Investigation of Various Machine Learning Techniques for Diabetes Analysis

Sangeeta Bairagi[1], Ankur Taneja[2]

[1,2] *Department of Computer Science and Engineering, Sam College of Engineering and Technology, Bhopal (M.P)*

---***---

Abstract - *Diabetes is a high-risk medical condition characterized by abnormally high blood sugar levels. It is a leading cause of death worldwide. According to increasing morbidity in recent years, the number of diabetic patients worldwide will reach 642 million by 2040, implying that one out of every ten adults will have diabetes. It is undeniable that this needs a great deal of attention. A variety of data mining and machine learning techniques have been used on the diabetes dataset to predict disease risk. The aim of this proposed work is to examine these machine learning techniques using output metrics and method features. The research includes the Pima Indian diabetes dataset, which includes 768 patients, 268 of whom are diabetic and 500 of whom are not. Diabetes es is a high-risk medical condition characterized by abnormally high blood sugar levels. It is a leading cause of death worldwide. According to increasing morbidity in recent years, the number of diabetic patients worldwide will reach 642 million by 2040, implying that one out of every ten adults will have diabetes. It is undeniable that this needs a great deal of attention. A variety of data mining and machine learning techniques have been used on the diabetes dataset to predict disease risk. The aim of this paper is to examine these machine learning techniques using output metrics and method features. The research includes the Pima Indian diabetes dataset, which includes 768 patients, 268 of whom are diabetic and 500 of whom are not.*

Key Words: Diabetes analysis; Glucose level prediction; Machine learning, SVM, Naive Bayes, Random forest

1.INTRODUCTION

Diabetes is a deadly killer that goes unnoticed. The presence of excessive levels of metabolites such as glucose is the primary cause of this disorder. Diabetes mellitus (DM) is the most common form of diabetes. It is a dangerous and complicated illness. When cells and/or the pancreas fail to produce enough insulin, blood sugar rises, affecting various organs, including the skin, kidneys, and nerves [1, 2]. Diabetes is one of the most common diseases that affects the elderly around the world. According to the International Diabetes Federation, there were 451 million diabetics worldwide in 2017. In the next 26 years, this figure is expected to rise to 693 million people [1]. The three forms of diabetes that exist in the human body as shown in fig 1.

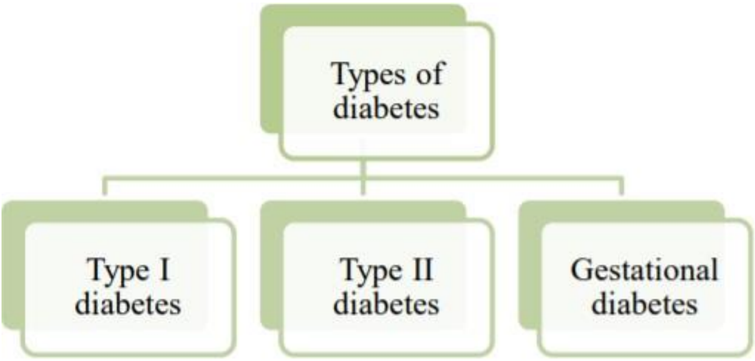

Figure 1 Types of diabetes

1.1 Risk Factors for Diabetes

One of the most important risk factors for type 1 diabetes is family history. Apart from that, type 1 diabetes is linked to environmental causes and certain viral infections [3]. Type-2 diabetes is due to the following risk factors.

1. Stress
2. The family history of diabetes
3. Age
4. Obesity
5. Lack of physical activity
6. Hypertension
7. Unbalanced diet
8. Gestational diabetes
9. Race/Ethnicity

Diabetes mellitus has long-term consequences. A diabetic also faces a high risk of developing a variety of health issues. Although the exact cause of diabetes is unclear, scientists agree that genetic factors as well as environmental lifestyle play a significant role. Even though it is incurable, it can be managed with therapy and medication. Diabetes patients are at risk of developing secondary health problems including heart failure and nerve damage. As a result, early diagnosis and treatment of diabetes will help to avoid complications and lower the risk of serious health issues.

Many bioinformatics researchers have attempted to solve this disease by developing systems and methods that will aid in diabetes prediction. They either used different types of machine learning algorithms, such as classification and association algorithms, to build prediction models.The most common algorithms were Decision Trees, Support Vector Machines (SVM), and Linear Regression [3-5].

Another form of machine learning technique is the Artificial Neural Network (ANN). It's well-known for its high efficiency and precision. Deep Learning (DL) has also been implemented as an update to ANN due to the growing size and complexity of the data. Recent DL-based studies have yielded impressive results [6,7]. These methods provided a range of accuracy rates. This has prompted researchers to try to improve accuracy by either constructing models with previously unseen classifiers or combining various classifiers [8-10]. The public Pima Indian Dataset collected from the UCI repository was used in most studies in the field of diabetes prediction.

The main goal of this research is to look at a dataset of diabetic patients and apply various machine learning algorithms to it. The accuracy and efficiency metrics are the main objective of the comparison of various algorithms. Comparative studies of this kind have been conducted in the past.The aim of this study is to use various machine learning algorithms to prepare for the analysis of diabetes.

2. DATASET

The Pima Indians Diabetes (PID) dataset can be found in [11] and was donated by Vincent Sigillito, a member of the Johns Hopkins University's Applied Physics Laboratory, in 1990. It's a compilation of 768 patients' medical diagnostic reports. Many of the patients are Pima Indian women who are at least 21 years old and live near Phoenix, Arizona, in the United States.

There are 268 cases in class 1 (positive diabetes test) and 500 cases in class 0 (negative diabetes test), respectively, accounting for 34.9 percent and 65.1 percent of the total dataset. The following are the eight attributes (plus class) that can be described:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)ˆ2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

3. MODEL BUILDING

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K-Nearest Neighbor and Gaussian Naïve bayes Algorithms

3.1 Algorithm :

Diabetes Prediction using various machine learning algorithms Generate training set and test set randomly. Specify algorithm that are used in model mn=[LogisticRegression()]

```
Take the value of n as input;
for(i=0; i<n; i++)

do
Model= min[i];
Model.fit();
model.predict();
print(Accuracy(i), classification_report);
End
```

Evaluation

This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as equation 1:

$$Accuracy = \frac{(Number of Correct Predictions)}{(Total number of predictions Made)}$$

(1)

4. RESULTS

Logistic regression can be used also to solve problems of classification. In general, logistic regression classifier can use a linear combination of more than one feature value or explanatory variable as argument of the sigmoid function. The corresponding output of the sigmoid function is a number between 0 and 1. The middle value is considered as threshold to establish what belong to the class 1 and to the class 0. An input producing an outcome greater than 0.5 is considered belongs to the class 1. Conversely, if the output is less than 0.5, then the corresponding input is classified as belonging to 0 class. After applying various machine Learning Algorithms on dataset, accuracies found are as mentioned in table 1. Logistic Regression gives highest accuracy of 82%. Table 2 represents the statistical analysis of Pima Indian data set.

Table 1Accuracy of various algorithms

Algorithms	Accuracy with PIMA Dataset
Logistic Regression	82%
Gradient Boost Classifier	77%
LDA	77%
AdaBoost Classifier	77%
Extra Trees Classifier	76%
Gaussian NB	67%
Bagging	75%
Random Forest	72%

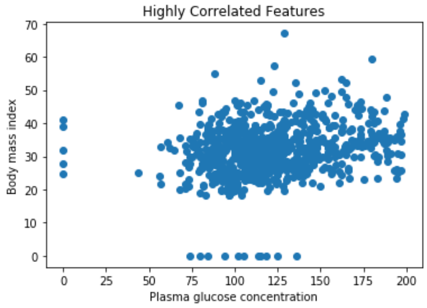

Fig. 2 Correlation between Plasma glucose and BMI

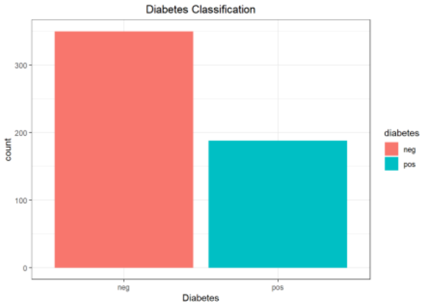

Fig. 3 Diabetes Classification

Table 2 Pima Indian dataset statistical analysis

	Number of times pregnant	Plasma glucose concentration	Diastolic blood pressure	Triceps skinfold thickness	2-Hour serum insulin	Body mass index	Diabetes pedigree function	Age	Outcomes
Count	768.00	768.00	768.00	768.00	768.00	768.00	768.0	768.00	768.00
Mean	3.84	120.89	69.10	20.53	79.79	31.99	0.471	33.24	0.348
Standard deviation	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76	0.47

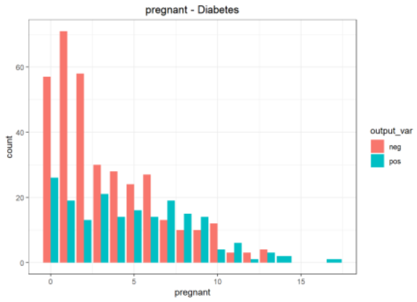

Fig. 4Pregnant women diabetes classification

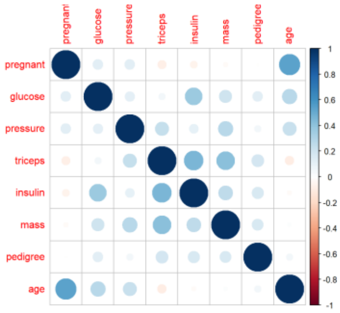

Fig. 5. Correlation map

5. Conclusions and Future Scope

A review of various machine learning methods is presented in this thesis with their evaluation at single place. Authors can understand machine learning methods and usage of these methods with domain dependent or non-domain dependent features for diabetes dataset. Evaluations of these methods are also shown, from which methods of best result can be used in future for other applications. According to this evaluation, logistic regression performed best compared to decision tree and random forest for Pima Indian dataset. The accuracy of logistic regression is 82% for Pima Indian dataset which got best score as compared to previous work. We have seen comparison of machine learning algorithm accuracies. The model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely non-diabetic people can have diabetes in next few years.

References

[1] Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; Fernandes, J.D.R.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pr. 2018, 138, 271–281.

[2] Sanz, J.A.; Galar, M.; Jurio, A.; Brugos, A.; Pagola, M.; Bustince, H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Appl. Soft Comput. 2014, 20, 103–111.

[3] Kandhasamy, J.P.; Balamurali, S. Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Comput. Sci. 2015, 47, 45–51.

[4] Iyer, A.; Jeyalatha, S.; Sumbaly, R. Diagnosis of Diabetes Using Classification Mining Techniques. Int. J. Data Min. Knowl. Manag. Process. 2015, 5, 1–14.

[5] Razavian, N.; Blecker, S.; Schmidt, A.M.; Smith-McLallen, A.; Nigam, S.; Sontag, D. Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors. Big Data 2015, 3, 277–287. [CrossRef]

[6] Ashiquzzaman, A.; Kawsar Tushar, A.; Rashedul Islam, M.D.; Shon, D.; Kichang, L.M.; Jeong-Ho, P.; Dong-Sun, L.; Jongmyon, K. Reduction of overfitting in diabetes prediction using deep learning neural network. In IT Convergence and Security; Lecture Notes in Electrical Engineering; Springer: Singapore, 2017; Volume 449.

[7] Swapna, G.; Soman, K.P.; Vinayakumar, R. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Comput. Sci. 2018, 132, 1253–1262.

[8] Rahimloo, P.; Jafarian, A. Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them. Bull. Société R. Sci. Liège 2016, 85, 1148–1164.

[9] Gill, N.S.; Mittal, P. A computational hybrid model with two level classification using SVM and neural network for

predicting the diabetes disease. J. Theor. Appl. Inf. Technol. 2016, 87, 1–10.

[10] NirmalaDevi, M.; Alias Balamurugan, S.A.; Swathi, U.V. An amalgam KNN to predict diabetes mellitus. In Proceedings of the 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, India, 25–26 March 2013; pp. 691–695.

[11] http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data