IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Early Heart Disease Prediction Using Data Mining Techniques: A Case Study

Shafquat Perween, Dr. Md. Masroor Ahmed
Research Scholar, Associate Professor
Magadh University Bodh Gaya, Bihar

ABSTRACT

Data mining is successful in e-business, publicizing, and marketing and it is now being applied in various other engineering divisions. Healthcare is one of the sectors that is only now established to be explored. The healthcare sector is typically abundant in information, yet not all of the necessary data is extracted for uncovering hidden patterns and making effective decisions. The discovery of these patterns and relationships is often overlooked. Sophisticated data mining modeling approaches can assist in improving this situation.

This study plans to utilize data mining Classification Modeling Techniques including Decision Trees, Naïve Bayes, and Neural Networks, in addition to the weighted association Apriori algorithm and MAFIA algorithm for predicting Heart Disease. By utilizing medical information like age, gender, blood pressure, and blood sugar levels, it is possible to forecast the chances of individuals developing heart disease.

Keywords: ***Data mining, healthcare, classification modeling, decision trees, Naïve Bayes, neural networks, Apriori algorithm, MAFIA algorithm, predictive analytics***

1. INTRODUCTION

Data mining consist of detection new patterns and leanings in databases that were previously unknown and using this information to create predictive models. Data mining uses statistical examination, machine learning, and database technology to uncover concealed patterns and relationships in extensive databases.

The 2015 WHO statistics report reveals that a third of adults globally suffer from high blood pressure, leading to just about 50% of deaths from stroke and heart attacks. Cardiovascular disease (CVD), or heart disease, covers various conditions that affect the heart beyond just heart attacks. Heart disease was the primary reason for loss of life in various countries, India included.

Heart disease results in the death of one person globally every 36 seconds. Coronary heart disease, Cardiomyopathy, and Cardiovascular disease are a few types of heart conditions. The term "cardiovascular disease" encompasses various conditions that impact the heart, blood vessels, and blood circulation throughout the body. Diagnosing is a complex and crucial job that must be done with precision and effectiveness. The doctor's experience and knowledge are typically used to make the diagnosis.

This results in conclusions that are not desired and high medical expenses for the treatments given to patients. Therefore, a system for medical diagnosis that operates automatically would be very advantageous. Our goal is to provide a comprehensive exploration of various data mining methods that can be utilized in automated systems Data mining involves uncovering new

patterns and trends in databases that were previously unknown and using this information to create predictive models. Data mining uses statistical analysis, machine learning, and database technology to disclose covered patterns and relationships in extensive databases.

2. METHODOLOGY

The goal of this paper is to investigate different data mining methods for precise identification of heart disease. Due to limitations in resources and the paper's focus, the main approach is to carry out a thorough review of literature from medical, computer science, and engineering journals and publications.

Review of Literature: Search and collect applicable research from platforms including PubMed, IEEE Xplore, and ACM Digital Library. Use search terms like "heart disease," "data mining," and "diagnosis" to gather up-to-date studies on data mining techniques in the healthcare field.

Criteria for inclusion: Choose high-impact articles from the last 5-10 years that concentrate on data mining methods for diagnosing heart disease. Incorporate research involving Classification Modeling Techniques such as Decision Trees, Naïve Bayes, Neural Networks, and association algorithms like Apriori and MAFIA.

Data Extraction: Gather relevant details from chosen articles, such as methods utilized, datasets employed, performance measures (e.g., precision, recall, specificity), and main outcomes related to predictive modeling for heart disease.

Combine and assess: Examine and contrast results from various studies to pinpoint patterns, advantages, and disadvantages of diverse data mining methods in predicting heart disease. Emphasize techniques that demonstrate potential for precise diagnosis through documented results.

3. RESEARCH FINDINGS

Data Mining in the Heart Disease Prediction.

Data taking out for the prediction of heart disease through data mining.

Various supervised machine learning algorithms such as Naïve Bayes, Neural Network, weighted association Apriori algorithm, and Decision algorithm were utilized to analyze the dataset in [1]. The experimentation involves utilizing the data mining software Weka 3.6.6. Weka is a set of machine learning algorithms designed for data mining resolutions. The algorithms can be used either directly on a dataset or invoked from your Java code. Weka has features for preprocessing data, classifying, predicting, grouping, identifying associations, and displaying information. It is also highly appropriate for creating new machine learning systems.

Decision Tree

Is a common classifier that is straightforward and simple to execute. No domain knowledge or parameter setting is needed, and it can handle high-dimensional data. It generates outcomes that are simpler to understand and analyze. The ability to access in-depth patients' profiles through drill through is exclusively found in Decision Trees.

Naïve Bayes

Is a type of statistical classifier that operates under the assumption of no correlation between attributes? This classification algorithm assumes that an attribute value for a specific class is independent of the values of other attributes, based on conditional independence. One benefit of Naïve Bayes is the ability to utilize the Naïve Bayes model without relying on Bayesian techniques

Neural Networks

Artificial intelligence systems based on the mechanism of the human brain.

