Author's Name: Sajal Sinha

Course Title: Cardio-Vascualar Disease Prediction Model

7 September 2021

CardioVascular Disease Prediction

Abstract:

In this project we were supposed to make a classifier models where the dataset is from an ongoing cardiovascular study on residents of the town of Framingham,

Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). 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, behavioural, and medical risk factors.

PROBLEM STATEMENT:

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham,Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

INTRODUCTION:

A lot of people suffer from various diseases. Some diseases like Sugar, High BP, Diabetes etc have become very common in our society. On an average around 30% youth in each state has its BMI greater than 25, which is considered as obsess. This small things which sometimes aren't considered much as threat lay foundation to major heart diseases in future. So, our model is aim to classify people goal is to predict whether the patient has a 10-year risk of future coronary heart disease which is based on various behavioural trait, medical history etc.

METHODOLOGY:

1. Data Cleaning (Outlier detection, Checking Null Values)
2. Exploratory Data Analysis
3. Data Processing (Scaling and Feature Selection)
4. Data Splitting
5. Model Training - Used various Models
6. Evaluation Metrics

DISCUSSION:

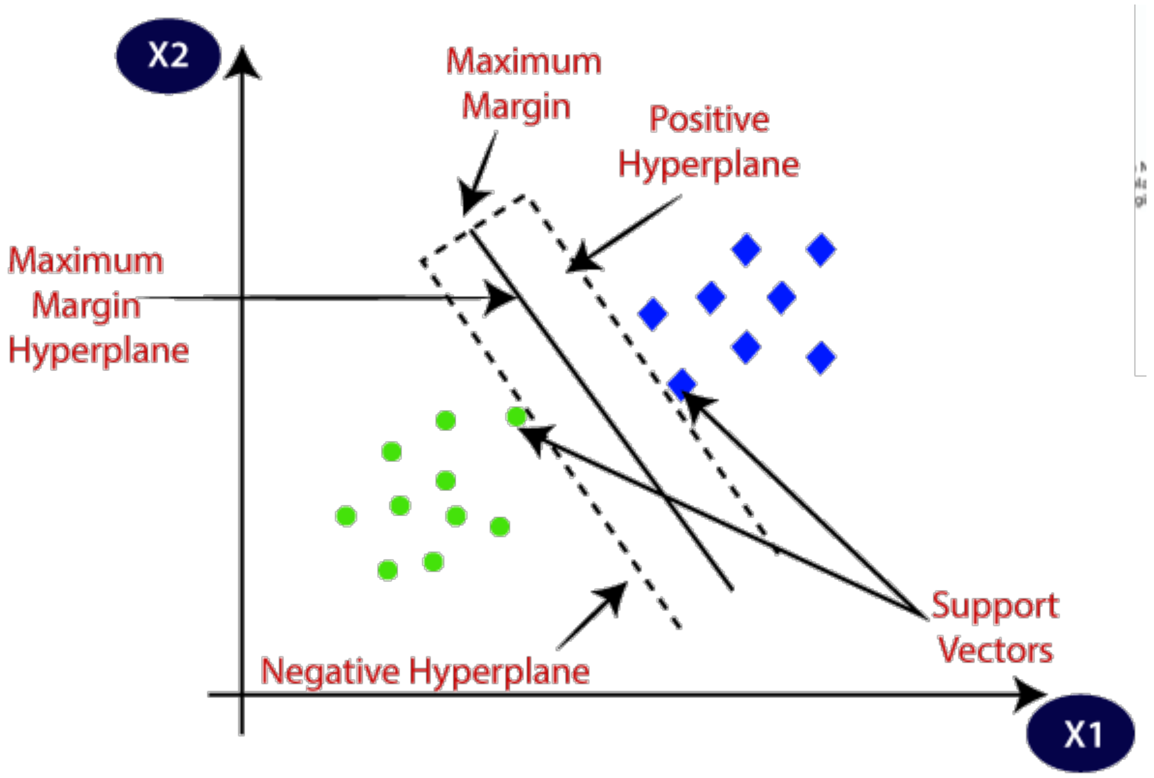
- **Logistic Regression**: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. It is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. It is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

- **KNN Classifier** : K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. 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. It 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. It can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. 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

Support Vector Machines: 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. It 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 above diagram in which there are two different categories that are classified using a decision boundary or hyperplane.



SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example *linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.*

Introduce Kernel functions for sequence data, graphs, text, images, as well as vectors. The most used type of kernel function is RBF. Because it has localized and finite response along the entire x-axis.

The kernel functions return the inner product between two points in a suitable feature space. Thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces.

Examples of SVM Kernels:

Let us see some common kernels used with SVMs and their uses:

A. Polynomial kernel

It is popular in image processing.

Equation is:

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$$

Polynomial kernel equation

where d is the degree of the polynomial.

B. Gaussian kernel

It is a general-purpose kernel; used when there is no prior knowledge about the data. Equation is:

$$k(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right)$$

C. Gaussian radial basis function (RBF)

It is a general-purpose kernel; used when there is no prior knowledge about the data.

Equation is:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma||\mathbf{x_i} - \mathbf{x_j}||^2)$$

Gaussian radial basis function (RBF)

, for:

$$\gamma > 0$$

Gaussian radial basis function (RBF)

Sometimes parametrized using:

$$\gamma = 1/2\sigma^2$$

D. Laplace RBF kernel

It is general-purpose kernel; used when there is no prior knowledge about the data.

Equation is:

$$k(x, y) = \exp\left(-\frac{||x - y||}{\sigma}\right)$$

E. Hyperbolic tangent kernel

We can use it in neural networks.

Equation is:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$$

Hyperbolic tangent kernel equation for some (not every) k>0 and c<0.

F. Sigmoid kernel

We can use it as the proxy for neural networks. Equation is

$$k(x, y) = \tanh(\alpha x^T y + c)$$

G. Bessel function of the first kind Kernel

We can use it to remove the cross term in mathematical functions. Equation is :

$$k(x, y) = \frac{J_{v+1}(\sigma||x - y||)}{||x - y||^{-n(v+1)}}$$

H. ANOVA radial basis kernel

We can use it in regression problems. Equation is:

$$k(x, y) = \sum_{k=1}^{n} \exp(-\sigma(x^k - y^k)^2)^d$$

<u>Naive Bayes Classifier:</u>

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem

Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

Types of Naive Bayes Classifier:

Multinomial Naive Bayes:

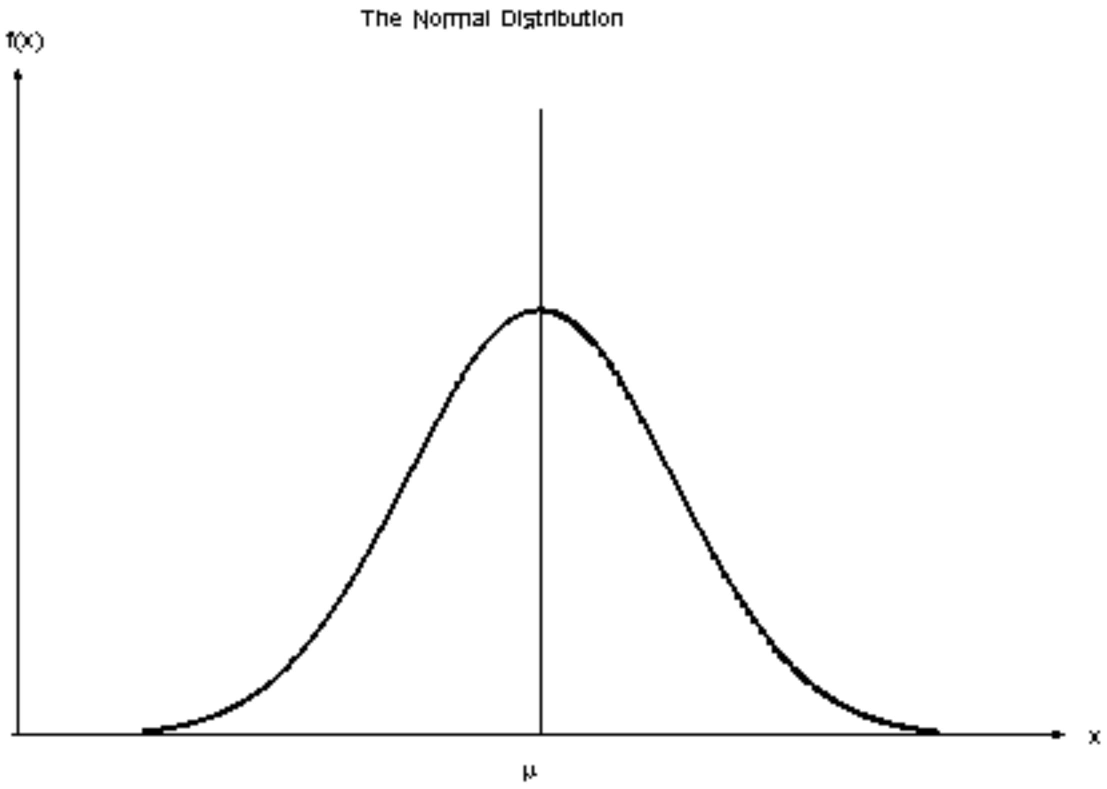
This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

