# HEART DISEASE PREDICTION USING RANDOM FOREST

**Project Report submitted in partial fulfillment of the**

**Requirements for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**by**


**SHAIK MOHAMMAD SOHAIL (1012002019)**

**M.HARSHITHA (1012002011)**

**G. CHARAN KUMAR REDDY (1012002004)**

**K.VIVEK (1012102908)**


Under the esteemed guidance of

T.MUKTHAR AHAMED , M.Tech

Academic consultant



## Department of Computer Science and Engineering

## Y.S.R ENGINEERING COLLEGE

## OF YOGI VEMANA UNIVERSITY

**PRODDATUR-516360, Y.S.(Dt.),A.P.**


**(2023-2024)**

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

## Y.S.R. ENGINEERING COLLEGE OF YOGI VEMANA UNIVERSITY

## PRODDATUR -516360, Y.S.R. (Dt.), A.P.

## CERTIFICATE

This is to certify that the project report entitled "HEART DISEASE PREDICTION USING RANDOM FOREST " is submitted by **SHAIK MOHAMMAD SOHAIL , M. HARSHITHA, G.CHARAN KUMAR REDDY, K.VIVEK ,** in partial fulfillment of the requirement for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING, Y.S.R. ENGINEERING COLLEGE OF YOGI VEMANA UNIVERSITY ,PRODDATUR , is a record of bonafide work carried out by them under my guidance and supervision.

**PROJECT GUIDE** **PROJECT CO-ORDINATOR** **HEAD OF THE DEPARTMENT**

Examiners:

**1.** **2.**

# ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without mention of the people, who made it possible, whose constant guidance and encouragement crowned our efforts with success. We take this opportunity to express my deepest gratitude and appreciation to all those who have helped us directly or indirectly towards the successful completion of this project.

It is a great pleasure to express a deep sense of gratitude and veneration to our guide Sri. T.MUKTHAR AHAMED, academic consultant in the Department of Computer Science and Engineering for his valuable guidance and thought-provoking discussion throughout the course of project work.

We extend our profound gratefulness to our professors **Dr. S. KIRAN**, Associate professor and Head of the Computer Science and Engineering and **Dr. R.PRADEEP KUMAR REDDY,** Project Coordinator and Associate professor for their encouragement and support throughout the project.

We take this opportunity to offer gratefulness to our Honorable Vice-Chancellor **Prof . CHINTA SUDHAKAR Garu**, our Dean of Engineering **Prof. K. VENKATA RAMANAIAH**, our Principal **Prof. C. NAGARAJU** for providing all sorts of environment during the project work.

We express our thanks to all our college teaching and non-teaching staff members who encouraged and helped us in some way or other throughout the project work.

**SHAIK MOHAMMAD SOHAIL (1012002019)**
**M.HARSHITHA (1012002011)**
**G. CHARAN KUMAR REDDY (1012002004)**
**K.VIVEK (1012102908)**

# TABLE OF CONTENT

## LIST OF FIGURES

# ABSTRACT

# ABSTRACT

Cardiovascular diseases are a leading cause of morbidity and mortality globally, underscoring the importance of accurate and reliable predictive models. This study focuses on the application of the Random Forest algorithm to predict heart disease using the well-known Cleveland Heart Disease dataset. The dataset encapsulates a diverse array of patient attributes, clinical markers, and diagnostic outcomes.

Random Forest model is trained on a subset of the Cleveland Heart Disease dataset, leveraging an ensemble of decision trees to make robust predictions. Hyperparameter tuning is performed to optimize the model's performance, and the final model is rigorously evaluated on an independent test set.

The study underscores the vital importance of accurate heart disease prediction, emphasizing the potential for early detection to mitigate adverse health outcomes. Leveraging the Random Forest algorithm offers a promising avenue due to its ability to handle complex relationships within the data and provide interpretable insights. Evaluation metrics including accuracy, precision, recall, and F1-score provide a comprehensive assessment of the model's predictive capabilities.

This study not only highlights the importance of heart disease prediction but also ultimately reduce the burden of cardiovascular diseases on global public health and demonstrates the practical application of advanced machine learning techniques, exemplified by the Random Forest algorithm, on a well-established dataset. The insights garnered hold promise for improving healthcare outcomes and advancing predictive modeling in the realm of cardiovascular health.

# CHAPTER 1

# INTRODUCTION

## 1.1 HEART DISEASE

symptoms such as chest pain (angina), shortness of breath, or in severe cases Heart disease, also known as cardiovascular disease (CVD), refers to a range of conditions that affect the heart. These conditions can include coronary artery disease (narrowing or blockage of the blood vessels that supply the heart muscle), heart failure (a condition where the heart can't pump blood effectively), arrhythmias (irregular heartbeats), and various congenital heart defects.

The most common type of heart disease is coronary artery disease, which is often caused by the buildup of plaque (made up of cholesterol, fat, and other substances) in the arteries that supply blood to the heart muscle. This buildup, known as atherosclerosis, can restrict blood flow to the heart and lead to, heart attack.

Other types of heart disease can affect the heart's valves, its electrical system, or the heart muscle itself. Risk factors for heart disease include high blood pressure, high cholesterol, smoking, obesity, diabetes, a sedentary lifestyle, and a family history of heart disease.

Prevention and management of heart disease often involve lifestyle changes such as eating a healthy diet, getting regular exercise, maintaining a healthy weight, not smoking, and managing stress. In some cases, medication or surgical procedures may also be necessary to treat or manage heart disease. Early detection and treatment are crucial for preventing complications and improving outcomes for individuals with heart disease.

Fig 1.1 Heart

.

## 1.2 PREDICTION OF HEART DISEASE

Prediction of heart disease involves assessing an individual's risk factors, clinical markers, and other relevant information to estimate their likelihood of developing cardiovascular problems in the future. This predictive approach is crucial for several reasons:

1. Early Intervention:

Predictive models help identify individuals at higher risk of developing heart disease before symptoms manifest. Early intervention, such as lifestyle changes or medication, can be initiated to prevent or delay the onset of cardiovascular events like heart attacks or strokes.

2. Risk Stratification:

Not all individuals face the same risk of heart disease. Predictive models allow healthcare providers to stratify patients based on their risk levels, enabling personalized preventive strategies. High-risk individuals may require more intensive monitoring and interventions compared to those at lower risk.

3. Resource Allocation:

Healthcare resources are finite, and prioritizing care for those at highest risk can optimize resource allocation. Predictive models assist healthcare systems in targeting interventions to the individuals who stand to benefit the most, thereby maximizing the efficiency of healthcare delivery.

4. Patient Empowerment:

Knowledge of one's risk for heart disease empowers individuals to take proactive steps to protect their cardiovascular health. This may include lifestyle modifications such as improving diet, increasing physical activity, quitting smoking, and managing underlying health conditions like hypertension or diabetes.

5. Population Health Management:

By identifying populations at higher risk of heart disease, predictive models inform public health initiatives aimed at reducing overall disease burden. These initiatives may include community-based education campaigns, policy changes to promote healthier environments, and screening programs for at-risk populations.

6. Research and Development:

Predictive models contribute to ongoing research efforts aimed at better understanding the complex factors underlying heart disease. By identifying novel risk factors or refining existing predictive algorithms, researchers can improve the accuracy of risk assessment tools and develop more effective preventive strategies.

Overall, the prediction of heart disease is essential for guiding clinical decision-making, optimizing healthcare resources, empowering individuals to take control of their health, and advancing our collective efforts to reduce the burden of cardiovascular disease on a population level. It represents a cornerstone of preventive cardiology and public health initiatives aimed at improving cardiovascular outcomes and enhancing overall well-being.

## 1.3 RANDOM FOREST

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It belongs to the ensemble learning methods, which combine multiple individual models to improve predictive performance. Here's a brief documentation for Random Forest:

Random Forest Algorithm:

Overview:

Random Forest is a supervised learning algorithm that constructs a multitude of decision trees during training and outputs the mode (for classification) or the average prediction (for regression) of the individual trees.

Key Features:

1.Ensemble Method:

Random Forest builds multiple decision trees and combines their predictions to produce a final output.

2. Bootstrap Aggregating (Bagging):

Each tree in the Random Forest is trained on a bootstrap sample (a random sample with replacement) of the training data.

3. Random Feature Selection:

At each node of the decision tree, a random subset of features is considered for splitting, rather than considering all features. This helps in decorrelating the trees and reducing overfitting.

4. Voting (Classification) / Averaging (Regression):

For classification tasks, the mode of the predictions from individual trees is taken as the final output. For regression tasks, the average of the predictions is used.

5. Parallel Training:

Random Forest can be trained in parallel since each tree is independent of the others.

Parameters:

1. Number of Trees (n_estimators):

The number of decision trees to be included in the forest.

2. Maximum Depth (max_depth):

The maximum depth of each decision tree in the forest.

3. Minimum Samples Split (min_samples_split):

The minimum number of samples required to split an internal node.

4. Minimum Samples Leaf (min_samples_leaf):

The minimum number of samples required to be at a leaf node.

5. Maximum Features (max_features):

The number of features to consider when looking for the best split at each node.

6. Bootstrap Sample (bootstrap):

Whether bootstrap samples are used when building trees.


Fig 1.2 Random Forest

## 1.4 WORKING OF RANDOM FOREST

1. Bootstrap Sampling:

Random Forest begins by creating multiple decision trees, each trained on a different subset of the original dataset. These subsets are created using a technique called bootstrap sampling, where random samples of the original dataset are drawn with replacement.

2. Decision Tree Construction:

For each tree in the Random Forest, a decision tree is constructed using the bootstrap sample. At each node of the tree, a random subset of features (selected typically using the square root or logarithm of the total number of features) is considered for splitting. This random feature selection helps in reducing the correlation between trees and prevents overfitting.

3. Tree Training:

Each decision tree is grown to its maximum depth or until the number of samples in each leaf node falls below a specified threshold (controlled by parameters like `max_depth`, `min_samples_split`, and `min_samples_leaf`).

4. Voting (Classification) / Averaging (Regression):

Once all trees are constructed, predictions are made for new data points. For classification tasks, the mode (most frequent class) of the predictions from individual trees is taken as the final output. For regression tasks, the average of the predictions is used.

