# Predicting Heart Disease using Data Mining and Machine Learning Techniques

Vaishali Sarde
Govt. J. Yoganandam Chhattisgarh College, Raipur (C.G.) India

Pankaj Sarde
Rungta College of Engineering & Technology, Bhilai (CG), India

## ABSTRACT
Heart-Disease has become a very common and serious problem among people worldwide in recent years. Now days many advanced technologies have evolved in treating heart disease. Medical practitioner's required reliable automated system which can accurately and efficiently major the problem belong to the patient. Early diagnosis of the disease can help the medical persons in treating the patient and saving the life. There are many techniques available which can be used to predict the heart disease by analysing the health-related parameters. This paper emphases on five different techniques from datamining and machine learning to predict heart disease. Comparative study among these techniques has been presented. Five techniques K-means, Decision Tree, Support Vector Machine, Naive Bayes and Artificial Neural Network are implemented. The dataset is taken from Kaggle repository. It combines 5 different datasets, with 11 features and detail of 1190 persons, which makes it the largest heart disease dataset.

## Keywords
SVM, K-means, Decision Tree, Navie Bayes, Artificial Neural Network, Machine Learning

## 1. INTRODUCTION
Heart disease is a general term that includes many types of heart problems. It's also called cardiovascular disease, which means heart and blood vessel disease. Heart disease is the leading cause of death of large populations, but there are ways to prevent and manage many types of heart disease. There are many different factors that can make more likely to develop heart disease. Some of these factors are age, Family history and genetics, life style, medical conditions etc. Medical Diagnosis is tough and lengthy process to do perfectly and effectively. Some kind of automated system is required which can analyse the parameters perfectly and provide the accurate result. Medical practitioners use their experience and knowledge in diagnosis of disease but there may be lacking in analysis of parameters. This lacking can be covered using the different techniques. This paper focuses on the implementation of five different algorithms k-means, Decision Tree, Support Vector Machine, Naive Bayes and Artificial Neural Network to identify the heart disease using heart related parameters. 11 important parameters are used and algorithms are compared in terms of accuracy, f1-score, precision, recall and support values.

## 2. LITERATURE REVIEW
Hazra, A., et. al [1] proposed algorithm was validated using two widely used open-access database, where 10-fold cross-validation is applied in order to analyze the performance of heart disease detection. An accuracy level of 97.53% accuracy was found from the SVM algorithm along with sensitivity and specificity of 97.50% and 94.94%respectively.

Napa K. et. al. [2] Intends to look at the presentation of machine learning tree classifier. In this investigation of foreseeing cardiovascular disease the random woodland machine learning classifier accomplished a higher precision of 85%, Roc AUC score of .8675 and execution time of 1.09 sec.The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed

Velmurugan T. et. al.[3] Proposed the method uses Named Entity Recognition (NER) algorithm to discover the equivalent words for the coronary illness content to mine the significance in clinical reports and different applications. The Heart sickness text information given by the physician is taken for the preprocessing and changes the text information to the ideal meaning, at that point the resultant text data taken as input for the prediction of heart disease.

Patil C et. al [4] stated that it is very important to monitor various medical parameters and post operational days. Hence the latest trend in Health care communication method using IoT is adapted. In this work the AVR-328 microcontroller (Arduino board) is used as a gateway to communicate to the various sensors such as temperature sensor, heartbeat sensor, ECG sensor, sensor for keeping a track of drip levels(blood or saline) and a sensor to keep track of motion. The micro-controller picks up the sensor data and sends it to the network through WiFi and hence provides real time monitoring of the health care parameters for doctors.

Beyene C.et. al. [5] proposed the methodology to predict the occurrence of heart disease for early automatic diagnosis of the disease within retrieve result in short time This p lay vital roles for healthcare experts to treat their patients based on accurate decision-making and give qualities of services to the people. The proposed methodology is also critical in healthcare Organization with experts that have no more knowledge and skill.

Raihan M et. al. [8] presented a structure of the work exercised through the DM methods which helps in the prediction and treatment of various diseases. This examines two problems (elongated work based on prediction of RCT and Heart Disease) with classification technique based on cross validation, decision tree to discover RCT, split validation and model to detect heart disease before probing advice from the doctor. The important aim of this paper is to treat whether there is a need of RCT or not with different input values of an attributes. All these symptoms or attributes are examined to arrive at a decision that person with an age greater than eighteen and having sensitivity issues have greater probability of root canal treatment. There are more chances of patient being prone to a heart problem through attributes like age, root canal treatment, diabetes and smoking.

