Voting Based Classification Method for Diabetes Prediction

Harwinder Kaur, Gurleen Kaur

Abstract- This research work is based on the diabetes prediction analysis. The prediction analysis technique has the three steps which are dataset input, feature extraction and classification. In this previous system, the Support Vector Machine and naïve bayes are applied for the diabetes prediction. In this research work, voting based method is applied for the diabetes prediction. The voting based method is the ensemble based which is applied for the diabetes prediction method. In the voting method, three classifiers are applied which are Support Vector Machine, naïve bayes and decision tree classifier. The existing and proposed methods are implemented in python and results in terms of accuracy, precision-recall and execution time. It is analyzed that voting based method give high performance as compared to other classifiers.

Index Terms: Voting based method, Support Vector Machine, Naïve bayes, decision tree, diabetes prediction..

I. INTRODUCTION

The process through which important information can be extracted from raw data is called data mining. Extraction of huge amount of data is necessary such that important information can be acquired. Large amount of data is available in every field and huge amount of time is consumed when analyzing this complete data [1]. For extracting the knowledge, data mining process is used such that no extra raw material is included. Mining is defined as the process through which important data is extracted. A prediction technique is used to discover the relationship which exists among the independent and dependent variables [2]. For predicting the future profits, prediction analysis techniques have been used in several fields. A major chronic disease that is being of major health concern all across the world is called diabetes.

When minimal amount of insulin required to maintain the rate of glucose is not outgrown, diabetes can arise [3]. A healthy diet, regular exercise as well as insulin injections are some of the basic methods to control diabetes. Several other problems like heart disease, kidney disease, blindness or blood pressure related problems can arise due to diabetes [4].

Revised Manuscript Received on July 22, 2019.

Harwinder Kaur, University Of Engineering, Chandigarh University,Mohali, India.
Gurleen Kaur, University Of Engineering, Chandigarh University, Mohali,India.

A group of metabolic disorders which result in causing abnormal insulin secretion or action is called Diabetes Mellitus (DM). Around 200 million people all across the globe are affected by DM which is thus known as the most common endocrine disorders. Over the years, it is assumed that the rate of growth of this disease will rise. Diabetes mellitus is broadly categorized into three types [5]. When a body fails to provide insulin, Type 1 Diabetes Mellitus is caused. An insulin resistance that is caused when the cells fail to use insulin properly results in causing Type 2 Diabetes Mellitus. When a pregnant women which was previously diagnosed on diabetes develops a high blood glucose level, Type 3 diabetes called Gestational diabetes occurs [6]. The most commonly found diabetes form which is also known as insulin resistance is T2D. The type of lifestyle, dietary habits and heredity are certain causes of T2d. Weight loss, polyurea [7], and polydipsia are some of the symptoms of DM. Further, depending upon the blood glucose levels, the diagnosis can vary. Mostly because of the chronic hyperglycemia [8], there are various complications seen in case of DM progression. There are several heterogeneous patho-physiological conditions covered by DM [9]. There are micro- and macro-vascular disorders within which mostly all the common complications have been categorized. It is very important to provide prevention and treatment because of the high DM mortality [10], morbidity and relevant disorders [11].

Several data mining techniques can be used to train the medical data [12]. For predicting diabetes, it is important to preprocess the dataset initially and then the missing values are filled. The categorization task is then performed by using the data mining supervised learning algorithms [13]. The hidden patterns can be predicted from the previous history by data mining technique. In medical data mining, the most commonly used technique is classification [14]. The estimation of predictive accuracy of a classifier is done. The number of tests needed to detect the disease can be

reduced with the application of data mining technique. Since diabetes causes various organs to affect when not cured for longer time, it is very important to provide treatments [15].

To ensure that proper treatments are made, efficient techniques are important to be designed. Several diabetes prediction techniques have been proposed by different researchers over the time.

