Prediction of Diabetes Disease using Data Mining Classification Techniques

Shahzad Ali
Department of Computer Science
National Textile University
Faisalabad, Pakistan
Shazadali039@gmail.com

Dawood Saddique
Department of Computer Science
National Textile University
Faisalabad, Pakistan
dawoodsaddique1997@gmail.com

Muhammad Usman Aslam
Department of Computer Science
National Textile University
Faisalabad, Pakistan
usmanaslam402@gmail.com

Muhammad Usman
Department of Computer Science
National Textile University
Faisalabad, Pakistan
nahaing44@gmail.com

Umair Maqbool
Department of Computer Science
National Textile University
Faisalabad, Pakistan
umairmaqbool.007@gmail.com

Shoaib Ejaz
Department of Computer Science
National Textile University
Faisalabad, Pakistan
shoaibejazabc@gmail.com

***Abstract*— Diabetes is one of the chronic diseases in which the blood sugar or blood glucose level is above a certain amount in the body. It is often known as the silent killer because of its easy-to-miss symptoms of the Diabetes Disease (DD). Gestational diabetes is a type of diabetes which occurs in women during their pregnancy and can cause potential health issues for both the mother and the child. The classification of the DD is essential to improve the quality of life of patients suffering from the disease. The primary objective of this research work is to identify the most dominant feature for the DD and to classify the DD for its early diagnosis. Data mining and machine learning (ML) techniques including Naive Bayes, Artificial Neural Network (ANN), Decision Tree (DT), Logistic Regression, and Support Vector Machine (SVM) are used to predict the DD. Pima Indian Diabetes (PID) dataset is used in this experimental investigation, and the performance of the developed models is evaluated using various performance evaluation matrices. The results indicate that the proposed methodology successfully classifies the DD as compared to techniques used in the past. The result also revealed that the proposed methodology could be successfully used in medical and healthcare centers for the classification and early diagnosis of the DD.**

Keywords*— *Diabetes Disease; Classification; Data mining; Logistic Regression

I. INTRODUCTION

Diabetes throughout the years in one of the chronic and significant issues of today's society health care problems. In the diabetes condition, the amount of glucose is above a certain amount in the body. In most industrialized nations, there is substantial evidence that diabetes is the fourth leading cause of death [1]. Diabetes disease occurs typically when a person's body is not able to respond to insulin or exceed the limit of insulin required to maintain the glucose rate in the body. Diabetes has different stages, and every stage has its side effects. DD leads to several other diseases, i.e., blood pressure, heart disease, blindness, kidney failure, and nerve damage [2]. The data attributes studied for the research purpose is to contain the data of pregnant women having diabetes. Pregnant women with insulin-dependent diabetes mellitus have a high risk of getting a chronic disease. The study is carried out to extract the factor which women more is pregnancy [3]. Disorder of glucose tolerance is gestational diabetes, which diagnosed in women during their pregnancy period. There is no role of insulin in this scenario. This disease is playing a cardinal role in health issues throughout the world. Gestational diabetes mellitus (GDM) affects up to 1% to 25% of all pregnancies globally [4], and it has a rapidly increasing rate. While the high blood glucose of GDM usually resolves after delivery, women with GDM have an increased risk of further episodes of GDM [5] and are seven times more likely to develop type 2 diabetes mellitus [6]. This concept is highlighted by the World Health Organization (WHO) [7]. The working done was not only to treat the physical symptoms but also instilling the positive mental state [8].

Machine learning approaches are used to find useful patterns within the datasets. Using the approaches, the primary goal is to discover the knowledge which is not valid and accurate but is also comprehensible and can be used for well fair of society. A medical diagnosis is always a classification problem. Classification is one of the most widely used data mining and machine learning (ML) technique in the medical and healthcare centers. There is an extensive hub of different algorithms and techniques used in data mining and machine learning approach specifically for supervised ML techniques. Thus, the selection of the most suitable algorithm or techniques has been a challenge for investigators in implementing the DD-detection and early diagnosis systems [9].