# 3. DATASET

## 3.1 Dataset Source

Heart disease dataset (Comprehensive) from Kaggle repository [9] is used for experiment. This dataset is formed by combining 5 different datasets over 11 common features which makes it the largest heart disease dataset available for research purposes. The five datasets used for its creation are:

**Database instances:**

1. Cleveland: 303
2. Hungarian: 294
3. Switzerland: 123
4. Long Beach VA: 200
5. Stalog (Heart) Data Set: 270

**Total 1190 records**

## 3.2 Dataset Features Description

**Table 1: Description of the features of dataset**

| S. No. | Feature | Description |
|---|---|---|
| 1 | age | Patient's Age in years |
| 2 | sex | Patient's Gender Male as 1 Female as 0 |
| 3 | chest pain type | Type of chest pain categorized into 1 typical, 2 typical angina, 3 non-anginal pain, 4 asymptomatic |
| 4 | resting bp s | Level of blood pressure at resting mode in mm/HG |
| 5 | cholesterol | Serum cholestrol in mg/dl |
| 6 | fasting blood sugar | Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false |
| 7 | resting ecg | result of electrocardiogram while at rest are represented in 3 distinct values 0 : Normal 1: Abnormality in ST-T wave 2: Left ventricular hypertrophy |
| 8 | max heart rate | Maximum heart rate achieved |
| 9 | exercise angina | Angina induced by exercise 0 depicting NO 1 depicting Yes |
| 10 | Old peak | Exercise induced ST-depression in comparison with the state of rest (Numeric) |
| 11 | ST slope | ST segment measured in terms of slope during peak exercise 0: Normal 1: Upsloping 2: Flat 3: Downsloping |
| 12 | target | Heart Risk 1 means heart disease 0 means normal |

