Implementing Classification Algorithms for Predicting Chronic Diabetes Diseases

M. Kavitha, S. Subbaiah

Abstract--- Now a day Chronic Diabetes Disease is increasing due to many reasons like changes in life style, food habit. It causes an increase in blood sugar levels. If Diabetes Disease remains untreated or unidentified, many different types of complications may be occurred. The doctors have the problem to identify these kinds of diseases easily. The machine learning algorithms helps the doctor to solve these types of problems. In this paper, we implemented three algorithms namely logistic regression, Naive Bayes and Decision tree algorithms to predict diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Dataset, which is from UCI machine learning repository. The performance of all the three algorithms is evaluated using measures on Accuracy. Results obtained showed logistic regression displays 75.3%, Decision tree displays 77.9% and Naive Bayes classifier displays the accuracy value is 76.6%.

Keywords--- Diabetes, Logistic Regression, Decision Tree, Naive Bayes, Machine Learning.

I. INTRODUCTION

Classification algorithms are used mostly in medical field for classifying data into different number of classes based on constraints. Diabetes is a one of the illness which affects the body in producing the insulin and levels of glucose in the blood.Diabetes has some symptoms like intensify hunger, frequent urination caused due to high blood pressure. Many problems occur if diabetes disease remains untreated. Diabetic ketoacidosis and nonketotic hyperosmolar coma are the most complications include in diabetes [1].

Diabetes is analyzed seriously during the measure of sugar substance cannot be controlled. Diabetes disease is not only affected due to some factors like height, weight, hereditary and insulin but the major factor is considered is sugar concentration among all these factors. The early prediction of chronic disease is only one way to recover from these complications [2].

Most Researchers are conducting experiments for diagnosing the chronic diabetes diseases using different kinds of classification algorithms of machine learning approaches like SVM, KNN, Decision tree, Naïve bayes etc.They proved that the machine learning algorithms works better for diagnosing different kinds of diseases. [3], [4], [5]. Data mining and machine learning algorithms increase in power due to capacity of managing a large amount of data to merge data from different sources and integrating the information [6].

This research works focuses on Naïve Bayes, Decision tree and logistic regression algorithms. These algorithms are used and evaluated on Pima Indians Diabetes Dataset to predict the diabetes disease. The performance of these algorithms is discussed and achieved good accuracy. The remaining of the research discussion is organized as follows: Section 2 specifies related work of different types of classification algorithms for predicting diabetes diseases, Section 3 briefs discussion of dataset used, Section 4 describes experimental results and Section 5 shows the conclusion of the research work.

II. RELATED WORK

Different classification and clustering algorithms have used for prediction and diagnosis of diabetes [7], [16]. In [8], cardiovascular disease is predicted using support vector machines.

They got 85% results accuracy.In [9], the author mentioned the Classification of Diabetes Disease Using Support Vector Machine. They got the results 78%.VeenaVijayan V [10] et al used different types of machine learning algorithms for predicting diabetes disease. They mentioned the decision tree, support vector machine, Naivebayes and decision stump. They got the result accuracy is 77%, 79%, 79% and 80% respectively.

Abdul AzisAbdullah [11] et al mentioned the Diagnosis of Diabetes using Support Vector Machines with Radial Basis Function Kernels.

They used different number of data sets like 100 data, 200 data, 300 data and 400 data and 500 data. They got accuracy value is 83%, 72%, 84%, 76% and 82%.Murat Pojan [12] predicts the student performance using machine learning algorithms.

The author used three algorithms. They are linear regression, decision tree and naïve Bayes classification. To improve the prediction valueused feature engineering. He concluded that naïve Bayes classification gave best results comparing with other algorithms. The results show 98% accuracy in naïve Bayes algorithms to predict student performance.

III. METHODOLOGY USED

A. Model Diagram

The model diagram is summarized in figure 1. The figure shows the sequence of research conducted in developing the model.

Revised Manuscript Received on 14 August, 2019.

M. Kavitha, Assistant Professor &Ph.D Part Time Research Scholar, PG & Research Department of Computer Science and Applications, Vivekanandha College of Arts and Science for Women, Tiruchengode, Tamilnadu, India.(e-mail: sankavis@gmail.com)

Dr.S. Subbaiah, Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamilnadu, India.(e-mail:subbunaren@yahoo.com)

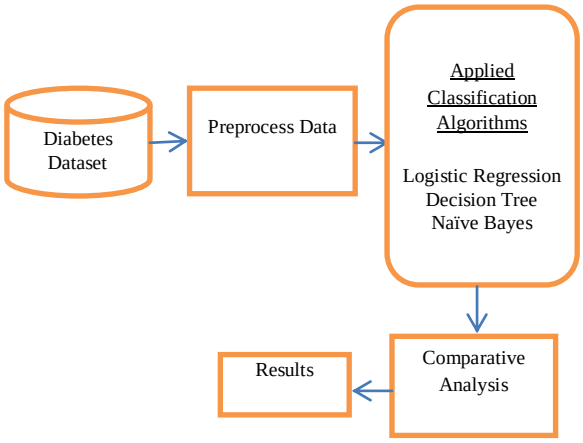

Fig.1: Proposed model diagram

B. Brief Description of Algorithm Used

a. Logistic Regression

Logistic regression is technique used in machine learning field of statistics.It is the method for binary classification problems with two class values. The logistic function, also called the sigmoid function. It was developed by mathematicians to define properties of population growth in ecology, increasing rapidly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Here is the base of the natural logarithms.

Below is an example logistic regression equation:

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x))

Here y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The Accuracy of the experiment is evaluated using R tool. Fig.2. shows that the matrix of scatterplots is produced in this experiment.

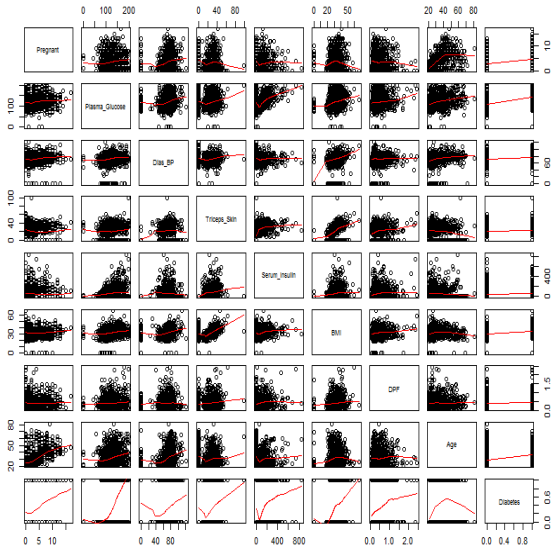

Fig. 2: Matrix of scatterplot in logistic regression

Then we apply the matrix of correlations between the variables. Fig. 3.Displays the correlations values among variables.

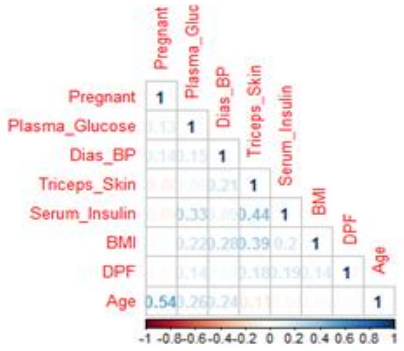

Fig. 3: Matrix of Correlations Between the Variables

Next the results give variables statiscally significance. Plot the new model developed using R. It displayed in Fig.4.

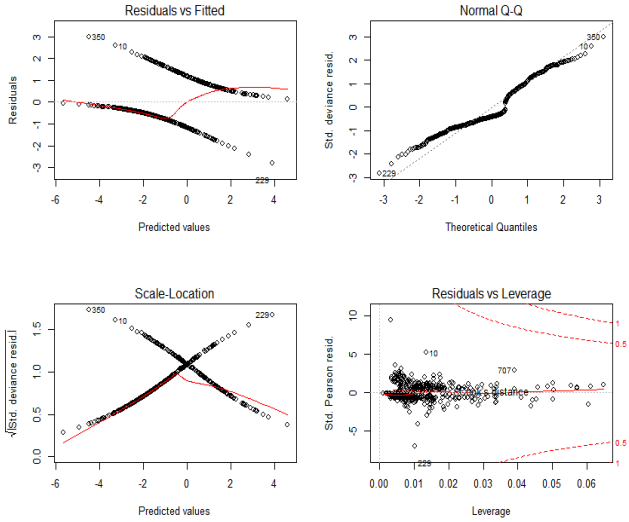

Fig.4: Logistic Regression Model

The Fig.5 Shows a brief description of the evaluated performance of Logistic Regression technique using Confusion Matrix is as follows:

