

# HEART DISEASE PREDICTION





# INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease.

The load of cardiovascular disease is rapidly increasing all over the world from the past few years.

Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk

The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.



# PROBLEM DEFINITION

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human.

Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients everyday in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.



# DATA SETS

The data set used is Heart Disease UCI which is taken from kaggle.

The dataset contains 14 columns such as :

- Age
- Sex
- Chest Pain Type
- Resting Blood Pressure
- Serum Cholesterol in mg/dl
- Fasting Blood Sugar
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect



# Essential Libraries:

numpy

pandas

matplotlib

Seaborn

warnings



# Steps in Prediction:

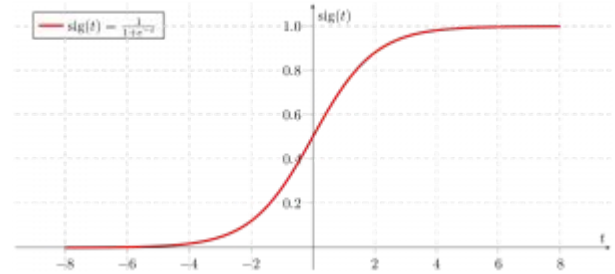
- Importing essential libraries
- Importing and understanding our dataset
- Exploratory Data Analysis (EDA)
- Train Test split
- Model Training
- Model Evaluation
- Building a Predictive System



## ALGORITHMS USED:

1. Logistic Regression (Scikit-learn)
2. Naive Bayes (Scikit-learn)
3. Support Vector Machine (Linear) (Scikit-learn)
4. K-Nearest Neighbours (Scikit-learn)
5. Decision Tree (Scikit-learn)
6. Random Forest (Scikit-learn)
7. XGBoost (Scikit-learn)
8. Artificial Neural Network with 1 Hidden layer (Keras)

# Logistic Regression

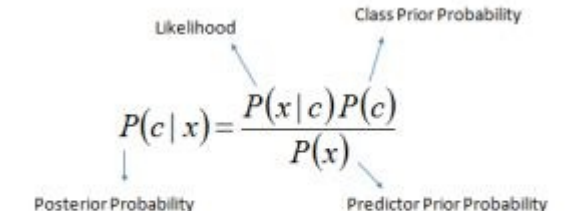


Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output),  $y$ , can take only discrete values for a given set of features(or inputs),  $X$ .

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.



# Naive Bayes



The diagram shows the Naive Bayes formula with labels pointing to its components:

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

Labels and arrows:

- Likelihood** points to  $P(x | c)$
- Class Prior Probability** points to  $P(c)$
- Posterior Probability** points to  $P(c | x)$
- Predictor Prior Probability** points to  $P(x)$

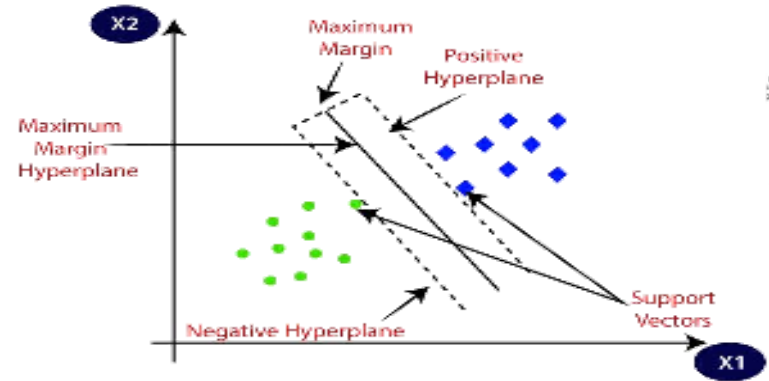
$$P(c | \mathbf{X}) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

The most important assumption that Naive Bayes makes is that all the features independent of each other. You might also need to convert the continuous variables discrete variables.

It is fast, intuitive, and is used for text classification tasks. Since it can be used for classification as well, it is considered a very versatile and flexible classifier.

# Support Vector Machine



The goal of **Support Vector Machine(svm)** the algorithm is to 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 in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane



## k-nearest neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slower as the size of that data in use grows.

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).



# XG BOOST

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems

XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems. The evidence is that it is the go-to algorithm for competition winners on the Kaggle competitive data science platform.