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on the clustering and classification. The diabetes mellitus is the diabetes disorder which is caused due to increase in sugar level in the blood. The machine learning algorithms are the most popular algorithms which are applied for the diabetes prediction. The SVM is the most common and widely used classification algorithm for the diabetes prediction. The diabetes disease dataset are very complex in nature means it has number of attributes due to which SVM classifier give less accuracy. The techniques needs to propose which give high accuracy for diabetes predication

II. LITERATURE REVIEW

Mobile Fikirte Girma Woldemichael, et.al (2018) proposed a study in which the data mining techniques were used to predict the patients suffering from diabetes [16]. J48, naïve bayes, back propagation and support vector machine were some commonly known classifiers used to predict diabetes from patients in the recent research studies. Large value learning rate was included in the neural networks used in these systems such that the performance of systems could be improved. Around 83% of accuracy, 76% of specificity and 86% of sensitivity were achieved as per the performance results achieved when using back propagation algorithm. It showed that in comparison to previously proposed studies, the outcomes of this study were better.

Ioannis Kavakiotis, et.al (2017) reviewed the various machine learning based Diabetes mellitus (DM) detection techniques proposed by different authors over the time [17]. From several clinics and biological fields, the data was collected to create the datasets on which the experiments could be performed. The supervised learning techniques were used in around 85% of the experiments and the remaining ones preferred unsupervised learning techniques. The most widely used and highly successful algorithm was Support Vector Machine (SVM). An in-depth exploration of the various techniques proposed to detect DM was presented in this research which aimed to perform an analysis as to what kinds of improvements could be made in future.

Yu-Xuan Wang, et.al, (2017) proposed a method to design operating system that used the support of data mining and machine learning [18]. With the help of this it becomes possible to discover a new, automatized way by which optimization of the complex algorithms become simple and easy to use. For the validation of the proposed method cache design was utilized that automatically control the replacement of cached contents to make decisions. All the collected data from the system was analyzed when reply is obtained from a data miner. As per performed experiments, it is concluded that proposed method provides effective results.

Zhiqiang Ge, et.al, (2017) presented a review on existing data mining and analytics applications by the author which is used in industry for various applications [19]. To the semi-supervised learning algorithms an application status was given in this paper. In the process of industry both the methods unsupervised and supervised machine learning is widely used for approximately 90%-95% of all applications. In the recent years, the semi-supervised machine learning has been introduced. Therefore, it is demonstrated that an essential role is played by the data mining and analytics in the process of industry as it leads to develop new machine learning technique.

Bayu Adhi Tama, et.al (2016) presented in this paper a chronic disease that causes major causalities in the worldwide that is Diabetes. As per International Diabetes Federation (IDF) around the world estimated 285 million people are suffering from diabetes [20]. This range and data will increase in nearby future as there is no appropriate method till date that minimize the effects and prevent it completely. The major issue was the detection of TTD as it was not easy to predict all the effects. Therefore, data mining was used as it provides the optimal results and help in knowledge discovery from data. In the data mining process, support vector machine (SVM) was utilized that acquire all the information extract all the data of patients from previous records. The early detection of TTD provides the support to take effective decision.

Jahin Majumdar, et.al, (2016) proposed a model in which SVM techniques were used as it provided the accuracy and heavy in the computational functions [21]. The accuracy level of SVM is measured with the help of dataset. In order to improve the data classification and pattern recognition in Data Mining

mainly feature selection various existing approaches were focused and experimented.

As per performed experiments, it is concluded that comparison between the existing possibilities are averaged for predicting the class labels in "soft" voting.

