Machine Learning-based Prognosis for Early Detection of Heart Disease – A Comparative Study

Hemasri R[1], Rajeshwari N[2]

[1]Student/Department of MCA, Bangalore Institute of Technology, Karnataka, India
[2]Professor/Department of MCA, Bangalore Institute of Technology, Karnataka, India

---***---

Abstract - Medical analysis is occasionally cited as a valuable source of insightful data. Coronary heart disease (CHD), a common and serious complication of diabetes mellitus type 2 (T2DM), is one of the most common chronic conditions characterized by an imbalance in insulin secretion. T2DM commonly has poor outcomes and even fatalities as a result of these complications. Identification of those who have an a higher risk of CHD problems is becoming more and more important due to the enormous number of people with T2DM, but a predictive method is still lacking. Early detection of CHD can help reduce mortality rates because it is one of the main causes of death in the globe. The challenge arises from the complexity of the data and relationship prediction using conventional methods. The purpose of this project is to use machine learning (ML) technologies and historical medical data to predict CHD. This study's primary objective is to identify correlations in CHD data using supervised learning methods such as Support Vector Machines (SVM), Decision Trees and Ensemble Classifiers. The researched ML techniques produce intelligent models. Empirical results demonstrate that probabilistic approaches are promising in diagnosing CHD using a variety of performance assessment parameters.

***Key Words*:** heart, health, machine learning, ensemble

1. INTRODUCTION *(Size 11, Times New roman)*

In this era of technology and digitization, data has established itself as the fuel for enterprises and sectors. The healthcare industry is not lagging in this regard. Today, almost all medical centres and hospitals maintain patient data electronically. Their medical background, symptoms, diagnosis, course of illness, instances of recurs and any fatalities are covered in this. As a result of this, the volume of medical data generated each day is continuously growing. However, this abundance of data is typically unused due to a lack of effective analytical tools, processes, and people to find insights and hidden connections in it. Utilizing existing data to create screening and diagnosis algorithms will not only reduce the need for medical personnel but would also aid in the early detection and treatment of patient's, greatly enhancing the health care system. Additionally, it can help in the creation of a surveillance and preventive program for patients who, based on their health conditions and family histories, may be prone to CHD.

Coronary arteries are crucial for supplying the heart muscle with oxygen. The chronic build-up of fat or cholesterol that is harmful within these artery walls, reported to the Southern Cross Healthcare Society of Zealand, results in arterial wall constriction and finally blockage, which leads to CHD.

A little impediment might simply result in momentary discomfort and alterations to the person's way of life when the coronary arteries' ability to carry oxygen is impaired. However, if significantly impeded, it could be fatal. The risks for CHD that can be changed by lifestyle choices are listed below. Uncontrollable risks for CHD include things like ethnic background, medical history, and others. Early identification of CHD signs can assist the patient in controlling some of those risk factors through dietary changes and/or medications, preventing the illness from getting worse and being fatal. ML algorithms are frequently used in the Data Science age to gain insightful knowledge and use the data acquired to influence decisions. They have contributed significantly to the automation and simplification of countless processes, as well as the optimization of business in many other industries. Intelligent methods are used in the field of machine learning (ML) to extract descriptive and prescriptive models from data. Machine learning (ML) is an automated strategy used by machines to learn from data, uncover pertinent patterns, and minimize human participation in decision-making, according to Thabtah et al. (2010). Supervised learning along with unsupervised learning is the two different categories of machine learning algorithms.

