Heart Disease Detection using Machine Learning

A Project Report

Submitted in the partial fulfilment of the requirements for the

award of the degree of

Bachelor of Technology

in

Department of Electronics and Communication Engineering

By

Ashwitha Reddy(2010040043)
Sai Nikitha (2010040082)
Sahitya Gouri M S (2010040094)
Rakshitha V(2010040123)

Under the supervision of

Dr. Ravi Boda

Assistant Professor

Department of ECE

Department of Electronics and Communication Engineering

K L University

Hyderabad,

Aziz Nagar, Moinabad Road, Hyderabad – 500075, Telangana, India

May 2022

DECLARATION

The Project Report entitled "Heart Disease Detection using Machine Learning" isa record of bonafide work of Ashwitha Reddy (2010040043), Sai Nikitha (2010040082), Sahitya Gouri M S (2010040094) and Rakshitha V (2010040123) submitted in partial fulfilment for the award of B.Tech degree in the Department of Electronics and Communication Engineering K L University, Hyderabad. The results embodied in this report have not been copied from any other departments/universities/institutes.

Ashwitha Reddy (2010040043)

Sai Nikitha (2010040082)

Sahitya Gouri M S (2010040094)

Rakshitha V (20100400123)

Certificate

CERTIFICATE This is to certify that the Project Report entitled “Heart Disease Detection using Machine Learning” submitted by Ashwitha Reddy (2010040043), Sai Nikitha (2010040082), Sahitya Gouri M S (2010040094) and Rakshitha V (2010040123) in partial fulfilment for the award of B.Tech degree in Department of Electronics and Communication Engineering K LUniversity, Hyderabad is a record of bonafide work carried out under our guidance and supervision. The results embodied in this report have not been copied from any other departments/universities/institutes.

Signature of theSupervisor

Dr. Ravi Boda

(Assistant Professor)

Signature of the HOD

Signature of the External Examiner

ACKNOWLEDGEMENT

First and foremost, we thank the lord almighty for all his grace & mercy showered upon us For completing this project successfully.

We take grateful opportunity to thank our beloved Founder and Chairman who has given constant encouragement during our course and motivated us to do this project.We are grateful to our Principal **Dr. Ramakrishna** who has been constantly bearing the torch for all the curricular activities undertaken by us.

We pay our grateful acknowledgment & sincere thanks to our Head of the Department **Dr. M. Goutham,** Associate Professor & Head of the Department, Department of ECE, K L University Hyderabad**.**for his exemplary guidance, monitoring and constant encouragement throughout the course of the project. We pay our sincere thanks to our Project Coordinator **Dr. Ravi Boda** who has been constantly inculcating logistic support and has given constant encouragement during our project. We thank Dr. Ravi Boda of our Department who has supported ted throughout this project holding a position of supervisor.

We wholeheartedly thank all the teaching and non-teaching staff of our department without whom we wouldn't have made this project a reality. We would like to extend our sincere thanks especially to our parents, our family members and friends who have supported us to make this project a grand success.

1. ABSTRACT

In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuition to doctors, who can then modify their diagnosis and approach per patient. The heart plays a significant role in living organisms.

Diagnosis and prediction of heart related diseases requires more precision, perfection, and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. To deal with the problem there is an essential need for a prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence(AI), it provides prestigious support in predicting any kind of event which take training from natural events We are attempting to use machine learning algorithms to predict potential heart conditions in humans. In this project, we compare the performance of various classifiers, including Extra Tree Classifier, SVC, KNN and Random Forest Classifier.

We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples. In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuition to doctors, who can then modify their diagnosis and approach per patient. We are attempting to use machine learning algorithms to predict potential heart conditions in humans. In this project, we compare the performance of various classifiers, Including Extra Three classifier, SVC, KNN and Random Forest Classifier. We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples.

2. Contents

List of Figures

3. INTRODUCTION

1.1. Machine Learning

Machine Learning is a branch of the broader field of artificial intelligence that makes use of statistical models to develop predictions. It is often described as a form of predictive modellingor predictive analytics and traditionally, has been defined as the ability of a computer to learn without explicitly being programmed to do so.

In basic technical terms, machine learning uses algorithms that take empirical or historical data in, analyze it, and generate outputs based on that analysis. In some approaches, the algorithms work with so-called "training data" first and then they learn, predict, and find waysto improve their performance over time.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed.ML is one of the most exciting technologies that one would haveever come across. As it is evident from the name,it gives the computer that makes it more similar to humans: The ability to learn.Machine learning is actively being used today, perhapsin many more places than one would expect

How does machine learning works

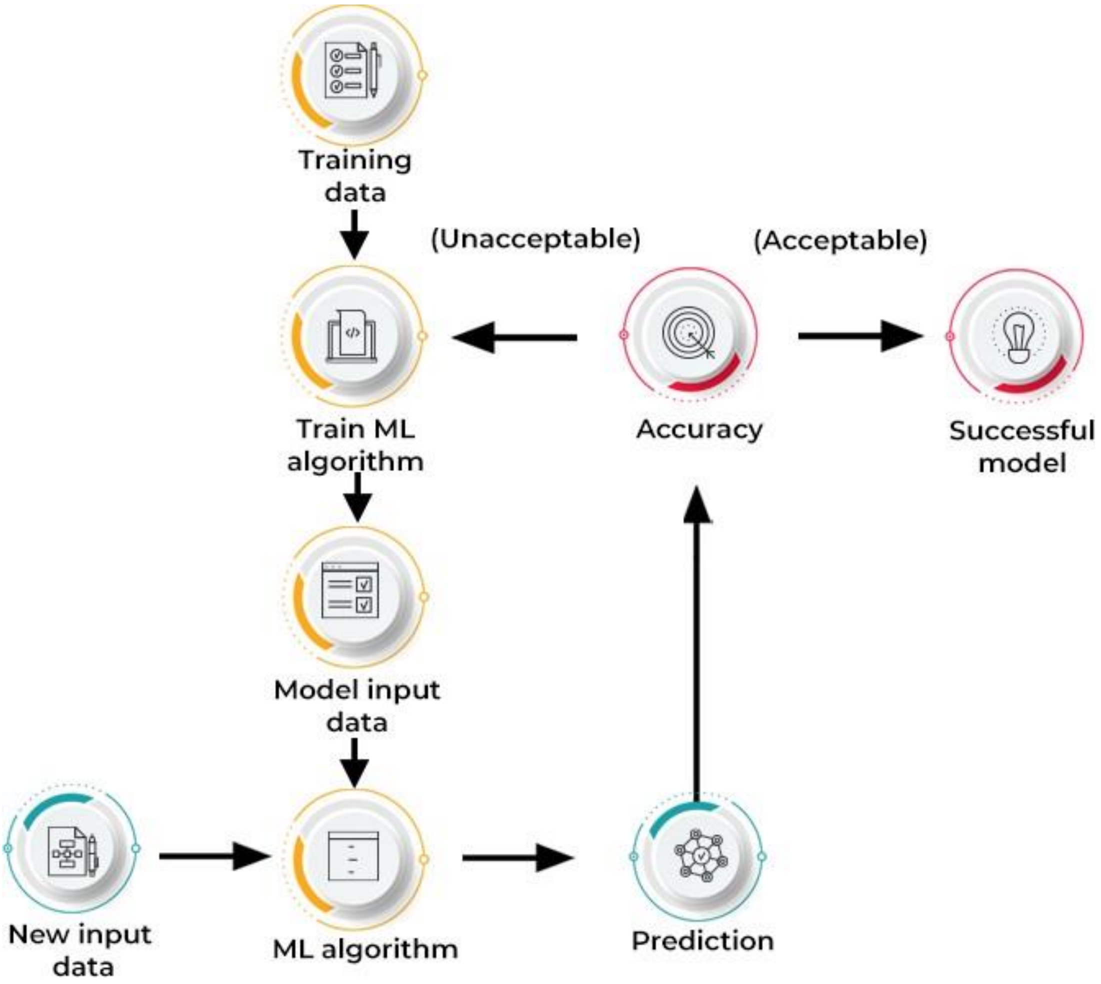


Figure 1 Machine Leaning

Type of Machine Learning

There are three main approaches to machine learning: supervised, unsupervised, and reinforcement learning. There are also hybrid approaches including semi-supervised learning, which can be tailored to the problem a researcher is seeking to solve. Each approach has specific strengths and weaknesses, and some techniques are better suited to particular types of problems than others.

Machine learning has a wide range of applications across various domains. Here are some of the notable applications of machine learning:

1. **Image and Video Analysis:**

- Image Classification: Identifying objects, scenes, or patterns in images.
- Object Detection: Locating and classifying objects within images or videos.
- Facial Recognition: Recognizing and verifying individuals' faces for security and authentication.
- Video Action Recognition: Recognizing and classifying actions or events in videos.

2. **Natural Language Processing (NLP):**

- Text Classification: Categorizing text data into predefined categories.
- Sentiment Analysis: Determining the sentiment (positive, negative, neutral) of text.
- Machine Translation: Automatically translating text from one language to another.
- Speech Recognition: Converting spoken language into written text.

3. **Recommendation Systems**:

- Content Recommendations: Suggesting products, movies, music, or articles to users based on their preferences and behavior.
- Collaborative Filtering: Recommending items by analyzing the behavior and preferences of

similar users.

4. **Anomaly Detection:**

- Identifying unusual patterns or outliers in data, which is valuable for fraud detection, network security, and quality control.

5. **Predictive Analytics:**

- Forecasting future trends and making predictions based on historical data, such as sales forecasting, demand forecasting, and stock price prediction.

6. **Healthcare:**

- Disease Diagnosis: Assisting in the diagnosis of diseases and medical conditions.
- Drug Discovery: Accelerating the process of discovering new drugs and compounds.
- Personalized Medicine: Tailoring treatments and medications to individual patients.

7. **Autonomous Systems:**

- Self-Driving Cars: Using machine learning for perception, decision-making, and control in autonomous vehicles.
- Drones and Robotics: Enabling robots and drones to navigate, recognize objects, and perform tasks autonomously.

8. **Financial Services:**

- Credit Scoring: Assessing the creditworthiness of individuals and businesses.
- Algorithmic Trading: Making investment decisions based on data analysis and predictive modeling.
- Fraud Detection: Identifying fraudulent transactions and activities.

9. **Manufacturing and Quality Control:**

- Quality Inspection: Inspecting and ensuring the quality of products on the production line.
- Predictive Maintenance: Anticipating equipment maintenance needs to minimize downtime.

10. **Gaming:**

- Creating intelligent and responsive non-player characters (NPCs) and opponents in video games.
- Generating game content and levels.

11. **Environmental Monitoring:**

- Analyzing climate data for weather forecasting and climate modeling.
- Monitoring and protecting wildlife through image and audio analysis.

12. **Social Media and Advertising**:

- Targeted advertising based on user behavior and preferences.
- Social network analysis for user engagement and content recommendation.

13. **Energy Management:**

- Optimizing energy consumption in buildings and industrial processes.

14. **Human Resources:**

- Automating the initial screening of job applicants.
- Predicting employee turnover and optimizing workforce management.

15. **Agriculture**:

- Precision farming, including crop yield prediction and pest detection.

These are just a few examples, and machine learning is continuously evolving and finding applications in new domains. Its ability to analyze vast amounts of data and make data-driven predictions and decisions makes it a powerful tool for solving a wide range of complex problems.

1.2 Supervised learning:

the computer is trained on a set of data inputs and outputs, with a goal of learning a general rule that maps the given inputs to the given outputs. Supervised learning is effective for a variety of business purposes, including sales forecasting, inventory optimization, and fraud detection.

Some examples of use cases include:

1.2.1.1 Text categorization

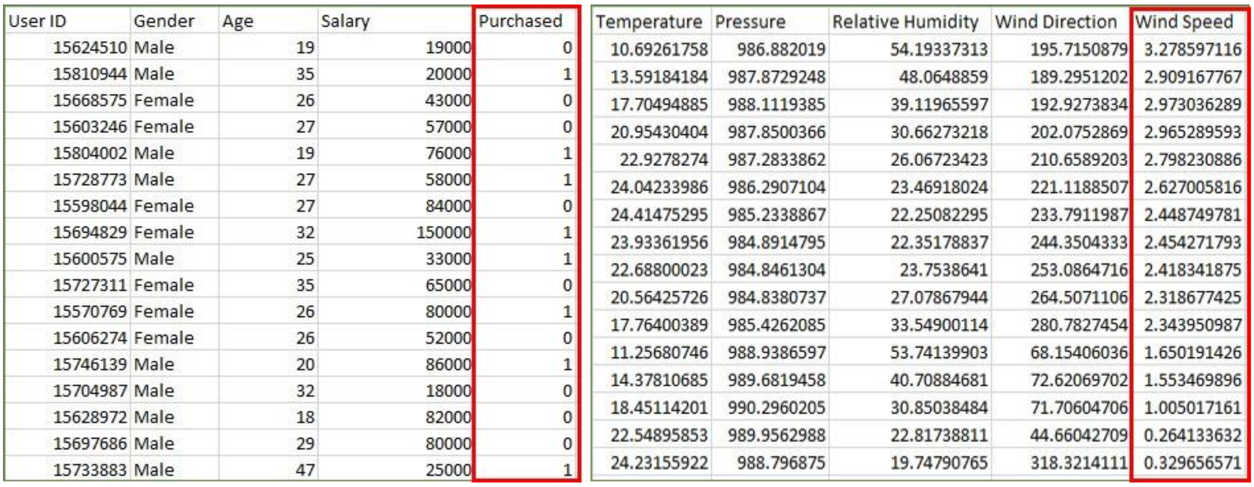
1.2.1.2 Face Detection

1.2.1.3 Signature recognition

1.2.1.4 Spam detection

1.2.1.5 Weather forecasting

1.2.1.6 Stock price predictions, among others



User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure 2 Classification & Regression

Both the above figures have labelled data set as follows:

1.2.1.6.1 **Figure A:** It is a dataset of a shopping store that is useful in predicting whether acustomer will purchase a particular product under consideration or not based on his/ her gender, age, and salary.

Input: Gender, Age, Salary

Output: Purchased i.e. 0 or 1; 1 means yes the customer will purchase and 0 means that the customer won't purchase it.

1.2.1.6.2 **Figure B:** It is a Meteorological dataset that serves the purpose of predictingwind speed based on different parameters.

Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction

Output: Wind Speed

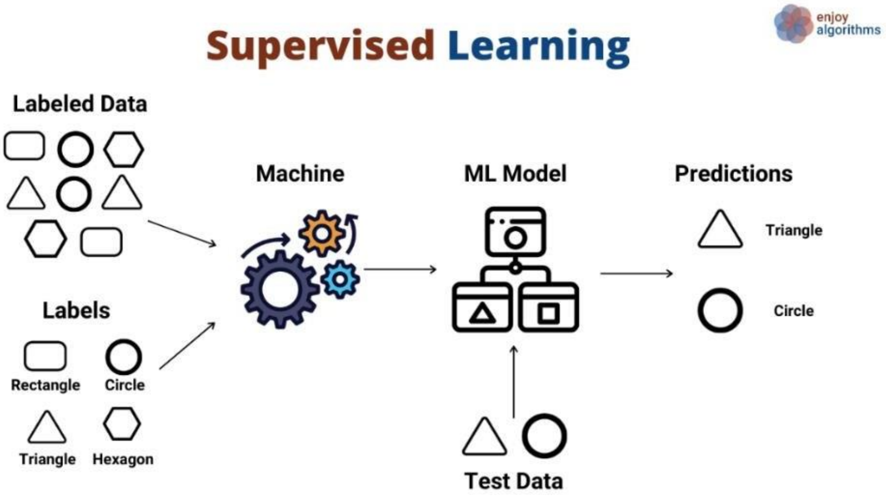


Figure 3 Supervised Leaning

Two main types of supervised learning are:

1.2.1.7 **Classification**, which entails the prediction of a class label, and

1.2.1.8 **Regression**, which entail the prediction of a numerical value.

1.3 Unsupervised learning:

The learning algorithm is not given this type of guidance; instead, it works to discover the pattern or structure in the input on its own. This is widely used to create predictive models. Common applications also include clustering, which creates a model that groups objects together based on specific properties, and association, which identifies the rules existing between the clusters.

A few example use cases include:

1.3.1.1 Creating customer groups based on purchase behavior

1.3.1.2 Grouping inventory according to sales and/or manufacturing metrics

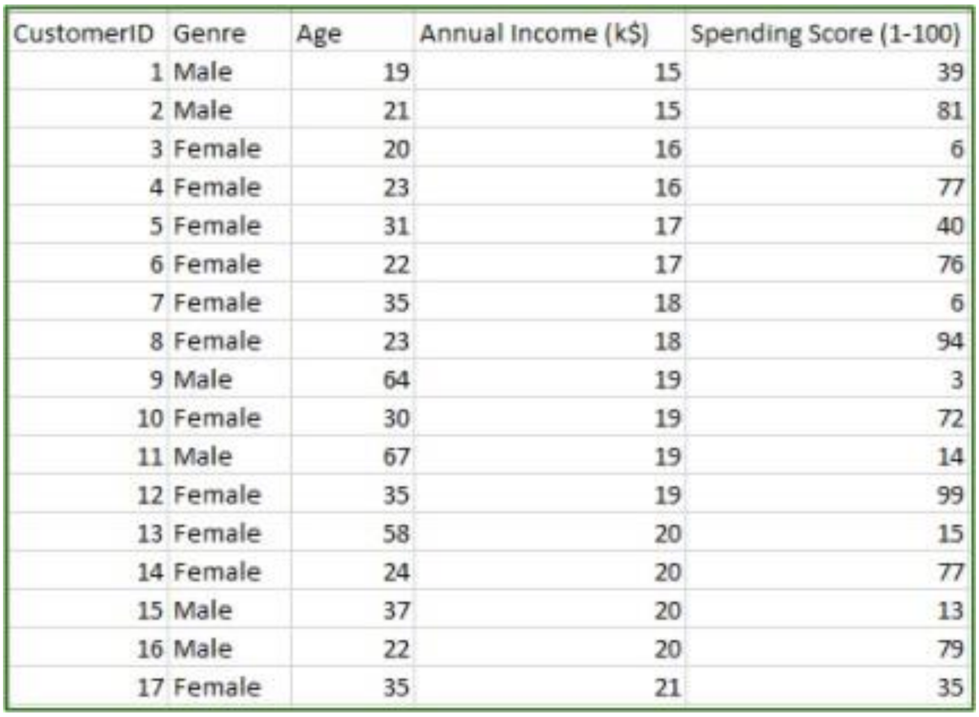
1.3.1.3 Pinpointing associations in customer data (for example, customers who buy a specific style of handbag might be interested in a specific style of shoe)

1.3.1.4 Fraud detection

1.3.1.5 Malware detection

1.3.1.6 Identification of human errors during data entry

1.3.1.7 Conducting accurate basket analysis, etc.



CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure 3 Unsupervised Leaning Model

The input to the unsupervised learning models is as follows:

1.3.1.7.1 **Unstructured data**: May contain noisy(meaningless) data, missing values, or unknown data

1.3.1.7.2 **Unlabeled data**: Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to the labeled one in the Supervised approach.

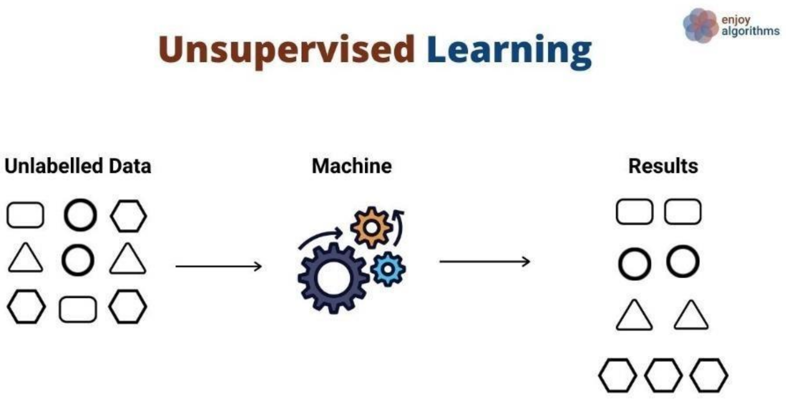


Figure 3 Unsupervised Leaning

Two main types of unsupervised learning are:

1.3.1.8 **Clustering**, which involves discovering groups within the dataset that share similar characteristics, and

1.3.1.9 **Density Estimation**, which involves evaluating the statistical distribution of the data set.

Unsupervised learning methods also include visualization with the data and projection, which reduces the dimensions of the data, a form of simplification.

1.4 Semi-supervised learning:

Semi-supervised learning is an important category that lies between the Supervised and Unsupervised machine learning. Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for the corporate purpose, it may have few labels.

The basic disadvantage of supervised learning is that it requires hand-labeling by ML specialists or data scientists, and it also requires a high cost to process. Further unsupervised learning also has a limited spectrum for its applications. To overcome these drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced. In this algorithm, training data is a combination of both labeled and unlabeled data. However, labeled data exists with a very small amount while it consists of a huge amount of unlabeled data. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labeled data. It is why label data is a comparatively, more expensive acquisition than unlabeled data.

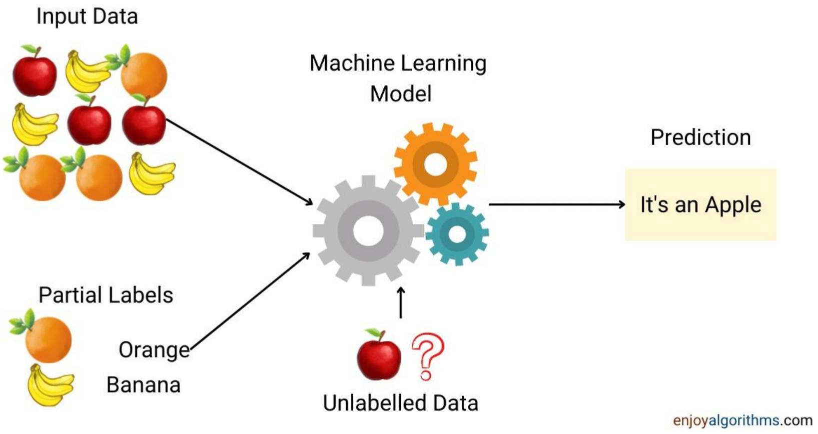


Figure 5 Semi Supervised Leaning

A few example use cases include:

1.3.1.10 **Speech Analysis-** It is the most classic example of semi-supervised learning applications. Since, labeling the audio data is the most impassable task that requires many human resources, this problem can be naturally overcome with the help of applying SSL in a Semi-supervised learning model.

1.3.1.11 **Web content classification-** However, this is very critical and impossible to label each page on the internet because it needs mode human intervention. Still, this problem can be reduced through Semi-Supervised learning algorithms. Further, Google also uses semi-supervised learning algorithms to rank a webpage for agiven query.

1.3.1.12 **Protein sequence classification-** DNA strands are larger, they require active human intervention. So, the rise of the Semi-supervised model has been proximate in this field.

1.3.1.13 **Text document classifier-** As we know, it would be very unfeasible to find a large amount of labeled text data, so semi-supervised learning is an ideal model to overcome this.

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is heart disease. One of the most crucial topics in the data analysis area is predicted cardiovascular disease. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Many studies have been carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk.

Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. Cardiovascular disease must be detected early. aiding high-risk patients in making decisions regarding lifestyle changes will help to reduce the difficulties. Making choices and predictions from the vast amounts of data generated by the healthcare sector is made easier with the help of machine learning. By evaluating patient data that uses a machine-learning algorithm to categorize whether a patient has heart disease or not, this study hopes to predict future cases of heart disease.

Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. By gathering information from numerous sources, organizing it into categories that make sense, and then performing analysis to get out the desired information based on statistics, we may conclude that this technique is quite adaptable.

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010). Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients' databases can be compared to real-life applications. Doctors' knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction.

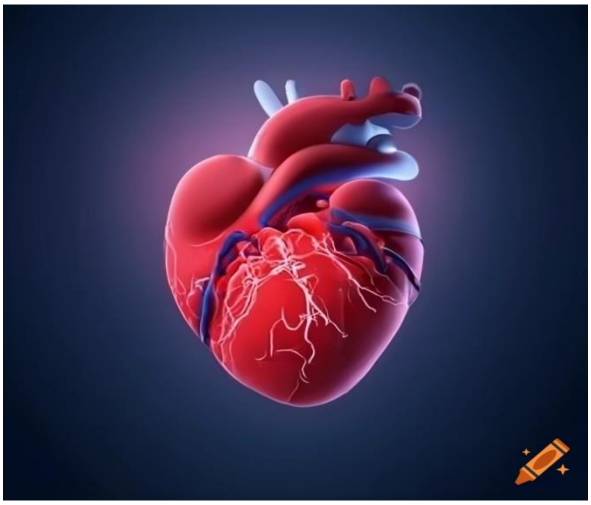


Fig 6 Heart Diagram

Therefore, it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the Diagnosis process. It also provides healthcare professionals with an extra source of knowledge for making decisions. The healthcare industry collects large amounts of health-care data and that needs to be mined to discover hidden information for effective decision making.

Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease (Helma, Gottmann et al. 2000). Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Lee, Liao et al. 2000).

2. Methodology

2.1 BLOCK DIAGRAM:

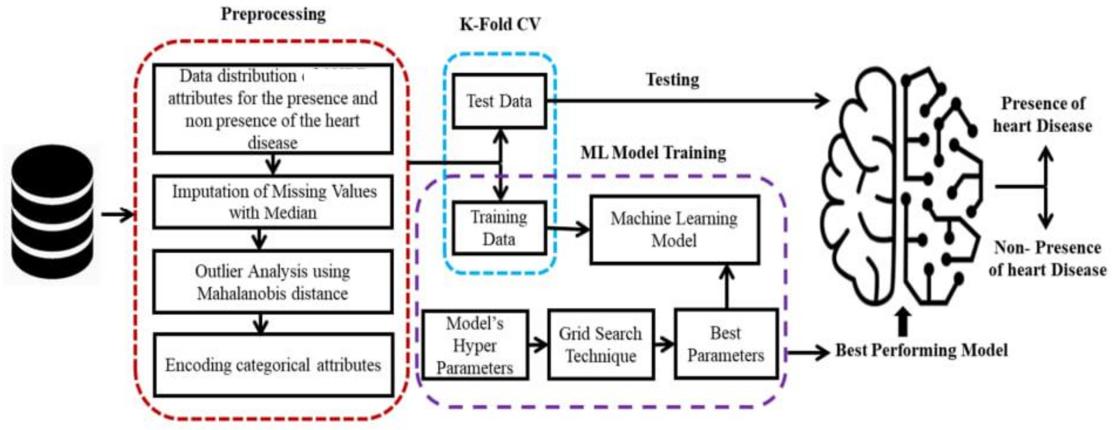

Preprocessing
Data distribution of attributes for the presence and non presence of the heart disease
Imputation of Missing Values with Median
Outlier Analysis using Mahalanobis distance
Encoding categorical attributes
K-Fold CV
Test Data
Training Data
Testing
ML Model Training
Machine Learning Model
Model's Hyper Parameters
Grid Search Technique
Best Parameters
Best Performing Model
Presence of heart Disease
Non- Presence of heart Disease

Fig 7 Block Diagram

Data distribution :

Gather a dataset that contains historical health records, patient information, and diagnostic data related to heart disease. This dataset should include features like age, gender, blood pressure, cholesterol levels, ECG results, and more.

Identifying missing values using median:

Identify Missing Values: First, identify the columns or features in your dataset that contain missing values.
Calculate the Median: For each of the columns with missing values, calculate the median of the available data points. The median is the middle value in a sorted list of values. If you have an even number of data points, the median is the average of the two middle values.
Impute Missing Values: Replace the missing values in each column with the calculated median value. This can be done for each feature with missing data independently

Encoding categorical values:

Encoding categorical values is an essential step in preparing your data for machine learning models, as most algorithms require numerical input data.

Separating Testing and Training Models

Machine Learning Models:

KNN

SVC- Support Vector Classification
Random Forest
Extra Tree Classifier

Removing Outliers:

Detecting outliers is an important step in data preprocessing and analysis to identify data points that deviate significantly from the majority of the data. Outliers can have a significant impact on statistical analysis and machine learning models, so it's important to identify and handle them appropriately.

2.2 Algorithms:

Support Vector Classifier Algorithm:

Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. It works by mapping data points to a high-dimensional space using the kernel function and then finding the optimal hyperplane that divides the data into two classes.

The function used in svc algorithm are exceptions targeted at software and operating systems for generating system function calls. They are sometimes called software interrupts. Instead of allowing user programs to directly access hardware, an operating system may provide access to hardware through an SVC.