Bernoulli Naive Bayes:

This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

Gaussian Naive Bayes:

When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.



Gaussian Distribution(Normal Distribution)

Since the way the values are present in the dataset changes, the formula for conditional probability changes to,

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

<u>Random forest classifier:</u>

Random forest classifiers fall under the broad umbrella of ensemble-based learning methods [30]. They are simple to implement, fast in operation, and have proven to be extremely successful in a variety of domains. The key principle underlying the random forest approach comprises the construction of many “simple” decision trees in the training stage and the majority vote (mode) across them in the classification stage. Among other benefits, this voting strategy has the effect of correcting for the undesirable property of decision trees to overfit training data. In the training stage, random forests apply the general technique known as bagging to individual trees in the ensemble. Bagging repeatedly selects a random sample with replacement from the training set and fits trees to these samples. Each tree is grown without any pruning. The number of trees in the ensemble is a free parameter which is readily learned automatically using the so-called out-of-bag error.

Much like in the case of naïve Bayes– and k-nearest neighbor–based algorithms, random forests are popular in part due to their simplicity on the one hand, and generally good performance on the other. However, unlike the former two approaches, random forests exhibit a degree of unpredictability as regards the structure of the final trained model. This is an inherent consequence of the stochastic nature of tree building. As we will explore in more detail shortly, one of the key reasons why this characteristic of random forests can be a problem in regulatory reasons—clinical adoption often demands a high degree of repeatability not only in terms of the ultimate performance of an algorithm but also in terms of the mechanics as to how a specific decision is made.

Decision Tree Classifier:

Decision Tree Classifier is a simple and widely used classification technique. It applies a straitforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time time it receive an answer, a follow-up question is asked until a conclusion about the calss label of the record is reached. The decision tree classifiers organized a series of test questions and conditions in a tree structure. The following figure [1] shows a example decision tree for predictin whether the person cheats. In the decision tree, the root and internal nodes contain attribute test conditions to separate recordes that have different characteristics. All the terminal node is assigned a class lable Yes or No. Once the decision tree has been constructed, classifying a test record is straightforward. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test. It then lead us either to another internal node, for which a new test condition is applied, or to a leaf node. When we reach the leaf node, the class lable associated with the leaf node is then assigned to the record, it traces the path in the decision tree to predict the class label of the test record, and the path terminates at a leaf node labeled NO.

K-fold Cross Validation:

Cross-Validation is just a method that simply reserves a part of data from the dataset and uses it for testing the model(Validation set), and the remaining data other than the reserved one is used to train the model. In k-fold CV, the dataset is split into ‘k’ number of subsets, k-1 subsets then are used to train the model and the last subset is kept as a validation set to test the model. Then the score of the model on each fold is averaged to evaluate the performance of the model.

