



MINOR PROJECT PRESENTATION

# HEART DISEASE PREDICTION

Using Machine Learning Algorithms.

---

**Presented by:**

Praphooll Markndey (16111018)  
Chandrika Rani Tudu (18111018)  
Durgesh Kumar (18111023)  
Jitendra Rathore (18111028)  
MD Samar Siddiqui (18111035)  
Prachi Dewangan (18111041)  
Sarita Kanwar (18111047)  
Surjeet Singh (18111053)

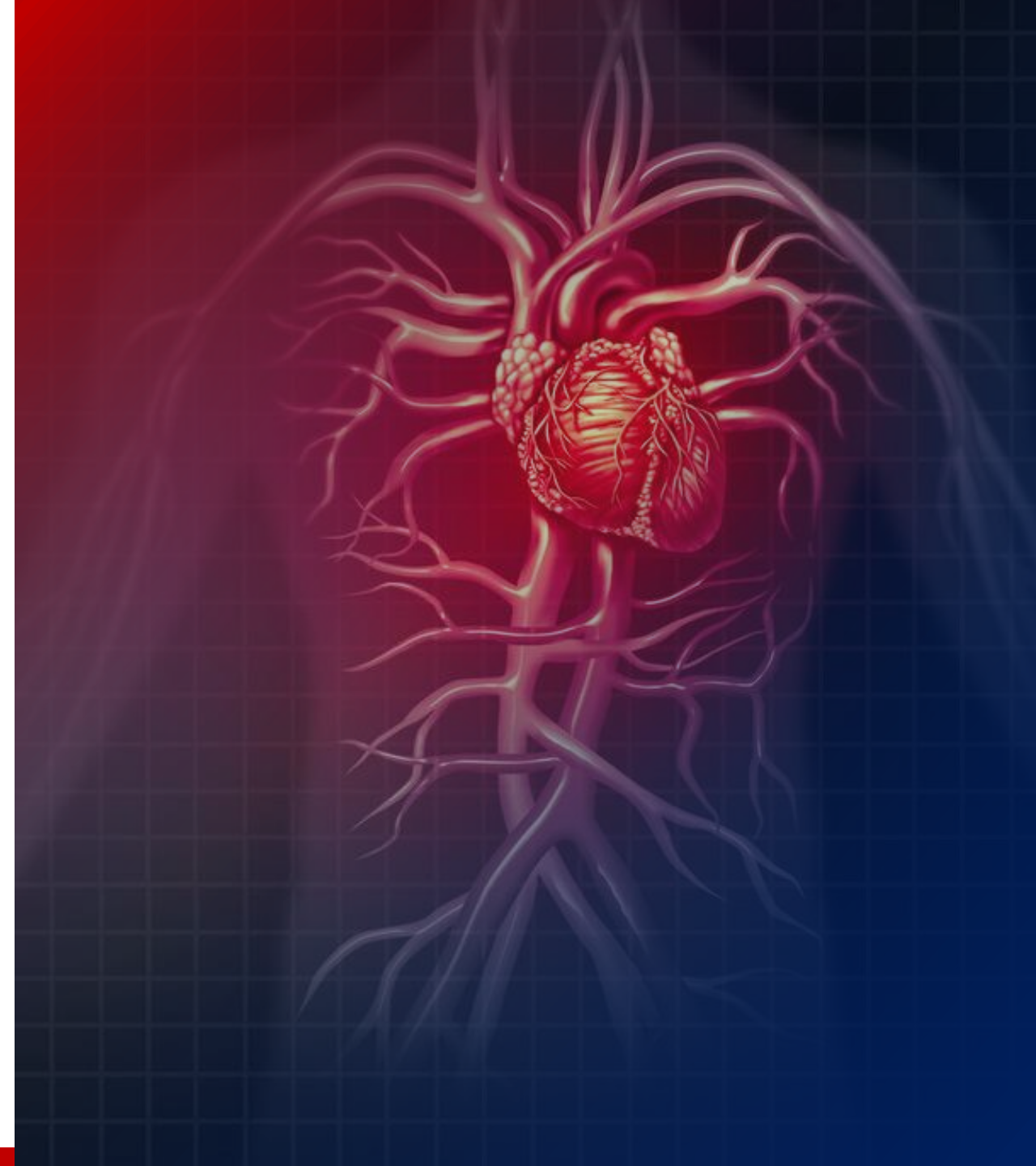
**Guided By:**

Dr. Neelam Shobha Nirala

# CONTENT

---

1. Introduction
2. Risk Factors
3. Dataset
4. About Machine Learning
5. Types of ML
6. Algorithm(methods)
7. Methodology
8. Output
9. Implementation In Real Life
10. Advantage/Drawback
11. Conclusion
12. References



# 1. INTRODUCTION

- Cardiovascular diseases (CVDs) are the leading cause of death globally.
- An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths.
- Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better.
- Heart disease is also referred to as a "silent killer" because it causes death without causing noticeable symptoms.
- Early detection of cardiac disease is critical for implementing lifestyle modifications in high-risk people and, as a result, reducing consequences.
- This project tries to predict future heart illness by evaluating patient data and using machine-learning algorithms to classify whether they have heart disease or not.

## CARDIOVASCULAR DISEASE THE WORLD'S NUMBER 1 KILLER

Cardiovascular diseases are a group of disorders of the heart and blood vessels, commonly referred to as **heart disease** and **stroke**.

**18.6** deaths every year from CVD  
**MILLION**

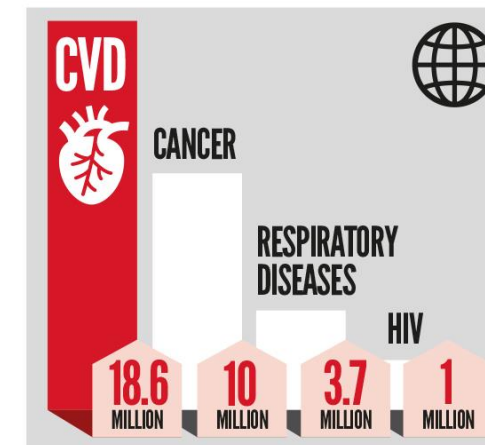


**33%** of all global deaths



**>75%** of CVD deaths take place in low- and middle-income countries

### GLOBAL CAUSES OF DEATH



### RISK FACTORS FOR CVD



Sources: World Health Organization; IHME, Global Burden of Disease

info@worldheart.org  
www.worldheart.org

f worldheartfederation  
worldheartfed  
worldheartfederation



## 2. RISK FACTORS

---

Risk factors for developing heart disease include:

- **Age** - Growing older increases your risk of damaged and narrowed arteries and a weakened or thickened heart muscle.
- **Sex** - Men are generally at greater risk of heart disease. The risk for women increases after menopause.
- **Family history** - A family history of heart disease increases your risk of coronary artery disease.
- **Smoking** - Heart attacks are more common in smokers than in nonsmokers.
- **Diabetes** - Diabetes increases your risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure.
- **High blood cholesterol levels** - High levels of cholesterol in your blood can increase the risk of plaque formation and atherosclerosis.
- **Obesity** - Excess weight typically worsens other heart disease risk factors.
- **Stress** - Unrelieved stress may damage your arteries and worsen other risk factors for heart disease.

# 7 STEPS TO A HEALTHIER HEART

You don't have to make big changes to reduce your heart attack and stroke risk. Here are 7 healthy habits that could save your life:



### 3. DATASET

---

This dataset consists of 12 attributes and 1189 rows.

- **Age:** represent the age of a person.
- **Sex:** describe the gender of person.  
(0-Female, 1-Male)
- **CP:** represents the severity of chest pain patient is suffering.
- **Trestbps:** resting blood pressure.
- **Chol:** It shows the cholesterol level of the patient.
- **FBS:** It represent the fasting blood sugar in the patient.
- **Restecg:** resting electrocardiograph results.
- **Thalach:** shows the max heartbeat of patient.
- **Exang:** used to identify if there is an exercise induced angina.
- **Oldpeak:** describes patient's depression level.
- **Slope:** describes patient condition during peak exercise. It is divided into three segments.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
5	39	1	3	120	339	0	0	170	0	0.0	1	0
6	45	0	2	130	237	0	0	170	0	0.0	1	0
7	54	1	2	110	208	0	0	142	0	0.0	1	0
8	37	1	4	140	207	0	0	130	1	1.5	2	1
9	48	0	2	120	284	0	0	120	0	0.0	1	0
10	37	0	3	130	211	0	0	142	0	0.0	1	0

