Heart Disease Prediction using Machine Learning Algorithms

By
Neha Tuniya
Department of Electronics and Telecommunication Engineering
College of Engineering, Pune

Under the Guidance of
Mahesh Parihar
Project Manager,
Smart Energy Informatics Laboratory,
Department of Computer Science & Engineering,
Indian Institute of Technology Bombay, Powai.

INTRODUCTION:

Daily increasing development in information technology caused in significant growth in sciences. One of the sciences is medical science. Using Machine learning techniques in all subjects of this branch of science especially cardiovascular diseases made it possible to design medical assistant systems. By taking attention to increase in new diseases and also extension of technologies, the diagnosis of diseases gone beyond the internal treatment style, and the most efforts of doctors and specialists is focused on early prediction of diseases using available signs. Medical information retrieval system is the best system for managing clinical data. This system is capable to healthcare operations in diagnosing diseases and has an important role in clinical decision making.

Cardiovascular diseases is one of the most spreading causes of death in worldwide. One main type of this disease is “coronary artery disease” (CAD), which about 25% of population without any previous signs, are suddenly subject of this disease, and experience severe heart attack and die. At the moment, Angiography uses for determining the amount and location of narrowing of the arteries of the heart, which has high price and several side effects. Using data mining for diagnosis of heart diseases may be very lower in price and very faster.

Based on the announced statistics by the World Health Organization in 2005, there was 17.5 million victims from cardiovascular diseases, which is 30% of all death in worldwide, and it was predicted that this value increase to 23 million people up to 2030. Examinations made in Iran showed that 38% of all death subjects caused by cardiovascular diseases, which is increasing in future. Based on the statistics obtained from the evaluation of cardiovascular diseases, it was shown that 16.1% of people have high blood pressure, 43.9% have extra weight, 38.9% have low physical activity, and also 10.8% use tobacco which is of the most important factor s of heart diseases.

Diagnosis of heart diseases is a significant and boring task and also an important duty in medical science, which requires extreme attention. However there is some tools for data extraction and analysis. Also existence of huge set of medical data leads to correct diagnosis of disease. Using medical data including age, sex, blood pressure, and blood sugar, it is possible to increase the possibility of heart diseases prediction. These data must be collected in organized manner, which could be used for integrating the prevention system.

Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis. Therefore using data mining and discovering knowledge in cardiovascular centers could create a valuable knowledge, which improves the quality of service provided by managers, and could be used by doctors to predict the future behavior of heart diseases using past records. Also some of the most important applications of data mining and knowledge discovery in heart patients system includes: diagnosing heart attack from various signs and properties, evaluating the risk factors which increases the heart attack.

LITERATURE SURVEY:

Research papers used for survey are listed as follows:

1. Predicting Heart attacks in patients using Artificial Intelligence Methods
2. A fuzzy clustering neural network architecture for classification of ECG arrhythmias.
3. Detection of Brain Tumor in MRI Images, using Combination of Fuzzy C-Means and SVM.
4. Hybrid Multistage Fuzzy Clustering System for Medical Data Classification.
5. Heart disease prediction using effective machine learning techniques - International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019
6. Heart disease prediction using machine learning techniques - A Survey International Journal of Engineering and technology
   DOI - 10.14419/ijet.v7i2.8.10557
7. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms.

Key points from the above mentioned papers:

1. *Predicting Heart attacks in patients using Artificial Intelligence Methods:* The diagnosis of heart patients is carried out using a combination of Fuzzy clustering algorithm and genetic algorithm, to gain more precise diagnosis in this disease. Traditional clustering methods like K-Means often judge on data using the distance between them. But in this study the highest objection is using this property, because available data about heart patients includes binary and nominal data. So using various improved clustering methods, we

can use other metrics (may include Jaccard distance, hamming distance, etc) instead of focusing on distance between data, to focus on qualitative properties, to increase the precision and gain more correct diagnosis.

Use of fuzzy clustering algorithm with genetic algorithm: In clustering, the data splits to some clusters, in such a way that the data in every cluster have maximum similarity with each other and minimum similarity with data of other clusters. So using clustering data will show that every cluster that has the patient, could help us in predicting if he/she is under heart attack risk or not.

In Machine learning, selecting parameters and then getting optimised result based on those parameters is a difficult task. But it is very important to do optimization because a classifier may produce a bad classification accuracy not because, for example, the data is noisy or the used learning algorithm is weak but due to the bad selection of the learning parameters initial values. To avoid this problem and get optimised results, genetic algorithms are used. In this paper, there are two goal functions which are considered parallel. They are global compressing and resolution functions. Resolution function is to be maximised and compression is to be minimised. Therefore, we minimise both compression and (1/resolution) function using genetic algorithms.

Data selection and normalisation: There were a total 76 attributes out of which 14 were selected. Data was normalised before applying the algorithm. Data is normalised using min-max normalisation method.
Need of normalisation: The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. Normalization is also required for some algorithms to model the data correctly.
For example, assume your input dataset contains one column with values ranging from 0 to 1, and another column with values ranging from 10,000 to 100,000. The great difference in the scale of the numbers could cause problems when you attempt to combine the values as features during modeling.

Fuzzy clustering algorithm: There are total n discrete objects each having p attributes, K clusters are to be formed. There are K modes corresponding to K clusters. Loss function which is to be minimised is measuring dissimilarity between every object and all cluster modes. This is an iterative process. Start with K random modes, the fuzzy membership values are calculated and based on membership values, the cluster modes are recalculated.

The algorithm is evaluated using compression and resolution and validated using Silhouette index and Dunn index. Silhouette index is defined as average distance between every sample of a cluster with all available samples of that cluster, and average distance of total current available samples in other clusters with a specific cluster. Based on this, the amount of diversity and data correlation of data is determined. The value of the Silhouette validation index is between -1 and +1. The higher values of this index (near to +1) shows that clustering is made correctly. If the index is near zero, means that we can assign a sample to a nearer cluster, and the sample is located in the same distance from both clusters. If the index becomes -1, it means that the clustering was not made correctly.

S(C) is an OSI (Overall Silhouette index) index for a clustering is the average value of Silhouette for all clusters: The Silhouette index used for evaluating the level of similarity between suggested algorithm and heart patients classifying.

2. *A fuzzy clustering neural network architecture for classification of ECG arrhythmias:*

   The structure proposed in this paper is composed of two subnetworks: fuzzy classifier and neural network. The fuzzy self-organizing layer performs the pre classification task and the following multilayer perceptron works as a final classifier.

   The fuzzy stage is responsible for the analysis of the distribution of data and grouping them into clusters with different membership values. On the basis of these membership values, the MLP network classifies the applied input vector, representing the heartbeat to the appropriate class. On the other hand, a number of segments in training patterns are reduced using FCM clustering in a fuzzy self-organizing layer before inputs are presented to the MLP. The obtained new training data whose number of segments is decreased using fuzzy clustering are presented to the MLP. Therefore, the training period of the neural network is decreased. This method proposed in this paper was used for electrocardiographic beat recognition and classification.

   Training patterns were clustered. Two processes were implemented in this part of study. Firstly, a number of segments in each type of arrhythmia were

reduced by using FCM clustering. Secondly, a number of segments in each type of arrhythmia were increased by using FCM clustering.
Results showed that the reduced number of segments performed better and had the least training error.

3. *Detection of Brain Tumor in MRI Images, using Combination of Fuzzy C-Means and SVM:*
   The proposed algorithm is a combination of support vector machine (SVM) and fuzzy c means, a hybrid technique for prediction of brain tumor. In this algorithm the image is enhanced using enhancement techniques such as contrast improvement, and mid-range stretch. Double thresholding and morphological operations are used for skull striping. Fuzzy c-means (FCM) clustering is used for the segmentation of the image to detect the suspicious region in brain MRI image. Grey level run length matrix (GLRLM) is used for extraction of feature from the brain image, after which SVM technique is applied to classify the brain MRI images, which provide accurate and more effective result for classification of brain MRI images.

4. *Hybrid Multistage Fuzzy Clustering System for Medical Data Classification:*
   In the proposed system, two fuzzy clustering algorithms specifically FCM and GK were initially employed to obtain the membership values. These weights are then used in the second stage of the system as additional informative features to improve the classification process completed by the SVM algorithm.

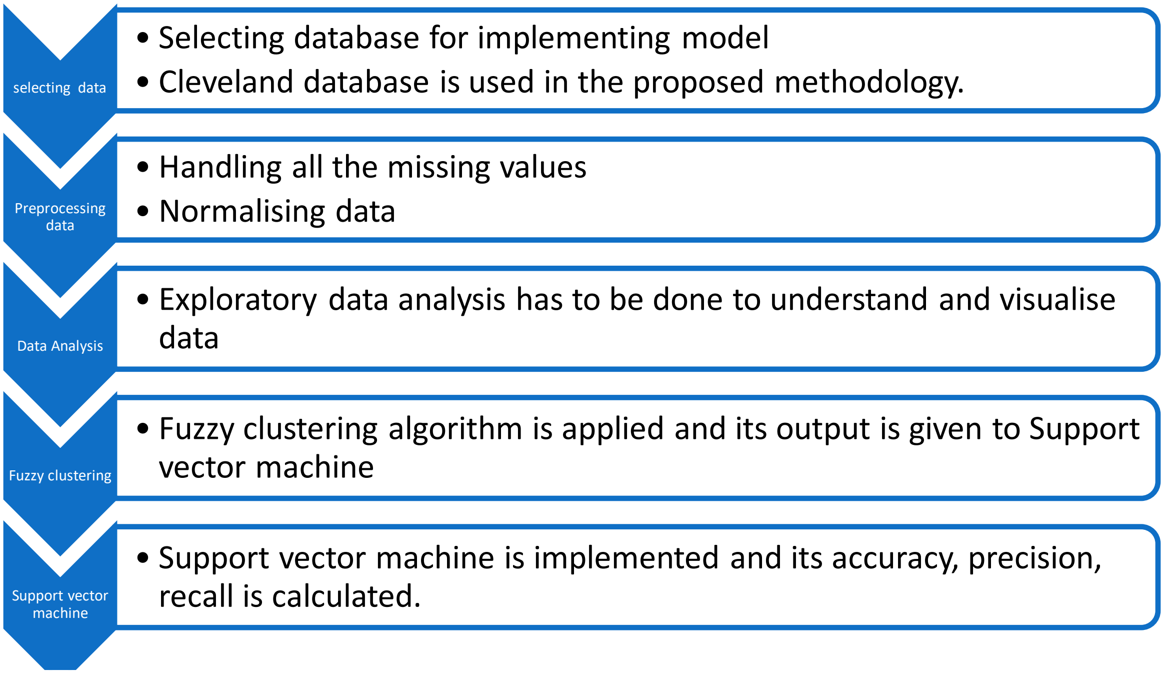
   Clustering phase: The experimental results have shown that a better performance was obtained by FCM over GK with 95% classification accuracy for the former and 91% achieved by the latter. This outcome demonstrates that FCM is more suitable for this particular dataset. In addition, based on these results one more assumption can be made regarding the data distribution. Since FCM searches for spherical clusters, the results achieved by FCM indicate that the data could have Gaussian normal distribution. This outcome leads to the conclusion that some of the fuzzy clustering methods are found to fit some cancer data more than other techniques. Since the data has Gaussian distribution, SVM machine with a linear kernel function is preferable for the second stage. In the second stage, the weights resulted in the earlier phase are added to the data as additional informative features. This is expected to result in better performance.