5. Importance of Trees:

The importance of each feature in the Random Forest can be calculated based on how much the tree nodes that use that feature reduce impurity (e.g., Gini impurity for classification or mean squared error for regression) on average across all trees. This provides insights into which features are most informative for making predictions.

6. Parallel Training:

Random Forest can be trained in parallel, as each tree is independent of the others. This makes it scalable and efficient, especially for large datasets.

7. Hyperparameter Tuning:

Random Forest has several hyperparameters that can be tuned to optimize performance, such as the number of trees (`n_estimators`), maximum depth of trees (`max_depth`), minimum samples required to split an internal node (`min_samples_split`), and others.

8. Bagging and Aggregation:

The ensemble technique used in Random Forest, known as bagging (Bootstrap Aggregating), ensures that the final prediction is robust by aggregating the predictions of multiple trees, reducing the variance of the model.

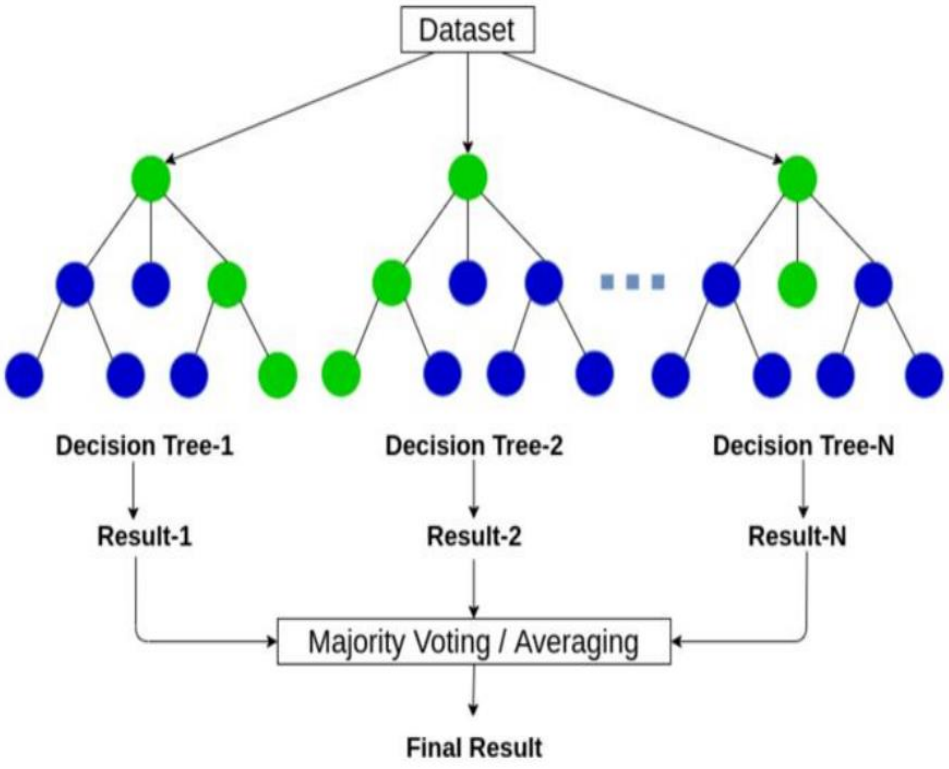


Fig 1.3 working of random forest

# CHAPTER 2

# LITERATURE REVIEW

Cardiovascular Disease is the primary justification behind death inside the world throughout the past ten years. Almost one individual passes on from heart condition concerning every moment inside the U. S. alone. To scale back the quantities of passings from heart sicknesses there should be a quick and prudent discovery strategy double-dealing information handling. By examining the exploratory outcomes, it's finished that the J48 tree procedure clads to be best classifier for heart condition forecast because of it contains a ton of precision and the most un-allout chance to make. A Random Forest Machine Learning Algorithm is incorporated with the Flask Web structure for anticipating of Heart Disease. Artery Blockage demonstrates the presence of heart disease. The higher the blockage, higher is the phase of heart disease.The Data expected for the prediction contains parameters, for example, Age, Sex, Blood Pressure, Sugar levels. Trial results say that predictions by utilizing the Random Forest Machine Learning Algorithm which is integrated with the Flask Web system is reliably better than those acquired utilizing different strategies. For the accurate detection of the coronary illness, a proficient ML strategy ought to be utilized which had been gotten from a distinctive examination among several machine learning algorithms in a Java Based Open Access Data Mining Platform, WEKA. to screen the heart disease patient nonstop by his/her caretaker/doctor, a constant patient observing framework was created and introduced by utilizing Arduino. The probabilities of heart condition and classification of patient's risk level by implementing different data processing techniques like Naive Bayes, Decision Tree, Logistic Regression and Random Forest was performed by making use of heart condition dataset available in UCI machine learning repository.

