Classification Technique for Heart Disease Prediction in Data Mining

Janmejay Kumar Tiwari[1], Manmohan Singh[2]
[1]Student M.Tech, [2]Assistant Professor
RKDF School of engineering, INDORE M.P.

***Abstract* -** The data mining is the technique to analyze the complex data. The prediction analysis is the technique which is applied to predict the data according to the input dataset. In the recent times, various techniques have been applied for the prediction analysis. In this work, the k-means clustering algorithm and SVM (support vector machine) classifier based prediction analysis technique is used for clustering and classification of the input data. In order to increase the accuracy of prediction analysis, the back propagation algorithm is proposed to be applied with the k-means clustering algorithm to cluster the data. 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, fault detection rate and execution time.

***Keywords* -** *SVM, Back propagation, Prediction*

I. INTRODUCTION

Data mining is defined as the process in which useful information is extracted from the raw data. In order to acquire essential knowledge it is essential to extract large amount of data. This process of extraction is also known as misnomer. Currently in every field, there is large amount of data is present and analyzing whole data is very difficult as well as it consumes a lot of time. This present data is in raw form that is of no use hence a proper data mining process is necessary to extract knowledge. The process of extracting raw material is characterized as mining [1]. For the analysis of the simple data there are various cheaper, simpler and more effective solutions are present. The main objective of using data mining is to discover important information that is available in distorted manner. With the help of databases, data mining tools can sweep and can identify previously hidden patterns. Data entry key errors are represented by the patterns discovery problems such as network system and fraudulent credit card transactions detection. Therefore the result must be presented in the way human can understand possible with the help of network supervisor and marketing manager domain expert. The predictive information can be extracted from various applications with the help of efficient data mining tools [2]. Nowadays satellite pictures, business transactions, text-reports, military intelligence and scientific data are the major source of information that needs to be handled. For decision-making no appropriate results are provided by the information retrieval. It is required to invent new methods in order to handle large amount of data that helps in making good decisions. In the raw data it is required to discover new patterns and important information is extracted in order to summarize all the data. In many applications a great success is provided by the data mining and various companies such as communication, financial, retail and marketing organizations are utilizing this technique in order to minimize their work pressure [3]. For the development of product and its promotion, retailers used data mining approach as by which they can build a record of every customer such as its purchase and reviews. An essential role is played by data mining process when it is impossible to enumerate all application. In the cluster analysis, image processing, market research, data analysis and pattern recognition are some major application of this technique. In the clustering technique, customer categorized group and purchasing patterns has been done in order to discover their customer's interest by the marketers [4]. It is also utilized in biology as it derives the plant and animal taxonomies and also categorizes genes with similar functionality. In geology this technique is used to identify the similar houses and lands areas. The group of metabolic diseases in which a person has high blood sugar is commonly referred as diabetes and in the scientific term as Diabetes mellitus. There are two reasons for the presence of high blood sugar in the body: (1) enough insulin is not produced by the pancreas, (2) no response by cells to the produced insulin. Hence, it is the condition that occurs in the human body due to absence of appropriate insulin. There are various types of diabetes exists such as diabetes insipidus [5]. There is a cycle among clinical research, outcomes and concepts, guidelines, quality indicators, performance measures that is provided by the medical management for the prediction of diabetes. Data generated by the by healthcare transactions are complex in nature and high in volume that can be analyzed by traditional methods. For the knowledge acquisition, Medical data mining has been utilized as it contains all the information from research reports, medical reports, flowcharts, evidence tables.

All this information is useful for decision making whether patient is suffering from diabetes or not. In India, Diabetes is a major health problem [6]. There are various impacts of this disease ion human body such as risk of kidney failure and eye issues. In order to avoid all these complications early detection of the disease and proper care management is required. The main objective of inventing a diabetes data system helps the diabetic patients during the disease. It is necessary for the diabetic patient to have daily glucoses rate and insulin dosages that can be possible by diabetes data system as it care the daily dosages a person inhale. This system is not only for the diabetic patient but also for those who suspect if they are diabetic. In order to map the input space to output space, fuzzy logic is an optimal way [7]. A wide range of problems in different applications are solved with the help of this method. This methods functioning depends up on the learning and adaptation capability. In the process of diagnosis of diabetes there are various set of framework in the fuzzy that has been utilized. The popular computing framework is a Fuzzy Inference System (FIS).

II. LITERATURE REVIEW

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 [8]. 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. Type 2 diabetes (TTD) is the most common type of diabetes. 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.

Yu-Xuan Wang, et.al, (2017) analyzed various applications that provide significance of the data mining and machine learning in different fields [9]. Research on the management de-signs of different components of the system is proposed as most of the work is done on the characteristics of the system that varies from time to time. The performance of the system with static or statically adaptive is optimized with the help of proposed method in order to design system. Author in this paper proposed a method to design operating system that use the support of data mining and machine learning. 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. For the data mining and analytics eight unsupervised and ten supervised learning algorithms were considered for the investigation purpose [10]. 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.

P. Suresh Kumar, et.al (2017) proposed a model that overcome all the problems such as clustering and classifications from the existing system by applying data mining technique. This method is used to diagnose the type of diabetes and from the collected data a security level for every patient. There are various affects of this disease due to which most of the research is done in this area [11]. All the collected data of the 650 patient's was used in this paper for the investigation purpose and its affects are identified. In the classification model, this clustered dataset was used as input that is used for the classification process such as patient's risk levels of diabetes as mild, moderate and severe. In order to diagnose diabetes, performance analysis of different algorithms was done. On the basis of obtained result the performance of each classification algorithm is measured.

Han Wu, et.al (2018) proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). The main objective of this paper is to improve the accuracy of the prediction model and to more than one dataset model is made adaptive in nature. Proposed model comprised of two parts based on a series of preprocessing procedures [12]. These two parts are improved K-means algorithm and the logistic regression algorithm. In order to compare the results with other methods the Pima Indians Diabetes Dataset and the Waikato Environment was utilized for Knowledge Analysis toolkit. As per performed experiments, it is concluded that proposed model show netter accuracy as compared to other methods and also provide the sufficient dataset quality. In order to evaluate the performance of the model it is applied to other diabetes dataset, in which good performance is shown by both the methods.

Jahin Majumdar, et.al, (2016) presented the most popular research areas in computer science that is data mining and machine learning is utilized in order to provide essential data or information [13]. The SFS and SBS approaches are the optimal approaches and preferred as it use with forward selection. SVM techniques are used by the proposed heuristic model as it provides the accuracy and heavy in the computational functions. 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 techniques was done in order to find out the best method. The theoretical limitations of existing algorithms were overcome by proposed method.

III. RESEARCH METHODOLOGY

This research work is based on the prediction analysis of heart diseases. The prediction analysis is the technique in which future possibilities can predicted based on the current dataset. In this research work, technique of SVM is applied previously for the prediction analysis. One of the simplest algorithms amongst all the learning machine algorithms is the SVM algorithm. Since there are no assumptions made on the underlying data distribution, decision tree is known to be a non-parametric supervised learning algorithm. Here, on the basis of nearest training samples present within the feature space, the samples are classified. The feature vectors are stored along with the labels of training pictures within the training process. Towards the label of its k-nearest neighbors, the unlabelled question point is doled out during the classification process. Through majority share cote, on the basis of labels of its neighbors, the object is characterized. The object is classified essentially as the class of the object that is nearest to it in the event when k=1. k is known to be an odd integer in case when there are only two classes present. During the performance of multiclass categorization, there can be tie in case when k is an odd whole number.

IV. EXPERIMENTAL RESULTS

The proposed approach is implemented in Python and the results are analyzed by showing comparisons amongst proposed and existing approaches in terms of accuracy and execution time.

1. *Accuracy:* Accuracy is defined as the number of points correctly classified divided by total number of points multiplied by 100, as shown in eqn. 1.