Classification Phase : For the second stage of the system, in continuation and based on the first stage, support vector machine is used on the same dataset. By adding the additional outcomes obtained from the fuzzy classifiers, SVM was trained and tested on three data sizes. The datasets were divided into several ratios 50%-50%, 60%-40%, 70%-30% and 80%-20%, respectively. Linear kernel functions were used for SVM in the experiments.

PROPOSED METHODOLOGY

The proposed methodology consists of fuzzy clustering followed by Support vector machine. In clustering, the data splits to some clusters, in such a way that the data in every cluster have maximum similarity with each other and minimum similarity with data of other clusters. Uncertainty or fuzziness lying in the data may be reduced as the unsupervised learning model is first applied to the input data and generates certain number of clusters. So using clustering data and original features, SVM is implemented which could help us in predicting that if he/she is under heart attack risk or not.

APPROACH OF IMPLEMENTATION



selecting data

- Selecting database for implementing model
- Cleveland database is used in the proposed methodology.

Preprocessing data

- Handling all the missing values
- Normalising data

Data Analysis

- Exploratory data analysis has to be done to understand and visualise data

Fuzzy clustering

- Fuzzy clustering algorithm is applied and its output is given to Support vector machine

Support vector machine

- Support vector machine is implemented and its accuracy, precision, recall is calculated.

DETAILS OF METHODOLOGY

1. *Selecting data*: Dataset for predicting heart disease is taken from UCI Machine learning repository. There are several dataset available in UCI which includes Cleveland dataset, Hungarian Institute in cardiovascular diseases, Long-Beach medical center and Medical Sciences University of Switzerland. The proposed system is implemented using Cleveland dataset.

This database has 76 attributes, but only 14 attributes used. All the attributes have numerical value and data-sets were used. Data-sets contain the following attributes:

1. age: The person's age in years
2. sex: The person's sex (1 = male, 0 = female)
3. cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
5. chol: The person's cholesterol measurement in mg/dl
6. fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. thalach: The person's maximum heart rate achieved
9. exang: Exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.
11. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12. ca: The number of major vessels (0-3)
13. thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. target: Heart disease (0 = no, 1 = yes)

2. *Normalising data*

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. Normalization is also required for some algorithms to model the data correctly.

For example, assume your input dataset contains one column with values ranging from 0 to 1, and another column with values ranging from 10,000 to 100,000. The great difference in the *scale* of the numbers could cause problems when you attempt to combine the values as features during modeling.
Normalization avoids these problems by creating new values that maintain the general distribution and ratios in the source data, while keeping values within a scale applied across all numeric columns used in the model.

Various Methods of normalization.

- *Z-score*: Converts all values to a z-score. The values in the column are transformed using the following formula:

  $$\mathrm{z} = \frac{x - \mathrm{mean}(x)}{std\_deviation(x)}$$

  Mean and standard deviation are computed for each column separately.

- *MinMax*: The min-max normalizer linearly rescales every feature to the [0,1] interval. Rescaling to the [0,1] interval is done by shifting the values of each feature so that the minimal value is 0, and then dividing by the new maximal value (which is the difference between the original maximal and minimal values).

  The values in the column are transformed using the following formula:

  $$\mathrm{z} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- *Logistic*: The values in the column are transformed using the following formula:

  $$\mathrm{z} = \frac{1}{1 + \exp(-x)}$$

- *LogNormal:* This option converts all values to a lognormal scale. The values in the column are transformed using the following formula:

  $$z = Lognormal.CDF(x; \mu, \sigma)$$

  Here μ and σ are the parameters of the distribution, computed empirically from the data as maximum likelihood estimates, for each column separately.

- *TanH*: All values are converted to a hyperbolic tangent. The values in the column are transformed using the following formula:

  $$p(k|x; \theta) = \frac{[E(Y|x)]^k e^{-E(Y|x)}}{k!}$$

In this model, min-max normalization is used. The formula for min-max normalization is given below. Here, x is the original value of a particular feature of some example. Min(x) is the minimum value of the feature and max(x) is the maximum value of the feature. Z is the normalized value.

$$\mathrm{z} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3. *Exploratory data analysis*

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Details with figures are listed below.

4. *Fuzzy clustering algorithm*

Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster.

Clustering or cluster analysis involves assigning data points to clusters such that items in the same cluster are as similar as possible, while items belonging to different clusters are as dissimilar as possible. Clusters are identified via similarity measures. These similarity measures include distance, connectivity, and intensity. Different similarity measures may be chosen based on the data or the application.

Types of similarity measures:

1. *Adjusted Rand Index*: The Rand Index computes a similarity measure between two clustering by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering. The raw RI score is then "adjusted for chance" into the ARI score using the following scheme:

$$ARI = \frac{RI - Expected\ RI}{\max(RI) - \ Expected\ RI}$$

   The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clustering are identical (up to a permutation). ARI is a symmetric measure.

2. *Mutual Info Score*: Mutual Information between two clustering. The Mutual Information is a measure of the similarity between two labels of the same data. Where |Ui| is the number of the samples in cluster Ui and |Vj| is the number of the samples in cluster Vj, the Mutual Information between clustering U and V is given as:

$$MI(U,V) = \sum_{I=1}^{U}\sum_{J=1}^{V}\frac{|U_i \cap V_j|}{N} log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

   This metric is independent of the absolute values of the labels: a permutation of the class or cluster label values won't change the score value in any way.

3. *Fowlkes – Mallows index*: Measure the similarity of two clusterings of a set of points. The Fowlkes-Mallows index (FMI) is defined as the geometric mean between of the precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP + FP) * (TP + FN)}}$$

4. *Davies –Bouldin score*: The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is

the ratio of within-cluster distances to between-cluster distances. Thus, clusters which are farther apart and less dispersed will result in a better score. The minimum score is zero, with lower values indicating better clustering.

5. *Silhoutte index*: Silhouette index is defined as average distance between every sample of a cluster with all available samples of that cluster, and average distance of total current available samples in other clusters with a specific cluster. Based on this, the amount of diversity and data correlation of data is determined. The value of the Silhouette validation index is between -1 and +1. The higher values of this index (near to +1) shows that clustering is made correctly. If the index is near zero, means that we can assign a sample to a nearer cluster, and the sample is located in the same distance from both clusters. If the index becomes -1, it means that the clustering was not made correctly. S(C) is an OSI (Overall Silhouette index) index for a clustering is the average value of Silhouette for all clusters.

Mathematical formulation:

Assume that X = {x_1, x_2, x_3,….. ,$x_{n,}$} is a set of n objects in discrete categorical zone. Every object, x_i , i=1, 2,…,n, describes by a set of p properties: A_1, A_2 ,..., A_p . Assume that DOM(A_j) , 1 ≤ j ≤ p is a domain of jth property and includes q_j different groupings, such that (A_j) = { a_j^1 , a_j^2,…. a_j^{qj} } So ith discrete object defined as x_i = [x_{i1}, x_{i2} ,..., x_{ip}] which1 ≤ j ≤ p, x_{ij} ∈ DOM(A_j). The defined centers of clusters in FCM are as follows: assume that Ci is a set of discrete objects which belongs to ith cluster. The center of C_i is a vector, m_i = [m_{i1},m_{i2} ,...,m_{ip}] , which 1 ≤ j ≤ p, m_{ij} ∈ DOM(A_j) , in such a way that minimize the following criterion:

$$D(m_i, x_i) = \sum_{x \in Ci} D(m_i, x)$$

Where D(m_i , x) is the amount of dissimilarity between m_i and x. Necessarily m_i is not one of the members of C_i set. Fuzzy C-means clustering algorithm is on data x with C clusters for minimizing the following criterion:

$$J_m(U, Z : X) = \sum_{k=1}^{n} \sum_{i=1}^{C} u_{ik}^m D(z_i, x_k)$$

There are some conditions on probabilistic fuzzy clustering are as follows:

$$0 \leq u_{ik} \leq 1, \ \ 0 \leq i \leq C, \ \ 0 \leq k \leq n$$

$$\sum_{i=1}^{C} u_{ik} = 1 , \qquad 0 \ \leq k \leq \ n$$

$$0 < \sum_{k=1}^{n} u_{ik} < n , \ \ 0 \ \leq i \ \leq \ C$$

The fuzzy power is m, and K × n fuzzy clustering matrix is $U = [u_{ik}]$, and kth discrete object's membership degree in ith cluster is u_{ik} . $Z = \{z_1, z_2 ,..., z_C \}$ shows the centers of clusters.

The Fuzzy algorithm is a part of periodic optimizing strategy, which contains the repetition of estimate of clustering matrix, and calculating the new cluster centers. It starts with C random primary centers, and then in each repetition, the fuzzy membership on every data point in every cluster calculates by the following equation.

$$u_{ik} = \frac{1}{\sum_{j=1}^{C} (\frac{D(z_i, x_k)}{D\left(z_j, x_k\right)})^{\frac{1}{m-1}}}$$

The finish condition of algorithm is when there was no significant improvement in J_m value. Finally, every object assigns to a particular cluster with the membership value it contains.