Senthilkumar Mohan et al [4], has used a hybrid machine learning algorithm to predict heart disease.The dataset used is the Cleveland dataset. Columns of age and sex from the database are not used as the authors assume that the information is personal and does not affect the prediction. They developed their own Hybrid Random Forest Linear Method (HRFLM), a combination of both Random Forest (RF) and Linear method (LM) algorithms. The authors move on to suggest that they have implemented a set of four algorithms. They concluded that the combination of these two algorithms produced better results than individual classifiers. Authors promotes continuous improvement in accuracy by using a combination of various machine learning algorithms.

Nagaraj M Lutimath, et al [5], worked on a prediction model for heart disease using the Naïve bayes classification and Support Vector Machine.The performance measurements used in the analysis are Mean Absolute Error, Root Mean Squared Error and Sum of Squared Error [1]. They found that SVM emerged as a much better performer in terms of accuracy than Naive Bayes algorithm.

They analysed the algorithms of Random Forest, Decision Tree, Naive Bayes and Logistic Regression classifiers on the basis of accuracy, precision, recall and f score. Then identified the best classification algorithm that can be used in predicting heart disease

Veshvendra K Singh, et al [6], worked with various machine learning algorithms such as Random Forest, SVM(Support Vector Machine), Linear Regression, Logistic Regression, Decision Tree with 3, 5 and 10 cross validation techniques. The authors used different splits, different tree numbers for each reference and a different number of cross validation fold. In the Random Forest, 85.81% accuracy is achieved by 20 splits, 75 trees and 10 folds.

Fahd Saleh Alotaibi [7] worked on proposing a ML model that compared five different algorithms. Rapid Miner tool has been used which has resulted in higher accuracy compared to the Matlab and Weka tool. In the study, accuracy of the Naïve Bayes, Decision Tree, Logistic Regression, Random Forest and Support Vector Machine classifiers are made to compare. The Decision Tree algorithm showed the highest accuracy on the Rapid Miner tool.

Tsien et al [3] in their study indicated that classification trees, which have certain advantage over logistic regression models, with patients having Myocardial Infarction (MI). The results shown that the occurrence of MI has been noticed in male than the female. Age, Systolic blood pressure, smoking has been found to be the important risk factor in the patients with MI.

Rea et al [4] concluded that smoking has been associated with an elevated risk for recurrent coronary events such as Angina,Acute Myocardial Infarction (AMI). In some cases, smoking has been associated with cholesterol for AMI. The subjects can be extended with various other events related to CHD.Since the risk factors have different degrees of impact, the population-specific risk function is needed for the prediction of CHD.

Karaolis et al [6] developed a data mining system for the assessment of heart related risk factors using association analysis based on Apriori algorithm. The results with 369 cases shown that smoking is one of the main risk factor that directly affect the coronary heart disease for all the events.

Kunc et al [7] presented simulation results which can be used for evaluation of patients with coronary heart disease, congestive heart failure, end-stage renal disease in Slovenia. At the same time also year treatment costs were calculated regarding each of observed diseases. The presented results

enable the estimation of potential savings resulting from more intensive chronic diseases treatment.

Srinivas et al [8] focused on using different algorithms for predicting combinations of several target attributes and presented automated and effective heart attack prediction methods using data mining techniques such as Decision trees, Neural networks and Bayesian models. Firstly, they provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weightage, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals are defined based on data exploration. For predicting heart attack significantly 15 attributes have been chosen. The future work signifies the usage of other attributes such as financial status, stress, pollution and previous medical history. Other data mining techniques, Time Series, Clustering and Association Rules can also be used to analyze patients‟ behavior.

Soni et al [9] provided a survey of current techniques in Data mining for heart disease prediction. Experiments has been conducted with various sorts of techniques using the same dataset out of which Decision tree shown high accuracy than that of the Bayesian classification, KNN, neural networks.The accuracy has been further improved by applying genetic algorithm with Decision trees. The work can be extended by using real dataset from health care organizations for the automation of Heart Disease prediction.

Rafiah et al [10] using Decision Trees, Naive Bayes, and Neural Network techniques developed a system for heart disease prediction using the Cleveland Heart disease database and shown that Naïve Bayes performs well followed by Neural Network and Decision Trees. The relationship between attributes produced by Neural Network is more difficult to understand than that of the other models used to predict heartdisease. Continuous data can be used instead of categorical data and text mining methods can be incorporated to mine vast amount of unstructured data available in healthcare databases.

Karl berg and Elo [11] calculated the burden of Ischemic Heart Disease (IHD) and coronary risk factors in a defined population using data from all public providers of healthcare. Calculation of the actual burden of disease in the population showed that when hospital discharge data were combined with the outpatient data, there were no or slight difference in the

age-specific rates of Acute Myocardial Infarction (AMI), while the rates of angina were between two-fold and four-fold higher, and unspecified IHD was between three-fold and tenfold higher in individuals aged greater than 50 years compared with using hospital discharge data alone. These findings suggest that hospital discharge data should be combined with outpatient care data to provide a more comprehensive estimate of the burden of IHD and its riskfactors.Meanwhile, this paper deals with the investigation of various events with their impacts and its risk factors associated with CHD and to improve the overall prediction accuracy using the Random forest classification algorithm.