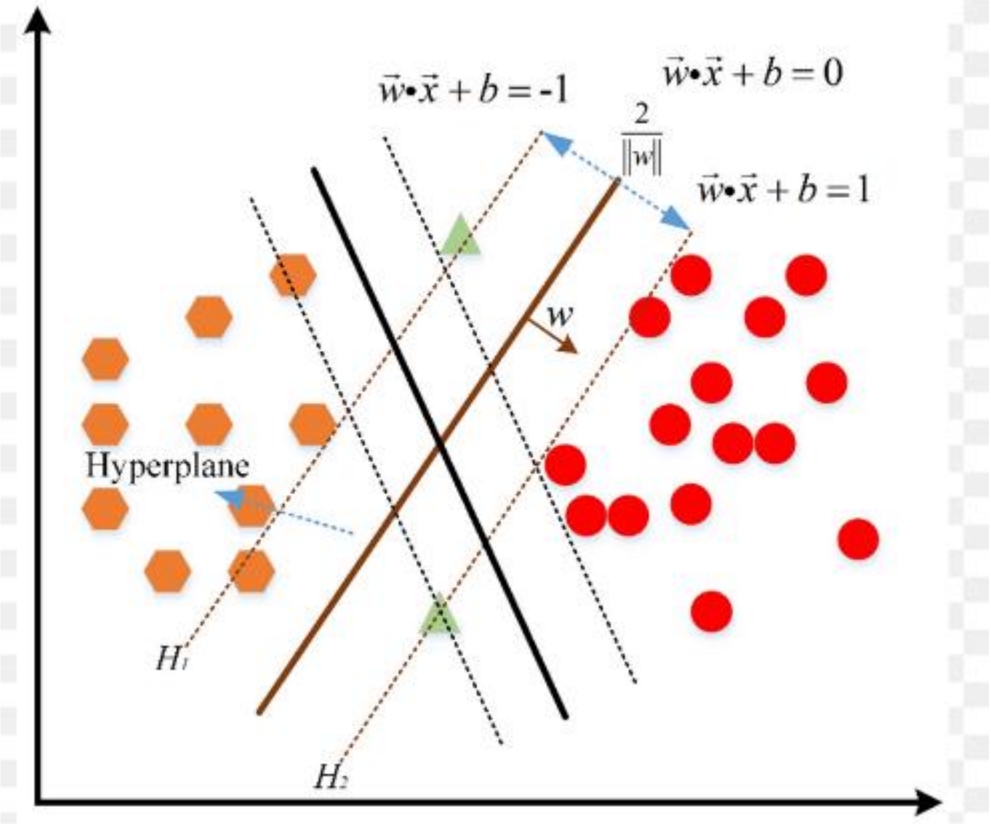

$\vec{w}\bullet\vec{x}+b=-1$
$\vec{w}\bullet\vec{x}+b=0$
$\frac{2}{\|w\|}$
$\vec{w}\bullet\vec{x}+b=1$
w
Hyperplane
H_1
H_2

Fig 8 Support Vector Classification

The purpose of using SVC is to find a hyperplane in a feature space that best separates the data points into different classes. It does this by maximizing the margin between the closest data points (support vectors) of different classes. The margin is the distance between the hyperplane and the nearest data points.

In a two-class classification problem, the hyperplane is the decision boundary that separates the data points of one class from the other. SVC can make use of the kernel trick, which allows it to transform the input features into a higher-dimensional space. This is particularly useful when the data is not linearly separable in the original feature space.

While SVC is originally designed for binary classification, it can be extended to handle multi-class problems through various techniques, such as one-vs-one or one-vs-all strategies.

K-Nearest Neighbors(KNN) Algorithm:

The k-nearest neighbors algorithm, also known as KNN is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. In KNN classification, the class label of the majority of the K nearest neighbors is assigned to the target data point.

KNN makes predictions based on the average value of the K nearest data points to the point you want to classify. "K" is a user-defined hyperparameter that determines the number of neighbors to consider. The optimal K value usually found is the square root of N, where N is the total number of samples. Use an error plot or accuracy plot to find the most favourable K value

The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.

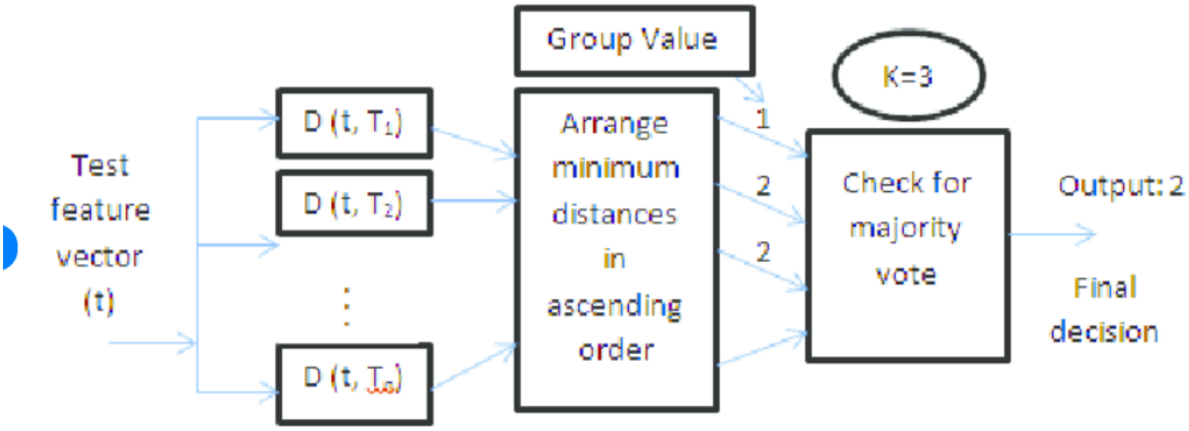


Fig 9 KNN Mechanism

Steps to implement the K-NN algorithm:

- Data Pre-processing step
- Fitting the K-NN algorithm to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

Random Forest Algorithm:

Random Forest is a machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression. It is based on which is a process of multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

The combination of multiple trees in the random forest enables the prediction of the dataset's class. It is conceivable that certain decision trees may accurately predict the output, while others may not. However, when considered collectively, all the trees successfully predict the correct output.

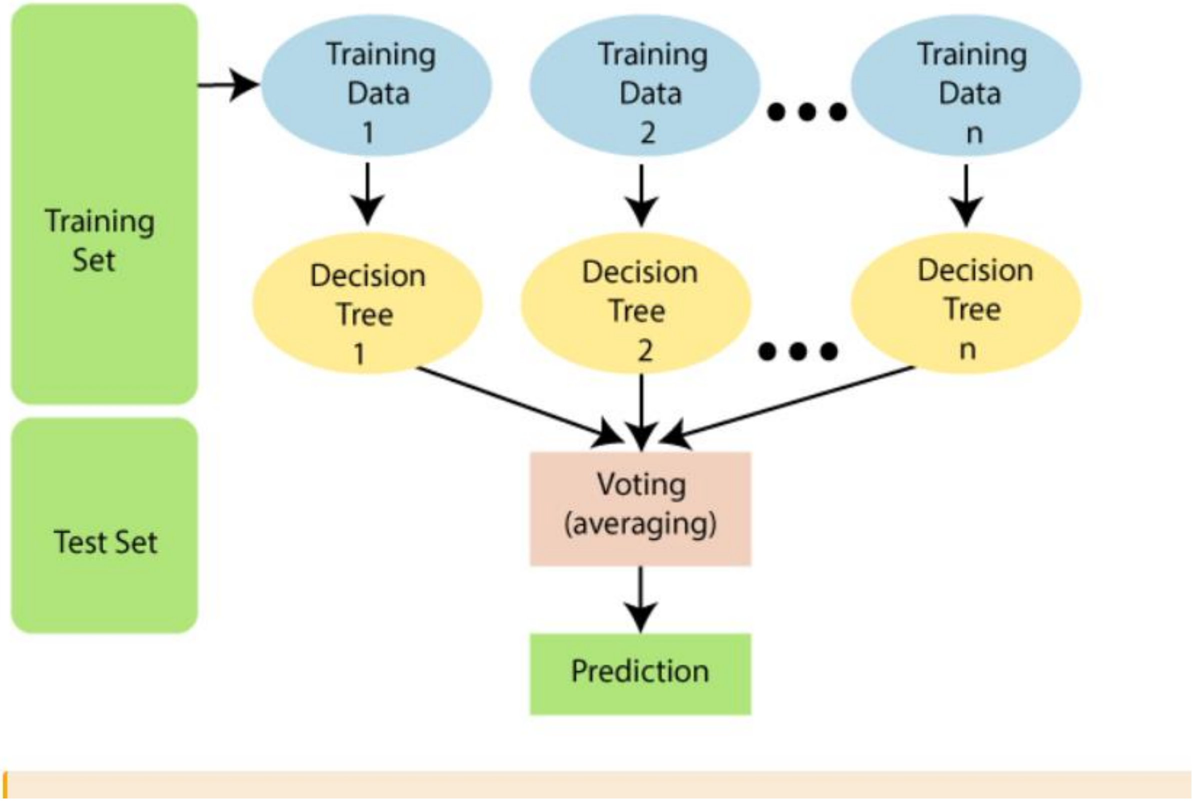


FIG 10 Random Forest

The feature variable of the dataset should contain real values to enable the classifier to generate precise predictions instead of relying on estimated outcomes. The predictions from each tree should exhibit minimal correlations.

In comparison to other algorithms, it requires less time for training. It demonstrates high accuracy in predicting output, even when handling large datasets, and operates efficiently. Additionally, it can uphold accuracy even when a significant portion of the data is missing.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is known for its ability to handle high-dimensional data and noisy datasets effectively.

The model's accuracy and generalization have been enhanced. Overfitting is mitigated through the use of an ensemble of decision trees and feature randomization, ensuring robustness. The model is capable of handling both categorical and numerical data. Additionally, it provides an out-of-the-box feature importance ranking, aiding in the identification of significant features within your dataset.

Implementation Steps are given below:

- Data Pre-processing step
- Fitting the Random forest algorithm to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

Extra Trees Classifier:

Extra Tree Classifier is a machine learning algorithm used for classification problems. It is an extension of the Random Forest algorithm and works by creating multiple decision trees using a random subset of features and data points. However, unlike Random Forest, Extra Tree Classifier selects the splitting point randomly instead of searching for the best split point. This random selection helps to reduce overfitting and increase the diversity of the trees.

Understanding Decision Trees:

Before we dive into Extra Trees, it's important to understand the basics of decision trees. Decision trees are a type of supervised learning algorithm that can be used for classification or regression tasks. They work by recursively splitting the data into subsets based on the values of the input features, until a stopping criterion is met. The result is a tree-like structure where each internal node represents a test on an input feature, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value

How Extra Trees Work:

Extra Trees work by creating a large number of decision trees, each on a random subset of the training data and a random subset of the input features. The trees are constructed using a random selection of splits at each node, rather than the best split, which makes them less prone to overfitting and more robust to noisy data. The class label of a test point is then determined by aggregating the predictions of all the trees, either by majority voting or by averaging the predicted values.

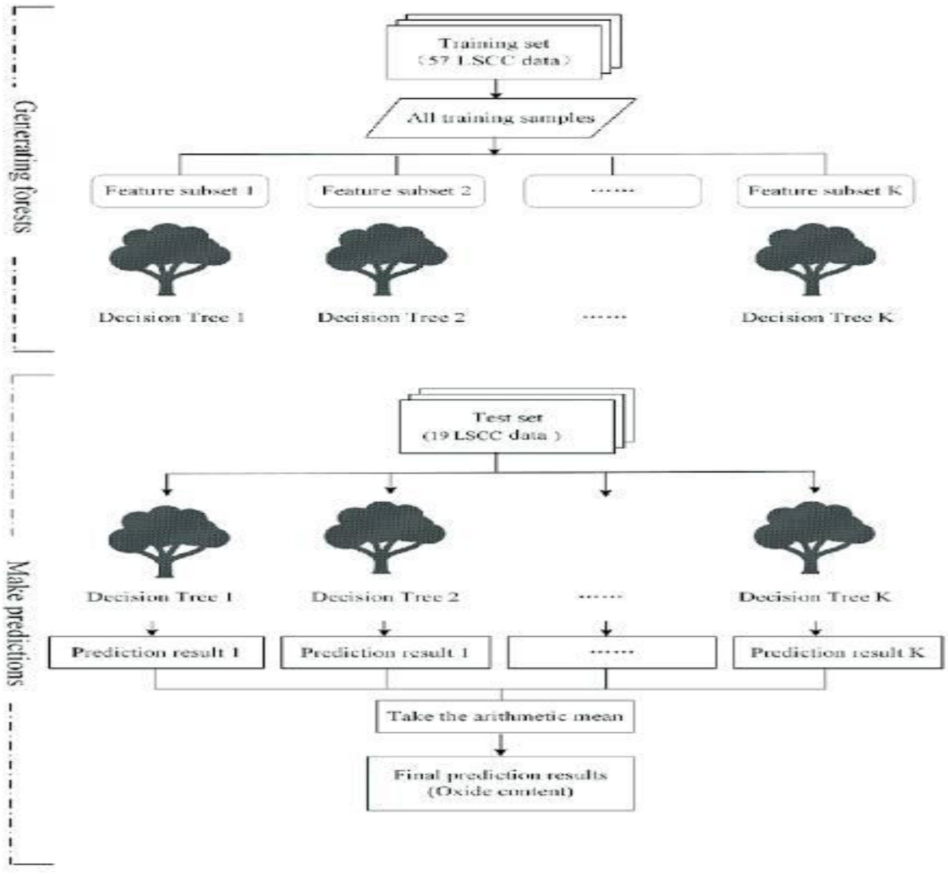


FIG 11 Extra Tree Classifier

2.3 ABOUT DATASET:

Dataset is taken from GitHub and this dataset has both types of attributes categorical and numerical. The total count of the samples are 1189. In Data set we found 446 records against 0(Female) target and 496 records against target 1(Male) for training .Dataset is balanced.

This dataset consists of 11 features and a target variable. It has 6 nominal variables and 5 numeric variables. The detailed description of all the features are as follows:

1. **Age:** Patients Age in years (Numeric)
2. **Sex:** Gender of patient (Male - 1, Female - 0) (Nominal)
3. **Chest Pain Type:** Type of chest pain experienced by patient categorized into 1 typical, 2 typical angina, 3 non-anginal pain, 4 asymptomatic (Nominal)
4. **resting bp s:** Level of blood pressure at resting mode in mm/HG (Numerical)
5. **Cholestrol :** Serum cholestrol in mg/dl (Numeric)
6. **fasting blood sugar:** Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
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 (Nominal)
8. **max heart rate:** Maximum heart rate achieved (Numeric)
9. **exercise angina:** Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
10. **oldpeak:** 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

Target Variable

target: It is the target variable which we have to predict 1 means patient is suffering from heart risk and 0 means patient is normal.

