International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Data Mining Technique for Diabetes Diagnosis using Classification Algorithms

Priya. M, M. Karthikeyan

Abstract*: Diabetes mellitus is defined as a one of the chronic and deadliest diseases which combined with abnormally high level of sugar (glucose) in the blood. The classification technique helps in diagnosis the symptoms at starting stages. This paper focused to prognosticate the chance of diabetes in patients with extremely correct classification of Diabetes. The classification algorithms viz., Naïve Bayes, Logistic Regression, and Decision Tree can be used to detect diabetes at an early stage. The algorithm performances are evaluated based on various measures like Recall, Precision, and F-Measure. Experiments are conducted where the time complexity of each of the algorithm is measured. Accuracy is also measured over correct classification and misclassification instances, observed that a Logistic Regression algorithm has much better performance when compared to the other type classifications. Using Receiver Operating Characteristic curves the results are verified in a systematic manner.

Keywords: Classification, Data Mining, Decision Tree, Diabetes Mellitus, Logistic Regression, Naïve Bayes.

I. INTRODUCTION

Data mining is to extract hidden knowledge from large data. Data mining techniques are major strength as a result to the capability of organizing a large amount of data into discovery of new information from many different sources and integrating the background of information [1]. Several classification algorithms are used diagnosis the symptoms of diabetes. Moreover, in this paper a proposed classification algorithm in data mining approaches for diabetic diagnosis.

Diabetes mellitus has a metabolic disease that causes high blood sugar. Diabetes is a major cause of concern and its prevalence of diabetes and impaired glucose tolerance were 2.5% and 3.2%, respectively. The hormone insulin moves sugar from the blood into your cells to be stored or used for energy. Untreated high blood sugar from diabetes can damage your nerves, eyes, kidneys and other organs. Diabetes is analyzed as an essential serious health problem during which the measure of blood sugar substance cannot be controlled. The early stages of disease identification are the only treatment to stay away from the complications.

The individual person a diabetic has dangerous of holds alternate illnesses in vein mischief, optical deficiency, coronary diseases, and kidney infection and nerve harm. Diabetes is experimented as a critical health issues during the measure of sugar (glucose) substance cannot be controlled. Diabetes is affected not only by various factors like hereditary factor, height, weight, and insulin but the main reason is considered as glucose concentration of all factors. The early diabetic identification is only solution to stay away from the complications [2]. The researchers are to diagnosing the diseases for conducting experiments using various classification methods of Machine Learning approaches like Naive Bayes, J48, SVM, Decision Tree, etc. As a researchers have research and to prove that the Machine Learning algorithms works better for diagnosing in different diseases [3],[4].

II. RELATED WORK

Iyer et al. focuses on pregnant women affecting from diabetes. In this work, the machine learning classification methods are used to evaluate on the Pima Indian Diabetic dataset to find the prediction of patient diabetics. [5]. Orabi et al. proposed a system for diabetes prediction, major aim is to predict the diabetes of a candidate whose is suffering at a particular age [6].

Priyam et al. proposed a system is based on the machine learning concept, by using decision tree. A result was obtained satisfactory and works well in prediction of the diabetes factors at a particular age, with higher accuracy [7]. Kumar et al. discussed the method provides to prevention control and awareness of diabetes and effect of relevant other diseases [8]. Fatima et al. focused the Data Mining algorithms to gain its strength due to the managing capability of a large amount of data is to combined from several different sources and integrating the background information [9].

Priya and Karthikeyan have discussed different clustering algorithms and compared for outlier detection. The comparison is made to detect outliers in the health care datasets by using clustering algorithms [10]. Rashid et al. in designed to predict diabetes chronic disease with two sub-modules. The first module is ANN (Artificial Neural Network) and the second module is FBS (Fasting Blood Sugar) [11]. Priya and Karthikeyan have proposed a method, it focused to the performance analyze of outlier detection algorithm using feature bagging technique for health care application.

