IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Ensemble Classification Model for Diabetes Prediction in Data Mining

Munendra Kumar[1*], Anuj Kumar[2]

[1,2] IEC College of Engineering & Technology, AKTU, Greater Noida, India

Corresponding Author: munedrakumar86@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.16431647 | Available online at: www.ijcseonline.org

Accepted: 27/May/2019, Published: 31/May/2019

Abstract**—** The prediction analysis is the approach which can predict the future possibilities based on the current information. The diabetes prediction is the approach which is applied to predict the diabetes based on the various attributes. The diabetes dataset has various attributes and based on that attributes diabetes can be predicted. In the previous year's approach of SVM is applied for the diabetes prediction. To improve accuracy of diabetes prediction voting based classification is applied in this paper. The proposed model is implemented in python and results are analyzed in terms of accuracy, execution time.

Keywords**—** Diabetes, SVM, Voting

I. INTRODUCTION

The procedure which is used for the extraction of useful information from the unprocessed information is called data mining. The extraction of huge amount of information is required to attain necessary knowledge. In the recent times, huge amount of data exists in almost every field. The inspection of this data is extremely complex because it needs a lot of time. The existing data is in rough format and therefore it cannot be used directly. A proper data mining technique is required for the extraction of knowledge [1].

The procedure used for the extraction of raw matter is named as mining. In this world, a lot of data signifies authority and achievement. This knowledge or information can be obtained with help of some refined technologies such as satellites, computers etc. The expansion of technology in the field of digital storage and computers had made the handling of massive data easy with different kinds of stored information. The prediction technique detects the correlation among independent variables and dependent variables in the data mining process. This method can be used in several regions for the prediction of prospect benefit. Thus, dependent variable is denotes profit whereas independent variable represents sale. With the help of a fitted regression curve, past sale and profit information has been used to predict profit [2].

The other name given to diabetes disease is Diabetes Mellitus (DM). In this disease, human body cannot process food appropriately for its utilization as power. In data mining, a significant role is played by data mining. In this study, diabetes predictions are analyzed. Generally, four types of Diabetes Mellitus are present i.e. Type1, Type2, Gestational diabetes, congenital diabetes. Type 1 diabetes initiates in early days. In this situation, the immune system of body damages the insulin excretion cells and thereby eliminates the generation of insulin from body. Without insulin, body cells are not able to absorb sugar in the absence of insulin which is required for the production of power [3].

Type 2 diabetes can start at any age and generally diagnosed in middle age. This type of diabetes can be eliminated or belated with a healthy way of life, by maintaining healthy weight with habitual work out. The gestational diabetes generally stars during pregnancy. It is frequently detected in middle or delayed pregnancy phase. In a mother, high blood sugar levels are distributed via placenta to the baby. This type of diabetes must be restricted for the protection of baby's development and maturity. This diabetes causes huge danger to mother and unborn infant [4].

Classification is the process of identifying a new observation category set on the basis of training set of data that contains observation whose category is known. The process used to identify a new monitoring group by means of training data suite is called classification. The objects are grouped on the basis of their resemblance using Cluster analysis method. K-means clustering algorithm is described as the fundamental partitioning dependent algorithm. This algorithm performs several clustering tasks for the execution of operation using low dimensional data suites. K represents parameters and the k clusters are generated through the partitioning of n objects. Identical objects are clustered in single cluster while different objects are located in the separate clusters. The cluster centres can be identified using this algorithm. It is essential to decrease the sum of the squared distances to the closest centre of cluster at every data point**.** Quadratic discrimination is the most widespread approach which is used for the supervised parametric classifiers hypothesis. These classifiers get the decision edges for dealing with d-dimensions in the last and make the process complicated. In order to generate computations, whole discriminant functions are performed in log-off mode. This approach is affected further because of dimensionality. This approach has quadratic discriminant in various parameters and their proper handling is imperative. The small training samples severely affect the performance of this approach. The multi-layer perceptron classifier is the fundamental step in the artificial neural network. In order to simple the process, a single hidden layer has been utilized. Primarily, a single concealed layer is used for the simplification of this procedure. After this, two hidden layers are used for the improved performance of classification. The hidden elements are selected in different way for every data suite. The hidden neurons are discovered after several attempts. The thumb rule is used for the detection of total number of hidden neurons. The obtained total net weight is approximately n/10, where n represents total number of training points. The neural network has been trained using back-propagation algorithm [5].

Support vector machine is also called as SVM. This is a classification algorithm. This algorithm relies on optimization hypothesis. This algorithm is also known as binary classifier because it maximizes the margin. Most appropriate hyperplane is used for the separation of all data points of an individual class. This can be detected with the help of the classification offered by support vector machine classifier. In SVM, the largest margin describes the biggest and finest hyperplane between two classes. The maximum width of the slab equivalent to the hyperplane having no interior data points is called margin. The SVM algorithm separates the maximal margin in hyperplane. Fuzzy logic is considered the best method for the mapping of input space to output space. This functioning of these techniques depends upon the learning and variation ability. The fuzzy system comprises several suites of structure for the diabetes diagnosis procedure. The most common computing framework is called Fuzzy Inference System (FIS). This system relies on the basic fuzzy set theory, fuzzy if- then rules and fuzzy reasoning. This fuzzy inference system was designed according to the past acquaintance of the targeted system. Thus, the fuzzy system replicates the behavior of the targeted system. This rule becomes complicated because of the utilization of huge quantity of inputs and outputs [6].

II. Literature Review

Han Wu, et.al (2018) proposed a new approach for the prediction of type 2 diabetes mellitus (T2DM). The proposed model was based on data mining methods [7]. On the basis of sequence of preprocessing measures, the model was made up of two parts. These parts included enhanced K-means algorithm and the logistic regression algorithm. The obtained outcomes and outcomes gathered from different researchers were compared using Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit. The comparative analysis showed that the proposed approach achieved 3.04% larger forecasting accuracy in comparison with other approaches. Furthermore, the proposed approach ensured the sufficiency of data suite quality. For further evaluation of the performance of proposed approach, this approach was used in two more diabetes data suites. Good performance was shown through both tests.

Prova Biswas, et.al (2018) proposed a novel approach for the evaluation of glucose masses and insulin masses in the different organs of body through just plasma glucose estimation [8].

The proposed approach used a combination of glucose-insulin homoeostasis model (in the presence of meal intake) and plasma glucose measurement with a Bayesian non-linear filter. The process noise was added to the homoeostasis model for including the ambiguity of the model over particular variations. The evaluation was performed for the fit human beings and type II diabetes mellitus patients. A comparison was performed among the performances of two linear filters on the basis of root mean square error. In future, this approach can be used to compute the medicine dosage for any hyperglycaemic patients and the development of a closed-loop controller for automatic insulin release scheme.

Ioannis Kavakiotis, et.al (2017) stated that various techniques like data mining and machine learning proved very beneficial in the prediction of diabetes. These approaches could prove beneficial in many other ways as well [9].

A lot of researches have already been carried out in the field of Diabetes mellitus research and especially in biomarker identification and prediction-diagnosis. Support vector machine classifier was considered best among the existing machine learning algorithms. Clinical data suites could be utilized on different data types. Various investigations conducted on DM proved that new methodologies were extremely helpful for the prediction of diabetes mellitus patients.

Zhiqiang Ge, et.al, (2017) presented an analysis of earlier data mining and analytics applications. This analysis was carried out on the process industry over some previous decades [10].

Eight unsupervised learning algorithms, ten supervised learning algorithms and the application status of semi-supervised learning algorithms was used for the inspection of state-of-the-art of data mining approaches. In this study, numerous viewpoints were considered. Further, these view points were conferred for future studies on data mining and diagnostics in the process industry.

Alexis Marcano-Cede˜no, et.al (2016).stated that diabetes caused several damages to the different organs of human body [11]. It was identified that diabetes records were used by means of appropriate understanding for the analysis of diabetic disease and this was the main issue of classification. Various methods were implemented on the basis of artificial intelligence for the removal of diabetes concerns. The detection of diabetes using artificial metaplasticity on multilayer perceptron (AMMLP) in the form of data mining approach was the major aim of this study. The Pima Indians diabetes data suite was utilized for evaluating the performance of proposed approach. The outcomes of proposed approach were compared with various algorithms and classifiers. The strength of proposed approach was measured in terms of certain parameters such as accuracy, sensitivity, specificity, confusion matrix and 10-fold cross-validation technique. The tested outcomes depicted that proposed approach showed better performance in comparison with existing approaches.