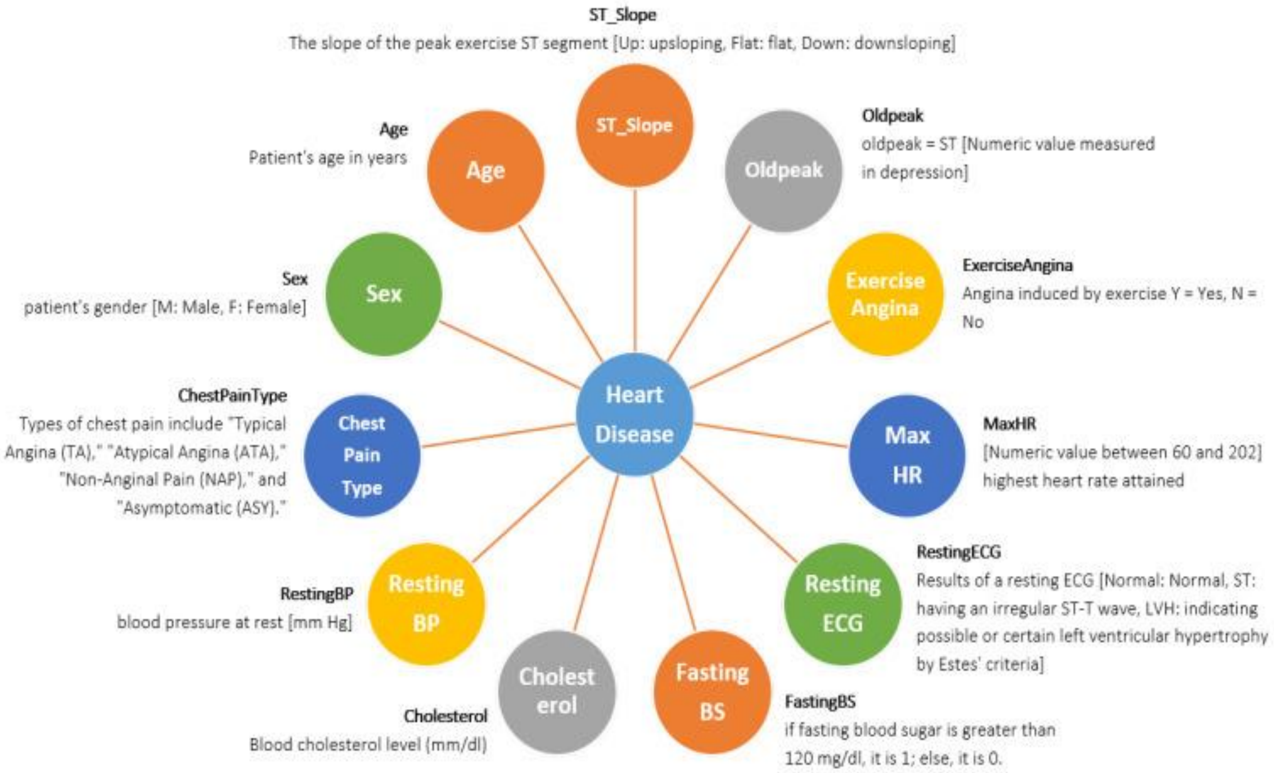

ST_Slope
The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
ST_Slope
Age
Patient's age in years
Age
Oldpeak
Oldpeak
oldpeak = ST [Numeric value measured in depression]
Sex
patient's gender [M: Male, F: Female]
Sex
Exercise Angina
ExerciseAngina
Angina induced by exercise Y = Yes, N = No
Heart Disease
ChestPainType
Types of chest pain include "Typical Angina (TA)," "Atypical Angina (ATA)," "Non-Anginal Pain (NAP)," and "Asymptomatic (ASY)."
Chest Pain Type
Max HR
MaxHR
[Numeric value between 60 and 202] highest heart rate attained
RestingBP
blood pressure at rest [mm Hg]
Resting BP
Resting ECG
RestingECG
Results of a resting ECG [Normal: Normal, ST: having an irregular ST-T wave, LVH: indicating possible or certain left ventricular hypertrophy by Estes' criteria]
Cholesterol
Blood cholesterol level (mm/dl)
Cholest erol
Fasting BS
FastingBS
if fasting blood sugar is greater than 120 mg/dl, it is 1; else, it is 0.

Fig 11 Dataset

2.4 INSTALLATIONS:

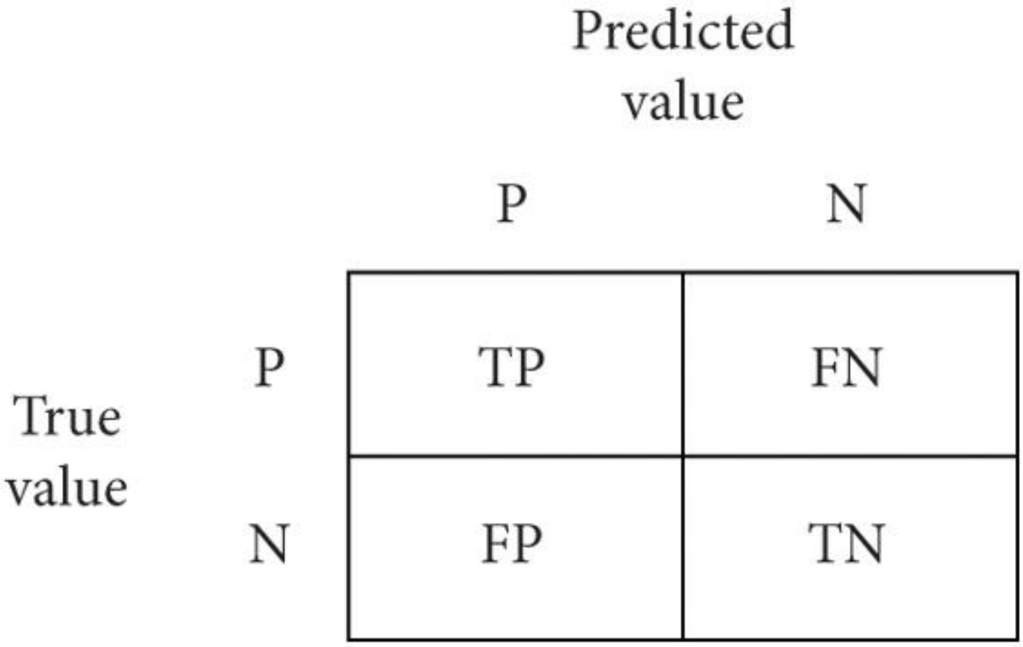
This project requires Python 3.x and the following Python libraries should be installed to get the project started:

- Numpy
- Pandas
- matplotlib
- scikit-learn
- seaborn

Algorithm:
Input: symptoms
Output: predict heart disease present or not present
1. If (the model has not been trained), then
2. Dataset load;
3. Correlation of data;
4. Check outliers;
5. Remove outliers;
6. Split x and y;
7. Train (80%), test (20%);
8. Load pre-trained model;
9. Educate the model;
10. Save the model that has been trained.
11. Loads trained model if everything else fails;
12. Validate the model using the test data set;
13. Confusion metrics and plot graphs.

The overall performance of the pre-trained models is evaluated using four criteria: true positive = *TP*, true negative = *TN*, false positive = *FP*, and false negative = *FN*. The system's performance is assessed by using the Equations (1)–(4)

Considering that when the balance of the samples is adequately predicted, the class of matter is genuinely positive and in the case of the class of matter is a genuine negative, the balance of the samples is not adequately predicted. The dimension of units mislabeled as a class of interest is known as false positive. The fraction of samples mislabeled as non-class of interest is false negative



		Predicted value	
		P	N
True value	P	TP	FN
	N	FP	TN

EVALUTION METRICS:

The performance of the each algorithm classifier model was evaluated using various metrics such as accuracy, precision recall, F1-score, and ROC-AUC score.

ACCURACY:

The accuracy of the model was found to be 0.90, indicating that the model was able to correctly predict the presence or absence of heart disease in 90% of cases

PRECISION:

The precision and recall scores were found to be 0.88 respectively, indicating that the model was able to correctly identify a higher proportion of true positives while minimizing false positive.

F1-score:

The F1-score of the model was found to be 0.90, which represents thee harmonic mean of precision and recall and provides a balanced measure of the model's performance.

ROU-AUC Score:

The ROUC-AUC score was found to be 0.91, indicating that the model has a high discriminatory power in distinguishing between positive and negative cases of heart disease.

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}},$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}},$$

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}},$$

$$F1 - \text{score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision+recall}}$$

Fig 12 Evalution Metrics

3 IMPLEMENTATION :

1.IMPORTING LIBRARIES

We have used 4 algorithms to predict the heart disease.

Random Forest

Extra tree classifier

K-Nearest Neighbor

SVC-Support Vector Classification

The performance of each of this model was evaluated using various metrics such as accuracy, precision recall, F1-score and ROC-AUC score.

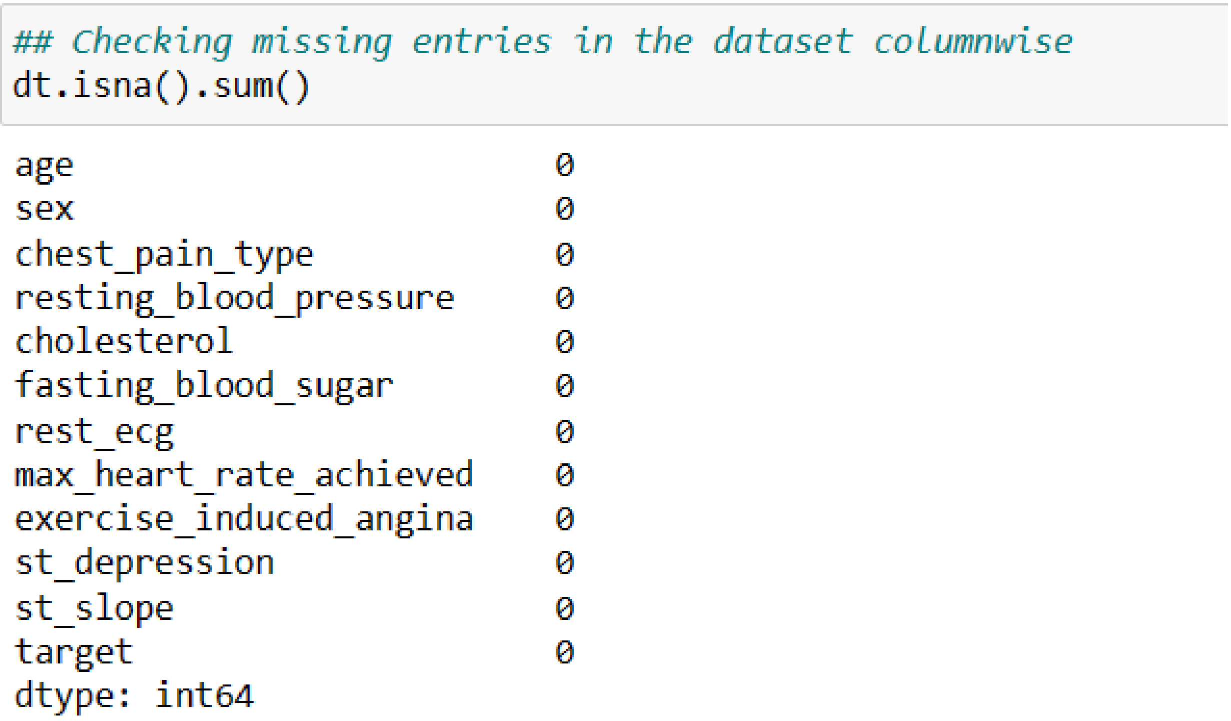
3.1 Loading Dataset:

As we can see from above dataset entries some of the features should be nominal and to be encoded as their category type. In the next step we will be encoding features to their respective category as per the dataset description.

3.2 Datacleaning and Preprocessing

In this step we will first change the name of columns as some of the columns have weird naming pattern and then we will encode the features into categorical variables

```
'age', 'sex', 'chest_pain_type', 'resting_blood_pressure',
'cholesterol', 'fasting_blood_sugar', 'rest_ecg',
'max_heart_rate_achieved','exercise_induced_angina', 'st_depression',
'st_slope','target'
```



```
## Checking missing entries in the dataset columnwise
dt.isna().sum()
```

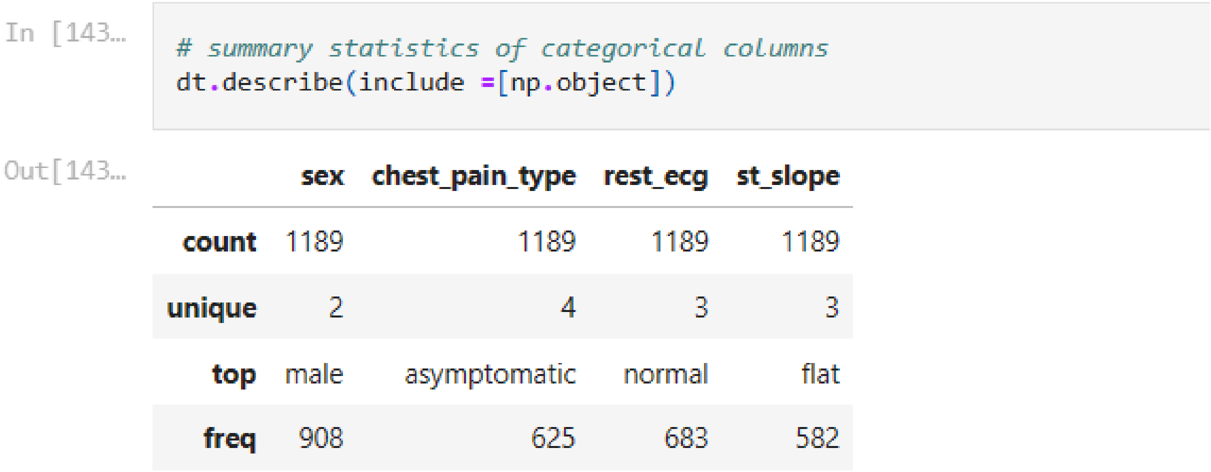
```
age                        0
sex                        0
chest_pain_type            0
resting_blood_pressure     0
cholesterol                0
fasting_blood_sugar        0
rest_ecg                   0
max_heart_rate_achieved    0
exercise_induced_angina    0
st_depression              0
st_slope                   0
target                     0
dtype: int64
```

Missing values in a data set can be handled in one of two ways: they can be ignored or they can be accounted for. Dropping missing values results in a loss of information and precision for the data collected, as well as a reduction in the number of data available for training the model. Missing data is usually not fixed; rather, it is dealt with in order to deal with it. Impu does the missing data accounting

So, there are no missing entries in the dataset thats great. Next we will move towards exploring the dataset by performing detailed EDA

3.3 Exploratory Data Analysis (EDA)

As we can see from above description resting_blood_pressure and cholestrol have some outliers as they have minimum value of 0 whereas cholestrol has outlier on upper side also having maximum value of 603.



```
In [143…  # summary statistics of categorical columns
          dt.describe(include =[np.object])
```

Out[143…

	sex	chest_pain_type	rest_ecg	st_slope
count	1189	1189	1189	1189
unique	2	4	3	3
top	male	asymptomatic	normal	flat
freq	908	625	683	582

Among the total count of 1189 among male and female Chest pain type-4(count)

Asymptomatic, non-anginal pain ,typical, angina, atypical angina.