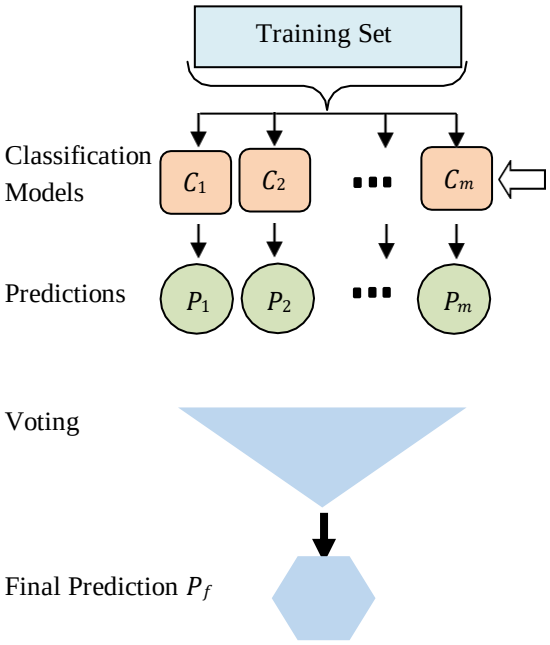


Fig 1: Voting Based Classification Method

III. PROPOSED METHODOLOGY

The diabetes prediction problem is complex in nature due to available of large number of attributes in the dataset. The voting is the classification method which is the combination of several classification methods. Several classifiers are collected to generate one ensemble classifier called the voting based classifier. For exploiting the various peculiarities of each algorithm, they are trained and evaluated in parallel individually. Either "hard" or "soft" voting is implemented in this classifier. The final class label is predicted as the class label that has been predicted most frequently by the classification models in "hard" voting. However, the class The following are the various steps which are performed for the diabetes prediction using voting based method:-

A. Data collection and pre-processing:- In the first phase, the data of diabetes is collected from the UCI repository. The collected data is in structured form and technique of pre-processing is applied to transform the data. The method of random sampler is applied for the data transformation.

B. Feature Extraction:- In the second step, the technique of feature extraction is applied which can establish relationship between the attribute the target set. The feature approach can help to identify the attributes which has maximum impact on the target set.

C. Model Building: - The model building is the third phase, in which whole dataset will be divided into training set and test set. The training set will be more as compared to test set. The model of classification is applied which take input training and test set. The method of voting based classification is applied which is combination of decision tree, support vector machine and naïve bayes classifier for the final prediction. The decision tree is flow-chart-like tree structure in which a test on an attribute is represented by internal node, outcome of test by branch and classes by leaf nodes is called a decision tree classifier. The SVM is statistical learning based classifier in which the decision boundaries are represented by support vectors is called SVM. The training data is represented by identifying the number of support vectors. The model is trained using the only portion of data. Binary classification is used to design the SVM originally. The naïve bayes classifier that is based on the assumption of class conditional independence which states that there are no dependencies among the attributes is called Naïve bayes classifier. It is called "Naïve" since it simplifies the computations involved. The three classifier which are applied individually and prediction result of each classifier is given as input to voting method. The soft voting based method is applied which can generate the final result which maximum accuracy for the diabetes prediction.

The Gini Index can be calculated as:

$$Gini = \sum i \neq j \; p(i)p(j) \text{ ---[1]}$$

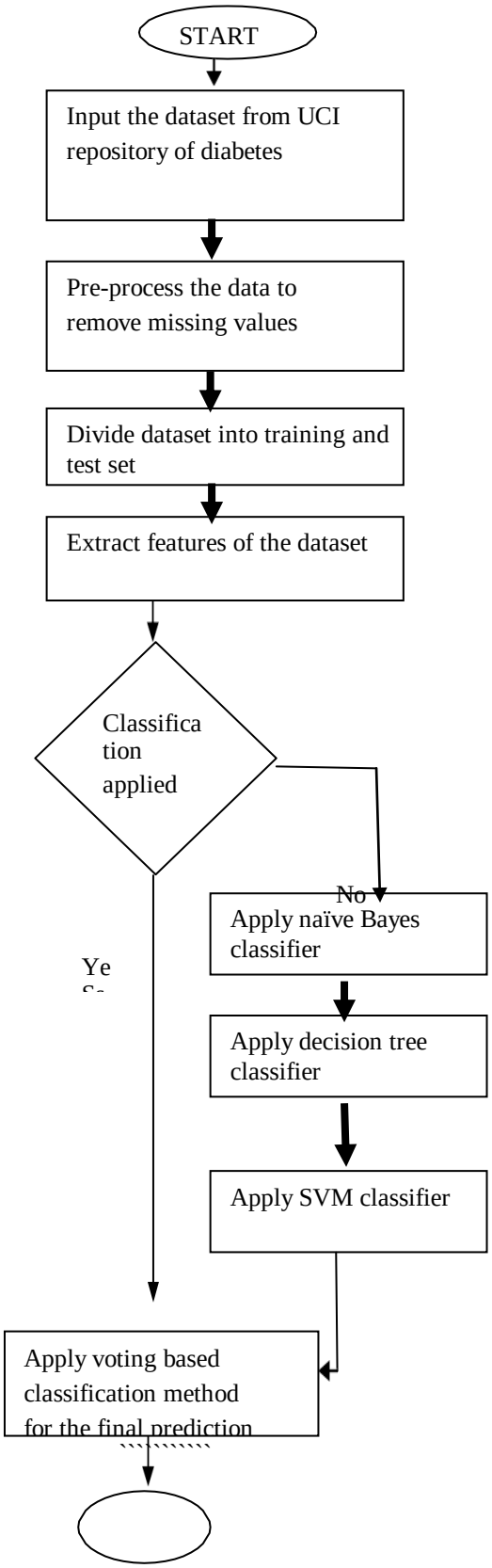
Here, i and j are levels of target values. The input dataset is denoted here as p.

