Diabetes Prediction Approach using Voting Based Classification Method

KAVITI HARIKA[1], S.Vizianandа Row[2]
[1]*Mtech Scholar,* [2]*Assistant Professor*
[12]*Andhra University,Visakhapatnam*

Abstract- In order to extract knowledge and patterns in large datasets, data mining can be used. The data mining tools can work and analyze different types of datasets irrespective of being structured or unstructured. In this work, the technique of KNN is applied for the heart disease prediction. The voting based classify data more accurately which improve accuracy of heart disease prediction. The proposed algorithm performance is tested in the heart disease dataset which is taken from UCI repository. There are 76 attributes present within a database. However, a subset of 14 amongst them is required within all the published experiments. Specifically, machine learning researchers have used Cleveland database particularly at all times. The proposed work will also be compared with the existing scheme (using arithmetic mean) in terms of accuracy, precision and recall

Keywords- Random forest, Diabetes prediction, SVM, Voting Method

I. INTRODUCTION

The process of extraction of interesting knowledge and patterns to analyze data is known as data mining. In data mining there are various data mining tools available which are used to analyze different types of data. Decision making, market basket analysis, production control, customer retention, scientific discovers and education systems are some of the applications that use data mining in order to analyze the collected information [1]. The multimedia, object relational, relational and data ware houses are some of the databases for which data mining has been studied. Data is generated on regular basis on large scale through various applications. However, there are some cases in which the useful information cannot be extracted from the applications due to which this application remains unutilized. There are both structured and unstructured types of data present within various systems amongst different channels [5]. Diabetes Mellitus is a chronic disease. This disease has no recognized treatment. This disease can be controlled using some particular situation management methods are like keeping blood sugar levels as close to normal as possible without causing hypoglycemia. This disease can be prohibited with diet, exercise and use of suitable medicines. Diabetes Mellitus (DM) is a set of related diseases. In this disease, the body is not able to control the sugar level in the blood. Some hormones involving insulin regulate the sugar level of blood in a normal human being. The pancreas, a small organ amid the stomach and liver generates insulin. Some other main enzymes for food digestion are secreted by the pancreas. The glucose from the blood into liver, muscle, and fat cells is transmitted by insulin. Here, it is utilized as fuel.

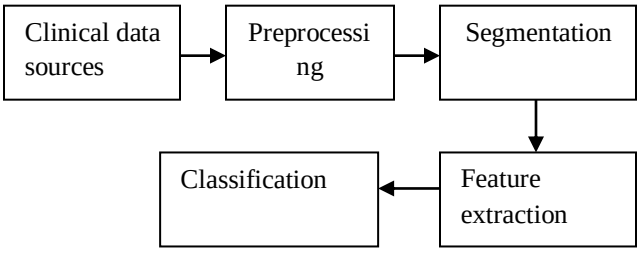

Fig.1: General steps involved in diabetes prediction

The clinical data sources are gathered from various sources. These sources include sensors, web warehouses and physical artificial datasets. The preprocessing step is applied after the efficient gathering of data. This process involves data cleaning, data integration, etc. Data cleansing is the other name given to the data cleaning process. In this process, the noisy, unrelated and missing facts form the gathered data are detected and removed. The next stage of cleansing is data integration. In this stage, data from numerous differ data sources is mixed in a general source set-up. The unidentified and hidden predictive information from heavy and dynamic databases is extracted in clinical data mining. The next step in diabetes prediction is called segmentation. In this process, the obtained data is divided into its essential parts or objects. The subdivision level is based on the issue being resolved. In order to extract features, Feature extraction techniques are evolved. High-level features required to carry out target classification are extracted by this technique. A set of things that describe target in a unique way are called features. The features include size, shape, composition, location etc. The last step in diabetes prediction is called Classification. This is one of the important processes of diabetes prediction. Classification is one of the most popular operations of data mining. The classification is generally involved in the huge number of business and medical data sets. Classification is a data mining function that can distribute the things in a compilation form to target different classes. Support vector machine is a supervised

learning algorithm. This algorithm characterizes instances of data in form of points in space. Later, this algorithm constructs a model for assigning novel instances to one class or another. Every data point is characterized as a n-dimensional vector. Further, SVM builds an n-1-dimensional separating hyperplane for discriminating two classes with maximum distance amid the hyperplane and data points on every side. The main purpose of this algorithm is to detect the optimum hyperplane to separate both classes. Data is characterized as:

$$(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$$

In the above equation, yi either 1 or -1 indicates the class of xi. Every xi is p-dimensional vector. This vector represents whole characteristic values (variables) of xi. The hyperplane that optimally separates the group of xi vectors where yi = 1 from the group of vectors where yi = -1 is

