**Cardiovascular Risk prediction**

**Saurabh Shinkar**

**Data science trainee,**

**Alma Better, Bangalore**

**Abstract:**

### The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

### The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

We will analyse each feature and understand about the different factors in the data which affects our target variable. We will apply different machine learning algorithms for classification to our data which will help to understand which model fit the best and then from that model we can estimate target variable.

**Problem Statement**

### The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

**Data Description**:

### Demographic:

• **Sex**: male or female ("M" or "F")

•**Age**: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) Behavioural

• **is smoking**: whether or not the patient is a current smoker ("YES" or "NO")

• **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day. (Can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

### Medical(history):

• **BP Meds**: whether or not the patient was on blood pressure medication (Nominal)

• **Prevalent Stroke**: whether or not the patient had previously had a stroke (Nominal)

• **Prevalent Hyp**: whether or not the patient was hypertensive (Nominal)

• **Diabetes**: whether or not the patient had diabetes (Nominal)

### Medical(current):

• **Tot Chol**: total cholesterol level (Continuous)

• **Sys BP**: systolic blood pressure (Continuous)

• **Día BP**: diastolic blood pressure (Continuous)

• **BMI**: Body Mass Index (Continuous)

• **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

• **Glucose**: glucose level (Continuous)

### Predict variable (desired target)

• 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”) - DV

* **Steps involved:**
* **Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. That’s what we have tried to done.

* **Making data in proper format**

We made the Date column in proper format. We removed less correlated features with target variable. Also removed independent feature which are strongly correlated and kept only one of them.

* **Analyzing each feature separately**

For numerical feature we look at the distribution of each feature, through Boxplot and Distribution Plot.

* **Fitting different models**

For modelling we tried various classification algorithms like:

➢ Logistic Model

➢ Decision Tree

➢ Decision Tree with Hyperparameter Tuning

➢ Random Forest

➢ Random Forest with Hyperparameter Tuning

➢ XG Boost

➢ XG Boost with Hyperparameter Tuning

➢ KNN

➢ KNN With Hyperparameter Tuning

➢ Naive Bayes Classifier

➢ Naive Bayes Classifier with Hyperparameter Tuning

➢ SVM With Hyperparameter Tuning

* **Fitting different plots**

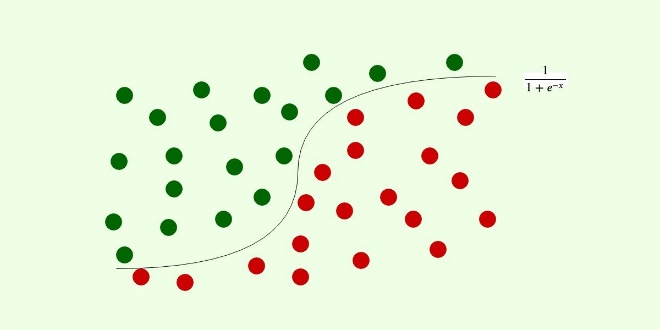
For modelling we tried various classification algorithms like:

1. **Box Plot**
2. **Pie Chart**
3. **Density Plot**
4. **Correlation Heatmap**
5. **Count Plot**
6. **Line Plot**

**3. Models:**

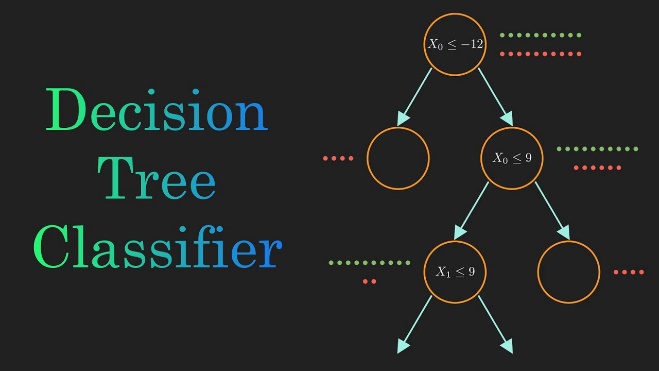
1. **Logistic:**

 Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.



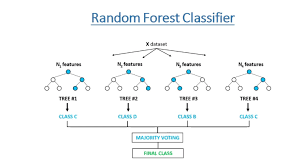
1. **Decision Tree:**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**.



1. **Random Forest:**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

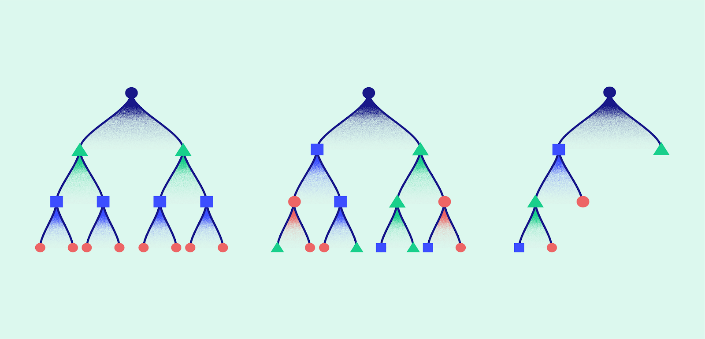


1. **XG Boost-**

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm.