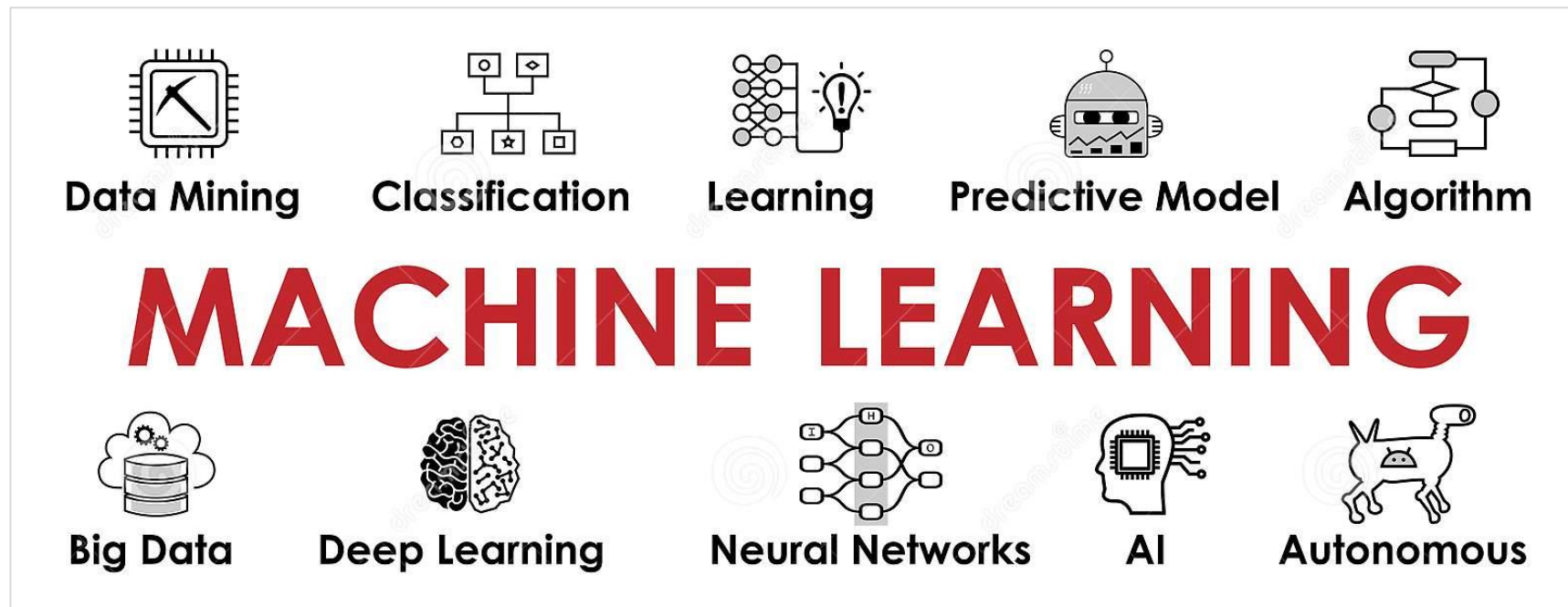
Heart Disease Prediction Data Set

## 4. MACHINE LEARNING

---

**Machine learning is an application of artificial intelligence that involves algorithms and data that automatically analyses and make decision by itself without any human intervention.**

In machine learning method we fed data to model, or a machine and model analyses the data by its own and while learning it improves it's accuracy by its own experiences and for better accuracy, we need large amount of data.

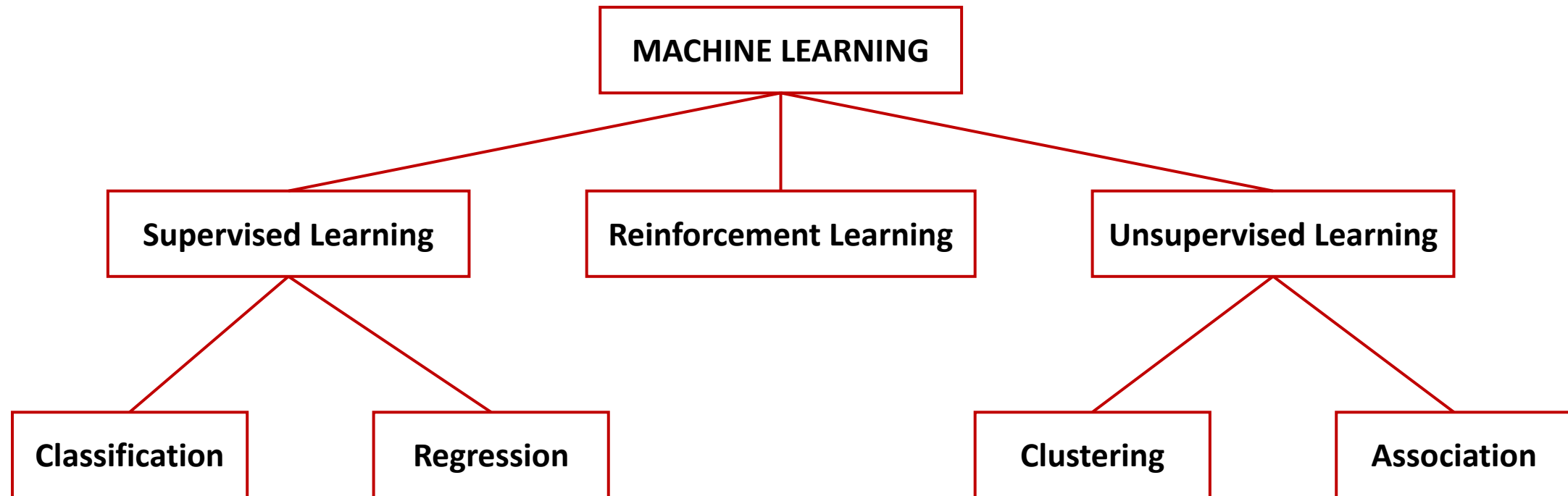


## 5. TYPES OF MACHINE LEARNING

---

**MACHINE LEARNING** algorithms are trained using three prominent methods.

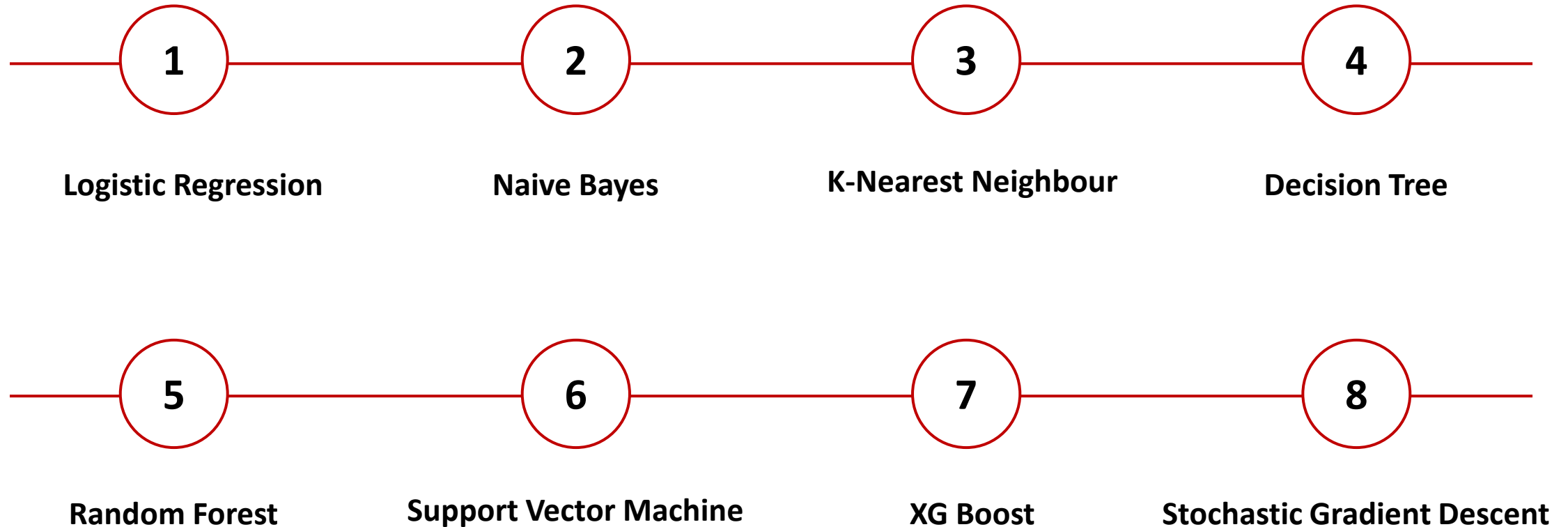
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning



## 6. ALGORITHM(METHODS)

---

**Total 8 Algorithms(models)** have been used for this Heart Disease Prediction.





## 6.1 LOGISTIC REGRESSION

- **Logistic regression** is a supervised learning algorithm used in machine learning to predict the probability of a binary outcome.
- **Sigmoid function** to map predicted prediction and their probability.
- Binary logistic regression , multinomial logistic regression , ordinal logistic regression when dependent variable are categorical then we use logistic regression .