In this investigation, we proposed different data mining and ML classification algorithm to train the model. We will check the efficiency between the different algorithms and proposed the best one who extracts the most accuracy for the classification of the type of diabetes. Comparing these efficiencies provide us to deal with the disease in a better way to improve the quality of life of the patient suffering from diabetes. To extract the maximum efficiency, the correlation between attribute is measured, and the attributes interlinked are

analyzed. The interlinkage of attributes extends the view to have a look at all the variables and analyzing whether these dependencies affect the results or not. The purpose of the whole research work is to improve the quality of life.

The remaining part of the paper is organized as follows: Section II reviews the literature work. Section III describes the dataset used in this investigation. Section IV briefly describe the proposed methodology including data preprocessing, feature selection, and classification techniques for DD-detection and classification. Section V have some results and outcomes of the algorithms and discussion on the algorithms, and the conclusion of the paper is given in Section VI.

II. Literature Review

Diabetes was first documented by Hey-Ra an Egyptian physician. As people learn more about this disease, they move toward the cure and with the growth of information technology and its application in medical science it becomes easy to deal with such problems. Classification of diabetes using different algorithms has made us know to more knowledgeable results. Polat et al. [10] used SVM and discriminant analysis for the classification of diabetes disease. Kayaer and Yldrm [11] used Regression using Artificial Neural Network (ANN) for the classification of diabetes.

Dogantekin et al. [12] developed an intelligent diagnosis system, the LDA-ANFIS, for diabetes using the Neuro-Fuzzy system and Linear Discriminant Analysis (LDA) classification method. Some researchers have applied Multilayer neural network (ML-NN) as a training algorithm and give efficient results in the classification of the DD [13].

Pasi Luukka [14] used Fuzzy entropy measure (FEM) and similarity classifier for the classification of diabetic disease. A fuzzy entropy based feature selection are used for the classification in that selected feature. H. Hasan Orkcu, Hasan Bal [15] compares the performance of back propagation and GA for the classification of data. Since Back propagation is used for the efficient training of data in Artificial neural network (ANN) but contains some error rate, hence GA is implemented for the binary and real-coded so that the training is efficient and some features can be classified.

III. Data Description

The dataset used in this analysis is the Pima Indian Diabetes (PID) dataset which is obtained from the UCI machine learning repository. This dataset is used several times for experiment purposes. The Pima is one of the most studied population regarding diabetes, not only in American Indian but in the entire world [16]. The dataset consists of a total of nine attributes including eight input attributes and one target attribute. The eight input attributes included in the dataset are Pregnancies, Glucose, blood pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, age, and one class attribute. Table I. illustrates the description of the attributes in the PID dataset. There is a total of 768 instances in the dataset. Several constraints are applied to the dataset for the selection of instance from a larger dataset.

The dataset is divided into a training dataset and test dataset with a proportion of 70% of training data and 30% of the test dataset. The training data is further divided into validation dataset using 10-fold cross-validation to avoid the overfitting problem in the training of the data mining classification algorithms. The class variable is divided into two classes 0 (means negative; healthy person) or 1 (means positive; diabetic person) with the indicating value of negative for a healthy person and positive for a diabetic patient.

Fig. 1. Illustrates the interdependences of class attribute on other attributes of the PID dataset. The graph propagates that the occurrence of diabetes is highly dependent on the glucose level test of patient women. However, BMI is the other factor which causes diabetes as well, and the age factor also plays a role in having diabetes in pregnant women. The main concern of diabetes depends on glucose with the highest correlate on the factor of 0.4621.

TABLE I. Description Of Attributes Of PID Dataset Used In This Investigation

Sr.	Attribute	Description
1.	Pregnancies	Pregnancies attribute has the number of times women get pregnant. This data attribute is numeric ranges from 1-17.
2.	Glucose	This attribute is about Plasma glucose concentration 2 hours in an oral glucose tolerance test.
3.	Blood Pressure	Blood Pressure contains the diastolic blood pressure of the pregnant woman in mm Hg and ranges from 0-122 with a mean of 69.1.
4.	Skin Thickness	This attribute has measures of body fats on right arm halfway
5.	Insulin	Insulin is a hormone produced by a beta cell of pancreatic. The value of Insulin ranges from 0-846.
6.	BMI	The measure of body mass divided by the square of body height. It is measured in Kg/ (height in m)2. The value of BMI for the data set ranges from 0-67.1.
7.	Diabetes Pedigree Function	It is the data on diabetes patient history.
8.	Age	Age attributes contain the data of pregnant women elder than 21 years. The value of age ranges between 21-81 years old.
9.	Class (outcome)	Class attribute (i.e., Healthy or diabetic person)