Manuscript published on November 30, 2019.
* Correspondence Author
M. Priya*, Assistant Professor, Department of Computer Science, PSPT MGR Government Arts & Science College, Sirkali – Puthur -609108, Tamilnadu, India .
Email: mpriyaau@gmail.com
M. Karthikeyan, Assistant Professor, Division of Computer and Information Science, Faculty of Science, Annamalai University, AnnamalaiNagar, Tamilnadu, India. Email: karthiaucse@gmail.com

Ensemble classifiers have been effective in improving overall performance and stability of machine learning techniques. [12]. Priya and Karthikeyan have been proposed an algorithm which is used to identify objects as outlier and outlier clusters in a database. It is based on mutual nearest neighbor graph clustering. This algorithm can be used to find the outlier value factor in the database and to detect the outliers and outlier clusters efficiently [13]. The recommendation of World Health Organization (WHO) is to develop a simple strategy to identifying those at risk of diabetes and to provide them with early lifestyle circumstance [14]. Mandal et al. discussed hierarchical clustering technique to find models for diabetics mellitus [15]. Ioannis K et al. have discussed a systematic review of the machine learning applications, and data mining tools which is used in data mining approaches in the field of diabetes research with respect to Diagnosis and Prediction, Diabetic obstacle, Genetic Background, and Health Care Management to be appearing the most popularly with the first category [16].

III. PROPOSED SYSTEM

In the real world databases are highly susceptible to missing, noisy, and inconsistent data due to their large size and likely origin from multiple, diversified sources. For that reason, data preprocessing phase is necessary for classification purposes. Data preprocessing must include cleaning of data, reduce dimensionality of data, and data transformation that is data normalization, data binning followed by classification. The modal diagram for the proposed system is summarized in flow chart. The Fig. 1 shows that the flow of constructing the model in which the research conducted.

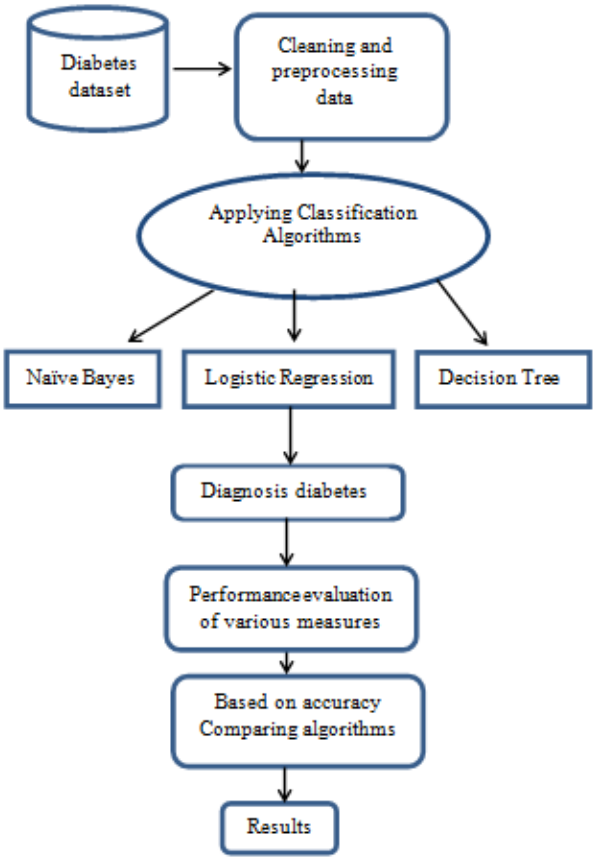

Fig. 1 Proposed Model Diagram

A. Naïve Bayes Classification

Naïve Bayes is a classification algorithm which defines all the features is independent and separated to each other. Naive Bayes is one of machine learning classifier which defines the principles of Bayes theorem that is the status of a particular feature in a class does not affect the status of another feature. Moreover it is based on to calculating conditional probability and it is considered as a dominant algorithm use for classification purpose. It has been works well for the misbalancing data problems and missing values. Bayes theorem is used by calculating the posterior probability P (H|X), from P(C), P(X) and P (X|H). Therefore, the Bayes theorem is in equation (1)

$$P(H \mid X) = \frac{(P(X \mid H)(P(H))}{P(X)} \quad (1)$$

Where,

P (H|X) = Posterior Probability of target class, P (X|H) = Probability of Predictor Class,

P (H) = H's class Probability being true, and P(X) = predictor Prior Probability.

