Classification Technique for Heart Disease Prediction in Data Mining

Mohini Chakarverti, Rajiva Ranjan Divivedi

ABSTRACT--- The difficult information is analyzed through an approach named data mining while prediction analysis approach is used for the prediction of information on the basis of input data suite. Currently, a lot of methods have been implemented for prediction analysis. In the proposed study, clustering and classification of the input information for heart disease forecasting is executed with the help of k-means clustering algorithm and SVM (support vector machine) classification model on the basis of prediction analysis methods. The back propagation algorithm along with k-means clustering algorithm is applied for the clustering of information. These algorithms support to enhance the precision of prediction analysis. A heart disease data suite obtained from the UCI repository is used for judging the performance of proposed algorithm. This data suite comprises total 76 features. But, whole tests require a subset of 14 features. Particularly, Cleveland database particularly is utilized by the machine learning researchers throughout the tests.A comparison between proposed study and earlier method (using arithmetic mean) is performed in terms of certain parameters such as accuracy, error recognition rate and execution time.

Keywords– SVM, KNN, Heart Disease Prediction

INTRODUCTION

The procedure which is used for the extraction of valuable data from the unprocessed information is called data mining. This arbitrarily retrieved information can be set in the structured format. This data can be utilized as significant information in several applications. This procedure of data extraction is identified as misnomer too. In the recent times, huge quantity of information is present in almost each domain. The inspection of entire accessible information is extremely complex and requires a lot of time too. This accessible unprocessed information cannot be utilized for any purpose. Therefore, a suitable data mining technique is required for the extraction of useful information. [1]. A lot of inexpensive, easier and more effectual techniques occur for inspection of the uncomplicated information.The major purpose of data mining is the detection of valuable data. This data often remains present in indistinct way. Data mining devices can sweep and recognize earlier concealed samples using database. The patterns detection issues like network structure and fake credit card transactions discovery represent the information entry key faults. Thus, the outcomes should be provided in comprehensible format. Proficient data mining devices are used for the extraction of predictive data from different applications [2]. At the present time, satellite images, business transactions, text-reports, army intelligence and technical information are the main source of data. This data should be handled in proper way. The data extraction procedure does not provide any suitable outcomes for decision making. In order to make good decisions, discovery of novel data handling techniques is imperative. In unprocessed information, discovery of novel patterns and vital data is necessary for the summarization of whole retrieved information. Data mining techniques provide huge achievements in a number of applications. Different business like communication, financial, retail and marketing associations are using data mining for minimizing their work load [3]. With the help of data mining, retailers can make a record of all clients on the basis of their purchase and reviews. Data mining approach plays a fundamental role when it is not viable to count all applications. In the cluster scrutiny, some main applications of data mining are image processing, market exploration, and information analysis and pattern detection. The users are classified into several classes and patterns during the clustering procedure. The marketers describe the interests of a user [4]. Clustering approach is applied in biology too. This technique performs the classification of plant and animals and also classifies genes with identical performance. This approach is utilized for the identification of analogous homes and land regions in geology. The data mining approach uses two methods i.e. Supervised and unsupervised machine learning for the prediction of heart diseases [5]. For learning the parameters of the model, a training suite is used in supervised machine learning for the learning of model parameters. The unsupervised machine learning does not use any training set such as k-means clustering. Two major aims of data mining are classification and prediction.The classification schemes classify the information process rate and unordered values. On the other hand, prediction models predict the continuous value. Decision trees and Neural Networks are some examples of classification models whereas Regression, Association Rules and Clustering are the examples of prediction algorithm. Classification models use Decision trees, neural networks and Naive Bayes Classifier in data mining to predict the heart diseases. The decision tree algorithm is the most prevailing approach [6]. In this approach, all models are created in shape of a tree. Data suites are divided into little suites and supports in the formulation of a related decision tree. A lot of elements are arranged in different number of interlinked layers in the neural network approach.

Revised Manuscript Received on May15, 2019.

MohiniChakarvertiResearch Scholar, IEC College of Engineering and Technology, Greater Noida, India. (E-mail: mchakarverti@gmail.com)

RajivaRanjanDivivediAssistant Professor, IEC College of Engineering and Technology, Greater Noida, India.(E-mail: rajiv.ranjan.0077@gmail.com)

Through this approach, the adaptive non-linear information processing algorithms are applied with the help of this approach for providing help in the integration of every multi-processing unit. In this approach, the networks are characterized according to self-organization and natural adjustment [7]. The Naive Bayes classifier is a probabilistic classifier which depends upon Bayes theorem. This algorithm is too identified as the independent feature model. In Naive Bayes classifier it is assumed that the existing features of a particular class do not relate to the existing feature of some other class. Naive Bayes classifiers are prepared to work in supervised learning.