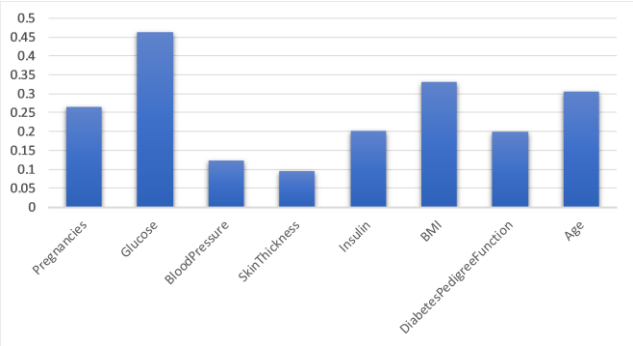

Fig. 1. Interdependence of each attribute on the Class attribute

IV. Methodology

The section describes the fundamental aspects of some of the most popular supervised data mining and machine learning classifiers including Naive Bayes (NB), Decision Tree (DT), artificial neural network (ANN), Support Vector Machines (SVM), and Logistic regression (LR). These algorithms have different capabilities in term of classification accuracy and speed and time of computation. The results of these algorithms have their significance and accuracy levels, and their comparison is essential to check which one to use and in what conditions. The schematic diagram of the proposed methodology is summarized in Fig. 2.

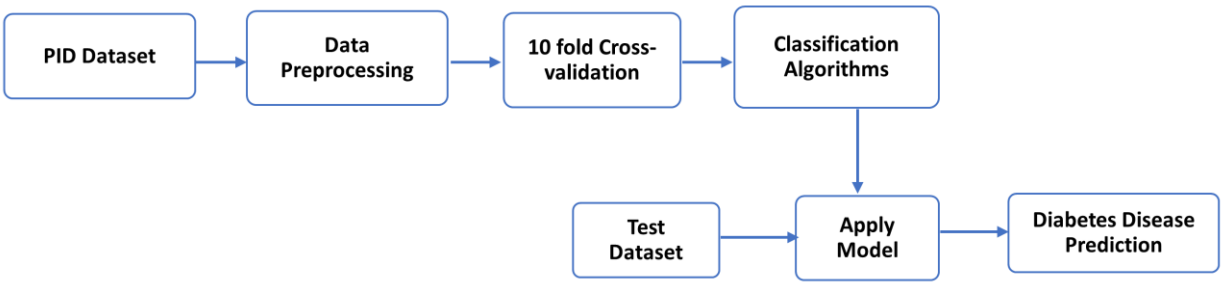

Fig. 2. The flow of the proposed methodology

A. *Naive Bayes*

Naive Bayes is a probabilistic learning technique. It is based on the Bayes theorem which is particularly suited for high dimensional inputs and deal with classification methods. Naive Bayes require some linear parameters against variables and classifier in a learning problem. It provides the easiest and fastest way. For independence dataset, Naive Bayes is better than logistic that need less training data. It is not a single algorithm but is a collection of algorithms where all of them share a universal principle. It provides a better result in the case of a categorical input variable as compared to numeric [17].

B. *Random Forest*

Random forest as the name suggests is a decision-based algorithm which is one of the famous learning technique used as both classification and regression problems. It helps in making a decision tree and prevents the habit of data to get overfit. Random forest tree algorithm also controls the performance of non-linearity. This technique deals with both dependent and independent data. Both regression trees and classification tree are used to deal with variables. It can deal with the missing value for missing data [18].