B. Decision Tree Classification

Decision Tree is one the supervised machine learning algorithm and it is used to solve classification problems. This algorithm is to construct tree and by using divide and conquer method. Hence, it is the prediction of destination class which is taken from prior data using some decision rule. It uses root nodes and internodes for the classification and prediction. Root level nodes classify the objects with different features. Root level nodes having two or more branches which called the leaf nodes to represent classification. In the Decision tree chooses each and every node by evaluating the maximum information gain for all the attributes at every stage.

The training data is used to construct the tree and then it is pruned to avoid the over fitting problem. The process of pruning is staring from the leaf nodes to root node and evaluate the error rate at leaf node. It calculates different error evaluations by changing branch with leaf or sub tree leaf or branch with other branch and selects the best combination which produce minimum error rate.

C. Logistic Regression Classification

Logistic regression is another technique in machine learning from the area of statistics. It is a nonlinear regression technique for predicting a categorical dependent variable. Logistic regression was observed well to identifying the risk factors for many diseases by using patient characteristics, history, and risk factors.

A functional relationship of a dependent variable Y with the independent variables X_1, X_2,…X_k and which involving parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ of the type. The regression mode defined as equation (2).

$$Y = \Psi (X_1, X_2,\ldots X_k \mid \beta_0, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon \quad (2)$$

Where Ψ is the form of the equation and ε indicates the random variable distributed with mean 0 and variance σ_ε^2 is called as the residual error term.

The logistic model is to computes formula for the probability of the selected disease y (that is y = 1 if the object affected from the disease; otherwise, y = 0) as the predictive risk factors of a function of the values. If the individual object affects from the disease, the conditional probability is given by p(y = 1 | X) = p(X), and the logistic model shown the form of equation (3).

$$\log[p(x)/1 - p(x)] = \beta_0+\beta_1X_1+\beta_2X_2+\ldots+ \beta_kX_k \quad (3)$$

where X = $(x_1,x_2,\ldots,x_k)$ the vector of k's risk factors by the logistic regression method. The LR model has been built based on the set of training data and it has tested using the testing dataset.

IV. RESULTS AND DISCUSSIONS

In data mining most of the research work related to the field of health care system diagnosis has consolidation of effort on the data set. Almost the machine learning approaches was designed and categorized for analyzing the data. Data analyzing include data cleaning, prediction and visualization.

A. Environment

Many tools and software's are designed for Data analysis. In this work, WEKA tool is used for testing the performance of the experiment. Waikato Environment for Knowledge Analysis (WEKA) tool is software and it is developed by University of Waikato in the country of New Zealand. It contains a collection of various data mining tasks especially for data preprocessing, classification, regression, clustering, visualization and feature selection. The main objective of this work is the prediction of the patient who is affected by diabetes by using health care data set.

B. Dataset Description

The data set which is used for this work has the patients who are affected by diabetics. The dataset consists of 370 samples of medical records (data objects). Each record has 9 attributes with respect to the diabetic diagnosis factors. The output of class variable labeled as 0 or 1 that is class 0 is for tested negative (non- diabetes) and class 1 is for tested positive (with diabetes). The data set description as shown below in table- I.

Table- I: The Data Set description

Attribute	Description
Age	Set as min=30 and max =70
Gender	Set as Male = 1 Female = 0
Mass	BMI Set as min=1 and max=200
Pedigree	Set as 0= no family history and 1= family history
Insulin dependent	Set as min = 50 and max = 500
Plasma	Set as min = 2 and max = 11
Systolic	blood pressure (Systolic) Set as min = 30 and Max = 370
Diastolic	blood pressure (Diastolic) Set as min=60 and max=350
Pregnancy	Set as 1= yes 0= no
Class	0 for tested negative (non- diabetes) and class 1 is for tested positive (with diabetes).

