JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction System With Data Mining & Machine Learning

Jitendra Singh Bedi, MSc Computer Science 4th Semester, Rungta College of Science and Technology
Dr. Ashish Tamrakar Assistant Professor, Rungta College of Science and Technology
Khusboo Sao Assistant Professor, Rungta College of Science and Technology
MD. Arif Khan Assistant Professor, Rungta College of Science and Technology

Abstract

The medical industry is information-rich, but not all the data is mined that are required for discovering hidden patterns & effective decision-making. In this paper, we developed a low-level prototype proficient "Heart Disease Prediction System" (HDPS) using data mining and machine learning models, namely Decision Trees, Naïve Bayes, and Artificial Neural Networks. The software can predict the particular patient getting a heart disease with the help of some key attributes of patients such as age, gender, Blood Pressure, etc. It can serve as a training tool to train nurses and technical staff to diagnose patients with heart disease so that the patient gets early treatment for a particular disease. Data mining and artificial intelligence calculations help experts make better predictions and diagnoses of the disease. The goal of this research is to use statics of the patient and predict or detect the disease or explore the condition of a heart.

Keywords: Data mining, Decision support, Decision Tree, Heart Disease, Naïve Bayes, Machine Learning.

1. Introduction

Data mining refers to extracting or "mining" knowledge from large amounts of data. It is a supervised process of containing accurate, original, novel, potentially, essential, and knowledgeable arrangements in data with the wide use of databases. Data mining is the search for the relationships and global patterns that exist in large directories that are hidden among huge amounts of information (Data). The essential process of discovering new information is converting data into knowledge to aid in decision-making, known as data mining.

A major challenge facing healthcare organizations (Hospitals, medical centers) is providing quality services at affordable costs. The World Health Organization estimated that 12 million deaths occur worldwide every year due to different types of cardiac conditions such as Heart Valve Disease, Heart Arrhythmias, Heart Failure, Heart Valve Disease, Pericardial Disease, Cardiomyopathy (Heart muscle disease), congenital heart disease. Number of deaths is huge and dangerous. numbers of casualties only in the United States and other developed countries as compared to developing countries. Accurate patient diagnosis and efficient therapy delivery are prerequisites for providing quality care. Destitute medical decisions can serve to destructive consequences which are therefore unacceptable. This practice leads to unwanted error and excessive medical costs which affects the quality of service provided to patients. They can achieve these results by employing appropriate computer-based information and/or decision-making systems.

Hospitals must also minimize the cost of clinical tests. The system is automation of this system is efficient and advantageous. The software-based data prediction is highly configurable and efficient in the age of

technology it provides accurate results based on the patient's dataset that saves the user time and costs of having different medical tests for the same.

2. Literature Review

1. Sellappan Palaniappan, Rafiah Awang: – "Intelligent Heart Disease Prediction System", https://ieeexplore.ieee.org/abstract/document/4493524 /2008.

2. Mrs. Subbalakshmi, Mr. K. Ramesh: - "Decision Support in Heart Disease Prediction System using Naive Bayes" – Decision Assistance in the Heart Disease Prognosis System is developed using the Naïve Bayesian classification technique. The system extracts covered knowledge from a previous heart disease database of patients so that it helps the patient's early diagnosis of the heart condition to take action on it.

3. Patricia Rufes, J Sorna Jenita, Divya: - "Heart Disease Prediction Using Machine Learning", Making a step forward to lowered the patients of cardiac conditions from early identification of the disease from AI generative and ML mechanism. (IRJAEH) 2024

4. Ilias Touguis, Abdelilah Jillbab, Jamal El Mhamdi: - **"**Heart disease classification using data mining tools and machine learning techniques", Extraction of the Cardiac conditions dataset from the UCI machine learning repository. Pre-processing of these data to remove the missing instances with the help of different machine learning & data mining techniques.(IUPESM and Springer-Verlag GmbH Germany/2023)

5. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta: - "Heart Disease and Prediction Using Machine Learning and Data Mining Techniques: A Review" This paper aims to summarize some of the current research about heart prediction diseases using data mining methods, analyze the various combinations of mining algorithm used and conclude which techniques are sharp and durable.

