ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 11, November 2019

Heart Disease Prediction Approach Using Machine Learning

Er. Kumari Samriti[1], Er. Poonam Chaudhary[2]

M.Tech Student, Department of Computer Science & Engineering, SIRDA Institute of Engineering Technology Sunder Nagar, Mandi, India[1]

Assistant Professor, Department of Computer Science &Engineering, SIRDA Institute of Engineering Technology Sunder Nagar, Mandi, India[2]

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 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.

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. First step is data cleaning which is used to remove noise and irrelevant data. Second, step is data integration which is used to combine multiple data sources. In third step data are retrieved from the database that comes under the step of data selection. When aggregations and summary operations are applied, the transformation or consolidation of data is done such that appropriate data can be generated within the fourth step. For the extraction of data patterns, data mining is an important process which applies different intelligentmethods then knowledge based interesting patterns is identified using pattern evaluation. With the help of knowledge representations and visualization approaches, the users are presented with the mined knowledge within the final step [2]. The data mining process extract large amount of data in order to acquire knowledge which is termed as a misnomer.It became overwhelming as the massive collections of data stored and it was difficult to handle such huge information. Therefore, to solve these issues database management system (DBMS) and structured databases are created. Whenever, it is required to retrieve the particular information from large amount of data an important role is played by efficient database management systems. It became easier to gather all sorts of information due to increase usage of database

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 11, November 2019

management systems [3]. The data warehouses are the houses where large amount of data is stored from the multiple sources. Data mining acts as a powerful tool to most of the companies as it helps them to reduce information present in their data warehouses. For the differentiation of data mining tools, an automated analysis process is utilized such that with the help of historical data, new information can be discovered [4]. Within reasonable time, data mining is used such that the large set of data can be analyzed in reasonable time. When analysis can be done for several fields or variables, this method can be applied to small amount of data. This approach provides easy and effective solutions for small relatively simple data analysis. Important data available within an unorganized manner can be discovered through the useful view of using data mining.Few of the applications in which cluster analysis is applied include pattern recognition approach, image processing and data analysis. The customer categorized group and purchasing patterns done by clustering can be used by marketer to discover their customer's interest [5]. For generative taxonomies of plants and animals, classifying the genes that function in similar manner and study the structures inherent within the populations, this approach is today applied in biological research areas. In a city, similar houses and lands area can be identified by employing clustering in geology. To discover new theories, information clustering can be used that classify all documents available on Web. The unsupervised data clustering classification method create clusters, group of objects in such a way that objects in different clusters aredistinct and that are in same cluster are very similar to each other. One of the traditional topics that are also known to be the initial step for knowledge discovery within data mining is known as cluster analysis. In order to cluster the tasks being performed in low dimensional data sets, the k-means clustering algorithm is applied. K is utilized as a parameter here and the k clusters are generated by partitioning n objects [6]. It ensures that the similar types of objects are grouped within one cluster and dissimilar objects are placed in separate clusters. The cluster centres are identified here with the help of this algorithm.Supervised and unsupervised learning are the two methodologies utilized by the data mining. In order to learn the parameters of the model, a training set is utilized in supervised learning while in case of unsupervised learning no training set is utilized, for example k-means clustering. Classification and prediction are the main objective of the data mining [7]. Data process rate and unordered values or data are classified by the classification models while continuous value is predicted by the prediction models. A binary classifier through which the margin is increased is known as SVM classifier. This algorithm helps in performing classification in which all the data points present in individual class are separated by the best hyperplane. The best hyperplace of SVM can be presented in the basis of highest margin present in the two classes.

II. LITERATURE REVIEW

Min Chen, et.al (2017) proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm [8]. 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 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.

Akhilesh Kumar Yadav, et.al (2013) presenteddifferent analytic tools used to extract information from large datasets such as in medical field where a huge amount of data is available [9]. The SGPGI real data set has been used that are always linked with different challenges. The classification becomes inefficient due to noise, high dimensional and missing values. Due to the different challenges have to face while performing data analytics clustering is used in replace of it. The foggy k-mean clustering based novel technique need to be developed is the main focus of authors. The proposed algorithm has

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 11, November 2019

been tested by performing different experiments on it that gives excellent result on real data sets. In real world problem enhanced results are achieved using proposed algorithm as compared to existing simple k-means clustering algorithm.