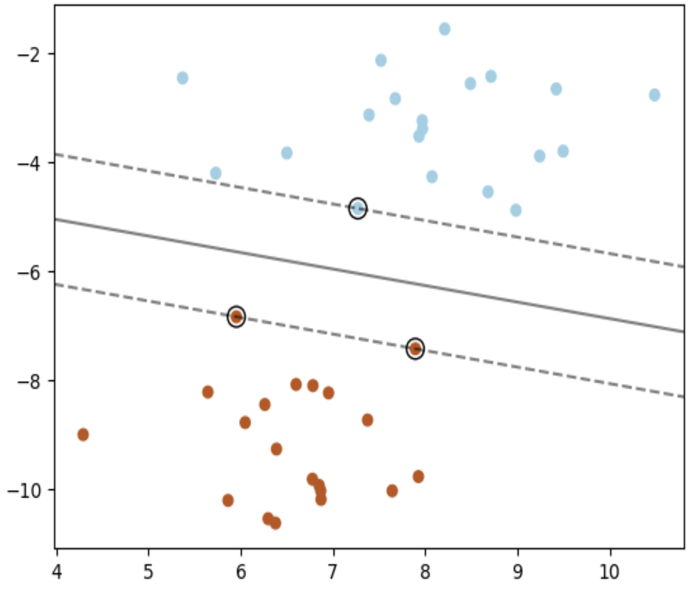
6. *Support vector machine*

The output of fuzzy clustering is given to Support vector machine to improve the accuracy of the model. Clustering algorithm results in the membership matrix *u* of dimension C*n which signifies the membership values of each data samples in C different clusters. This matrix *u* is added to the data set as additional features. Thus the number of features will be equal to the sum of the original features present in the dataset (in this case 13) and number of clusters (C).

In machine learning, support-vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.



In the above figure, Data set consist two attributes which are separated a hyperplane created by SVM and is at the maximum margin. It shows the decision function for a linearly separable problem, with three samples on the margin boundaries, called “support vectors”

In general, when the problem isn’t linearly separable, the support vectors are the samples within the margin boundaries.

Mathematical formulation

Given training vectors $x_i \in R^p$, $i=1,\ldots, n$, in two classes, and a vector $y \in \{1,-1\}^n$, our goal is to find $w \in R^p$ and $b \in R$ such that the prediction given by $\text{sign}(w^T\phi(x)+b)$ is correct for most samples.

SVC solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{Subject to } y_i(w^T\phi(x_i)+b) \geq 1-\zeta_i$$

$$\zeta_i \geq 0,\ i=1,2\ldots,n$$

Intuitively, we’re trying to maximize the margin by minimizing $\|w\|^2 = w^T w$, while incurring a penalty when a sample is misclassified or within the margin boundary. Ideally, the value $y_i(w^T\phi(x_i)+b)$ would be ≥ 1 for all samples, which indicates a perfect prediction. But problems are usually not always perfectly separable with a hyperplane, so we allow some samples to be at a distance ζ_i from their correct margin boundary. The penalty term C controls the strength of this penalty, and as a result, acts as an inverse regularization parameter.

The dual problem to the primal is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

$$\text{subject to } y^T\alpha = 0;\ \ 0 \leq \alpha_i \leq C,\ i = 1,\ldots.,n$$

where e is the vector of all ones, and Q is an n by n positive semidefinite matrix, $Q_{ij} = y_i y_j K(x_i,x_j)$, where $K(x_i,x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel. The terms α_i are called the dual coefficients , and they are upper bounded by C. This dual representation highlights the fact that training vectors are implicitly mapped into higher dimensional space by the function ϕ.

Once the optimization problem is solved, the output of devsion function for a given sample x becomes:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b$$

And the predicted class correspond to its sign. We only need to sum over the support vectors because the dual coefficients α_i are zero for the other samples.

Linear SVC

The primal problem can be equivalently formulated as

$$\begin{matrix} min \\ w, b \end{matrix} \frac{1}{2} w^T w + C \sum_{i=1} \max(0, y_i (w^T \varphi(x_i) + b)$$

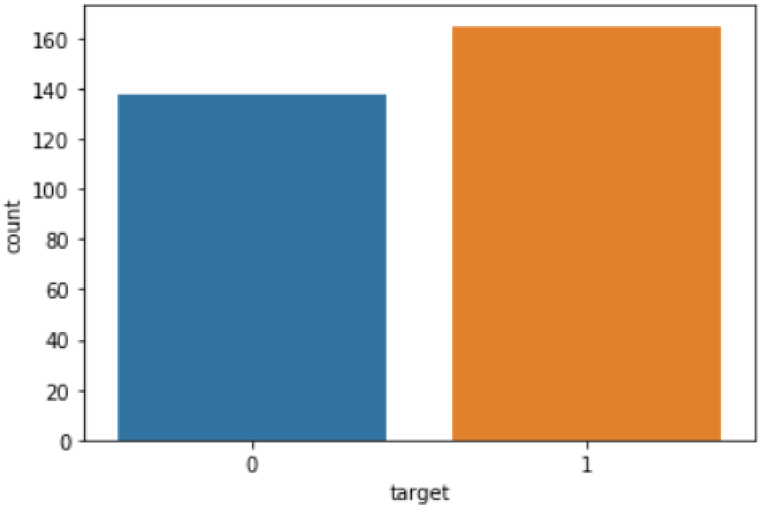
Where we make the use of hinge loss. This is the form that is directly optimized by Linear SVC, but unlike the dual form, this one does not involve inner products between samples, so the famous kernel trick cannot be applied.

EXPLORATORY DATA ANALYSIS

Data-sets contain the following attributes:

1. age: The person's age in years
2. sex: The person's sex (1 = male, 0 = female)
3. cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
5. chol: The person's cholesterol measurement in mg/dl
6. fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. thalach: The person's maximum heart rate achieved
9. exang: Exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.
11. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12. ca: The number of major vessels (0-3)
13. thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. target: Heart disease (0 = no, 1 = yes)

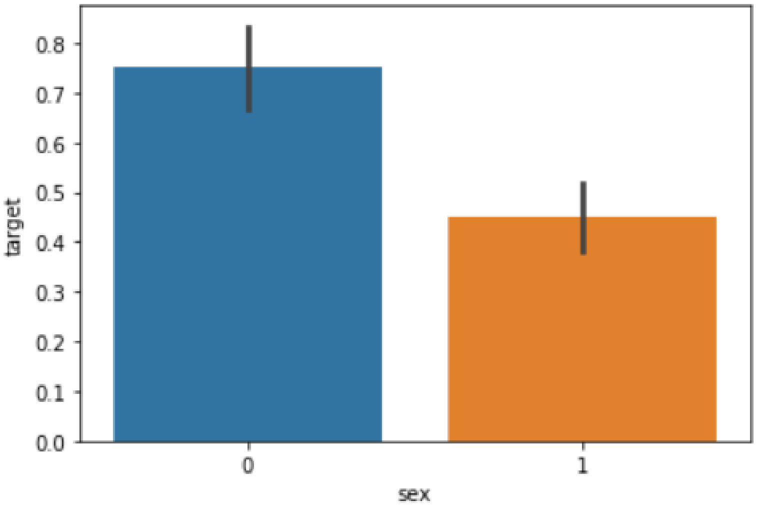
- Plot showing the number of infected and uninfected people and their percentages



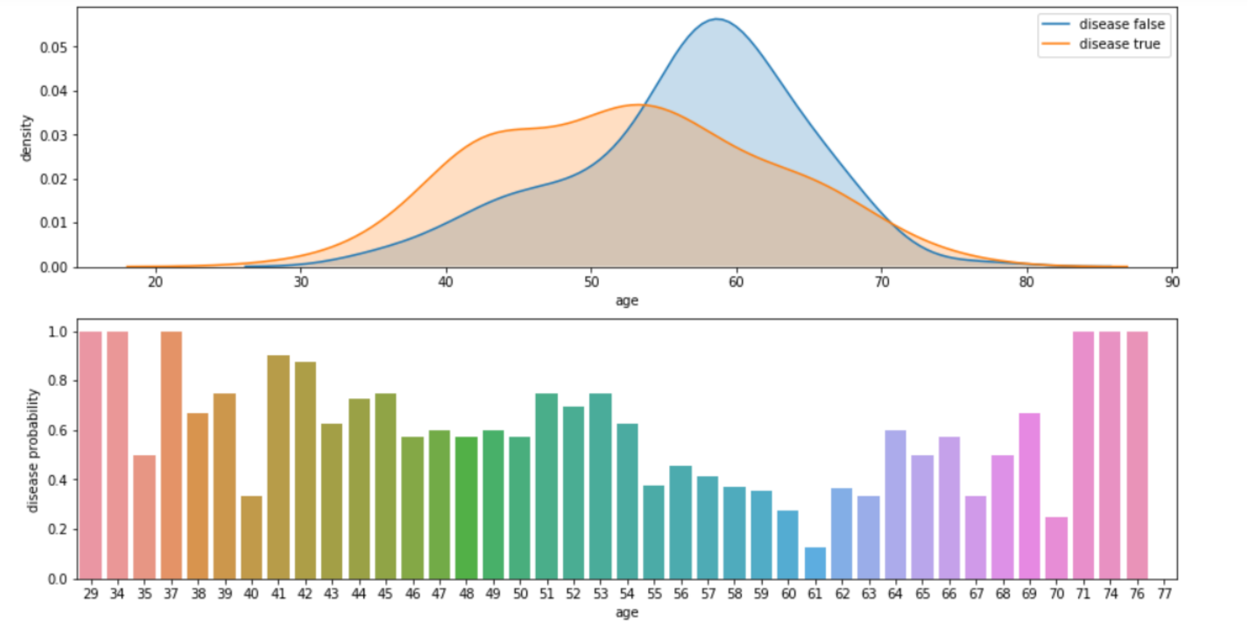
Number of patients having heart disease is 165 and the people who are not having heart disease is 138. Therefore, percentage of patience having heart disease is (165/303)*100 = 54.46%

Percentage of patience without heart disease is 45.54
Percentage of patience with heart disease is 54.46

- Plot showing the probability of person having disease versus the sex of the person

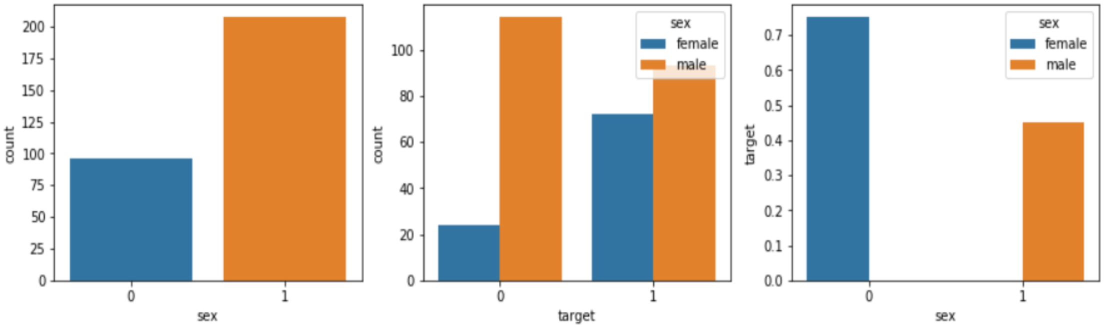


- Plot showing the probability of person having the disease versus the age of the person