Accuracy =

$$\frac{\text{Number of points correctly classified}}{\text{Total Number of points}}*100 \text{ ---1}$$

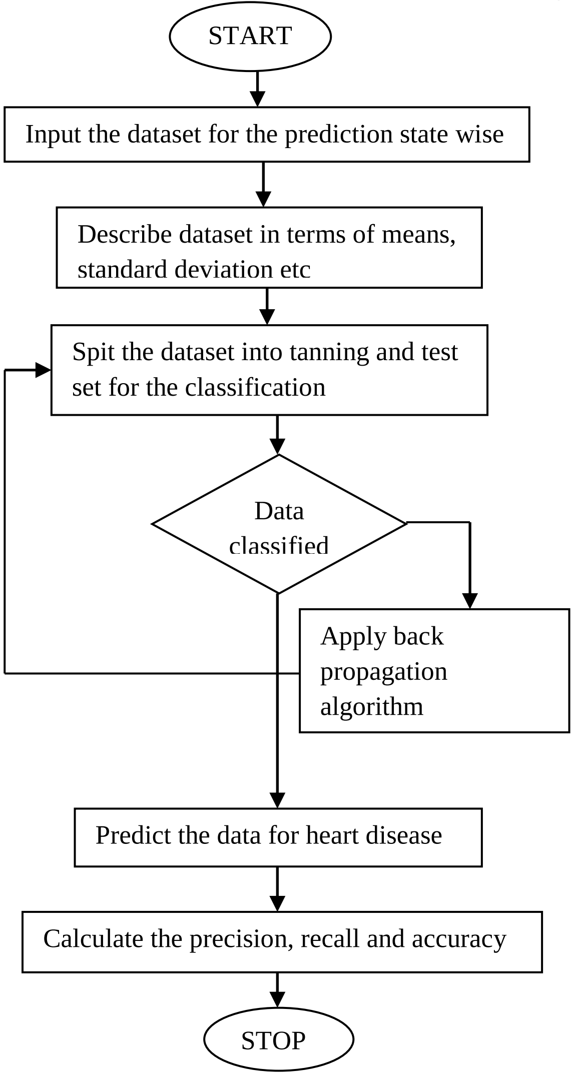
START
Input the dataset for the prediction state wise
Describe dataset in terms of means, standard deviation etc
Spit the dataset into tanning and test set for the classification
Data classified
Apply back propagation algorithm
Predict the data for heart disease
Calculate the precision, recall and accuracy
STOP

Fig 1: Proposed Methodology

As shown in figure 2, the accuracy comparison of existing and proposed algorithm is shown. The accuracy of proposed algorithm is high as compared to existing algorithm.

2. *Execution Time:* Execution time is defined as difference of end time when algorithm stops performing and starts time when algorithm starts performing as shown in eqn. 2.

   Execution time = End time of algorithm- start of the algorithm --2

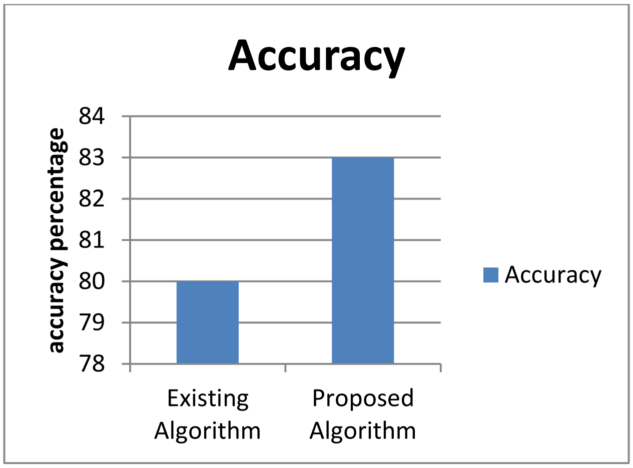

Fig 2: Accuracy Comparison

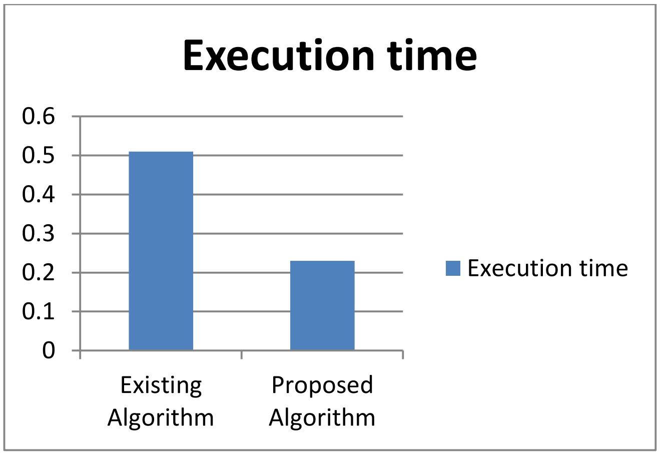

Fig 3: Execution time

As shown in figure 3, the execution time of proposed and existing algorithm is shown. The execution time of proposed algorithm is less as compared to existing algorithm.

3. CAP Analysis: -A canonical analysis on the principal coordinates for any resemblance matrix, including a permutation test. CAP takes into consideration the structure of the data. So, it is more likely to separate your different levels if there is no strong difference and is good to show the interaction between factors

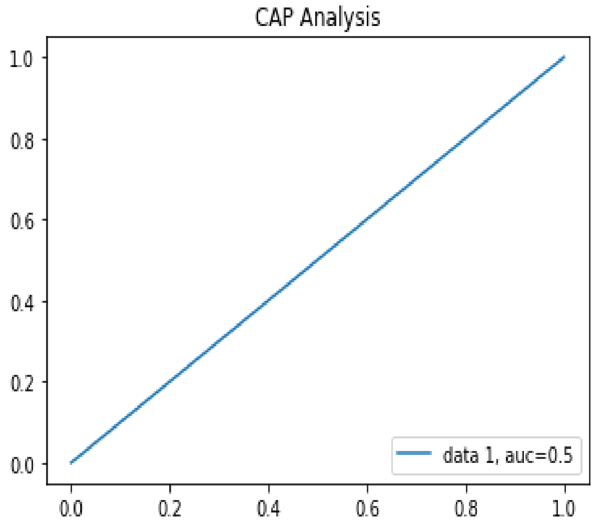

Fig 3: CAP Analysis

As shown in figure 3, the CAP analysis is shown in this figure. On the axis of this cure the training dataset is given as input and on the y-axis the test data is given as input. The blue line shows that CAP curve which represents accuracy of the classifier.

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. The SVM 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 this research work back propagation algorithm is applied with SVM classifier to increase accuracy of prediction. The proposed algorithm performs well in terms of accuracy and execution time. In future proposed technique will be further improved to design hybrid classifier for the heart disease prediction

VI. REFERENCES

[1] Yanhui Sun, Liying Fang and Pu Wang, Improved k-means clustering based on Efros distance for longitudinal data, 2016 Chinese Control and Decision Conference (CCDC), Vol. 11, issue 3, pp. 12-23, 2016.

[2] Shunye Wang, Improved K-means clustering algorithm based on the optimized initial centroids, 2013 3rd

International Conference on Computer Science and Network Technology (ICCSNT), Vol. 11, issue 3, pp. 12-23, 2013.

[3] Phattharat Songthung and Kunwadee Sripanidkulchai, Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification, 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Vol. 11, issue 3, pp. 12-23, 2016.

[4] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, vol. 3, pp. 1-31, 2000.

[5] Gary M. Weiss, Brian D. Davison, “Data Mining”, To appear in the Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, vol. 3, pp. 121-140, 2010.

[6] 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.

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

[8] Bayu Adhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, “Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine”, Vol. 11, issue 3, pp. 12-23, 2008.

[9] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, “Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization”, Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.

[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] 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.

[12] 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.

[13] Jahin Majumdar, Anwesha Mal, Shruti Gupta, “Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining”, 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), vol. 3, pp. 73-77, 2016.