Heart Disease Analysis Research Using K-Nearest Neighbor: a Review

Neelamani Samal, Manjinder Kaur, Rohit Kumar Singhal and Jai Sukh Paul Singh

July 27, 2023

Heart Disease Analysis research using K-Nearest Neighbor : A Review

[1]Neelamani Samal, [2]Manjinder Kour, [3]Rohit Kumar Singhal, [4]Jai Sukh Paul Singh

[1]Department of Computer Science and Engineering, Chandigarh University, India, neelamani.samal@gmail.com
[2]Department of Computer Science and Engineering, Chandigarh University, India, kour.manjinder897@gmail.com
[3]Department of Computer Science and Engineering, Chandigarh University, India , rohit.e13497@cumail.in
[4]Department of Computer Science and Engineering, LPU,India, khalsa128@gmail.com

ABSTRACT

Analyzing complex data is called data mining. The process of deciding what will happen next is called forecasting. Many techniques for predictive analytics are used these days. Predictive analytics are performed using the SVM method. This technique splits the data into her two phases: testing and training. The 1st type of test data is primarily on people who have less or not at all of developing heart disease. The possible getting heart disease is over 50% in his second class of test data. In this work, we propose to improve existing methods using decision tree classifiers. This suggestion improves accuracy while reducing execution time.

Keywords: decision tree , SVM, KNN,

1.1 Introduction

Data mining (DM) is the process of sifting through the data and identifying the specifics and trends that may be utilized to infer important information about the user. Data mining software comes in a variety of forms and can be used to analyze various kinds of data. This approach of use the information ingrained in the many fields has developed into a significant one which can be applied to many different fields. To use only useful information, one of the key purposes of the EDI is to analyze the stored information [1]. We have some expertise in relational database data analysis, including object relational databases, multimedia assets, and personal information (medical records).

Processing should first clean the data to remove noise and unnecessary data in order to acquire a good rate. Data integration is the next step, which combines various data sources into a single, coherent entity. In the second step you choose which data to retrieve from the database and the computer records the information [2]. The collection of vast amounts of data enables the analysis of how the general public is adapting to social change, regardless of the specific method used to collect the data. Exploring the data may require some manipulation of the data to extract the exact results. However, because gold is typically extracted from rock or sand, the term "mining" is more frequently used to describe gold extraction than rock or sand extraction.

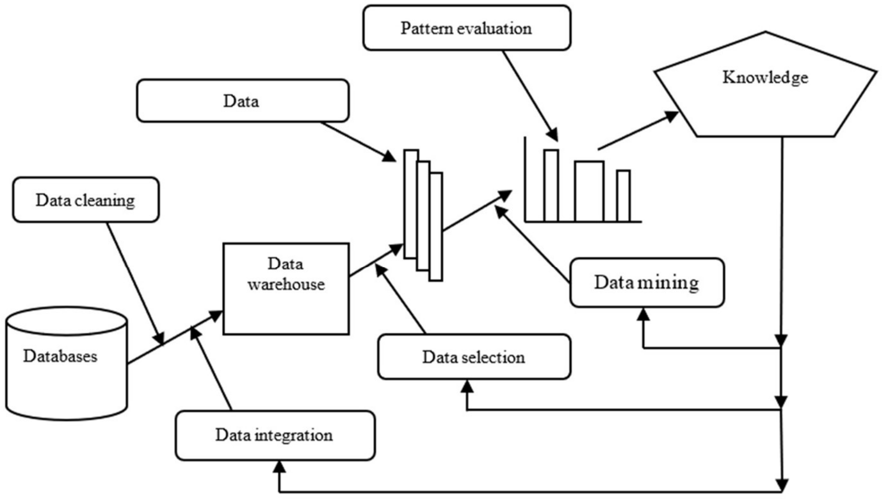

Fig. 1.1 Phases of Data Mining

The process of getting raw materials is known as mining, which is defined as the act of pulling something out of the ground. A number of names are used to explain about knowledge extraction and its utility. Data archeology, knowledge mining from databases, data mining. In a world full of information, satellites, computers, and other advanced technological devices make that information available only to the most powerful. A central computer system that houses a variety of data, including the medical data required to follow each patient, provides quick access to a wide range of information.

1.1.2 Approaches to data mining

a) **Association:** Associations are among the best data mining techniques. For transmission of similar data, patterns are discovered based on relationships, using associations of specific items to other items. This association technique is informative, used to predict heart disease, and used to analyze relationships between various attributes. All patient risk factors are identified and used to predict disease.

b) **Classification:** On the basis of machine learning, classification is a well-known data mining method. Each element in the dataset is classified into a existing collection of classes or groups during data classification. Decision trees, linear programming, neural networks, and statistics are some of the mathematical methods used in classification systems [6].

c) **Clustering:** An approach to data mining called clustering finds groups of objects. Having similar properties, this can be done automatically. Unlike classification, where objects are allocated to already defined classes, clustering techniques define classes and place objects into them. A cluster set is obtained using a clustering method for predicting heart disease and contains a list of patients who share the same risk factors. As a result, another patient list is created using this technique.

d) **Prediction:** To find connections between independent and dependent variables, data mining techniques such as forecasting are applied. This approach can be used in many areas, including sales, to predict future profits. Therefore, profit is called the dependent variable and sales is called the independent variable. Using historical sales and profit data, you can create a adjusted regression curve and use it to predict profit.

Data Mining Issues

DM algorithms are methods, some of which have been around for a long time. However, it can later be used as a reliable and scalable tool, outperforming previous traditional statistical methods. Despite its youth, DM has gained widespread attention and become ubiquitous. Before DM to be regarded as a conventional, mature, and reputable field, a number of concerns need to be resolved. They are divided into the following categories:

a) **Social and security concerns:** This is a significant subject in gathering data to be shared and suggested to make strategic decisions. In addition, a large amount of confidential and personal information about individuals or companies is collected and stored after the data has been collected for many purposes. The sensitivity of this data, as well as the unauthorized access to information, is notorious. In addition, new and implied information about individuals or groups is disclosed by DM. This may violate your privacy policy, especially if the discovered information could be disseminated.

b) **User interface issues:** The value of the information gleaned through DM tools depends on how entertaining and clear it is to the user [7]. The fine data visualization makes it simple to understand the results of the DM and helps users comprehend their needs. The fundamental issues with user interfaces and graphics are screen space and interaction.

c) **Mining methodology issues:** These problems are associated with the application of DM techniques and their shortcomings. The choice of mining methods is based on factors including the flexibility of mining approaches, estimation of the researched knowledge, utilization of background information and the data about the data , controlling and managing of noise in this kind of data, etc. To give an example, it is necessary to have access to a variety of DM techniques because different approaches can be used in different ways depending on the data. Additionally, many strategies are appropriate and capable of handling user requirements in various ways [8].

d) **Performance issues:** Various AI and statistical techniques are available for data analysis and interpretation. However, the development of these techniques has not been done for the massive datasets that DMs are dealing with these days. TB size is very common. Problems arise regarding the scalability and effectiveness of DM techniques when dealing with large amounts of data. Other topics are incremental updates and parallel programming. Only when the dataset can be partitioned and the results afterwards combined can parallelism help with size difficulties. When new data becomes available, incremental updates are required for combining parallel mining results and updating DM results without reanalyzing the entire dataset [9].

Machine Learning

Computer systems may directly learn from examples, information, and experience thanks to a branch of artificial intelligence known as "machine learning". Computers can now carry out specialized jobs intelligently thanks to machine learning systems that can carry out complicated procedures by learning from data rather than by following pre-programmed instructions. In recent years, the remarkable progress is made in the ML has expanded its potential for various application. Increased data availability has made it possible to train machine learning systems on large sample pools. Furthermore, the analytical capabilities of these systems have been enhanced by greater processing power [10]. There have also been advances in algorithms in this area that increase the power of machine learning. These developments have allowed systems that were much below human levels just a few years ago to currently execute some jobs better than humans.

Literature Review

Tülay Karayilan, et.al, “Prediction of Heart Disease Using Neural Network”, 2017 Heart Disease is a deadly illness that many people currently suffer because its detection and prevention are very important and early diagnosis is necessary. It's a sickness. The diagnostic process for these diseases is complicated by the need for proper monitoring. Therefore, accurate and required early disease detection is needed. This disease causes the greatest number of causalities [11]. Conventional methods have various limitations as they are analyzed in experiments, so an improved method is proposed in the article.

Ms. Tejaswini U. Mane, et.al "Smart Heart Disease Prediction System Using Improvement K-Means and ID3 on Big Data", 2017. The WHO's global heart disease survey, which kills more than 12 million people each year, Presented here from this deadly disease. Therefore, detection of this disease is required, resulting in maximum losses. The Hadoop Map platform is being used to condense huge data utilized in cardiology, which is frequently referred to as a big data strategy [12].

"Prediction of Heart Disease Using Hybrid Technique For Selecting Features," Kanika Pahwa et al., 2017. There is a significant amount of data available in the healthcare industry that must be discovered using hidden patterns based on requirements. Advances in data mining techniques are required in this area to enable effective decision making. The proposed method's functionality can be used with random forest and naive Bayes methods. The method's performance level can be improved based on the results [13]. The SVM-RFE and gain ratio algorithms were used to select the features from the dataset. After using this technique, each feature is given a specific weight. The studies' findings support the assertion that the suggested method offers the best accuracy while taking the shortest amount of time to compute.

"Heart Disease Diagnosis Using Data Mining Technique," Sarath Babu et al., 2017. Data integration is the next step, which combines various data sources into a single, coherent entity. It is the process of gathering useful information for use in implementation. In the medical dataset or medical sector, the mining, which has the huge capability to uncover the unseen patterns, is crucial. These patterns have found applications in clinical diagnosis, where the collection of data needs to adhere to a structured formats. To evaluate whether a patient has heart disease, characteristics including age, gender, blood pressure, and blood sugar levels are taken from medical profiles[14].

"A Non-Invasive Technique of Early Heart Disease Prediction from Photoplethysmography Signal," Monira Islam et al., 2017. A Non intrusive technique for detecting heart rate from a photoplethysmography signal (PPG) has been proposed[15]. This proposed technique will be extremely useful in identifying heart disease. Because conventional sensors can cause tissue damage during cardiac signal extraction, this results in significantly less discomfort for each patient. Using PPG and extracting from videos of human faces, a convenient cardiac detection mechanism is used. This could replace expensive ECG devices used to detect heart disease. Your heart rate can be determined using FFT, which can then be compared to the ECG machine's baseline heart rate.

Tahira Mahboob et.al, "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics",2017 The author examined different learning methods to support the identifying myriad heart diseases [16]. Data mining, support vector machines, computational intelligent classifiers, hidden Markov models, and other specific methods were applied. These types of advanced techniques have been developed to overcome this problem because the treatment of heart disease is very expensive and inaccessible to normal people. It also helps with early-stage predictions. Avoid all other future illnesses by making small changes to your routine. Thus, the authors' conclusion is that the proposed approach offers several benefits and is highly valuable.

Procheta Nag et.al, “A simple acute myocardial infarction prediction system using clinical data and data mining techniques,” 2017 Patients with cardiac arrest have chest pain or other early symptoms [17]. Clinical data from hospitalized patients with acute myocardial infarction were segmented to create a prototype (AMI). Heat stroke is accompanied by several symptoms such as chest pain, breathing problems, heart palpitations, nausea, vomiting, and continuous sweating. The investigation of heart attack incidence and classification of the whole range of heart attack symptoms using decision trees, a subset of data mining, and random forests leads to the conclusion that people are becoming more reliant on computer technology in the medical and healthcare areas. Data mining results are very informative and used to better support.

PROBLEM FORMULATION

Predictive analytics is a technique that utilizes current data to predict future outcomes, based on clustering and classification tools. The paper under review uses medical data from Central China between 2013-2015 to prepare modal data for predictive analytics. The neural network-based modal data is grouped and split into test and train sets using a classification algorithm, and SVM classifiers are employed to classify data into distinct classes. However, the accuracy of predictive analytics may be impacted by the k-mean clustering approach, which relies on arithmetic averaging of the entire dataset to determine center points. This method may not be effective for complex datasets where establishing relationships between attributes can be challenging. To classify wheat production in this study, a decision tree classifier was used to divide it into multiple classes. Other classifiers could be used instead of the decision tree to improve classification accuracy.

Research Methodology

This paper is based on a heart disease prediction model. Predictive analytics is a technique that uses current data sets to forecast future possibilities. We apply the decision tree method to predictive analytics in advance in this study. The decision tree algorithm is one of the most basic machine learning algorithms. As no assumptions are made about data distribution, decision trees which are classified as nonparametric supervised learning algorithms. Samples are classified based on the closest training sample in the feature space. During the training process, feature vectors are saved along with the training image labels. Below are the techniques used in our research methodology to label the k-nearest neighbors based on certain initial predictions so that certain outcomes can be run through the classification process.

The training set and each result have another set of attributes called target or prediction attributes so that the processed algorithm predicts the results. Algorithms discover relationships between attributes that help predict outcomes. This algorithm provides a data set called the prediction set. This data set contains the same set of attributes, but no unknown predictive attributes. In order to aid in the prediction process, algorithms examine the input. The predictability of an algorithm is what determines its accuracy .

a) Genetic programming

Genetic programming (GP) is widely used in research to solve classification problems in data mining. The ability of genetic programming to effectively predict the rules naturally expressed in GPs is the most important factor that has led to its widespread use. Optimal results are produced by the GP with global search problems like classification. Several 'peaks' are present in the search space for classification this cause local search algorithms also known as simulated annealing that performs badly.

b) Neural networks

Neural networks are the interconnectivity between the processing elements also called units, nodes, or neurons. These networks are designed after the cognitive processes of the brain. These networks are used to predict new outcomes from the previous observations. In order to produce an output function all the present neurons within the network work together. The collective neurons performed the computational functions within a neural network still it is capable of producing the output function even some of the individual neurons are malfunctioning. Instead network remains robust and fault tolerant. Within a neural network, each neuron has an associated activation number and also a weight associated in each connection between the neurons .

c) Ant colony

Ant Colony algorithms is the natural inspired technique and by the behavior of ants as they help in finding the optimal path from the colony to food. They use the good paths within a graph in order to find optimal ways. Chemical called pheromones are deposited by the ant on their trails when they are travelling from one place to another in the search of food. Ants used these trails to find their way back to the colony without distracting and it is followed by other ants as well if they find path safe. Due to this effect more pheromone are deposited on the trail which cause the effect of reinforcing. If for the long way, these trails have not been utilized than the pheromone starts to evaporate. The density of pheromone remains high as the short and optimal paths are utilized again and again that provides faster rate to find food. Therefore, large number of ants travelling on the shortest path due to which density of the pheromone is increased that is followed by all ants as well. This behavior of the ants is copied by the Ant Colony algorithms by which an optimal within a graph is determined. In the initial stage small amount of pheromone is deposited on the trails of the ant randomly.

d) Statistical algorithms: ID3 AND C4.5

The ID3 algorithm was developed by J. Ross Quinlan at the University of Sydney and published in the 1975 book Machine Learning. The ID3 algorithm generates a classification model from your data. This method, also known as a supervised learning algorithm for different classes, has also been trained to facilitate class prediction for new elements. An ID3 identifies attributes and differentiates one class from another. All of these attributes must be known in order to be selected from a known set of values. Temperature and nationality are both valid attributes. ID3 used entropy as a statistical property to determine attribute importance. The entropy measure is used to determine an attribute's information content. So this is how the decision tree that will be used to test future cases is built.

Conclusions

Data mining is used to extract relevant information from unprocessed datasets. After calculating the similarity between the input datasets, similar and dissimilar data are clustered. SVM is utilized to categorize diverse data types, both alike and different, through a center point determined by calculating the mean of a dataset. The similarity between different data points is calculated using the calculated Euclidean distance of the center points. In the final step, the clustered dataset is classified using a decision tree classification scheme based on the type of input dataset. The decision tree classifier in this work takes the place of the SVM classifier. The decision tree's quick processing time and great classification accuracy are evident.

Future Scope

SVM classifiers are used in place of decision tree classifiers. We can use hybrid classifier types for predictive analytics in the future to improve the proposed algorithm. You can also compare its reliability to that of other classifiers.

References:

[1] Monali Dey, Siddharth Swarup Rautaray, 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] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (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. Fal Dessai, 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] Abhishek taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, vol. 5, issue 4, pp. 959-963, 2013.

[6] Kajal C. Agrawal and Meghana Nagori (2013), "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm", International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, issue 3, pp. 546-552, 2013.

[7] Swain Sunita, Badajena J Chandrakanta and Rout Chinmayee, A Hybrid Approach of Intrusion Detection using ANN and FCM, European Journal of Advances in Engineering and Technology, vol. 3, issue 2, pp. 6-14, 2016.

[8] Tetiana Gladkykh, Taras Hnot and Volodymyr Solskyy, Fuzzy Logic Inference for Unsupervised Anomaly Detection, IEEE First International Conference on Data Stream Mining & Processing vol. 4, issue 1, pp. 42-47, 2016.

[9] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Elsevier, vol. 8, issue 1, pp. 23-48, 2013.

[10] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur Diagnosis of valvular heart disease through neural networks ensembles, Elsevier,vol. 4, issue 1, pp. 23-48, 2009.

