Name: - Nisarg Patel

Roll No.: - 18BCE136

Branch: - Computer Science & Engineering

Division: - C

Batch: - C1

Heart Disease Prediction Using Machine Learning Algorithms

Introduction

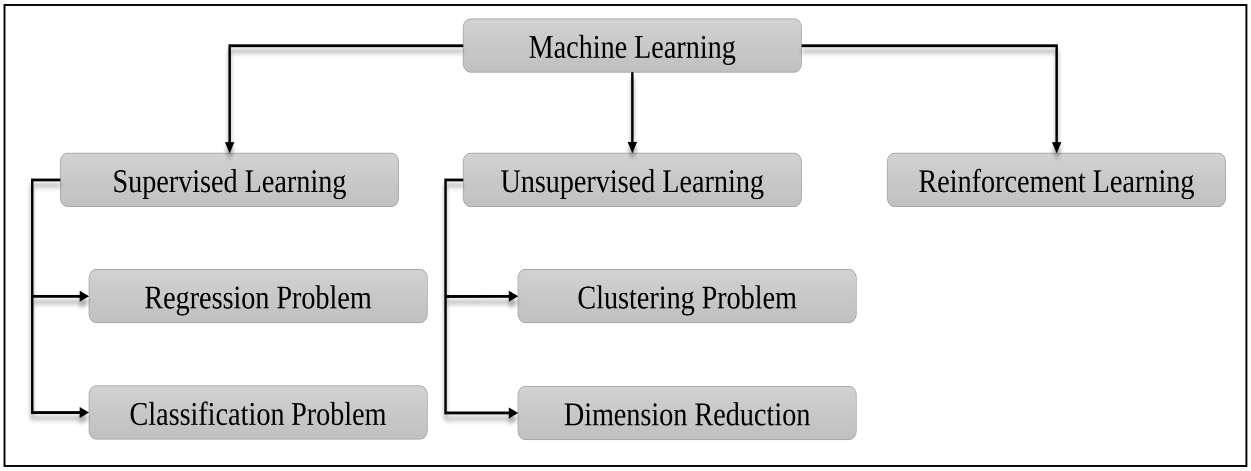
Heart plays a very important role in every part of our anatomy. Heart is an important organ of the human body and to keep the body healthy everyone should keep their heart healthy. Heart disease is the most common disease nowadays. Any kind of heart related problem is not at all desirable. If it fails to function properly, including the brain to various other body organs will not be able to work properly and due to this it leads to paralyses to brain death sometimes. To keep the heart healthy, everyone should make a habit of exercise and change their lifestyle. CVD - Cardiovascular Diseases or Heart related diseases are the main reason for the death of the significant number of people in the world not only in India over the last few decades and become the most threatening disease in the all-around. According to the World Health Organization, coronary heart associated illnesses are liable for the taking 17.7 million lives each year, 31% of all worldwide deaths. In India too, heart related diseases have ended up the main cause of mortality. Heart diseases have killed 1.7 million Indians in 2016, in line with the 2016 Global Burden of Disease Report, launched on September 15, 2017. Heart associated diseases boost the spending on fitness care and also reduce the productivity of a man or woman. So there is a need for a feasible and accurate diagnosis system to predict such kinds of disease in the human system and the system must be more precise, perfect and correct because a little error in the system may lead to fatigue problems like death of patience or Coma/Paralysis.

Medical organizations, all over the global, gather statistics on numerous health associated troubles. These data may be exploited by the usage of numerous machine mastering strategies to advantage beneficial insights. But the records obtained are very large and, many a time, these records can be very noisy. These datasets, which are too overwhelming for human minds to understand, can be without problems exploring the use of diverse system learning techniques. Thus, these algorithms have grown to be very beneficial, these days, to predict the presence or absence of heart associated illnesses appropriately. Machine learning is one of the most efficient technologies for predicting, classifying which is basically based on training and testing the model. Machine learning (ML) is the branch of the AI-Artificial Intelligence and Deep learning (DL) is a subset of the Machine Learning techniques. Artificial Intelligence is a wide range of learning algorithms where machines mimic the human abilities and from this, machine learning is the branch of specific types of functions. Machine learning systems are trained to learn the data and how to process it hence it also can be called machine intelligence. In this paper there are

5 machine learning algorithms performed and the data is processed and also comparison of all is also done. Machine learning definition says that it learns from the natural phenomenon and in for testing, there are many biological parameters available like blood pressure, cholesterol, age, sex, etc.

