Early Stage Disease Detection System

[1]Pooja Hande, [2]Anil Kadam
[1]Student, [2] Assistant Professor
[1] Department of Computer Engineering ,
[1] All India Shri Shivaji Memorial Society's College of Engineering, Pune-411001, Maharashtra, India.

Abstract : Data Mining is the method of tacit, previously unknown and potentially useful data extraction. A pattern is important when it is true with some degree of certainty, novelty, potentially useful and easily understood by humans for a given test data. The enormous amount of data created for disease prediction is too complex and voluminous for traditional methods to process and analyze. Advanced data mining tools by discovering hidden patterns and useful information from complex and voluminous data overcome this problem. Researchers reviewed literature on disease prediction using data mining techniques and stated that vector machine help overcomes all other techniques with higher accuracy levels. Applying data mining techniques to data on healthcare will help predict the probability that patients will become sick. This paper highlights the important role played in predicting and diagnosing disease through data mining techniques in analyzing huge volumes of healthcare related data.

Index Terms - Data mining, support vector machine, prediction.

1. INTRODUCTION

Data mining is the process of finding previously unknown patterns and secret healthcare databases information. Data mining incorporates statistical analysis, algorithms for machine learning and database technology to retrieve secret patterns and relationships from large databases. Data mining is now becoming popular in the healthcare sector, as efficient analytical methodology is required to detect unknown and useful healthcare knowledge. Cardiovascular disease (also known as heart disease) is a form of heart or blood vessel disease (arteries, capillaries, and veins). Cardiovascular disease is the world's leading cause of death, with cardiovascular mortality rates declining in many high-income countries since the 1970s. At the same time, in low- and middle-income nations, cardiovascular deaths and disease have increased rapidly. While disease typically affects older adults, the history of disease, especially atherosclerosis; begins early in life making primary prevention efforts from childhood essential. Hence, expanded accentuation on forestalling atherosclerosis by changing danger factors, proof recommends various hazard factors for coronary illness, for example, age, sexual orientation, hypertension, high serum cholesterol levels, smoking, unreasonable liquor utilization, sugar utilization, family ancestry, heftiness, absence of physical movement, psychosocial factors, diabetes mellitus, air contamination and utilizing tobacco.

The World Health Statistics 2012 report edifies the way that one out of three grown-ups worldwide has raised circulatory strain– a condition that causes around half of the passings from stroke and coronary illness. Coronary illness is the significant reason for losses in the various nations including India. Coronary illness kills one individual in at regular intervals in the United States. Determination is confounded and significant assignment that should be executed precisely and productively. The analysis is regularly made, in light of a specialist's understanding and information. This prompts undesirable outcomes and extreme medicinal expenses of medications gave to patients. In this manner, a programmed restorative conclusion framework would be exceedingly helpful. This examination work is an endeavor to introduce the definite investigation about the various information mining strategies which can be sent in these robotized frameworks.

2. LITERATURE SURVEY

This paper targets breaking down the different information mining strategies presented as of late for coronary illness expectation. Various information mining methods have been utilized in the analysis of CVD over various Heart sickness datasets. A few papers utilize just a single strategy for conclusion of coronary illness and different specialists utilize more than one information digging procedures for the analysis of coronary illness.

Nidhi Bhatla et al. [1] perceptions uncovered that the Neural Networks with 15 traits performed better in correlation with other information mining procedures [1]. The examination study presumed that Decision Tree procedure demonstrated better execution with the assistance of hereditary calculations utilizing included subset determination. This examination work additionally proposed a model of Intelligent Heart Disease Prediction framework utilizing information mining methods specifically Decision Tree, Naïve Bayes and Neural Network. A sum of 909 records were gotten from the Cleveland Heart Disease database. The outcomes revealed in the exploration work advocated the better execution of Decision Tree methods with 99.6% exactness utilizing 15 characteristics. Be that as it may, Decision tree procedure in mix with hereditary calculation the exhibition detailed was 99.2% utilizing 06 characteristics.

V. Manikandan et al. [2] suggested that affiliation rule mining is utilized to remove the thing set relations. The information order depended on MAFIA calculations which brought about better precision. The information was assessed utilizing entropy based cross approval and parcel strategies and the outcomes were analyzed. MAFIA (Maximal Frequent Itemset Algorithm) utilized a dataset with 19 characteristics and the objective of the exploration work was to have exceptionally exact review measurements with more significant levels of accuracy.

Chaitrali S. Dangare and Sulabha S. Apte [3] demonstrated that Artificial Neural Network outflanks other information mining methods, for example, Decision Tree and Naïve Bayes. In this exploration work, Heart malady forecast framework was created utilizing 15 characteristics [3]. The exploration work included two additional characteristics weight and smoking for proficient determination of coronary illness in creating powerful coronary illness expectation framework.