## 3.3 Dataset Description

### *3.3.1 Counting of Each Target Class*

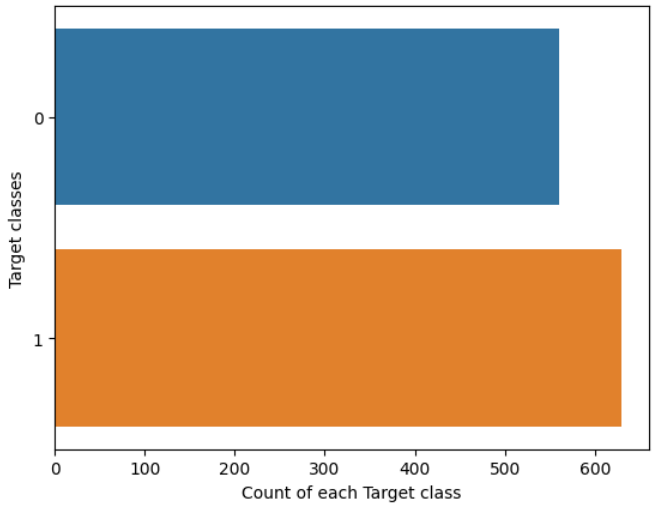


**Fig 1: Counting of each target class**

### *3.3.2 Feature Distribution*

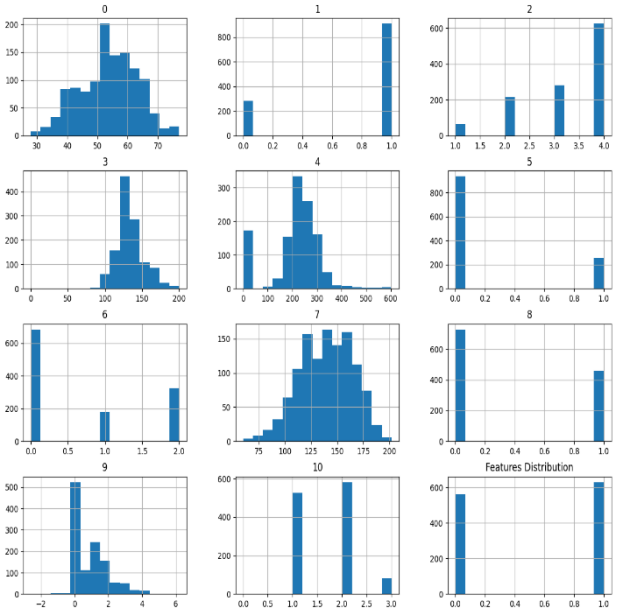


**Fig2: Feature Distribution**

### *3.3.3 Heat Map*

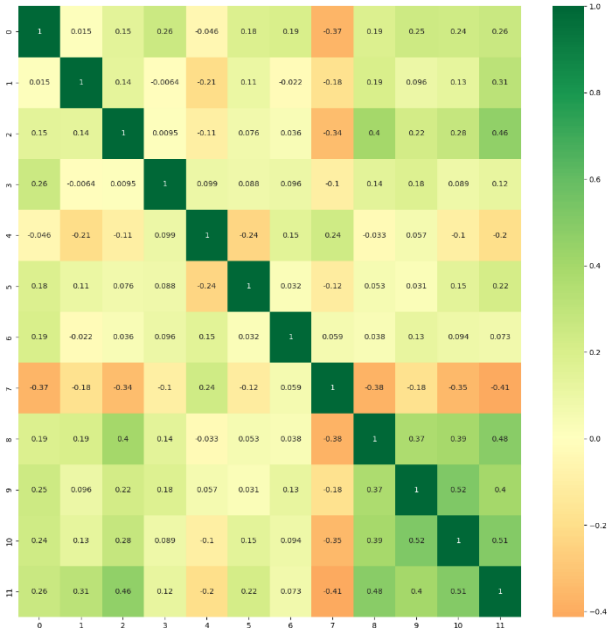

**Fig3: Heat Map**

## 3.4 Splitting Dataset

Dataset is divided between training and testing data set. 67% data is used for training purpose and remaining 33% data is used for testing purpose.

# 4. METHODOLOGIES

## 4.1 K-Means Clustering

K-Means Clustering is an Unsupervised Learning algorithm. It groups the unlabeled dataset into different clusters. Here K is the number of pre-defined clusters. if K=2, clusters will be created, and if K=3, three clusters will be created, and so on.

## 4.2 Decision Tree

Decision Tree is a Supervised learning technique. It is used for classification and Regression problems. It is a tree-like structure, where internal nodes are represented by the features of a dataset, branches are represented by the decision rules and leaf nodes are represented by the result.

It is a graphical way to represent the all possible solutions of given problem on the basis of the feature values.

## 4.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification. SVM algorithm is used to find the optimal hyperplane to separate the data points in different classes. SVM tries to maintain gap between the nearest points of different classes to be as large as possible.

## 4.4 Navie Bayes

Naïve Bayes algorithm is used for supervised learning. It is based on Bayes theorem. It is basically apply for classification problems.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence

## 4.5 Artificial Neural Network(ANN)

Artificial Neural Networks consists of artificial neurons also called as units. These units are arranged in different layers that together form a artificial neural network. ANN has three types of layers, input layer, hidden layers and output layers. Input layer takes the real world data and passed it to multiple hidden layers which process and transform data and passed it to the output layer. Output layer holds the units belongs to the different classes.

The proposed model for this paper is as follows

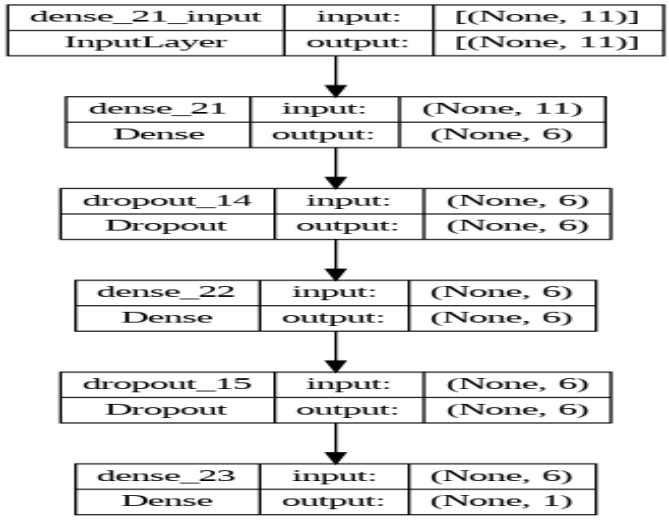


**Fig4: Proposed ANN Model**

# 5. RESULTS AND ANALYSIS

## 5.1 k-Means Clustering

**Table2: Result k-means**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.12 | 0.04 | 0.06 | 561 |
| 1 | 0.47 | 0.75 | 0.57 | 629 |
| accuracy | | | 0.41 | 1190 |
| macro avg | 0.29 | 0.39 | 0.32 | 1190 |
| weighted avg | 0.30 | 0.41 | 0.33 | 1190 |