# CHAPTER 3
# EXISTING METHOD

## 3.1 EXISTING METHOD

Random forest algorithm is one of the most effective ensemble classification approach. The Random forest algorithm has been used in prediction and probability estimation . Random forest consists of many decision trees .Each decision tree gives a vote that indicate the decision about class of the object. Random forest item was first proposed by Tin kam HO of bell labs in 1995. Random forest method combines bagging and random selection of features. There are three important tuning parameters in random forest1) No. of trees (n tree) 2) Minimum node size 3) No. of features employed in splitting each node 4) No. of features employed in splitting each node for each tree (m try).

Random forest algorithm advantages are listed below.

1) Random forest algorithm is accurate ensemble learning algorithm.

2) Random forest runs efficiently for large data sets.

3) It can handle hundreds of input variables.

4) Random forest estimates which variables are important in classification.

5) It can handle missing data.

6) Random forest has methods for balancing error for class unbalanced data sets.

7) Generated forests in this method can be saved for future reference .

8) Random forest overcomes the problem over fitting.

9) In training data, Random forest is less sensitive to outlier.

10) In Random forest, parameters can be set easily and eliminates the need for tree pruning.

11) In Random forest accuracy and variable importance is automatically generated .

When constructing individual trees in random forest, randomization is applied to select the best node to split on. This value is equal to $\sqrt{A}$, where A is no. of attributes in the data set . However Random forest will generate many noisy trees, which affect classification accuracy and wrong decision for new sample.

## 3.2 DISADVANTAGES OF EXISTING SYSTEM:

- Existing approaches are low accuracy and high computation time and these might be due the use of irrelevant features in dataset. In order to tackle these problems new methods are needed to detect HD correctly. The improvement in prediction accuracy is a big challenge and research gap.
- Accuracy is very low.
- Computationally complex.
- More execution time required to generate results.

```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier(n_estimators=500,criterion='entropy',max_depth=8,min_samples_split=5)
model3 = rfc.fit(X_train, y_train)
prediction3 = model3.predict(X_test)
cm3=confusion_matrix(y_test, prediction3)
```

**Output:**

```
80.26
```

**Output:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.79 | 0.78 | 34 |
| 1 | 0.83 | 0.81 | 0.82 | 42 |
| accuracy | | | 0.80 | 76 |
| macro avg | 0.80 | 0.80 | 0.80 | 76 |
| weighted avg | 0.80 | 0.80 | 0.80 | 76 |

# CHAPTER 4

# PROPOSED METHOD

## 4.1 PROPOSED METHOD

The proposed technique uses random forest algorithm for prediction of heart disease. Feature subset selection is a process that selects a subset of original attributes and reduces feature space.

We applied, Random forest algorithm for the Cleveland heart disease dataset , this Cleveland heart disease dataset is taken from UCI data respository. In our proposed work ,we used Parameter Tuning to select attributes and keep only attributes which contribute more towards the diagnosis of heart disease.

### Dataset description

14-Attributes

303-Rows

15-Columns

### Columns description

age : age in years

sex : (1 = male; 0 = female)

cp : chest pain type

trestbps : resting blood pressure (in mm Hg on admission to the hospital)

chol : serum cholestoral in mg/dl

fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg : resting electrocardiographic results

thalach : maximum heart rate achieved

exang : exercise induced angina (1 = yes; 0 = no)

oldpeak : ST depression induced by exercise relative to rest

slope : the slope of the peak exercise ST segment

ca : number of major vessels (0-3) colored by flourosopy

thal : 3 = normal; 6 = fixed defect; 7 = reversable defect

target : refers to the presence of heart disease in the patient (1=yes, 0=no)

## Splitting the Dataset into the Training set and Test set:

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

**Training Set:** A subset of dataset to train the machine learning model, and we already know the output.

**Test set:** A subset of dataset to test the machine learning model, and by using the test sets model predicts the output.

## 4.2 Random Forest Algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

## Important Features of Random Forest :

- Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.

- Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.

- Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

- Train-Test split- In a random forest we don’t have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

- Stability- Stability arises because the result is based on majority voting/ averaging.