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 128 34

1 23 46

Accuracy : 0.7532

95% CI : (0.6924, 0.8074)

No Information Rate : 0.6537

P-Value [Acc> NIR] : 0.0007155

Kappa : 0.4368

Mcnemar's Test P-Value : 0.1853263

Sensitivity : 0.8477

Specificity : 0.5750

PosPredValue : 0.7901

NegPredValue : 0.6667

Prevalence : 0.6537

Detection Rate : 0.5541

Detection Prevalence : 0.7013

Balanced Accuracy : 0.7113

'Positive' Class : 0

Fig. 5: Confusion Matrix of Logistic Regression

b. Decision Tree

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes [11].

The Fig.6 shows the evaluated performance of Decision Tree technique using Confusion Matrix is as follows:

Confusion Matrix and Statistics
Reference
Prediction 0 1
0 127 26
1 25 53
Accuracy : 0.7792
95% CI : (0.7201, 0.831)
No Information Rate : 0.658
P-Value [Acc> NIR] : 4.001e-05

Kappa : 0.508
Mcnemar's Test P-Value : 1

Sensitivity : 0.8355
Specificity : 0.6709
PosPredValue : 0.8301
NegPredValue : 0.6795
Prevalence : 0.6580
Detection Rate : 0.5498
Detection Prevalence : 0.6623
Balanced Accuracy : 0.7532

'Positive' Class : 0

Fig.6: Confusion Matrix of Decision Tree

c. Naive Bayes

Naive Bayes is a classification technique with a notion which describes all features are independent and unrelated to each other. It describes that position of a precise feature in a class does not disturb the position of another feature. Since it is founded on conditional probability it is measured as a commanding algorithm working for classification purpose. It works well for the data with unbalancing problems and missing values. Naive Bayes [13] is a machine learning classifier which employs the Bayes Theorem.

Using Bayes theorem posterior probability P (C|X) can be calculated from P(C), P(X) and P (X|C) [14].

Therefore, P (C|X) = (P (X|C) P(C))/P(X)

Where, P (C|X) = target class's posterior probability.

P (X|C) = predictor class's probability.

P(C) = class C's probability being true.

P(X) = predictor's prior probability.

The Fig 7. Shows the evaluated performance of Naive Bayes algorithm using Confusion Matrix is as follows:

Confusion Matrix and Statistics
Reference
Prediction no yes
no 128 30
yes 24 49

Accuracy : 0.7662
95% CI : (0.7063, 0.8192)
No Information Rate : 0.658
P-Value [Acc> NIR] : 0.0002359

Kappa : 0.4709
Mcnemar's Test P-Value : 0.4962425
Sensitivity : 0.8421
Specificity : 0.6203
PosPredValue : 0.8101
NegPredValue : 0.6712
Prevalence : 0.6580
Detection Rate : 0.5541
Detection Prevalence : 0.6840
Balanced Accuracy : 0.7312

'Positive' Class : no

Fig.7: Confusion Matrix of Naive Bayes

d. Dataset Used

In this work R tool is used for performing the experiment. R is free software with collaborative project with many contributors. It includes a collection of various machine learning techniques for data classification, clustering, regression, visualization etc. The main goal of this work is the prediction of the patient affected by diabetes using the R tool by using the medical database PIDD. Table-1 shows a brief description of the dataset.

Table 1: Dataset Description Database

Database	No.of Attributes	No.of instances
Pima Indians Diabetes Dataset(PIDD)	9	768

The proposed methodology is assessed on Diabetes Dataset namely (PIDD) [15], which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are feminine patients. The dataset also contains numeric-valued 9 attributes where value of one class '0' preserved as tested negative for diabetes and value of another class '1' is preserved as tested positive for diabetes.