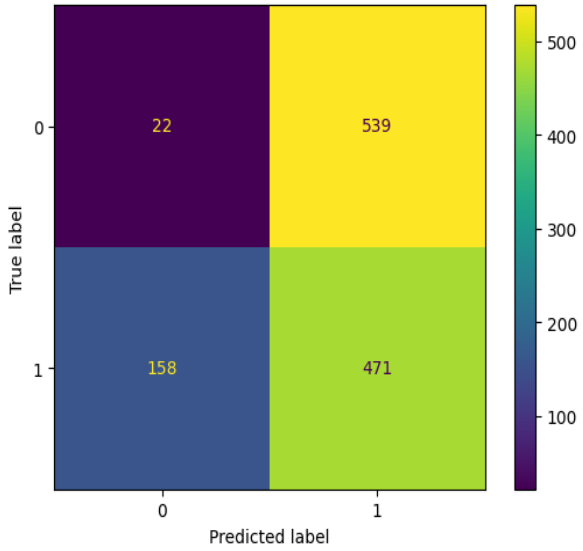


**Fig 5: Confusion Matrix K-Means**

## 5.2 Decision Tree

**Table3: Result Decision Tree**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.81 | 0.85 | 0.83 | 170 |
| **1** | 0.88 | 0.85 | 0.87 | 223 |
| **accuracy** | | | 0.85 | 393 |
| **macro avg** | 0.85 | 0.85 | 0.85 | 393 |
| **weighted avg** | 0.85 | 0.85 | 0.85 | 393 |

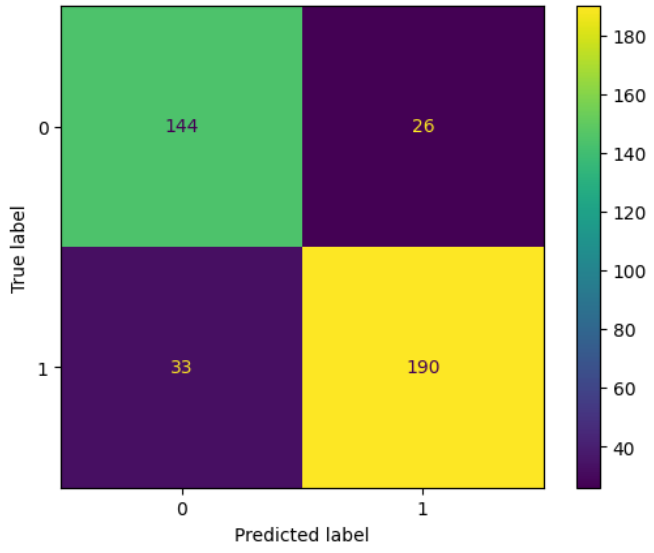


**Fig 6: Decision Tree Confusion Matrix**

## 5.3 Support Vector Machine

**Table 4: Result SVM**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.82 | 0.84 | 0.83 | 170 |
| **1** | 0.87 | 0.86 | 0.87 | 223 |
| **accuracy** | | | 0.85 | 393 |
| **macro avg** | 0.85 | 0.85 | 0.85 | 393 |
| **weighted avg** | 0.85 | 0.85 | 0.85 | 393 |

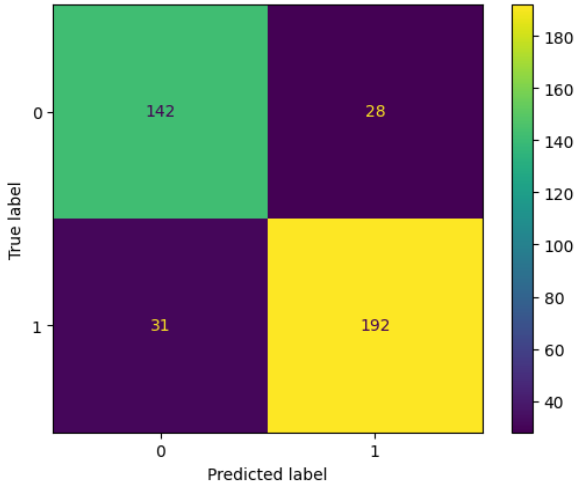


**Fig 7: SVM Confusion Matrix**

## 5.4 Navie Bayes

**Table 5: Result Navie Bayes**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.82 | 0.84 | 0.83 | 170 |
| **1** | 0.87 | 0.86 | 0.87 | 223 |
| **accuracy** | | | 0.85 | 393 |
| **macro avg** | 0.85 | 0.85 | 0.85 | 393 |
| **weighted avg** | 0.85 | 0.85 | 0.85 | 393 |

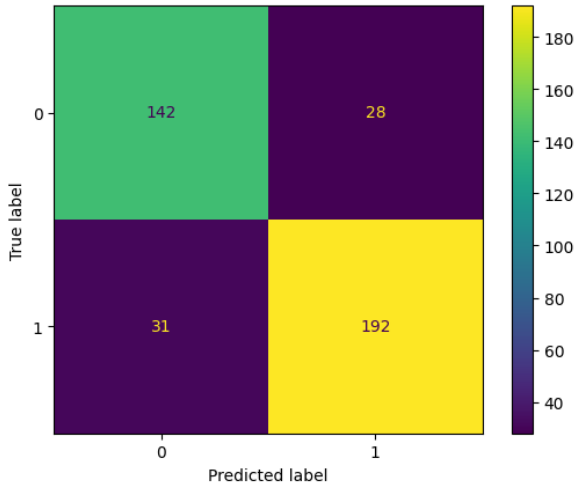


**Fig 8: Navie Bayes Confusion Matrix**

## 5.5 Artificial Neural Network(ANN)

**Table 6: Result ANN**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.77 | 0.89 | 0.83 | 170 |
| **1** | 0.90 | 0.80 | 0.85 | 223 |
| **accuracy** | | | 0.84 | 393 |
| **macro avg** | 0.84 | 0.84 | 0.84 | 393 |
| **weighted avg** | 0.85 | 0.84 | 0.84 | 393 |

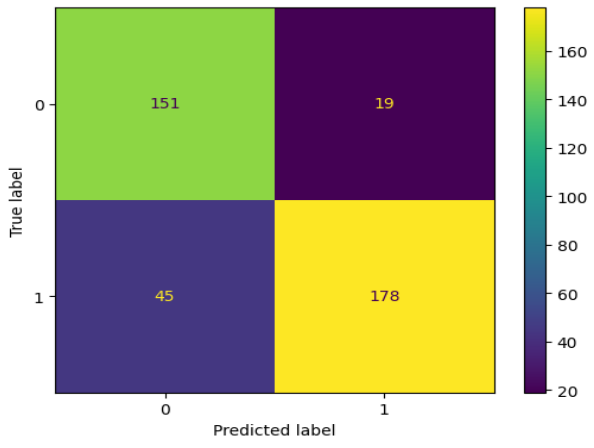


**Fig 9: ANN Confusion Matrix**

## 5.6 Comparison among five algorithms

**Table 7: Accuracy for five algorithms**

| S. No. | Accuracy |
|---|---|
| K-Means | 41.42% |
| Decision Tree | 84.98% |
| Support Vector Machine | 85.01% |
| Navie Bayes | 85.03% |
| Artificial Neural Network | 83.71% |

Table 7 shows that the all the algorithms are performing good except k-means on heart disease dataset. The accuracy of support vector machine and Navie Bayes are almost same. ANN model has the scope of improvement to increase the accuracy.

# 6. CONCLUSION

The objective of this study is to apply various datamining and machine learning techniques for heart disease prediction and determining the effective and accurate technique to predict heart disease. Five different techniques from data mining and machine learning, K-means, Decision tree, Support Vector Machine, Navie Bayes and ANN are used for prediction of heart disease. The dataset from Kaggle repository is used for experimental purpose. As result shows Support Vector Machine and Navie Bayes performance are almost same and best among others but the performance of ANN model can be improved by varying the number of neurons and the number of hidden layers.

# 7. REFERENCES

[1] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, A. 2017. Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology, 10, 2137-2159.

[2] Napa K, Sarika Sindhu G., Krishna Prashanthi D,. Shaeen Sulthana, Aril 2020. A. Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers, 2020 6th International Conferenc on Advanced Computing & Communication Systems (ICACCS.

[3] Velmurugan T., Latha U. January 2021Classifying Heart Disease in Medical Data Using Deep Learning Methods Journal of Computer and Communications. Vol.9 No.1.

[4] Patil C, A. B. and Sonawane, P. 2017 To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients. International Journal on Emerging Trends in Technology (IJETT), 4, 8274-8281.

[5] Beyene C., Kamat P., January 2018, Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques, International Journal of Pure and Applied Mathematics 118(8):165-173

[6] Methaila A., Kansal P., Arya H., and Kumar P.,

2014. Early heart disease prediction using data mining techniques, Computer Science & Information echnology Journal , pp. 53 -59, 2014.

[7] Chauhan R, Bajaj P, Choudhary K, Gigras Y.2015 Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[8] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K., December 2016. Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

[9] Siddhartha M. Kaggle dataset repository https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final

[10] Nagamani T., Logeswari S., B.Gomathy, January 2019 Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3.