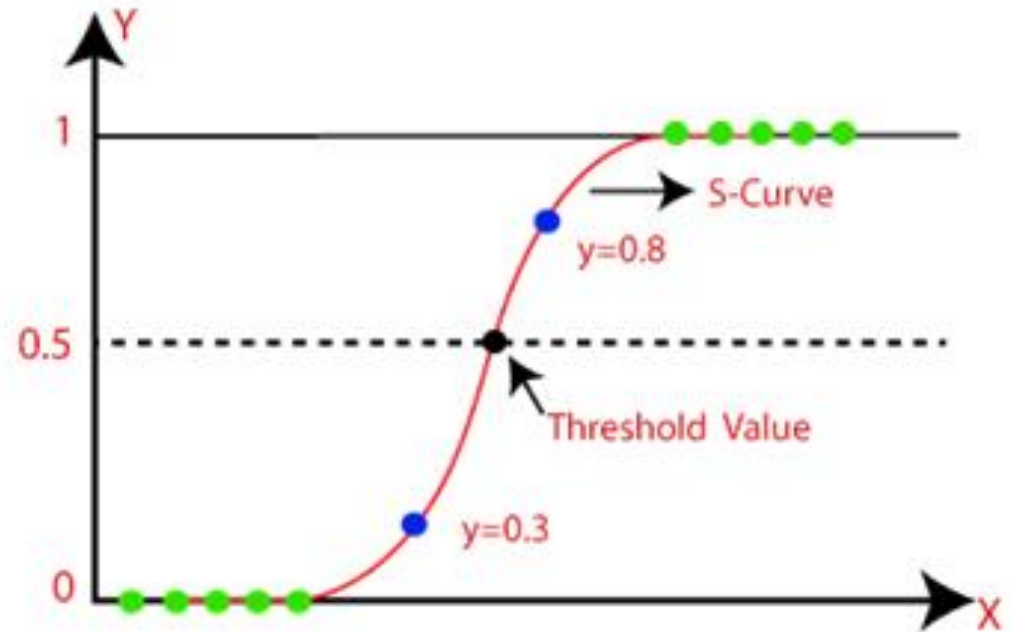


Fig. Logistic Regression

## 6.2 NAÏVE BAYES

- **Naïve Bayes** classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. In this project we have used gaussian NB algorithm.
- It is a probabilistic supervised machine learning method used for classification. It is used to determine the probability of hypothesis with prior knowledge. It depends on the **conditional probability**.
- **Bayes theorem formula is given as:**

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

Here you can find the complete PPT on NAÏVE BAYES - [link](#)

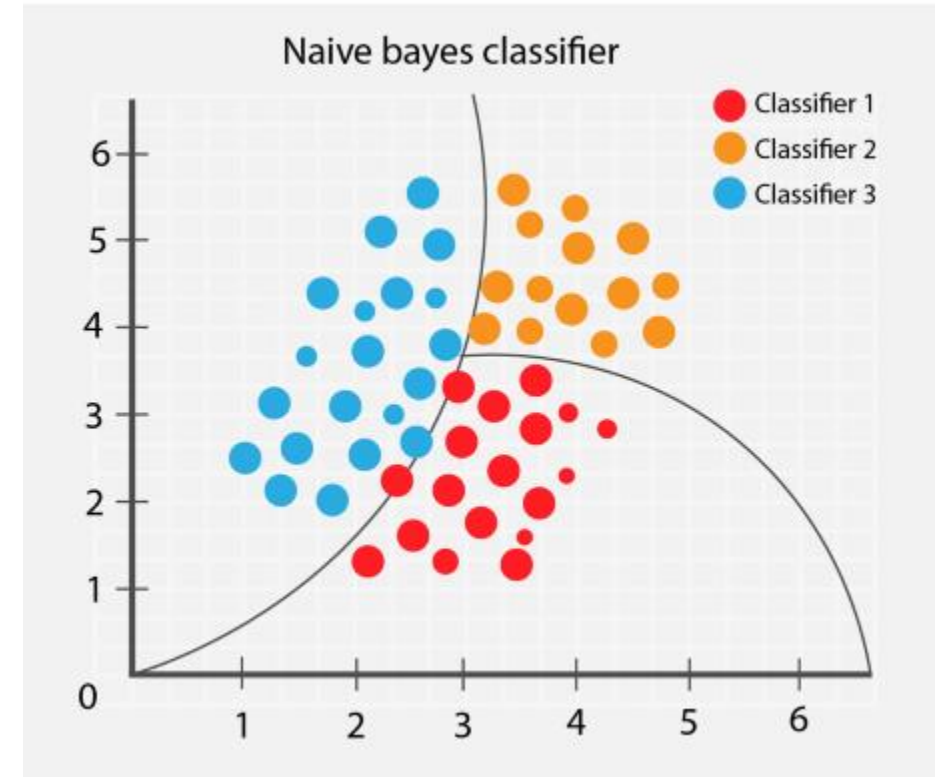
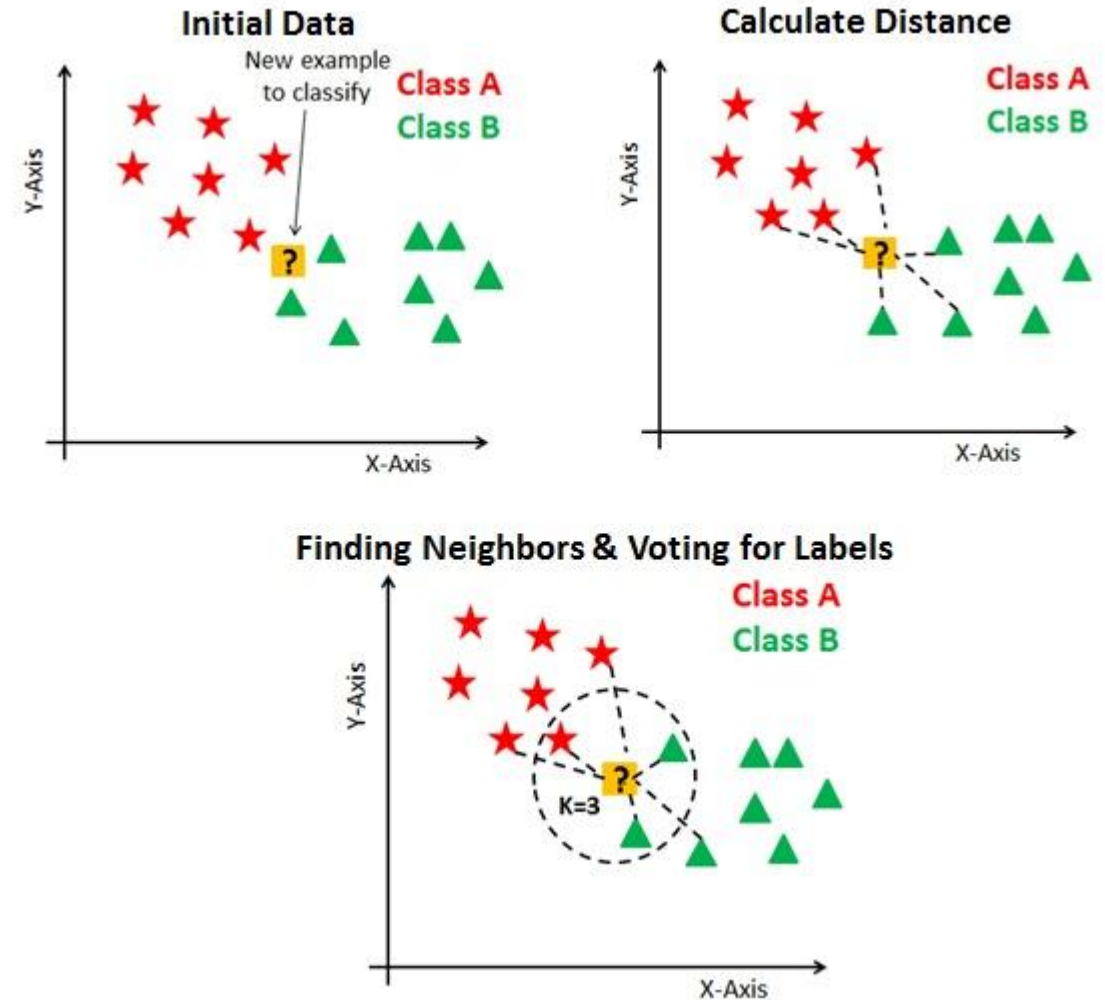


Fig. NAÏVE BAYES

## 6.3 K-NEAREST NEIGHBOUR

- **K-NN algorithm** is supervised machine learning algorithm that can be used to solve both classification and regression problems.
- It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- **Principle** - KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).
- **The algorithm's learning is**
  - Instance-based learning
  - Lazy Learning
  - Non -Parametric



Here you can find the complete PPT on KNN - [link](#)

Fig. K-NEAREST NEIGHBOUR

## 6.4 DECISION TREE

- **Decision tree** is a tree like model of decisions and their possible consequences, a way of breaking down of a complicated situation to easier to understand scenarios.
- A decision tree is a flowchart- like structure in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represent a class label (decision taken after computing all attributes). The path from root to leaf represent classification rules.

**There are mainly two algorithms to control the splitting conditions in a decision tree.**

- Entropy
- Information gain
- Gini index

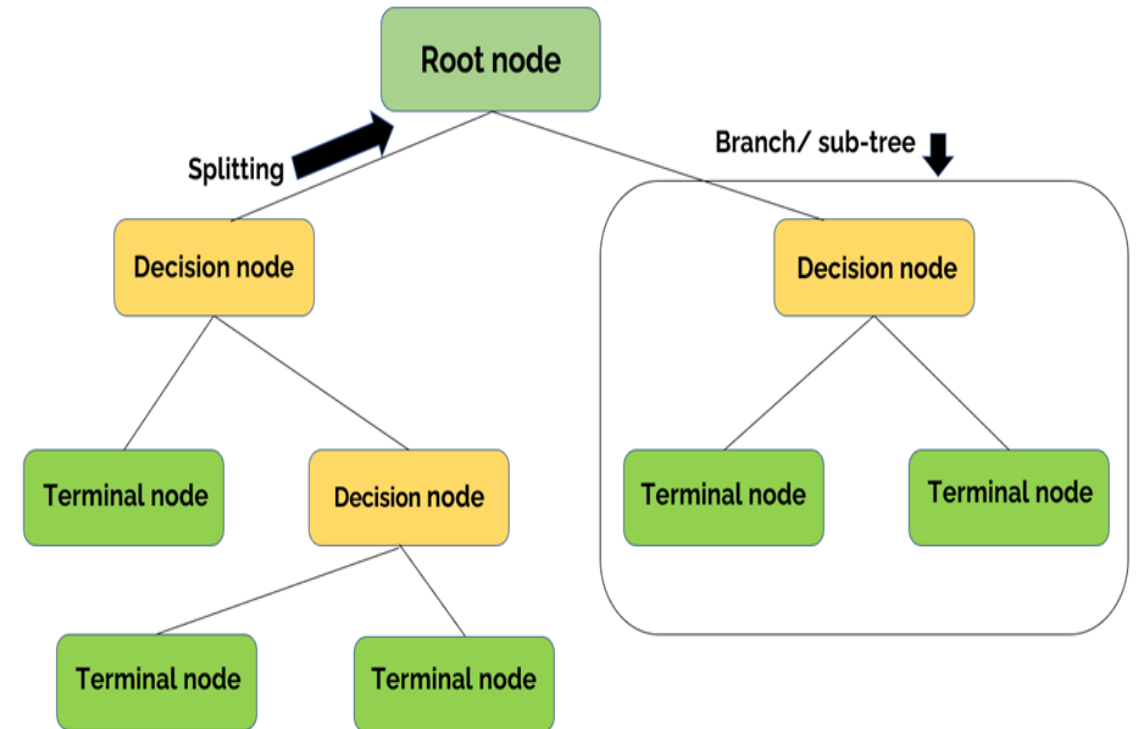


Fig. DECISION TREE

Here you can find the complete PPT on DECISION TREE - [link](#)