Machine Learning

Training and testing is the heart of Machine learning. Training is to be done based on the type of the data. System takes training directly from the data or from experience and according to that testing is to be done and accuracy is being measured. Machine learning algorithms are classified in three ways as follow:



A. *Supervised Learning*

Supervised Learning can be defined as the learning under the guidance or in the presence of the teacher. Our train dataset acts like a teacher for the model and through this dataset model learns and predicts the output for the test dataset and then accuracy of the model is measured. “Train Me” concept is applied in supervised learning. There are mainly two technique or you can say problem style on which supervised learning can be applied.

- Regression Problems
- Classification Problems

To perceive examples and measure likelihood of uninterruptible results, is a phenomenon of regression. Whereas classification problems can be considered as a grouping the train data and try to classify the test data accurately. Here is the list of the supervised machine learning algorithms:

- Linear Regression
- Logistic Regression

- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

B. *Unsupervised Learning*

In unsupervised learning there’s neither a guide nor teacher like supervised learning. In an unsupervised learning dataset is given to the model and the model itself tries to recognize some pattern and relation between data and make clusters and distribute itself. When a test dataset is given it tries to arrange in the most suitable cluster. We can say that unsupervised learning is based on the concept of “Self-Sufficient”.

For example suppose there’s cow, buffalo and zebra and when unsupervised learning is applied to this dataset it classify all three in the three different clusters on basis of their characteristics and when test dataset is given to same model it can classify them into any one of the best suitable cluster. Where supervised learning can say that there are three groups named cow, buffalo and zebra where on the same data unsupervised data can say that there’s three different clusters. Unsupervised learning can be applied to

- Clustering Problem
- Dimensionality Reduction

Unsupervised learning have following algorithms:

- k-means clustering
- PCA
- t-SNE

C. *Reinforced learning*

Reinforced learning is the specialist capacity to connect with the climate and discover the result. It depends on the "hit and trial" idea. In Reinforced learning every specialist is granted with positive and negative focuses and based on certain focuses Reinforced learning gives the dataset yield that is based on sure honors it prepared and based on this preparation plays out the testing on datasets.