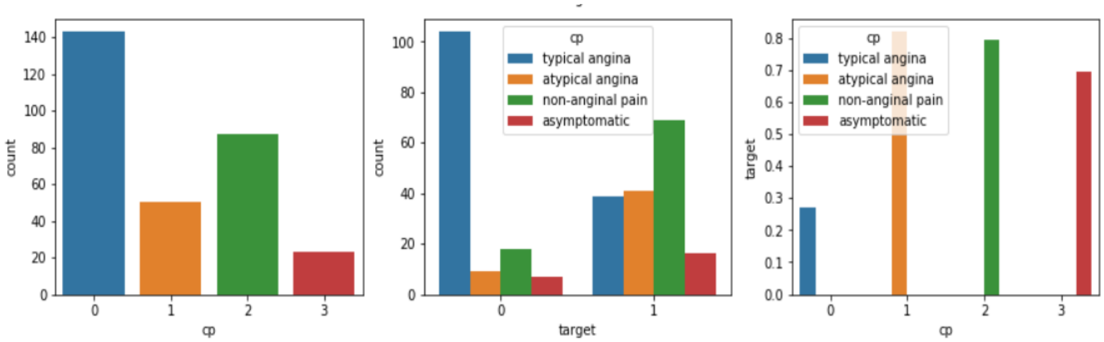
From the above graph, it is seen that people with age around 60 are less prone to heart disease, while people with age between 71 to 76 and 29 to 37 are more prone to disease.

- Details of categorical attributes
  - Sex



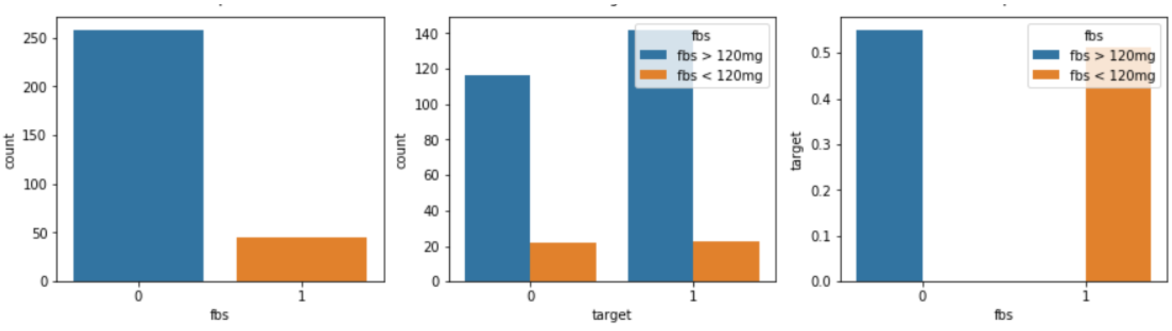
The number of the male are approximately double than the female in the data set. The probability of female having heart disease is approximately is 0.75 and male having heart disease is 0.45. Therefore, female patients are more prone to heart disease.

- Chest pain type(cp)



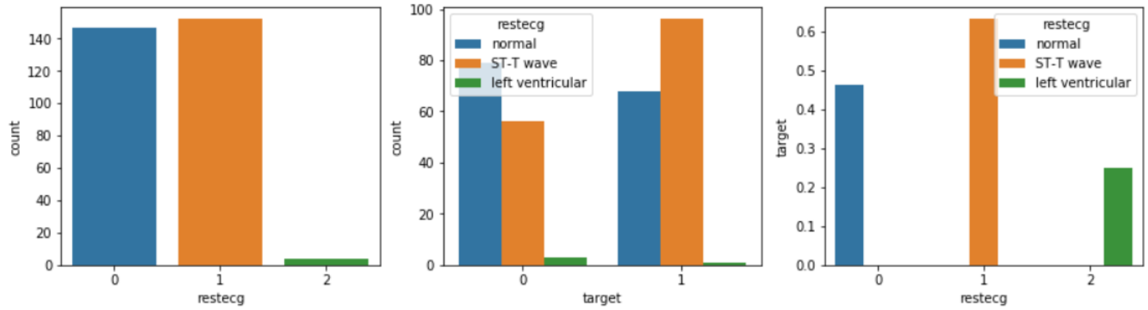
The number of people with chest pain type typical angina is highest and that asymtomatic is least in the data set. People with chest pain type atypical angina and non-anginal pain are most prone to having Heart Disease.

- Fasting blood sugar(fbs)



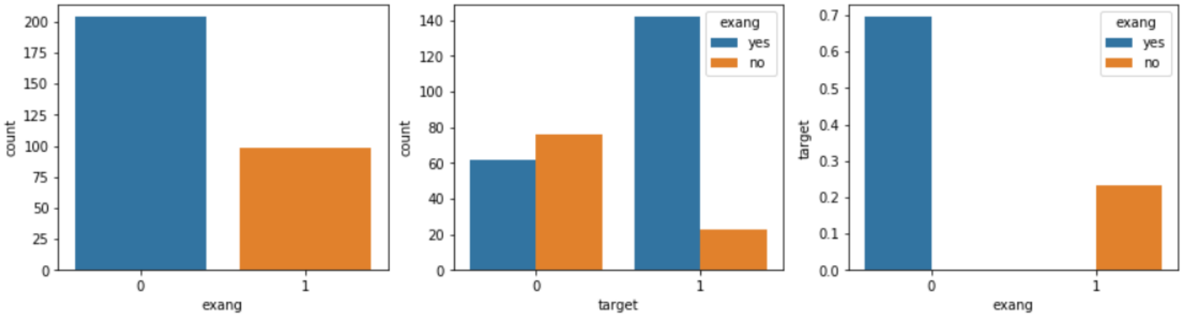
People having fasting blood sugar more than 120mg are nearly five times more than the people having fasting blood sugar less than 120mg in the dataset. Probability of having heart disease is equal for both type (i.e having fast blood sugar less than 120mg and more than 120mg)

- Resting electrocardiographic measurement(restecg)



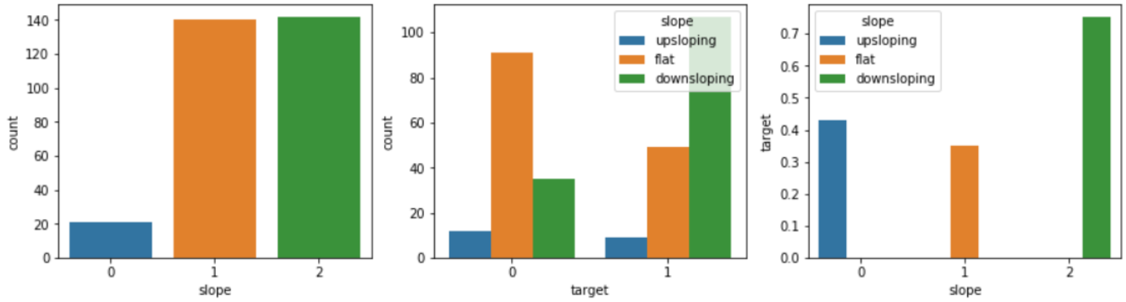
People having resting ecg normal and ST-T wave abnormality are nearly equal while peolpe showing probable or definite left ventricular hypertrophy by Estes' criteria are very less in the dataset. People showing ST-T wave abnormality are more prone to heart disease and showing probable or definite left ventricular hypertrophy by Estes' criteria are very less prone to heart disease.

- Exercised induced angina (exang)



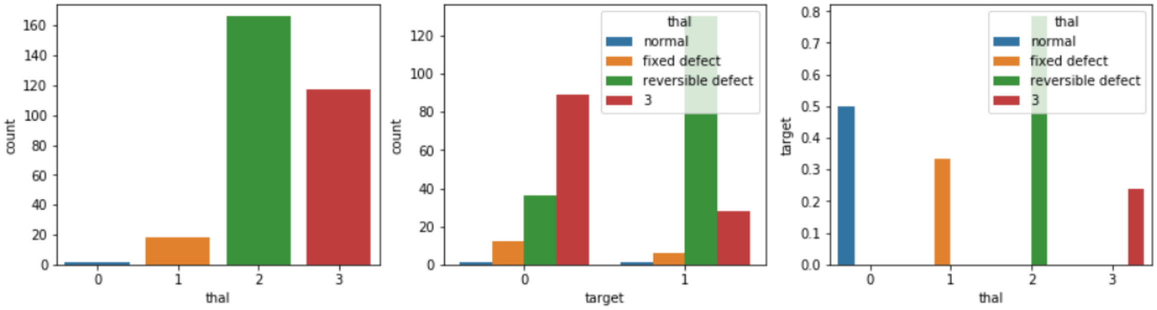
People having Exercised induced angina are half than the people not having exercised induced angina in the dataset. Probabilty of people having exercised induced angina are more susceptible to hear disease.

- Slope of the peak exercise ST segment (slope)



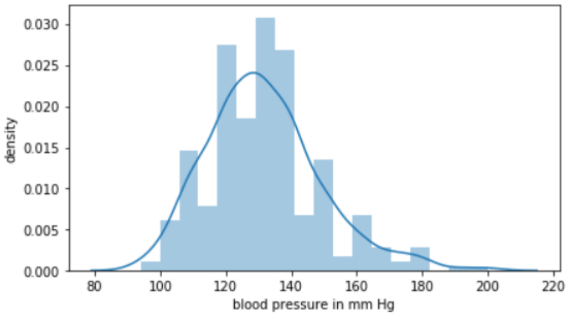
People having ST segment slope flat and downsloping are nearly equal while people having upsloping are less in the dataset. Probabilty of peak exercise ST segment slope downsloping are more prone to heart disease while people having flat and upsloping are nearly equally prone to heart disease.