[11] Tülay Karayilan, et.al, "Prediction of Heart Disease Using Neural Network", IEEE Xplore: 02 November DOI: 10.1109/UBMK.2017.8093512, 2017

[12] Ms. Tejaswini U. Mane, et.al "Smart Heart Disease Prediction System Using Improvement K-Means and ID3 on Big Data", DOI: 10.1109/ICDMAI.2017.8073517, 2017

[13] Kanika Pahwa et al., Prediction of Heart Disease Using Hybrid Technique For Selecting Features," **DOI:** 10.1109/UPCON.2017.8251100 ,2017

[14] Sarath Babu et al., "Heart Disease Diagnosis Using Data Mining Technique," ,DOI: 10.1109/ICECA.2017.8203643 , 2017

[15] Monira Islam et al.,"A Non-Invasive Technique of Early Heart Disease Prediction from Photoplethysmography Signal," **DOI:** 10.1109/EICT.2017.8275222 ,2017

[16] Tahira Mahboob et.al, "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics", **DOI:** 10.1109/ITECHA.2017.8101920 , 2017

[17] Procheta Nag et.al, "A simple acute myocardial infarction (heart attack) prediction system using clinical data and data mining techniques", DOI: 10.1109/ICCITECHN.2017.8281809, 2017

[18] Karna Vishnu Vardhana Reddy et al, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators", Appl. Sci. 2021, 11, 8352. https://doi.org/10.3390/app11188352 , https://www.mdpi.com/journal/applsci

[19] Harshit Jindal et al, "Heart disease prediction using machine learning algorithms", IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 ,IOP Publishing doi:10.1088/1757-899X/1022/1/012072

[20] Dr. Poonam Ghuli et. al, " Heart Disease Prediction using Machine Learning ", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 9 Issue 04, April-2020

[21] M.Kalaivani, Dr.S.Anitha et.al, "An Efficient Heart Disease Prediction System based on Supervised Machine Learning Methods", International Journal of Computing Algorithm Volume: 10 Issue: 01 June 2021 ISSN: 2278-2397, Article · November 2022 DOI: 10.20894/IJCOA.101.0010.001.001

[22] Ch. Anwar ul Hassan et. al, "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers", Sensors 2022, 22, 7227. https://doi.org/10.3390/s22197227 https://www.mdpi.com/journal/sensors Sensors 2022, 22, 7227

[23] Jasmine R. Eddinger et .al. "Alcohol Use and Drinking Motives Among Suddenly Bereaved College Students", JOURNAL OF DUAL DIAGNOSIS https://doi.org/10.1080/15504263.2018.1531185
[24] Valdo Henriques And Reza Malekian ," Mine Safety System Using Wireless Sensor Network", Digital Object Identifier 10.1109/ACCESS.2016.2581844

[25] PRIYANGA eT AL, "A hybrid recurrent neural network-logistic chaos-based whale optimization framework for heart disease prediction with electronic health records ", Computational Intelligence. 2020;1–29, wileyonlinelibrary.com/journal/coin © 2020 Wiley Periodicals LLC

[26] Sibo Prasad Patro et al, "Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning ", Evolutionary Intelligence © Springer-Verlag GmbH Germany, part of Springer Nature 2020

[27] Jared C. Van Hooser et al, "Knowledge of heart attack and stroke symptoms among US Native American Adults: a cross-sectional population-based study analyzing a multi-year BRFSS database ", BMC Public Health (2020) 20:40 https://doi.org/10.1186/s12889-020-8150-x

[28] Sheng-Feng Sung et al, "Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning", International Journal of Medical Informatics 152 (2021) 104505, www.elsevier.com/locate/ijmedinf

[29] Dr. M. Kavitha et al, "Heart Disease Prediction using Hybrid machine Learning Model ", Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021] IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9

[30] Ankanksha Kumari et al, “A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms ”, Asian Conference on Innovation in Technology (ASIANCON), 2021

[31] Shuge Ouyang et. al, “Research of Heart Disease Prediction Based on Machine Learning”, 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2022.

[32] Kuldeep Vayadande et al, “Heart Disease Prediction using Machine Learning and Deep Learning Algorithms” ,International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022.

[33] Yu Lin Et al, “Prediction and Analysis of Heart Disease Using Machine Learning ” IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI), 2021.

[34] Gnaneswari G et. al, “Analysis of The Diagnostic Parameters of Heart Diseases and Prediction of Heart Attacks ”, IEEE 3rd Global Conference for Advancement in Technology (GCAT), 2022.