## 6.5 RANDOM FOREST

- It's an **Ensemble learning technique**
- Basic element of Random Forest is **Decision trees**.
- So, we can say Random forest classifier, is an extension to bagging which uses multiple decision trees to predict the output.
- **Bagging** = bootstrap + aggregation
- **Bootstrapping** means randomly draw datasets **with** replacement from the training data, each sample the same size as the original training set.

Here you can find the complete PPT on RANDOM FOREST: [LINK](#)

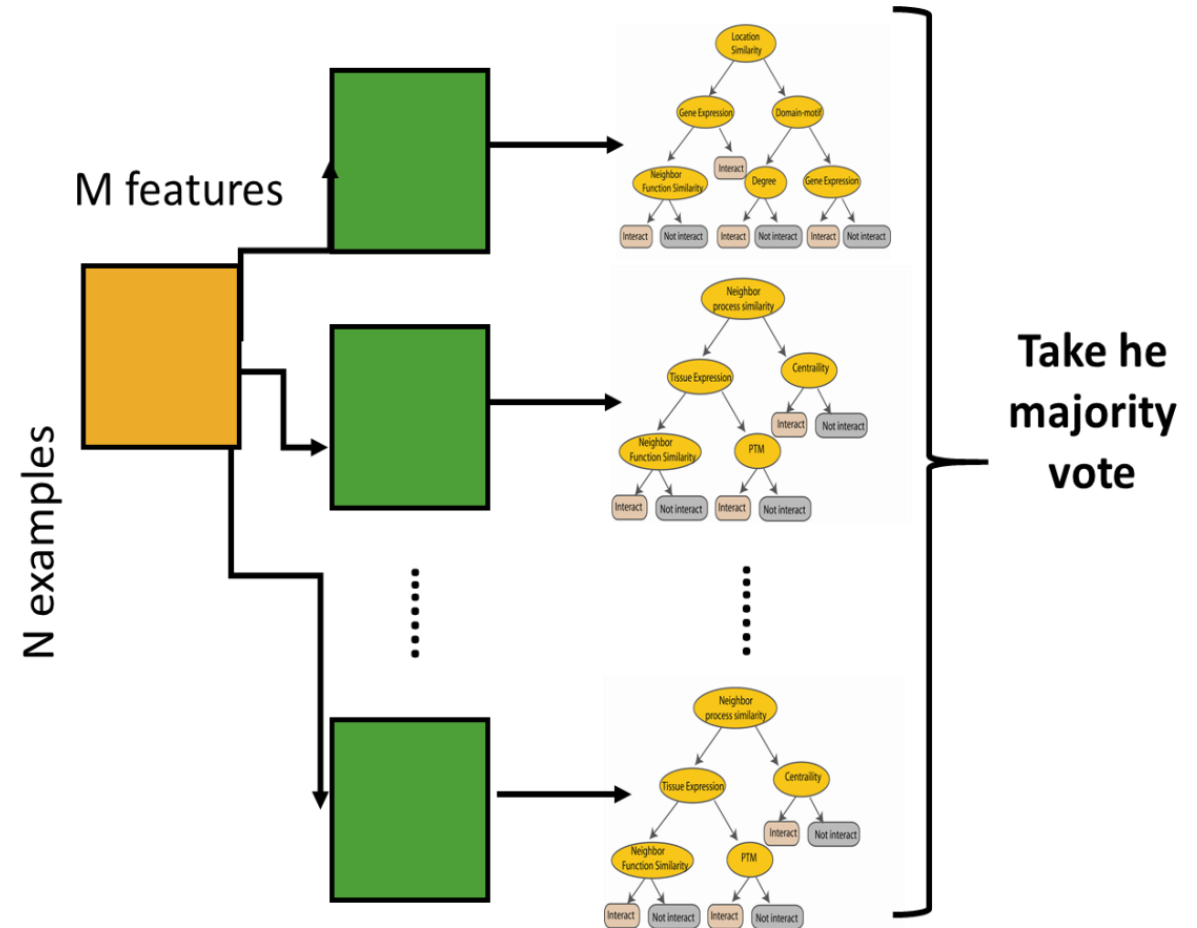


Fig. RANDOM FOREST



## 6.6 SUPPORT VECTOR MACHINE

**SVM** algorithm create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category. This best decision boundary is called a **hyperplane**. The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as **Support Vector**.

### 1. Types of Support Vector Machine:

- Linear SVM: used for linearly separable data.
- Non-linear SVM: used for non-linearly separated data

### 2. Tuning Parameters

- Margin: distance between the vectors and the hyperplane.
- Regularization: prevent the model from overfitting.
- Gamma: gamma is a parameter for non-linear hyperplanes.
- Kernel: mathematical functions for transforming data.

**3. Pros & Cons of SVM:** effective in high dimensional spaces but Takes higher time for large data set.

**4. Applications:** Face detection, classification of images, text and hypertext categorization

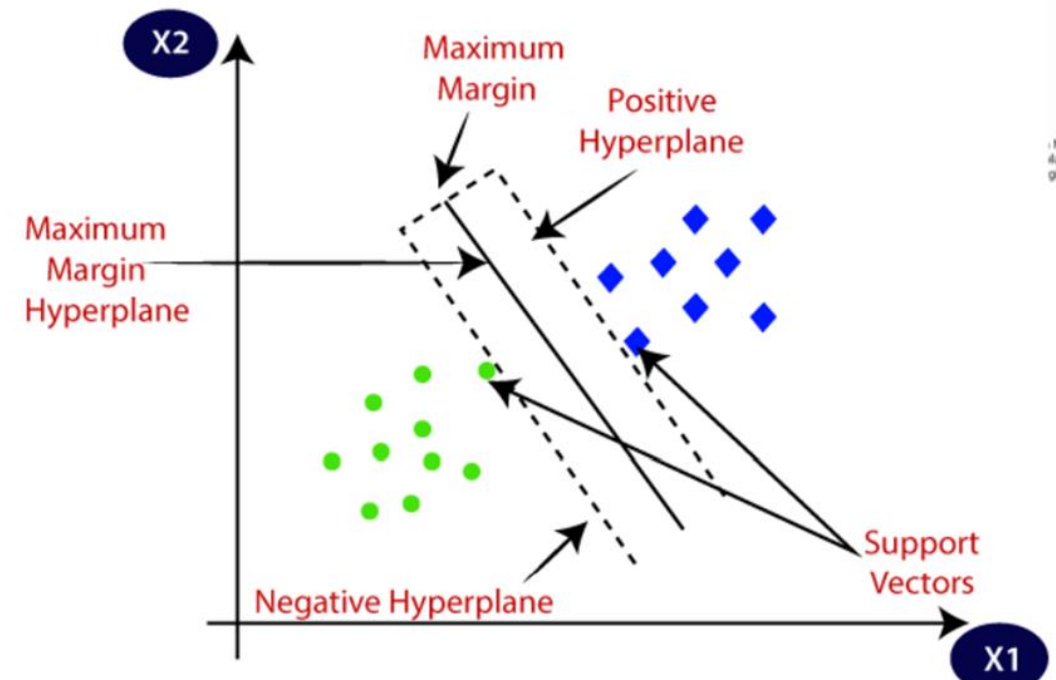


Fig. SUPPORT VECTOR MACHINE

Here you can find the complete PPT on SVM- [link](#)

## 6.7 XG BOOST

---

- **XG Boost** is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XG Boost models majorly dominate in many Kaggle Competitions.
- In this algorithm, decision trees are created in sequential form. Weights play an important role in XG Boost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these the variables are then fed to the second decision tree.
- These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

**XG Boost Features:** The library is laser-focused on computational speed and model performance, as such, there are few frills.

**Three main forms of gradient boosting are supported:**

- Gradient Boosting
- Stochastic Gradient Boosting
- Regularized Gradient Boosting

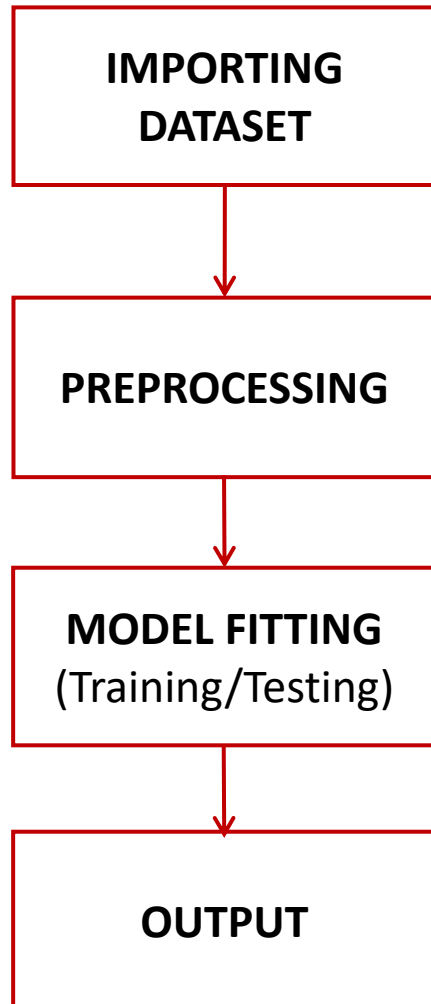
## 6.8 STOCHASTIC GRADIENT DESCENT

---

- **Stochastic gradient descent** is an optimization algorithm often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs. It's an inexact but powerful technique.
- Stochastic gradient descent is widely used in machine learning applications. Combined with back propagation, it's dominant in neural network training applications.
- **Gradient descent** is an iterative algorithm, that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function.
- Stochastic gradient descent (often abbreviated SGD) is an iterative method for optimizing an objective function with suitable smoothness properties (e.g. differentiable or sub differentiable). It can be regarded as a stochastic approximation of gradient descent optimization, since it replaces the actual gradient (calculated from the entire data set) by an estimate thereof (calculated from a randomly selected subset of the data).
- Especially in high-dimensional optimization problems this reduces the computational burden, achieving faster iterations in trade for a lower convergence rate.

## 7. METHODOLOGY

---



**Importing Dataset:** The first step in the Machine Learning process is getting data.

**Preprocessing:** Real-world data often has unorganized, missing, or noisy elements. Therefore, for Machine Learning success, after we chose our data, we need to clean, prepare, and manipulate the data.

**Model Fitting:** measure of how well a machine learning model generalizes data similar to that with which it was trained.

**Output:** Final result

## 7.1 LIBRARY USED

---

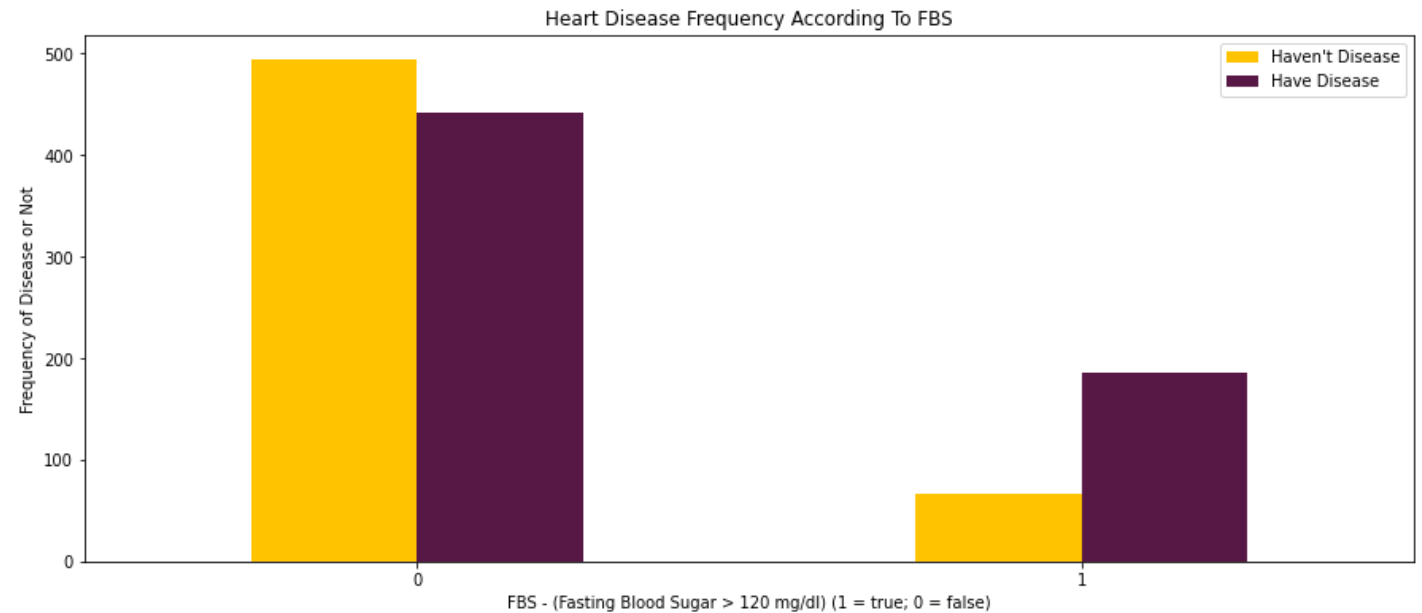
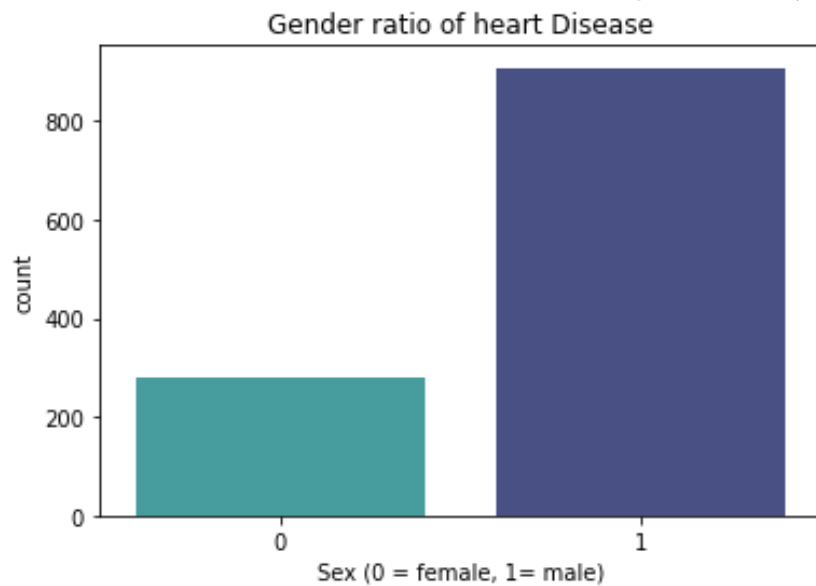
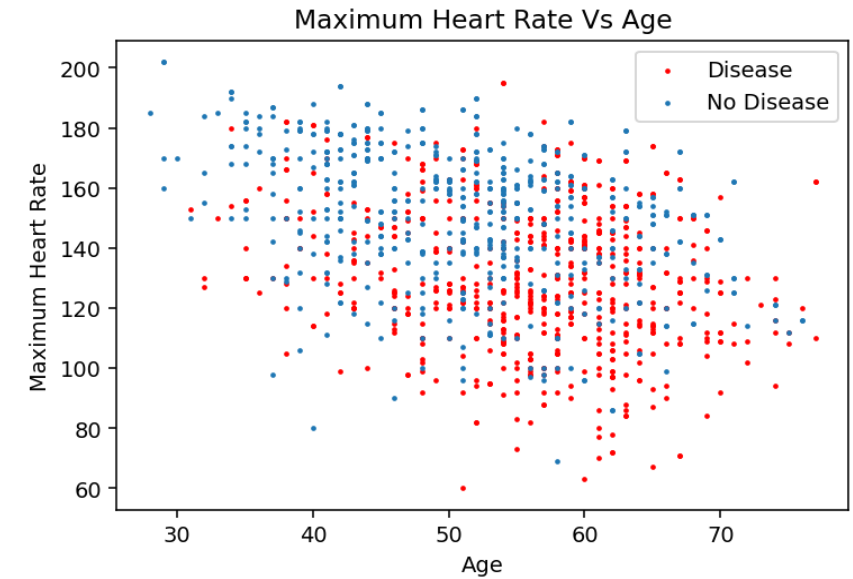
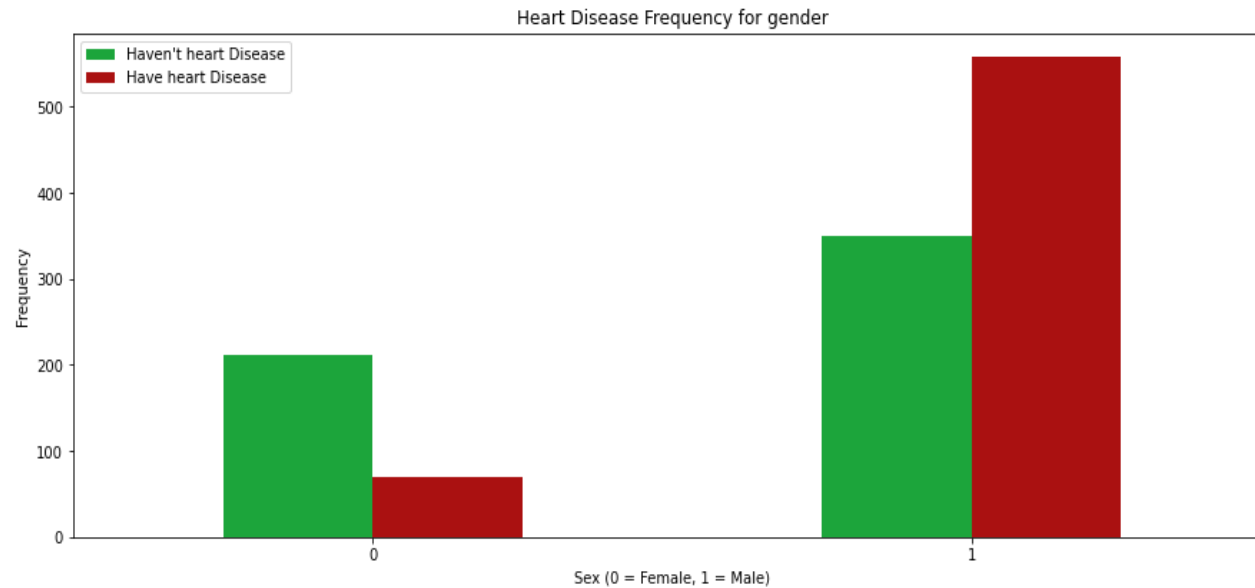
- We have used **python, pandas, matplotlib & seaborn** libraries to perform heart disease prediction for the data obtained from the Kaggle.
  - **Numpy** - for faster numerical calculation
  - **Pandas** - for data manipulation and analysis
  - **Matplotlib & Seaborn** - for Data Visualization (i.e. plot different kinds of graph)
- It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics.
- ML process starts from a pre-processing data phase followed by feature selection based on data cleaning, classification of modelling performance evaluation.

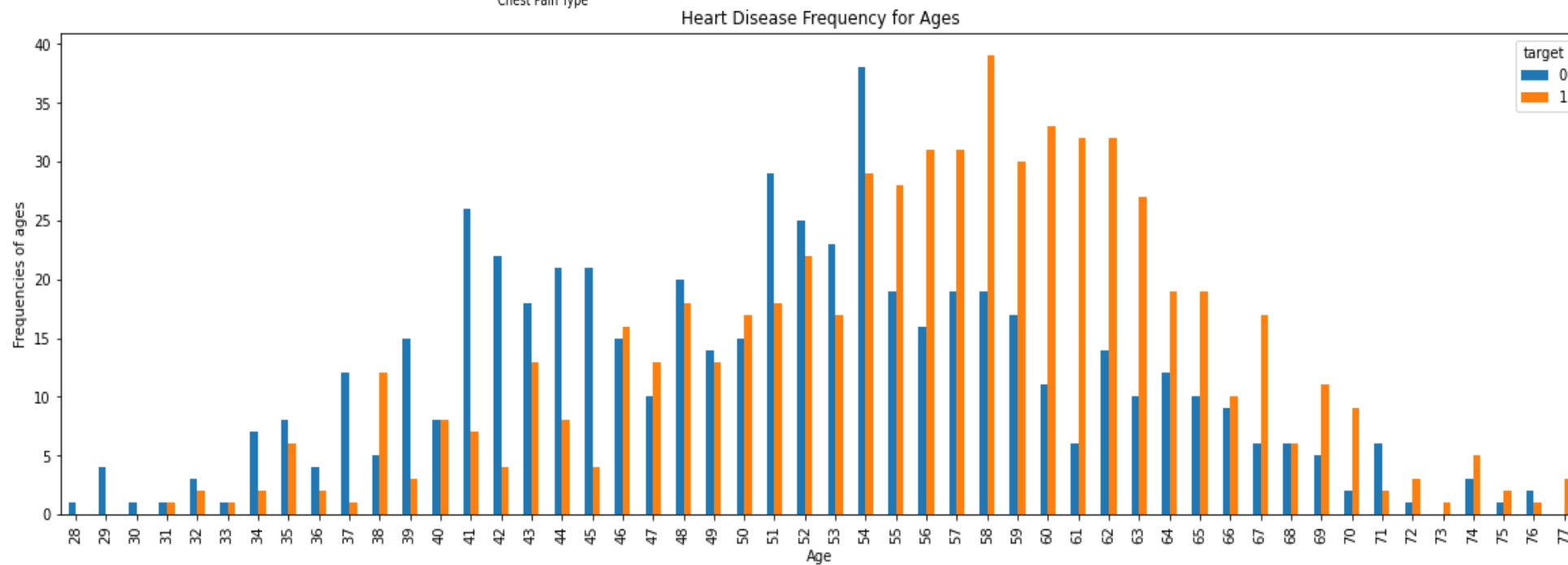
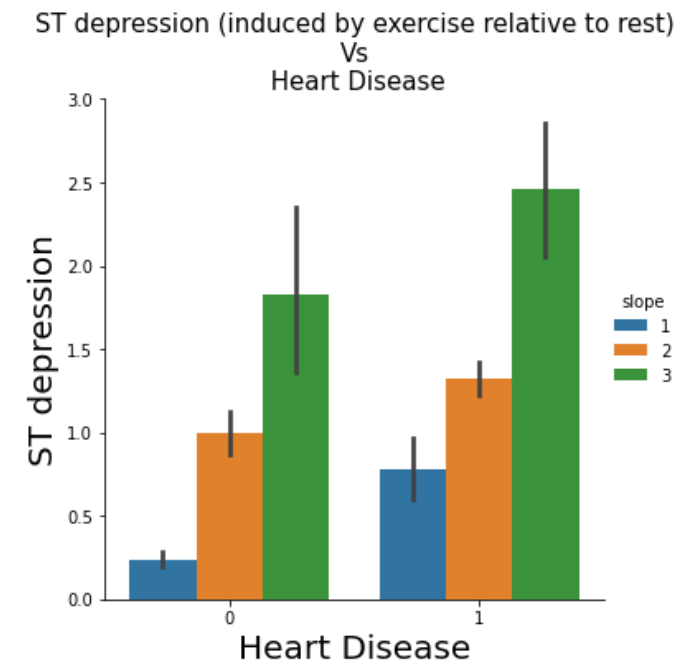
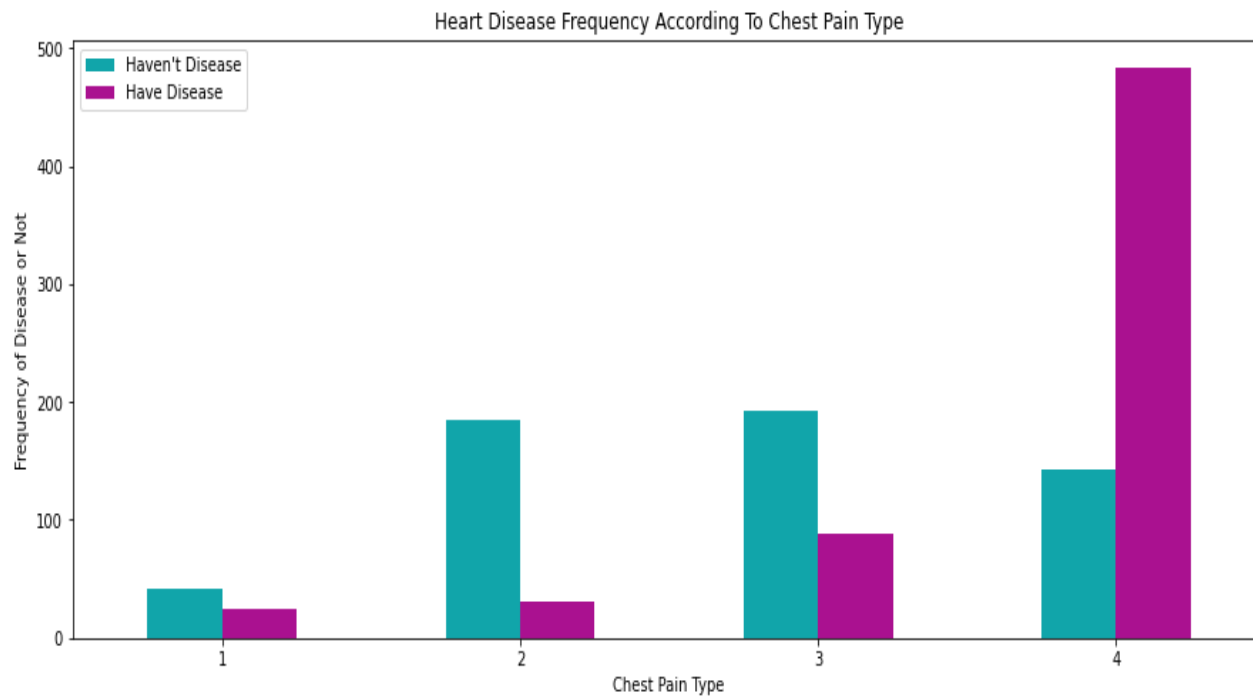
*Here you can explore our dataset which we used in this project - [dataset](#)*