Williamjeet Singh and Beant Kaur [5] distributed an examination paper in IJRITCC "Survey on Heart Disease utilizing Data Mining Techniques". The creator referenced crafted by huge number of specialists and thought about different information mining methods dependent on execution and precision.

Jyoti Sonia, et.al. [6] in year 2011 exhibited three classifiers Decision Tree, Naïve Bayes and Classification by means of grouping to analyze the nearness of coronary illness in patients. Order by means of bunching: Clustering is the way toward gathering comparable components. This system might be utilized as a preprocessing step before sustaining the information to the characterizing model. Investigations were led with WEKA 3.6.0 apparatus. Informational index of 909 records with 13 distinct characteristics. All traits were made absolute and irregularities were settled for effortlessness. To improve the forecast of classifiers, hereditary hunt was consolidated. Perceptions display that the Decision Tree information mining method beats other two information mining systems in the wake of fusing highlight subset choice yet with high model development time.

Vikas Chaurasia, et.al. [9] In their examination work utilized three well known information mining calculations CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and choice table (DT) separated from a choice tree or rule-based classifier to build up the forecast models utilizing a bigger dataset. Perception demonstrated that presentation of CART calculation was better when contrasted and other two order techniques.

Abhishek Taneja [10] explore work was meant to structure a prescient model for coronary illness identification utilizing information mining strategies from Transthoracic Echocardiography Report dataset that is equipped for improving the unwavering quality of coronary illness determination utilizing echocardiography. The models were based on the preprocessed Transthoracic Echocardiography dataset with three diverse managed AI calculations J48 Classifier, Naïve Bayes and Multilayer Perception utilizing WEKA 3.6.4 AI programming. The presentation of the models was assessed utilizing the standard measurements of exactness, accuracy, review and F-measure. The best model to foresee patients with coronary illness seemed, by all accounts, to be a J48 classifier executed on chosen properties with a grouping precision of 95.56%. From a sum of 15 characteristics that were accessible, 8 properties that were profoundly pertinent in foreseeing coronary illness from Transthoracic Echocardiography dataset were chosen in the examination work.

3.METHODOLOGY

3.1 Proposed System:

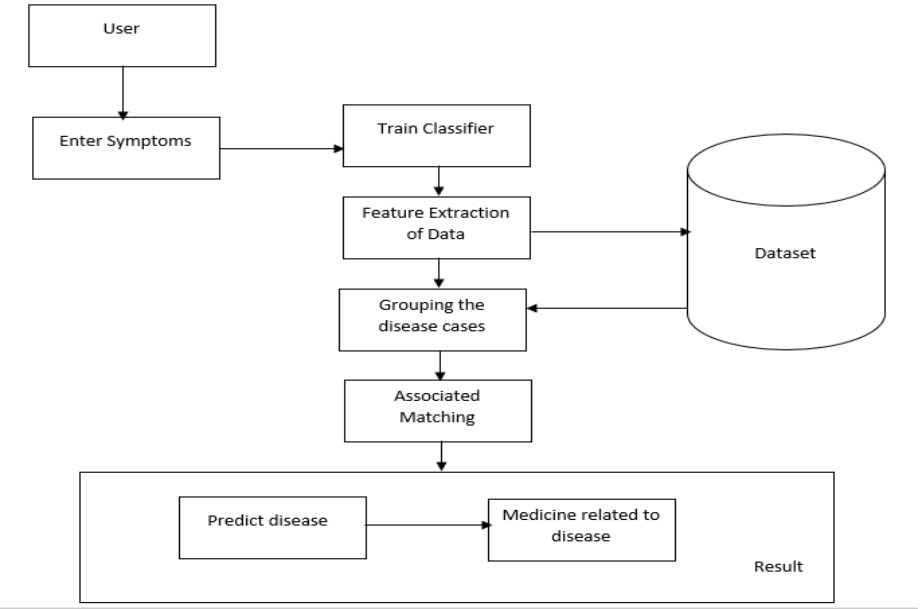

Fig 1: System Architecture.

Machine Learning (ML) offers methodologies, methods, and tools that can help solve analytical and predictive hitches in a number of medicinal fields. ML is used to examine the wild of managed edges and their mixtures for forecasting, e.g. disease production forecasting, elimination of medicinal knowledge for outcome analysis, direction and provision of care, and for the overall sustainability organization. ML is used to examine the wild of managed edges and their mixtures for forecasting, e.g. disease production forecasting, elimination of medicinal knowledge for outcome analysis, direction and provision of care, and for the overall sustainability organization. It is contended that the successful presentation of ML attitudes can help the tally of computer-based structures in the healthcare setting providing chances to ease and enhance the exertion of medical boffins and eventually to recover the competence and excellence of medicinal repair. Below, it précises some main ML requests in medicine. Machine Learning learns the data and produces the result.

Classification is one of the well-known data mining issues. Data / objects should be divided into different classes or categories. For example, data can be broken down by file type, average file size, gigabytes, and megabytes of topical content. Classification is the learning process of a method that can be used to assign data objects to a subset of a class set. Some types of classification goals, first finding a good general that can predict with high accuracy the class of but far unknown data objects. Second, to consider each other's compact and easy to understand class pattern.

Cluster is a set of items. For example, data elements in different similarity groups between the data set in cluster classes in a single group cluster partitions. Every object nearby is a part of the neighborhood. There are two cluster targets. First, an intra-class is an inter-class second. Inter-cluster implies increasing the distance from the cluster. Intra-cluster implies reducing cluster distances. In this process of selecting a subset of appropriate features for use in model design, feature selection also known as variable selection attribute selection and variable subset selection.

3.2 Algorithm: Support Vector Machine

Support vector machine (SVM) is a machine learning algorithm which is used for solving classification and regression problems. SVM was first introduced by Vladimir Naumovich Vapnik and his colleagues in 1992. SVM regression is considered a nonparametric technique because it relies on kernel functions. It uses the maximum margin algorithm: a non-linear function is leaned by linear learning machine mapping into high dimensional kernel induced feature space.

1. Prepare the pattern matrix
2. Select the kernel function to use
3. Select the parameter of the kernel function and the value of C
    i. use the values suggested by the SVM software, or you can set apart a validation set to determine the values of the parameter.
4. Execute the training algorithm and obtain the $\propto i$
5. Unseen data can be classified using the $\propto i$ and the support vectors.

4 . RESULTS AND DISCUSSION:

Author	Technique	Accuracy
Jyoti Sonia, et.al.	Naïve Bayes, Decision Tree, KNN	77%
K.Srinivas et.al.	Naïve Bayes, knn and D.L.	81.11% 81.48% 81.11% 80.96%
Nidhi Bhatla et.al.	Naïve Bayes, Decision Tree, Neural Network	84.5%
Chaitrali S. dangare & Sulabha S. Apte, Abhishek Taneja	Naïve Bayes, Decision Tree, Neural Network	87.0%
R. Chitra et. al.	Naïve Bayes,J48 unpruned tree, Neural Network	87.0%
Vikas Chaurasia, et.al.	Hybrid Intelligent Techniques	78.9% 81.41%
Proposed System	**SVM**	**94.5%**

Table 1:Result

5 . CONCLUSON

The goal of this research work is to provide insight into various data mining techniques that can be used in automated prediction systems for disease. Disease is one of the leading causes of death worldwide and it is very difficult to predict early different disease. As a method for diagnosing illness, the computer-aided disease prediction program helps the doctor. Some of the classification systems for diseases have been analyzed in this report and it has been concluded that data mining plays a major role in the classification of diseases based on various research studies. Supporting Vector Machine with offline preparation is a good tool for early-stage disease prediction. The system's good performance can be accomplished by preprocessed and structured data set. By reducing features and using different techniques, the classification accuracy can be improved. After evaluating the different results published in the checked research studies, supporting vector-based techniques with 15 attributes showed better performance using 15 attributes of 99.62 percent. The various data mining techniques will assist a specialist in successful decision-making and better diagnosis, contributing to optimal disease prevention in the healthcare sector.

ACKNOWLEDGMENT

This paper would not be possible without the contributions of numerous researchers in this ever growing field. Thus we would like to thank all of them.

REFERENCES

[1] Nidhi Bhatla, Kiran Jyoti, "An Analysis of Disease Prediction using Different Data Mining Techniques" International Journal of Engineering and Technology Vol.1 issue 8 2012.

[2] V. Manikandan and S. Latha, "Predicting the Analysis of Disease Symptoms Using Medical Data Mining Methods "International Journal of Advanced Computer Theory and Engineering", Vol. 2, Issue. 2, 2013.

[3] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques." International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[4] Shadab Adam Pattekari and Asma Parveen, "Prediction system for heart disease using naïve bayes", International Journal of Advanced Computer and Mathematical Sciences, 2012.

[5] Beant Kaur and Williamjeet Singh., "Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC ,October 2014.

[6] Jyoti Soni et.al. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[7] B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014.

[8] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014.

[9] Vikas Chaurasia, et al, "Early Prediction of Heart Diseases Using Data Mining Techniques", Caribbean Journal of Science and Technology ISSN 0799-3757, Vol.1,208-217, 2013.

[10] Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques" Oriental Journal of computer science & Technology ISSN: 0974-6471 December2013

S