Accuracy of different models:

k fold Cross Val Score

logistic=85.01%

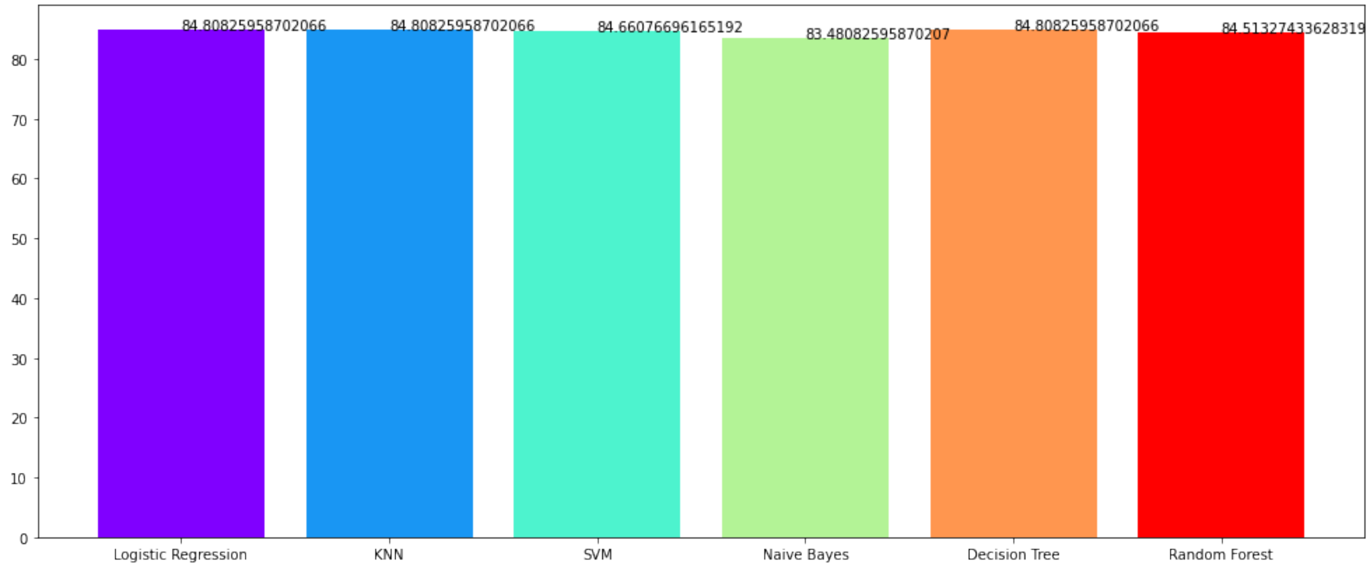
KNN =84.07%

Gaussian Naive Bayes=83.22%

SVC=84.93%

Decision Tree Classifier=83.86%

Random Forest Classifier=83.27%



Conclusion:

- Logistic Regression got better result than any model.
- Highest Number of cigratte smoked in a day is 50.
- Males consume more cigrattes than females in a day.
- People with less education are more prone to have heart disease after 10 years.
- People with less education are more prone to get addicted to smoking.
- More males are suffering from diabetes than female.
- Those who have high BP are more prone to heart disease.
- Those who have low BP are less prone to heart disease.
- Non- diabetic people smokes more

Some measures that can be taken to prevent Heart disease includes:

1. No smoking
2. Maintain Healthy daily life.
3. BMI should be checked regularly inorder to have a note of ourselves.

4. As we get older, selection of food must be done properly, so that one can control cholesterol, glucose etc.

5. Having proper cardio routine should be must. Yoga, walking or jogging are good enough.

Reference:

- https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74
- https://www.geeksforgeeks.org/naive-bayes-classifiers/
- https://www.askpython.com/python/examples/k-fold-cross-validation/\