## 4.3 Assumptions for Random Forest:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

## Types of Ensembles :

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

**Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

**Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

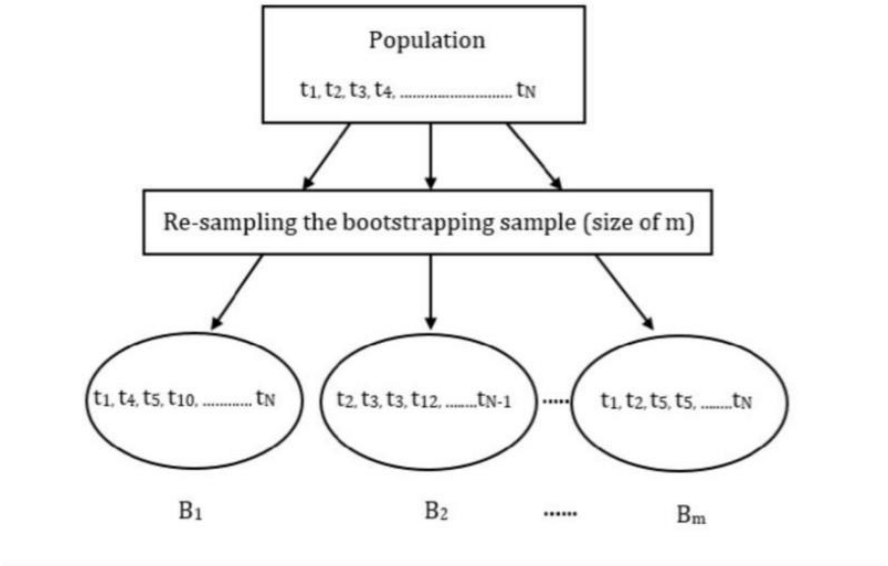


Fig 4.1 bootstrapping sample

## Hyperparameter Tuning:

Hyperparameter tuning relies more on experimental results than theory, and thus the best method to determine the optimal settings is to try many different combinations evaluate the performance of each model.

## Hyperparameters:

Parameter tuning is crucial for optimizing the performance of a Random Forest model. Here are some key parameters to consider tuning in a Random Forest classifier:

**1.n_estimators**: number of trees in the forest. Increasing the number of trees generally improves performance, but it also increases computational costs. You can try different values and choose the one that balances performance and computational efficiency.
Range: between 50 and 200

**2.max_depth**: Maximum depth of the trees. Deeper trees may capture more complex patterns, but they are also more prone to overfitting. Cross-validation can help you find an optimal value.
Range: between 5 and 30

**3.min_samples_split**: Minimum number of samples required to split an internal node. Increasing this value can lead to a more robust model against noise but may result in underfitting.
Range: between 2 and 20

**4.min_samples_leaf**: Minimum number of samples required to be at a leaf node. Increasing this value can smooth the model and reduce overfitting.
Range: between 1 and 20

**5.max_features**: The number of features to consider when looking for the best split. You can experiment with different values such as 'sqrt', 'log2', or an integer.
**6.criterion**: The function used to measure the quality of a split. 'Gini' and 'entropy' are common choices. Try both and see which one works better for your data.

**7.bootstrap**: Whether to use bootstrapped samples when building trees. Turning it off (**bootstrap=False**) may lead to a more diverse set of trees.

**8.random_state**: Set a seed for reproducibility.
Range: Any integer for reproducibility.

## 4.4 Advantages of proposed system:

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

Criterion – gini

We usually use the gini index since it is computationally efficient ,it takes a shorter period of time for execution because there is no logarithmic term like there is in entropy here. Usually, if you want to do logarithmic calculations it takes some amount of time.So we use the Gini index as their parameter.

```
# Train the model with the best hyperparameters on the entire training set
best_rf_model = RandomForestClassifier( criterion='gini',n_jobs=-1,
    min_samples_leaf=1,
    min_samples_split=2,
    n_estimators=100,
    random_state=123)
```

```
Sensitivity (Recall): 0.8947368421052632
Specificity: 0.9142857142857143
Accuracy: 90.74%
```

# CHAPTER 5

# RESULT ANALYSIS AND DISCUSSION

## 5.1 PERFORMANCE EVALUATION METRICS OF HEART DISEASE PREDICTION

Machine learning plays a key role in case of classifying and predicting the information. Accuracy is the most popular parameter for classifier algorithms, but accuracy only not able to judge the performance of the model. In addition to the accuracy, other evaluation metrics like, precision, Recall, specificity, False Positive Rate, and F-measure are used.

Generally, a binary classifier is used to classify the information into two types positive and negative. Positive indicates when correct classification or prediction has been made, whereas negative indicates the objective not belonging to a particular instance. Based on these two binary patterns, again, information is represented with four values which are termed as

- True Positive (TP)
- False Positive (FP)
- True Negative (TN)
- False Negative (FN)

This kind of plot representation against classification or prediction is termed a confusion matrix.

In the representation of the confusion matrix, two kinds of values are used: actual and predicted values. Further, these two values are classified based on binary classification representation.

Case 1: If the given predicted positive value matches with an actual positive value, then it is referred to as TP.

Case 2: If the given predicted positive value matches the actual negative value, it is referred to as FP.

Case 3: If the given predicted negative value matches with an actual positive value, then it is referred to as FN.

Case 4: If the given predicted negative value matched with the actual negative value, then it is referred to as TN.

Based on the above four parameters, the following parameters are derived.

- Accuracy
- Precision
- Recall
- FPR (False positive rate)

- F-Measure or F1-Score or F-Score

### *5.1.1 Accuracy*

This is the base metric for any model of evaluation, which is defined as the number of correct predictions divided by overall predictions. Accuracy equal to 1.0 is considered best, whereas 0.0 is considered worst.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ size\ of\ the\ dataset}$$

### *5.1.2 Precision*

Precision is used to estimate the number of correctly predicted positive instances. Further, which is defined as

$$Precision = \frac{No\ of\ correctly\ predicted\ positive\ instances}{Total\ No\ of\ positive\ predictions\ observed}$$

Precision value 1.0 is considered to be as best precision, whereas 0.0 is the worst

### 5.1.3 Recall

Recall or Sensitivity represents the same measure. It is evaluated by using the number of correctly predicted positive instances divided by the number of total positive instances in the dataset. Recall value 1.0 is considered to be an as best recall, whereas 0.0 is the worst.