Architecture of Prediction System

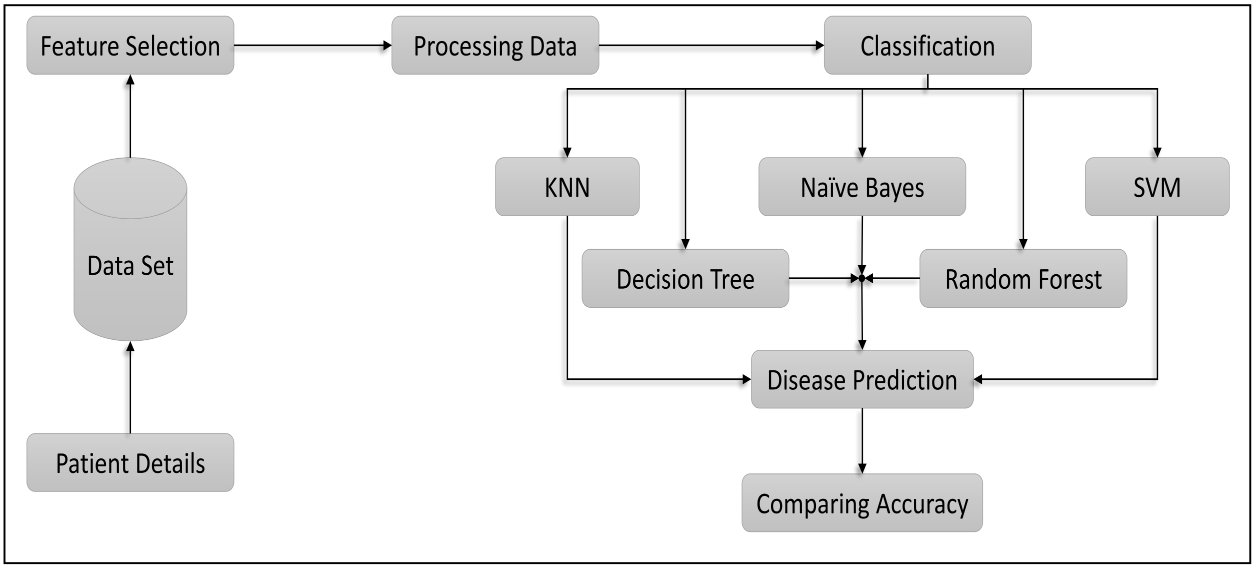


Figure shown above represents the system which is implemented on the dataset. For this demonstration we have the dataset of 1025 people from the real world and features are biological so it helps to predict the heart disease of the person. So we have a dataset and after that we are going to have a look at the feature selection part but fortunately our dataset has all the necessary data and with less correlation between each pair. After the feature selection part we have focused on the data processing part as it is the most important part of the system. We have to transform the data in such a way that the model can understand each feature easily and its effect on the output. For the same objective we use dummy variables and separate the data according to the need. After getting a dummy variable now we have to split our data into two part train dataset and the other one test dataset. Now the implementation part comes. In this particular system we are going to use 5 different algorithms on the same dataset to predict whether the patient has heart disease or not. We are using GridSearchCV to get better results of each algorithm. After training our model we are going to have a look for training accuracy and then we are going to perform our prediction on the test dataset. After getting prediction of the test dataset again we are going to compare the train accuracy and test accuracy for all algorithms.

Details of the Dataset

There are a total 14 columns in this dataset for 1025 people and here is the description of the same 14 columns and values they have with their meaning and effect to our heart.

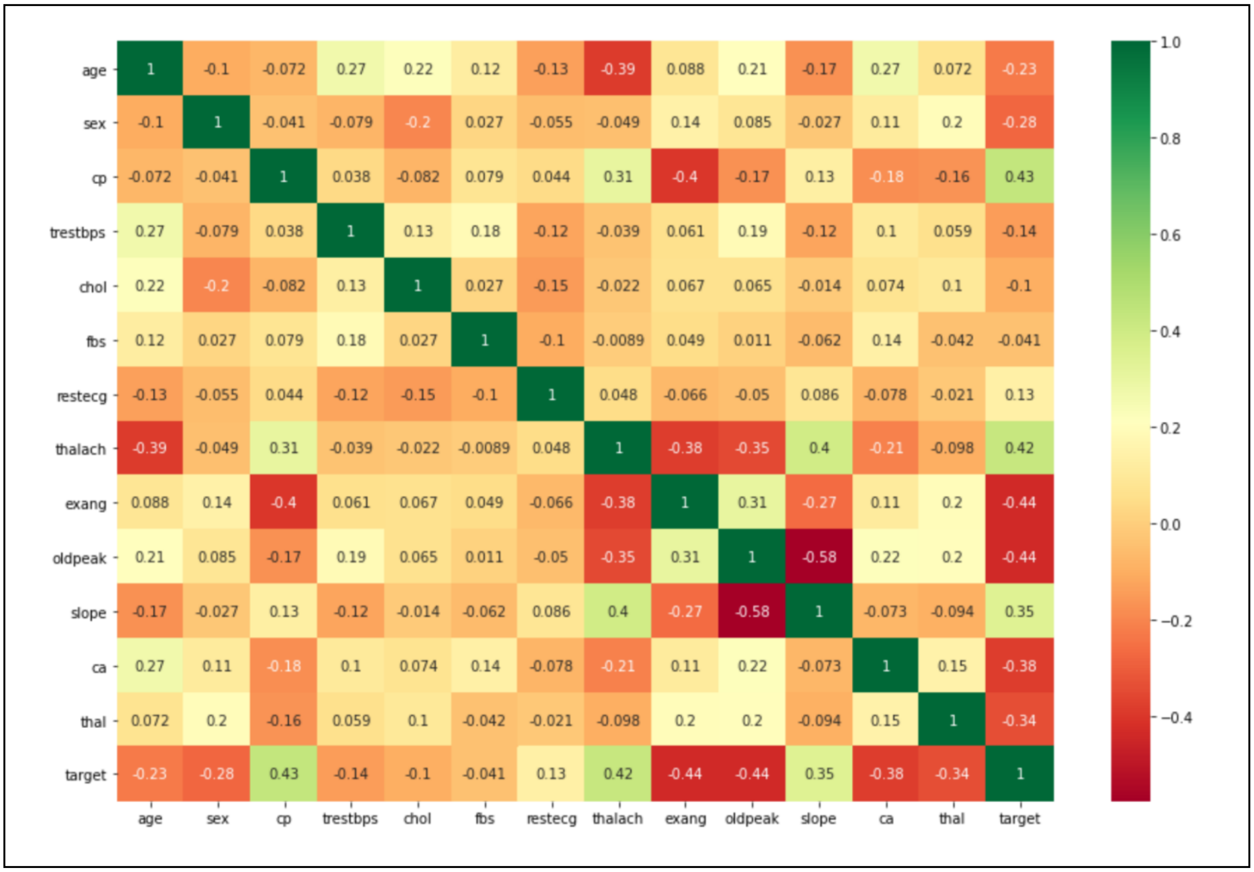
1. age - age in years
2. sex - (1 = male; 0 = female)
3. cp - chest pain type