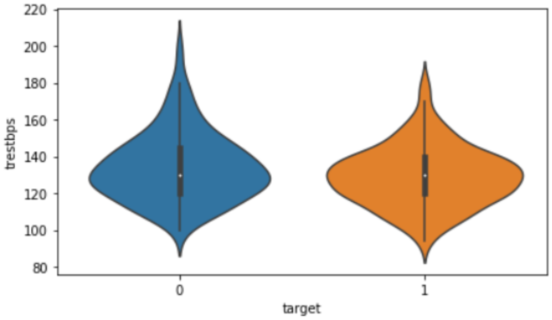
- A blood disorder called thalassemia (thal)



People having reversible defect are highest while people having fixed defect are least in the dataset. Probabilty of people having reversible defect are most prone to heart disease.

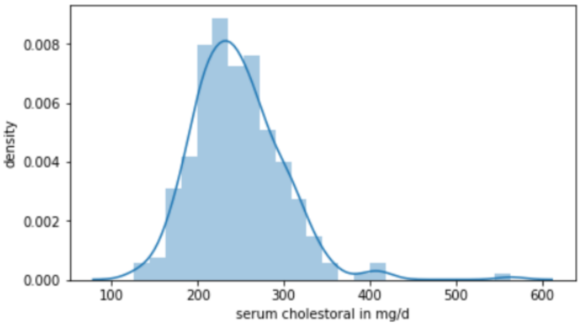
- Details of continuous attributes
  - The person's resting blood pressure (trestbps)

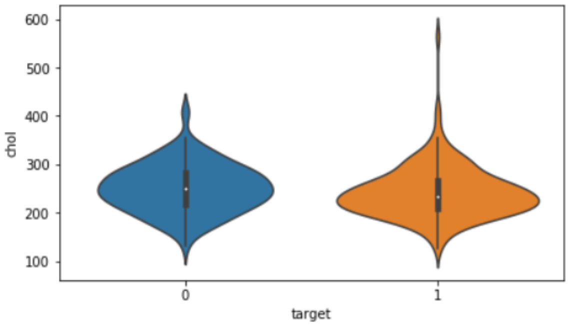




Resting blood pressure is nearly normally distributed around 130mm Hg in the data set. Median is nearly same for all people (i.e people having heart disease and not having heart disease)

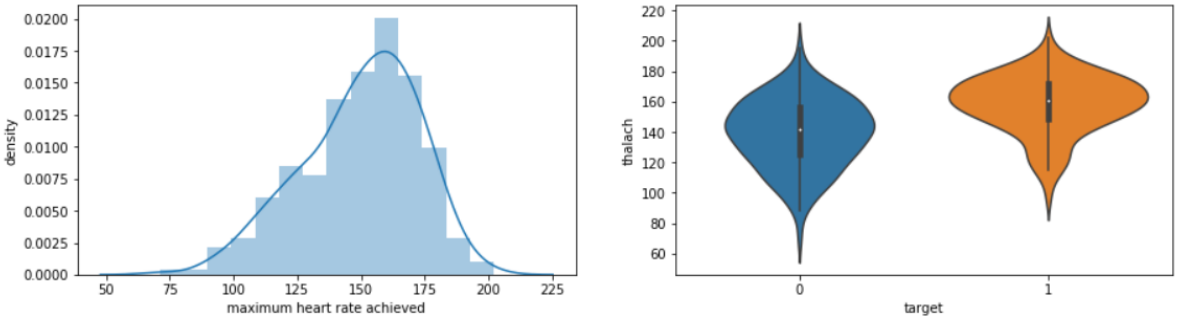
  - The person's cholesterol measurement in mg/dl (chol)





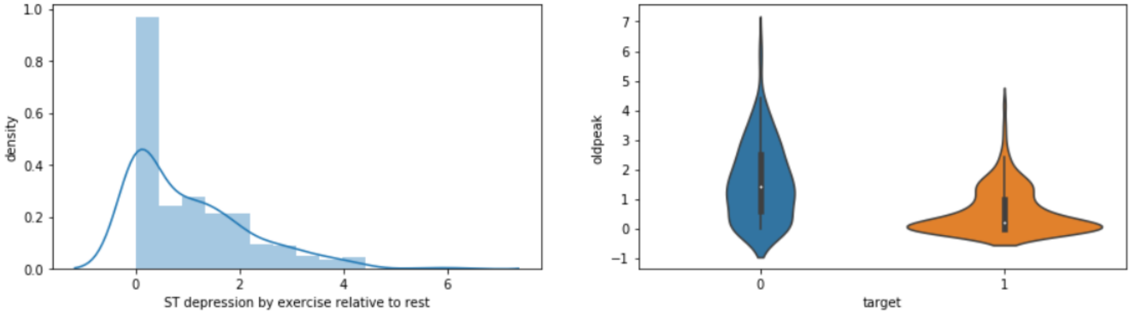
Cholesterol is nearly normally distributed around 220mg/dl in the dataset. People having heart disease have cholesterol varying over a wide range.

- The person's maximum heart rate achieved (thalach)



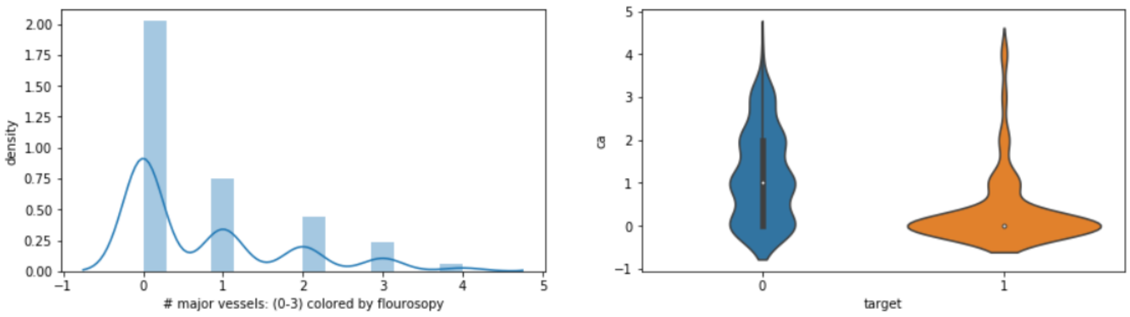
Maximum heart rate achieved is around 160 for most of the people in the dataset and it approximately varies normally around 160. Maximum heart rate for people having heart disease is more than the people not having heart disease.

- ST depression induced by exercise relative to rest (oldpeak)



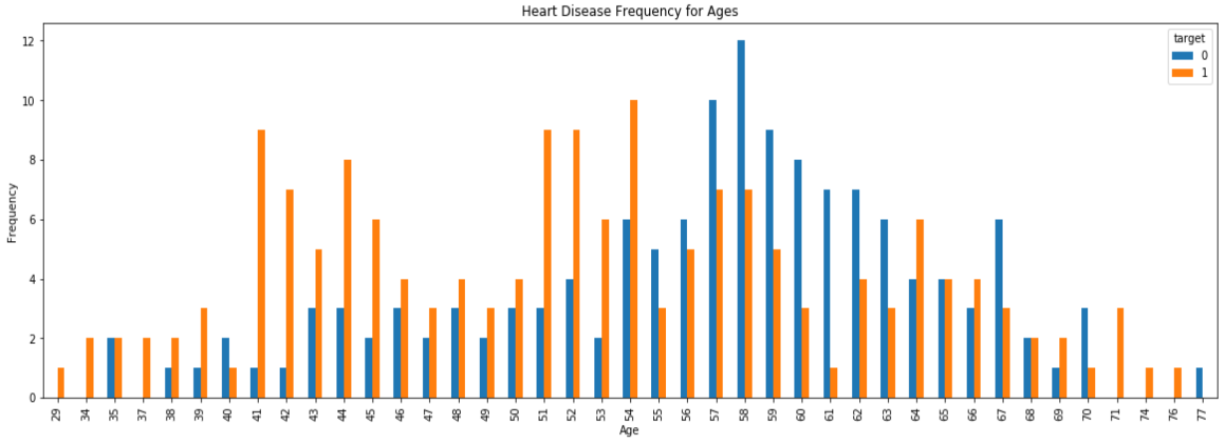
People having ST depression induced by exercise relative to rest nearly equal to zero is highest in the dataset. People having heart disease mostly have oldpeak value less than people not suffering from heart disease.

- The number of major vessels (ca)

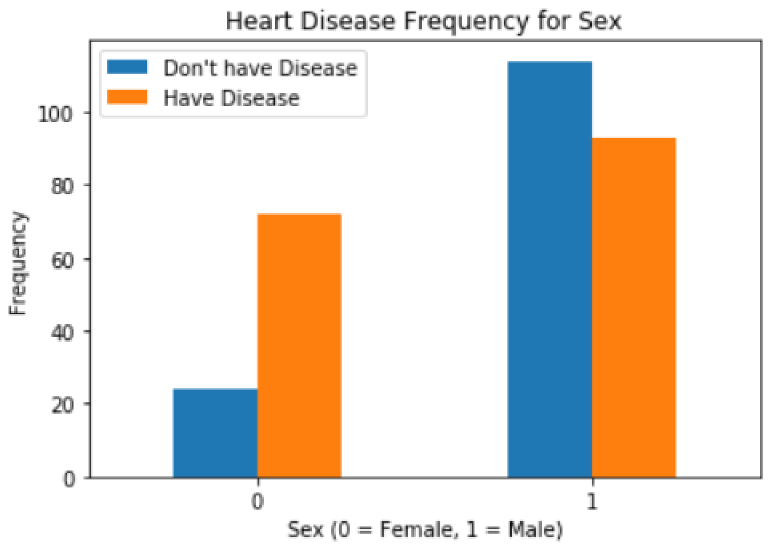


People having major vessels least are more than the people having major vessels highest in the dataset. Very few people having heart disease have more number of major vessels.