*Here you can find the link of code – [project code](#)*



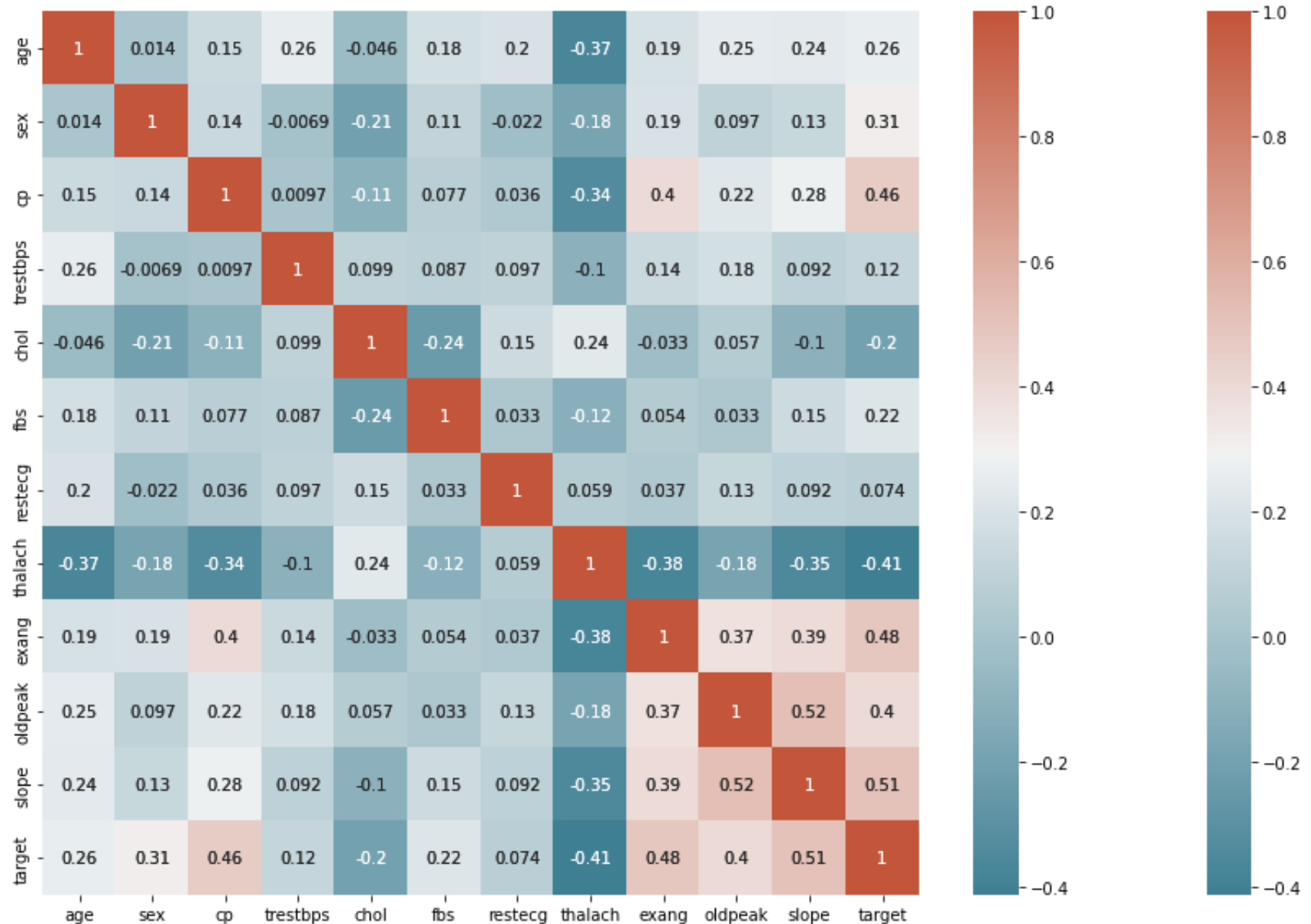
## 7.2 ANALYSIS OF DATASET





## CORRELATION MATRIX

- It is simply a table which displays the correlation between **attributes and target**.
- Positive correlation means variable and target are **directly proportion**.
- We can see there is a positive correlation between **slope & target** (our predictor).



### 7.3 Criteria to Measure performance of models

Several standard performance metrics such as accuracy, precision, F1 score and Recall in classification have been considered for the computation of performance efficiency of this model.

- **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall (Sensitivity)** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.
- **F1 score** - F1 Score is the weighted average of Precision and Recall.
- **Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

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

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

**True Positives (TP)** : actual class is yes and the value of predicted class is also yes.

**True Negatives (TN):** actual class is no and value of predicted class is also no.

**False Positives (FP):** actual class is no and predicted class is yes

**False Negatives (FN):** actual class is yes but predicted class is no.

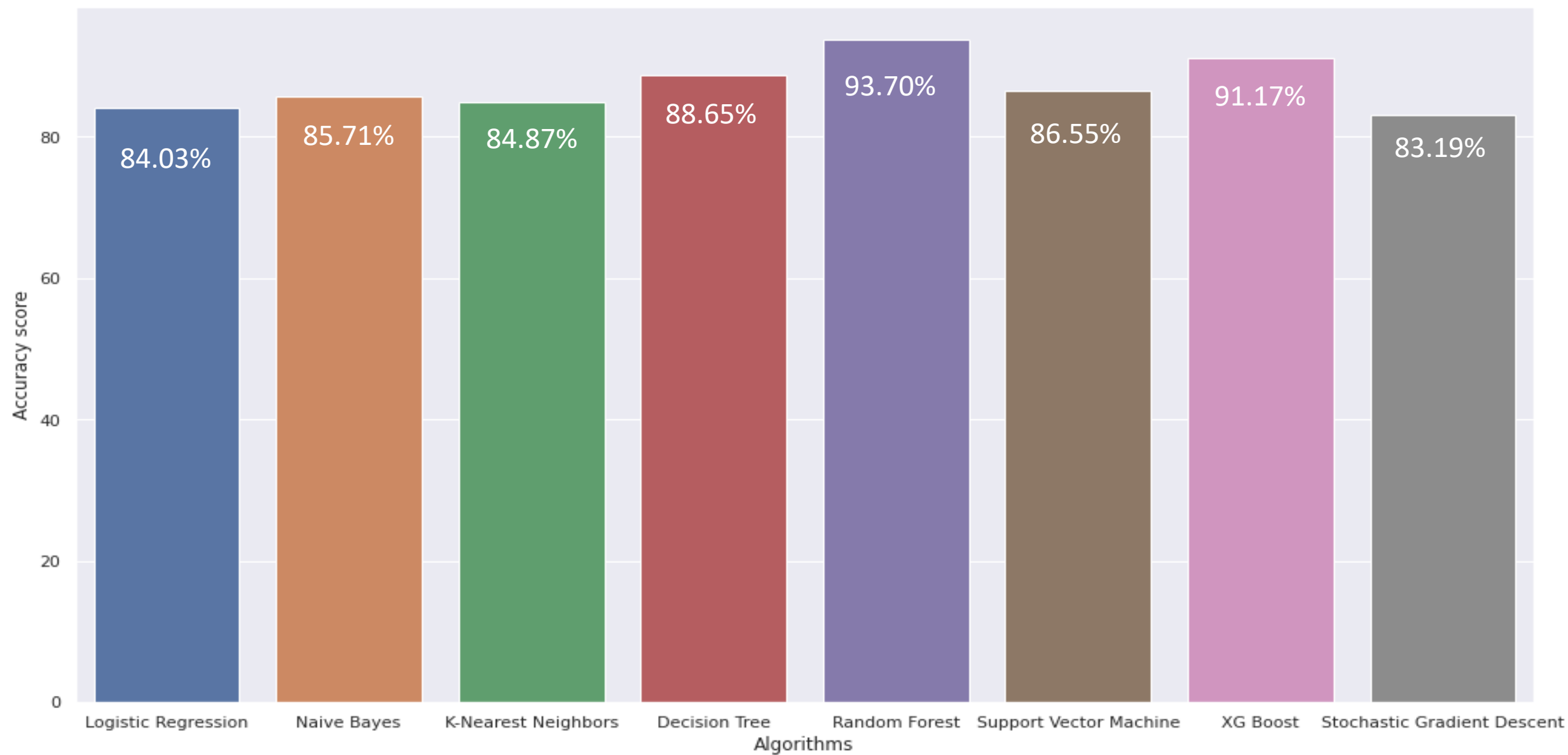
## 8. OUTPUT

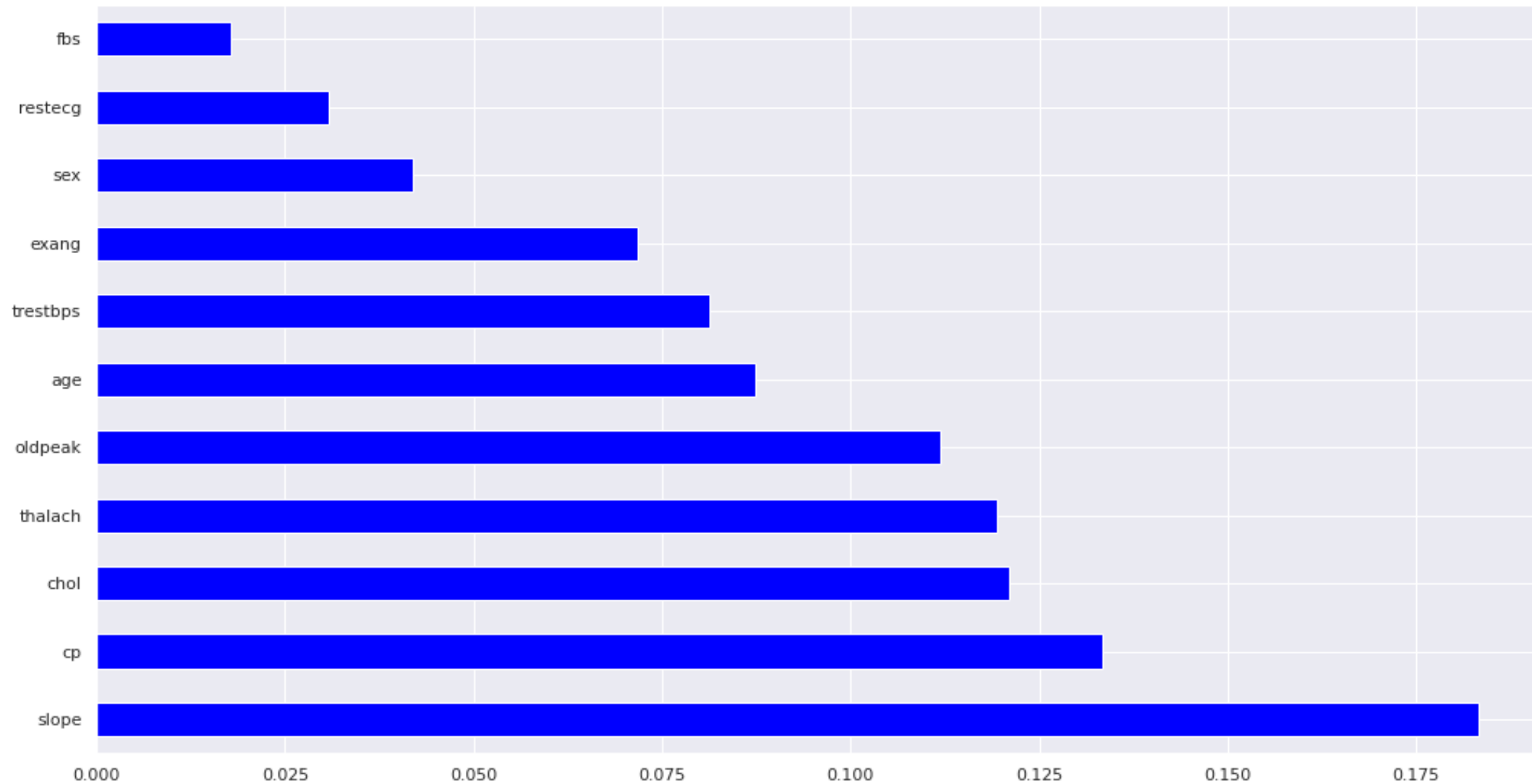
---

- This table shows analysis of different machine learning algorithms which are used in the project.
- By using **Random Forest method**, we can find
  - highest accuracy (93.70%) and
  - highest precision (94%).

Algorithm	Precision	Recall	F1-score	Accuracy(in %)
Logistic Regression	0.84	0.84	0.84	84.03
Naive Bayes Classifier	0.86	0.86	0.86	85.71
KNN	0.85	0.85	0.85	84.87
Decision Tree	0.89	0.89	0.89	88.66
Random Forest	0.94	0.94	0.94	93.70
SVM	0.87	0.87	0.87	86.55
XG Boost	0.91	0.91	0.91	91.18
SG Descent	0.83	0.83	0.83	83.19







**Feature Importance** provides a score that indicates how helpful each feature was in our model.

— HEART DISEASE PREDICTION —  
Know your disease, save your life

Please enter your Age:

⇒ 52

Please Enter Your Sex

- 0. Female
- 1. Male

⇒ 1

Enter Your Chest Pain Type:

- 1. Typical angina
- 2. Atypical angina
- 3. Non-anginal pain
- 4. Asymptomatic

⇒ 4

Enter Your Resting systolic blood pressure(in mm Hg)  
(Normal range <120 mm Hg)

⇒ 112

Enter Your Serum cholesterol in mg/dl  
(Normal Range <200 mg/dl)

⇒ 342

Enter Your Fasting blood sugar>120 mg:/dl

- 0. No
- 1. Yes

⇒ 0

Enter Your Resting electrocardiograph:

- 0. Normal
- 1. Having ST-T wave abnormality
- 2. Left ventricular hypertrophy

⇒ 1

Enter Your Maximum heart rate achieved  
(Normal range between 60-100 beats/minute)

⇒ 96

Enter Your Exercise-induced angina:

- 0. No
- 1. Yes

⇒ 1

Enter Your ST depression  
(Normal Range )

⇒ 1

Enter Your slope of the peak exercise ST segment:

- 1. Upsloping
- 2. Flat
- 3. Down sloping

⇒ 2

Select Model

- 1. Logistic Regression
- 2. Naive Bayes
- 3.K-Nearest Neighbors
- 4.Decision Tree
- 5.Random Forest
- 6.Support Vector Machine
- 7.XG Boost
- 8.Stochastic Gradient Descent

⇒ 5

You have Heart Disease

FINAL CODE OUTPUT

## 9. REAL LIFE IMPLEMENTATION

To implement this project in real life we need following portable devices.

- portable glucometer,
- cholesterol meter,
- automated BP machine,
- portable automated
- ECG machine.



**Glucometer**




**Cholesterol meter**



**Automated BPS**



**Portable ECG**

**HEART DISEASE PREDICTION**  
Know your disease, save your life

**Age:**

**Sex:** ☐ Male ☐ Female

**Chest pain type:** ☐ typical angina  
☐ atypical angina  
☐ non-anginal pain  
☐ asymptomatic


**Resting blood pressure:**

**Serum cholesterol:**

**Fasting blood sugar > 120 mg/dl:** ☐ Yes ☐ No

**Resting electrocardiographic:**  
☐ Normal  
☐ ST-T wave abnormality  
☐ Left ventricular hypertrophy

**Maximum heart rate:**

**HEART DISEASE PREDICTION**  
Know your disease, save your life

**Maximum heart rate:**

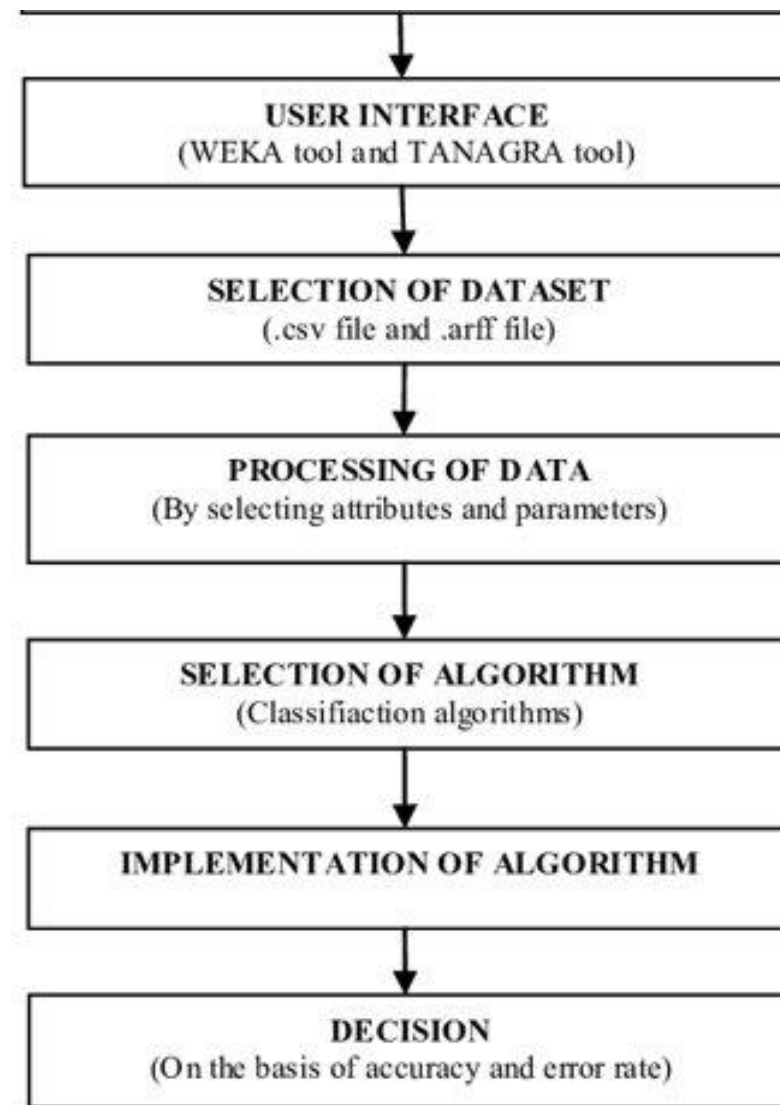
**Exercise induced angina:** ☐ Yes ☐ No

**ST depression:**

**Slope of the peak exercise ST segment:**  
☐ Upsloping  
☐ Flat  
☐ Down Sloping

**Check Now**

Result will shown here





## 10. ADVANTAGES AND DRAWBACKS

---

### ADVANTAGES

- Early prediction of heart disease can be done.
- The cost of medication will be minimized.
- Better performance.

### DRAWBACKS

- Accuracy Issues because a computerized system alone does not ensure accuracy.
- The system is not fully automated, it needs data from user for full diagnosis.
- We need more data to accurately predict heart disease .

## 11. CONCLUSION

---

- Different machine learning algorithms have been learnt and we discovered how several factors influence **cardiovascular disease (CVD)**.
- We picked eight algorithms to test on a data set, and the results were excellent, and the methods were quite accurate in predicting heart disease.
- Some of the methods have low accuracies ,to improve accuracy, we hope to require more data set because 1189 instances of data set are not sufficient to do an excellent job.
- In the future, We will try to improve our machine learning model by adding additional attributes so that we can precisely predict heart disease.
- We'll attempt to create a web application that we can use in real life.

## 12. RESOURCES

---

- <https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.114.008729>
- <https://www.giroadmedical.eu/colson-cardiopocket-cms-8-single-channel-portable-ecg.html>
- [https://www.who.int/health-topics/cardiovascular-diseases#tab=tab\\_1](https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1)
- <https://www.who.int/india/health-topics/cardiovascular-diseases>
- <https://tools.acc.org/ascvd-risk-estimator-plus/#!/calculate/estimate/>

THANK YOU

---