3. Methodology

The methodology used in it is Naïve Bayes and Decision Tree for data mining. The decision tree approach is more powerful for classification problems and regression tasks. There are a few steps in this technique building a tree & applying the tree to the databases. There are many decision tree algorithms such as CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses the pruning method to build a tree. Pruning is a technique that compresses the size of the structured tree by removing overfitted data which leads to poor accuracy in predictions. J48 method accurately classified data as possible this method provides a higher level of precision on data the dataset. The overall concept is to be a tree that provides a balance of flexibility & accuracy. In the Medical industry, AI can show promising results. A 2012 research in a journal claimed that Machine Learning will play a crucial role in radiological applications by clarifying complex patterns and helping radiologists and technical staff make intelligent and promising decisions. Furthermore, in a study in 2015, the researcher demonstrated the relevance of ML methods in improving our perception of cancer progression in inclusion to their effectiveness and accuracy in decision-making.

4. Inputs that are necessary to diagnose the heart:

Input Attributes

1. Age
2. Gender
3. Chest Pain type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
5. ECG (electrographic results) (value 0:normal; value 1: ST-T wave abnormality; value 2: probable or definite left ventricular hypertrophy)
6. Slope (Peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. Exang - (exercise-induced angina (value 1: yes; value 0: no)
8. CA - No. of huge vessels colored by fluoroscopy (value 0 – 3)
9. Blood Pressure (mm Hg at the time of admission to the hospital)
10. Thalach (upper limit heart rate achieved)
11. Serum Cholesterol - (mg/dl)
12. Oldpeak – (ST depression induced by exercise relative to rest)

5. Data Collections

To make this research genuine and progressive we ensure that the information is correct and genuine for the calculation and probability of disease and condition of heart. 909 records are registered with 12 medical attributes contained from an addition of the dataset on cardiac disease, and Hungarian and Switzerland forecast of heart disease datasets. Figure 1 represents the attributes taken in the prediction system. The attribute Diagnosis with a value of "1" for patients with cardiac conditions and a value of "0" for patients with no cardiac conditions. There are 6 computer machine learning techniques tested on the databases using particular tools: Logistic regression, Support-Vector-Machine, K Nearest Neighbors, Neural Network, Naïve Bayes, and, Random Forest.

6. Methods Use in Study:

6.1 Naïve Bayes Algorithm

It's a guided education algorithm, which is founded on Bayes Theorem and used for solving classification problems. It is used in categorization & restoration tasks that include vast datasets. The Naïve Bayes algorithm is an effective and simple and efficiently driven algorithm which is effective in building fast machine-learning product models. The algorithm can easily be handled by a less experienced researcher or

student. The main algorithm's goal is to form quick predictions as it is to give data. It can applied in real-time predictions as well.

Why preferred Naive Bayes implementation:

1) Vast Dataset.

2) When the attributes are independent.

3) want more effective output, as a comparison to individual other methods output.

Bayes Theorem

Bayes theorem generally known as Bayes 'Rule or Bayes 'Law, resolves the potential of a hypothesis from its prior abstract knowledge. It depends on the conditional possibility.

The formula for Bayes' theorem is:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Post Possibility: The possibility of a hypothesis A on the observed event B.

P(B|A) is Likelihood probability: The amount of outcome of the clue given the potential for a hypothesis being true.

P(A) Preliminary possibility: The potential for the hypothesis before observing the clue(evidence).

P(B) is a marginal possibility: The potential for a clue.

6.2 Decision Tree

A Decision Tree kind a tree structure algorithm that helps us to decisions about a tree structure algorithm. In this model, we build a structure called a tree wherein a particular individual inner node denotes its features, its branches denote its regulations and the leaf denotes the results.

This is an adaptable machine-learning algorithm, that operates to retrogression tasks and classification. It is an efficient and very powerful algorithm. It shapes the Random Forest's most advanced algorithm in the artificial intelligence module. Decision Tree, CART (Regression & Classification) algorithm is employed.

The Categorization and Retrogression Tree-based Algorithm for Classification:

Q_m is the data at node m and has n_m samples and t_m is the threshold for node m. then,
The categorization and retrogression tree-based algorithm for grouping can be written as:

$$G(Q_m, t_m) = \frac{[n_m^{left}]}{n_m} \mathrm{H}(Q_m^{left}(t_m) + [\frac{n_m^{right}}{n_m}]H(Q_m^{left}(t_m)$$

It's fact obtain. To be able to make a decision tree.

$$t_m = t_m H(Q_m, t_m)$$

The Categorization and Retrogression Tree-based Algorithm for Regression:

Let the databases present at node m be Qm and it has nm samples. and tm as the threshold for node m. then, The categorization and retrogression tree-based algorithm for regression can be written as :

$$G(Q_{m,}t_m) = \frac{n_m^{Left}}{n_m} \mathrm{MSE}\ (Q_m^{Left}(t_m)) + \frac{n_m^{Right}}{n_m} MSE\ (Q_m^{Right}(t_m))$$

6.3 Random Forest

The Random forest is an ensemble learning method (also thought of as a form of nearest neighbor predictor) for categorization and retrogression techniques. It constructs several During training, decision trees produce a class that is the average of the classes that each tree produces.