- Plot showing the frequency of the number of people versus age with target values

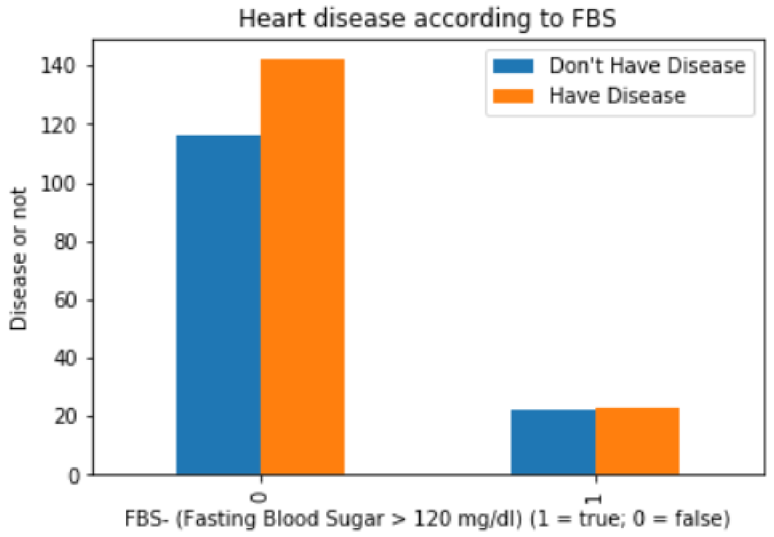


- Plot showing the frequency of the number of people versus sex with target values



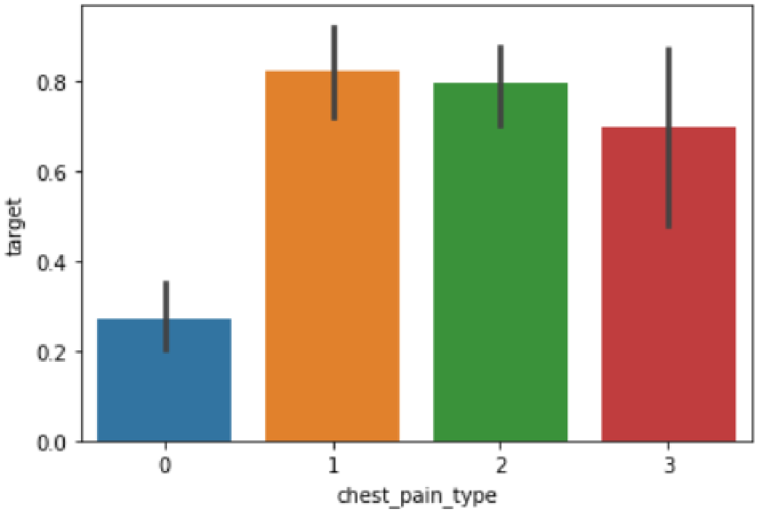
The number of male are more in the dataset than female. Male are less prone to heart disease than female.

- Plot showing the frequency of the number of people versus fbs with target values



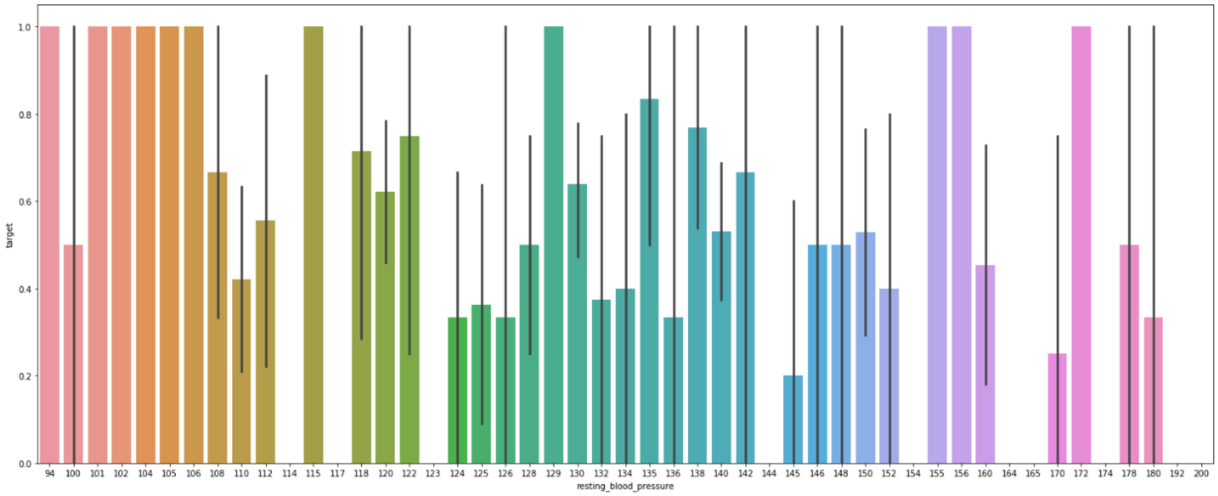
People having fasting blood sugar greater than 120mg/dl are equally probable to heart disease. Number of people having fasting blood sugar less than 120mg/dl are more in the dataset.

- Barplots of attributes
  - Probability of person having disease versus chest pain type



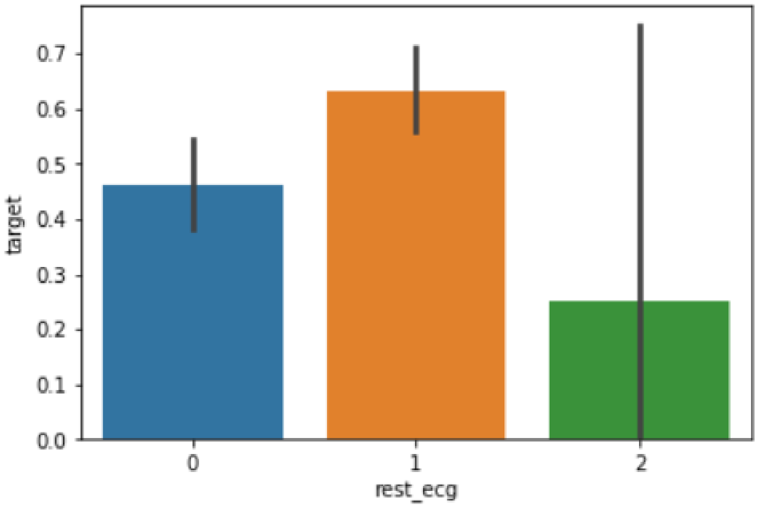
People with chest pain type 1 are more prone to heart disease and people with chest pain type 0 are least prone to heart disease.

- Probability of person having disease versus resting blood pressure



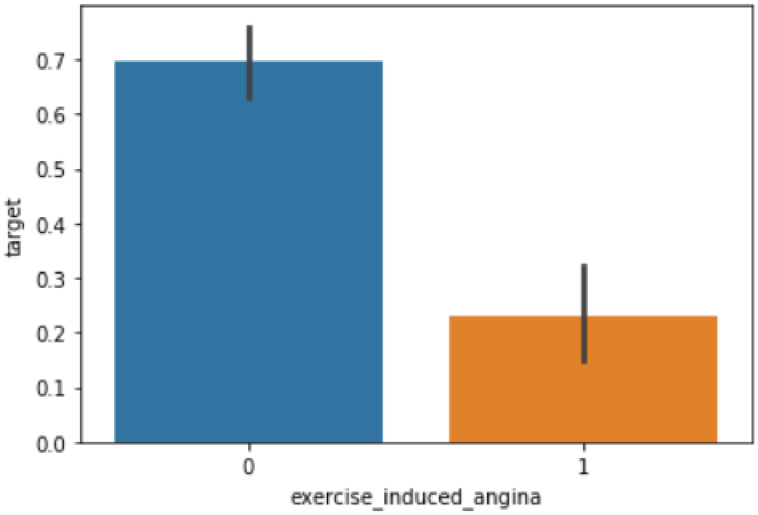
People with high or low blood pressure are more suspectible to heart disease.

- Probability of person having disease versus restecg



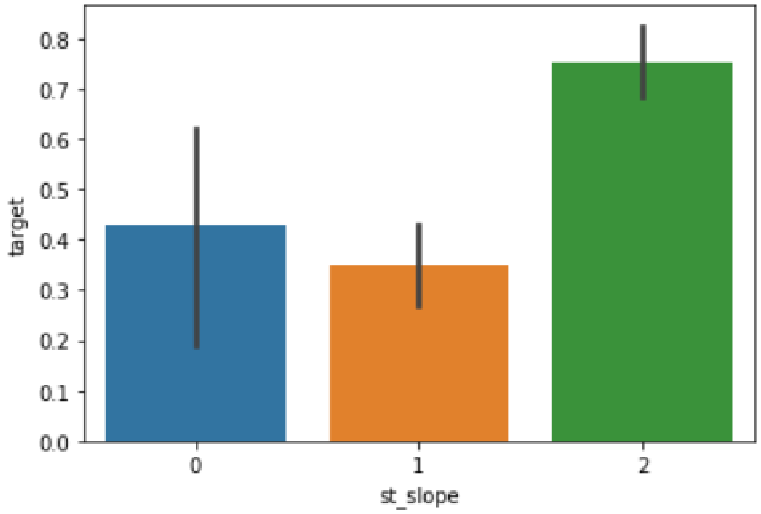
People with resting electrocardiographic measurement as ST-T wave abnormality are more prone to heart disease while people with showing probable or definite left ventricular hypertrophy by Estes' criteria are very less prone to heart disease.

- Probability of person having disease versus exercise induced angina



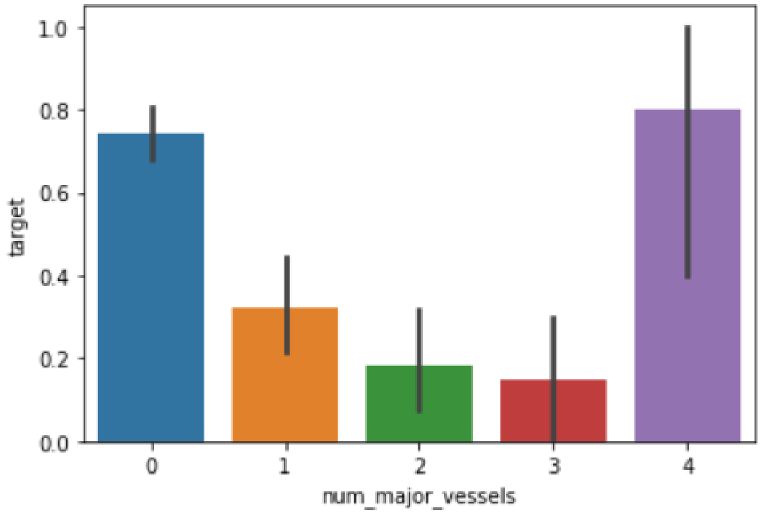
People not having exercised induced angina are prone to heart disease than people having exercise induced angina.

- Probability of person having disease versus st_slope



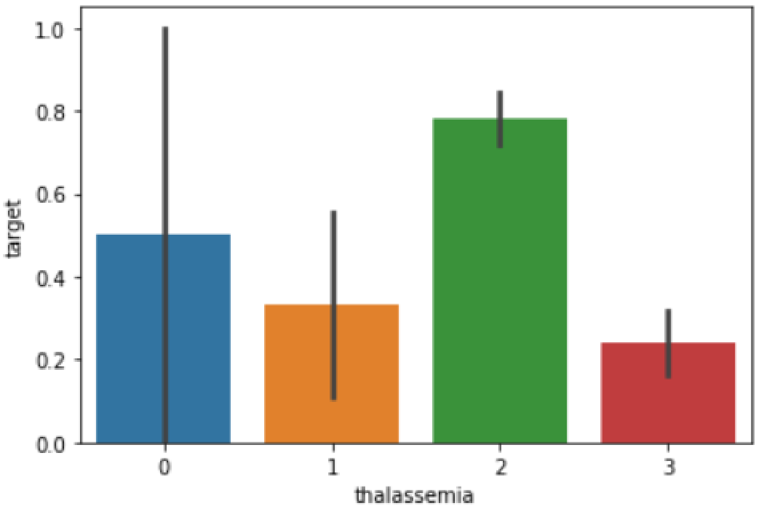
People with the slope of the peak exercise ST segment as downsloping are more prone to heart disease and people with the slope of the peak exercise ST segment as flat are least prone to heart disease.

- Probability of person having disease versus num of major vessels



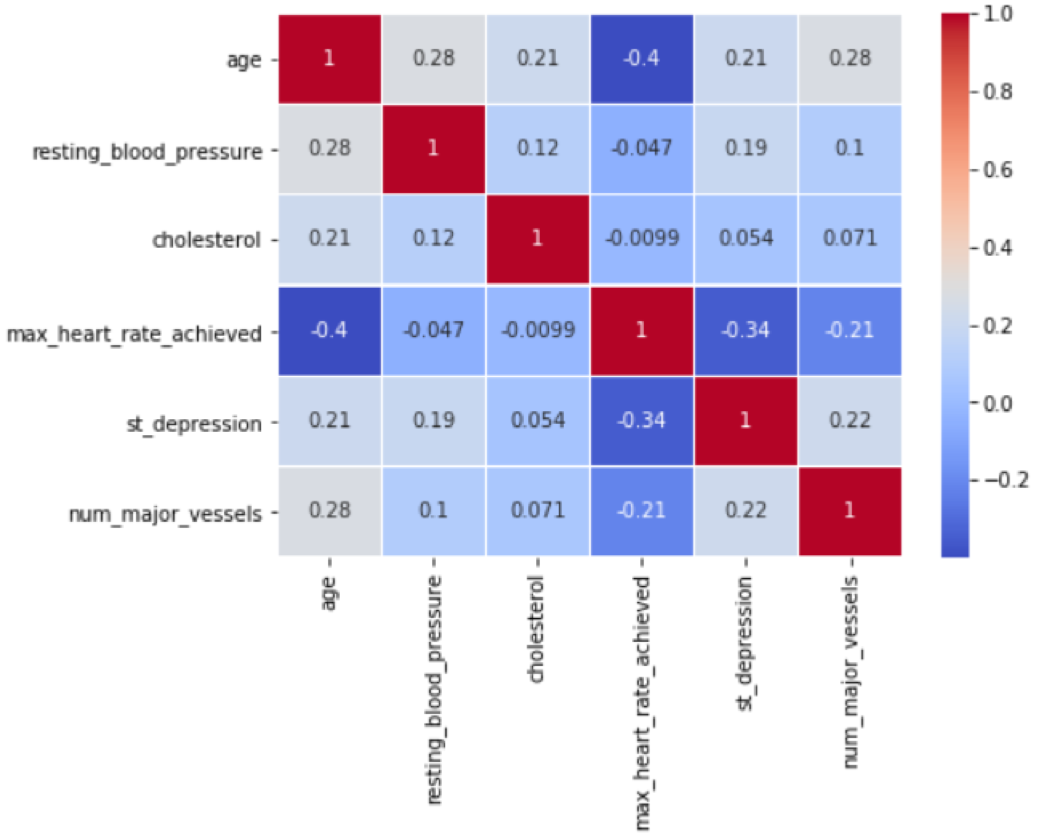
People with number of major vessels 0 and 4 are more suspectible to heart disease and people with number of major vessels are least prone to heart disease.

- Probability of person having disease versus thalassemia



People with reversible defect are more prone to heart disease.

- Correlation plot



Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.
The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.
If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

From the above correlation plot, it can be seen that resting blood pressure and age are highly correlated. Maximum heart rate achieved and ST depression are inversely correlated maximally. Cholesterol and maximum heart rate achieved are least correlated.

DATA SPLIT

- Normalized data is split into test data and training data. The size of training data is 80% and the size of test data is 20%.
- This training data will help the machine to connect the patterns in the data to the right answer. Once trained in this way, a machine can now be given test data that has no answers. The machine will then predict the answers based on the training it received.
- Data can also be divided into three portions: training data, cross-validation data and testing data. The training data is used to make sure the machine recognizes patterns in the data, the cross-validation data is used to ensure better accuracy and efficiency of the algorithm used to train the machine, and the test data is used to see how well the machine can predict new answers based on its training.

Significance of Silhoutte index

Silhouette index is defined as average distance between every sample of a cluster with all available samples of that cluster, and average distance of total current available samples in other clusters with a specific cluster. Based on this, the amount of diversity and data correlation of data is determined. The value of the Silhouette validation index is between -1 and +1. The higher values of this index (near to +1) shows that clustering is made correctly. If the index is near zero, means that we can assign a sample to a nearer cluster, and the sample is located in the same distance from both clusters. If the index becomes -1, it means that the clustering was not made correctly.

S(C) is an OSI (Overall Silhouette index) index for a clustering is the average value of Silhouette for all clusters: The Silhouette index used for evaluating the level of similarity between suggested algorithm and heart patients classifying.

RESULTS

1. Fuzzy clustering algorithm

   When fuzzy clustering algorithm is applied for various values of fuzzy power (m) and number of clusters using Grid Search technique, the best results obtained are as follows:

   - Maximum Silhouette index obtained is 0.2594 when number of clusters is 3 and fuzzy power is 1.5
   - Maximum Fuzzy partition coefficient is 0.7059 when number of clusters is 2 and fuzzy power is 1.5

2. Support vector machine

   - Linear SVC and Non-linear SVC performance difference is explained with the help of ROC curves, accuracy, f1-score.
   - The following figure compares the performance of both the technique(linear SVM and non-linear SVM) with the help of ROC curve.

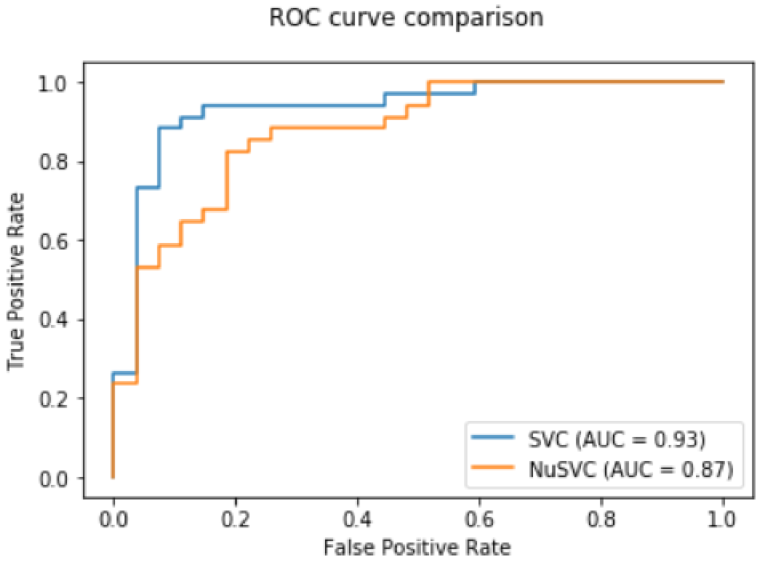


fig: Comparsion of Linear SVM and nonlinear SVM with polynomial kernel

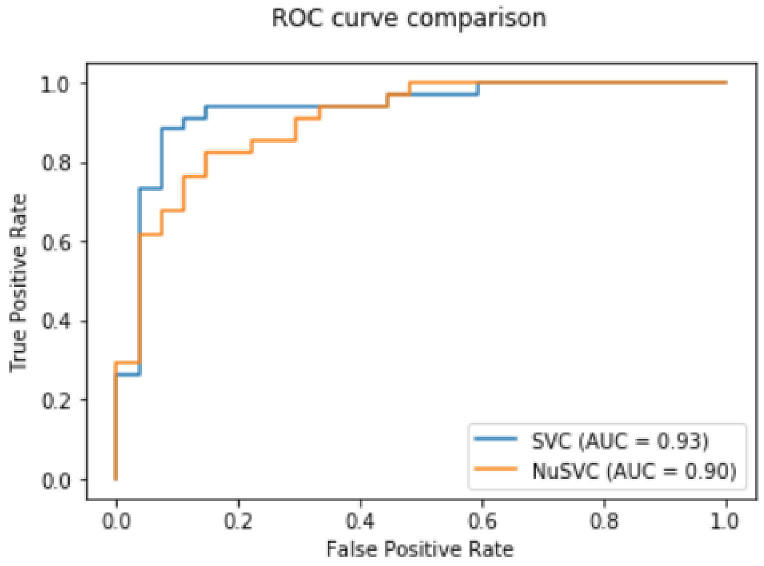
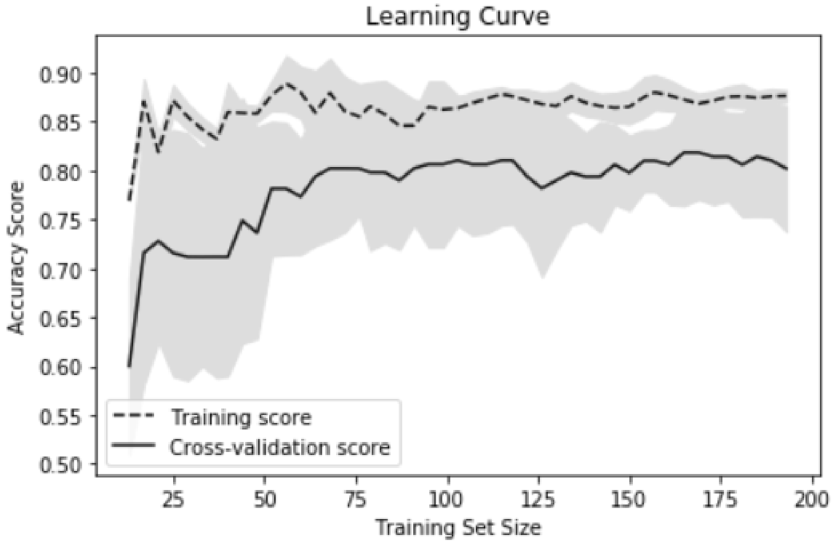


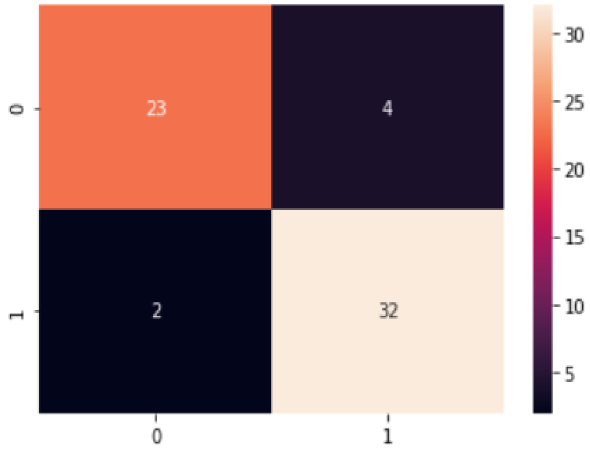
fig: Comparsion of Linear SVM and nonlinear SVM with rbf kernel

It can be seen that as AUC of linear SVC is greater than that of non-linear, linear SVC performs better.

- Learning curve is shown below



- Confusion matrix of linear SVC on test data is as follows:



From the above confusion matrix, it is seen that

I. False positives = 2

II. False negatives = 4

III. True positives = 23

IV. True negatives = 32

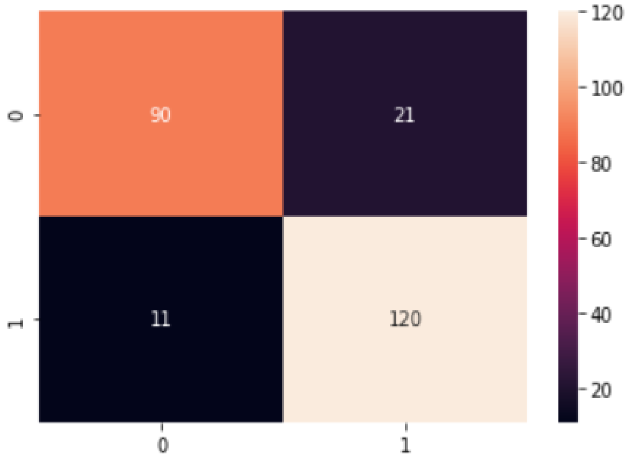
The number of people actually having no heart disease and predicted positive are 2.

The number of people actually having heart disease and predicted negative are 4.

The number of people actually having heart disease and predicted positive are 23.

The number of people actually having no heart disease and predicted negative are 32.

- Confusion matrix of linear SVC on training data is as follows:



From the above confusion matrix, it is seen that

I. False positives = 11

II. False negatives = 21

III. True positives = 90

IV. True negatives = 120

The number of people actually having no heart disease and predicted positive are 11.

The number of people actually having heart disease and predicted negative are 21.

The number of people actually having heart disease and predicted positive are 90.

The number of people actually having no heart disease and predicted negative are 120.

- Comparison with nonlinear SVM

a. Non-linear SVM with RBF kernel

Training accuracy = 80.99%

Test accuracy =80.33%

b. Non-linear SVM with polynomial kernel

Training accuracy = 72.73%

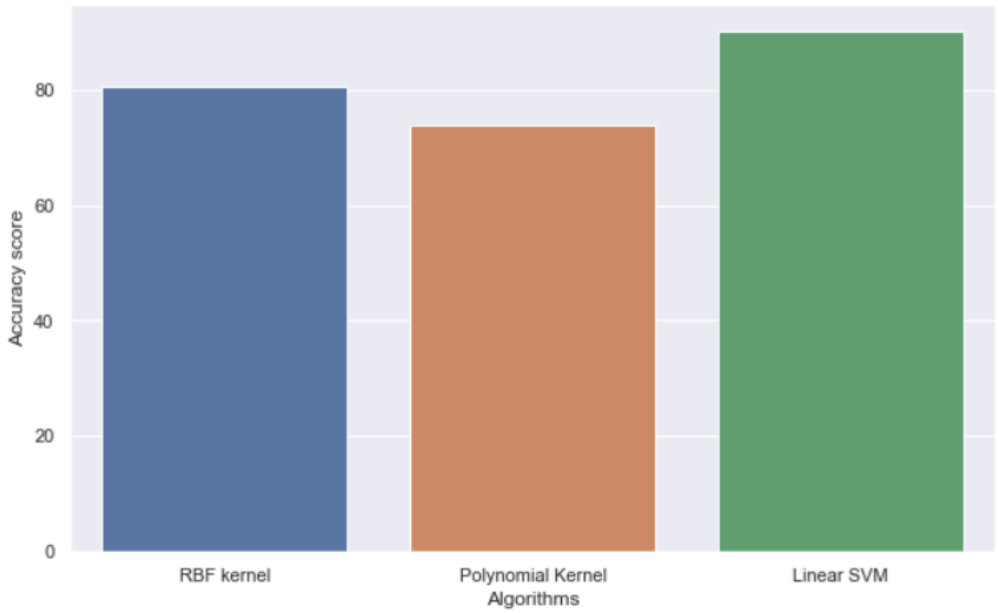
Test accuracy =73.77%

c. Linear SVM

Training accuracy = 86.78%

Test accuracy =90.16%

- Comparsion through Barplot



- Final performance measures on test data are:

  Accuracy = 90.16%

  Precision = 88.88%

  Recall = 94.11%

FUTURE SCOPE

In fuzzy clustering algorithm, selecting parameters like fuzzy power, number of clusters and the initialization of 2d array is very important for the performance of the algorithm. Therefore, various algorithm can be applied for the initialization of 2d array. Complete model can still be worked on to increase the accuracy.

The model can be implemented on real time data. Different health monitoring system can be implemented which shall involve collection of data from the patient wearing sensors or other devices that captures the values of vital parameters like heart rate, blood pressure, respiratory rate, etc. A mechanism has to be developed where the system is able to predict health parameter values by observing the values from the sensors and learning from them. This ensures that any an anomaly that is likely to occur is predicted in advance. Accordingly the doctor can be alerted in advance and some preventive action can be taken to prevent such anomalies. This requires the development of an algorithm to predict the values and observe any abnormal changes in their trend, thus alerting the health professionals in advance.

Now a days, many sites are developed for health monitoring system which provides symptoms and precautions to be taken for many diseases. This can be useful in our project as follows:
We can develop a web platform to predict the occurrences of disease based on various symptoms. The user can select various symptoms and can find the diseases with their probabilistic figures. The project can be improved by implementing medicine suggestion to the patient along with the results. We can implement a feedback from the experienced doctors who can give their views and opinions about certain medicines /practices done by the doctor on the patient. We can implement a live chat option where the patient can chat with a doctor available regarding medication for the respective result for their symptoms. The project could be used as a training tool for Nurses and Doctors who are freshly introduced in the field related to heart diseases. The patient can have a choice in choosing the medicines he/she should take in order to have a healthier life. Moreover, if implemented on a large scale it can be used in medical facilities like hospital,

clinics where a patient wouldn’t have to wait in long queues for treatment if he is feeling symptoms related to heart disease.

CONCLUSION

The overall objective of the project is to predict accurately with less number of tests and attributes the presence of heart disease. In this project, fourteen attributes are considered which form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but the goal is to predict with less number of attributes and faster efficiency to predict the risk of having heart disease at a particular age span.

Combination of supervised and unsupervised learning is used which involves fuzzy clustering followed by support vector machine.

In clustering, the data splits to some clusters, in such a way that the data in every cluster have maximum similarity with each other and minimum similarity with data of other clusters. Uncertainty or fuzziness lying in the data may be reduced as the unsupervised learning model is first applied to the input data and generates certain number of clusters. So using clustering data and original features, SVM is implemented which could help us in predicting that if he/she is under heart attack risk or not.

If only SVM is implemented, uncertainty or fuzziness lying in the data cannot be reduced, therefore the accuracy of model will reduce.

The Algorithms used in the project does not give a 100% accuracy, so the prediction is not 100% feasible. This is because, as seen from confusion matrix there are some false negatives and false positives present. The algorithm will give 100% accuracy only if number of false positive and false negative are reduced to zero. Clinical diagnosis and diagnosis using our project may differ slightly because the prediction is not 100% accurate. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Health care monitoring involves the patient wearing sensors or other devices that capture the values of vital parameters like heart rate, blood pressure, respiratory rate. A mechanism has to be developed where the system is able to predict the health parameter values by observing the values from the sensors and learning from them. This ensures that any an anomaly that is likely to occur is predicted in advance.

Accordingly, the doctor can be alerted in advance and hence some preventive action can be taken to prevent such anomalies.This requires the development of an algorithm to predict the values and observe any abnormal changes in their trend, thus alerting the health professionals in advance.

CHALLENGES FACED DURING THE PROJECT.

- Implementing the model in Python was quiet challenging as python was completely new to learn.
- Choosing the proper model to implement.

REFERENCES.

1. Predicting Heart attacks in patients using Artificial Intelligence Methods
2. A fuzzy clustering neural network architecture for classification of ECG arrhythmias.
3. Detection of Brain Tumor in MRI Images, using Combination of Fuzzy C-Means and SVM.
4. Hybrid Multistage Fuzzy Clustering System for Medical Data Classification.
5. Heart disease prediction using effective machine learning techniques - International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019
6. Heart disease prediction using machine learning techniques - A Survey International Journal of Engineering and technology
   DOI - 10.14419/ijet.v7i2.8.10557
7. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms.
8. A Novel fuzzy clustering Neural Network by Pradeep M. Patil, Manish P. Deshmukh and P.M. Mahajan.
9. Online Incremental Learning Algorithm for Anomaly Detection and Prediction in Health Care by Kirthanaa Raghuraman, Monisha Senthurpandian, Monisha Shanmugasundaram, Bhargavi, V.Vaidehi Department of Information Technology,Madras Institute of Technology, Anna University.
10. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms.
11. A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases by Priyan Malarvizhi Kumar, Usha Devi Gandhi, School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, India.