Cardiac Care: Heart Disease Prediction Using Machine Learning Techniques

Dr. S.S. Dhotre [1], Mayank Ramani[2], Priyam Maheshwari [3], Subhanshu Tripathi[4]
[1]*Associate Professor, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune*
[2,3,4]*Student, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune*

***Abstract* - Currently, cardiovascular disease has a significant impact on the death rate. Heart disease is the biggest cause of death all over the world. The procedure of predicting a heart attack can be quite difficult. Heart disease is one of the most difficult diseases, and it has affected many people around the world. There is an abundance of patient data that can be utilized to train a model to predict cardiac disease. Researchers proposed a model that predicts a patient's heart health based on their medical state using various data mining and machine learning techniques. The model is trained on the Cleveland database, which is available on the UCI data repository. It features 303 patient health data records, each of which has 14 different qualities that affect the human heart. The goal of this study is to use supervised learning algorithms like K-nearest and Decision Tree to design a model that can predict the likelihood of patients developing heart disease.**

***Index Terms* - cardiovascular disease prediction, machine learning, UCI Dataset. K-Nearest, Decision Tree.**

I.INTRODUCTION

Coronary heart disease occurs when the coronary arteries get clogged, causing symptoms like angina, chest discomfort, and heart attacks [1]. The blood supply to the heart muscle is provided via arteries. Plaques are formed when fatty deposits, including cholesterol, accumulate over time and restrict the arteries. The majority of heart attacks are caused by this disease. A complete blockage indicates that the patient has had a ST elevation myocardial infarction (STEMI), whereas a partial blockage indicates that the patient has had a partial blockage [2]. One of the leading causes of sickness and mortality among the world's population is heart disease. One of the most important topics in the domain of clinical data analysis is cardiovascular disease prediction. In the healthcare industry, there is a massive amount of data. The healthcare business accumulates massive volumes of data, which is regrettably not "mined" for hidden information that can help decision-makers make better decisions. Hidden patterns are frequently overlooked[3]. Advanced data mining techniques can assist in resolving this issue. Data mining converts a significant amount of raw healthcare data into information that may be used to make better decisions and forecasts [4]. It has a lot of promise for improving health-care systems. It identifies best practices for improving treatment and lowering costs using data and analytics. .[14][15]

The researchers employed data mining and machine learning approaches to predict heart health by analyzing key variables that are thought to be the most heart-affecting characteristics. The dataset from the UCI repository is first used to train the model with the best suited algorithm, and then the model is tested on more data to ensure the model's prediction quality.

Predisposing factors for heart disease include behavioral risk factors such as smoking and caffeine, stress, and physical inactivity, as well as physiological variables such as obesity, hypertension, high blood cholesterol, and pre-existing cardiac problems[5]. The ability to diagnose cardiac disease quickly, accurately, and accurately plays a critical role in adopting preventative actions to avoid death. The extraction of features impacting the heart and causing all the risk factors stated can be done via data mining. This model is based on the UCI dataset, which has several attributes that are important in heart disease. It contains age, sex, chest pain, rest blood pressure, Serum cholesterol fasting blood sugar, Rest Electrocardiograph, MaxHeart Rate, Exercise induced angina, ST depression, Slope, no. of vessels, Thalassemia.

The major goal of the project is to develop a model that can process the different human elements that affect the heart, forecast heart health, and provide prior knowledge about a patient's heart health, whether it is good or bad. This would keep the patient constantly informed about their health status, as well as reduce the expense of numerous testing and procedures.

II. MOTIVATION

Heart disease is on the rise, and it affects people of all ages. Recently, a 29-year-old young football player fainted on the field from cardiac arrest[6]. This indicates that heart disease has an impact on people of all ages, not just the elderly. Most people should become more knowledgeable about their heart health. There has also been a significant amount of money wasted for people whose hearts are completely OK, all because of the doctors' reckless and sloppy behavior. Doctors recommend a variety of tests to patients to assess their heart health. However, what if the patient already knew about their heart health? This would save a lot of money. Apart from that, hospitals generate a large amount of data, which is stored in logs unused. A significant amount of information can be extracted from those data, and patient health files contain high-quality data that can be used to train and predict not only heart disease but a variety of other diseases prior to their onset.

III. PROBLEM STATEMENT

Heart disease can be effectively managed with a combination of lifestyle changes, medication, and surgery in some circumstances[7]. The symptoms of heart disease can be lessened, and the heart's function enhanced with the correct treatment. The projected outcomes can be utilized to prevent and thereby minimize the cost of surgery and other costly treatments. Individuals with prior awareness of their health state may be able to avoid serious threats to their hearts and may also be able to avert future sickness.