Distribution of Heart disease (target variable)

The distribution of the data plays an important role when the prediction or classification of a problem is to be done. We see that the heart disease occurred 53% of the time in the dataset, whilst 47% was the no heart disease. So, we need to balance the dataset or otherwise it might get overfit. This will help the model to find a pattern in the dataset that contributes to heart disease and which does not as shown in Figure.

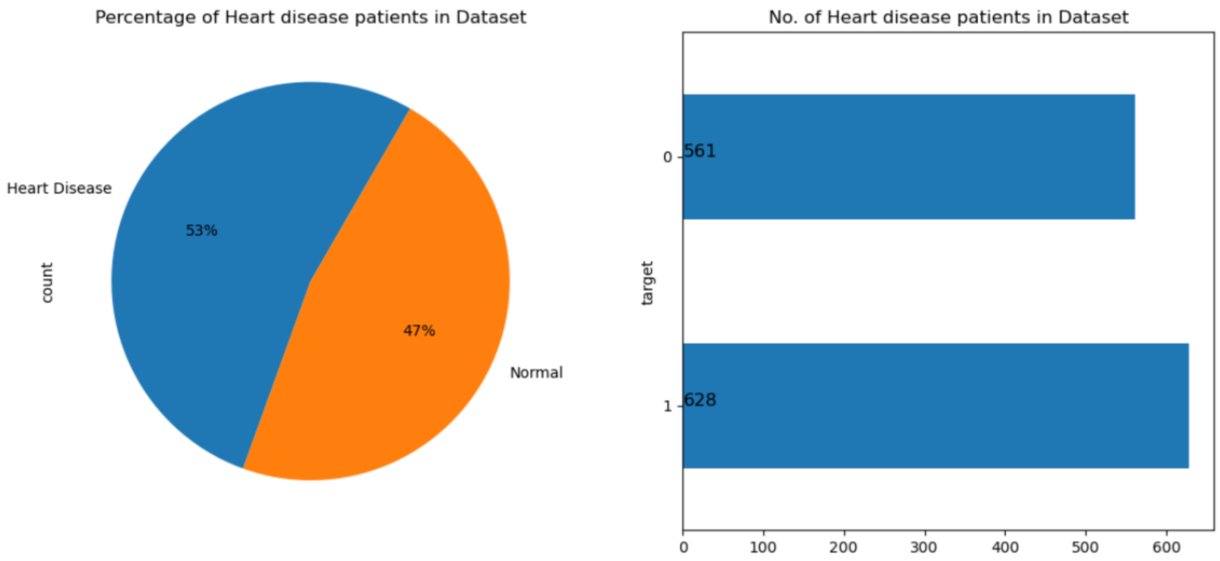


Fig 14 Heart Disease and Normal patients

The dataset is balanced having 628 heart disease patients and 561 normal patients

Checking Gender & Agewise Distribution

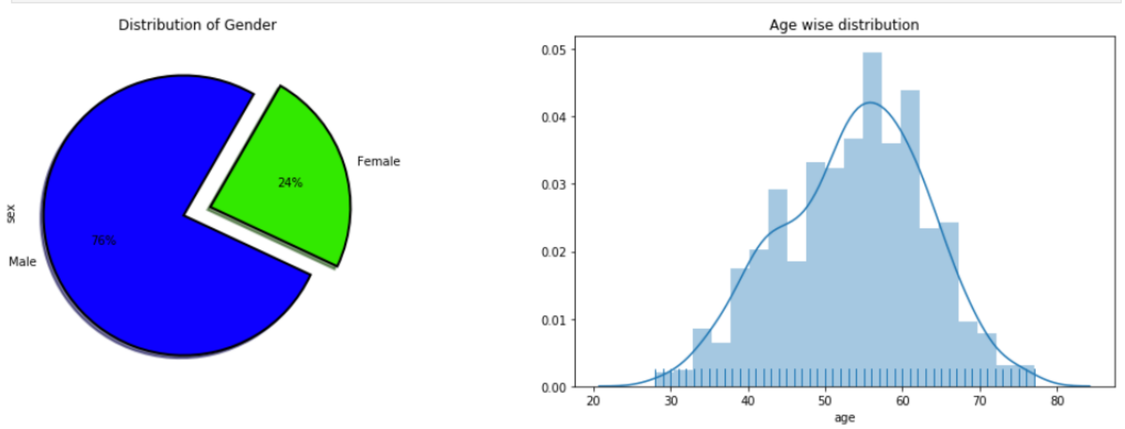


Fig 15 Female and male age distribution

As we can see from above plot, in this dataset males percentage is way too higher than females where as average age of patients is around 55.

Distribution of Chest Pain Type:

CHEST PAIN OF NORMAL AND HEART DISEASE PATIENT:

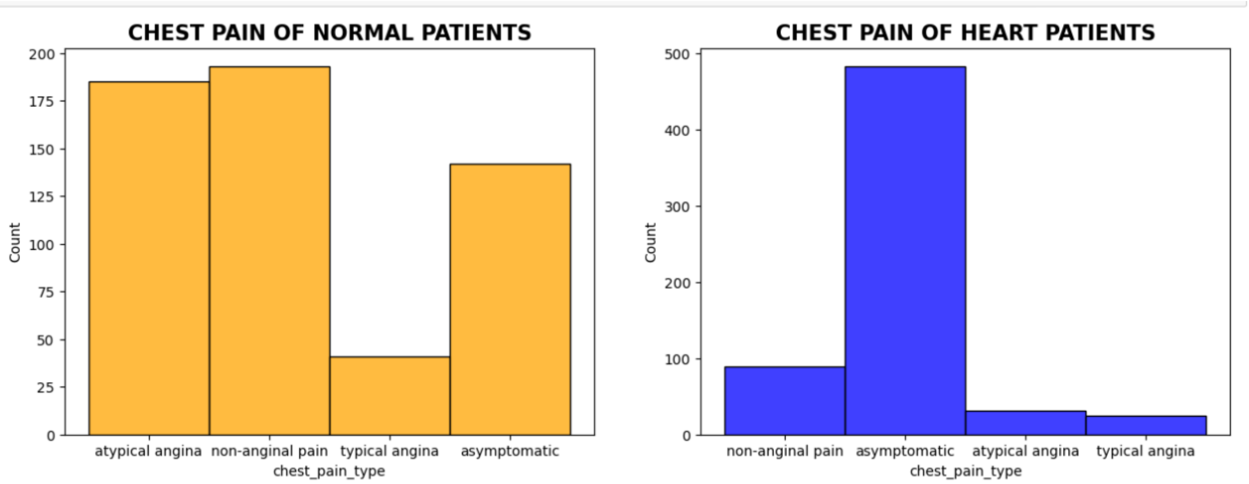
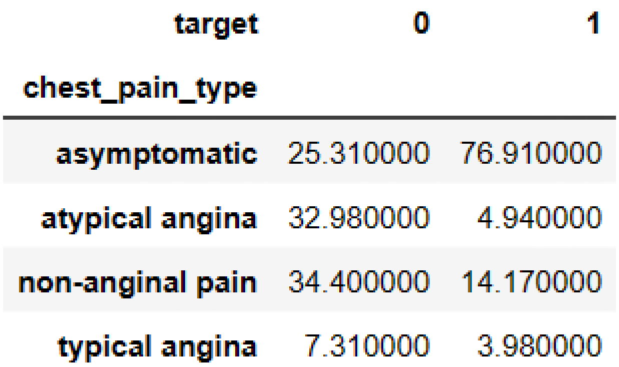


Fig 16 Chest pain of heart disease



target	0	1
chest_pain_type		
asymptomatic	25.310000	76.910000
atypical angina	32.980000	4.940000
non-anginal pain	34.400000	14.170000
typical angina	7.310000	3.980000

As we can see from above plot **76%** of the chest pain type of the heart disease patients have asymptomatic chest pain.

ASYMPTOMATIC:

Asymptomatic is a medical term used to describe a condition or disease in which a person is infected or affected, but they do not display any noticeable symptoms or signs of the illness. Essentially, an asymptomatic individual is carrying or has the condition but feels perfectly healthy and does not exhibit any of the typical manifestations of the disease

Atypical angina:

Atypical angina, also known as "atypical chest pain," refers to chest discomfort or pain that does not fit the typical or classic pattern of symptoms associated with angina, a condition caused by reduced blood flow to the heart muscle. Angina is often described as a chest pain or discomfort that occurs when the heart muscle doesn't receive enough oxygen-rich blood, usually due to coronary artery disease. It's essential for healthcare professionals to carefully evaluate individuals with atypical chest pain to rule out serious cardiac conditions and consider other potential causes, such as gastrointestinal issues, musculoskeletal pain, or anxiety.

Non-anginal pain:

Non-anginal pain, sometimes referred to as "non-cardiac chest pain," is chest discomfort or pain that is not related to angina or heart-related causes. In other words, it is chest pain that does not originate from the heart or coronary arteries. the chest can result from muscle strain, rib cage injuries, or conditions like costochondritis, which is inflammation of the cartilage that connects the ribs to the breastbone. non-anginal chest pain from cardiac chest pain (angina or other heart-related pain) as the causes and management differ significantly.

Distribution of Rest ECG Type:

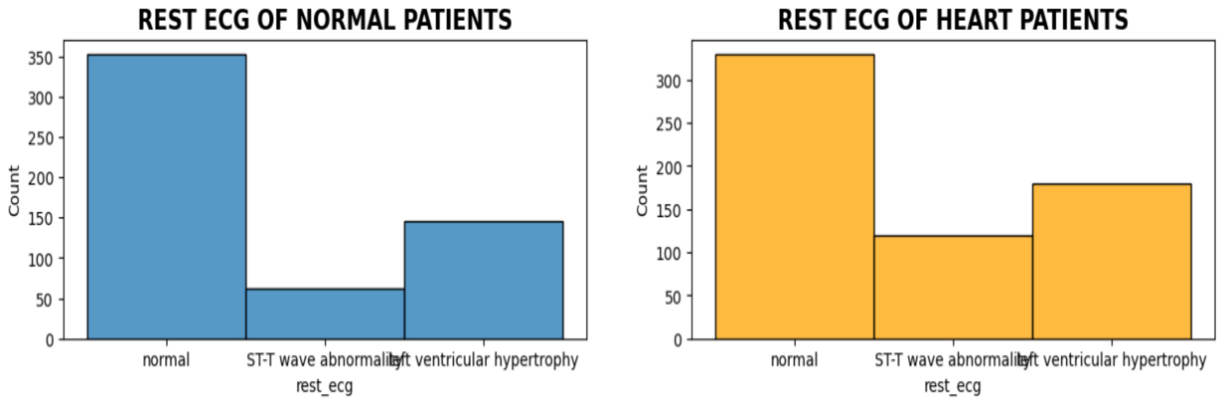
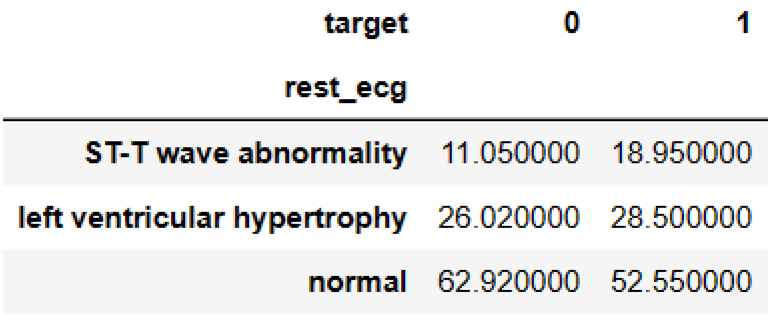
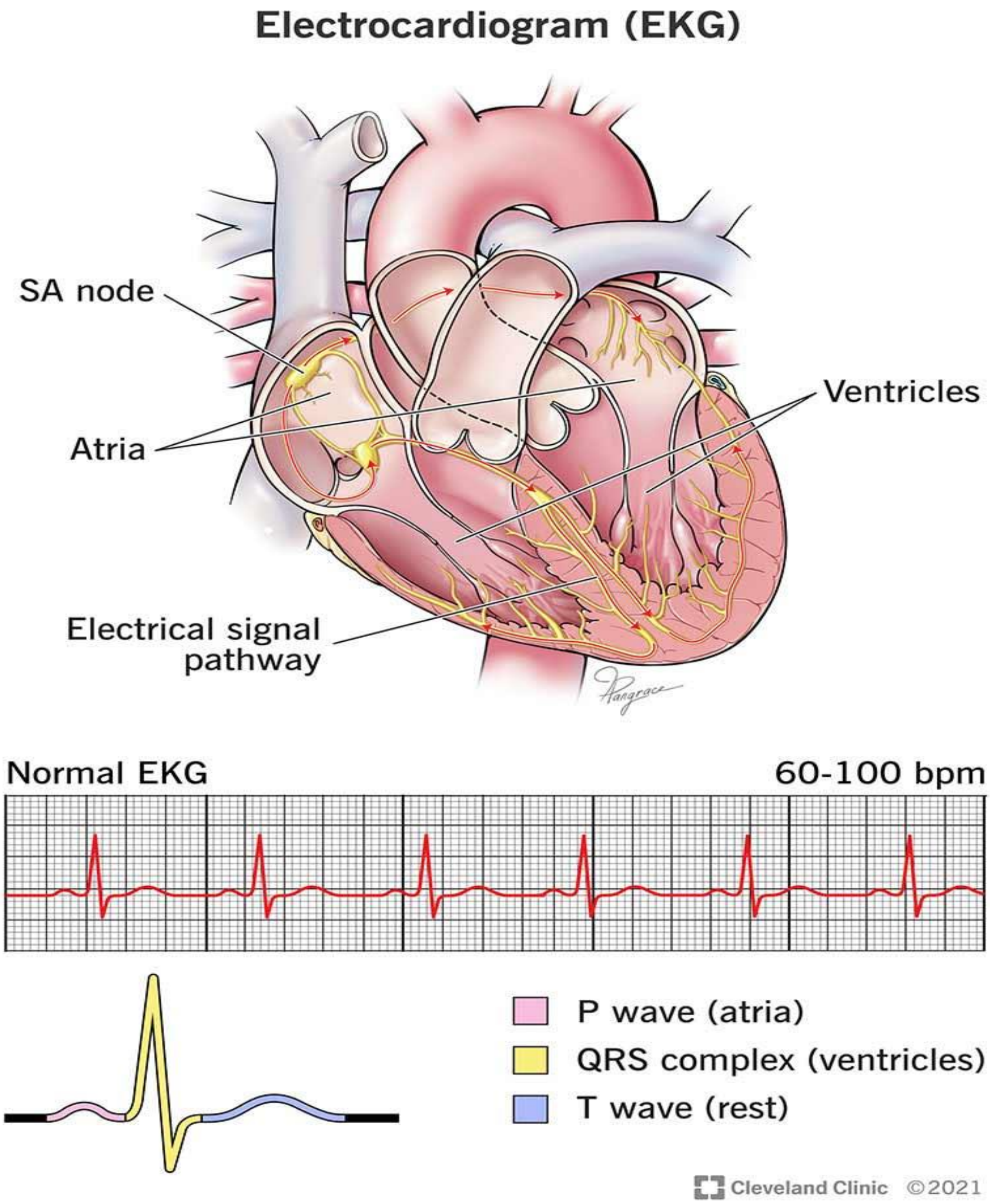


Fig 17 Rest ECG



target	0	1
rest_ecg		
ST-T wave abnormality	11.050000	18.950000
left ventricular hypertrophy	26.020000	28.500000
normal	62.920000	52.550000



Cleveland Clinic ©2021

ST-T wave abnormality:

An ST-T abnormality on an (ECG) is known to independently predict subsequent morbidity and mortality from cardiovascular diseases. The interpretation of ST- and T-wave changes is contingent upon the clinical context and the existence of comparable findings on previous electrocardiograms.

