

Heart Disease Prediction Using Machine Learning Techniques

GONDI NAGA VAMSI KRISHNA

Gondi Naga Vamsi Krishna & Department of computer science and Engineer,
PVKK Institute of technology, Anantapur, Andhra Pradesh.

ABSTRACT

Chronic heart failure (CHF) affects over 26 million of people worldwide, and its incidence is increasing by 2% annually. Despite the significant burden that CHF poses and despite of sensors in our lives methods. The Heart Disease prediction is a machine learning based system designed to predict the likelihood of heart disease in individuals using the widely adopted UCI Heart Disease Dataset. The project implements a complete ML pipeline that include data preprocessing, feature selection, model training, evaluation, and deployment, enabling users to generate real-time predictions through an interactive web interface. The dataset cleaned, normalized. Multiple classification algorithm such as logistic regression and random forest classifiers to determine the most accurate predictor of heart disease risk. Model performance is evaluated using accuracy, precision, and other classification metrics to identify the best-performing approach.

INDEX TERMS: Python, Naive Bayes, Random Forest Classification,

I. INTRODUCTION

Chronic heart failure (CHF) is a progressive condition underscored by the hearts in ability to supply enough to target tissues and organs at the physiological filling pressures to meet their metabolic demand. CHF has reached epidemic proportions in the population, as its incidence is increasing by 2% annually. In the developed world, CHF affects 1-2% of the total population and 10% of people older than 65 years. Currently, the diagnosis and treatment the CHF use approximately 2% of the annual healthcare budget. In absolute terms.

the other countries spent approximately 35 billion. Over the last decade, heart disease remains the leading killer globally, with concerning trends showing rising risk factors, especially in younger adults, while overall death rates saw temporary dips followed by recent increases, particularly in older groups, driven by lifestyle (diet, inactivity, stress) and genetic predispositions, notably in India where age of onset is decreasing. Key drivers include urbanization, poor diets (unhealthy fats, sugar), inactivity, rising obesity/diabetes, high blood pressure, and stress.

major global. In this research, we aim to investigate the effectiveness of various machine learning (ML) algorithms in predicting heart disease. To achieve this goal, we employed a variety of techniques, including random forest decision, multilayer perceptron to build predictive models.

The age of person in years, there are checking the person sex like whether male or female. The chest pain contains the 4 types; the cp values are code as:

Cp=1 ->Typical angina causes STABLE ANGINA Chest pain, discomfort, tightness, occurs when the heart muscles do not receive enough oxygen-rich blood.

Cp=2 -> Atypical angina causes chest pain that is not typical for classic angina, but stall possibly heart relates.

Cp=3-> non-anginal pain chest pain is unlikely to be caused by heart disease. dataset was changed from Numeric to Nominal in order to make the dataset compatible with the ML Techniques used.

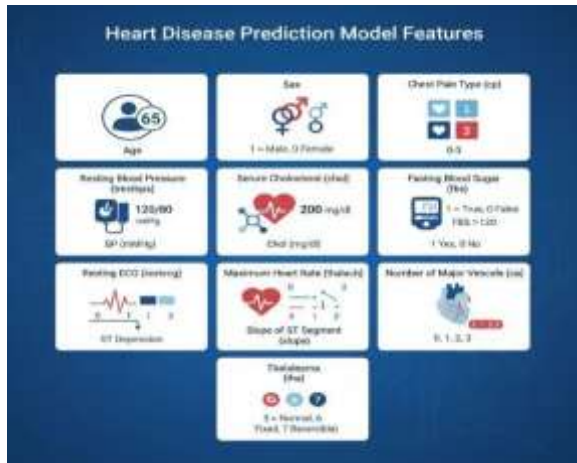


FIGURE 1: Heart Disease Prediction Model Features for the UCI Heart Disease Datasets

Cp=4 -> Asymptomatic the patient has no chest pain at all, but other tests like ECG or angiography show evidence of heart disease.