C. Performance Evaluation

According to the strength of the proposed system, performance evaluation is measured based on real life datasets. David et al. proposed an evaluation measures for evaluating results of Machine Learning experiments [17]. In this work, the evaluation measures like accuracy, recall, precision and F-Measure, and ROC curves are used to classifying the results. In WEKA, many measures are used. There are,

- *Kappa Statistic*: It is a measured value that compares an observed agreement with expected agreement (random chance). The value varies from 0 to 1, where 0 for the agreement chance is equivalent and 1for perfect agreement. Equation 4 is used to calculate kappa statistic value.

$$K = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \quad (4)$$

where, p_o is the observed related agreement, and p_e is probability chance of agreement.

- *Confusion Matrix:* It is a method used for summarizing the classification algorithms performances. Confusion matrix can be used to tells your classification model recognize records of different classes. Table - II shows the confusion matrix. Then the measures are defined as in Table- III.

Table- II: Confusion Matrix

		Actual Class	
		Positive	***Negative***
Predicted Class	***Positive***	TP	FP
	Negative	FN	TN

TP - True Positive, FP - False Positive,
FN - False Negative, TN - True Negative

Table- III: Evaluation Measures

Measures	Definitions	Formula
Accuracy (A)	Accuracy determines the predicting instance accuracy of the algorithm	$A = \frac{TP + TN}{\text{Total no. of samples}}$
Precision (P)	Determine exactness accuracy of the Classifier	$\text{Precision}(P) = \frac{TP}{TP+FP}$
Recall (R)	To measure the completeness of the classifiers or sensitivity	$\text{Recall}(R) = \frac{TP}{TP+FN}$
F-Measure	F-Measure can be weighted average or harmonic mean of precision and recall	$F_{Measure} = 2*\left[\frac{(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}\right]$
ROC	Using ROC (Receiver Operating Curve) curves to compare the effectiveness of tests	

D. Result Analysis

In this work, confusion matrix is used to estimate the performance of the three classification techniques for diagnosis of diabetes. From the training and testing datasets are represented in confusion matrices, the prediction details are produced. A confusion matrix is a matrix which is represents the classification results. Confusion matrix of classification models as shown in Table- IV.

Table- IV: Confusion matrix

Classification Models		A	B
Logistic Regression	*A*	186	0
	B	184	0
Decision Tree	*A*	170	16
	B	165	19
Naïve Bayes	*A*	110	76
	B	112	72

A-Tested Negative B-Tested Positive

The three algorithms are analyzed on the basis performance measures such as kappa statics, precision, recall, F- measure, ROC, and accuracy percentage. The comparative Performance measures of three classification algorithms values are listed in Table- V.

Table- V: Performance measures of classification algorithms

Classification Algorithms	Kappa Statistic	Precision	Recall	F-Measure	Accuracy %	ROC
Logistic Regression	0.983	0.992	0.992	0.992	99.18	0.999
Decision Tree	0.952	0.935	0.938	0.936	95.17	0.95
Naïve Bayes	0.879	0.859	0.863	0.860	86.30	0.819

From the Table- V shows that the Logistic regression classifier gets maximum accuracy. Hence the Logistic regression classification algorithm has been diagnosis of diabetes with more accuracy than other algorithms. The performance of models based on the valuation measures are shown via graphical representation in Fig. 2. In Fig. 3 shows the ROC area of three classifications methods.