LEFT VENTRICULAR HYPERTROPHY:

Left ventricular hypertrophy is a condition characterized by the thickening of the muscular wall of the left ventricle of the heart. The left ventricle is the chamber responsible for pumping oxygen-rich blood to the rest of the body.

NORMAL:

A normal electrocardiogram (ECG or EKG) is a graphical representation of the electrical activity of the heart that shows a specific pattern of waves and intervals. The ECG is a valuable tool used to assess the heart's electrical activity and can help detect various heart-related conditions.

An electrocardiogram records the electrical signals in your heart. It's a common test used to detect heart problems and monitor the heart's status in many situations. Electrocardiograms — also called ECGs or EKGs. but ECG has limits. It measures heart rate and rhythm—but it doesn’t necessarily show blockages in the arteries.Thats why in this dataset around 52% heart disease patients have normal ECG

Distribution of ST Slope:

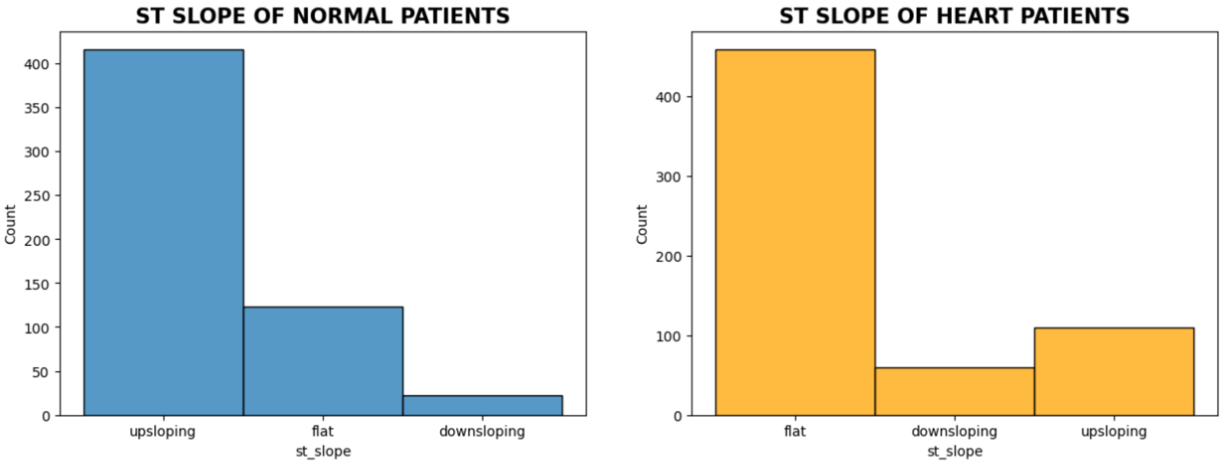
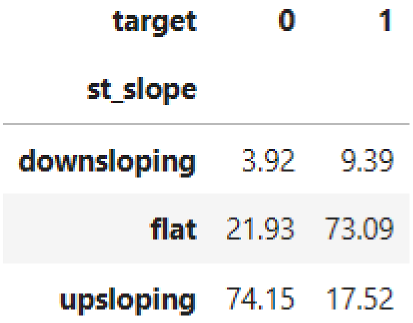


Fig 19 ST Slope



target	0	1
st_slope		
downsloping	3.92	9.39
flat	21.93	73.09
upsloping	74.15	17.52

Flat:

A "flat" segment suggests that there are no abnormal electrical patterns or deviations from the expected normal electrical activity of the heart during that specific part of the ECG

Upsloping:

In Upsloping is refers to a specific pattern of ST-segment changes in the ECG recording. The ST segment is the flat line on an ECG that occurs between the end of the QRS complex and the beginning of the T wave.

Downsloping:

The downsloping is often used in the context of electrocardiography or stress testing to describe a particular pattern of ST-segment changes, which can be indicative of cardiac ischemia or heart-related conditions. In a standard ECG, the ST segment is a flat line that occurs between the S wave and the T wave

The ST segment /heart rate slope (ST/HR slope), has been proposed as a more accurate ECG criterion for diagnosing significant coronary artery disease (CAD) in most of the research papers.

As we can see from above plot upsloping is positive sign as 74% of the normal patients have upslope where as 72.97% heart patients have flat sloping.

The distribution of age and sex, the distribution of chest pain and trestbps, the distribution of cholesterol and fasting blood, the distribution of ecg resting electrode and thalach, the distribution of exang and oldpeak, the distribution of slope and ca, and the distribution of thal and target all are analyzed.

Distribution of Numerical features

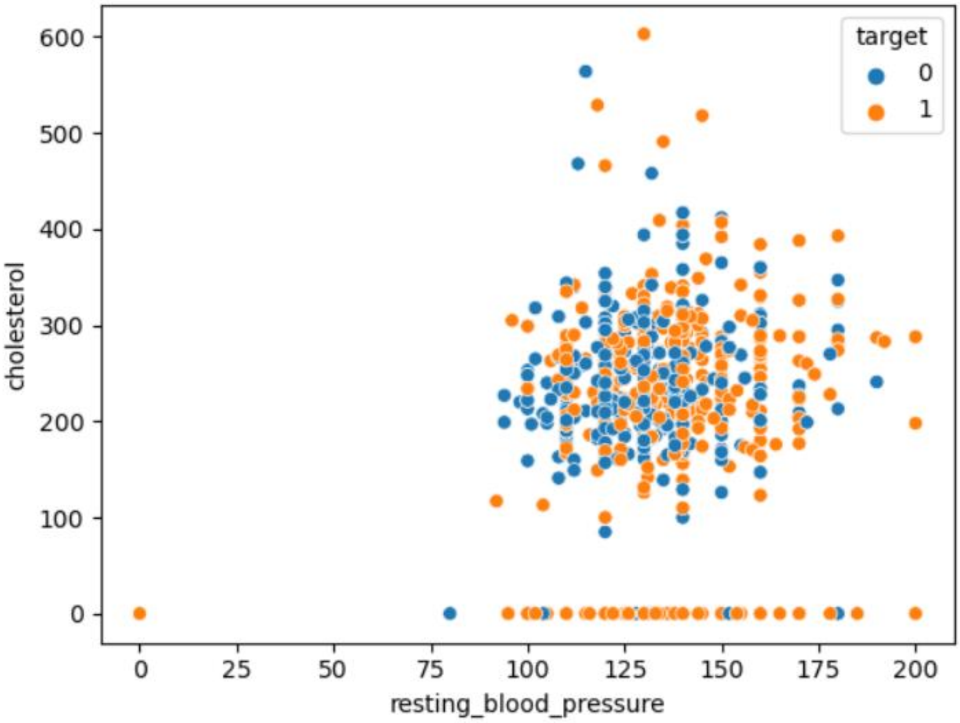
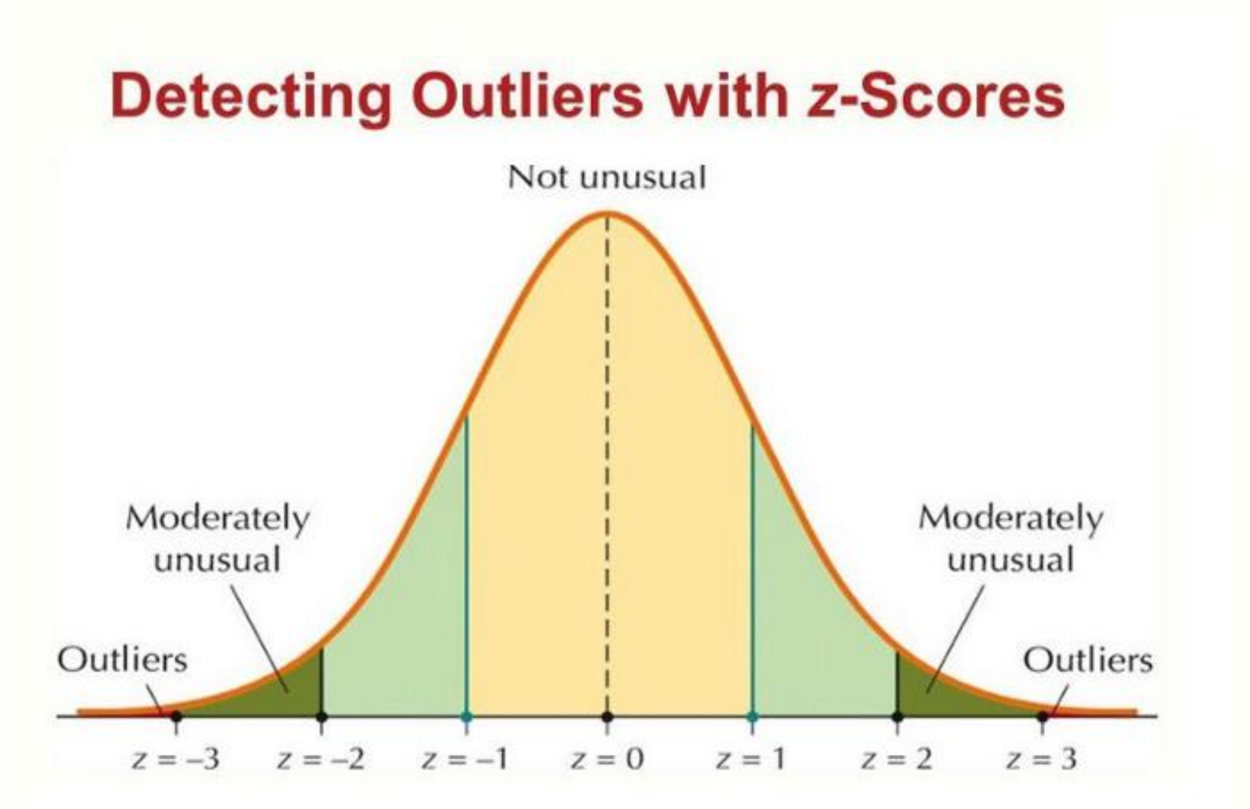
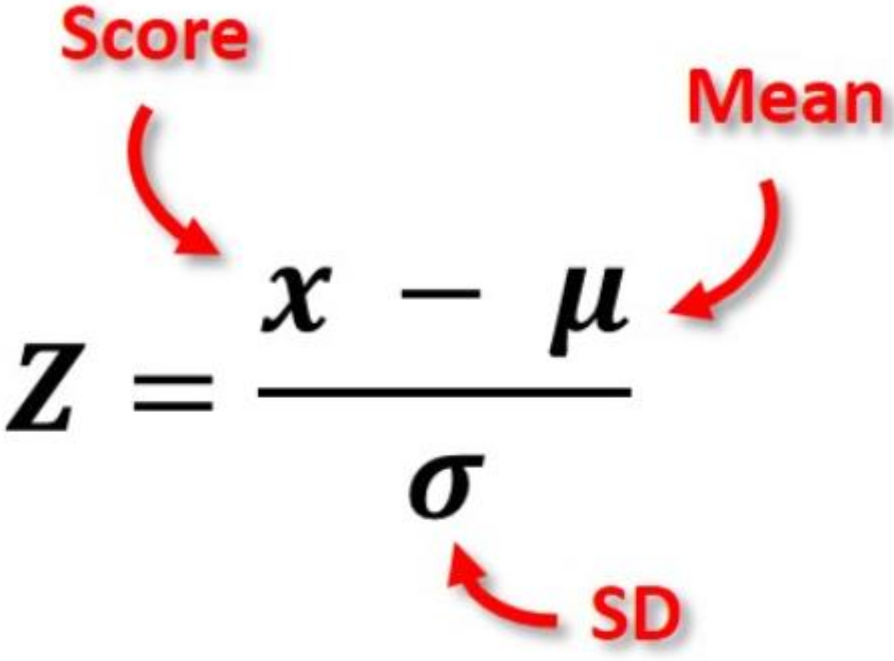


Fig 20 distribution of numerical features

From the above plot we can see outliers clearly as for some of the patients cholestrol is 0 whereas for one patient both cholestrol and resting bp is 0 which is may be due to missing entries we will filter these ouliers later.

3.4 Outlier Detection and Removal:



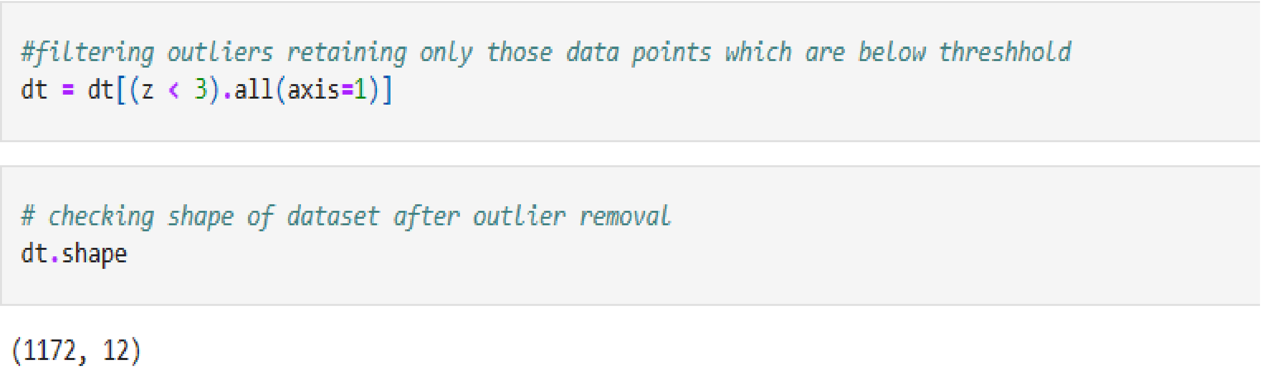


Outliers Outliers are measurements that are markedly different from the rest of the data. These outliers are highly sensitive to data distribution and have a direct impact on the learning process of machine learning classifiers

Before removing the outlier:

(1189, 12)

After removing the outlier:



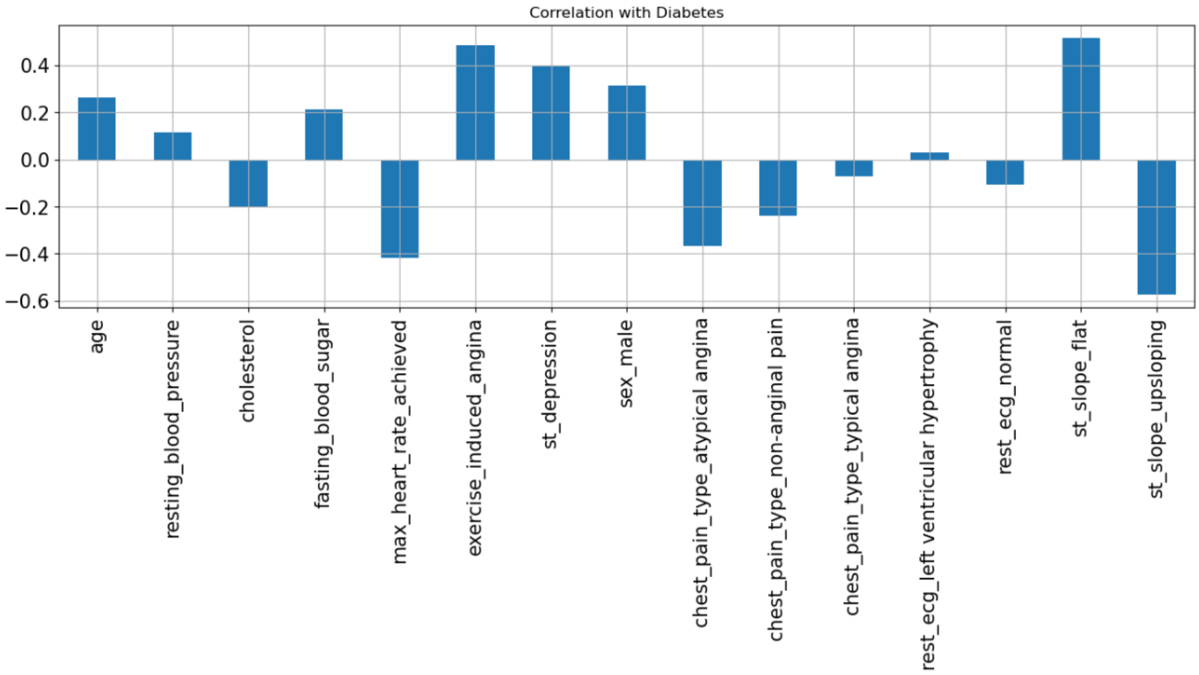
```
#filtering outliers retaining only those data points which are below threshhold
dt = dt[(z < 3).all(axis=1)]
```

```
# checking shape of dataset after outlier removal
dt.shape
```

```
(1172, 12)
```

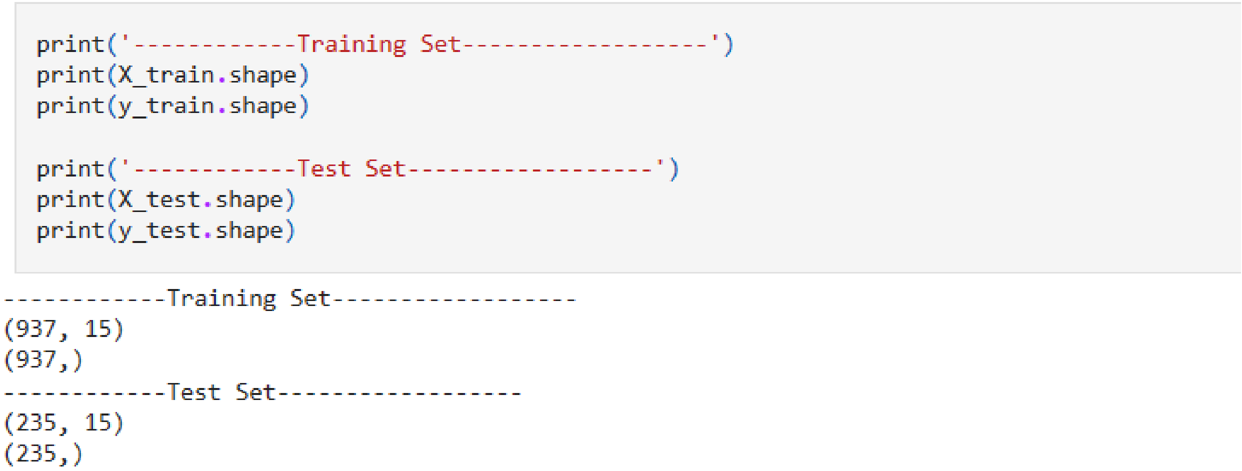
Now before splitting dataset into train and test we first encode categorical variables as d ummy variables and segregate feature and target variable.

Checking correlation:



3.5 SPLITING TRAINING AND TESTING:

```
Distribution of traget variable in training set
1    491
0    446
Name: target, dtype: int64
Distribution of traget variable in test set
1    123
0    112
```



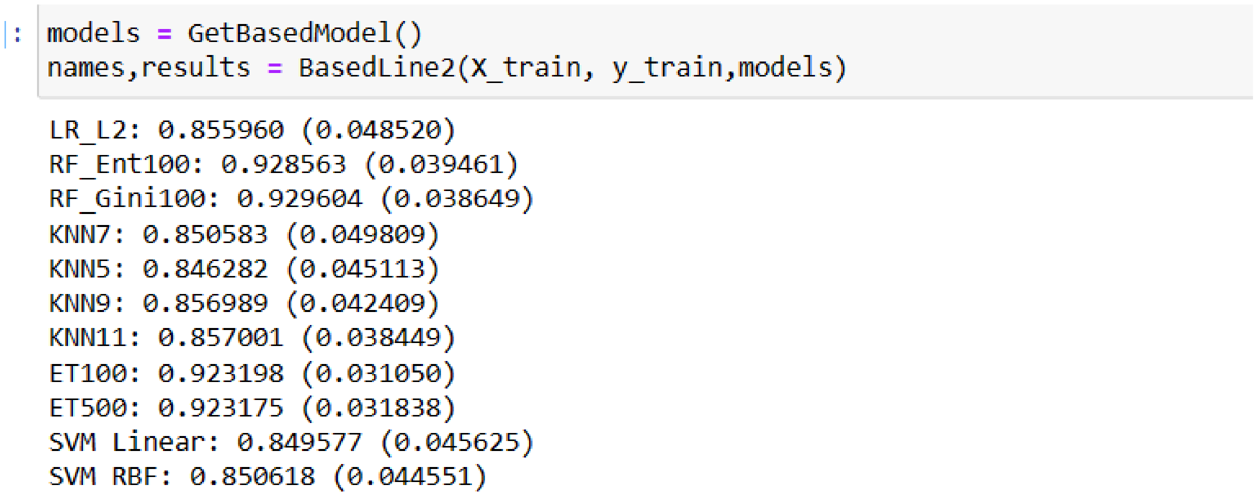
```
print('------------Training Set------------------')
print(X_train.shape)
print(y_train.shape)

print('------------Test Set------------------')
print(X_test.shape)
print(y_test.shape)
```

```
------------Training Set------------------
(937, 15)
(937,)
------------Test Set------------------
(235, 15)
(235,)
```

3.6 FEATURE NORMALIZATION:

CROSS VALIDATION:

In this step, we will build different baseline models and perform 10-fold cross validation to filter top performing baseline models to be used in level 0 of stacked ensemble method.



```
models = GetBasedModel()
names,results = BasedLine2(X_train, y_train,models)
```

```
LR_L2: 0.855960 (0.048520)
RF_Ent100: 0.928563 (0.039461)
RF_Gini100: 0.929604 (0.038649)
KNN7: 0.850583 (0.049809)
KNN5: 0.846282 (0.045113)
KNN9: 0.856989 (0.042409)
KNN11: 0.857001 (0.038449)
ET100: 0.923198 (0.031050)
ET500: 0.923175 (0.031838)
SVM Linear: 0.849577 (0.045625)
SVM RBF: 0.850618 (0.044551)
```

4 RESULTS:

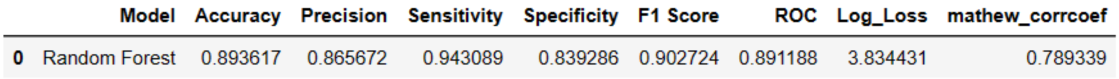
- Using the First Approach (without Doing Feature Selection and Outliers Detection We didn’t get better results.

- Using the Second Approach (Doing Feature Selection and No Outliers Detection)
- After selecting the features (feature selection) and scaling the data as there are outliers, the robust standard scalar is used; it is used when the dataset is having certain outliers. In the second approach, the accuracy achieved by Random Forest is 87% and extra tree classifier has 88.

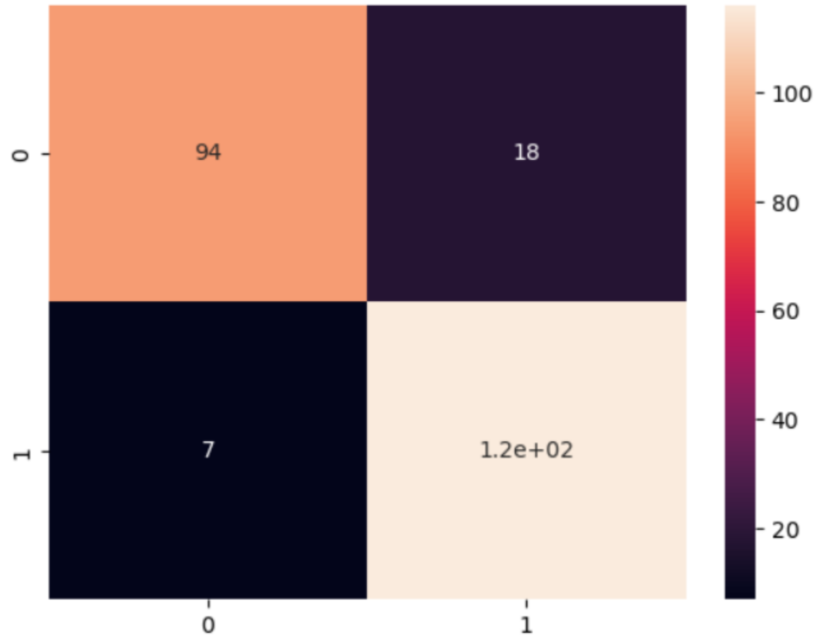
- Using the Third Approach (by Doing Feature Selection and Also Outliers Detection)

- In this approach, the dataset is normalized and the feature selection is done and also the outliers are handled using the Isolation Forest. The correlation comparison can be seen in Figure . The accuracy of the Random Forest is 89.3%, KNeighbors is 80.4%, Support Vector Machine is 81%, and the extra tree classifier as 90%.

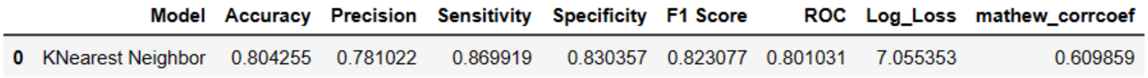
Random Forest



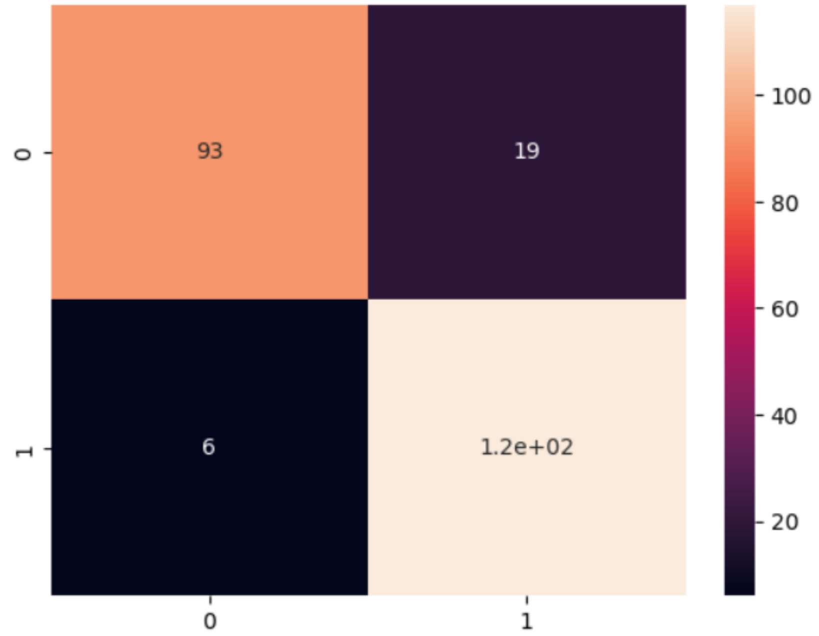
	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.893617	0.865672	0.943089	0.839286	0.902724	0.891188	3.834431	0.789339



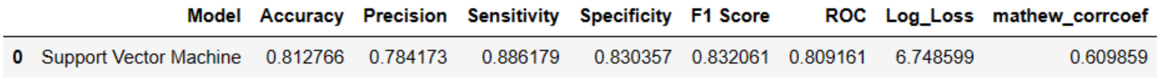
K Nearest Neighbor:



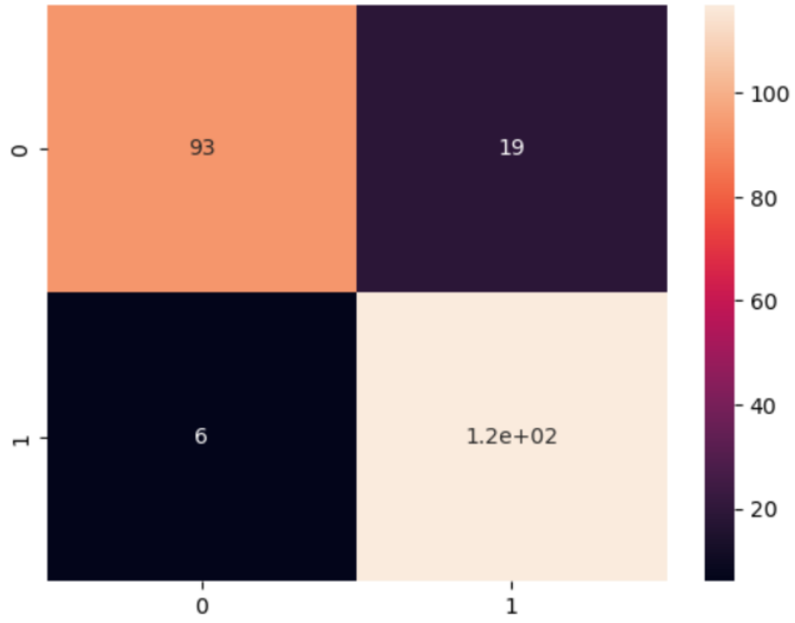
	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	KNearest Neighbor	0.804255	0.781022	0.869919	0.830357	0.823077	0.801031	7.055353	0.609859



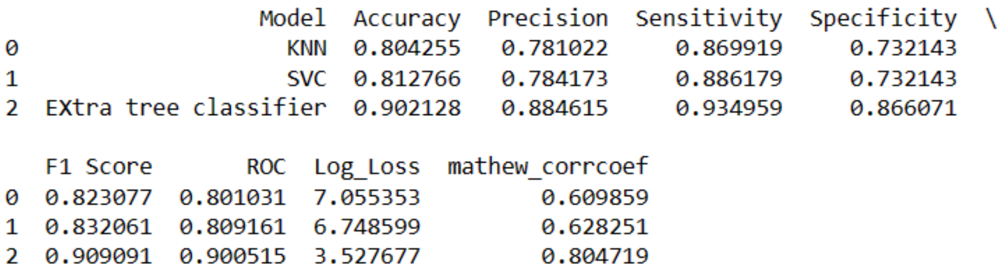
Support Vector Machine:



	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Support Vector Machine	0.812766	0.784173	0.886179	0.830357	0.832061	0.809161	6.748599	0.609859



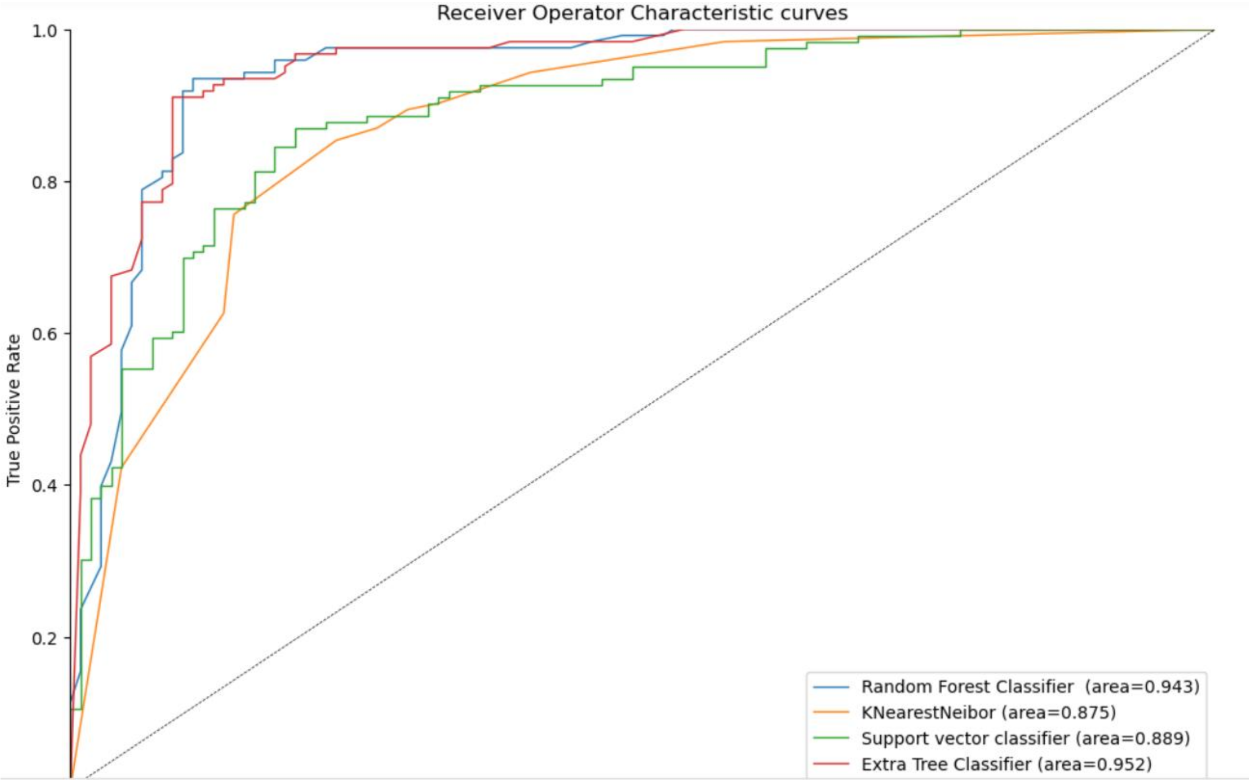
Extra Tree Classifier



```
                    Model  Accuracy  Precision  Sensitivity  Specificity  \
0                     KNN  0.804255   0.781022     0.869919     0.732143   
1                     SVC  0.812766   0.784173     0.886179     0.732143   
2  EXtra tree classifier  0.902128   0.884615     0.934959     0.866071   

   F1 Score       ROC  Log_Loss  mathew_corrcoef  
0  0.823077  0.801031  7.055353         0.609859  
1  0.832061  0.809161  6.748599         0.628251  
2  0.909091  0.900515  3.527677         0.804719  
```

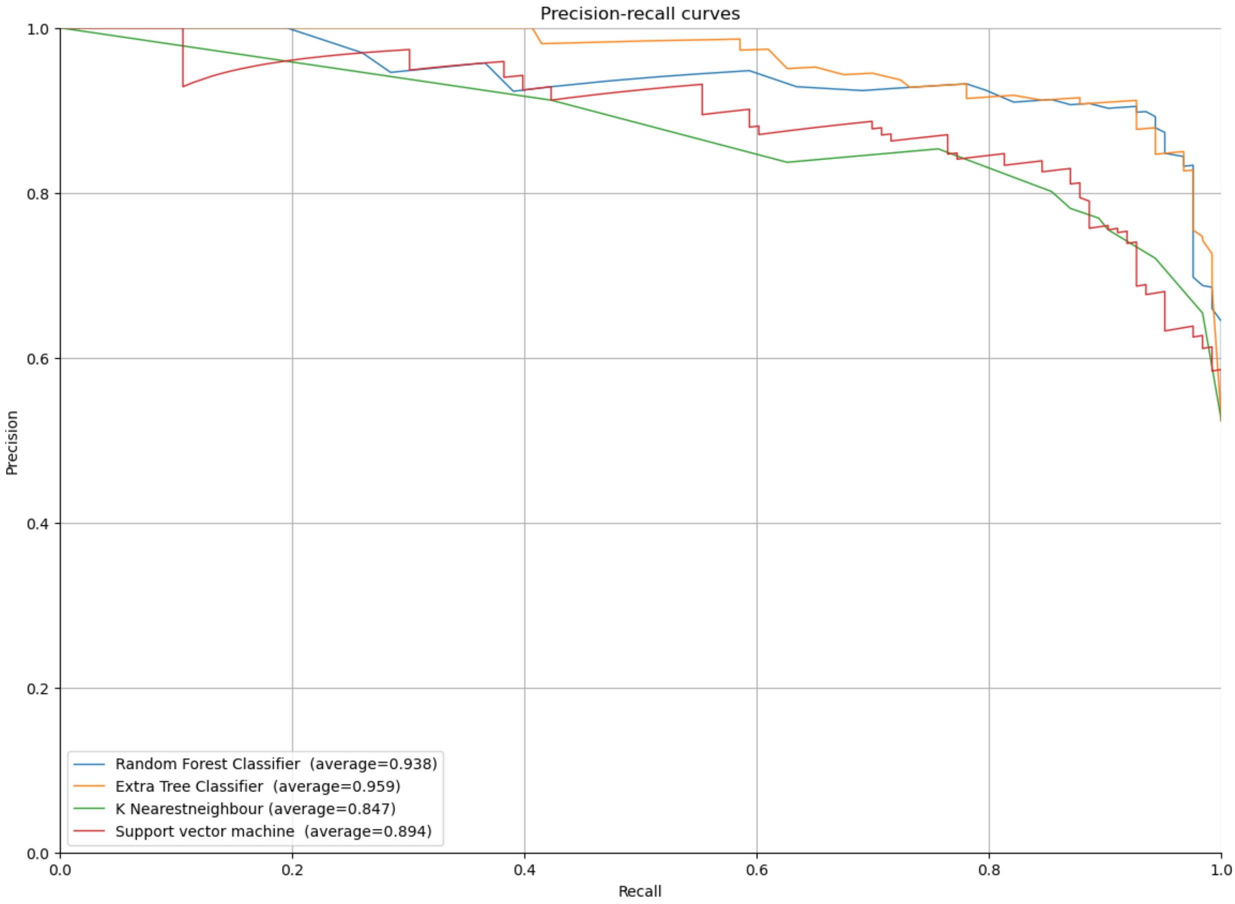
BASED ON THE ABOVE RESULTS THE EXTRA TREE CLASSIFIER HAS HIGHER ACCURACY THAN OTHER MODELS.

ROU-AOU CURVE:



In addition, we use ROC and AUC to evaluate the performance of the model. Receiver operating characteristic curve (ROC) is a curve drawn based on a series of different boundary values with a true positive rate on the ordinate and a false positive rate on the abscissa . AUC is the area under the ROC curve, which represents the probability of the calculated score of the positive sample higher than that of the negative sample when the samples are randomly selected, which can measure the pros and cons of the prediction model. Results show that the average AUC value of our model is 0.952 and the ROC curve of an experiment.

PRECISION-RECALL CURVE:



4.1 CONCLUSION:

The computational time was also reduced which is helpful when deploying a model. It was also found out that the dataset should be normalized; otherwise, the training model gets overfitted sometimes and the accuracy achieved is not sufficient when a model is evaluated for real-world data problems which can vary drastically to the dataset on which the model was trained. It was also found out that the statistical analysis is also important when a dataset is analyzed.

If a large dataset is present, the results can increase very much in machine learning.

- As we have seen, stacked ensemble of power machine learning algorithms resulted in higher performance than any individual machine learning model.
- We have also interpreted second best performing algo i.e., random forest algorithm
- The top 5 most contribution features are:

1. **Max heart Rate achieved**
2. **Cholestrol**
3. **st_depression**
4. **Age**
5. **exercise_induced_angina**

REFERENCES

R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586

L. KN, N. R, N. K, R. Kumari, S. N and V. K, "Heart Disease Detection using Machine Learning Technique," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1738-1743, doi: 10.1109/ICESC51422.2021.9532705

A.Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5, doi: 10.1109/WITS.2019.8723839.

R. Aggarwal and S. Kumar, "MLPPCA: Heart Disease Detection using Machine learning," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 457-461, doi: 10.1109/PDGC56933.2022.10053266.

P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298354.

R. Buettner and M. Schunter, "Efficient machine learning based detection of heart disease," 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom), Bogota, Colombia, 2019, pp. 1-6, doi: 10.1109/HealthCom46333.2019.9009429.

S. Dhar, K. Roy, T. Dey, P. Datta and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.

A Kumari and A. K. Mehta, "A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544544.

N. Basha, A. K. P.S., G. K. C. and V. P., "Early Detection of Heart Syndrome Using Machine Learning Technique," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 387-391, doi: 10.1109/ICEECCOT46775.2019.9114651.

FUTURE SCOPE:

The future scope of heart disease prediction is promising, as advancements in technology and medical research continue to improve our understanding and ability to predict and prevent heart disease. Here are some areas of development and future possibilities in heart disease prediction:

Genetic Risk Assessment: Genetic testing can provide valuable insights into an individual's predisposition to heart disease. Future advancements may lead to more comprehensive and accessible genetic risk assessments, allowing for personalized preventive strategies.

Wearable Devices: The use of wearable devices like smartwatches and fitness trackers to monitor heart rate, blood pressure, and other health metrics is becoming more common. The data collected by these devices can be used to detect irregularities and assess cardiovascular health over time.

Telemedicine: Telemedicine and remote monitoring technologies are making it easier for individuals to receive ongoing cardiovascular care and for healthcare providers to monitor patients at risk for heart disease, even from a distance.

Biomarkers: Research is ongoing to discover new biomarkers and molecular indicators that can predict heart disease with greater precision. Advances in proteomics and metabolomics may provide additional tools for early detection.

Big Data and Electronic Health Records: Access to extensive electronic health records and medical data can aid in predicting heart disease. These records can be mined for insights into risk factors, disease trends, and treatment outcomes.

Preventive Health Education: Public awareness and education about the risk factors for heart disease will continue to play a significant role in prevention. Future initiatives may leverage technology to deliver personalized health advice and education.

Targeted Therapies: As our understanding of the genetic and molecular basis of heart disease improves, targeted therapies may become more common, allowing for more effective prevention and treatment strategies.

Public Health Programs: Governments and healthcare organizations may implement large-scale public health programs to reduce risk factors for heart disease, such as smoking cessation, improved nutrition, and increased physical activity.

Community Health Initiatives: Local community-based initiatives can play a vital role in heart disease prevention. These may involve creating walkable neighborhoods, offering access to healthy foods, and providing support for individuals at risk.

AI-Enhanced Imaging: Advanced imaging techniques, such as MRI and CT scans, can be enhanced with AI algorithms to detect early signs of heart disease, including atherosclerosis and structural abnormalities.