D. Evaluation:- In the last phase, the results of the proposed model is evaluated in terms of certain parameters. The result evaluation takes input actual result and predicted result and return value of parameters like accuracy etc.

IV. RESULTS

The dataset is collected from the UCI repository. The two scenarios are

implemented which is existing scenario in which SVM classifier method is applied and in the proposed scenario voting method is applied for the diabetes prediction. The performance of the proposed method is tested in terms of accuracy, precision, recall and execution time. The dataset description is given in the table



V. RESEARCH METHODOLOGY

The dataset is collected from the UCI repository. The two scenarios are implemented which is existing scenario in which SVM classifier method is applied and in the proposed scenario voting method is applied for the diabetes prediction. The performance of the proposed method is tested in terms of accuracy, precision, recall and execution time. The dataset description is given in the table 1

Table 2: Dataset description

Table 1: Dataset Description				
Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	N/A	**Area:**
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**
Associated Tasks:	N/A	**Missing Values?**	N/A	**Number of Web Hits:**

The parameters for the performance analysis are described below:-

A. Precision: In pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.

Precision=(True Positive)/(True Positive+False Positive)

B. Recall: Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

Recall=(True Positive)/(True Positive+False Negative)

C. Accuracy: Accuracy is defined as the number of points correctly classified divided by total number of points multiplied by 10

g

2

:

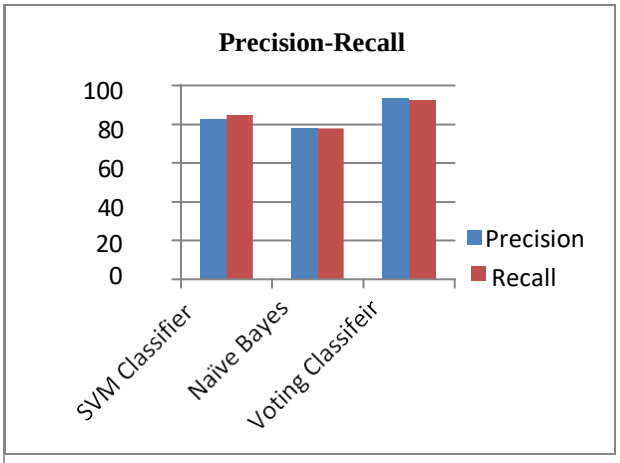
Accuracy = Number of points exact classified*100/Total Number of points

D. Execution Time: Execution time is defined as difference of end time when algorithm stops performing and start time when algorithm starts performing

Execution time = End time of algorithm- start of the algorithm

Table 1: Dataset Description

Data Set Characteristics :	Multivariate, Time-Series	**Number of Instances:**	N/A	**Area:**	Life
Attribute Characteristics :	Categorical, Integer	**Number of Attributes :**	20	**Date Donated**	N/A
Associated Tasks:	N/A	**Missing Values ?**	N/A	**Number of Web Hits:**	367821