Dataset description defined by Table-1 and the Table-2 represents Attributes descriptions.

In this experiment, 70% (537) of the data from Pima Indians Diabetes Dataset(PIDD) used for training process and 30% (231) of the data used for testing process.

Table 2: Attribute Description

Attribute	Abbreviation of Attributes
1. Number of times pregnant	pr
2. Plasma glucose concentration	pl
3. Diastolic blood pressure (mm Hg)	pr
4. Skin fold thickness (mm)	sk
5. 2-Hour serum insulin (mu U/ml) in	in
6. BMI (weightinkg/(heightinm)2) ma	ma
7. Diabetes pedigree function pe	pe
8. Age in years	ag
9. Class '0' or '1'	cl

IV. RESULTS

Table-3represents different performance values of all classification algorithms calculated on Accuracy measures.From Table-3 it is analyzed that Decision Tree showing the maximum accuracy. So the Decision tree machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers.

Table 3: Comparative Performances of all classifier's based on Accuracy

Classification Algorithms	Accuracy %
Logistic regression	75.3%
Decision tree	77.9%
Naive Bayes classifier	76.6%.

V. CONCLUSION

One of the significant real-world medical problems is the discovery of diabetes at its early stage. In this study, systematic efforts are made in modeling asystem that results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 77.9 % using the Decision Tree classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

REFERENCES

[1] Dr. D. Ashok Kumar, R. Govindasamy, "Performance and Evaluation of Classification Data Mining Techniques in Diabetes", *International Journal of Computer Science and Information Technologies*, Vol. 6 (2), 2015, 1312-1319

[2] V. VeenaVijayan, C. Anjali, "Prediction and diagnosis of diabetes mellitus A machine learning approach", *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 122–127.

[3] R. Aishwarya, P. Gayathri, N. Jaisankar , "A Method for Classification Using Machine Learning Technique for Diabetes", *International Journal of Engineering and Technology (IJET)*, Vol. 5(3), 2903–2908.

[4] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", *Computational and Structural Biotechnology Journal*, Vol. 15, 2017, 104–116.

[5] B. DhomseKanchan, M. MahaleKishor, 2016, "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE.* pp. 5–10.

[6] M.Fatima, M. Pasha,"Survey of Machine Learning Algorithms for Disease Diagnostic", *2017 Journal of Intelligent Learning Systems and Applications*, 09, 1–16.

[7] A. Iyer, R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", *2015 International Journal of Data Mining & Knowledge Management Process*, 1–14.

[8] S.R. Alty, S.C. Millasseau , P.J. Chowienczyc, A. Jakobsson, "Cardiovascular Disease Prediction Using Support Vector Machines", *2003 46th Midwest Symposium on Circuits and Systems,IEEE,*2014

[9] V. AnujaKumari, R.Chitra, "Classification of Diabetes Disease Using Support Vector Machine", *International Journal of Engineering Research and Applications (IJERA)* ISSN: 2248-9622, Vol. 3, Issue 2, March - April 2013, pp.1797-1801

[10] V. VeenaVijayan, C. Anjali, "Prediction and Diagnosis of Diabetes Mellitus –A Machine Learning Approach", *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* | 10-12 December 2015.

[11] Abdul Azis Abdullah, Suwarno, "Diagnosis of Diabetes using Support Vector Machines with Radial Basis Function Kernels", *International Journal of Technology* (2016) 5: 849-858 ISSN 2086-9614.

[12] Murat Pojon (2017), Using Machine Learning to Predict Student Performance.

[13] Irina Rish, 2001, "An empirical study of the Naïve Bayes classifier", *IJCAI2001 workshop on empirical methods in artificial intelligence*, IBM.pp.41–46.

[14] S. Ray, 2017, 6 Easy Steps to Learn Naïve Bayes Algorithm

[15] K. Kayaer, K.Tulay, 2003, "Medical diagnosis on Pima Indian diabetes using general regression neural Networks".

[16] DeeptiSisodia, Dilip Singh Sisodia,"Prediction of Diabetes using Classification Algorithms", *International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science* 132,(2018),pp.1578–1585