- 0: Typical angina: chest pain related decrease blood supply to the heart
- 1: Atypical angina: chest pain not related to heart
- 2: Non-anginal pain: typically esophageal spasms (non-heart related)
- 3: Asymptomatic: chest pain not showing signs of disease

4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholesterol in mg/dl
    - serum = LDL + HDL + 0.2 * triglycerides above 200 is cause for concern
6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
    - '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
    - 0: Nothing to note
    - 1: ST-T Wave abnormality can range from mild symptoms to severe problems signals non-normal heart beat
    - 2: Possible or definite left ventricular hypertrophy enlarged heart's main pumping chamber
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    - 0: Up sloping: better heart rate with exercise (uncommon)
    - 1: Flat sloping: minimal change (typical healthy heart)
    - 2: Downslopes: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by fluoroscopy
    - colored vessel means the doctor can see the blood passing through the more blood movement the better (no clots)
13. thal - thallium stress result
    - 1,3: ranging between 1 to 3 in increment of 1
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

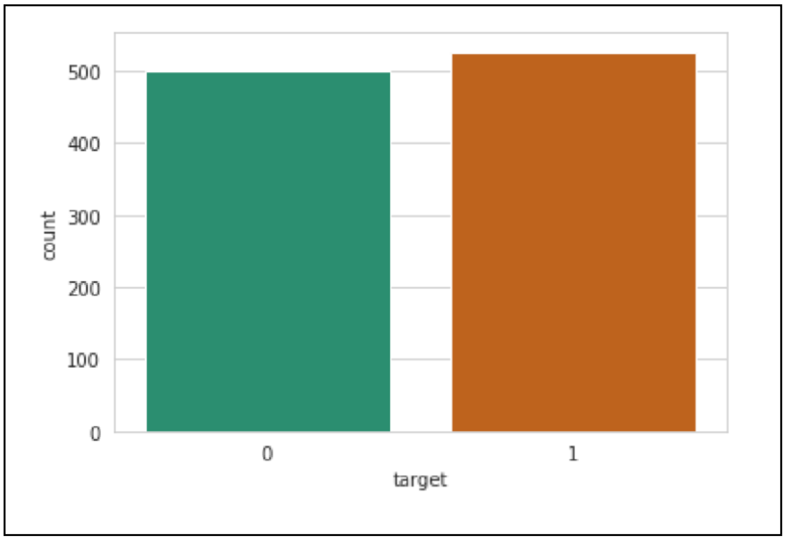
Feature Selection for Dataset

To get an accurate result on the dataset we have to make sure that the any two parameters must haven’t correlation in between. Correlation between any predictor variable is not at all desirable for a good machine learning model. We can know whether linearity is present or not in the model by “Pearson correlation coefficient” between the predicates. We can get the value by heat-map for the feature of the dataset and if any two feature has correlation coefficient value nearer to 1 then they are highly correlated and if nearer to -1 then highly not correlated and for good model building it is desirable not to get correlation between features. Fortunately in our dataset there is no correlation between predicate features as shown by heat-map below.