For the figure-1 **resttbps**-Resting blood pressure at rest. In the **chol** – Serum Cholesterol Indicates the level of total cholesterol in the blood, **fbs**-Fasting blood sugar > 120 mg/dl (1=true/0=false), A binary number indication showing the whether the patient's fasting blood glucose is above 120 mg/dl, **restecg** -

Resting electrocardiographic results 0-2. In the case-0 show the normal, Case-1=ST-T wave abnormality, Case-2 In case left ventricular hypertrophy in the **thalach** Maximum heart rate achieved. **Exang**-Exercise induced angina using the binary number 0 or 1. whether the patient experience chest pain during physical exertion. **Oldpeak**: ST depression induced by exercise relative to rest. **slope**: – Slope of the peak exercise ST segment (0–2). Describes the shape of the ST segment at peak exercise: Case=0 => upsloping, Case = 1 => flat, Case= 2 => down sloping. **CA** – Number of major vessels colored 0-3. Counts how many of the major coronary arteries show narrowing when viewed by fluoroscopy. **thal** – Thalassaemia the normal = 3; fixed defect = 6; reversible = 7. In the Finally, ML algorithms are trained to classify healthy vs unhealthy UCI Dataset based on those features.

II. RELATED WORK

Heart Disease Prediction is a standard using the UCI Cleveland dataset, it common machine learning

equations like the Logistic Regression, Random Forest / Decision Tree and using the Evaluation Metrics.

$$P(y=1 | \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{13} x_{13})}}$$

Where:

P (y=1 |x) =probability of heart disease (target =1) given features then the (x1, x2, x3 xn) are the 13 input features (age, sex, cp, resttbps, etc.), $\beta_0, \beta_1, \beta_2, \dots, \beta_{13}$ are the learned coefficients.

Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

F1-score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

III. DATA DESCRIPTION

This project using the Widely Studied UCI Disease schema, where each row is a patient and columns represent demographic, symptomatic and test-base attributes. In the attached data, for the Table 1. show the cleaned 14-feature subset as show in the Table 1. By using Methodology and data preparation To the feed this clinical variable into machine learning models, the performs several preprocessing steps and saves intermediate datasets in the provided CSV files. Numerical feature-selection. So, there are 1025 instances in total, which have been used in this research work. Table 1 shows the mean value of each attribute. The dataset used have missing values too, which is handled through data preprocessing using some statistical techniques. Class Value 1, is interpreted as "tested positive for the disease", and Class value 0, means "tested negative for the disease". The dataset has been divided into certain percentage, such as 80% data has been considered as training data and rest 20% as data. Accuracy of model. After that the values of the

IV.MACHINE LEARN TECHNIQUES

We have chosen four popular ML techniques to develop the heart disease prediction model. Details of these techniques are as follows:

A. Decision Tree

Decision Tree algorithm in Machine Learning is used to develop the Classification models. This classification model is based on the tree-like structure. This comes under the category of supervised learning, where the target result is categorical and numerical data can be applied on Decision tree algorithm. Decision tree consists of root node, branches and leaf nodes.

Data is evaluated on the basis of traversing path from the root to a leaf node. For our dataset - CHDD, a total of 283 tuples were assessed down the decision tree. They potentially came to a positive or negative assessment for the heart disease prediction. These were compared to the actual parameters to check for the false positives/false negatives which show the accuracy, specificity, and sensitivity of the model.

B. Naive Bayes

This supervised machine-learning algorithm is based on the Bayes' Theorem, which consider that features are statistically independent to each other's.

The Naïve Bayes Classifier is used with high dimensionality of inputs data. Naive Bayes method useful in computer vision application. In particularly it has proven itself to be a classifier with good results.

C. Naive Bayes classification

This supervised machine-learning algorithm is based on the Bayes' Theorem. Naive Bayes is a probabilistic classifier depending on Bayes' theorem.