Bayu Adhi Tama, et.al (2016) stated that diabetes was a chronic disease. This disease became the reason of several deaths all across the world [12].

Approximately 285 million people worldwide were suffering from diabetes according to a survey conducted by International Diabetes Federation (IDF). This range and data could increase in nearby future due to the absence of a suitable diabetes prevention technique. Type 2 diabetes (TTD) was the most common type of diabetes. The discovery of all effects of Type 2 diabetes was not a simple job and this was the main concern. Thus, data mining was utilized because it provided the finest outcomes and supported the knowledge discovery from data. In the data mining procedure, support vector machine (SVM) was used for the extraction of patients' data from earlier reports. The timely recognition of TTD provided help to take effectual decision.

Aiswarya Iyer, et.al (2015) offered a real world medical issue related to the automated diagnosis of diabetes. The timely recognition of diabetes was its main therapy [13].

In this study, some classifiers named as Decision Trees and Naïve Bayes were used for the real analysis of diabetes in modeling of local and systematic healing. The functioning of proposed approach was analyzed for the issues related to the diagnosis of diabetes. The tested outcomes depicted the sufficiency of the proposed approach. The proposed approach performed better in comparison with various other existing approaches.

III. Research Methodology

This research work is based on the diabetes prediction. The diabetes prediction technique has the three phase which are pre-processing, feature extraction and classification. The phases of the proposed technique are explained below:-

1. Data Collection and pre-processing:- The dataset of diabetes will be collected from the UCI repository. The dataset get pre-processed and missing, redundant data will be removed from the dataset

2. Feature Extraction- In the second phase, technique of feature extraction will be applied in the network. The feature extraction technique will establish relationship between the attribute and target set. The PCA algorithm is applied on the input dataset which will simply the dataset. The principle component analysis (PCA) algorithm will select the most frequent attribute from the dataset for the prediction

3. Classification: - The classification approach is the last phase of the prediction analysis. In the classification approach, input dataset will be divided into training and test sets. The training set will be 60 percent of the whole dataset and test set will be 40 percent. The approach of voting based classification is applied for the diabetes prediction. In the voting based classification, three classification methods are applied for the prediction analysis. The KNN is the first classification approach which is applied for the diabetes prediction. In K-means clustering algorithm is defined as the basic partitioning based method in which different clustering tasks has been utilized in order to perform function within the low dimensional data sets. K is referred as the parameters and by partitioning n objects the k clusters are generated. Within one cluster similar types of objects are grouped and in the separate clusters dissimilar objects are placed. With the help of this algorithm it is feasible to identify the cluster centres. At each data point it is necessary to reduce the sum of the squared distances to the nearest centre of cluster which is required. The second is the naïve bayes classifier in which The most generalized approach for the supervised parametric classifiers theory is quadratic discrimination. These classifiers obtained the decision boundaries when it is required to deal with d-dimensions at the end this process becomes complicated. All the discriminant function has been done off-line for the computational generation. Due to dimensionality this approach is more affected as there is quadratic discriminant in the large number of parameters which is necessary to manage. Its performance is affected drastically by the small training samples. SVM stands for support vector machine and it is a classification algorithm that is based on optimization theory. As it maximizes the margin it is also known as a binary classifier. All the data points of an individual class are separated by the best hyperplane, this can be identified through the classification provided by SVM. In the SVM the largest the best hyperplane is described by the largest margin between the two classes. There are no interior data points when there is maximum width between the slabs parallel to the hyperplane which is also known as margin. The maximum margin in hyperplane is separated by the svm algorithm. The voting based classification approach will vote between three classifiers which are KNN, naïve bayes and SVM. The classifier which give maximum accuracy is considered as final classifier for the prediction analysis.

