JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org Vol.3, Iss.1 Pages:09-15

# DIABETIC PREDICTION USING MACHINE LEARNING ALGORITHMS

**G Amirthavalli[1], V Amala Deepa[2]**

[1] M.Sc Computer Science, [2]Assistant Professor

Holy Cross College, Trichy, Tamilnadu, India

**Article Received: January 2022 Published: April 2022**

## Abstract

Diabetes is a disease that is caused by a high glucose level in the human body. It raises the chance of acquiring other diseases such as heart disease, renal illness, blood vessel damage, nerve damage, and blindness, and diabetes patients are more susceptible to become infected with the corona virus, as that is widely known. Patients with diabetes mellitus have poor immune-system that cannot fight back the spread of the virus thus aiding the transmission of infection faster. Diabetes can be treated if it is detected early enough. To attain this goal, we will use few machine learning techniques to do early diabetes prediction in a human. By developing models using Machine Learning Techniques with patient datasets, we can improve our prediction results. Using Machine Learning techniques like Decision Tree, Logistic Regression, Random Forest, we have predicted the diabetes and provided a comparative analysis of these machine learning used for the experiment prediction. A comparative analysis on these algorithms were made and the results showed that the accuracy of RF algorithm is higher. The objective of this paper is completely related to the prediction of diabetic disease via machine learning.

***Keywords: Diabetes, Machine, Learning, Prediction, Dataset, Supervised machine learning***

## INTRODUCTION

People's lifestyles are too hectic nowadays, the majority of them do not prioritise their health or know how to protect it. It has the potential to cause a lot of metabolic disorders; For example, diabetes mellitus is one of the diseases that is directly linked to our lifestyle. If it goes unnoticed, it can be a fatal disease. Obesity, high blood glucose levels, and other factors led to diabetes. It has an effect on the hormone insulin, causing abnormal metabolism and increasing blood sugar levels. When the body produces insufficient insulin, diabetes will occur. As per the World Health Organization, 422 million people worldwide are having diabetes and the majority of these people living in low- or middle-income countries. And over the next 8 years this could be increased to 490 billion. Diabetes is the world's leading cause of death. If diabetes is diagnosed early, it can be controlled and a human life will be saved. To do so, this research looks into diabetes prediction using a variety of diabetes-related attributes.

We used the Pima Indian Diabetes Dataset for this, and we predict diabetes using a variety of Machine Learning Algorithms. The term "machine learning" refers to a technique for training the computers or machines. By developing a model from acquired datasets, various Machine Learning Techniques that offers efficient results for gathering knowledge.

## RELATED WORK

K.VijiyaKumar et al. [2] proposed the Random Forest algorithm for diabetes prediction to design a system that can do early diabetes prediction for a patient with a better accuracy using machine learning techniques. The proposed method produces the best diabetic prediction results, demonstrating that the prediction system is capable of accurately, efficiently, and most importantly, instantaneously predicting diabetes disease.

Predicting diabetes onset: an ensemble supervised learning method was described by Nonso Nnamoko et al. [3]. The ensembles are built using five widely used classifiers, with the outputs aggregated using a meta-classifier. The findings are presented and compared to those of other studies in the literature that used the same dataset. It is demonstrated that diabetes onset prediction can be done with more accuracy using the proposed method.

N. Joshi et al. [4] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease.

Sarwar et al. [5] gave a comparative study for prediction of diabetes mellitus using various machine algorithms and also discuss various statistical measures during this study. According to this if the dataset is large in size and more balance then accuracy is improved. For predictive analysis, five classifiers used the name as logistic regression (LR), k-Nearest Neighbour (KNN), Support Vector Machine (SVM), random forest (RF), and Decision Tree(DT).

Sisodia et al. [6] discuss predicting the diabetes disease using three classifiers name as such as Naïve Bayes (NB), Support vector machine (SVM), and Decision tree (DT). An experiment is performed on the Pima Indian Diabetes Database (PIDD). The performance metric is measured in the term of Precision, Accuracy, Recall, and F-Score. Results obtained show Naïve Bayes outperformed among the three algorithms with 76.30 % Accuracy.

Wang et al. [7] discuss the predictive analysis of diabetes mellitus considering the role of imbalanced data with missing values. In their experiment, they used Naïve Bayes for data normalization by compensating the missing values. For addressing class imbalance problem oversampling with the ADASYN algorithm is used. Finally, the Random Forest (RF) is used for prediction. An experiment is performed Pima Indian Diabetes Database and performance is checked by using the combined approach of these classifiers than they individually work to improve the results.

## PROPOSED SYSTEM

Melanoma is the most dangerous type of skin cancer. Dermoscopy based early detection and recognition strategy is critical for melanoma therapy. So proposed system to extract the features based on color and texture. Segment the skin lesions using active contour based snake model. Finally classify the cancer using convolutional neural network algorithm with improved accuracy rate.