It assumes that all features are independent of each other given the class. It is fast, simple, and works well for text classification and spam detection.

D. Random Forest Classification

Random Forest is a troupe of unpruned classification- based trees. It gives amazing performance with concern to number of real-life problems, as it is non effective to noise in the dataset and risk of overfitting

Table 1. show the demographic, symptomatic, and test-base attributes.it will show the cleaned 14 - feature subset.

age	63	67	67	37	41	56	62	57	53
sex	1	1	1	1	0	1	0	0	1
cp	0	3	3	2	1	1	3	3	3
trestbps	145	160	120	130	130	120	140	120	140
chol	233	286	229	250	204	236	268	354	203
fbs	1	0	0	0	0	0	0	0	1
restecg	1	1	1	0	1	0	1	0	1
thalch	150	108	129	187	172	178	160	163	155
exang	0	1	1	0	0	0	0	1	1
oldpeak	2.3	1.5	2.6	3.5	1.4	0.8	3.6	0.6	3.1
slope	2	1	1	2	0	0	2	0	2
ca	0	3	2	0	0	0	2	0	0
thal	2	1	3	1	1	1	1	1	3
target	0	1	1	0	0	0	1	0	1

Table 2. indicators for coronary heart disease and Standard subset used UCI Heart disease dataset.

id	Person 1	Person 2	Person 3	Person 4
age	1.007386	1.432034	1.432034	-1.75283
sex	0.516931	0.516931	0.516931	0.516931
cp	-2.41815	0.806051	0.806051	-0.26868
trestbps	0.705176	1.518569	-0.65048	-0.10822
chol	0.303643	0.789967	0.266939	0.459634
fbs	2.380476	-0.42008	-0.42008	-0.42008
restecg	0.51201	0.51201	0.51201	-0.74748
thalch	0.489727	-1.18148	-0.34588	1.961979
exang	-0.76029	1.315283	1.315283	-0.76029
oldpeak	1.368109	0.611589	1.651804	2.502889
slope	2.234081	0.295067	0.295067	2.234081
ca	-0.3614	4.411152	2.820301	-0.3614
thal	0.006695	-1.53319	1.546578	-1.53319

For the Table 2. Research shows features like cp, thalach, slope, oldpeak, other features information for distinguishing patients with and without heart disease.



In X_scaled.csv the values are around -3 to +5, which means a standardization was applied so each feature has mean ≈ 0 and standard deviation ≈ 1 . the scaled features for table 2. data are fed into a classifier learn a mapping: **Inputs = [age, sex, cp, ..., thal]** --> **output = heart disease(yes/no)**. new patient's measurements and predict the probability.

For the table 3 is used the Data as the input features for the ML model; these are standard UCI heart- disease variables, In this project, each column's role is as an input feature; together they form X that the model uses to predict the target (heart disease yes/no). A typical roles mapping for :

- Cp, chol, thalach, age, oldpeak, exang, trestbps, thal, sex, restecg → predictor feature
- Target → output label.

Table 3. indicator for show the model on features with predictive power, improving performance and reducing overfitting

cp	-2.41815	0.806051	0.806051	-0.26868	-1.34342	-1.34342
chol	0.303643	0.789967	0.266939	0.459634	0.037541	0.331171
thalch	0.489727	-1.18148	-0.34588	1.961979	1.36512	1.603864
age	1.007386	1.432034	1.432034	-1.75283	-1.32818	0.264251
oldpeak	1.368109	0.611589	1.651804	2.502889	0.517024	-0.05037
exang	-0.76029	1.315283	1.315283	-0.76029	-0.76029	-0.76029
trestbps	0.705176	1.518569	-0.65048	-0.10822	-0.10822	-0.65048
thal	0.006695	-1.53319	1.546578	-1.53319	-1.53319	-1.53319
sex	0.516931	0.516931	0.516931	0.516931	-1.93449	0.516931
restecg	0.51201	0.51201	0.51201	-0.74748	0.51201	-0.74748

Table 4. Indication provides the input features and it provides the correct answer for each row.