Data Balancing

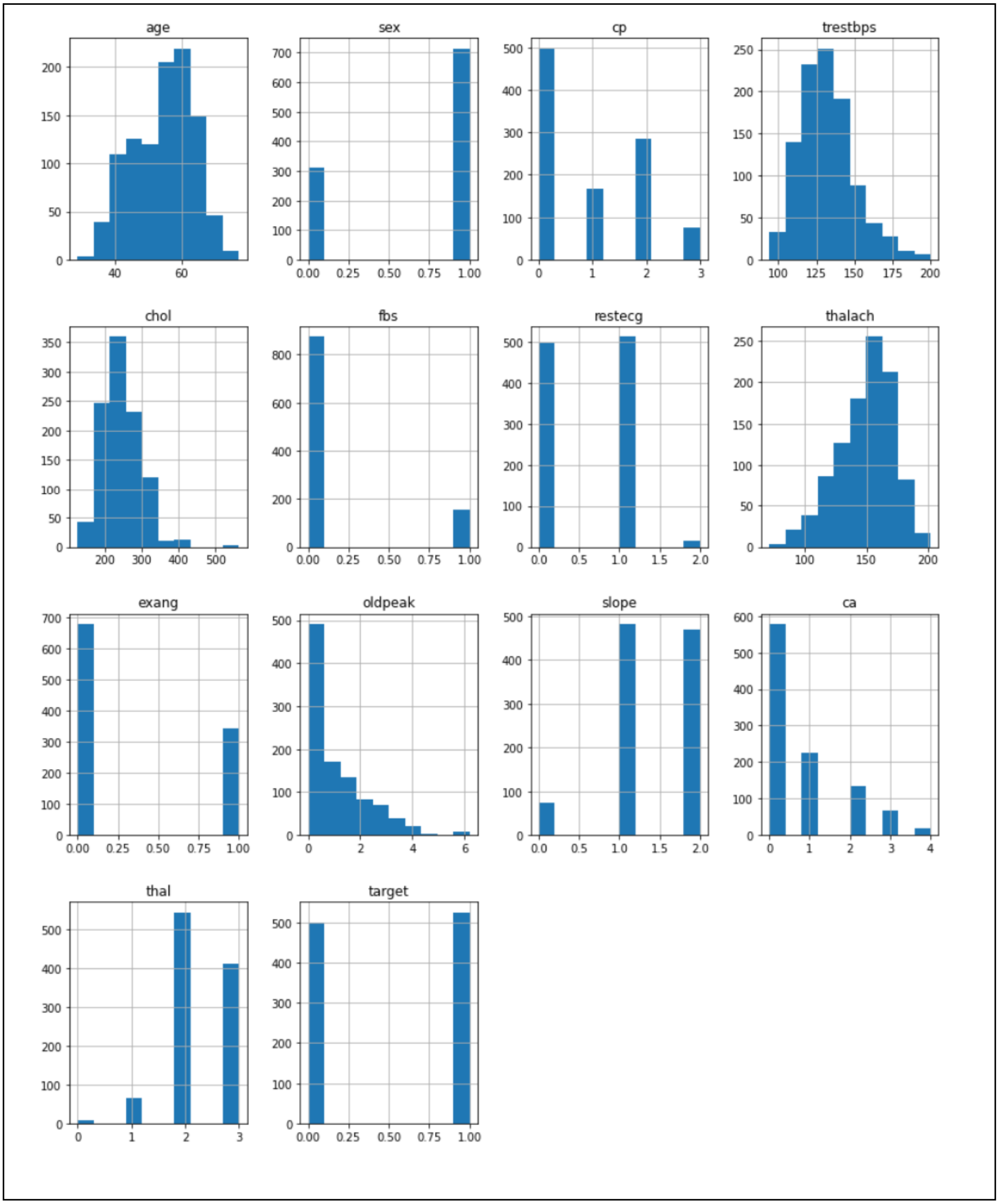
To get accurate results it is necessary to have a balanced dataset by target perspective. If our dataset has biased data for the target then the model will not be able to predict for the minor target dataset accurately. So below is the figure that shows that the data is biased or not.



Here clearly we can see in the graph of count vs. target that the number of patients having heart disease and number of patients not having heart disease is quite similar and there's not a biased data. If we unluckily got the biased data then we have to select some portion of the higher values part and make both parts unbiased and then train and test the model and definitely it'll give more accurate prediction that previous one.

Details of the data in terms of histogram

Histogram of attributes shows the range of dataset attributes and some other details too.



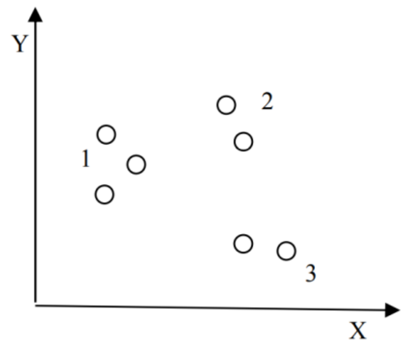
Data Pre-Processing

Preprocessing required for accomplishing esteemed outcomes from the machine learning algorithms. For instance, Random forest algorithm doesn't uphold invalid qualities dataset and for this we need to oversee invalid qualities from unique crude information. For our model we have transformed some categorized values in form binary 0 and 1with the help of dummy variables. After adding dummy variable into our dataset now our dataset transformed from 1025*14 to 1025*31. Here is the list of feature to whom we have to create dummy variables: 'sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'. Now we have scaled our features using standard scaling and after doing this we can get the final dataset through which we can get the maximum accuracy with less errors in the model. Here is the list of features to whom we have to scale: 'age', 'trestbps', 'chol', 'thalach', 'oldpeak'.

Algorithms and its performance

A. K–Nearest Neighbor

K-Nearest Neighbor is one of the most effective classification techniques and one of the elementary techniques too. Main property of this technique is that it doesn't make any assumptions about the data for classification. This technique is basically used when there is very less prior information available about the distribution of the data. This algorithm includes finding the k closest information focused in the training dataset to the target point for which an objective worth is inaccessible and allotting the normal estimation of the discovered information focuses to it. It makes the cluster of the data as shown in the figure here.



Classification in Three Cluster

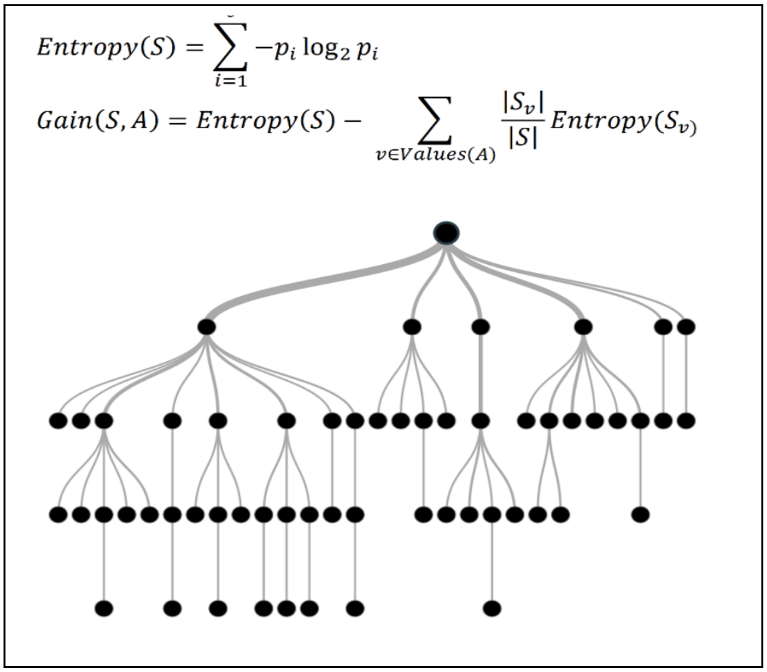
We are using two techniques of the same algorithm 1. Manually 2.GridSearchCV. When we tries to train our dataset manually it gives training accuracy of **89.49416%** with the neighbor value K=17 and when we train same model with GridSearchCV we got the training accuracy as **100%** and testing accuracy **98.44357%** with the neighbor value K=23. According to our objective, we are much interested in testing accuracy rather than training accuracy as we want to predict the result for the unknown dataset rather than existing ones.

B. Decision Tree

Decision tree is a supervised learning algorithm and generally used in classification type problems. In this algorithm the dataset is going to divide in two or more parts according to its effect on the target values. For example, most significant predicate variables can be seen near to the root node and likewise and the significance level is decided based on the minimum entropy or the maximum information gain of the predicate variables. These steps are performed recursively till we get the leaf node. At the point when the quantity of nodes are imbalanced then the tree creates the over fitting issue which isn't useful for the figuring and this is one of motivation behind why choice trees have less precision as contrast with linear regression.

Entropy can be measured by the formula given in the figure and the information gain is given by the formula below the formula of the entropy. In the most significant predicate variable we can see higher information gain and lower entropy.

We are going straight with the GridSearchCV method to get accurate results speedily. When we are going to train our Decision Tree model we get training accuracy as **99.86979%** and when our trained model is tested with the test dataset then it gives accuracy as **97.27626%.**

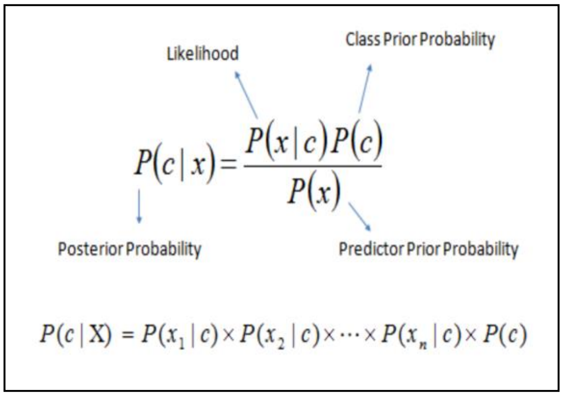


$$Entropy(S) = \sum_{i=1} -p_i \log_2 p_i$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_{v)}$$

C. Gaussian Naive Bayes

Gaussian Naïve Bayes doesn’t have hyper parameters to tune so we are not able to perform the GridSearchCV and go with a simple way and try to train and test with the same dataset as previous. This algorithm technique is quite simple but more accurate for some specific dataset. This algorithm is based on the “Bayes Theorem " and assumes that all predicate variables are independent from rest. Our feature selection and removing collinearity help here to get more accuracy than before one. And unfortunately in other cases there may be some linearity then also all features help independently to generate output and that’s the reason to name this algorithm Naïve Bayes Algorithm.



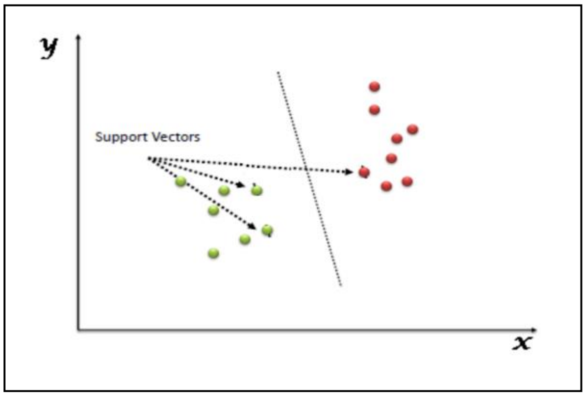
$$P(c\mid x) = \frac{P(x\mid c)P(c)}{P(x)}$$

$$P(c\mid X) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Here there’s not any hyper parameter to tune so no chance to perform GridSearchCV and it results in less accuracy than the rest algorithms but still this algorithm perform well on real life data. Gaussian Naïve Bayes shows the accuracy of **84.24479%** on training dataset where on testing dataset it shows the accuracy of **84.04669%**.

D. Support vector Machine

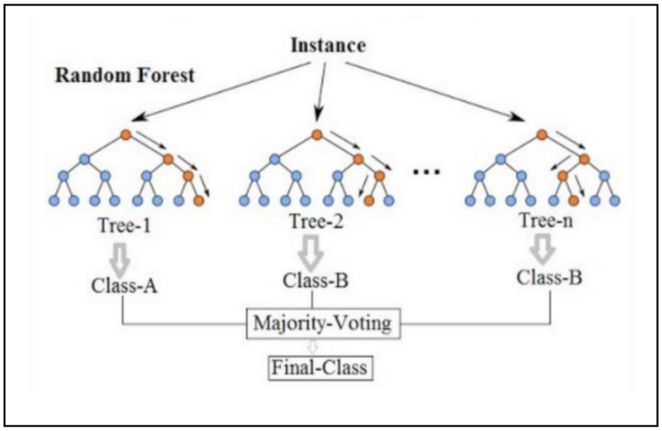
SVM: Support Vector Machine is one the most popular supervised machine learning algorithms. The reason this algorithm is popular is that it can be applied to problems of classification as well as problems of prediction. This algorithm is basically based on the concept of “hyper-plane”. For classification, it tries to find the hyperplane which separates the data on the feature space in such a way that the margin between two is the most. The points of the test datasets are then mapped into that same feature space and differentiate between the sizes of the margin they fall.