2. LITERATURE SURVEY

A number of research projects have been carried out by experts, academia, and the data analytics community with the goal of identifying and evaluating medical data for different diseases. These predictions have been made in earlier investigations using a variety of ML approaches. We will assess significant research publications before we start our dataset analysis. Researchers developed a model that uses data mining to predict CHD using 100 CHD cases and survival rate data in response to the healthcare community's demand for an
At the first occurrence of an acronym, spell it out followed by the acronym in parentheses, e.g., charge-coupled diode (CCD).
improved CHD prediction approach. The authors used 502

examples of support vector machines (SVM), artificial neural networks (ANN), and Decision Trees (DT) to evaluate model performance. They also used a confusion matrix and the 10-fold cross-validation approach. His investigation produced the following findings:

With a dataset of 13 variables, including important elements like gender, hypertension, and cholesterol, Apte et al. predicted heart disease. The scientists added smoking and obesity as two new traits. After using the sorting techniques ANN, DT, and NB, the findings indicated that ANN achieved the best projected accuracy for the provided dataset.

In order to establish a connection between important trends in a dataset of 14 features, Jenzi employed the technique of association rule mining. The authors developed a number of classifier models using DT, NB, and ANN classification techniques. On the Microsoft.NET a platform, a user interface with graphics (GUI) was created, with connections made through IKVM interface and Java libraries. The model results were displayed on the receiver.

3. EXISTING SYSTEM

- Mohammed Abdul Khaleel presented a presentation on Medical Techniques for Finding Frequent Diseases locally at the Survey of Techniques for Mining Data.
- This study focuses on dissecting information mining process those are helpful for medicinal information mining, specifically to locate locally visit ailments such as heart disease, lung cancer, bosom disease, and so on. Information mining is a method of extracting information with the purpose of locating inactive examples, which Vembandasamy et al. used to analyze and diagnose cardiac disease. The Naive Bayes technique was utilized in this case. In the Nave Bayes method, they applied the Bayes theorem. As a result, Naive Bayes has a high degree of independence in making assumptions. The data set utilized was taken from a major diabetes research facility in Chennai, Tamil Nadu. The dataset has almost 500 patients. Weka is the tool utilized, and the categorization is done using 70% Percentage Split. Naive Bayes has an accuracy of 86.419%.

DISADVANTAGES

Incorrect categorization outcomes.

Classification accuracy is poor.

4. PROPOSED SYSTEM

To fix every flaw in the existing system, the recommended model is presented. In this instance, a system based on machine learning was used to forecast the user's risk of developing chronic coronary heart disease from a dataset. The training data should be used to estimate the SVM, Decision tree, KNN, Ensemble, and ANN algorithms. This is accomplished through the importance of maximum-likelihood estimation. A lot of machine learning algorithms frequently use maximum-likelihood estimation as a learning strategy, despite the fact that it does make assumptions regarding the distribution of our data.

ADVANTAGES

- High performance; accurate forecast results.

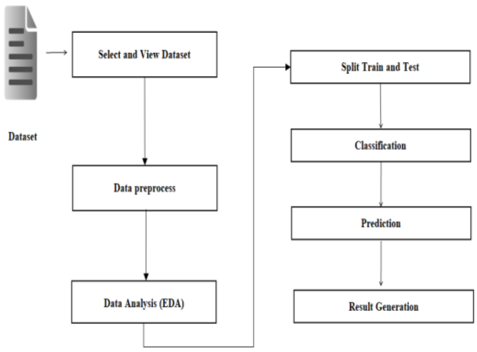


Fig: Proposed Architecture

5. IMPLEMENTATION

SELECTION AND LOADING OF DATA

The process of picking data for the purpose of detecting assaults is known as data selection. The coronary Heart disease dataset is utilized to detect disease in this study. In this proposed system we are taking Cleveland heart disease dataset which consist of 14 attributes.

PREPROCESSING OF DATA

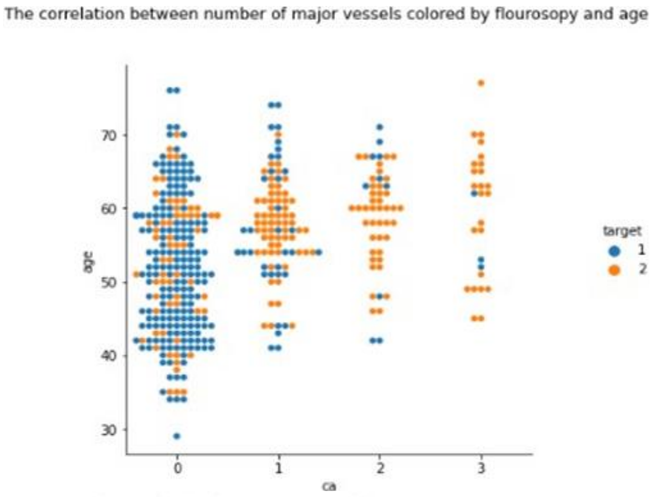
The process of deleting undesirable data from a dataset is known as data pre-processing. Removal of missing data Missing data removal: The imputer library is used in this procedure to remove null values such as missing values.

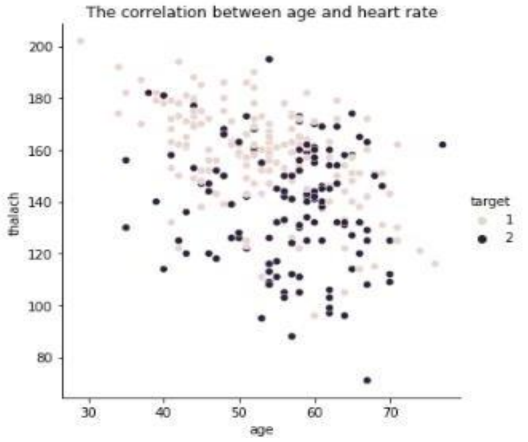
The first stage in this procedure is data pre-processing. The data is pre-processed to remove any undesirable information.

DATA EXPLORATION AND ANALYSIS:

In statistics, exploratory data analysis is the way of analyzing data sets to be able to summarize their primary features, which is frequently done using statistical graphics and other data visualization approaches. The main objective in EDA is to find out what the information can tell us past the traditional modeling or testing of hypotheses assignment. A model based on statistics could or might not be used. To motivate statisticians to study data and formulate ideas that could lead to additional data collection and experiments, exploration data analysis was promoted. EDA is different from the initial analysis of data (IDA), concentrating on correcting missing values and modifying variables as necessary, as well as evaluating hypotheses for the fitting of models and testing hypotheses. IDA is a part of EDA.

Charts





6. CONCLUSIONS

We infer that illness identification not only assists consumers by expediting medication, but it also assists medical facilities and authorities in better allocating money and devising measures to completely avoid or at least limit the occurrence of disorders. Early identification of deadly illnesses increases the likelihood of treatment in some circumstances. Many research techniques have been pursued in terms of illness prediction and screening utilising medical data. There are various machine learning (ML) algorithms present to aid in the detection of CHD. The goal of this study was to point out a few of the current prediction techniques and the performance metrics associated with them. From this the gradient boosting classifier has the best accuracy of 94% among the other algorithms.

REFERENCES

[1] Dulhare, U. N. (2018). "Prediction system for heart disease using naïve bayes and particle swarm optimisation." Biomedical Research, 29 (12), 2646-2649.

[2] BMC Public Health. doi: 10.1186/S12889-019-6721-5. Benjamin, H., David, F., & Belcy, S. A. (2018). "Heart disease prediction using data mining techniques." ICTACT Journal of Soft Computing, 9(1), 1824-1830

[3] Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). "Predicting coronary artery disease: a comparison between two data mining algorithms."

[4]Gawali, M., & Shirwalkar, N. (2018). "Heart disease prediction system using data mining techniques." International Journal of Pure and Applied mathematics, 120 (6), 499-506

[5] Haq, A. U., Li, J.-P., Memon, M. H., Nazir, S., & Sun, R. (2018). "A hybrid intelligentsystem framework for the prediction of heart disease using machine learning algorithms."
Hindawi Mobile Information System. Doi: 10.1155/2018/3860146

[6] Q. GU, Z. Cai, L. Zhu & B. Huang, data mining on imbalanced data sets, International Conference on Advanced Computer Theory and Engineering, 2008.