C. *Support Vector Machine*

SVM is a methodology based on supervised learning with abilities to classify data on similarity bases and make an algorithm better. It is most precise than other ML algorithms based on supervised classification. For text classification, it is the most accurate method among other technique. SVM is a linear machine learning methodology that has optimally recognized hyperplane to separate the data into classes. SVM maximize training point distance closest from either class and achieve improved generalization performances on test data. To make data separable linearly for linear SVM formulation, input data transformed into high dimensional feature spaces. Usually, the kernel function helps to achieve transformation [1].

D. *Logistic Regression*

Logistic regression is the best approach to use when the regression analysis is done on a binary dependent variable. Logistic regression is used for predictive analysis and is used to describe the data. It explains the relationship between the dependent variable of type binary and the independent variable of type nominal, ordinal, interval or ratio-level [20].

E. *Artificial neural network*

Artificial neural network (ANN) is part of an extensive family of ML techniques based on learning patterns, as opposed to problem-specific techniques and algorithms. ANN helps to achieve accuracy and exceed human performance level. It requires a large amount of labeled data and high computational powers. It is self-learning techniques where the model is trained by layered network architecture and data labeling. First of all, ANN is trained and tested until the error in prediction is minimized, and the different types of inputs are given to predict the output. This method helps to get rid of manual feature selections. One of the common ANN algorithms in a conventional neural network [21].

TABLE II. Confusion Matrix For Logistic Regression on The Testing Dataset

	True Negative	**True Positive**
Pred. Negative	128	37
Pred. Positive	17	52

V. Results and Discussion

Table II. Illustrates the comparison of the results of different classification algorithms applied to the dataset. The performance of a diagnostic problem is evaluated in term of Accuracy, Root Mean-Squared Error, correlation, squared correlation, Root Relative-Squared Error, kappa Statistics, and some other factors which are shown below. The values of this parameter are extracted using algorithms like Naive base, SVM, Decision Tree, Random Forest, Logistic regression and ANN.

These factors demonstrate the overall performance of data on the applied algorithm. So logistic regression is considered as the best machine learning algorithm in this experiment with an accuracy rate of 77.78% higher than other classifiers.

Table. III elaborates the prediction of the linear regression model. The model shows that 128 negative instances out of 145 are predicted as negative correctly and 17 instances are predicted as positive. While 37 instances out of 89 positive instances are predicted as negative and 52 are predicted as true positive.

TABLE III. Comparison Of The Results Of Different Classification Algorithms

Performance Measure	**Validation Set**					**Testing Set**				
	NB	***ANN***	***DT***	***LR***	***SVM***	***NB***	***ANN***	***DT***	***LR***	***SVM***
Accuracy	75.10%	74.53%	73.41%	76.97%	77.34%	74.36%	72.22%	74.36%	77.78%	76.92%
Kappa Statistics	0.430	0.453	0.341	0.454	0.461	0.439	0.436	0.327	0.509	0.488
Root-Mean-Squared Error	0.420	0.401	0.444	0.397	0.397	0.423	0.413	0.471	0.406	0.406
Root-Relative-Squared Error	0.420	0.602	0.667	0.598	0.597	1.112	1.085	1.239	1.068	1.067
Logistic_loss	0.410	0.411	0.422	0.413	0.420	0.410	0.415	0.428	0.415	0.421
correlation	0.433	0.457	0.360	0.461	0.469	0.443	0.444	0.338	0.517	0.497
Squared_correlation	0.185	0.208	0.129	0.212	0.220	0.196	0.`97	0.114	0.267	0.247
weighted_mean_recall	71.02%	73.64%	65.46%	71.46%	71.74%	71.28%	72.81%	65.45%	74.48%	73.35%

VI. Conclusion

Diabetes detection in its early stages is one of the world leading health problems. This study shows that systematic efforts are made in designing a system which results in the prediction of diabetes. In this analysis, five data mining and machine learning algorithms are used for diabetes disease classification purposes. Experimental results show that the accuracy of logistic regression on the PID dataset is 77.78% which is the best among the other algorithms applied in this analysis. The results of the analysis proved that the proposed methodology successfully classifies the diabetes disease as compared to techniques used in the past. The result also revealed that the proposed methodology could be successfully used in medical and healthcare centers for the classification and early diagnosis of the diabetes disease.