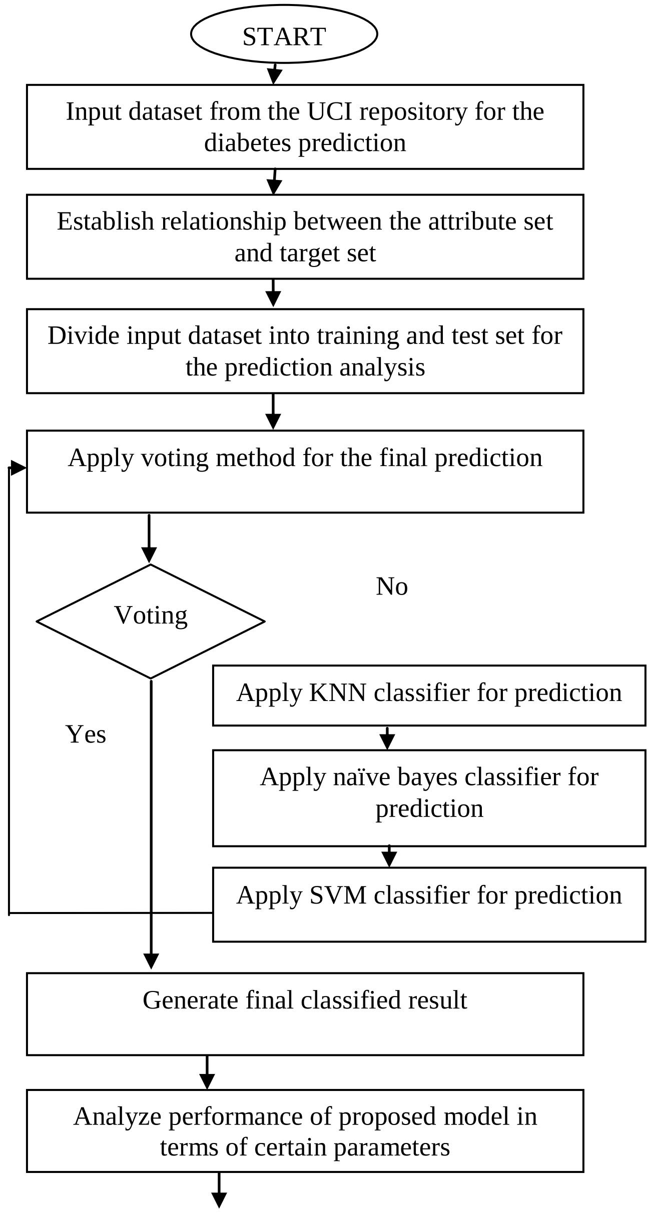
START
Input dataset from the UCI repository for the diabetes prediction
Establish relationship between the attribute set and target set
Divide input dataset into training and test set for the prediction analysis
Apply voting method for the final prediction
Voting
No
Yes
Apply KNN classifier for prediction
Apply naïve bayes classifier for prediction
Apply SVM classifier for prediction
Generate final classified result
Analyze performance of proposed model in terms of certain parameters

Fig 1: Proposed Flowchart

IV. RESULT AND DISCUSSION

The proposed technique for the diabetes prediction is based on the voting based classification. In the existing method, the approach SVM classifier is applied for the diabetes prediction. The results of the proposed and existing method are analyzed in terms of accuracy and execution time.

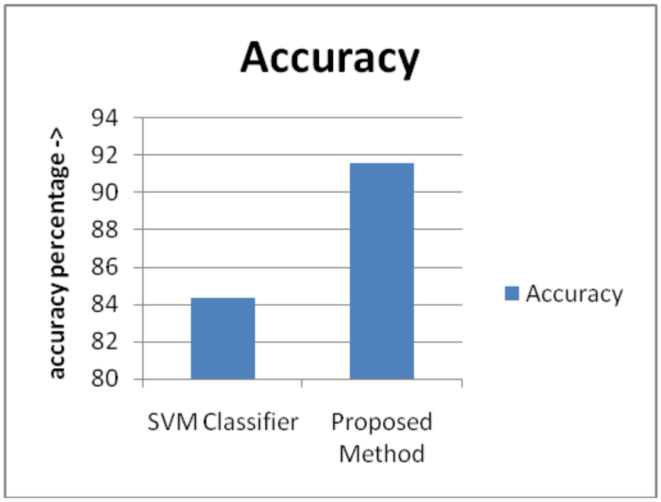

Fig 2: Accuracy Comparison

As shown in figure 2, the accuracy of proposed model is compared with the existing model. The accuracy of proposed model is high due to voting approach