$$\vec{w} \cdot \vec{x} - b = 0,$$

In the above equation, $\vec{w}$ is the normal vector to the hyperplane. The offset of the hyperplane from the source is represented by b. In case of linearly separable data points, the hard margin can be described as

$$\vec{w} \cdot \vec{x} - b = 1 \quad \vec{w} \cdot \vec{x} - b = -1.$$

The objects based on closest training examples are classified using KNN classifier in the feature space. This is the most fundamental category of instance-based learning or lazy learning. In n-dimensional space, this classifier imagines all instances as points. The "closeness" of instances can be determined using a distance measure. The n-dimensional numeric attributes generally describe training samples in KNN classifier. The n dimensional space is used to store the training samples. The 'k' training samples closest to the unknown sample or test sample are searched by this classification model in case of a given test sample. In general, Euclidean distance represents closeness. The equation expressed below represents the Euclidean distance between two points P and Q i.e. P (p1, p2… Pn) and Q (q1, q2,..qn)

$$d(P, Q) = \sum_{i==1}^{n} (P_i - Q_I)^2$$

II. LITERATURE SURVEY

In paper [6], the author has proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. The data was gathered from a hospital which included within it both structured as well as unstructured types of data. In order to make predictions related to the chronic disease that had been spread within several regions, various machine learning algorithms were streamlined here. A latent factor model was utilized in this method in order to reconstruct the incomplete type of data present within the gathered data. A regional chronic disease of cerebral infarction was utilized in order to perform various experiments to evaluate the performance of proposed method. It was seen through the various comparisons made amongst existing and the proposed technique that none of the previously existing methods dealt with both types of data that was gathered from medical fields. 94.8% of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms.

The author in paper [7] has presented that most of the deaths every year is caused due to heart disease, it is the fatal disease. It is necessary diagnose this disease at the early stage as maximum number of causalities are occurred from this disease. Higher knowledge and expertise researcher or doctors has been required for the prediction of disease, therefore it considered as the difficult task. There are various issues occur in the prediction of heart disease for which different attributes has been used. All this process is done on the basis of the data mining techniques. For the investigation of the heart disease various experiments were performed by the author. KStar, J48, SMO, Bayes Net and Multilayer Perceptron were used for this purpose that can be possible through Weka software. Data mining techniques performance is compared with the standard data setin terms of predictive accuracy, ROC curve and AUC value.The SMO and Bayes Net technique shows the optimal performance as compared to the performance of KStar, Multilayer Perceptron and J48 techniques.

The author has presented that coronary heart disease is the most fatal heart disease as large amount of the deaths occur due to this disease in the worldwide in paper [8]. The diagnosis process of this disease is complicated as it requires proper monitoring all the time. Therefore, invention of an intellectual decision support system is essential in order to predict heart disease. Author in this paper discussed the use of data mining techniques in the medical system. These techniques provide the idea to the doctors whether the patient is suffering from any heart disease or not. Hidden Naïve Bayes is the extended version of the traditional Naïve Bayes method in the data mining. The conditional independence assumption of traditional method, in the data mining is relaxed by using this model. For the classification and prediction of heart disease, Hidden Naïve Bayes has been utilized in accordance with the proposed model. On the basis of the performed experiments, it is concluded that Hidden Naïve Bayes (HNB) is superior to naïve bayes in terms of optimal accuracy.

III. RESEARCH METHDOLOGY

This research work is based on the prediction analysis of diabetes diseases. The prediction analysis is the technique in which future possibilities can be predicted based on the current dataset. The k-mean clustering is the clustering technique in which similar and dissimilar data is clustered together on the basis of their similarity. In the k-mean clustering, the dataset is considered and from that dataset arithmetic mean is calculated which will be the central point of the dataset. The Euclidian distance from the central point is calculated and points which are similar and dissimilar are clustered into different clusters. The Euclidian distance is calculated dynamically in this work to increase accuracy of clustering. The Euclidian distance is calculated dynamically using back propagation algorithm which clusters the uncluttered points and increase accuracy of clustering.

1. Pre-Processing:- In this phase, the data is given as input and data which is cleaned means missing values, redundant values are removed. The data set is described in terms of standard deviation, mean etc values are calculated.

2. Prediction Phase: - In the phase, The voting based classifier provides an ensemble solution by gathering a set of various classifiers and training and evaluating them parallel such that the various peculiarities of each of the classifiers can be exploited.

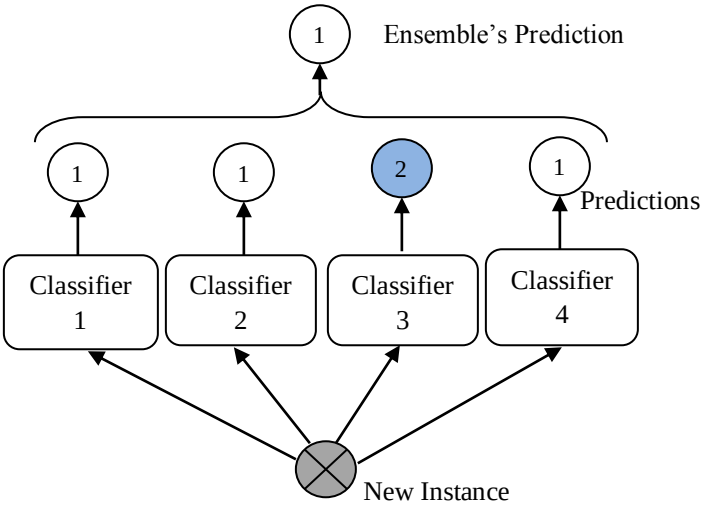

Fig.2: Voting Based Classifier

Various classifiers are combined to generate this classifier as shown above. A prediction method is defined here through which the majority rule of predictions is taken by the classifiers. For instance, for any given sample if the prediction is

- Classifier 1 - class 1
- classifier 2 -> class 1
- classifier 3 -> class 2

The sample is classified here as "class 1". Further, to assign a particular weight to each classifier, a "weights" parameter is added. The predicted class probability for each classifier is collected then multiplied with the classifier "weight" and then average is calculated to work with weights. The class label can then be assigned based on the weighted average probability.

A classifier that creates a set of decision trees from randomly chosen subset of training set is known random forest classifier. The votes from various decision trees are aggregated for deciding the final class of test object. The random forest algorithm introduces predictive models for deriving solutions of issues arising from classification and regression. Multiple learning models are used by ensemble methods to achieve improved predictive results. The random forest model creates an entire forest of randomly uncorrelated decision trees to provide one appropriate result. To resolve the correlation problem by random forest, only one subsample of feature space is selected at each split. To de-correlate and prune the trees for node splits, a stopping criterion is set. Mainly, with the increase in number of trees, the robustness of forest also increases. In the similar way, if the numbers of trees present in forest are high, the accuracy of the outputs of classifier is also increased.

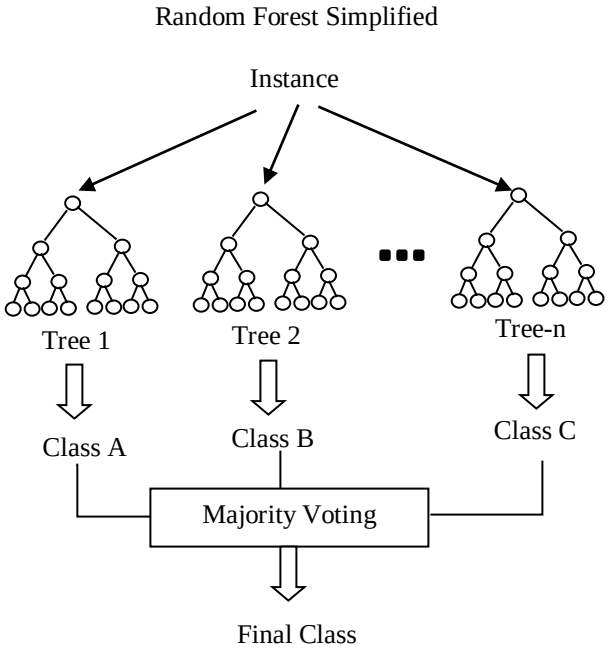

Fig.3: Random Forest Classifier

IV. RESULT AND DISCUSSION

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 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:**	3678

The parameters for the performance analysis are described below:-

1. 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)

2. 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)

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

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

4. *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

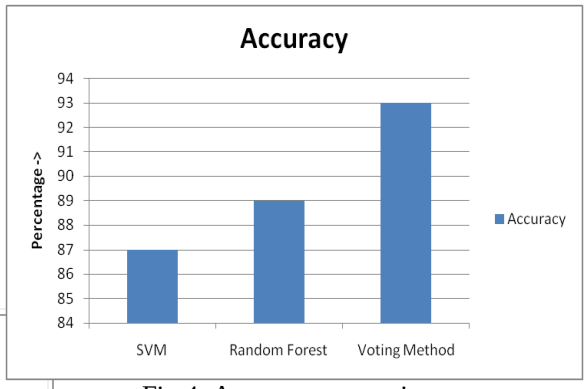

Fig.4: Accuracy comparison

As shown in figure 4, the accuracy of three classifier which are SVM, random forest 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.