Sanjay Chakrabotry, et.al (2014) presented the use of clustering for different forecasting [10]. The weather forecasting has been performed using proposed incremental K-mean clusteringgeneric methodology. The purpose behind this paper is to analyze air pollution for it they have used dataset of west Bengal. The clusters peak mean values are used to develop a weather category list and K-means clustering is applied on the dataset of air pollution. The weather category has been defined in different clusters and a new data is checked by incremental K means to group it into existing clusters. The future weather information is able to be predicted using proposed approach. The used data set contains the weather forecasting information of west Bengal that is able to reduce the air pollutions consequences. The weather events forecasting and prediction becomes easy using modeled computations.In the last the authors have performed different experiments to check the proposed approach correctness.

Chew Li Sa, et.al (2014) proposed Student Performance Analysis System (SPAS) for keeping track of the record of the performance of the students of a particular university [11]. The design and analysis has been performed to predict student's performance using proposed project on their results data. The data mining technique generated rules that are used by proposed system to gives enhanced results in predicting student performance. The student's grades are used to classy existing student using classification by data mining technique.

Qasem A. Al-Radaideh, et.al (2013) proposed that data analysis prediction acts as an important subject for forecasting stock return [12]. The data analysis future can be predicted through past investigation. The past historical knowledge of experiments has been used by stock market investors to predict better timing to buy or sell stocks. There are different available data mining techniques out of all a decision tree classifier has been used by authors in this work.

K. Rajalakshmi, et.al (2015) presented a study related to the fast growing medical field [13]. In this field every single day a large amount of data has been generated and to handle this much of large amount of data is not an easy task. So, this data need to be handled properly for it different technologies need to be used after that a data need to be mined to turn it into useful pattern. The medical line prediction based systems optimum results are produced by medical data mining. The K-means algorithm has been used to analyze different existing diseases. The cost effectiveness and human effects has been reduced using proposed prediction system based data mining.

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 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.

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

Prediction Phase: - In the phase, the input dataset is divided into training, test part. The KNN classifier is applied for the prediction analysis which take input test and training data and output in the form of predicted data. Since there are no

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 11, November 2019

assumptions made on the underlying data distribution, KNN is known to be a non-parametric supervised learning algorithm. 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 k nearest 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. The classification of samples on the basis of majority class of its nearest neighbor is the major task of KNN algorithms.

$$Class = arg_v max \sum_{(X_i, y_i) \in D_z} I(v = y_i)$$

A set of labeled objects, a distance or similarity metric that calculates the distance amongst objects and the number of nearest neighbors that is the value of k, are the three important elements within the KNN approach. In order to make the recognition task successful, the selection of an appropriate similarity function as well as value for parameter k is important.

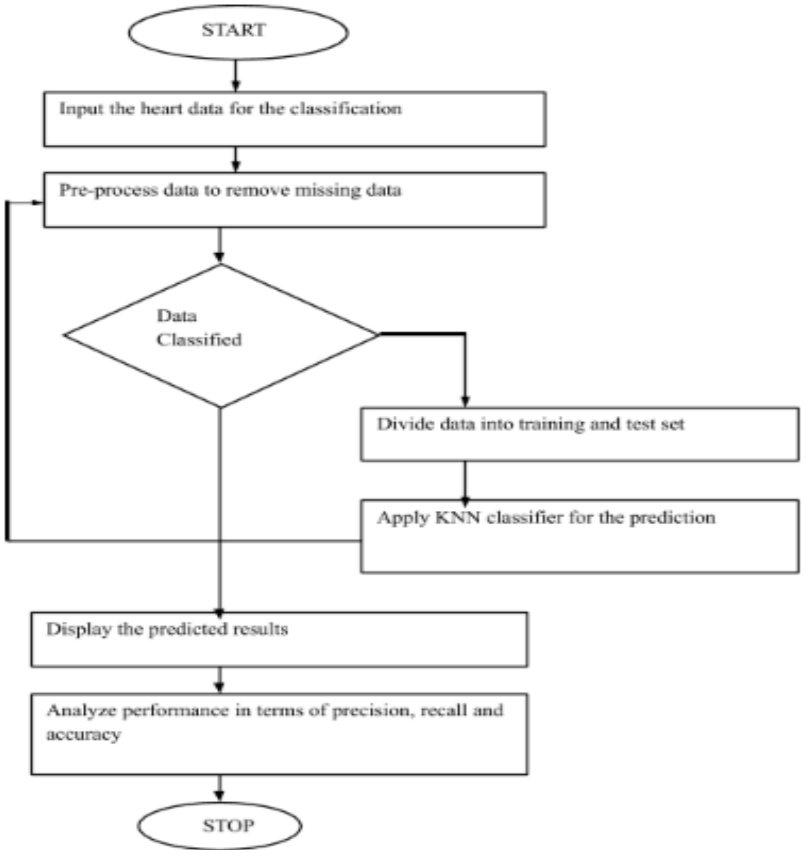

Figure 1: Proposed Flowchart

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 11, November 2019

IV. EXPERIMENTAL RESULTS

The proposed research is implemented in Python and the results are evaluated by making comparisons against proposed and existing techniques in terms of various parameters.

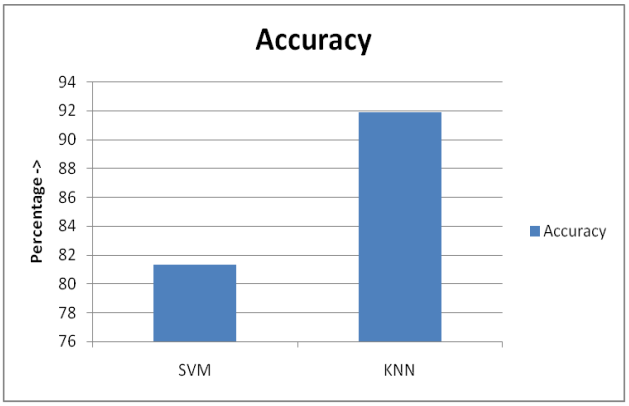

Fig 2 Accuracy Comparison'

As shown in figure2, the accuracy of SVM classifier is compared with the KNN classifier for the heart disease prediction. The accuracy of SVM is approx 81.35 percent and KNN has the accuracy of 91.87 percent

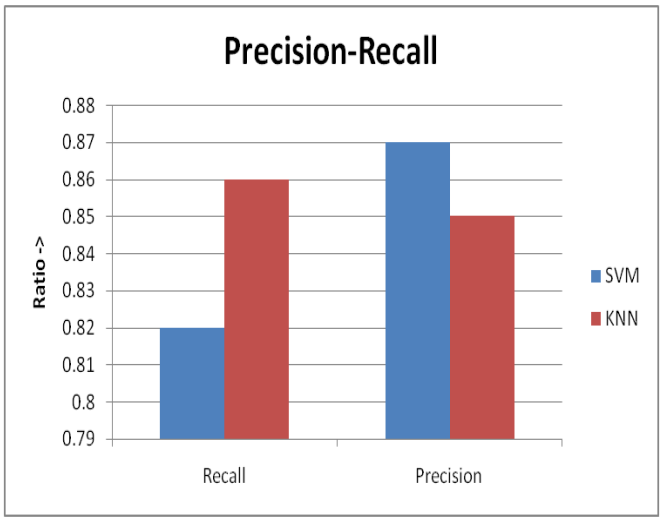

Fig 3. Precision-Recall Comparison'

As shown in figure 3, the precision-recall of SVM classifier is compared with the KNN classifier for the heart disease prediction. The recall of SVM is approx 82 percent and KNN has the precision value of 87 percent

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 11, November 2019

Table 1: Performance Analysis

Algorithm	Recall	Precision
SVM	0.82	0.87
KNN	0.86	0.85

V. CONCLUSION AND FUTURE WORK

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 KNN is applied for the heart disease 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

REFERENCES

[1] AnandBahety, "Extension and Evaluation of ID3- Decision Tree Algorithm", ICCCS, ICCC, vol. 4, issue 1, pp. 23-48, 2014.
[2] K. ZakirHussain, M. Durairaj, G. RabialahaniFarzana. "Criminal Behavior Analysis By Using Data Mining Techniques", IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012), vol. 4, issue 1, pp.30-31, 2012.
[3] Prashant K. Khobragade, Latesh G. Malik, "Data Generation and Analysis for Digital Forensic Application using Data mining", Fourth International Conference on Communication Systems and Network Technologies, vol. 4, issue 1, pp. 23-48, 2014.
[4] Tejaswini U. Mane, Mrs. Asha M. Pawar, "A Survey On Big Data And Its Mining Algorithm", IJIRCCE, Vol. 3, Issue 12, pp. 12-22, 2015.
[5] 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 12, pp. 123-128, 2010.
[6] 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 4, pp. 959-963, 2012.
[7] K.Srinivas, B.Kavihta Rani and Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", (IJCSE) International Journal on Computer Science and Engineering, vol. 4, issue 9, pp. 23-48, 2010.
[8] 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.
[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.