

Effective Heart Disease Prediction Systems Using Data Mining Techniques

1. Pavithra M.

Assistant Professor, Department of Computer Science Engineering, K. Ramakrishnan College of Technology, pavithram.cse@krct.ac.in

2. Sindhana. A. M

Department of Computer Science Engineering, K. Ramakrishnan College of Technology, sindhana20@gmail.com,

3. Subajanaki T.

Department of Computer Science Engineering, K. Ramakrishnan College of Technology, subajanaki27@gmail.com,

4. Mahalakshmi S.

Department of Computer Science Engineering, K. Ramakrishnan College of Technology, rsmahakrct@gmail.com

Abstract— The risk of heart disease is much high all over the world and is a major cause of death. Predicting the disease is difficult as it needs much knowledge and experience. It is a very difficult process even if the disease is predicted as it takes quiet a long time to cure. Predicting the disease early helps cure the patient with ease. Data mining and artificial intelligence techniques is a fast-growing technique used for collecting significant data and predicting the output. Necessary data from the patient like age, Chest Pain Type, Blood Sugar level etc., is collected and classified using data mining technique and the disease is easily identified. Thus necessary treatment can be taken at an early stage reducing the rate of death.

I. Introduction

Data mining is a process where important data's are extracted and turned into useful information. The important data are extracted and used to classify the information to predict the output. Heart Diseases are being predicted as it has very fewer symptoms, leading to death before proper treatment. C4.5 algorithm used in data mining is a decision tree classifier that classifies and predicts the output in the given data. We develop a website that provides information on whether the patient has heart disease or not. The user first registers his/her name along with the mail id and password. Then the user will enter the data's to predict heart disease. As soon as the user enters the details the data is collected and the disease is predicted. The user can also use the same credentials to log in to the account. Here HTML, CSS, and Javascript are used as frontend and Python is used as a backend language. MySql is used as a database language. If the patient has heart disease then the patient is suggested proper diet and exercise and advised to go to the hospital.

II. Literature Review

R. Indrakumari proposed a system for Heart Disease Prediction using Exploratory Data Analysis using the K-means Clustering algorithm with an accuracy of 94% [2]. Mohammed JasimA. Alkhafaji worked on Clean medical data and predict heart disease using the Decision Tree algorithm with an accuracy of 98% [3]. SinkonNayak, recommends a Naive Bayes

algorithm for Prediction of Heart Disease by Mining Frequent Items and Classification Techniques with 88.67% accuracy [4]. Anjan Nikhil Repaka proposed work for Design and Implementing Heart Disease Prediction Using Naive Bayesian with an accuracy of 98% [5]. Mamatha Alex P worked on Prediction and Diagnosis of Heart Disease Patients using the Artificial neural network with 92.21% accuracy [6]. MohiniChakarverti used a Classification Technique for Heart Disease Prediction in Data Mining called Support Vector Machine with 83% accuracy [7]. Dhara B. Mehta used a Newfangled Approach for Early Detection and Prevention of Ischemic Heart Disease using Data Mining technique Decision Tree algorithm with an accuracy of 93.4% [8]. RadhanathPatra worked on Predictive Analysis of Rapid Spread of Heart Disease with Data Mining technique Decision tree algorithm with an accuracy of 93% [9]. MafizurRahman proposed an approach for Effective Prediction On Heart Disease using a Random forest algorithm with 97% of accuracy [10].

III. Heart Diseases

Heart disease is one of the major cause of death all around the world. India is one of the high rise of load of heart disease all over the world. The annual number of deaths from heart disease in India is forecasted to increase from 2.26 million in 1990 to 4.77 million in 2020. Even children at a very young age are affected due to heart disease. Major risk factors include diabetes, abdominal obesity, hypertension, etc... Coronary is a heart disease that is caused due to inappropriate beating of the heart, whether asymmetrical, too fast or too slow. It happens when voltaic impulses in the heart don't work properly. This disease does not have any noticeable symptoms and hence it difficult to predict. Here we have used data mining techniques to predict whether the patient has heart disease or not.

IV. Data Mining

Data mining is a process of searching through a large amount of information to identify patterns, important data and predict the output. Data mining is also called Knowledge Discovery in Database (KDD). It is implemented using various kinds of data such as relational database, data warehouses, Data Repository and so on. Various techniques are being used in Data Mining and here we use the classification technique. Data Mining is a broadly used concept used in various fields such as healthcare, fraud detection, Market Basket Analysis, Customer Relationship Management, Education, Financial Banking etc... related with other applications data mining is cost-efficient. Data Mining helps the decision-making process of company. Data mining is more efficient due to the usage of a large amount of existing data in a short period of time.

V. Materials and Methods

The analysis is done using available data for heart disease. The dataset consists of 14 attributes. The dataset is analysed using the data mining technique C4.5 algorithm. After the data is analysed the patient is suggested with proper diet.

VI. Dataset Collection

The data are collected from the website called www.kaggle.com. The data consists of 13 columns such as Age, gender, Resting blood sugar, chest pain, serum cholesterol, fasting blood sugar, resting electrographic results etc...

Attributes	Description	Type
age	Age in years	Numeric
sex	Patients gender(1=male; 0=female)	Nominal
cp	Chest pain type	Numeric
trestbps	Resting blood pressure(in mm Hg)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
Fbs	Fasting blood sugar > (120 mg/dl) (1=true; 0=false)	Nominal
Restecg	Resting electrocardiographic results	Numeric
Thalach	Maximum heart rate achieved	Numeric
Exang	Exercise induced angina (1=yes, 0=no)	Nominal
Oldpeak	ST depression induced by exercise relative to rest	Numeric
Slope	The slope of the peak ST segment.	Numeric
Ca	Number of major vessels (0-3) colored by fluroscopy	Numeric
Thal	1 = Nominal, 2 = fixed defect, 3 = reversible defect	Numeric
Target	1 or 0	Nominal

Table 1: Features and information on heart disease dataset. (Source: <https://www.kaggle.com/johnsmith88/heart-disease-dataset>)

VII. Registration Process

The client and server play a major role in the website. The server side is used by the admin to insert delete or modify the data in the database. On the other side, the user wants to register

their details to access the portal. The user first enters the name, mail id and password to register on the website. Once the account is verified, the users can access the portal.

VIII. Classification

In Classification, various steps are involved. First, the classification algorithms build the classifier. The classifier is built from the training set by the use of database attributes and associated class labels. Each attribute that constitutes the training set is mentioned to as a category or class. These attribute can also be mentioned to as sample, object or data points. The classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable. We use a c4.5 algorithm which is a decision tree classifier technique.

IX. C4.5 based Classification

In this technique, we use C4.5 algorithm which can be waged to produce a decision, based on a definite sample of data. First, the entropy of the data is calculated. Entropy is nothing but the measure of disorder.

- The Formula for entropy is given below,
 Next, the information gain is calculated. Information Gain is the decrease or increase in Entropy value when the node is split.

- The Formula is given below,
- Next, the split information is calculated.

$$SI(Y,X) = -\sum E(Y/X) * \log_2 E(Y/X)$$

- Then the information Gain Ratio is being calculated.

$$IGR = IG(Y,X) / SI(Y,X)$$

Pruning is the process of cancelling noise in the data. Here a part of the obtainable labeled data for validation is being withheld. The validation set is not used during training. Once training finished next testing will take place. In a validation process. If the error rate of the original decision tree in the validation set exceeds the error rate of a pruned version of the tree then the pruning operation is carried out, reducing the over fitting, which reduces the amount of data available for training.

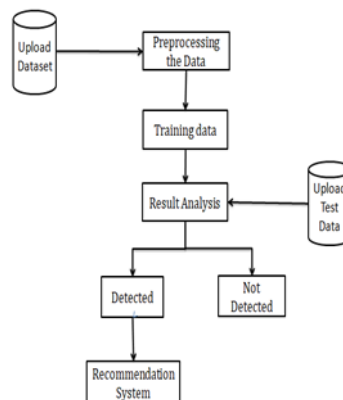


Fig1 : Workflow of C4.5 Algorithm

X. Results and Discussion

Once the details of the user are collected the output of the disease is executed and the result is shown to the user along with some suggestion details for proper diet and exercise. Thus the heart disease is predicted using the C4.5 algorithm. The proposed heart disease prediction system identification system achieves 89% accuracy.

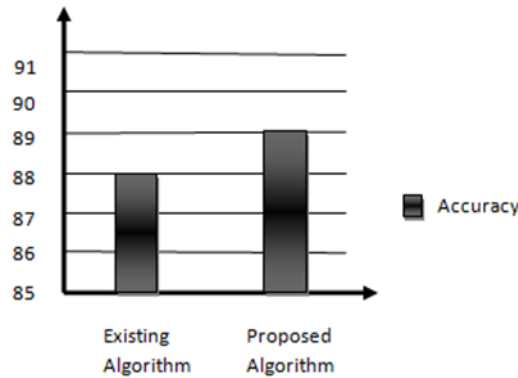


Fig2 : Accuracy Diagram

XI. Conclusion

Identifying Heart disease using raw health data will help to save human lives for the long term. These techniques were used in this work in order to process data and provide a new and effective result in heart disease. Thus, the mortality rate can be drastically reduced if the disease is detected at an early stage and treatment can be provided as soon as possible. The result shows that data mining techniques can be widely adopted as they give better accuracy. Our future work may be to explore the location of the hospital using MapSDK.

XII. Reference

- [1] Senthilkumar Mohan, ChandrasegarThirumalai, AndGautamSrivastava (2019) “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”
- [2] R. Indrakumari, T. Poongudi, And SoumyaRanjanJena (2020) “Heart Disease Prediction Using Exploratory Data Analysis”
- [3] Mohammed JasimA.Alkhafaji, Abbas FadhilAljuboori, Abdullahi Abdu Ibrahim (2020) “Clean Medical Data And Predict Heart Disease” Information Technology Altinbas University Istanbul, Turkey. College Of Engineering. University Of Information Technology And Communications Baghdad, Iraq. Computer Engineering Altinbas University Istanbul, Turkey.
- [4] SinkonNayak, ManjushaPandey, Mahendra Kumar Gourisaria, SiddharthSwarupRautarav (2019) “Prediction Of Heart Disease By Mining Frequent Items And Classification Techniques” Proceedings Of The International Conference On Intelligent Computing And Control Systems (Iciccs 2019)
- [5] Anjan Nikhil, Repaka, Ramya G, Franklin, Sai Deepak Ravikanti (2019) “Design And Implementing Heart Disease Prediction Using Naïve Bayesian”
- [6] Mamatha Alex P, Shaicy P Shaji (2019) “Prediction And Diagnosis Of Heart Disease Patients Using Data Mining Technique” International Conference On Communication And Signal Processing, April 4-6, 2019, India

- [7] MohiniChakarverti, Rajiv Rajan, SaumyaYadav (2020) "Classification Technique For Heart Disease Prediction In Data Mining" 2019 2nd International Conference On Intelligent Computing , Instrumentation And Control Technologies (Icicict)
- [8] Dhara B Mehta, Nirali C, Varnagar (2019) "Newfangled Approach For Early Detection And Prevention Of Ischemic Heart Disease Using Data Mining" Proceedings Of The Third International Conference On Trends In Electronics And Informatics (Icoei 2019)IeeeXplore Part Number: Cfp19j32-Art; Isbn: 978-1-5386-9439-8
- [9] RadhanathPatra, BonomaliKhuntia (2019) "Predictive Analysis Of Rapid Spread Of Heart Disease With Data Mining"
- [10] MafizurRahman, Maryam MehzabinZahin, Linta Islam (2020) "Effective Prediction On Heart Disease: Anticipating Heart Disease Using Data Mining Techniques" Second International Conference On Smart Systems And Inventive Technology (Icssit 2019)
- [11] Imran Mirza, ArnavMahapatra, Daryl Rego, Kenneth Mascarenhas (2019) Department Of Computer Engineering, Don Bosch Institute Of Technology, Mumbai-70,India. "Human Heart Disease Prediction Using Data Mining Techniques"
- [12] Saravanan M, Aramudhan M, PandiyanSs, AvudaiappanT.Priority Based Prediction Mechanism For Ranking Providers In Federated Cloud Architecture. Cluster Computing. 2019 Jul;22(4):9815-23.
- [13] Avudaiappan, T., Balasubramanian, R., Pandiyan, S. S., Saravanan, M., Lakshmanaprabu, S. K., & Shankar, K. (2018).Medical Image Security Using Dual Encryption With Oppositional Based Optimization Algorithm. Journal Of Medical Systems, 42(11), 208.
- [14] Sivakumar M, Reddy Us. Aspect Based Sentiment Analysis Of Students Opinion Using Machine Learning Techniques. In2017 International Conference On Inventive Computing And Informatics (Icici) 2017 Nov 23 (Pp. 726-731). Ieee.
- [15] MallikarjunanKn, ShalinieSm, Sundarakantham K, AarthiM.Evaluation Of Security Metrics For System Security Analysis. Incomputational Intelligence: Theories, Applications And Future Directions-Volume I 2019 (Pp. 187-197). Springer, Singapore.