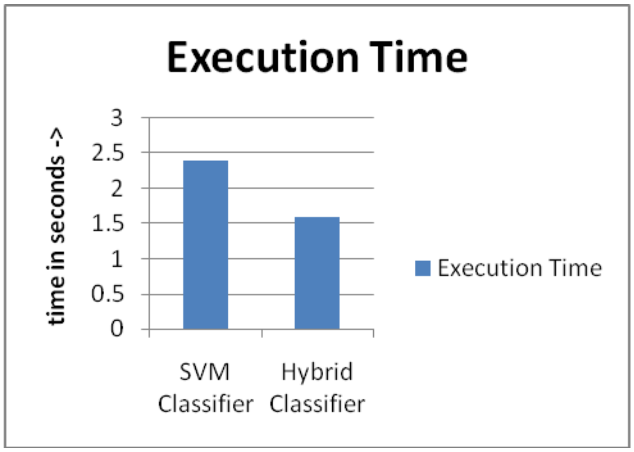

Fig 3: Execution Time

As shown in figure 3, the execution time of proposed model is compared with the existing mode. The execution time of proposed model is less as compared to existing model.

V. CONCLUSION

In this work, it is concluded that diabetes prediction is the complex problem of prediction analysis. The diabetes prediction approach has three phases which are pre-processing, feature extraction and classification. In the previous system SVM classifier is applied for the diabetes prediction. In the proposed technique of voting based classification is applied for the diabetes prediction. The proposed model is implemented in MATLAB and results are analyzed in terms of accuracy and execution time. The accuracy of proposed system is high and execution time is low as compared to existing system.

REFERENCES

[1] Abdelghani Bellaachia and Erhan Guven (2010), “Predicting Breast Cancer Survivability Using Data Mining Techniques”, Washington DC 20052, vol. 6, 2010, pp. 234-239.

[2] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity”, Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.

[3] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, 2017, IEEE, vol. 15, 2017, pp- 215-227

[4] Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), “Clustering of Lung Cancer Data Using Foggy K-Means”, International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.

[5] Kajal C. Agrawal and Meghana Nagori (2013), “Clusters of Ayurvedic Medicines Using Improved K-means Algorithm”, International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, 2013, pp. 546-552.

[6] [10] Chew Li Sa, Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M. (2014), "Student performance analysis system (SPAS)", in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, vol.15, 2014, pp.1-6.

[7] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, “Type 2 diabetes mellitus prediction model based on data mining”, ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.

[8] Prova Biswas, Ashoke Sutradhar, Pallab Datta, “Estimation of parameters for plasma glucose regulation in type-2 diabetics in presence of meal”, IET Syst. Biol., 2018, Vol. 12 Iss. 1, pp. 18-25, 2018.

[9] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal 15 (2017) 104–116

[10] Zhiqiang Ge, Zhihuan Song, Steven X. Ding, Biao Huang, “Data Mining and Analytics in the Process Industry: The Role of Machine Learning”, 2017 IEEE. Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.

[11] Alexis Marcano-Cede˜no, Diego Andina, “Data mining for the diagnosis of type 2 diabetes”, IEEE, Vol. 11, issue 3, pp. 9-19, 2016.

[12] Bayu Adhi Tama, Afriyan Firdaus, Rodiyatul FS, “Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine”, Vol. 11, issue 3, pp. 12-23, 2008.

[13] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, “DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, 2015.

Authors Profile

“Munendra Kumar”

B.Tech(CSE) Graduated in 2011 from UCER United College of Engineering & Research Gr. Noida affiliated to Gautam Buddha Technical University Lucknow (U.P.). and now pursuing his Master degree in Computer Science & Engineering from IEC College of Engineering & Technology Gr. Noida affiliated to Dr. APJ Abdul Kalam Technical University Lucknow (U.P.). He is now active in writing papers and joining conferences.

“Prof. Anuj Kumar”

Presently working as a Assosiate Professor in IEC College of Engineering and Technology, Greater Noida since 2009. He Has 9 Years teaching experience. He has done MCA from GBTU Lucknow. He has done M.Tech(CSE) from Gautam Buddha University Noida. He is pursuing his P.hd(Thesis Submitted) AKTU Lucknow. He has published many research papers in the field of Software Engineering, Software Testing in IEEE