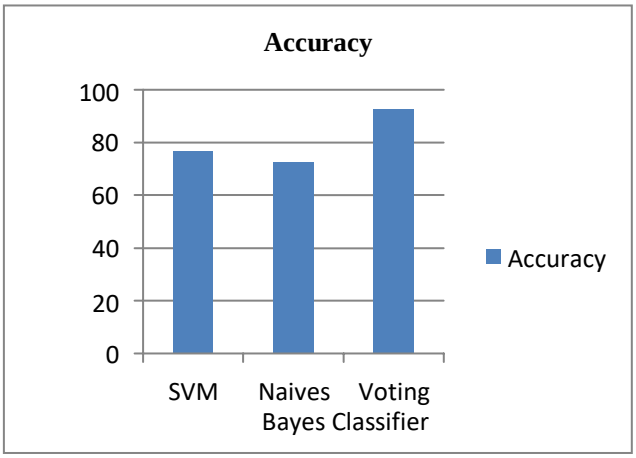


Fig 4: Accuracy comparison

As shown in figure 3, the performance of three classification methods which are SVM, naïve bayes and voting method are compared in terms of precision-recall values. The voting based method is the ensemble classifier due to which it has maximum value of precision-recall as compared to SVM and naïve bayes As shown in figure 4, the accuracy of three classifier which are SVM, naïve bayes and voting based methods are compared for the performance analysis. It is analyzed that voting classifier give maximum accuracy as compared to other classification methods.

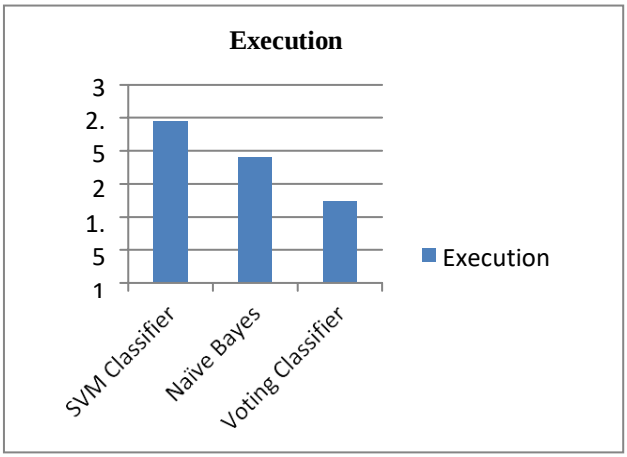


Fig 5: Execution Time

As shown in figure 5, the execution time of the SVM, naïve bayes and voting based method. It is analyzed that voting based classification has least execution time as compared to SVM, naïve bayes and voting method.

Accuracy of Voting Classifier is 0.8398268398268398

	precision	recall	f1-score	support
0	0.86	0.91	0.88	151
1	0.80	0.71	0.75	80
avg / total	0.84	0.84	0.84	231

Fig 6: Classification report

As shown in figure 6, the classification report of precision, recall and f-measure are shown for the class into 0 and 1.

VI. CONCLUSION

In this work, it is concluded that diabetes prediction is the complex problem due to high complexity of the dataset. The various

classification methods are designed for the prediction which is SVM and naïve bayes.

In this work, voting based classification method is designed for the prediction analysis. The voting based method is the combination of SVM, naïve bayes and decision tree. It is analyzed in terms of accuracy, precision-recall and execution time that voting based method high performance as compared to other classification methods.

The hypotheses were tested by using R and MS excel. Correlation Analysis was performed on R and regression analysis was done using MS excel. The results for correlation analysis are shown in Table 7 and the results of regression analysis are shown in the results section according to hypotheses.

REFERENCES

[1] Alexis Marcano-Cede˜no, Diego Andina, "Data mining for the diagnosis of type 2 diabetes", IEEE, Vol. 11, issue 3, pp. 9-19, 2016.
[2] B. M. Patil, R. C. Joshi, Durga Toshniwal, "Association rule for classification of type -2 diabetic patients", 2010 Second International Conference on Machine Learning and Computing, Vol. 8, issue 3, pp. 7-23, 2010.
[3] Prova Biswas1,2, Ashoke Sutradhar3, 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.
[4] MS.Tejashri n. Giri, prof. S.r.todamal, "data mining approach for diagnosing type 2 diabetes", international journal of science, engineering and technology, vol. 2 issue 8, 2014.
[5] P. Suresh Kumar and V. Umatejaswi, " Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017.
[6] M. Sharma, G. Singh, R. Singh, "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques", Elsevier, vol. 5, pp. 202- 222, 2017.
[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] Yan Luo, Charles Ling, Ph.D., Jody Schuurman, Robert Petrella, MD, "GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing", 2014 IEEE International Conference on Data Mining, Vol. 9, issue 8, pp. 12-23, 2014.
[9] Abdelghani Bellaachia and Erhan Guven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, 2010, pp. 234-239.
[10] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance"