Detection of Diabetes Using Various Machine Learning Algorithm

Rishav Karanjit
Undergraduate Student, Siddaganga Institute of Technology

Abstract **- Lot of the people around the globe are suffering from diabetics. Diabetes is caused by very high blood glucose. There are different tests available to test diabetes. All of these tests directly or indirectly require assistance from medical personnel. Machine learning can help individuals to detect diabetes without the need of medical personnel. In this paper, we have proposed the solution of detecting diabetes using Decision Tree, K-Nearest Neighbour, Random Forest and soft voting classifier which achieved the accuracy of 80.52%, 75.97%, 81.17%, 83.12% respectively.**

Index Terms **- Diabetes, Machine learning, Decision Tree, K-Nearest Neighbour, Random Forest, soft voting classifier.**

I.INTRODUCTION

Diabetes mellitus or simply Diabetes is a disease that occurs when blood glucose in the body is too high. In most of the cases, the cause of diabetes is due to pancreas not producing enough insulin or cells of the body not responding properly to the insulin produced. The most common symptoms in Diabetes are frequent urination, increased thirst and increased appetite. If diabetes is left untreated, it can cause many health complications. Some of the complication includes diabetic stroke, cardiovascular disease, foot ulcers, ketoacidosis, damage to the nerves, hyperosmolar hyperglycaemic state, chronic kidney disease, damage to the eyes, cognitive impairment or even death. There are three main types of Diabetes mellitus. They are:

- Type 1 diabetes: It is also known as juvenile diabetes or insulin-dependent diabetes. This is a type of diabetes which produces little or no insulin. In this diabetes, pancreas fails to produce enough insulin due to loss in beta cells.
- Type 2 diabetes: In this type of diabetes body is unable to use the insulin available in this body. The most common cause of this diabetes is excessive body weight and insufficient exercise.
- Gestational diabetes: This is the third form of diabetes which occurs in pregnant women without previous history of diabetes. For most women, this type of diabetes doesn't cause noticeable signs or symptoms.

According to International Diabetics Federations (IDF), there are 464 people living with diabetics and diabetics cause 4.2 million deaths. 1 in 5 of the people who are above 65 years old have diabetes. 1 in 2 people with diabetes were undiagnosed. Diabetes caused at least USD 769 billion dollars in health expenditure in 2019. More than 1.1 million children and adolescents are living with type 1 diabetes. More than 20 million live births (1 in 6 live births) are affected by diabetes. 374 million people are at increased risk of developing type 2 diabetes. [1]

According to WebMD [3], there are different test available for diagnosis of diabetes:

- Fasting plasma glucose test: This test measures the blood glucose after the patient have gone at least 8 hours without eating.
- Oral glucose tolerance test: This test measures the patient's blood sugar after the patient has gone at least eight hours without eating and two hours after you drink a glucose-containing beverage.
- Random plasma glucose test: In this test, blood sugar is checked without regard to when the patient ate his/her last meal.

There are different methods to check glucose level of our body but we also need to go to doctor for the diagnosis of diabetics. This paper will help to predict diabetes so that everyone can diagnosis diabetics without a doctor.

II. RELATED WORK

M. Deepika and Dr. K. Kalaiselvi [4] presents data mining solution for disease diagnosis. In this paper,

they had given solution for different diseases. For Diabetes diagnosis, they used artificial neural network (ANN), logistic regression and decision tree for which the accuracies achieved were 73.23, 76.13 and 77.87 respectively.

In [5] Muhammad Azeem Sarwar et. al proposed a 6-algorithm solution for the prediction of Diabetes. The 6-algorithm used were Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), K Nearest Network (KNN). LR achieved 74%, SVM achieved 77%, NB achieved 74%, DT and RF achieved 71%, KNN achieved 77% accuracy. In this paper, SVM and KNN achieved highest accuracy. Pima Indian Diabetes dataset from UCI machine learning repository was used in this paper.

Ridam Pal et. al [6] proposed a research paper titled "Application of Machine Learning Algorithms on Diabetic Retinopathy" where Naive Bayes Classifier, Decision Tree, K-Nearest Neighbors and Support Vector Machine algorithm were used. These algorithms were implemented using Python and WEKA simulator. In Python, Naive Bayes Classifier achieved 65.97%, Decision Tree achieved 65.45%, K-Nearest Neighbors achieved 67.71% and Support Vector Machine algorithm achieved 74.65%. In WEKA, Naive Bayes Classifier achieved 56.64%, Decision Tree achieved 63.51%, K-Nearest Neighbors achieved 60.03% and Support Vector Machine algorithm achieved 67.85%.

Priyanka Sonar and Prof. K. JayaMalini has discussed algorithm like Decision Tree, ANN, Naive Bayes and SVM algorithms. Among these algorithms, Decision Tree performed best. Dataset was taken from UCI machine learning repository. [7]

P. Suresh Kumar and V. Umatejaswi used Naive Bayes, Random Tree, C4.5 and simple logistic classifier. With these algorithms, type 1, type 2 and type 0 diabetes were individually predicted. In the dataset, discretize filter was used for obtaining good intervals of data which eliminated all invalid and null data from the dataset. [8]

Mamta Arora and Mrinal Pandey in [9] proposed a deep learning solution to detect diabetes. The deep learning algorithm automatically identifies the pattern and classifies the retina image into one of the five class based.

III. DATASET AND DATA PREPROCESSING

A. *Dataset*

In this research paper, dataset from UCI machine learning repository is used. The dataset consists of only female patient aged at least 21 years. The dataset consists the record of 768 patients with 9 attribute each. The attributes in this dataset are:

- No of pregnancies
- Glucose concentration
- Diastolic Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age
- Outcome

Table 1 contains the description of each of the attributes.

TABLE 1. ATTRIBUTE DESCRIPTION

Attribute	Description
No of pregnancies	Number of times patient had been pregnant
Glucose concentration	Plasma glucose concentration level
Diastolic Blood Pressure	Diastolic blood pressure in mm Hg
Skin Thickness	Triceps skin fold thickness in mm
Insulin	2-Hour serum insulin in mu U/ml
BMI	Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function	Diabetes pedigree function
Age	Age of the patient in years
Outcome	Diabetics or not

B. *Data Pre-processing*

In the dataset, we had to do some pre-processing before sending it to our machine learning model. The dataset contained many missing values. For example, in some records of patient the value of Blood Pressure was 0 which is not possible. So, we replaced these missing values with the mean of the attribute of all records. Table 2 describes the data statistics after pre-processing of data.

TABLE 2. DATA STATISTICS

Attribute	Count	Mean	STD	Min	Max
No of pregnancies	768	3.84	3.36	0	17
Glucose concentration	768	121.68	30.43	44	199

Diastolic Blood Pressure	768	72.25	12.11	24	122
Skin Thickness	768	26.60	9.63	7	99
Insulin	768	118.66	93.08	14	846
BMI	768	32.45	6.87	18.20	67.10
Diabetes Pedigree Function	768	0.47	0.33	0.078	2.42
Age	768	33.24	11.76	21	81
Outcome	768	0.34	0.47	0	1

IV. EXPERIMENTAL RESULTS

For the prediction of Diabetes Machine learning is used. Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data.[1] In this paper, we have used different supervised machine learning algorithm for the prediction of Diabetics. The algorithm we have used are Decision Tree, Random Forest, K Nearest Neighbors (KNN) algorithm and lastly, we have used a soft voting classifier using Decision Tree, Random Forest and KNN.

Figure 1 illustrates our proposed system for Diabetics prediction. First of all, the dataset was obtained from UCI repository. Then, data pre-processing was performed which is described in section III. After data pre-processing, feature section was done where 8 features were selected. Furthermore, the dataset was divided into train set and test set which consists of 80% and 20% of the dataset respectively. The training data is used to train the model along with the machine learning algorithm. Finally, this model performs prediction on the test set for the result analysis of the model which was created.

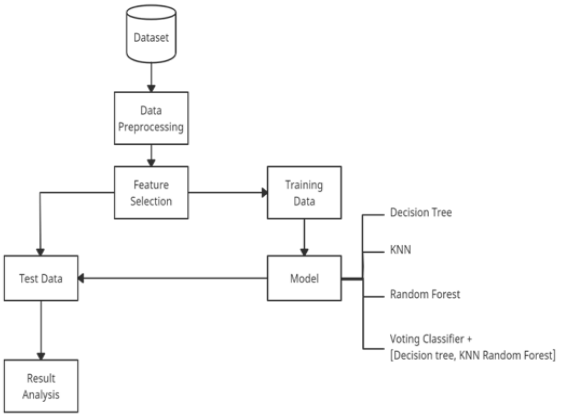

Figure 1. Proposed System

V. RESULT AND DISCUSSION

Every machine learning algorithm has its own pros and cons. An algorithm might be very efficient for some problem but the same algorithm might be very inefficient for some other problem. Decision Tree, KNN, Random Forest and soft voting classifier were used to predict Diabetes. Among these algorithms soft voting classifier had the highest accuracy of 83.12%. The accuracy was calculated using equation 1.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \quad (1)$$

The predicted values were divided into True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) for the calculation of accuracy of the machine learning algorithm. Table 3 shows the TP, FP, TN, FN and accuracy of each of the algorithm used.

Table 3. TP, FP, TN, FN and accuracy of each algorithm

Algorithm	TP	FP	TN	FN	Accuracy
Decision Tree	41	16	83	14	80.52
KNN	32	14	85	23	75.97
Random Forest	36	10	89	19	81.17
Voting Classifier	40	11	88	15	83.12

We can see that Voting classifier had the highest accuracy of 83.12% and KNN had the lowest accuracy of 75.93%. Remaining two algorithms namely Decision Tree and Random Forest had an accuracy of 80.52% and 81.17% respectively.

VI. CONCLUSION

We have used 4 machine learning algorithms (Decision Tree, KNN, Random Forest and soft voting classifier). Voting classifier had the best accuracy. We realized that when we take sum the prediction probability from Decision Tree, KNN and Random Forest and choose the prediction with largest sum we gain more accuracy than the prediction by these three algorithms individually. However, the main issue faced during the prediction was the missing values in the dataset. So, our future work will focus on improvement of the dataset by integration of different dataset to one for the prediction of Diabetics. We also need to explore new methods to eliminate the missing values present in the dataset.

REFERENCES

[1] International Diabetes Federation (IDF). (2017) IDF DIABETES ATLAS - 8TH EDITION ,2017. "http://www.diabetesatlas.org/"
[2] Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill.
[3] WebMD. Diagnosis of Diabetes. "https://www.webmd.com/diabetes/guide/diagnosis-diabetes"
[4] M. Deepika, Dr. K. Kalaiselvi, "An Empirical study on Disease Diagnosis using Data Mining Techniques", 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018); ISBN:978-1-5386-1974-2
[5] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid and Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare" International Conference on Automation and Computing (ICAC) 2018
[6] Ridam Pal, Dr.Jayanta Poray, Mainak Sen, "Application of Machine Learning Algorithms on Diabetic Retinopathy", IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, 2017.
[7] Priyanka Sonar, Prof. K. Jaya Malini, "diabetes prediction using different machine learning approaches", Third International Conference on Computing Methodologies and Communication (ICCMC) 2019.
[8] P. Suresh Kumar, V. Umatejaswi "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017
[9] Mamta Arora, Mrinal Pandey "Deep Neural Network for Diabetic Retinopathy Detection", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con) 2019
[10] Aishwarya Mujumdara, Dr. Vaidehi V "Diabetes Prediction using Machine Learning Algorithms", International Conference on Recent Trends in Advanced Computing (ICRTAC) 2019