$$\text{Recall} = \frac{\text{No of correctly predicted positive instances}}{\text{Total no of positive instances observed}}$$

### 5.1.4 False Positive Rate (FPR)

False positive rate (FPR) is evaluated using the number of positive predictions incorrectly divided by the total negative instances. FPR ranges between 0.0 to 1.0. The best FPR is 0.0, whereas 1.0 is the worst.

$$\boldsymbol{FPR = \frac{FP}{TN + FP}}$$

### *5.1.5 F-Measure or F1-Score*

*F-Measure or F1-Score* is described as a harmonic mean in between precision and recall. The following formula is used to calculate F1-Score. The low value of the F-Measure represents the best performance.

$$f - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

## Output:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.91      0.93        35
           1       0.85      0.89      0.87        19

    accuracy                           0.91        54
   macro avg       0.90      0.90      0.90        54
weighted avg       0.91      0.91      0.91        54
```

```
Confusion Matrix:
[[32  3]
 [ 2 17]]
```

```
Sensitivity (Recall): 0.8947368421052632
Specificity: 0.9142857142857143
Accuracy: 90.74%
```

# CHAPTER 6

# CONCLUSION & FUTURE WORK

## 6.1 CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal.

Our proposed approach achieved an accuracy of 90.74% for the Cleveland Heart Disease dataset. Applying Random forest has shown improved accuracy in prediction of heart disease.

This model purposes a dataset which have 14 attributes directed on various people. The dataset was split for training and testing.Therefore, the model was trained using random forest algorithm. Flask web application framework was utilised to create the web application where the user can enter the details. Based on the user inputs the result will be displayed. If the person is predicted to have heart disease, the alert message will be sent.

## 6.2 FUTURE WORK

The future course of this work can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature-selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

# CHAPTER 7

# BIBLIOGRAPHY

# BIBLIOGRAPHY

**[1]** A. A., M. S., D. R. D. P. G. Apurb Rajdhan, "Heart Disease Prediction using Machine Learning," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), vol. 09, no. 4, pp. -, 2020.

**[2]** D. a. G. C. Dua, "UCI Machine Learning Repository: Statlog (Heart) Data Set," University of California, Irvine, School of Information and Computer Sciences, - - 2017.

**[3]** D. Varghese, "Comparative Study on Classic Machine learning Algorithms," Towardsdatascience.com, 6 December 2018. [Online]. Available: https://towardsdatascience.com/comparative study-on-classic-machine-learning-algorithms 24f9ff6ab222. [Accessed - May 2022].

**[4]** C. T., G. S. Senthilkumar Mohan, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access 7, pp. 81542-81554, 2019.

**[5]** C. C. B. P. Nagaraj Lutimath, "Prediction of Heart Disease using Machine Learning," International Journal of Recent Technology and Engineering, vol. 8, no. 2S10, pp. 474-477, 2019.

**[6]** N. S. S. K. S. Yeshvendra K Singh, "Heart Disease Prediction System Using Random Forest," Advances in Computing and Data Sciences, vol. 721, no. -, pp. 613-623, 2017.

**[7]** F. S. Alotaibi, "Implementation of Machine Learning to Predict Heart Failure Disease," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 6, p., 2019.

**[8]** A. S. Abdullah and R. R. Rajalaxmi, ''A data mining model for predicting the coronary

heart disease using random forest classifier,'' in Proc. Int. Conf. Recent Trends Comput. Methods, Controls, Apr. 2012

**[9]** N. Al-milli, ''Backpropogation neural network for prediction of heart disease,'' J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

**[10]** C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ''Analysis of neural networks based heart disease prediction system,'' in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233 239.

**[11]** P. K. Anooj, ''Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,'' J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 10.1016/j.jksuci.2011.09.002.

**[12]** L. Baccour, ''Amended fused TOPSIS VIKOR for classification (ATOVIC) applied to some UCI data sets,'' Expert Syst. Appl., vol. 99, pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

**[13]** R. Das, I. Turkoglu, and A. Sengur, ''Effective diagnosis of heart disease through neural networks ensembles,'' Expert Syst. Appl., vol. 36, no. 4, pp. 7675 7680, May 2009.

**[14]** A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ''Prediction of heart disease using machine learning,'' in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018.

**[15]** B. S. S. Rathnayaka and G. U. Ganegoda, ''Heart diseases prediction with data mining and neural network techniques,'' in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018.

**[16]** N. K. S. Banu and S. Swamy, ‘‘Prediction of heart disease at early stage using data mining and big data analytics: A survey,’’ in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016.

**[17]** P.Shrama,k.saxena,“heart disease prediction system evaluation using c4.5 rules and partial trees “AISC,Springer,pp 285-294(2016)

**[18]** Hlaudi Daniel Masethe,“Prediction of heart disease using classification algorithms“, vol 11,pp1-4,WCECS2014

**[19]** Sheik abdullah,RR Rajalakshmi,“A data mining model for predicting the coronary heart disease using random forest classifier“,IJCA,PP 22-25(2012)

**[20]** Kemal polat, S.Gunes, S.Tosun, ” Diagnosis of heart disease using artificial immune recognition system and fuzzy weight preprocessing”, pattern recognition, 39,.

**[21]** Resul das ,Turkoglu,A Sengur,” Effective diagnosis of heart disease through network ensembles”, Expert System with Applications36,pp7675-7680(2009)

**[22]** PK Anooj,” Clinical decision support system: Risk level prediction of heart disease using Weighted fuzzy rules”, Journal of king saud university, CIS, 24, PP 27-40(2012)

**[23]** Detrano ,Janosi,W Stein burn,et.al,” International application of new probability algorithm for the diagnosis of CAD”. The American Journal of Cardiology, pp 304-310,64(5),(1989)

**[24]** A. Ramcharan, K. ] Mai Shouman, Turner, Stocker,” Using decision tree for diagnosing heart disease patients”, In 9th Australian data mining conference, Australia vol 121,ACM(2011)

**[25]** Tu et.al,” Effective diagnosis of heart disease through bagging approach” Biomedical Engineering and approach, pp 1-4, BMEI2009, IEEE (2009)

**[26]** Alaa Elsayad,Mahmoud Fakhr,"Diagnosis of cardiovasular diseases with bayesian classifier",Journal of Computer Science,vol 11(2),pp274-282(2015).

**[27]** M.A.Jabbar,B L Deekshatulu,Priti chandra,"heart disease classification using nearest neighbor classifier with feature subset selection"annals computer science series , 11th tome,1st fasc,pp 47-54(2013)

**[28]** ]Randa El-Bay,"Feature analysis of coronary artery heart disease data sets",Procedia Computer science,Elsevier,vol 65,pp 459-469(2015).

**[29]** ]M.A.Jabbar,B L Deekshatulu,Priti chandra,"Heart disease prediction system using associative classification and genetic algorithm",ICECIT 2012,VOL 1,PP 183-192(2012)

**[30]** Saaol times, Monthly magazine" Modifiable risk factors of heart disease", pp 6-10, July (2015)

**[31]** Khan MG,"Heart disease diagnosis and therapy", a practical approach,2nd Edition Springer,pp544(2015)

**[32]** M.A.Jabbar,B L Deekshatulu,Priti chandra ,"classification of heart disease using artificial neural network and feature subset selection",GJCST,Vol13, issue 3,2013

**[33]** Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel, "Heart Disease Prediction Using Machine Learning and Data Mining Technique", In International Journal of Computer Science & Communication (IJCSC), Volume 7, Number 1 Sept 2015- March 2016 Page No.129 – 137. Shi, D.-P.,Wu, C. "The influence of infrared temperature measurementbased on reflection temperature compensation and incident temperature compensation" Electron. Measur. Technol. 08, 2321–2326 (2015).

**[34]** Akram Ahmed Mohammed, Rajkumar Basa, Anirudh Kumar Kuchuru, Shiva Prasad Nandigama ,Maneeshwar Gangolla, "Random Forest Machine Learning technique to predict Heart disease" European Journal of Molecular & Clinical Medicine ISSN 2515-8260 Volume 7, Issue 4, 2020.

**[35]** Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System" World Journal of Engineering and Technology, 2018, 6, 854-873.

**[36]** RachitMisra , Pulkit Gupta , Prashuk Jain, "Rachit Misra1 , Pulkit Gupta2 , Prashuk Jain", July 2021| IJIRT | Volume 8 Issue 2 | ISSN: 2349-6002.

**[37]** Rahul Chaurasia, Saksham Gupta and Shipra Singh Siddhu, "Prediction Of Heart Disease Using Machine Learning Algorithm" 2018 IJCRT | Volume 6, Issue 2 April 2018.

**[38]** O.E. Taylor1, P. S. Ezekiel, F.B. Deedam-Okuchaba, "A Model to Detect Heart Disease using Machine Learning Algorithm" International Journal of Computer Sciences and Engineering Vol.-7, Issue-11, Nov 2019.

**[39]** Rahul Chaurasia, Saksham Gupta and Shipra Singh Siddhu, "Prediction Of Heart Disease Using Machine Learning Algorithm" 2018 IJCRT | Volume 6, Issue 2 April 2018.

**[40]** KillanaSowjanya, Dr. G. Krishna Mohan, "Predicting Heart Disease Using Machine Learning Classification Algorithms and Along With TPOT (AUTOML)" International Journal Of Scientific & Technology Research Volume 9, Issue 04, April 2020.

**[41]** Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6.

**[42]** M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies 10.1109/ICICT50816.2021.9358597.

**[43]** 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.

**[44]** Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y.

**[45]** C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020, doi: 10.1109/ACCESS.2020.2981159.

**[46]** Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, Future Generation Computer Systems, Volume 111, 2020, Pages 714-

722, ISSN 0167-739X, https://doi.org/10.1016/j.future.2019.09.056.

**[47]** KillanaSowjanya, Dr. G. Krishna Mohan, "Predicting Heart Disease Using Machine Learning Classification Algorithms and Along With TPOT (AUTOML)" International Journal Of Scientific & Technology Research Volume 9, Issue 04, April 2020.

**[48]** Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). "Heart disease prediction using data mining techniques". In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE.