# Survival Analysis

# Heart Disease –Stroke Prediction

# 

## TEAM: SMART LEARNERS

1. SASANALA SRI LEHA

2. TANGIRALA UDAYA LAKSHMI

3. SUSHMA REDDY NALLAMILLI

4. MADDU ESWARI

5. SAYYAPARAJU SAI DURGA VIAJAYA PREETHI

**Content page**

Topic……………………………………………………… PageNo.

1. Title of the project 3
   1. Introduction 3
   2. Objectives of research 4
   3. Problem Statement 4
   4. Industry Profile 4

1. Review of literature 5
2. Data Collection 5
3. Methodology 6
   1. Exploratory Data Analysis 6

4.1.1 Figures and tables 10

* 1. Statistical Techniques and Data Visualization 10
  2. Data Modeling using Supervised ML techniques 16

1. Findings and Suggestion 18
2. Conclusion 18
3. References 19

**Survival Analysis**

1. **Heart disease and Stroke prediction:**
   1. **Introduction:**

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc. The survival time can be measured in days, weeks, years, etc. For example, if the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack. Many machine learning algorithms are adapted to effectively handle survival data and tackle other challenging problems that arise in real-world data.

The use case we need to deploy is heart disease and stroke prediction.

**Heart disease:**

Heart disease is a general term that means that the heart is not working normally. It generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain or stroke.

* Babies can be born with heart disease. This is called congenital heart disease.
* If people get heart disease later, it is called acquired heart disease. Most heart disease is acquired.

**Symptoms of heart disease:**

* Pain in the chest🡪the heart muscle is not getting enough flow to keep it going.
* Trouble in breathing🡪blood may back up into the lungs.
* Palpitations 🡪a feeling that the heart is beating too fast, too hard, or not regularly.
* Swelling of feet or legs🡪blood is backing up from the heart into the lower body.
* Feeling weak because the body and brain are not getting enough blood to supply them with oxygen.
* Cyanosis (skin turning a blue colour)🡪 means that too little oxygen is in the bloodstream to supply the cells in the body.

**Causes of heart disease:**

* High blood pressure.
* Diabetes.
* Smoking.
* Excessive use of alcohol or caffeine.
* Drug abuse.
* Stress.

**How to prevent heart disease:**

* Control your blood pressure.
* Keep your cholesterol levels under control.
* Stay at a healthy weight.
* Eat a healthy diet.
* Get regular exercise.
* Manage stress.
  1. **Objectives Of Research:**

Machine Learning plays an essential role in predicting presence of heart disease.Heart disease is a general term that means that the heart is not working normally. It generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain or stroke.

X-variables taken are age, sex, cp ,chol, thalach ,exang, thal and Y is a target variable.

* 1. **Problem Statement:**

In order to choose the best classification model for the heart disease and stroke prediction, we should apply different classification algorithms such as K-Nearest Neighbour, Support vector machine, Naïve bayes, Random forest, Support Vector Machine Kernel, Logistic Regression, Decision tree.

The main goal is to predict the binary class heart-disease-present, which represents whether or not a patient has heart disease:

* 0 represents no heart disease present
* 1 represents heart disease present
  1. **Industry profile:**

Machine Learning plays an essential role in predicting presence of heart disease.Heart disease is a general term that means that the heart is not working normally. It generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain or stroke.

Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today’s time. This project aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction.

The available data mining techniques are used in the diagnosis of Heart disease over the heart disease dataset. Data mining is very useful tool for the diagnosis of diseases. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis.

1. **Review of literature:**

Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today’s time. This project aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction.

The available data mining techniques are used in the diagnosis of Heart disease over the heart disease dataset. Data mining is very useful tool for the diagnosis of diseases. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Mining tools and techniques, various fields of business get benefit by easily evaluate various trends and pattern of market and to produce quick and effective market trend analysis.

1. **Data Collection:**

The data is collected from the UCI machine learning repository. The data set is named Heart Disease Data Set and can be found in the UCI machine learning repository. The UCI machine learning repository contains a vast and varied amount of datasets which include datasets from various domains. These data are widely used by machine learning community from novices to experts to understand data empirically.

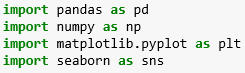
There are 14 columns in the dataset:

* Age- (type: integer): age in years
* sex -(type: binary): 0: female, 1: male
* cp-Chest-pain-type (type: integer): chest pain type (4 values).
* Trestbps-resting-blood-pressure (type: integer): resting blood pressure
* Chol-serum-cholesterol (type: integer): serum cholestoral in mg/dl
* Fbs-fasting-blood-sugar (type: binary): fasting blood sugar > 120 mg/dl
* Restecg-resting-ecg-results (type: integer): resting electrocardiographic results (values 0,1,2)
* Thalach-max-heart-rate-achieved (type: integer): maximum heart rate achieved (beats per minute)
* Exang-exercise-induced-angina (type: binary): exercise-induced chest pain (0: False, 1: True)
* Oldpeak-eq-st-depression (type: float): oldpeak = [ST depression](https://en.wikipedia.org/wiki/ST_depression) induced by exercise relative to rest, a measure of abnormality in electrocardiograms
* slope (type: integer): the slope is an electrocardiography read out indicating quality of blood flow to the heart
* ca-Num-major-vessels (type: integer): number of major vessels (0-3).
* Thal (type: categorical): results of [thallium stress test](https://www.ucsfbenioffchildrens.org/tests/007201.html) measuring blood flow to the heart, with possible values normal, fixed-defect, reversible-defect
* target: diagnosis of heart disease (angiographic disease status)   
  -- Value 0: < 50% diameter narrowing   
  -- Value 1: > 50% diameter narrowing

1. **Methodology:**

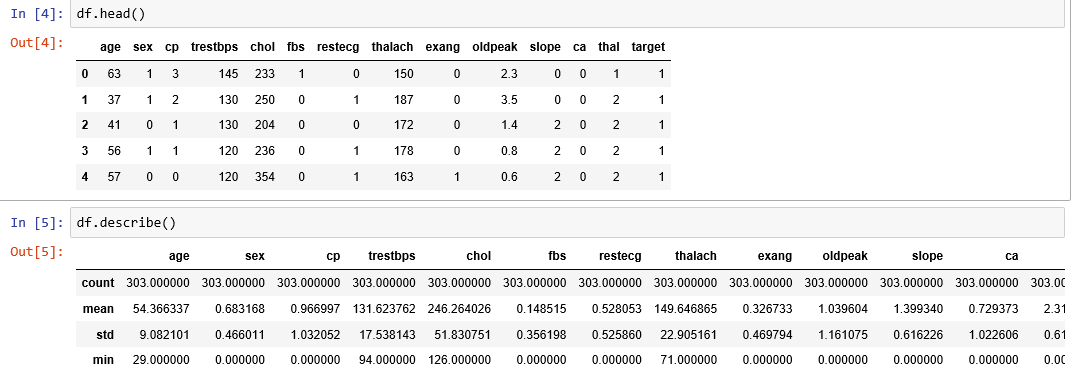
In this project, comparison of various machine learning methods is done for predicting the risk of coronary heart disease of the patients from their medical data. The heart disease data set is taken as input. It is then pre-processed by replacing non-available values with column means. Different methods were used in this project. The output is the accuracy metrics of the machine learning models. The model can then be used in prediction.

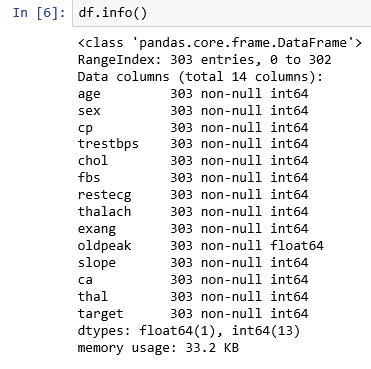
* 1. **Exploratory Data Analysis:**
* Importing the libraries

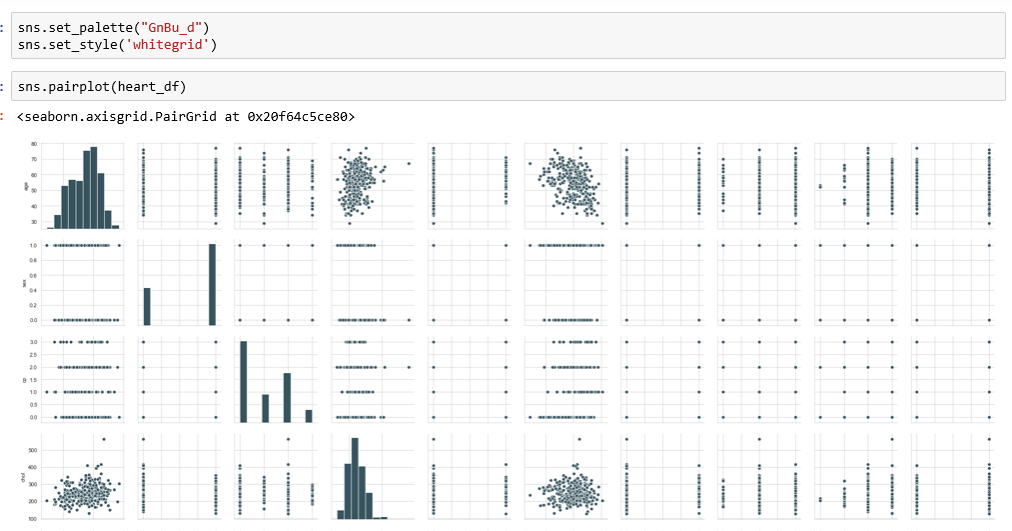


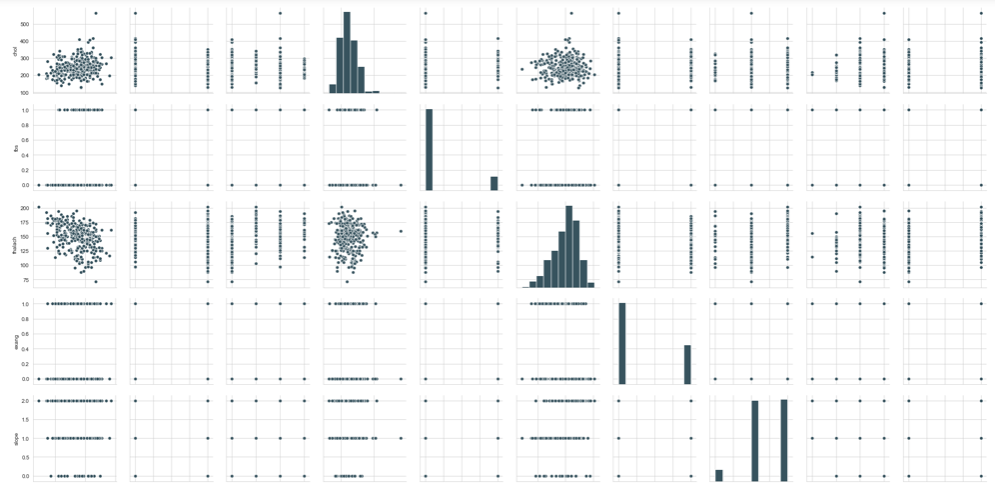
* Importing the Data set

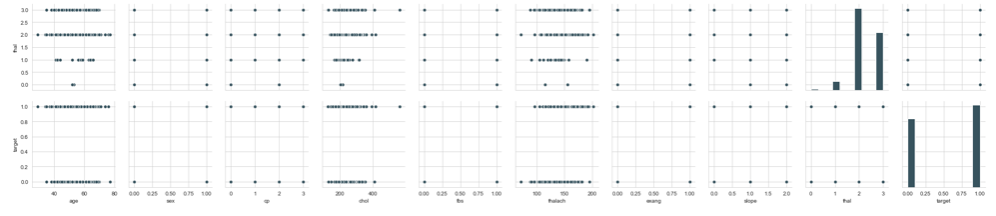








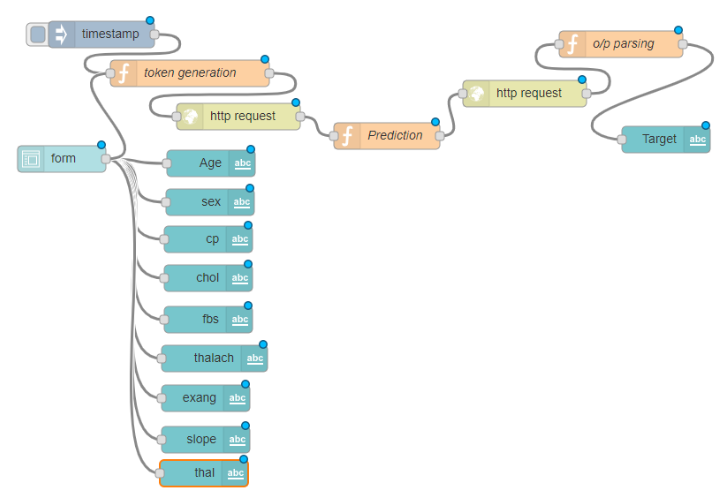


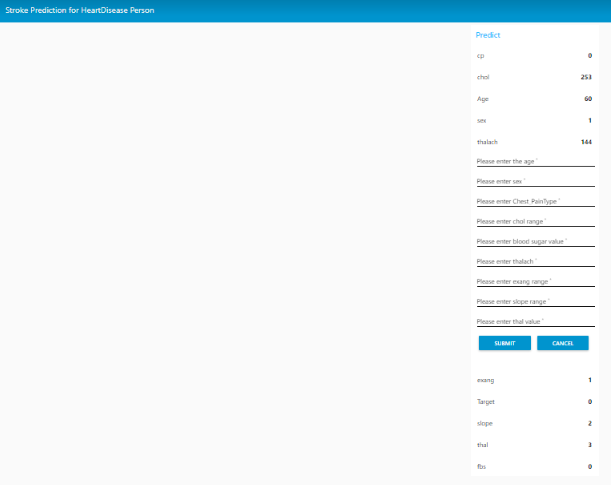
**Node Red Flow:**

Node-Red is a programming tool for writing together hardware devices,APIs and online services in new and interesting ways.

It provides a browser based editor that makes it easy to wire together flows using the wide range of nodes in the palette that can be deployed to its runtime in a single-click.

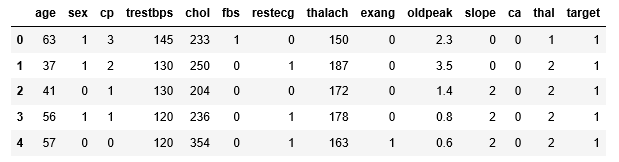
First the token should be generated and after generating the token it should be connected to the http request and all the needed attributes should be attached to the form and the http request wiil generate the url and the target variable will be predicted and parsed using the o/p parsing.



UI 

* + 1. **Figures and tables:**

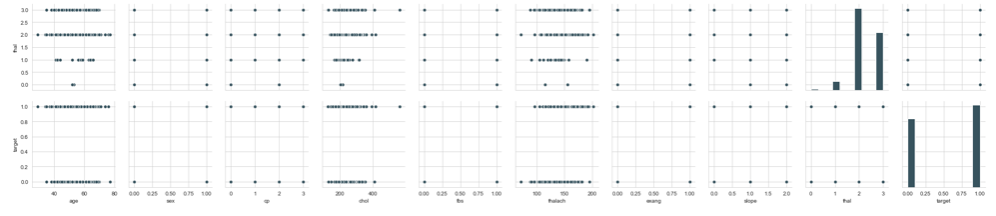
**Dataset: heart.csv**

****

**Evaluation metrics of different classification models:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sno | Model | Recall | Precision | F1 Score | Support | Accuracy |
| 1 | Navie Bayes | 0.71 | 0.71 | 0.71 | 76 | 0.71 |
| 2 | Logistic Regression | 0.76 | 0.76 | 0.76 | 76 | 0.76 |
| 3 | Decision Tree | 0.71 | 0.71 | 0.71 | 76 | 0.71 |
| 4 | Random Forest | 0.76 | 0.77 | 0.76 | 76 | 0.76 |
| 5 | KNN | 0.82 | 0.82 | 0.82 | 76 | 0.81 |
| 6 | SVM | 0.76 | 0.76 | 0.76 | 76 | 0.76 |
| 7 | SVM Kernel | 0.75 | 0.75 | 0.75 | 76 | 0.75 |

**Pair plot:**



* 1. **Statistical Techniques and Data Visualization:**

**#Scatter plot:**

A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis.

#scatter plot visualization

col=df.columns

col Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'], dtype='object')

for i in col:

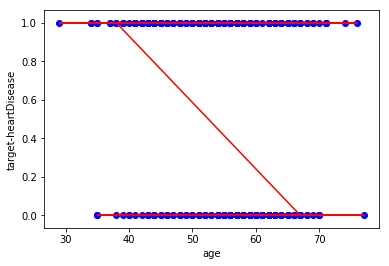
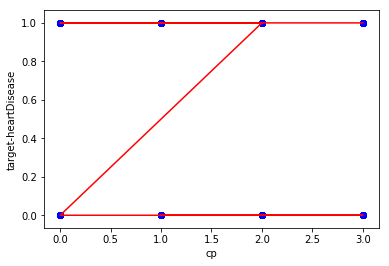
plt.scatter(df[i],df['target'],color='b')

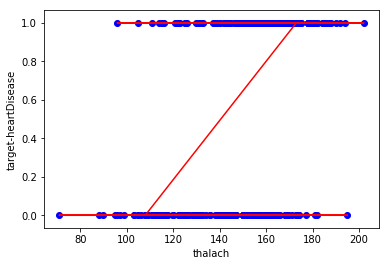
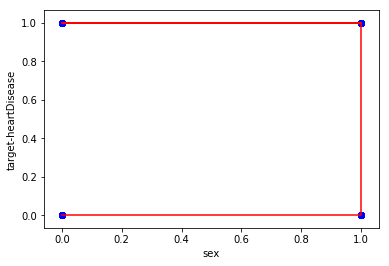
plt.plot(df[i],df['target'],color='r')

plt.xlabel(i)

plt.ylabel('target-heartDisease')

plt.show()

  
**#Histogram:**

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

#Histogram visualization

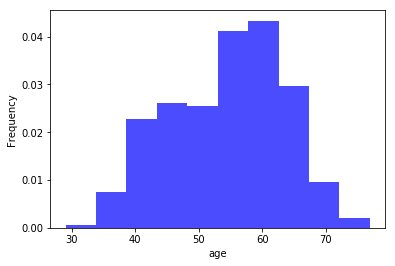
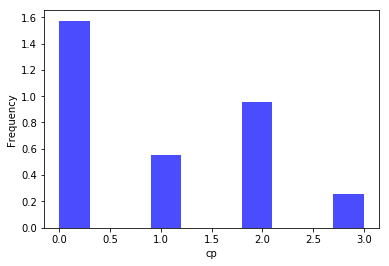
for i in col:

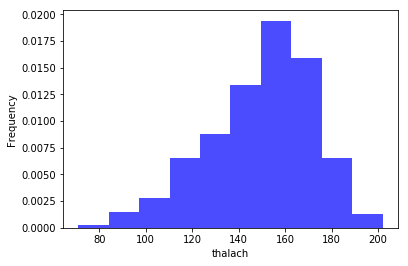
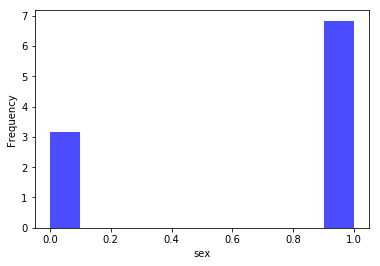
plt.hist(df[i],bins=10,normed=1, facecolor='blue', alpha=0.7)

plt.xlabel(i)

plt.ylabel('Frequency')

plt.show()



**#Bar plot:**

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph.

#Bar plot visualization

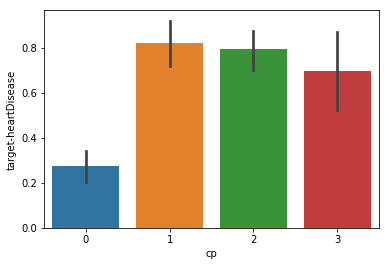
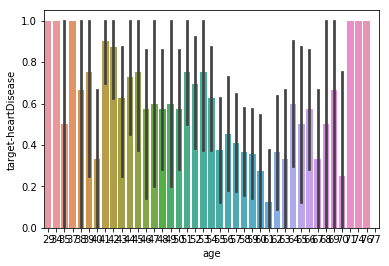
for i in col:

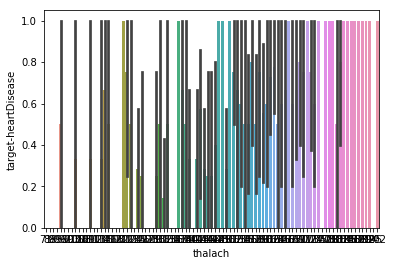
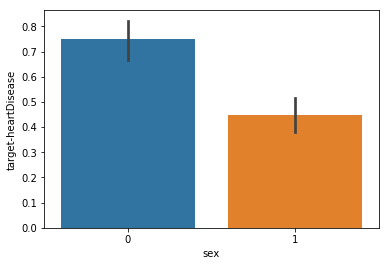
sns.barplot(df[i],df['target'])

plt.xlabel(i)

plt.ylabel('target-heartDisease')

plt.show()





**#Counter Plot:**

A contour plot is a graphical technique for representing a 3-dimensional surface by plotting constant z slices, called contours, on a 2-dimensional format.

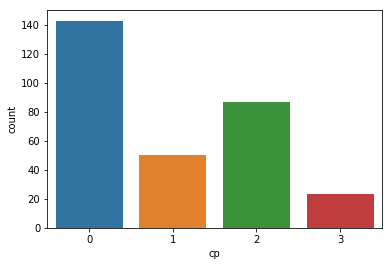
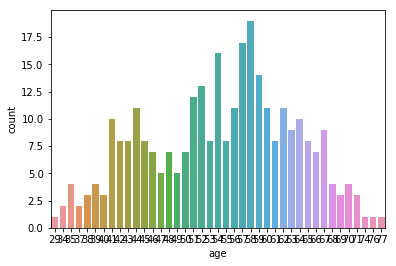
#counter plot visualization

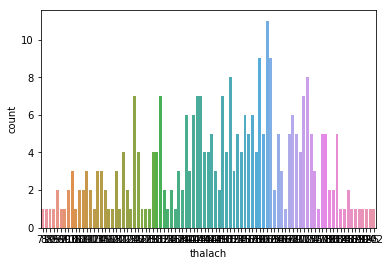
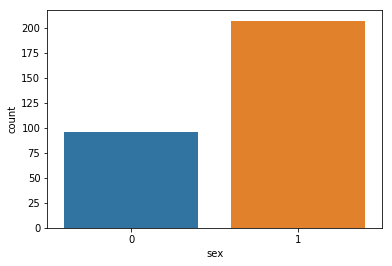
for i in col:

sns.countplot(x=i,data=df)

plt.xlabel(i)

plt.show()





**#Box Plot:**

A simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.

#Box plot visualization

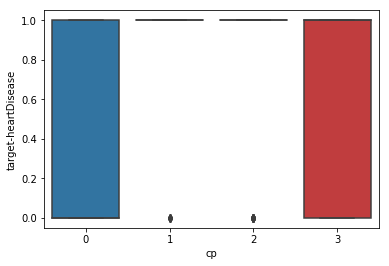
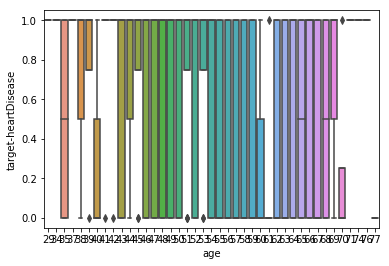
for i in col:

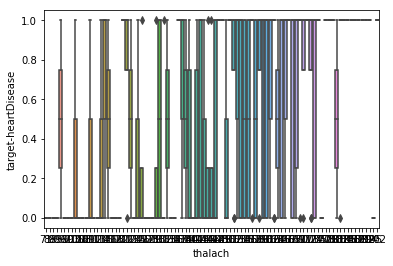
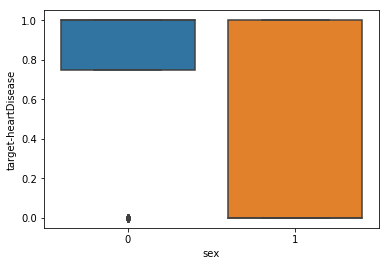
sns.boxplot(x=i,y='target',data=df)

plt.xlabel(i)

plt.ylabel('target-heartDisease')

plt.show()





**#Dist plot:**

Flexibly plot a univariate distribution of observations. This function combines the matplotlib hist function (with automatic calculation of a good default bin size) with the seaborn [kdeplot()](https://seaborn.pydata.org/generated/seaborn.kdeplot.html" \l "seaborn.kdeplot" \o "seaborn.kdeplot) and [rugplot()](https://seaborn.pydata.org/generated/seaborn.rugplot.html" \l "seaborn.rugplot" \o "seaborn.rugplot) functions.

#Dist plot visualization

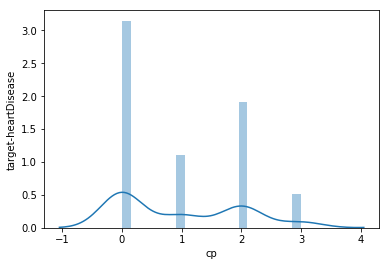
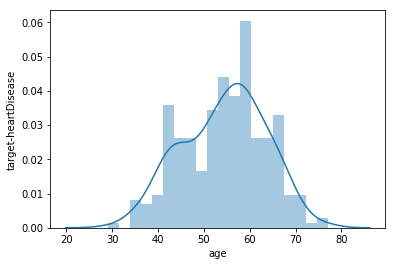
for i in col:

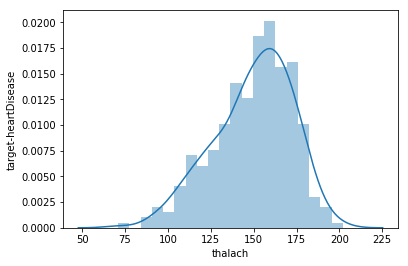
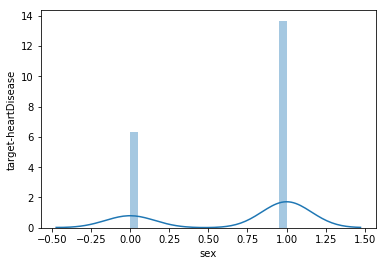
sns.distplot(df[i],bins=20,kde=True)

plt.xlabel(i)

plt.ylabel('target-heartDisease')

plt.show()



* 1. **Data Modeling using Supervised ML techniques:**

**Supervised ML techniques:**

There are two types of Supervised Learning techniques: Regression and Classification. Classification separates the data, Regression fits the data.

* **Regression:**

Regression is a technique that aims to reproduce the output value.

* **Classification:**

Classification is a technique that aims to reproduce class assignments. It can predict the response value and the data is separated into “classes”.

The data modeling techniques are as follows:

* **Logistic Regression:**

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

* **Support vector machine:**

The SVM algorithm is used to predict this disease by plotting the train dataset where a hyper plane classifies the points into two -presence and absence of heart disease. SVM works by identifying the hyper plane which maximizes the margin between two classes. SVM algorithms uses a set of mathematical functions called kernel.



* **Naive bayes algorithm:**

This is a classification algorithm which is used when the dimensionality of the input is very high. A Naive bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is based on bayes theorem. The bayes theorem is as follows:

P(Y/X) = P(X/Y) P(X)

* **Decision trees:**

Decision tree is one of the ways to display an algorithm. It is a classic machine learning algorithm. In heart disease, there are several factors such as cigarette, BP, Hypertension, age etc. The challenge of the decision tree lies in the selection of the root node. This factor used in root node must clearly classify the data. We make use of age as the root node. The decision tree is easy to interpret. They are non-parametric and they implicitly do feature selection.

* **K-Nearest Neighbour (KNN):**

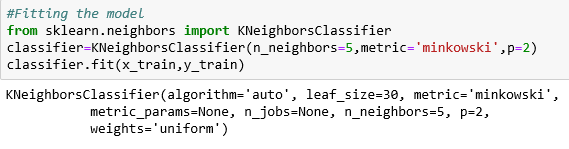
KNN is a non-parametric machine learning algorithm. The KNN algorithm is a supervised learning method. This means that all the data is labeled and the algorithm learns to predict the output from the input data. It performs well even if the training data is large and contains noisy values. The data is divided into training and test sets. The train set is used for model building and training. A k- value is decided which is often the square root of the number of observations. Now the test data is predicted on the model built. There are different distance measures. For continuous variables, Euclidean distance, Manhattan distance and Minkowski distance measures can be used. However, the commonly used measure is Euclidean distance. The formula for Euclidean distance is as follows:

K-Nearest Neighbour

**Steps for KNN:** 

* s1: choose k value of neighbours.
* s2: take k nearest neighbours of new data point, acc to Euclidian.
* s3: Among k neighbors, count the no of data points in each category.
* s4: Assign new data point to the category where you counted most neighbours.
* S5: Model is ready.

**Fitting the model to the Training and test sets**



Sklearn.neighbours should import the library KNeighbours and should best fit.

The no of neighbours should be 5 and the metric and p represents the Euclidian distance.

**Euclidian distance:**



1. **Findings and Suggestion:**

From the data set heart disease and stroke prediction, comparing the sex attribute i.e male and female, male is having the highest probability of getting the heart disease, due to high cholesterol level in the blood and bad diet, and also heavy weight so to control the heart disease they should eat a balanced diet and weight by doing so, the cholesterol level will also get balanced.

While training and modeling the data set more data values need to be given and trained by doing this when we give the raw input values we will get the correct predicted values.

1. **Conclusion:**

* Heart disease is one of the leading causes of death worldwide and the early prediction of heart disease is important.
* The computer aided heart disease prediction system helps the physician as a tool for heart disease diagnosis.
* The purpose of this work was to compare algorithms with different performance measures using machine learning.
* All data were pre-processed and used for test prediction. Each algorithm worked better in some situations and worse in others.
* The proposed methods are compared to supervised algorithms based on existing approximate sets and classification accuracy measurements are used to evaluate the performance of the proposed approaches.
* From the analysis it is concluded that, data mining plays a major role in heart disease classification.

1. **References:**

* “The Atlas of Heart Disease and Stroke”, [online]. <http://www.who.int/cardiovascular_diseases/resources/atlas/en>
* J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, “Big data analytics to improve cardiovascular care: promise and challenges”,Nature Reviews Cardiology, Vol.13, No.6, pp.350, 2016.
* W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama, and I. C.Paschalidis, “Prediction of hospitalization due to heart diseases by supervised learning methods”, International Journal of Medical Informatics, Vol.84, No.3, pp.189–197, 2015.
* <http://bigdata-madesimple.com/14-useful-applications-of-data-mining>
* <http://www.slideshare.net/bigml/big-ml-spring-2014-webinar-clustering>
* <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>
* K(x,x’) = exp((-||x-x’||2)/2σ2

Support Vector Machine (SVM)

Support Vector Machine (SVM)

Support Vector Machine (SVM)

Support Vector Machine (SVM)

Support Vector Machine (SVM)