ACKNOWLEGMENT

We are thankful to our instructor, Dr. Sohail Jabbar for his insight guidance, and suggestions during this research work that is part of our undergraduate subject. He motivated and encouraged us, helped us in gathering dataset to complete this work.

References

[1] Gan, D. (Ed.), 2003. Diabetes Atlas, 2nd ed. Brussels: International Diabetes Federation. (accessed 19.06.11).

[2] Mostafa Fathi Ganji, Mohammad Saniee Abadeh "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease" Proceedings of ICEE 2010, May 11-13, IEEE 2010.

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1595250/, Accessed on 18 Jan 2019.

[4] Zhu Y, Zhang C. Prevalence of gestational diabetes and risk of progression to type 2 diabetes: a global perspective. Curr Diab Rep. 2016;16(1):7.

[5] Kim C, Berger DK, Chamany S. Recurrence of gestational diabetes mellitus. Diabetes Care. 2007;30(5):1314–1319.

[6] Bellamy L, Casas J-P, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. Lancet. 2009;373(9677):1773–1779

[7] World Health Organization, Construction in Basic Documents, World Health Organization, Geneva, Switzerland, 1948.

[8] P. Bech, D. Carrozzino, S. F. Austin, S. B. Møller, and O. Vassend, "Measuring euthymia within the Neuroticism Scale from the NEO Personality Inventory: a Mokken analysis of the Norwegian general population study for scalability," Journal of Affective Disorders, vol. 193, pp. 99–102, 2016.

[9] O. Erkaymaz, M. Ozer, Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes, Chaos, Solitons, and Fractals 83 (2016) 178-185.

[10] K. Polat, S. Gene?, A. Arslan, A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine, Expert systems with Applications 34 (2008) 482-487.

[11] K. Kayser, T. Yildirim, Medical diagnosis on Pima Indian diabetes using general regression neural networks, Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), 2003, pp. 181-184.

[12] E. Dogantekin, A. Dogantekin, D. Avci, L. Avci, An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive-network-based fuzzy inference system: LDA-ANFIS, Digital Signal Processing 20 (2010) 1248-1255.

[13] Matlab documentation. (2004). Version 7.0, release 14. The MathWorks Inc. Mohamed, E. I., Linderm, R., Perriello, G., Di Daniele, N., Poppl, S. J., & De Lorenzo, A. (2002). Predicting type 2 diabetes using an electronic nose-base artificial neural network analysis. Diabetes Nutrition & Metabolism, 15(4), 215–221.

[14] P. Lukka, ―Feature Selection using fuzzy entropy measures with similarity classifier‖, Elsevier: Expert Systems with Applications, vol. 38, (2011), pp. 4600-4607.

[15] H. Hasan Orkcu and H. Bal, ―Comparing Performances of Backpropagation and Genetic Algorithms in the Data Classification‖, Elsevier: Expert Systems with Applications, vol. 38, (2011), pp. 3703-3709

[16] M. Awad, Y. Motai, J. N. Janne, H. Yoshida, A clinical decision support framework for incremental polyps classification in virtual colonoscopy, Algorithms 3 (2010) 1-20.

[17] Siddique, Aieman Quadir., Hossain, Md. Saddam. (2013) 'Predicting Heart-Disease from Medical Data by Applying Naive Bayes and Apriori Algorithm,' International Journal of Scientific and Engineering Research (IJSER), Vol. 4, Issue 10.

[18] Ziegler, A, and König, I.R. (2014). Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 1, pp. 55–63.

[19] M. Farhan, M. Aslam, S. Jabbar, and S. Khalid, "Multimedia based qualitative assessment methodology in eLearning: student teacher engagement analysis," Multimed. Tools Appl., pp. 1–15, 2016.

[20] A. Paul, A. Ahmad, M. M. Rathore, and S. Jabbar, "Smartbuddy: Defining human behaviors using big data analytics in social internet of things," IEEE Wirel. Commun., vol. 23, no. 5, pp. 68–74, 2016.

[21] M. Farhan et al., "A Real-Time Data Mining Approach for Interaction Analytics Assessment: IoT Based Student Interaction Framework," Int. J. Parallel Program., 2017.