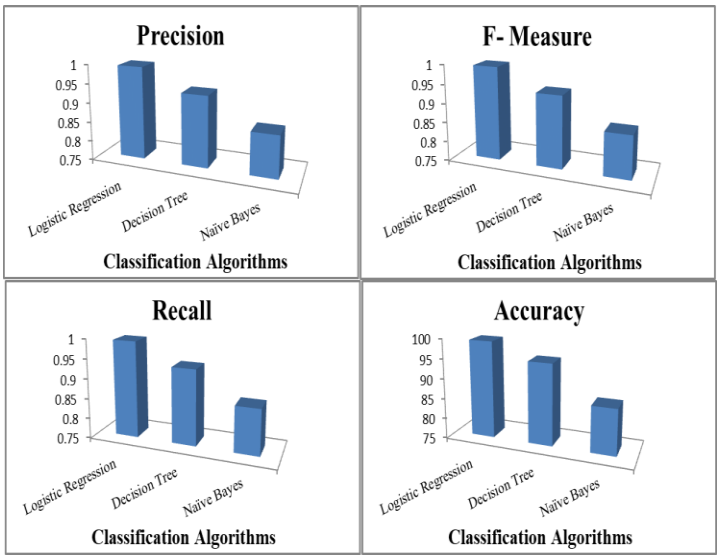

Fig. 2 Performance Measures of Classification Algorithms

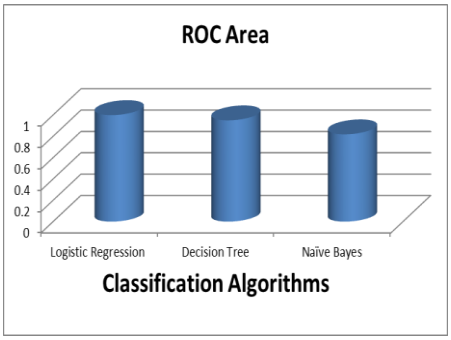

Fig. 3 ROC Area of Classification Algorithms

Moreover the time taken to build the model and classified instance based performance also evaluated and analyzed. The classified instances based performance and the time efficiency that is the time taken to build these models (in seconds) is defined in Table- VI.

Table VI: Time and Classified Instances Based Performance

Number of Instances	Classification Algorithms	Time (in Seconds)	Correct Classification Instances	Misclassification Instances
370	Logistic Regression	0.20	186	184
	Decision Tree	0.38	189	181
	Naïve Bayes	0.40	182	188

According to Table- VI, the Logistic regression model can do better performance as compared with other models. Fig. 4 shows the graphical representation of classification algorithms based time efficiency. From Table- V and Table- VI, we can observe that Logistic regression classification algorithm has better performance when compared to the other classifications.

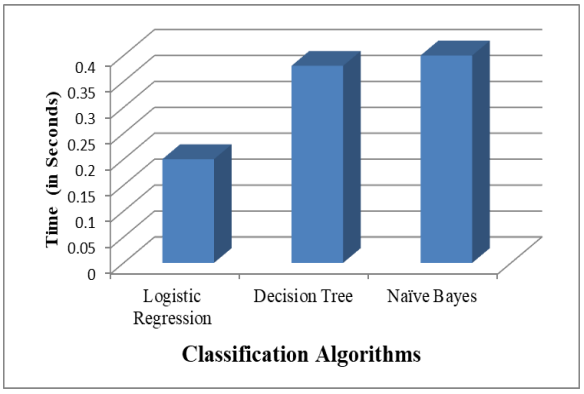

Fig. 4 Time Efficiency of Classification Algorithms

V. CONCLUSION

The diagnosis of diabetes mellitus is its early stage is an important in real life and challenges to the medical field. In this paper, an attempt has been made to classify the diabetic mellitus from other people using classification method adhere by different algorithms. Throughout this paper, the three classification algorithm has been studied and the performance can be evaluated based on various measures. As result determines the Logistic regression classification algorithm has achieved 99.18 % of accuracy and 0.20 seconds is take the time to build this model. Moreover it can be classified all instance in correctly. The overall conclusion, observed from the figure and tables, is that logistic model based algorithm is found to be more efficient for the correct classification when compared to that of other methods. Further, more perfect classification can be done by incorporating the risk level and optimal machine learning technique to disseminate the diabetes mellitus patients from the common people.