In all hospitals, there is an abundance of patient personal healthcare data, and the quantity of knowledge that can be gleaned from this data is quite useful[8]. All the information can aid in the early detection of many diseases. There will also be no human bias or hidden costs, and the prediction algorithm will offer the results regardless of who the user is. This model will pre-process a patient's health data before predicting heart health by comparing various features and gaining insightful knowledge using the best machine learning technique.

IV. LITERATURE REVIEW

In various papers, to diagnose Heart Disease, researchers have proposed various diagnostic algorithms based on machine learning. This paper offers several current machine learning-based diagnostic tools to clarify the significance of the proposed work. Several researches have shown the development of machine learning models for heart disease diagnostics in order to deliver better outcomes for an HDPM.

S. Mohan, C. Thirumalai and G. Srivastava [9] used numerous combinations of information and several established classification approaches to propose a Heart Disease Prediction Model using hybrid Machine Learning techniques. They achieve a higher degree of performance with an accuracy of 88.7% by using the prediction model for heart disease with the (HRFLM).

N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee [10] developed a model that can predict heart illness at an early stage and can then be utilized to detect heart disease status using the CDSS. DBSCAN is used to discover and remove outliers, a hybrid Synthetic Minority Oversampling Technique Edited Nearest Neighbor is used to balance the training data distribution, and XGBoost is used to forecast heart disease.

J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor [11] developed a model using Machine Learning classification algorithms such as Support vector machine, Logistic regression, K-nearest neighbor, Naïve bays, and Decision tree while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. They also proposed an unique fast conditional mutual information feature selection approach to solve the feature selection problem (FCMIM). The experimental results suggest that the proposed feature selection technique (FCMIM-SVM) can be used to create a high-level intelligent system that uses a classifier support vector machine to categorize heart disease.

C. Xiao, Y. Li and Y. Jiang [12] proposed a unique fast conditional mutual information feature selection approach to solve the feature selection problem (FCMIM). The experimental results suggest that the proposed feature selection technique (FCMIM-SVM) can be used to create a high-level intelligent system that uses a classifier support vector machine to categorize heart disease. With or without the centerline dataset, the model conducts experimental comparison results. The results reveal that the model training effect of centerline preprocessing is superior to the original results. The accuracy of the model's conclusion has been confirmed to be 82 percent.

S. Palaniappan and R. Awang [13] proposed an expert medical diagnosis system for HD identification. In development of the system the predictive model of machine learning, such as navies bays (NB), Decision Tree (DT), and Artificial Neural Network were used. The 86.12% accuracy was achieved by NB, ANN accuracy 88.12% and DT classifier achieved 80.4% accuracy.

A. J. A. Majumder, Y. A. ElSaadany, R. Young and D. R. Ucci [14] presented a multiple-sensory system. In order to detect and predict sudden heart attacks with greater precision, signal processing and ML techniques were implemented for sensor data analytics.

V. APPROACH

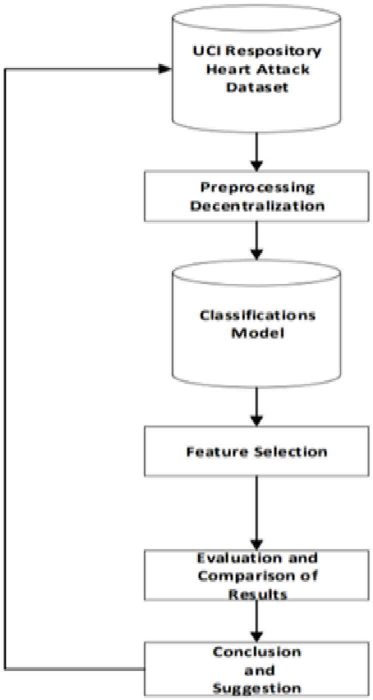
UCI Respository Heart Attack Dataset
Preprocessing Decentralization
Classifications Model
Feature Selection
Evaluation and Comparison of Results
Conclusion and Suggestion

Fig.1. Flowchart of Implementation

This research aims to estimate the possibility of acquiring heart disease as a likely cause of computerized heart disease prediction that will be valuable to doctors and patients in the medical field. To achieve the goal, this research study investigates the usage of numerous machine learning algorithms on the data set and dataset analysis. This study also reveals certain variables are more important than others in predicting higher precision.

5.1 Data source

For our research, we used a dataset from the UCI Machine Learning library. It contains a real dataset with 300 data samples and 14 various features (13 predictors; 1 class), such as blood pressure, type of chest discomfort, ECG result, and so on (Table 1). In this work, we used two methods to detect cardiac illness and build a model that was as accurate as possible.

5.2 Data Preprocessing

In real-life data, there are a lot of missing and noisy data. To eliminate these errors and produce credible predictions, the data is pre-processed. Because most machine learning algorithms demand integer values, attributes with category values were converted to numerical values. For variables with more than two categories, dummy variables were established.

The data is then classified and divided into training and test data sets, which are then exposed to various algorithms to determine accuracy scores.

We found out that among all the attributes, chest pain is the most important feature in our dataset.

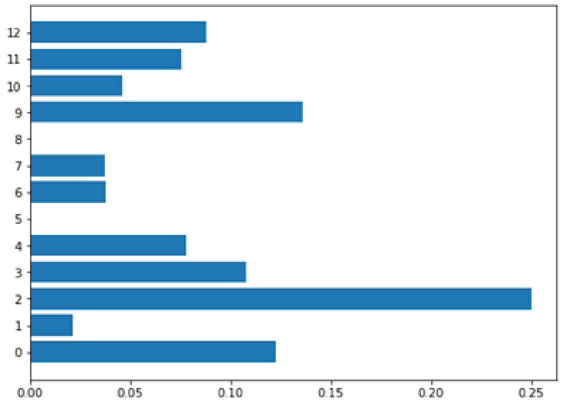

Fig.2. Plotting the most important feature

5.3 Graphical Representation

For a better understanding of the dataset, Graphical representation of all the attributes vs their respective counts, i.e., what are the number of people present in the dataset with that attribute value is done by using matplotlib library of the python.

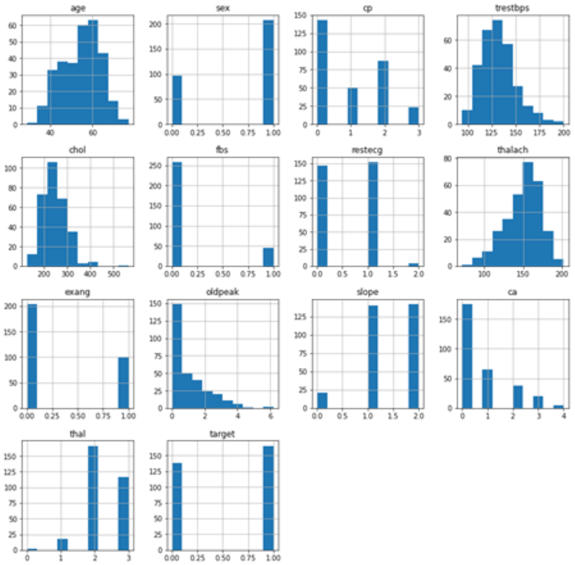

Fig.3. Plotting all features

A bar graph is shown between the goal and count attributes to highlight the association between persons who have heart disease and those who do not have it. It shows the number of people who have Target Value = 1 (Heart Disease) and Target Value = 0 (No Heart Disease) (Not Suffering From Heart Disease). This would provide a clearer picture of the dataset. Data visualization is handled by the seaborn library bar graph function, and data representation is handled by matplotlib.

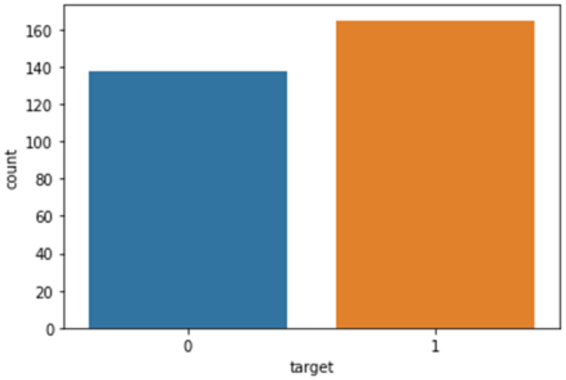

Fig.4. Comparing the target variable(class)

According to the above bar graph, there are 165 people suffering from heart disease and 138 persons in good heart health.

Plot a bar graph with the seaborn library of python to better comprehend and differentiate total persons based on their age and sex. This bar graph will display the results in two partitions, one for those with heart disease and one for people without heart disease. Then, in each partition, a male and female partition is created.

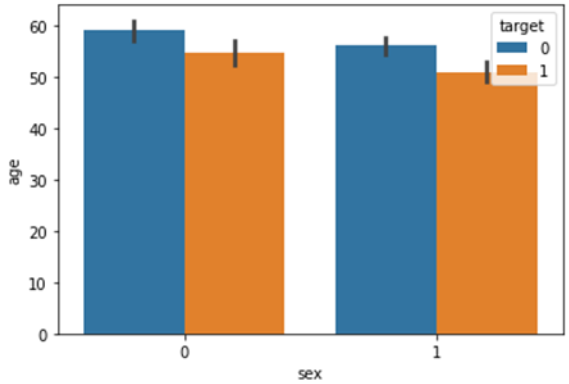

Fig.5. Comparison of target with sex

We can represent a proper distribution of persons according to their heart health and maximal heart rate using the distplot of the seaborn library. We used distplot to visualise the age distribution of patients with and without cardiac disease. A distribution graph for the maximum heartbeat rate of people with and without cardiac disease is also presented.

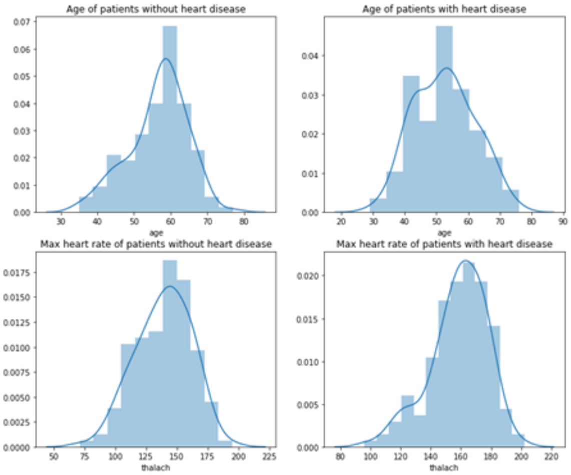

Fig.6. Distribution graph

VI. ALGORITHM OF WORKFLOW

6.1 Decision Tree

A decision tree is a classification algorithm for categorizing both categorical and numerical data. A decision tree is a structure that resembles a tree in appearance[11]. A decision tree is a simple and regularly used strategy for dealing with medical data. A tree-shaped graph's data is easy to implement and analyze. The decision tree model's analysis is based on three nodes.

- The main node, on which all other nodes are based, is known as the root node.
- Various properties are handled by the interior node.
- The result of each test is represented by the leaf node.

This program separates the data into two or more equivalent sets based on the most important indicators. The entropy of each attribute is calculated, and the data is separated into predictors with the largest information gain or the lowest entropy. The resulting findings are more straightforward to read and understand.

This technique is more accurate than other algorithms since it analyses the dataset in a tree-like graph. The data, on the other hand, could be overclassified, and only one attribute is considered at a time for decision-making. We obtained an accuracy of 76%(approx.) using decision tree on our dataset

6.2 KNN (K Nearest Neighbor)

The K-nearest-neighbors algorithm is a supervised classification approach. It categorizes items depending on how close they are to one another[11]. It is a good example of situational learning. The distance between an attribute and its neighbors is calculated using the Euclidean distance. It determines how to mark another point by using a set of named points.K-NN can be used to fill in the missing values in the data after the data has been grouped based on their similarity. The data set is then submitted to a variety of prediction algorithms after the missing values have been filled in. It is possible to enhance accuracy by using various combinations of these strategies.

The K-NN technique is simple to use and does not require the use of a model or any other assumptions. This algorithm can be used to perform classification, regression, and search[11]. Even though K-NN is the simplest method, it is influenced by noisy and irrelevant data. 86%(approx) of accuracy score was obtained when we performed KNN on our dataset

VII. RESULTS

The purpose of this research is to see if a patient will acquire heart disease. This research focused on supervised machine learning classification algorithms including decision trees and K-nearest neighbour using the UCI repository. A variety of tests using various classifier algorithms were conducted using the scikit learn library in python. The dataset was split into two parts: a training set and a test set.

The data is analyzed, pre-processed, and supervised classification algorithms including decision trees and K-nearest neighbour are used to get an accuracy score. Python programming was used to note the accuracy score outcomes of several classification approaches for training and test data sets.

We found out that both the algorithms worked pretty well but KNN outperformed Decision tree with 86%(approx.) accuracy results.

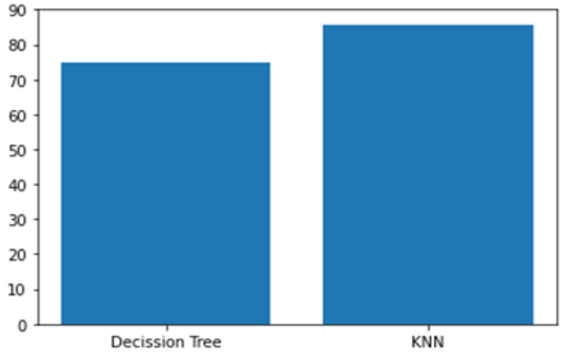

Fig.7. Comparing results- KNN vs Decision tree

VIII. PROJECT SCOPE

With the cooperation of numerous hospitals and patient healthcare data, the project can be taken to new heights. Also, if the right support is given, this project has the potential to make a big difference in the medical profession.

- This application can be integrated into a real-time system that incorporates sensors that track certain features. We will be able to obtain real-time forecasts and notifications based on the user's physical condition because of this.
- The program can be hosted on internet servers and made available to individuals all over the world.
- A master and slave database structure can be used to reduce database query overload.
- Set up a backup system to back up the codebase and databases on a regular basis.

IX. CONCLUSION

The major goal is to develop several data mining methods for properly forecasting heart illness. With fewer features and tests, we hope to give efficient and reliable prediction. In this analysis, only 14 key features are considered. The 2 machine learning categorization techniques we employed were K-nearest neighbour and decision tree.

Before being used in the model, the data was analyzed pre-processed. K-nearest neighbour, is the algorithm with the best results in this scenario. We determined that K-nearest neighbours had the highest accuracy of 86% after creating two algorithms. Given the limitations of this study, more complex and coupled models will be required to increase the accuracy of early heart disease prediction.

REFERENCES

[1] WHO- Cardiovascular Diseases(CVDs) https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Cardiovascular%20diseases%20(CVDs)%20are%20the,to%20heart%20attack%20and%20stroke. Last accessed 11 June 2021

[2] https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106#:~:text=If%20the%20clot%20is%20large,elevation%20myocardial%20infarction%20(STEMI).

[3] Mat Ghani, Ts. Dr. Mohd & Awang, Raflah. (2008). Intelligent heart disease prediction system using data mining techniques. 8. 108 - 115. 10.1109/AICCSA.2008.4493524.

[4] https://bigdata-madesimple.com/14-useful-applications-of-data-mining/

[5] https://www.webmd.com/heart-disease/risk-factors-for-heart-disease

[6] 6)https://www.indiatoday.in/sports/football/story/christian-eriksen-suffered-cardiac-arrest-confirms-denmark-team-doctor-euro-2020-denmark-vs-finland-group-c-1814377-2021-06-13

[7] https://www.nhs.uk/conditions/coronary-heart-disease/treatment/

[8] Singh P, Singh S, Pandi-Jain GS. Effective heart disease prediction system using data mining techniques. Int J Nanomedicine. 2018;13(T-NANO 2014 Abstracts):121-124. Published 2018 Mar 15. doi:10.2147/IJN.S124998

[9] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[10] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[11] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ ACCESS. 2020.3001149.

[12] C. Xiao, Y. Li and Y. Jiang, "Heart Coronary Artery Segmentation and Disease Risk Warning Based on a Deep Learning Algorithm," in IEEE Access, vol. 8, pp. 140108-140121, 2020, doi: 10.1109/ACCESS.2020.3010800.

[13] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., pp. 108-115, Mar. 2008.

[14] Kaviani, P., & Dhotre, S. (2017). Short survey on naive bayes algorithm. International Journal of Advance Engineering and Research Development, 4(11), 607-611.

[15] Devendra D. Borse , Prof. Dr. Suhas. H. Patil , Dr. Sunita Dhotre, Credit Card Fraud Detection Using Naive Bayes and Robust Scaling Techniques, of Research in Science, WARSE The World Academy, and Engineering. “Credit Card Fraud Detection Using Naive Bayes and Robust Scaling Techniques.” International Journal of Advanced Trends in Computer Science and Engineering (2021): n. page. Web.

[16] Prof. Dr.Mrs. Sunita Dhotre Shruti Rajendra Kudagi,Prof. Dr. S. H.Patil, Survey on prediction of crop yield by using different technique Journal of Emerging Technologies and Innovative Research (JETI)

[17] A. J. A. Majumder, Y. A. ElSaadany, R. Young and D. R. Ucci, "An energy efficient wearable smart IoT system to predict cardiac arrest", Adv. Hum.-Comput. Interact., vol. 2019, pp. 1-21, Feb. 2019.