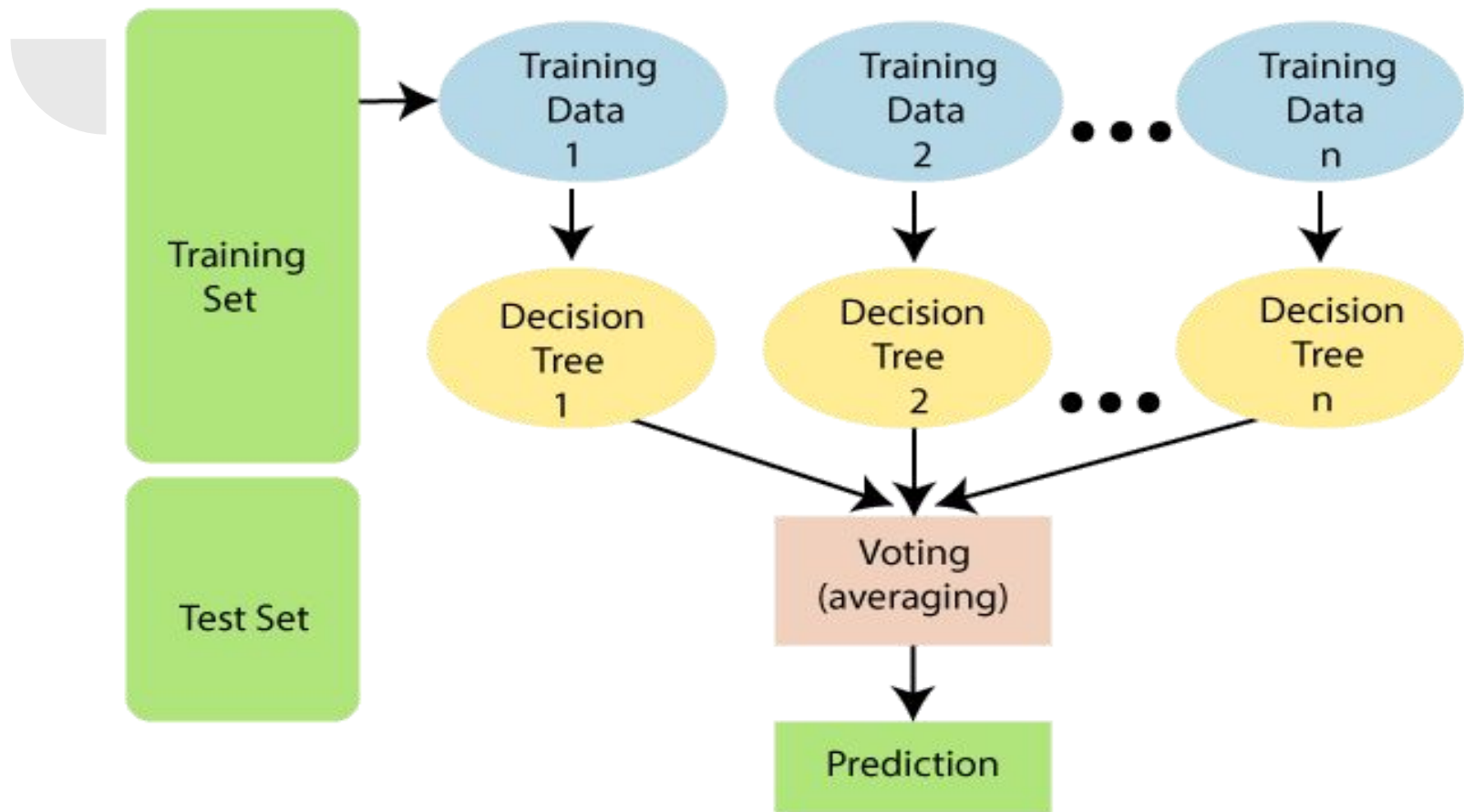
It is second most priority algorithm used in heart disease prediction



# Random Forest:

Random forest is a *Supervised Machine Learning Algorithm* that is *used widely in Classification and Regression problems*. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing *continuous variables* as in the case of regression and *categorical variables* as in the case of classification. It performs better results for classification problems.



# Result:

Random Forest Algorithm gives the Highest Accuracy of 95%