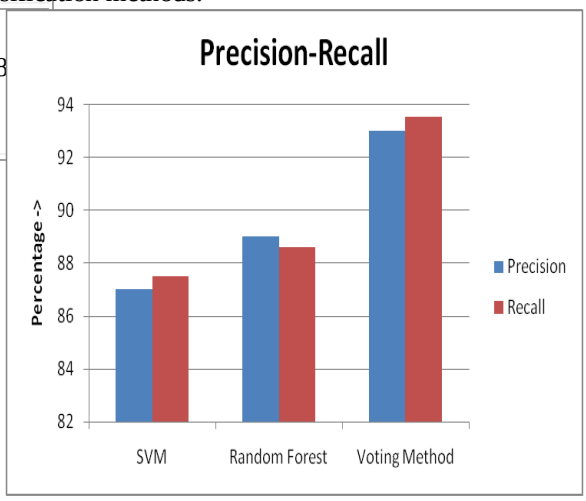

Fig.5: Precision-recall comparison

As shown in figure 5, the performance of three classification methods which are SVM, random forest 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 random forest

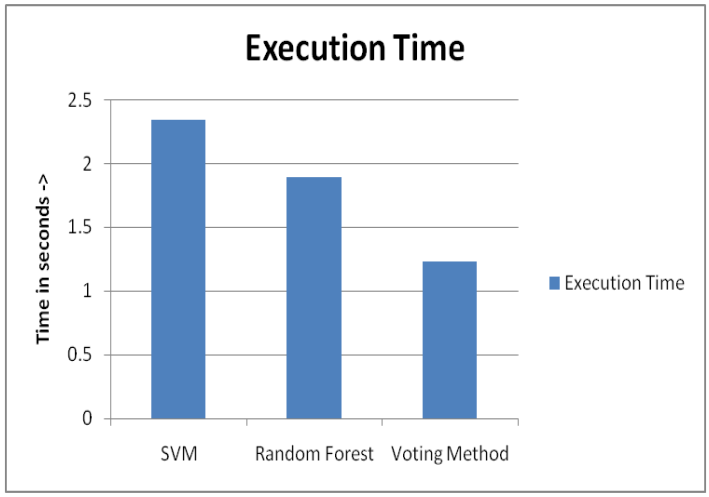

Fig.6: Execution Time

As shown in figure 6, 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, random forest and voting method

V. CONCLUSION

The relevant information is fetched from rough dataset using data mining technique. The similar and dissimilar data is clustered after calculating a similarity between input dataset. SVM is used to classify both similar and dissimilar data type in which central point is calculated by calculating an arithmetic mean of the dataset. The central point calculated Euclidian distance is used to calculate a similarity between different data points. According to the type of input dataset a clustered data is classified using SVM classifier scheme in the last step. In this research work, the voting based classifeir is applied for the diabetes prediction. The clustered result will be given as input for the classification. It is analyzed that improved technique has less execution time and high accuracy of classification as compared to existing technique

VI. REFERENCES

[1]. MonaliDey, SiddharthSwarupRautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies, vol. 6, issue 3, pp. 234-239, 2014.

[2]. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, issue 4, pp. 123-128, 2010.

[3]. AzharRauf, Mahfooz, Shah Khusro and HumaJaved (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, issue 6, pp. 959-963, 2012.

[4]. Indira S. FalDessai, Intelligent Heart Disease Prediction System Using Probabilistic Neural Network, International Journal on Advanced Computer Theory and Engineering, vol. 7, issue 4, pp-56-62, 2013.

[5]. Abhishektaneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, vol. 5, issue 4, pp. 959-963, 2013.

[6]. Min Chen, YixueHao, 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, issue 4, pp- 215-227, 2017.

[7]. Marjia Sultana, Afrin Haider and Mohammad ShorifUddin, "Analysis of Data Mining Techniques for Heart Disease Prediction", IEEE, vol. 14, issue 1, pp. 123-138, 2016.

[8]. M. A. Jabbar, Shirinasamreen, "Heart disease prediction system based on hidden naïve bayes classifier", vol. 4, issue 11, pp. 23-48, 2016.

[9]. Akhilesh Kumar Yadav, DivyaTomar and Sonali Agarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, issue 16, pp.121-126, 2013.

[10]. Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", vol. 8, issue 9, pp. 142-147, 2014.

[11]. Chew Li Sa, BtAbang 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),The 5th International Conference on, vol.15, issue 6, pp.1-6, 2014.

[12]. Qasem A. Al-Radaideh, Adel Abu Assaf and EmanAlnagi (2013), "Predicting Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013), vol. 23, issue 17, pp. 32-38, 2013.

[13]. K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin (2015), "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, issue 12, pp. 1023-1028, 2015.

[14]. BalaSundar V, T Devi and N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications, vol. 48, issue 15, pp. 423-428, 2012.

[15]. Daljit Kaur and KiranJyot (2013), "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, vol. 2 issue 12, pp. 724-729, 2013.