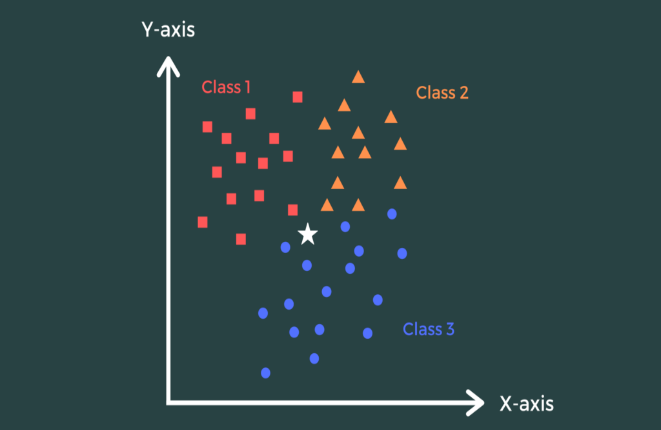
Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions.

Regression predictive modelling problems involve predicting a numerical value such as a dollar amount or a height. XG Boost can be used directly for regression predictive modelling.



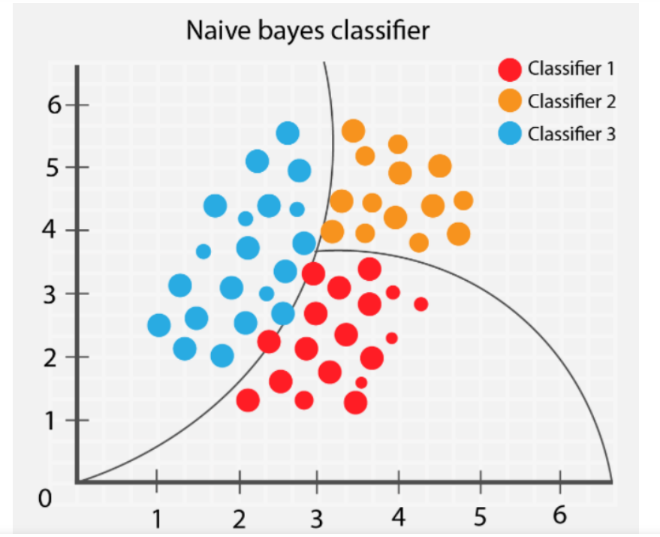
1. **KNN–**

The k-nearest neighbors’ algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.



1. **Naïve Bayes Classifier**

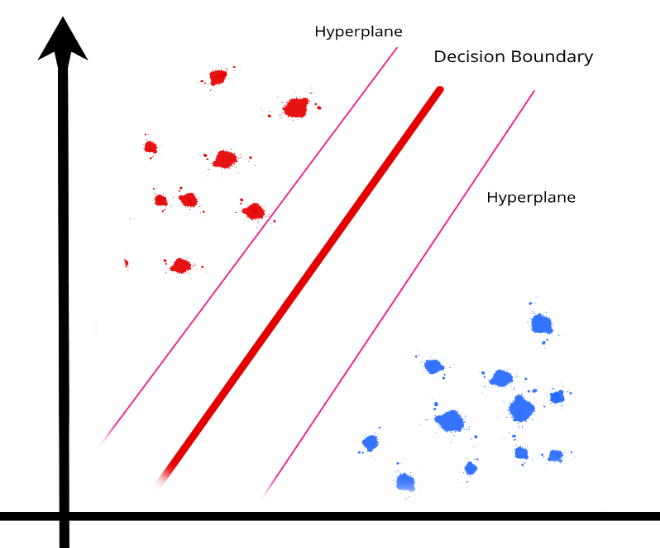
Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem.** It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

****

1. **SVM:**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM 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.

****

**Hyper parameter tuning:**

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter tuning. This also results in cross validation and in our case, we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

**Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**Conclusions-**

* Logistic Model Has Accuracy Of 74%
* Logistic with Hyperparameter Tuning Has Accuracy Of 74%
* Decision Tree Has Accuracy Of 81%
* Decision Tree with Hyperparameter Tuning Has Accuracy Of 81%
* Random Forest Has Accuracy Of 91%
* Random Forest with Hyperparameter Tuning Has Accuracy Of 92%
* XG Boost Has Accuracy Of 86%
* XG Boost with Hyperparameter Tuning Has Accuracy Of 91%
* KNN Has Accuracy Of 84%
* KNN with Hyperparameter Tuning Has Accuracy Of 84%
* Naive Bayes Has Accuracy Of 62%
* Naive Bayes with Hyperparameter Tuning Has Accuracy Of 62%
* SVM Has Accuracy Of 77%
* SVM with Hyperparameter Tuning Has Accuracy Of 85%
* From Above We Can Conclude That Random Forest with Hyperparameter Tuning and XG Boost with HyperparameterTuning Is the Best Fitted Model to Our Data.
* Random Forest with Hyperparameter Tuning has highest precision, recall and f1 score among all models.
* According to Random Forest Model SysBP, Age And education are the most important features which affects our Target variable.