6.4 Artificial Neural Networks

Artificial Neural Network (ANN) is a computational model based on the structure and functions of biological neural networks. The information that flows through the network
affects the building design of a neural network because a neural network changes or learns in a sense based on input and output. For that particular stage and consequently for each stage. ANNs are considered nonlinear statistical data-modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found ANNs have layers that are interconnected Artificial neural networks are fairly simple mathematical methodologies to improve operating data analysis technologies.

7. Future Scope

We all are living in an era of technology and advancements. Day by day we are getting more advanced by having such great things e.g. Artificial Intelligence, Robotics, Machine Generative Approaches, etc. A data mining mechanism is a procedure that turns an assortment of information into knowledge. In the future, an intelligent system may be developed that may guide the choice of proper treatment methods for a patient diagnosed with cardiac conditions. The data mining method can be a good method for categorization and data comparative tasks. Patient can easily check the condition of their heart on their smartphones or other devices. It may help them avoid heavy medical tests and checkups that are costly.

8. Benefit & Limitation

This Heart Disease anticipation System can function as a teaching tool for medical Staff and nurses to identify patients with cardiac disease. It can also provide decision support to assist doctors in making better clinical decisions to cure a particular patient. The current system is established on the 11 attributes listed in the table. Further, this list can be expanded to bring a more comprehensive diagnosis system.

9. Conclusion

A low-level heart disease anticipation system was developed with the aid of machine learning & data mining mechanisms. The system extracts the hidden information from the historical disease database of a single patient. The methods worn here we can extract patterns in response to the predictable state. The most efficient and durable model to predict patients with cardiovascular disease appears to be Naïve Bayes, followed by Artificial Neural Networks and Decision Trees. All three models might provide complicated answers, each having its strength. Naïve Bayes could answer four out of the five goals; Decision Trees, three, and Neural Network, two. Although not the most effective model, Decision Tree results are easier to understand and interpret. The drill-through function to grab detailed patients' profiles may be easily found in the Trees of Decisions. naive Bayes Fared was more accurate than Decision Trees as it could pinpoint every important indicator for medicine. The association between attributes produced by Neural Networks is more challenging to comprehend. It has a future to explore and expand. Can incorporate the other medicated characteristics besides the 15 listed in Fig.1 It can furthermore concatenate alternative methods of data mining techniques, e.g., Time Series, Clustering, and Association Rules. Ongoing data can furthermore as a substitute for categorical data. Another area is to mine the enormous amount of unstructured data present in healthcare databases using text mining. Integrating data would be an additional challenge in extraction and text-mining.

10. References

1. **Dj P Barker (1995) - "**Fetal origins of coronary heart disease".

(**2. Senthikumar Mohan, Chandrasegar Thirumalai (2019) – "**Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", (IEEE Transactions on cardiac conditions Prediction).

3. "**Sellappan Palaniappan, Rafiah Awang" – "**Intelligent heart disease prediction system".

4. Vijetha Sharmma – "Heart Disease Prediction using Machine Learning Techniques"(2020)

5. J. Thomas, R.Theresa Princy – "Human heart disease prediction system using data mining techniques"2016

6. Seth S. Martin, Aaron W. Aday – "A Report of US and Global Data From the American Heart Association".(2024)

7. Archana Singh, Rakesh Kumar – "Heart Disease Prediction Using Machine Learning Algorithms" (2020 Conference on International Electronics and Electrical Engineering (CIE3))

8. Norma Latif Fitriyani, Muhammad Syafruddin – "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System"(IEEE Access (Volume: 8)

9. Chintan M.Bhatt, Parth Patel, Tarang Getia – "Effective Heart Disease Prediction Using Machine Learning Techniques"

10. Kawsar Ahmed – "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison" (Computers in Biology and Healthcare Sector)