Support Vector Machine model when trained using same dataset as previous then it shows the accuracy of the **88.93229%** with the help of the GridSearchCV as I have passed very less parameter with less selection space to verify that it work with less selection space or not and it worked significantly well. When the test dataset predicts the value of the target it gives the accuracy of **86.77042%.**

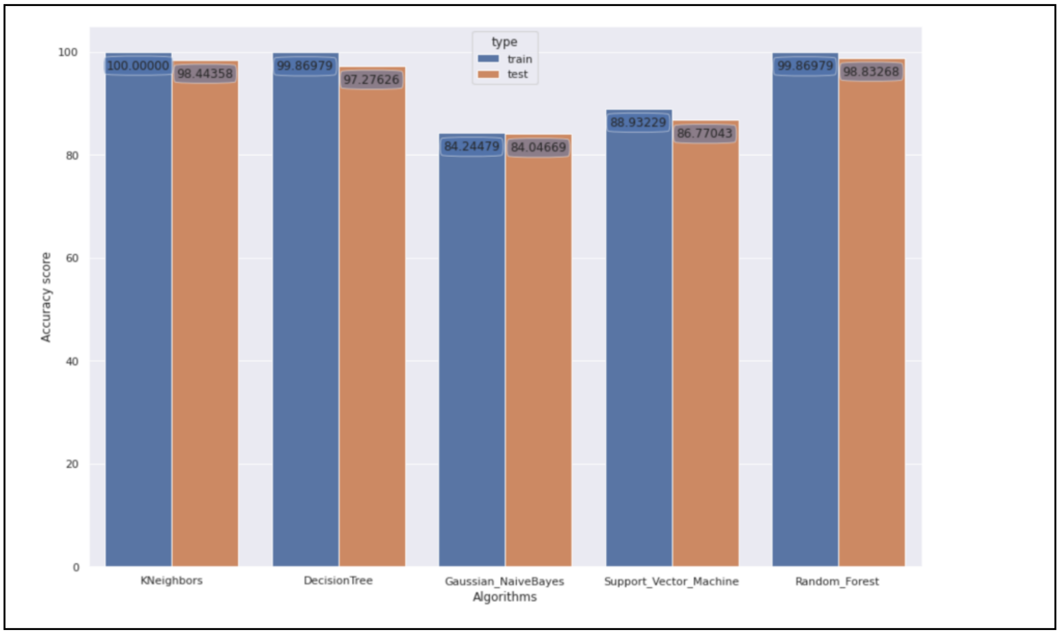
E. Random Forest

Random Forest is the most accurate and the most popular supervised machine algorithm nowadays. This algorithm is used for regression problems as well as for classification problems but generally it is used for classification as classification problems perform better than the regression problem. As per name conventional, Random Forest algorithm uses multiple decision trees and performs the training and testing before giving output. We can consider Random Forest as an ensemble of decision trees. This algorithm works on the belief that more trees lead to more accuracy and after converging gives more accuracy. For classification, it uses a voting mechanism and finds the maximum voted result and then predicts the output. For a regression model, it takes the mean of all decision trees’ output and predicts the final output for the same. It is noticed that it works well with high dimensionality and large datasets.



Random forest performs exceptionally well and accurately on this dataset also. With the same training dataset it gives accuracy of **99.86979%,** and the main important accuracy of the test dataset, we got the accuracy of the **98.83268%** which is highest till the time on this train and test dataset. We can get more accuracy than this also on some other ensemble model or on hybrid model but the scope of this paper restrict us to explore that part.

Comparison of Algorithms



Here is the graph of accuracy of different algorithms and we can clearly see that Random Forest Algorithm is working very well on this test dataset and K-Nearest Neighbor also work well on the train dataset.

Conclusion

Based on the review of this paper, we can conclude that there is huge scope in the field of anatomy and especially in heart disease or cardiovascular diseases prediction. All the supervised algorithms mentioned above perform well in some cases and poor in some other cases. After performing 5 supervised machine learning algorithms for testing and training we can get the result as Random Forest is much efficient as compared to rest 4 algorithms but accuracy is not only the parameter to judge the model. We have to consider the time of training the model as well. In this system we have noticed that Naïve Bayes Classifier is very fast in computation. Machine learning and strategies based systems have been exact in anticipating heart related diseases but at the same time there is a ton of examination to be done on the best way to deal with high dimensional information and over-fitting. A great deal of exploration should likewise be possible on the right group of calculations to use for a specific kind of information.

References

1. Heart disease prediction using machine learning techniques: a survey V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja
2. Heart Disease Prediction Using Machine Learning Algorithms: Archana Singh, Rakesh Kumar
3. Dataset