Advantages include the ability to extract all features, decreased dimensionality, improved classification accuracy, and automatic segmentation.

## METHODOLGY

This approach was implemented using the Microsoft Windows 10 operating system. This includes an Intel(R) Core (TM) i3-1005G1 CPU running at 1.20GHz and 8GB of RAM.

A. Dataset and attributes

The National Institute of Diabetes and Digestive and Kidney Diseases provided the dataset and that is used in this study. The dataset's objective is to diagnose whether a patient has diabetes by certain diagnostic measurements included in the dataset. All of the patients in this dataset, are Pima Indian women over the age of 21. This dataset comprises 768 Pima Indian women's patient records with 9 attributes. Table1 displays the dataset's attributes. The fundamental data statistics are listed in Table2. The datasets include various medical predictor variables as well as one outcome variable. The number of pregnancies the patient had, their BMI, insulin level, age, and other variables are all the predictor variables.

B. Data Pre-processing

Data pre-processing is most important cycle. In general, medical care-related information comprises missing value and various pollutions that can affect the information viability. Data cleaning refers to identifying incomplete, incorrect, inaccurate, or tangential sections of the knowledge and substituting, changing, or eliminating the filthy or coarse data from a record

set, table, or data [1]. After cleaning the data, the information is regularized in order to train and test the model.

C. Model Building

In this study, we have carried out various ML algorithms for predicting the diabetes. First, import the necessary libraries and the diabetes dataset. Data was examined for missing values using data analysis, and it was found that the number of instances with zero values was extremely high, so it was replaced with mean values. The data was then split into two groups, training data and testing data. Then, using the training data, a machine learning model was trained to make predictions. After the model has been trained using training data, testing data was used to predict outcomes and verify accuracy, and the model was finally evaluated Figure 1 shows the model-building process. This process was followed for all 3 machine learning algorithms used in this paper. Experiments have been performed and results were obtained.

Table 1: Data Description

| Predictor Variables | Description |
|---|---|
| Pregnancies | Number of Pregnancies. |
| Glucose | Plasma glucose concentration. |
| blood_pressure | Diastolic Blood Pressure (mm Hg). |
| skin_thickness | Triceps skin fold thickness (mm). |
| Insulin | 2-Hour serum insulin (mu U/ml). |
| bmi | Body Mass Index. |
| diabetes_pedigree_function | Diabetes Pedigree Function. |
| Age | Age. |
| Outcome | Diabetes or no diabetes (1/0). |

Table 2: Data Statistics

| | Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age |
|---|---|---|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | 0.471876 | 33.240885 |
| **std** | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 |
| **max** | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 |

D. Algorithms

Logistic Regression

Under the Supervised Learning method, one of the most prominent Machine Learning algorithms is Logistic Regression. It's a method to predict a dependent variable from independent variables [10]. From the Linear Regression equation, the Logistic Regression equation can be obtained. Hereunder are the mathematical steps to obtain Logistic Regression equations:

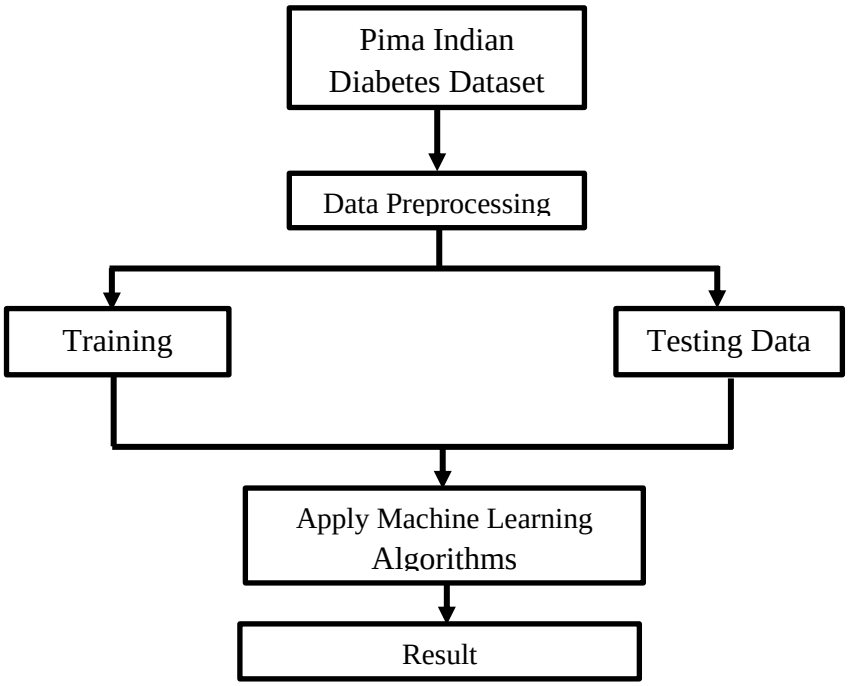


Figure 1: Machine Learning Model Building Process

- We know the equation of the straight line can be written as:
- y in Logistic Regression can only be between 0 and 1, Lets divide the previous equation by (1-y):
- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n$$

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n$$

Decision Tree

A simple classification approach is the decision tree. DT is supervised learning method. When the response variable is categorical, a decision tree is used. A decision tree is a model with a tree-like structure that depicts classification processes depending on input features. Input variables can be of any sort, including graphs, text, discrete values, and continuous values. Steps for Decision Tree Algorithm are followed below

• Create a tree using nodes as the input feature.

• Choose the feature with the best information gain to predict the output from the input feature.

• For each attribute in each tree node, the highest information gain is determined.

• Repeat step 2 to create a subtree utilising the feature not used in the previous node.

Random Forest

Leo Bremen is the developer of Random Forest. Random Forest rule is a supervised classification rule [8]. Aside RF assesses the significance of each attribute and selects the most critical predictor from a wide number of predictors [9]. The pseudo code for Random Forest is a. The first step is to pick the "R" features from a total “m” feature, where R<<m.

b. The node using the best split point out of all the "R" attributes.

c. Using the best split, split the node into child nodes.

d. Repeat steps a to c until the "l" number of nodes is attained.

e. Create Forest by repeating steps a to d for "n" number of times to obtain "n" number of trees.

## RESULTS AND DISCUSSION

In this paper, three machine learning algorithms such as Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) have applied on PIMA Indian diabetes dataset. Data was divided into two portions, training data and testing data, both these portions consisting 75% and 25% data respectively. All these three algorithms were applied on same and results were obtained. The key evaluation criterion we used in this study was prediction accuracy. The accuracy of algorithms was measured and shown in Figure 2. Logistic Regression gives 74% accuracy, 67% accuracy was achieved by using DT and RF achieved highest accuracy which is 80%. From the experimental results obtained, it can be concluded that the RF algorithm is appropriated for predicting the diabetes status of patients.

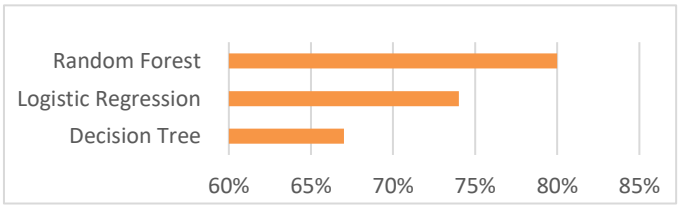


Figure 2: Comparison of RF algorithm with other two algorithms

## CONCLUSION

Predictive analytics in healthcare has the potential to transform the way medical researchers and practitioners analyse data and make choices. We employed three common machine learning algorithms for predictive analytics in this work. These algorithms include Logistic regression, Decision tree and Random Forest. Diabetes predictions were developed using the PIMA Indian dataset, which consist of 768 records. The prediction model was trained and tested using 8 attributes. It is clear from the experimental results that RF has the highest accuracy for predicting diabetes. RF algorithm provide 80% percent accuracy, which is the highest of the two methods employed in this paper. Therefore, it can be concluded that RF is appropriated for predicting the diabetes. As a future work a hybrid model (mixture of Decision Tree and Random Forest) can be designed to increase the prediction accuracy using hospital data.

## REFERENCES


1. D. Menon, K. Schwab, D.W. Wright, A.I. Maas, and the Demographics and Clinical Assessment Working Group of the International and Interagency Initiative toward Common Data Elements for Research on Traumatic Brain Injury and Psychological Health, Position statement: definition of traumatic brain injury, Arch. Phys. Med. Rehabil., vol. 91, pp. 1637– 40, Nov 2010.
2. 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.
3. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ".IEEE Congress on Evolutionary Computation (CEC), 2018.
4. 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
5. M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, 2018.," Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in: 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, United Kingdom, pp. 1-6, 2018.
6. Sisodia, D., Sisodia, D.S., "Prediction of Diabetes using Classification Algorithms," in: International Conference on Computational Intelligence and Data Science (ICCIDS 2018), ELSEVIER. Procedia Computer Science, ISSN 1877-0509, vol 132.
7. Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data with Missing Values," in IEEE Access, vol. 7, pp. 102232-102238, 2019.
8. Consistency Of Random Forests, By ErwanScornet Sorbonne University, UPMC Paris 06, F-75005, Paris, France, By Gerard Biau ´ Sorbonne Universities, UPMC Univ Paris 06, F-75005, Paris, France


**Cite this article:**

G Amirthavalli, V Amala Deepa, "Diabetic Prediction using Machine Learning Algorithms", Journal of Multidimensional Research and Review (JMRR), Vol.2, Iss.4, pp.09-15, 2022