REFERENCES

1. Han. J, Jian Pei, Micheline Kamber, "Data Mining Concepts and Techniques", Third Edition, Elsevier Inc, 2012.
2. Vijayan, V. V.,Anjali, C., "Prediction and diagnosis of diabetes mellitus a machine learning approach", *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2015, 122–127.
3. Aishwarya, R., Gayathri, P., Jaisankar, N., "A Method for Classification Using Machine Learning Technique for Diabetes", *International Journal of Engineering and Technology (IJET)*, 5, 2013, 2903–2908.
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research", *Computational and Structural Biotechnology Journal*, 15, 2017, 104–116.
5. Iyer, A., S, J., Sumbaly, R., "Diagnosis of Diabetes Using Classification Mining Techniques", International *Journal of Data Mining & Knowledge Management Process*, 5, 2015, 1–14.
6. Orabi, K.M., Kamal, Y.M., Rabah,T.M., "Early Predictive System for Diabetes Mellitus Disease", *Industrial Conference on Data Mining, Springer*, 2016, pp. 420 – 427.
7. Priyam, A., Gupta, R., Rathee, A., Srivastava, S., "Comparative Analysis of Decision Tree Classification Algorithms", *International Journal of Current Engineering and Technology*, Vol.3, 2013, 334–337.
8. Kumar, P. S., Umatejaswi, V., "Diagnosing Diabetes using Data Mining Techniques", *International Journal of Scientific and Research Publications*, 7, 2017, 705 –709.
9. Fatima, M., Pasha, M., "Survey of Machine Learning Algorithms for Disease Diagnostic", *Journal of Intelligent Learning Systems and Applications*, 09, 2017, 1–16.
10. M. Priya and M. Karthikeyan, "A Comparative Study Of Clustering Algorithms For Outlier Identification", *International Journal for Research in Engineering Application & Management (IJREAM)*, Vol-04, Issue-07, 2018, pp. 391-396.
11. Tarik A. Rashid, S. M. A., Abdullah, R. M., "An Intelligent Approach for Diabetes Classification, Prediction and Description", *Advances in Intelligent Systems and Computing*, 424, 2016, 323–335.
12. Priya. M and M. Karthikeyan, "Performance Evaluation of Ensemble Method Based Outlier Detection Algorithm", *International Journal of Research in Advent Technology*, Vol.7, No.3, 2019, pp. 1376 - 1380.
13. Priya. M and M. Karthikeyan, "An Efficient Cluster Based Outlier Detection Algorithm", *Journal of Engineering and Applied Sciences*, volume: 14, Issue 23, 2019, PP. 8699-8704.
14. World Health Organization, "Action plan for the global strategy for the prevention and control of non-communicable disease", Geneva: WHO; 2008 - 2013.
15. Mandal, S, Dubey V., "Implementation and evaluation of diabetes management system using clustering techniques", *Special issue, International Journal of Computer science and Informatics*, 2(2): 33 -36.
16. Ioannis K, Olga T, Athanasios S , Nicos Maglaveras, Ioannis V , Ioanna C, " Machine Learning and Data Mining Methods in Diabetes Research " *Computational and Structural Biotechnology Journal*, 15, 2017,104–116.
17. M. David, W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", *School of Informatics and Engineering Flinders University, Australia, Technical Report*, 2007.

AUTHORS PROFILE

M. Priya is **c**urrently working in as an Assistant Professor in the Department of Computer Science at PSPT MGR Government Arts & Science College, Sirkali – Puthur -609108. She has received her Bachelor's and Master Degree in Computer Science from Bharathidhasan University, Tiruchirappalli and M.phil., from Annamalai University. She is presently pursuing Ph.D in Computer Science in Annamalai University and area in the research of Data Mining. Her research interest includes Data Mining and Big Data analytics**.** She is published more than 5 research articles published in UGC and Scopus indexed journal**.** She has been participated more than 20 National and International seminars / conference / workshops and also presented more than 15 papers for various National and International level conferences.

Dr. M. Karthikeyan working as Assistant professor in Division of Computer and Information Science, Annamalai University, India. He completed his M.Sc. [Computer Science] from Bharathiar University and M.Phil [Computer Science] and Ph.D from Annamalai University in 2005 and 2014 respectively. He is having 19 years of teaching experience. His area of interest is Data Mining, Digital Image Processing, and Artificial Neural Networks. He has published more than 25 research papers in various reputed journals and conferences. And also more than 25 papers presented in various international and national level seminars and conferences. Under his guidance more than 10 students completed the M.Phil. Degree and is presently 8 scholars pursuing Ph.D. under his guidance.