International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

Machine Learning Approaches for Diabetes Classification

Rajesh Saxena

Assistant Professor, Department of Computer Science
Disha Bharti College of Management and Education, Saharanpur, India

---

Abstract - *In the last few years, Machine Learning and Artificial Intelligence are being used to upgrade our healthcare system and treat serious diseases. In today's world, diabetes is a major threat to our health and it is a major cause of death in the world. In this materialistic age, people are very busy in earning money so that they can fulfill their daily needs. But due to this busy lifestyle, they are not able to pay full attention to their health and for this reason the number of people suffering from various diseases is increasing continuously. As we know that the field of Artificial Intelligence (AI) is growing very fast and touching every aspect of our life. Machine Learning (ML) is a part of AI and it is possible that using machine learning techniques, we can create high-quality models with the help of which diabetes and its side effects can be predicted. In this paper we experimentally analyze the adoption of machine learning in diabetes care to examine how it can improve the accuracy of diagnosis and make life easier for patients and doctors.*

Key Words: Diabetes, Machine Learning, Random Forest, Multilayer Perceptron, KNN

1. INTRODUCTION

According to IDF Diabetes Atlas Tenth edition 2021 [1], approximately 537 million people worldwide are suffering from diabetes. It is believed that by 2030 this number will increase to 643 million and 783 million by 2045. In our country, in the year 2021, the estimated number of diabetic patients in the age group of 20 to79 years was 74.2 million and by 2025 this number is likely to increase to 124.8 million. Which means that, in 2021, one out of every 17 people in India suffers from diabetes and this number is increasing rapidly. Diabetes is a chronic disease, mainly due to the imbalance of our body's metabolism. When we eat carbohydrates, such substances become dextrose after digestion and they become glucose after being absorbed by the small intestine. Pancreas secretes insulin and in the body of a healthy person, glucose is converted into energy by insulin. People suffering from diabetes do not have enough insulin in their body or the insulin made by their body does not work completely. Therefore, the cells of the body do not absorb sugar and the amount of sugar in the blood increases. When level of glucose increase in the body, the metabolism of fats and proteins is disturbed and in this condition the chances of destruction of the body's systems and organs increase. There are three types of diabetes whose names are Type 1, Type 2 and Gestational. Type 1 diabetes, also called insulin dependent diabetes, begin due to the lack of complete production of insulin by the pancreas in the body. In this type of diabetes, patients need insulin to keep their disease under control and survive, and if they stop taking insulin, they can be at risk of death. Type 2 diabetes is also called diabetes in older people. In this type of diabetes, insulin is produced inside the body, but its quantity is not enough to meet the need of the body and the cells of the body are not able to use sugar as a source of energy. Due to this, the level of sugar in the body increases. During pregnancy, nutrients and water reach the unborn baby through the placenta. The placenta also produces a variety of hormones necessary for pregnancy, some of these hormones (estrogen, cortisol) inhibiting insulin. As the placenta grows, more of these hormones are produced, and the risk of insulin resistance also increases. Normally, the pancreas makes extra insulin to overcome insulin resistance, but when insulin production begins to decrease compared to placental hormones, Gestational diabetes takes place.

This paper is organized as follows: we have described in detail the work and literature related to our topic in section 2, we have described the proposed research and the machine learning algorithms used in this research in section 3, we discuss about implementation in section 4 and discusses the results from the performed experiments and future work in section 5.

2. LITERATURE REVIEW AND RELATED WORKS

The main objective of literature review is to get information about the work done in our domain in the past and on the basis of that, get an idea for our study. This increases our knowledge related to that domain and helps us to improve our project further. In this section we will try to know about the research work done in Healthcare and their assessment using Machine Learning and Data Mining techniques in the past. Minyechil, A., & Rahul, J. [2] they studied various techniques of data mining and various machine learning algorithms were applied in different medical datasets. They found that single algorithm provided less accuracy than group of algorithms. They found that decision tree algorithm

provided high accuracy. In this study, they used Weka and java as tools to predict diabetes dataset. Kumari Sonu and Archana Singh [3] proposed an intelligent and effective methodology for the automated detection of diabetes using neural network technique. A. Islam and N. Jahan [4], used the Pima Indian data set to study and predict the risk level of diabetes. They used many machine learning algorithms to classify the data and studied the outcome obtained from them. G Bansal and Singla [5] presented a hybrid method to detect diabetes. They selected the Pima Indian Diabetes Dataset to do their study. They used non-linear Support Vector Machine with partial least square method. Classifiers used for the model are the neural network, SVM, and DT. The accuracy rate of their method was 84.5%. Quan Zou et al [6] proposed a model for diabetes prediction using three different classifiers and the PIMA Indian Diabetes dataset [7]. Both the WEKA and MATLAB platforms were used for their study. Random Forest (RF) and Decision Tree (DT) were implemented in WEKA and Neural Network was implemented in MATLAB. A maximum accuracy of 80.84% was achieved using RF.

3. METHODOLOGY

Machine Learning (ML) represents computer algorithms which can observe and analyze data on their own. That means, first of all we should collect data related to a particular domain. Then we have to develop and algorithm through which we can represent the relationship between various elements in the dataset. This algorithm is called machine learning model. This model works like brain for the computer to understanding the data. We split the data generally into two parts: the train data and the test data. The train data is used to provide training the model. After the training, the model will be able to predict the future behavior with new data, called test data. The good part of Machine Learning is that many models were already developed by Computer Scientists using different types of logic. These model can be used without the need of redeveloping them.

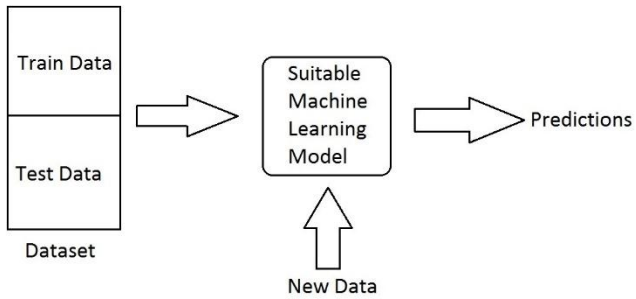

Fig - 1: How Machine Learning Models works

For this study, I use three classification model of ML. These models are:

3.1 Random Forest

Random Forest is an algorithm that operates based on several Decision Tree (DT). It uses multiple Decision Tree during training phase and the final result is decided based on the decision given by majority of Decision Tree. Random Forest is generally used in classification problems where we have to classify the digit written by a human being into the available 10 digits (0 to 9) or classify a patient into a diabetic or non-diabetic. The advantage of Random Forest is to reduce overfitting problem that generally occur in many Machine Learning models.

3.2 KNN classifier

K Nearest Neighbor (KNN) is an algorithm that classifies a new data point based on its nearest neighbors. A group of neighbors are selected (this is k value) and the data point is classified under that class to which majority of neighbors belong. Here k represents the number of selected nearest neighbors. Choosing k value properly will give more accuracy. Choosing right value for is called parameter tuning.

3.3 Multilayer Perceptron

Multi-layer Perception is also known as MLP. These are fully connected dense layers, which transform any input dimension to the desired dimension. The multi-layer perceptron defines the most complex architecture of artificial neural networks. It is largely made up of multiple layers of perceptron. Every node in the multi-layer perception uses a sigmoid function. The sigmoid function takes real values as input and converts this input to numbers between 0 and 1 with help of the sigmoid formula. Sigmoid function is written as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

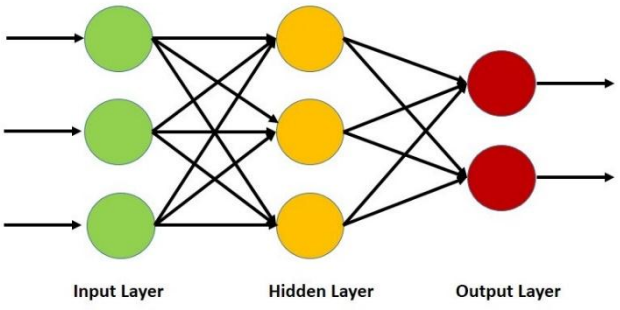

Fig - 2: A schematic diagram of a Multi-Layer Perceptron

This diabetes prediction system consists of two main steps and both steps work together to achieve the desired

results. The first step of this system is data preparation and second step is the classification.

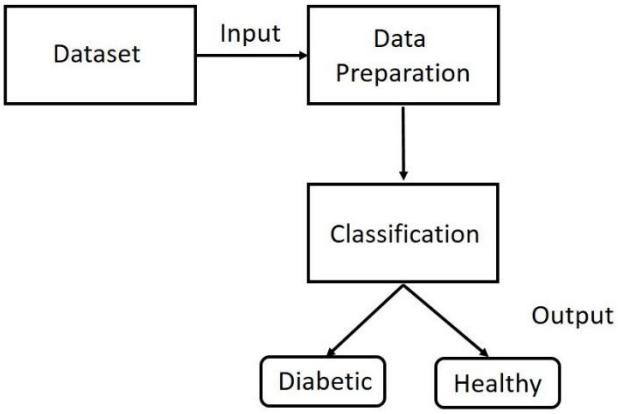

Fig - 3: Main stages of proposed system

4. EXPERIMENT DESIGN

In this study, Pima Indian Diabetes Dataset (PID) will be use. The Pima Indian Diabetes Dataset contains 768 instances. This dataset consists of one target variable (Outcome) and the 8 attributes: Pregnancies, OGTT (Oral Glucose Tolerance Test), Blood Pressure, Skin Thickness, Insulin, BMI (Body Mass Index), Age, Diabetes Pedigree Function. Each attribute is explained in Table 1. The manipulation of this dataset has been done in Jupyter Notebook with the help of Pandas library. First of all, loading the dataset with the pandas **read_csv()** method. Then use **isnull().sum()** method to check Null values in the dataset. Now check zero value in each column and after applying pandas library methods, we found 5 zero value in Glucose, 35 zero values in BloodPressure, 227 zero value in SkinThickness, 374 zero values in Insulin, 11 zero value in BMI. With the help **replace()** method of pandas, we will replace zero value with mean of that particular column. After completing data preprocessing, separate the dataset into training (70% of the data) and test datasets (30% of the data).

Table 1: Attributes

S. No.	Name of Attribute	Meaning
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2 – Hour serum insulin (mu U/ml)
6	BMI	Body Mass Index (kg/m^2)
7	DiabetesPedigreeFuction	Diabetes pedigree function
8	Age	Age (years)
9	Outcome	Class variable (0 or 1), 268 of 768 are 1, 500 of 768 are 0

We can use **train_test_split()** for this. Scikit-Learn library provides **fit()** and **predict()** methods for performing training and prediction, respectively. Now import the particular Classifier class, an instance of the this particular class was created and **fit()** and **predict()** methods were executed. After this, we again trained the machine using cross validation approach. We considered 5, 10, 15, 20 fold for training the machine. Now we will evaluate the performance of these classifiers. For this, find the value of TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). Now calculate Accuracy, Sensitivity, Specificity, Precision, F1 score, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Accuracy shows the number of correct predictions of all predictions. Sensitivity shows the number of correctly predicted patients with diabetes. Specificity shows the number of correct predictions for non-diabetics. Precision shows ratio of accurate positive observations. F1 score is used to measure the accuracy of a Machine Learning Model. The MSE is a measure of how close a fitted line is to data points and RMSE is just the square root of the mean square error. The model with highest Accuracy, Sensitivity and Specificity is the best machine learning predictive model.

5. RESULT ANALYSIS

A confusion matrix is a table of data that summarizes the performance of a Machine Learning model. It helps to know where the model is performing well and where it is failing. Generally, confusion matrix is created for the Classification models to know their performance on test data. For the two prediction classes of classifiers, the matrix is of 2x2 table and for 3 classes, it is 3x3 table. The confusion matrix is divided into two dimensions that are predicted values and actual values along with the total number of predictions. Basis structure of 2x2 confusion matrix:

TN (True Negative)	FP (False Positive)
FN (False Negative)	TP (True Positive)

TN (True Negative) = Model has given prediction No, and the actual value was also No.

TP (True Positive) = The model has predicted Yes, and the actual value was also Yes.

FN (False Negative) = The model has predicted No, but the actual value was Yes.

FP (False Positive) = The model has predicted Yes, but the actual value was No.

In this study, first we train the machine by splitting our dataset.

Total records = 768

Training dataset = 70% of 768 = 538

Testing dataset = 30% of 768 = 230

Confusion Matrix for Multi-Layer Perceptron

121	37
27	45

Confusion Matrix for Random Forest

140	18
31	41

Confusion Matrix for Random Forest

122	36
30	42

With the help of these confusion matrix we calculate the accuracy of used models.

Accuracy of Multi-Layer Perceptron = 72.17%

Accuracy of Random Forest = 78.70%

Accuracy of KNN = 71.30%

Now we trained the model using cross validation method and check the accuracy of the model.