Id	1	2	3	4	5	6	7	8	9
Target	0	1	1	0	0	0	1	0	1

During the process, the classifier will learn to map from x_scaled to y_target, allowing it to predict heart disease in the future. It also analyzes and visualizes confusion matrix, true positive, precision, recall, and false negative, etc. It is open source, as well as platform-independent, portable, GUI-based software, and it is packed with a lot of advanced machine learning techniques like deep learning, image processing, etc. Four measures of accuracy have been used to compare the performance of the four models, as follows: • Precision or Positive Predictive Value It is the average probability of relevant retrieval. Precision = Number of true positives/Number of true positives + False positives.

• Recall:

It is the average probability of complete retrieval. Recall= True positives/True positives + False negative.

• Accuracy:

The accuracy of the classifier is defined as the percentage of the correct prediction of the total number of correct predictions divided by the total number of instances.

Accuracy = [Number of True Positives + True Negatives]/[Total Instances]

• ROC Area

ROC curves illustrate the trade-off between the true positive rate and the false positive rate of a classification system.

True Positive Rate:

$$\frac{[\text{Number of True Positives}]}{[\text{Number of True Positives} + \text{False Negatives}]}$$

False Positive Rate:

$$\frac{[\text{Number of False Positives}]}{[\text{Number of False Positives} + \text{True Negatives}]}$$

10-Fold Cross Validation splits the data into two sections: the training dataset and the testing dataset

V.RESULT

At the end of our experiment, result shows that Random Forest performed very well in comparison to Gaussian Naive Bayes and Decision tree. Model developed with approximately 13% greater than Decision tree. In the same way, model build with Random Forest gives best prediction result with 99% accuracy, which is itself more accurate than our second best SVM model for heart disease prediction. Unfortunately, we did not find decision tree suitable for our data.

TABLE 5. PERFORMANCE MEASURE OF MODEL

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.995	0.995	0.995	0.995
Random Forest	0.997	0.997	0.997	1.00
(Any other model...)	0.904	0.904	0.904	0.904

VI. CONCLUSION

Through this research we have attempted to analyze the various machine learning techniques and anticipate if someone in particular, given different individual attributes and indications, will get coronary illness or not. The primary thought process of our report was to looking at the exactness and analyzing the reasons behind the variation of different algorithms. We have used Cleveland dataset for heart diseases which contains 1025 instances and used percent split to divide the data into two sections which are training and testing datasets. We have considered 14 attributes and implemented four different algorithms to examine the accuracy. By the end of the implementation part, we have discovered that Random Forest is giving the maximum accuracy level in our dataset which is 99 percent and Decision Tree is

playing out the least with an accuracy level of 85 percent. Probably for other instances and different datasets other algorithm may work in better manner however for our situation, we have discovered this outcome. Also, on the off chance that we increment the number of training data, maybe we can find more accurate result but it will take more time to process and the system will be slower than now as it will be more perplexing and will be handling more data.

REFERENCES

- [1] <https://pmc.ncbi.nlm.nih.gov/articles/>
- [2] <https://jmai.amegroups.org/article/view/9054/html>
- [3] https://en.wikipedia.org/wiki/Decision_tree_learning
- [4] https://en.wikipedia.org/wiki/Bayes27_theorem [5] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- [6] <https://towardsdatascience.com/understanding-random-forest58381e0602d2>
- [7] <https://archive.ics.uci.edu/ml/datasets/heart+disease>
- [8] <https://www.youtube.com/watch?v=AEZ-ooHbZVE>
- [9] <https://www.sciencedirect.com/science/article/pii/S110866524001221>
- [10] Zhiqiang, Zhiyuan Song, Steven X. Ding, Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE Transactions on content mining and analysis, vol. 5, pp. 20590-20616, 2017