An artificial neural network (ANN), commonly known as a "neural network" (NN), is a mathematical or computational model inspired by biological neural networks. In simpler terms, it replicates a biological neural system [9]. In feed-forward neural networks, neurons in the first layer pass their output to neurons in the second layer in one direction, showing that neurons do not receive input from the opposite direction. Each layer is linked to another layer through connections, with weights being assigned to each connection. Neurons in the input layer primarily function by distributing input xi to neurons in the hidden layer. The hidden layer neuron combines input signal xi with weights wji from connections in the input layer. Yj is dependent on

$$Yj = f (\Sigma wji\ xi)$$

Yj is equal to the function of the sum of wji multiplied by xi.

When f is a basic threshold function like sigmoid or hyperbolic tangent function.

4. DATA SOURCE

By analyzing medical data like age, gender, blood pressure, and blood sugar levels, it is possible to forecast the chances of individuals developing heart disease. It allows for substantial learning, such as identifying patterns and relationships among medical factors associated with heart disease. The heart disease database that is accessible to the public can be utilized to diagnose different types of heart conditions.

Risk factors	Description	General Symptom
Age	**Old people are more suffers from heart disease**	**Chain pain** **Shortness of breath** **Irregular heartbeat** **Fatigue** **Fainting** **Swollen feet**
Sex	**Males are at greater risk than females**	
Family history	**If relatives have heart disease the probability of a person to have cardiovascular disease is high**	
Smoking	**Heart disease higher in smokers than nonsmokers people**	
Poor diet	**Diet food is essential for development of heart**	
Blood pressure	**Blood pressure can effect in narrowing hardening arteries, as well as thickening blood vessels [1], [2].**	
High blood cholesterol levels	**It increases formation of plaques**	
Diabetes	**It is the disease as a result of sugar in our body**	
Obesity	**Overweight body is one of the cause for heart diseases**	
Physical inactivity	**Physical activity helps heart to function properly**	
Stress	**Damage arteries**	
Poor hygiene	**It increases heart disease**	

Table 1. Factors of Heart Disease

Key Attributes

1. Patient – Patient's identification number

909 records were picked up from the Cleveland Heart Disease database in total. During the analysis, it was noted that Naive Bayes demonstrates higher accuracy with a correct prediction rate of 81.53% for heart disease patients, followed by Neural Network at 85.53% and Decision Trees. However, Decision Trees seem to be most efficient when predicting individuals without heart disease.

i.e. (89%) as equated to other two models.

Techniques	Accuracy
Naive Bayes	81.53 %
Decision Tree	**89%**
ANN	85.53

Table 2. Sample of Data Mining methods used in Heart Disease

In this study, the amount of characteristics utilized for diagnosing heart disease was decreased. Previously, this prediction utilized 13 attributes; however, using Genetic Algorithm and Feature Subset Selection, only six attributes are now used in this research.

4.2Genetic Algorithm [6] utilizes the principles of natural evolution. The genetic search begins with no attributes and a starting population containing randomly created rules. Following the concept of survival of the fittest, a new population is formed to align with the strongest rules in the existing population and their offspring. Genetic operators, specifically cross over and mutation, were used to produce offspring. The cycle of reproduction persisted until it produced a population P in which each rule met the fitness criteria. Starting with a population of 20 instances, the generation progressed to the twentieth generation with a crossover probability of 0.6 and a mutation probability of 0.033. The genetic search found six out of thirteen attributes. After reducing 13 attributes to just 6, the dataset with these new attributes is tested using different classifiers to predict heart disease.

Table 1 displays the performance evaluation of these classifiers. It is evident from the table that the Decision Tree model has surpassed the others, showing the highest accuracy and the lowest mean absolute error.

Techniques	Accuracy
Naive Bayes	96.53 %
Decision Tree	99.2%
Classification via clustering	88.3

Table 3. Showing the highest accuracy

5. ASSOCIATIVE CLASSIFICATION

CLASSIFICATION THAT UTILIZES THE ASSOCIATION BETWEEN DIFFERENT ATTRIBUTES.

Utilizing association rule discovery techniques, associative classification mining is a promising approach in data mining for constructing classification systems, also referred to as associative classifiers. Association rule mining is utilized for discovering connections or relationships between different sets of items. It is an unsupervised learning method that does not use class attributes to discover association rules. However, classification involves a supervised learning approach in which the class attribute is utilized in building the classifier to classify or predict unknown data samples. Associative classification combines association rule mining with classification to create a predictive model that offers high accuracy, making it a valuable technique. Associative classifiers are well-suited for scenarios where the highest level of accuracy is needed in order to predict outcomes effectively. Different strategies that can be utilized include apriori algorithm, éclat algorithm, FP-growth algorithm. In this study, we will utilize the Apriori Algorithm to uncover intriguing connections related to heart diseases.

Apriori Algorithm:

Apriori is a technique used to mine frequent item sets and learn association rules from transactional databases. The process involves finding common individual items in the database and then expanding them into larger item sets as long as these sets are frequent enough. A frequent item set is a subset of frequent item set where both {P} and {Q} need to be frequent for {PQ} to be frequent.

1. Iteratively discover frequent itemsets with cardinality from 1 to k (k-item set).
2. Use the frequent itemsets produce association rules. Join Step: Ck is generated by joining Lk1with itself Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k item set Initialize: K: = 1, C1 = all the 1- item sets; read the database to count the support of C1 to determine L1. L1 := {frequent 1- item sets}; k: =2;
//represents the pass number//
While (Lk-1 ≠) do begin
Ck: = gen_candidate_itemsets with the given Lk-1 Prune (Ck) for all candidates in Ck do
count the number of transactions of at least k length that are common in each item Ck Lk := All candidates in Ck with minimum support; k := k + 1;
end

Frequent Pattern mining using MAFIA

Searching for significant relationships between items in databases is a popular research topic in data mining known as mining frequent item sets [11]. It is applicable to many issues like uncovering association rules, sequential patterns, correlations, and more. The suggested method employs a productive algorithm named MAFIA (Maximal Frequent Item set Algorithm) that integrates various traditional and modern algorithmic concepts to create a functional algorithm. The suggested algorithm is utilized to extract association rules from the clustered dataset and works effectively with databases containing lengthy item sets.

Pseudo code for MAFIA :
MAFIA(C, MFI, Boolean IsHUT)
{
name HUT = C.head C.tail; if HUT is in MFI
stop generation of children and return
Count all children, use PEP to trim the tail, and recorder by increasing support, For each item i in C, trimmed_tail
{
IsHUT = whether i is the first item in the tail newNode = C I
MAFIA (newNode, MFI, IsHUT)
}
if (IsHUT and all extensions are frequent)
Stop search and go back up subtree
If (C is a leaf and C.head is not in MFI)
Add C.head to MFI
}

The cluster containing the most relevant data on heart attacks is input into the MAFIA algorithm to discover frequent patterns within it. Next, the importance of each pattern is determined using the method outlined in the subsequent section.

Significance Weightage Calculation

Once the MAFIA algorithm is used to mine frequent patterns, the weightage of each pattern's significance is computed. The calculation takes into account the importance of each attribute in the pattern and how often the pattern occurs.

$$s_{\omega i} = \sum_{i=1}^{n} w_{ifi}$$

Where *Wi* represents the weightage of each attribute and *fi* denotes the frequency of each rule. Subsequently the patterns having significant weightage greater than a predefined threshold are chosen to aid the prediction of heart attack

$$SFP = \{x: Sw\ (x) = \Phi\}$$

In this context, SFP indicates important frequent patterns and Φ indicates the significant weight. This SFP is applicable for creating a heart attack prediction system.

5. CONCLUSION

This paper things to see the utilization of various algorithms and mixtures of multiple target attributes for accurate heart attack prediction through data mining. Utilizing 14 attributes, Decision Tree achieved a remarkable accuracy of 99.62%. Moreover, the Decision Tree and Bayesian Classification's accuracy increases when employing genetic algorithm to minimize the data size and obtain the most suitable attributes for predicting heart disease.

The apriori algorithm, a technique used for association classification, was utilized in conjunction with a newer algorithm called MAFIA.

Apriori-based algorithms directly calculate the quantity of 2k subsets for every k-item set, making them impractical for lengthy item sets. "Look a heads" are used to decrease the amount of item sets that need to be counted. MAFIA works well with databases that have lengthy item sets.

6. Future Work

Future research should focus on integrating real-time data streams for immediate diagnosis, enhancing feature selection with advanced techniques like genetic algorithms, and developing predictive models for preventive healthcare. Deep learning integration for complex pattern extraction from medical images and ethical considerations in data privacy are crucial for clinical adoption and validation.

REFERENCES

[1] P .K. Anooj, ―Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.

[2] Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques".International Journal of Engineering Research & Technology

[3] Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[4] 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)

[5] Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[7] Erickson, T., Smith, D., Kellogg, W., Laff, M., Richards, J and Bradner, E. Socially translucent conversations: Social proxies, persistent conversation, and the design of "Babble."Proc. ACM CHI (1999), 72–79.

[8] Hollan, J., Hutchins, E. and Kirsh, D. Distributed cognition: Toward a new foundation for human computer interaction research. ACM TOCHI, 7(2),(2000), 174–

[9] Shantakumar B.Patil, Y.S.Kumaraswamy, ―Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.

[10] Statlog database: http://archive.ics.uci.edu/ml/machinelearning- databases/statlog/heart

[11] Shantakumar B.Patil,Dr.Y.S.Kumaraswamy "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009

[12] Azhar Rauf, Mahfooz, Shah Kusro and Huma Javed (2012)."Enhanced K-Mean Clustering algorithm to Reduced Number of iteration and Complexity", Middle East Journal of Scientific Research vol.12, issue 6, pp.959-963,2012

[13] Pratikhsha Shetogoankar , Dr. shailendra Awale,"Heart Disease Prediction Data Mining Techniques", vol 10, issue0, February 2021.