Table 2: Performance Measurement

Cross Validation	Measure - ment Matrix	Multi-Layer Perceptron	Random Forest	KNN
5	TP Rate	0.759	0.757	0.689
	FP Rate	0.295	0.305	0.378
	Precision	0.758	0.754	0.687
	ROC Area	0.808	0.821	0.653
	Accuracy	**75.91 %**	**75.65%**	**68.80%**
10	TP Rate	0.753	0.763	0.677
	FP Rate	0.304	0.295	0.386
	Precision	0.751	0.761	0.678
	ROC Area	0.798	0.819	0.640
	Accuracy	**75.26%**	**76.30%**	**67.70%**
15	TP Rate	0.763	0.762	0.674
	FP Rate	0.297	0.294	0.386
	Precision	0.760	0.760	0.676
	ROC Area	0.809	0.824	0.638
	Accuracy	**76.30%**	**76.17%**	**67.44%**
20	TP Rate	0.733	0.753	0.682
	FP Rate	0.314	0.313	0.378
	Precision	0.735	0.749	0.683
	ROC Area	0.793	0.824	0.650
	Accuracy	**73.30%**	**75.26%**	**68.22%**

In this study by applying three machine algorithms we develop three model. After comparing these model on the basis of different parameters, we can say that Random Forest model is the best model.

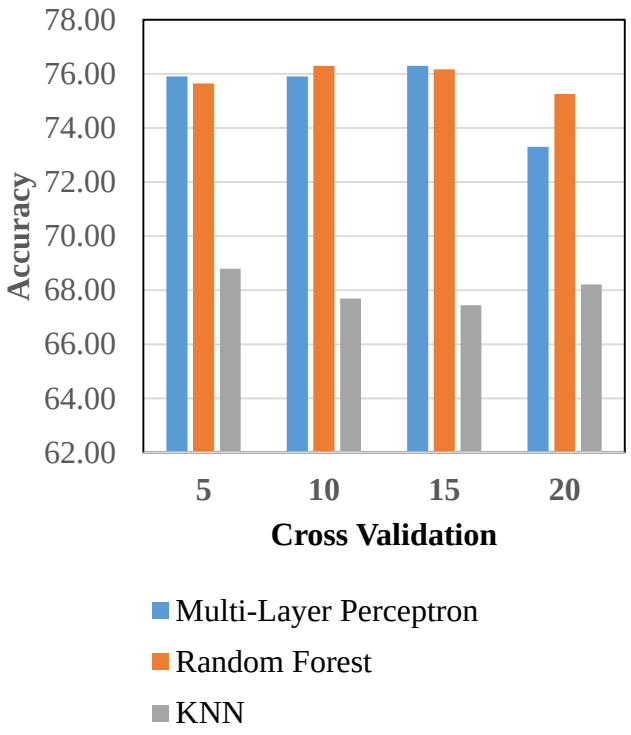

Fig - 4: Performance Measurement

6. CONCLUSIONS

In this research paper, we have applied commonly used machine learning techniques and datasets to predict the diabetes disease in a patient. In this study, we checked the performance and accuracy of each of the algorithms used on the basis of various parameters and tried to find the best algorithm out of the three. In this study, we found that out of the three algorithms used, Random Forest is the best compared to the other for the classification of diabetes dataset. The experimental results may aid in better clinical decision-making for early health care prevention and control of diabetes.

REFERENCES

[1] IDF Diabetes Atlas Tenth edition 2021, https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[2] Minyechil, A., & Rahul, J. (2017). Analysis and Prediction of Diabetes Diseases Using Machine Learning Algorithm: Ensemble Approach. Volume: 04, Issue 10. Retrieved from ww.irjet.net/archives/V4/i10/IRJET-V4I1077.pdf

[3] Kumari Sonu and Archana Singh, A data mining approach for the diagnosis of diabetes mellitus, 2013 7th International Conference on Intelligent Systems and Control (ISCO), DOI: 10.1109/ISCO.2013.6481182, https://ieeexplore.ieee.org/document/6481182

[4] Aminu, I., & Nusrat, J. (2017). Prediction of Onset Diabetes Using Machine Learning Techniques. International Journal of Computer Applications, Volume 180 – No.5. pdfs.semanticscholar.org/2c3a/6609a76762e3d40bdd90d8c07fa714d611fa.pdf

[5] G. Bansal and M. Singla, "Ensembling of non-linear SVM models with partial least square for diabetes prediction," Lecture Notes in Electrical Engineering, Vol. 569, pp. 731–739, 2020. https://assets.researchsquare.com/files/rs-1572946/v1/3c5e9981-c99a-43ab-aac7-25a6d347cfa2.pdf?c=1650486875

[6] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Prediction of diabetes mellitus with machine learning techniques," Frontiers in Genetics, Vol. 9, no. November, pp. 1-10, 2018.

[7] Pima Indians Diabetes Database, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[8] John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.

[9] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).