Predictive model for Diabetes the Silent Killer

Adarsh Jaiswal[1], Dishant Kumbhar[2], Samruddhi Patil[3]
[1,2,3] *Nutan College of Engineering and Research*

***Abstract*—Diabetes is a disease caused due to the increase level of blood glucose. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of machine learning technique. This project aims to predict diabetes via supervised machine learning methods Logistic regression. This project also aims to propose an effective technique for earlier detection of the diabetes disease.**

***Index Terms*— *KNN*, Logistics Regression, Machine Learning, SVM**

I. INTRODUCTION

Diabetes mellitus is an endless infection portrayed by hyperglycemia. It might cause numerous inconveniences. As per the developing bleakness as of late, in 2040, the world's diabetic patients will achieve 642 million, which implies that one of the ten grown-ups later on is experiencing diabetes. There is no uncertainty this disturbing figure needs extraordinary consideration. World Health Organization has assessed 12million passing happen around the world, consistently because of Heart maladies. A large portion of the passing in the United States and other created nations are expected to cardio vascular maladies. The early visualization of cardiovascular sicknesses can help in settling on choices on way of life changes in high hazard patients and thus decrease the intricacies. This exploration means to pinpoint the most significant/hazard elements of coronary illness just as anticipate the general hazard utilizing calculated relapse. Machine Learning has been connected to numerous parts of medicinal wellbeing. In this project, we have utilized different algorithms to anticipate diabetes mellitus.

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 neuralnetwork (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%, DecisionTreeachieved63.51%, K-NearestNeighbors achieved 60.03% and Support Vector Machine algorithm achieved67.85%.

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.

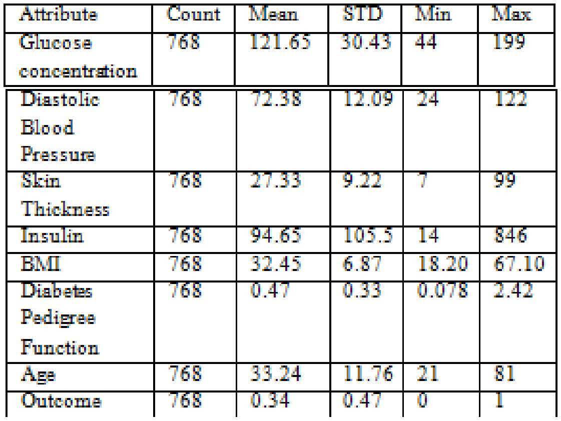
TABLE1. ATTRIBUTEDESCRIPTION

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



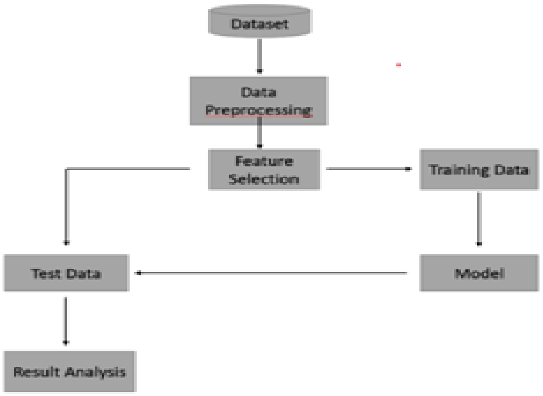
Attribute	Count	Mean	STD	Min	Max
Glucose concentration	768	121.65	30.43	44	199
Diastolic Blood Pressure	768	72.38	12.09	24	122
Skin Thickness	768	27.33	9.22	7	99
Insulin	768	94.65	105.5	14	846
BMI	768	32.45	6.87	18.20	67.10
Diabetes Pedigree Function	768	0.47	0.33	0.078	2.42
Age	768	33.24	11.76	21	81
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 machine learning algorithm for the prediction of Diabetics. The algorithm we have used are Logistic Regression, KNN, SVM.

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.



V. RESULTS 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. Logistic Regression, KNN, SVMwere used to predict Diabetes. Among these algorithms Logistic Regression classifier had the highest accuracy of 81.81%. The accuracy was calculated using below equation

$$\textbf{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

Algorithm	Logistic Regression	KNN	SVM
Accuracy	81.81	68.18	77.27

We can see that Logistic Regression has highest accuracy of 81.81and KNN has the lowest accuracy of 68.18 . SVM algorithm had an accuracy of 77.27.

VI. CONCLUSION

We have used 3 Machine Learning Algorithms (Logistic Regression, KNN, SVM). Logistic Regression had the best accuracy. However, the main issue faced during the prediction was the missing values in our 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

[11] Academia.eduhttps://en.wikipedia.org/wiki/Clinical_data_management

[12] http://shodh.inflibnet.ac.in:8080/jspui/bitstream/123456789/4170/3/03_lite rature%20review.pdf

[13] World Health Organization. Available online: http://www.who.int (accessed on 14 September 2018).

[14] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal