Grenze International Journal of Engineering and Technology, Jan Issue

Parameter findings in Healthcare Analysis using K-means and Decision Tree Algorithm

Divya S Hakkaladavar[1] and Dr. Rajashri Khanai[2]
[1]Student, Department of Computer Science Engineering Dr. M. S. Sheshgiri College of Engineering and Technology Belgaum, India,
[2]Head of Department of Computer Science Engineering, KLE Dr. M. S. Sheshgiri College of Engineering and Technology Belgaum, India
E mail: 389divyah@gmail.com, rajashrikhanai@klescet.ac.in

***Abstract*— Our heart is one of the most fundamental or significant parts of our body. The heart is responsible for maintaining blood flow and forming it. There are thousands of cases on the planet connected to heart disease. A growing number of people are committing suicide because of coronary illness. Different side effects like chest pain, fast heartbeat, etc are referenced. The medical industry tracks down information. This paper discusses anticipating coronary illness using computer-aided calculations. Here, we will utilize different AI calculations like a help vector classifier, K-means, and D-tree algorithms. Calculations are used based on highlights and to anticipate coronary illness.**

By applying data mining techniques, it takes less time to predict disease. Based on the highlights, computations forecast cardiac illness. Using data mining techniques like D-tree and K-means algorithms speeds up and enhances disease prediction precision.

***Index Terms*— Data mining, medical data, Heart disease, Decision Tree, K-means.**

I. INTRODUCTION

Stress at work, lifestyle changes, and poor eating habits all contribute to heart disease. Heart disease has become one of the leading causes of death Heart-related illnesses increase healthcare costs and lower productivity. Therefore, it is crucial to predict heart-related disorders.

Around the world, medical organizations gather information on health-related topics. To extract insightful knowledge from these data, multiple machine learning approaches can be used. But because so much data is captured, it is often noisy. These datasets can be easily investigated using a variety of machine learning techniques, even though they are too large for human minds to grasp. As a result, these algorithms have recently proven to be highly helpful in precisely predicting heart-related disorders.

This project uses biological parameters like cholesterol, blood pressure, sex, age, etc. In order to test data. Machine learning learns from natural phenomena. Our algorithm comparison is based on the accuracy of these metrics.

A. Problem Statement

Effective management of heart disease can be achieved through the use of medication, lifestyle modifications, and, occasionally, surgery. With proper care, heart disease symptoms can be lessened and heart function enhanced.

Grenze ID: 01.GIJET.10.1.522

Predicted outcomes can be used to avoid, and hence lower, expensive surgical procedures. The overarching goal of the project will be to accurately predict heart disease using a few tests and features. In most cases, tests and analyses give accurate results based on the attributes taken into account. There are many more input attributes that can be employed. However, our goal is to estimate the risk of heart disease as quickly and efficiently as possible with the fewest possible attributes. The data set and databases that doctors use are generally not as knowledge-rich as the instincts and experience doctors rely on to make their decisions.

B. Motivation

Healthcare institutions face a key problem: delivering high-quality services at reasonable prices, which entails accurately diagnosing patients and giving them appropriate treatments. Hospital information systems produce data as numbers, text, charts, and images.

Regrettably, these findings only partially support clinical decision-making. These files hold a wealth of unknown, secret information. How can we transform data into information that helps healthcare professionals make wise clinical decisions? Is a crucial topic raised. This is the primary driving force behind the study.

II. Literature Survey

Ankita R. Mokashi , et.al [1] Although there is a lot of data in the health care system, understanding is lacking. It's because there isn't a good analysis mechanism in place to find hidden relationships and data. Because of them, knowledge is extracted and analysed via data mining.

Purushottam, et.al [2] this research paper have demonstrated an effective system for predicting heart disease using data mining. This system can aid medical professionals in making effective decisions depending on the provided parameter.

Mirpouya Mirmozaffari , et.al [3] In this work, data mining is used as a method to uncover hidden patterns in a clinical dataset by applying it to a database. There are 209 occurrences and 8 attributes in the database. To attain the best accuracy, all clustering techniques that are currently available are compared. The dataset is pre-processed using a variety of supervised and unsupervised algorithms in order to further improve the accuracy of the answer. Weka was used to create the system, and we compared the prediction accuracy for 40 approaches using 40 steps.

Mr.Santhana Krishnan.J, et.al [4] In this study, two supervised data mining algorithms—Naive Bayes Classifier and Decision Tree Classifier—were applied to the dataset to estimate the likelihood that a patient will develop heart disease. To compare the accuracy of these two methods, the same dataset is used for both of them.

Aditi Gavhane, et.al [5] According to this study, alternating decision trees performed remarkably well when combined with PCA, however under other circumstances, decision trees performed dreadfully, probably as a result of overfitting. Due to the fact that they employ a number of methods to address the issue of overfitting, Random Forest and Ensemble models have demonstrated good performance (multiple Decision Trees in case of Random Forest).

V.V. Ramalingam, et.al [6] The effectiveness of various models developed utilising these techniques and approaches is examined in this study. In order to automate the study of huge and complicated data, machine learning techniques and approaches have been used on a variety of data sets. Several researchers have recently aided the medical community and specialists in the detection of heart-related disorders by using a number of methods based on machine learning.

Fahd Saleh Alotaibi [7]] To improve the precision of HF prediction, this work makes use of information from the UCI heart disease dataset. In order to understand the data and predict the chance of heart failure in a medical database, a variety of machine learning approaches were used. The results and comparison analysis made it clear that the current work improved the accuracy rating for predicting heart disease as well. The medical information systems could be combined with the machine learning model developed in this paper to forecast HF or any other disease using real-time patient data.

Galla Siva Sai Bindhika, et.al [8] This study found that consumers are very concerned about the likelihood of developing heart disease, hence one of the criteria used to assess an algorithm's performance is algorithm accuracy. How accurate machine learning algorithms are depends on the dataset used for both training and testing.

G. Jignesh Chowdary, et.al [9] This study uses a Python-based Jupiter Notebook to build the model, which was then trained on the Kaggle standard dataset. The findings demonstrate that using ensemble modelling produces better forecasts than using individual models, which can lead to better predictions and potentially save live

III. Dataset Description

Dataset by Heart Disease UCI [14] is used in this work. This dataset is a multivariate type of dataset, which means that multivariate numerical data analysis was conducted and/or involved a number of distinct mathematical or statistical variables. A total of 206 cases are taken into consideration for training, while 160 cases are used for testing. It consists of the following characteristics.

TABLE I: ATTRIBUTE DESCRIPTION

1	Age	Age In Years
2	Gender	1 = Male 0 = Female
3	(CP) Chest Pain	1. Typical Angina 2. Atypical Angina 3.Non-Anginal Pain 4.Asymptomatic
4	ABP	Absolute Blood Pressure
5	Cholesterol	A waxy molecule called cholesterol is present in your blood.
6	FBS(Fasting Blood Sugar Is > 120 Mg/Dl)	1 = Yes 0 =No
7	ECG: Electrocardiographic Results	0: Normal 1: Having ST-T Wave Abnormality 2: Left Ventricular Hypertrophy By Estes' Criteria
8	HR	Maximum Heart Rate
9	Angina: Exercise Induced Angina	1 = Yes 0 = No
10	ST1	ST Depression Induced
11	SST2 : Slope Of The Peak Exercise ST Segment	1: Upsloping 2: Flat 3: Down sloping
12	Class	1: Healthy 2: Non- Healthy

IV. System Overview

This approach is based on mainly following steps: Input Symptom Data, Data Quantization, Features Analysis, Classifiers and Result.

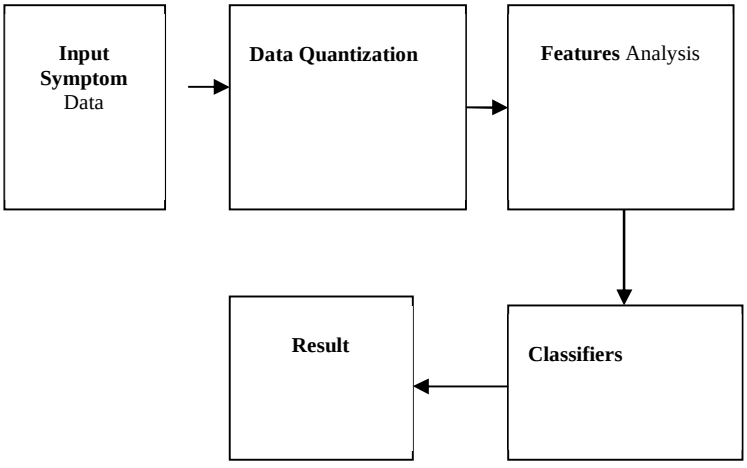


Fig -1: System Overview

The first step is the collection and formation of a dataset of different cases, which could be used as input. The initial quantization result was better. Features Analysis is done to allot values in different ranges. Using a variety of machine learning algorithms i.e., Decision tree algorithm, K-Means Clustering have been used for analyzing the dataset. System architecture is a theoretical model that defines the structure, performance, and view of the system. Here in the above diagram are two sections: the training and testing phases. These sections include the

following steps: A total of 197 healthy and 169 unhealthy cases were found, out of which 80 healthy and 80 unhealthy cases were taken for testing and those remaining were kept for training purposes.

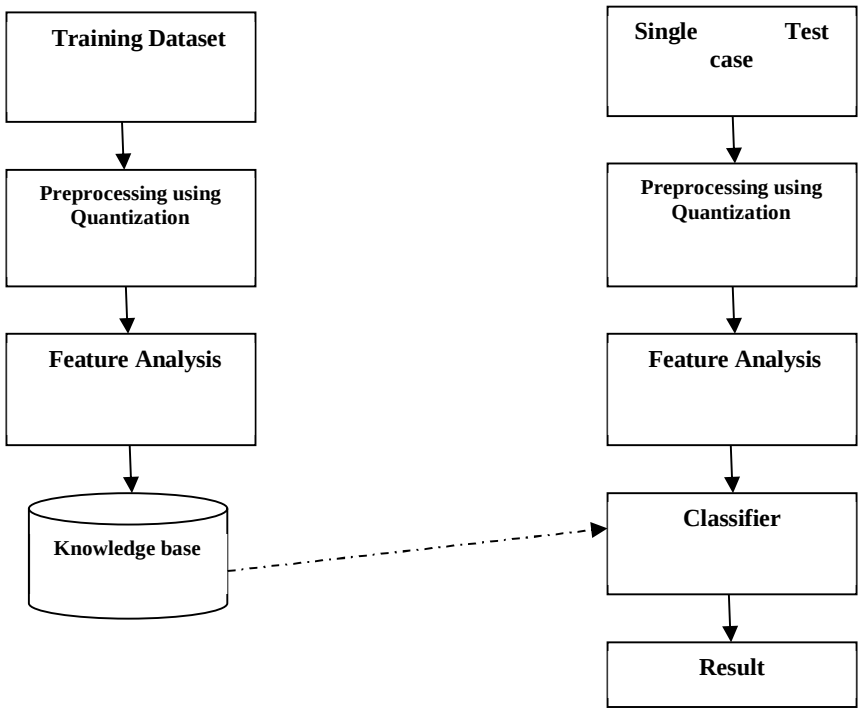


Fig -2: System Details

- Pre-processing using Quantization: Each clinical data parameter quantizes i.e. for example Male is 1, Female is 0 etc.
- Feature Analysis: By Using some laboratory clinical data sample, found that the values fall mainly under 3 ranges normal zone, moderate zone, danger zone, according to this study data are divided into ranges.
- Classification: Here the result will be displayed.

V. MEDICAL TERMS AND MEANING

A. Chest Pain

1. Typical Angina: Angina, a specific kind of chest pain, is caused on by a decreased blood supply to the heart.
2. Atypical Angina: Atypical chest pain is defined as chest discomfort that doesn't fit the angina criteria.
3. Non-Anginal Pain: Non-cardiac chest pain (NCCP) is a phrase used to describe chest discomfort in people who do not have heart disease that resembles heart pain (also known as angina).
4. Asymptomatic: Asymptomatic means neither causing nor exhibiting symptoms of disease

B. ABP

Absolute Blood Pressure: low (<130 mm Hg), medium (130–160 mmHg) and high (>160 mmHg)

C. Cholesterol

A waxy, fatty substance that is found in all the cells in human body. Good (< 200), medium (200-240) and high (>240)

D. Maximum Heart Rate

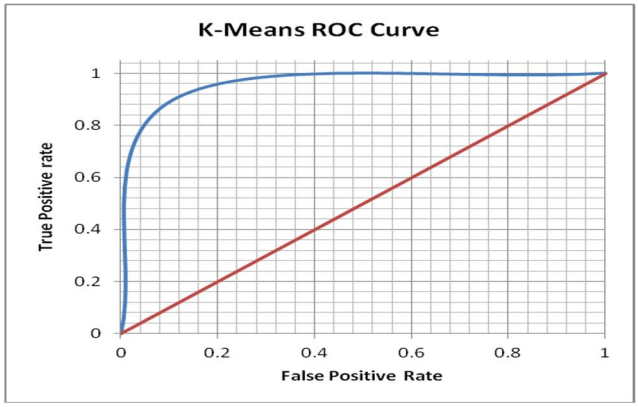
The age can be subtracted from 220 to determine the maximal heart rate. For instance, if you were 45 years old and your age was deducted from 220, your maximum heart rate would be 175. During working exercise, your heart should not beat any more frequently than this on average per minute.

VI. Results and Analysis

A. K-means Analysis

```
Confusion Matrix - Kmeans
                Predicted
              ------------
              | No   | Yes |
       ----------------------
       | No   |  71 |   9 |
Actual ----------------------
       | Yes  |   2 |  78 |
       ----------------------

Accuracy = 93.13 %

True Positive Rate = 0.98

False Positive Rate = 0.11

Precision = 0.90

Time for Testing per Case 31.37  millsec
```

Chart -1: K means ROC curve



B. D-tree Analysis

```
Confusion Matrix - Dtree
                Predicted
              ------------
              | No   | Yes |
       ----------------------
       | No   |  59 |  21 |
Actual ----------------------
       | Yes  |  23 |  57 |
       ----------------------

Accuracy = 72.50 %

True Positive Rate = 0.71

False Positive Rate = 0.26

Precision = 0.73

Time for Testing per Case 9.12  millsec
```

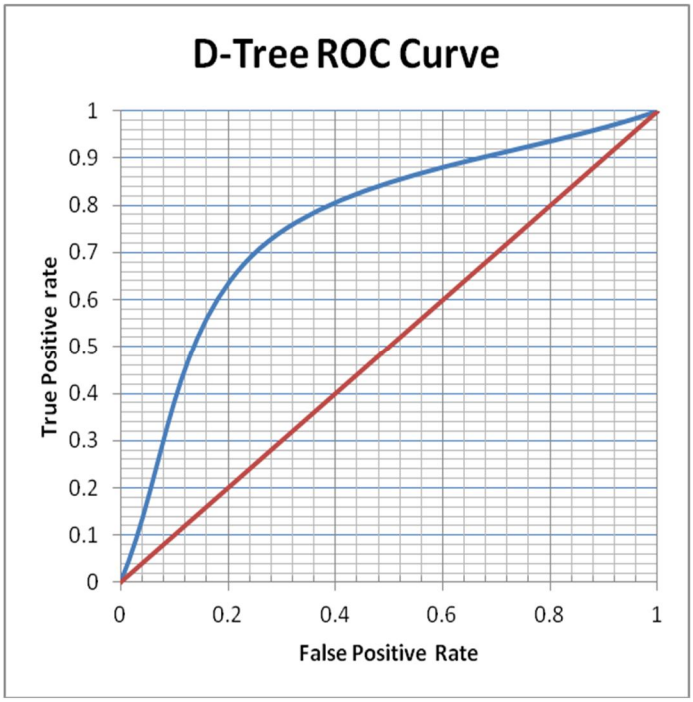


Chart -2: D tree ROC curve

C. Performance Comparison

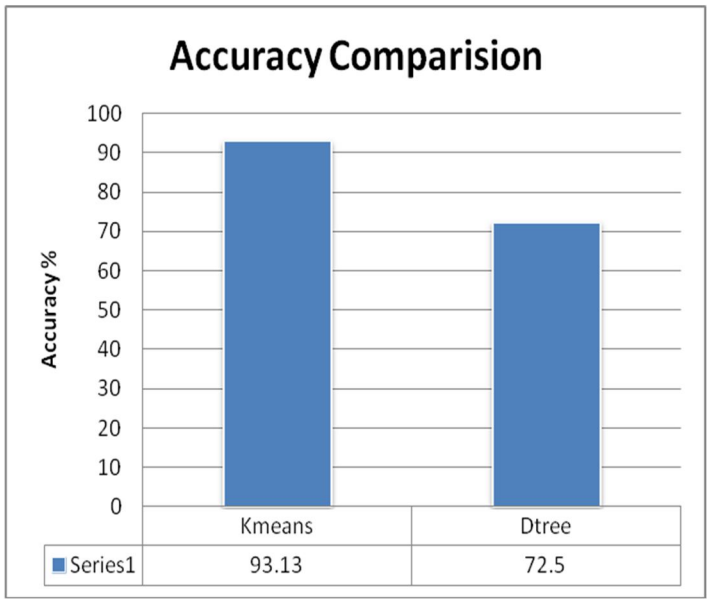


Chart – 3 Comparative graph K-means Vs D-tree

VII. CONCLUSIONS

Here, to determine the most reliable clustering method for predicting heart disease, data mining algorithms were tested. Several filters combine to form a distinctive model. To determine the most efficient algorithm, numerical evaluation and multilayer filtering are used. Regarding accuracy, error functions, and building times, clusters are contrasted. Based on the obtained results, K-Means had a maximum accuracy of 93.13% and took 22.22 milliseconds to complete, whereas D-tree had a maximum accuracy of 72.50 and required 9.12 milliseconds.

REFERENCES

[1] Ankita R. Mokashi, Madhuri N. Tambe, Pooja T. Walke , "Review on Heart Disease Prediction using Ann and Classifier " , International Journal of Engineering Research & Technology (IJERT)
September-2015

[2] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System", 1877-0509, 2016 The Authors. Published by Elsevier B.V.

[3] Mirpouya Mirmozaffari, Alireza Alinezhad, and Azadeh Gilanpour, "Heart Disease Prediction with Data Mining Clustering Algorithms", Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 4, Issue 1 (2017).

[4] Mr.Santhana Krishnan.J , Dr.Geetha.S , "Prediction of Heart Disease Using Machine Learning Algorithms"

[5] Aditi Gavhane,Isha Pandya,Gouthami Kokkula,Prof. Kailas Devadkar (PhD), "Prediction of Heart Disease Using Machine Learning" Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

[6] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja , "Heart disease prediction using machine learning techniques : a survey" , International Journal of Engineering & Technology, 7 (2.8) (2018).

[7] Divya Yadav*, Prof. Gayatri Pandi (Jain), "A Survey Techniques Used for Prediction of Heart Attack with Machine Learning and Medical Text Mining", 2019 IJSRSET.

[8] Fahd Saleh Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[9] Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy, Rajalakshmi, "Heart Disease Prediction Using Machine Learning Techniques", International Research Journal of Engineering and Technology (IRJET) Issue: 04 | Apr 2020.

[10] G. Jignesh Chowdary, Suganya. G, Premalatha. M "Effective Prediction of Cardiovascular Disease Using Cluster of Machine Learning Algorithms", Journal Of Critical Reviews, Issue 18, 2020.

[11] Archana Singh, Rakesh Kumar "Heart Disease Prediction Using Machine Learning Algorithms", IEEE2020, International Conference on Electrical and Electronics Engineering (ICE3-2020).

[12] Shi Na, Liu Xumin, Guan yong , "Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm" , 2010 , IEEE

[13] He Zhang; Runjing Zhou, "The analysis and optimization of decision tree based on ID3 algorithm", 2017 9th International Conference on Modelling, Identification and Control (ICMIC)

[14] UCI, ― Heart Disease Data Set. [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci.