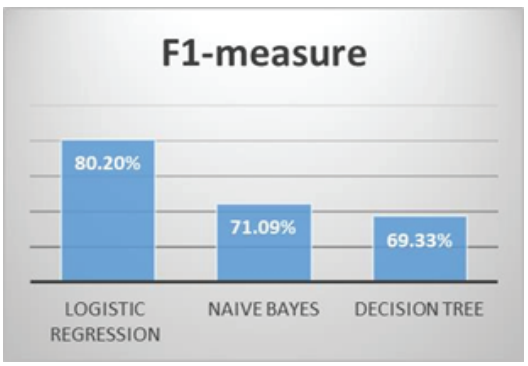


*Figure 4: F1 score-measurement*

## 6. CONCLUSION

Diabetes is a serious medical condition affecting many people worldwide and leading cause of death. Early prediction of diabetes can help to reduce its severity and associated risks. In this study, we proposed a model that can accurately forecast whether a person has diabetes or not. To evaluate the model's effectiveness, we applied three different machine learning classification techniques to the Pima Indian Diabetes dataset and compared their performance using various evaluation metrics. Our results showed that Logistic Regression was the most effective algorithm for predicting diabetes, with an accuracy of 71.39%.

## REFERENCES

[1] Z.Pu P. Katz, Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome, Can. J. diabetes 42, S10–S15, 2018 .

[2] R.Vaishali, R.Sasikala, S.Ramasubbareddy, S.Remya, S.Nalluri. Genetic feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset, in 2017 international conference on computing networking and informatics (ICCNI), pp. 1–5, 2017.

[3] S.K.Dey, A.Hossain, M.M.Rahman. Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm, in 2018 21st international conference of computer and information technology (ICCIT), pp. 1–5, 2018.

[4] N.H.Cho, J.E.Shaw, S.Karuranga, Y. Huang, J.D.da Rocha Fernandes, A.W. Ohlrogge, B. Malanda, IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Res. Clin. Pract. 138, 271–281, 2018.

[5] P. Saeedi et al., Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, Diabetes Res. Clin. Pract. 157, 107843, 2019.

[6] Yahyaoui, A.Jamil, J.Rasheed, M. Yesiltepe. A decision support system for diabetes prediction using machine learning and deep learning techniques, in 2019 1st International Informatics and Software Engineering Conference (UBMYK), pp. 1–4, 2019.

[7] C.E.Sapp, Preparing and architecting for machine learning, Gart. Tech. Prof. Advice, 1–37, 2017.

[8] Ezzati, M.Prediabetes prediction using machine learning: A retrospective cohort study. Journal of Medical Internet Research,20(9), e256, 2018.

[9] Zhang, J.Predictive modeling of diabetes mellitus based on machine learning techniques. PLOS ONE,14(10), e0223779, 2019.

[10] Pereira, T., Machine learning applications in diabetes: A systematic review. Diabetes & Metabolic Syndrome: Clinical Research & Reviews,12(3), 823-832, 2018.

[11] https://www.kaggle.com/uciml/pima-

indians-diabetes-database? select=diabetes.csv.

[12] T.Phaladisailoed, T.Numnonda. Machine learning models comparison for bitcoin price prediction, in 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 506–511, 2018.

[13] C.Xiong, J.Callan. Query expansion with freebase, in Proceedings of the 2015 international conference on the theory of information retrieval, pp. 111–120, 2015.

[14] D.Sisodia, D.S.Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132, 1578–1585, 2018.

[15] Iyer, A., S, J., Sumbaly, R., Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.5101, arXiv:1502.03774, 2015.

[16] S.N.Pasha, D.Ramesh, S.Mohmmad, A. Harshavardhan, Shabana, Cardiovascular disease prediction using deep learning techniques, IOP Conf. Ser. Mater. Sci. Eng. 981, 022006, https://-doi.org/10.1088/1757- 899X/ 981/2/022006, 2020.

[17] M.Toepfer, C.Seifert. Content-based quality estimation for automatic subject indexing of short texts under precision and recall constraints, in International Conference on Theory and Practice of Digital Libraries, pp. 3–15, 2018.