LITERATURE REVIEW

BayuAdhi Tama, et.al (2016) proposed a research related to the chronic disease identified as diabetes. This disease was considered extremely common and caused main causalities. Approximately 285 million people all around the world were suffering from diabetes according to a survey conducted by International Diabetes Federation (IDF) [8]. These numbers could increase in nearby future due to the inexistence of a suitable technique in the absolute minimization and prevention of this disease. The most common type of diabetes was Type 2 diabetes. The main problem was the discovery of TTD since the prediction of all its effect was not an easy task. Thus, data mining was utilized because it provided the finest outcomes and helped in the discovery of information from accessible data. A classifier named support vector machine (SVM) was used in data mining procedure for the extraction of valuable information of all patients from the earlier records. The timely recognition of TTD provided help in the taking of efficient decision.

Yu-Xuan Wang, et.al, (2017) scrutinized a number of applications that provided importance of the data mining and machine learning in different domains [9]. Study conducted on the organization de-signs of different devices was proposed since most of the work was performed on the features of the system that varied timely. The functionality of the structure was modified using proposed approach for designing a system. In this study, a novel technique was proposed for the designing of an operating system. The proposed approach utilized data mining and machine learning techniques. After obtaining a response from a data miner, the entire data gathered from the structure was examined. On the basis of conducted tests, it was identified that proposed approach provided efficient outcomes.

ZhiqiangGe, et.al, (2017) presented a study on earlier data mining and analytics applications. These techniques were utilized in business for different perspectives. Eight unsupervised and ten supervised learning algorithms were used for the research in data mining and analytics [10].In this study, an application status was presented for the semi-supervised learning algorithms. In industry procedure, both unsupervised and supervised machine learning techniques were utilized in approximate 90%-95% applications. The semi-supervised machine learning has been proposed in the recent times. Thus, it was depicted that the data mining and analysts played an essential role in the designing of novel machine learning approaches in the industry related applications.

P. Suresh Kumar, et.al (2017) proposed a novel approach for the removal of several problems being experienced in clustering and classification techniques in data mining structures. This technique was utilized for analyzing the diabetes type. The fitness level of each patient was measured from the gathered information. This disease caused several effects because of which several researches were conducted in this field [11]. In this study, all gathered information about 650 patients was utilized for research perspective and its affects were detected as well. This clustered data suite was utilized as input in the classification system. This classification system further classified information into definite classes such as patient's risk levels of diabetes were classified as mild, moderate and severe. The performances of different algorithms were analyzed for the correct diagnosis of this disease. The performance of every classification algorithm was measured according to the attained outcomes.

Han Wu, et.al (2018) proposed a new model on the basis of data mining processes to predict type 2 diabetes mellitus (T2DM). The major aim of this study was the improvement in the precision of prediction model. The creation of an adaptive data suite model was the other aim of this study. Proposed model was made up of two parts these parts were based on the sequence of preprocessing procedures [12]. These two sections were identified as enhanced K-means algorithm and the logistic regression algorithm. For comparing the outcomes of proposed and existing techniques, the Pima Indians Diabetes Dataset and the Waikato Environment was utilized along with Knowledge Analysis toolkit. The comparative outcomes depicted that proposed model showed better precision in comparison with other techniques and provided the adequate dataset quality as well. The performance of proposed model was evaluated through its usage in several other diabetes datasets. It was identified that both the approaches showed good performance.

JahinMajumdar, et.al, (2016) provided an analysis of data mining and machine learning approaches. These two approaches were the most popular investigational areas in computer discipline [13]. The SFS and SBS techniques were the most advantageous approaches and applied with forward selection method.The proposed heuristic model utilized SVM classifier. These classification model provided accuracy and computational operations. A data suite was used to measure the accuracy rate of SVM classifier. A number of tests were conducted for improving the data classification and pattern detection in Data Mining. In these tests, feature selection was the prime focus. A number of comparisons were performed to find out the superlative technique. The tested outcomes indicated that the proposed approach removed the limitations of existing algorithms.

PROPOSED METHODOLOGY

This study is based on the prediction analysis of heart diseases. In prediction analysis approach, future potentials can be forecasted on the basis of present data suite. In this

study, earlier SVM classifier is used for prediction analysis. The SVM algorithm is considered one of the simplest algorithms among all the machine learning algorithms. Decision tree is identified as a non-parametric supervised learning algorithm as no suppositions are made on the underlying information distribution.In this study, the samples are classified according to the closest patterns existing inside the feature space. During the training procedure, feature vectors are stored with training pictures' labels. The unlabelled question point is doled out for the duration of the classification process in the direction of its k-nearest neighbors' label. The object is characterized using majority share cote according to the labels of its adjacent. The object is classified basically since the class of the object that is nearest to it in the event when k=1. K is identified as an odd integer in such a situation, when just two classes remain present. When k is an odd whole number, then there can be a tie during the performance of multiclass classification.

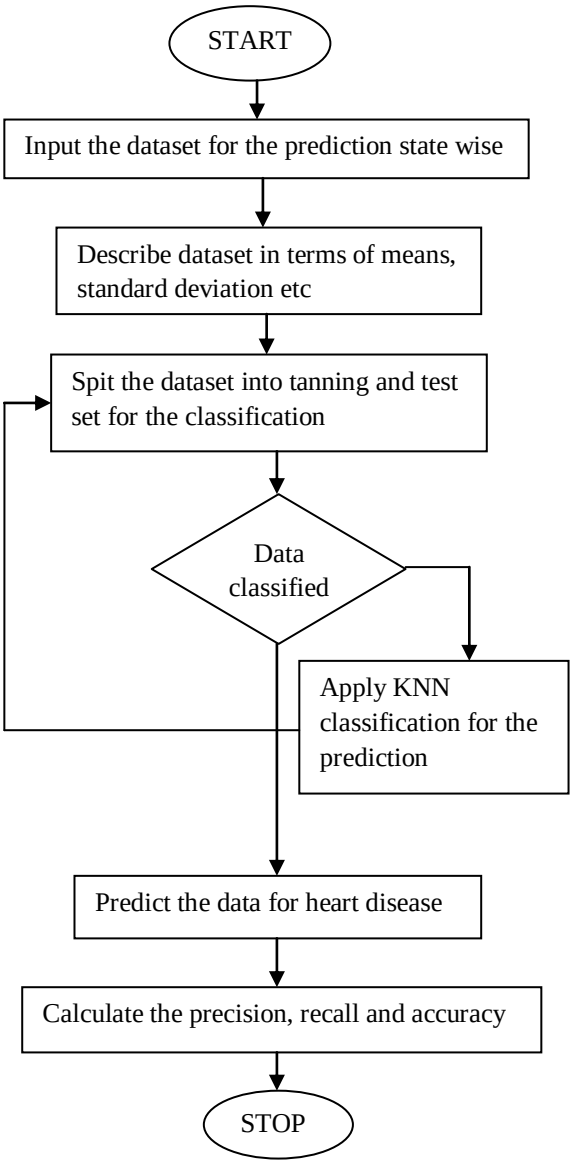

Fig 1: Proposed Methodology

RESULT AND DISCUSSION

A programming language named as Python is used for the execution of proposed technique. A number of comparisons were performed between proposed and existing approaches for the inspection of outcomes. These comparisons were performed on the basis of accuracy and execution time.

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

Accuracy =

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

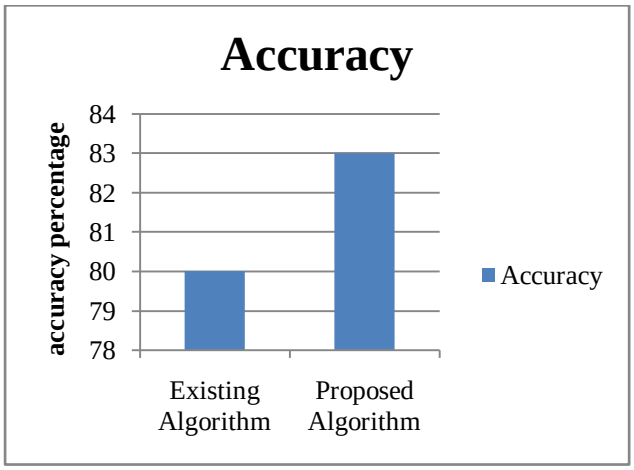

Fig 2: Accuracy Comparison

The earlier and proposed algorithm is compared in terms of accuracy as depicted by the figure 2. The proposed algorithm shows higher accuracy rate in comparison with earlier algorithm.

2. *Execution Time:* The difference of end time when algorithm stops performing and starts time when algorithm starts performing is identified as execution time as depicted by the eqn. 2.

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

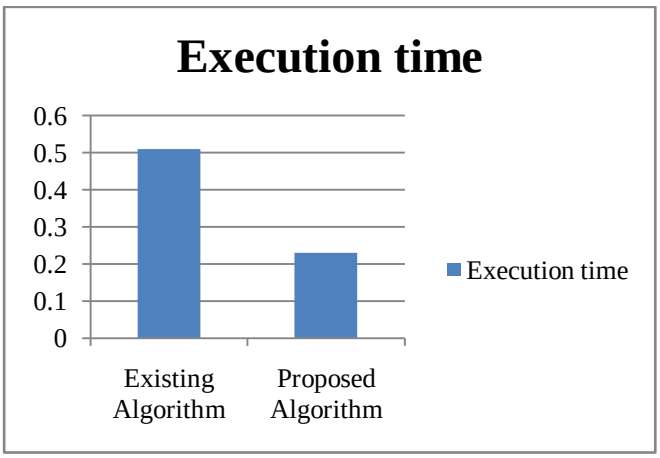

Fig 3: Execution time

The earlier and proposed algorithm is compared in terms of execution time as depicted by the figure 3. The proposed algorithm takes less execution time in comparison with earlier algorithm

3. CAP Analysis: -A canonical analysis on the primary coordinates for any similarity matrix involves a permutation test. CAP considers data structure. Therefore, different levels can be separated in the absence of some strong difference. This approach is good for depicting the communication between dynamics

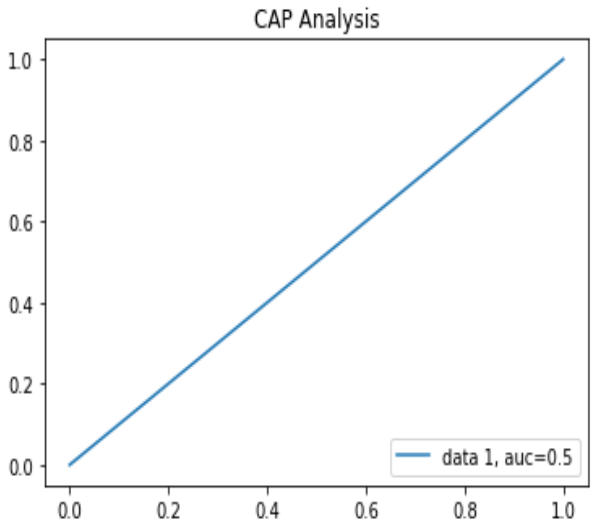

Fig 3: CAP Analysis

The figure 3 depicts the CAP analysis. The training data suite is specified as input on the x-axis of this curve. The test data is given as input on y-axis. The blue line depicts that CAP curve. The accuracy of classifier is depicted through this curve.

CONCLUSION

The useful data is extract from raw dataset with the help of data mining approach. The alike and unlike information is clustered after measuring a resemblance among input dataset. Both alike and unlike data types are classified using SVM classifier. In this classifier, arithmetic mean of the dataset is computed for measuring the data point. The central point computed as Euclidian distance is utilized foe computing the resemblance present among different data points. A clustered data is classified with the help of SVM classifier on the basis of input data type. In this study, a combination of back propagation algorithm and SVM classifier is implemented for increasing the prediction accuracy. The proposed algorithm shows good in terms of accuracy and execution time. In future, the proposed approach will be further enhanced to design hybrid classifier for the prediction of heart diseases.

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. PhattharatSongthung and KunwadeeSripanidkulchai, 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, MichelineKamber, “Data Mining: Concepts and Techniques”, vol. 3, pp. 1-31, 2000.
5. Ms. Tejaswini U. Mane, “Smart heart disease prediction system using Improved K-Means and ID3 on Big Data”, 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), vol. 8, issue 11, pp. 123-148, 2017.
6. SellappanPalaniappan, RafiahAwang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, vol. 5, issue 1, pp. 13-28, 2008.
7. KanikaPahwa, Ravinder Kumar, “Prediction of Heart Disease Using Hybrid Technique For Selecting Features”, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), vol. 4, issue 5, pp. 23-48, 2017.
8. BayuAdhi Tama,1Afriyan 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. ZhiqiangGe, 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. JahinMajumdar, 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.