**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM,

APPROVED BY AICTE & GOVT.OF KARNATAKA)

Logo

Description automatically generated

**Learning Activity Project Report**

on

**Title: HEART DISEASE PREDICTION**

*Submitted in partial fulfilment of the requirement for the award of Degree of*

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

**Faculty In-charge: Dr. VANI VASUDEVAN**

**Submitted by:**

D S SAI JAHNAVI 1NT19CS062

JALPA C J 1NT19CS088

LEENA JENNIFER EDWIN 1NT19CS100

MAHATHI D N 1NT19CS104

## Abstract:

# The heart is the most crucial & critical organ of the human body. Life is completely dependent on the efficient working & functioning of our heart. It is one of the major causes of mortality in today's world. Heart disease remains one of the most serious health issues of our day. It is said to be the primary motive in death globally. Many times it's difficult for medical professionals to expect a heart disease on time. Nowadays, the health sector contains a lot of precious hidden facts & information which could prove to be very helpful in making predictive decisions especially in the field of medicine. Data mining is a method or technique used to analyze vast datasets and then derive significant and useful results with the use of extraordinary AI-based techniques. This article attempts to use three of these data mining techniques namely Decision Tree, Naive Bayes, & KNN for forecasting cardiovascular or heart disease. All of these methods will be evaluated based on different unique & parameters with optimizations for better accuracy. The accuracy of each method will then be compared depending on accuracy based on various parameters. The best & accurate technique is then implemented for predicting whether or not a man or a woman will have coronary heart disease. This technique can be used by medical practitioners for early prediction of the disease so that timely care can be taken by the patient.

# 1. Introduction

## 1.1 Motivation:

## With the rampant increase in the heart stroke rates at juvenile ages, we need to put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc.

## 1.2 Problem domain:

## Today, data mining in healthcare is used mainly for predicting various diseases, assisting with diagnosis and advising doctors in making clinical decisions. But, the potential of data mining is much bigger – it can provide question-based answers, anomaly-based discoveries, provide more informed decisions, probability measures, predictive modeling, and decision support

## Data mining optimizes physician decision-making and predicts patient problems

## Data mining is used for detection and prevention of the disease.

## data mining providers can develop smart methodologies for treatment, best standards of medical and care practices

## 1.3 Aim and Objectives:

* The aim of this is to employ and analyze different data mining techniques for the prediction of heart disease in a patient through extraction of interesting patterns from the dataset using vital parameters.
* We determine it by using data mining technique like classification, in which we use various types of classification algorithms like KNN algorithm, Decision Tree algorithm, Naïve bayes to predict the output of categorical variables.

# 2. Data source and Data Quality

## 2.1 Dataset Used

* This dataset consists of features that can be used to predict which patients have a high risk of heart disease.
* The used attributes are
  + Age: age in years
  + Sex: sex (1 = male; 0 = female)
  + ChestpainType: chest pain type

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomatic

* + BP: resting blood pressure (in mm Hg on admission to the hospital)
  + Cholesterol: serum cholestoral in mg/dl
  + FBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
  + Ecg: resting electrocardiographic results

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

* + MaxHR: maximum heart rate achieved
  + ExerciseAngina: exercise induced angina (1 = yes; 0 = no)
  + Stdepression
  + SlopeofST: the slope of the peak exercise ST segment

-- Value 1: upsloping

-- Value 2: flat

-- Value 3: downsloping

* + Nov: number of major vessels (0-3) colored by flourosopy
  + Thalium: 3 = normal; 6 = fixed defect; 7 = reversable defect
  + Target: 0=heart disease absence ;1=heart disease presence

* We collected the data form the link mentioned below.

## <https://www.kaggle.com/rishidamarla/heart-disease-prediction>

## 2.2 Data Pre-processing:

* Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model.
* The steps involved in this are
  + Getting the dataset

## dataset = pd.read\_csv("Heart\_Disease\_Prediction (3).csv")

## Importing libraries

## import numpy as np

## import pandas as pd

## import matplotlib.pyplot as plt

## import seaborn as sns

## from sklearn.model\_selection import train\_test\_split

## from sklearn.tree import DecisionTreeClassifier

## rom sklearn.naive\_bayes import GaussianNB

## from sklearn.neighbors import KNeighborsClassifier

## 

## 

# Here, In our Dataset there are no missing values, null values, redundancy in the values.

# 3. Methods and Models

# 3.1 Data Mining Questions

# Question 1:

# Calculate the percentage of the population having heart disease and percentage of population not having heart disease.

# Question 2:

# With the help of the dataset given, Find out and plot the graph for among the males and females, who are suffering with heart problems more?

# Question 3:

# With the given dataset, plot the graph for age vs bp and find out which age group more likely to have more bp?

# Question 4:

# With the given dataset, plot the graph for sex vs cholesterol and find out which gender more likely to have cholesterol?

# Question 5:

# From the dataset, plot ST Depression vs age graph and analyse the result.

# Question 6:

# From the dataset, plot Exercise Angima vs Heart Disease graph and analyse the result.

# Question 7:

# Plot heatmap for the correlation data of the given dataset.

# Question 8:

# Find Accuracy score of test dataset using Decision Tree

# Question 9:

# Find the Accuracy Score of the test dataset using Naive Bayes.

# Question 10:

# Find the Accuracy score of the test dataset using K Nearest Neighbors.

## 3.2 Data Mining Algorithms

* An “algorithm” in machine learning is a procedure that is run on data to create a machine learning “model.”
* The data mining Algorithms that are used in this project is
* **Decision Tree algorithms**: - It is a supervised learning Algorithm, this algorithm builds decision trees using a top-down, greedy approach.
* **KNN algorithm**: - Used to solve both classification and regression problems. It's easy to implement and understand.
* **Naive Bayes** is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

## 3.3 Data Mining Models

* A “model” in machine learning is the output of a machine learning algorithm run on data.
* A model represents what was learned by a machine learning algorithm.
* The data mining model that are used in this project is Predictive model, in which we used classification.

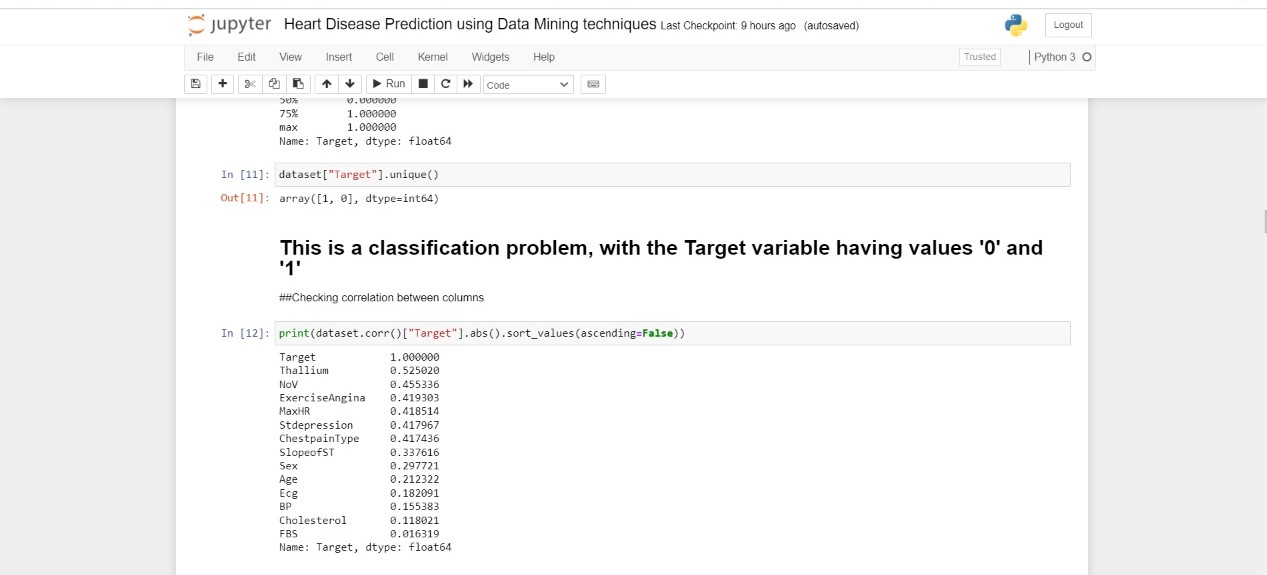
A predictive model constitutes prediction concern values of data using known results found from various data. Predictive modelling may be made based on the use of variant historical data. Predictive model data mining tasks comprise regression, time series analysis, classification, prediction.

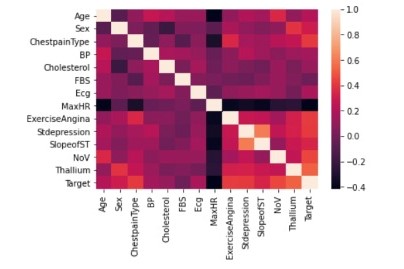
The Predictive Model is known as Statistical Regression. It is a monitoring learning technique that Incorporates an explication of the dependency of few attribute values upon the values of other attributes In a similar item and the growth of a model that can predict these attribute values for recent cases.

Classification – Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

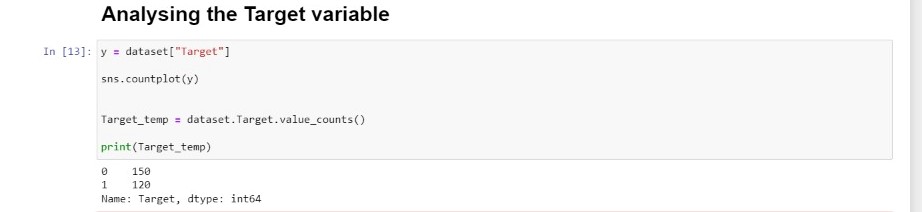
# 4. Model Evaluation & Discussion (with necessary visualizations)

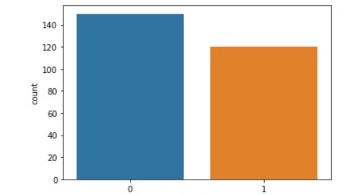
* Model Evaluation is the process through which we quantify the quality of a system's prediction.

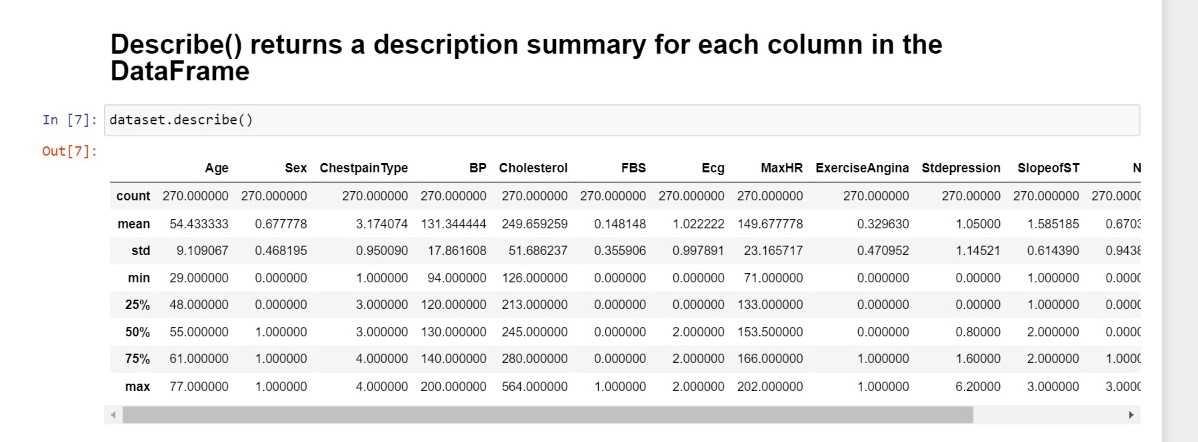


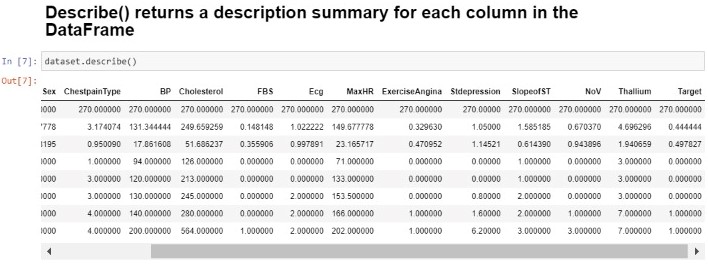


A heatmap contains values representing various shades of the same colour for each value to be plotted. Usually the darker shades of the chart represent higher values than the lighter shade

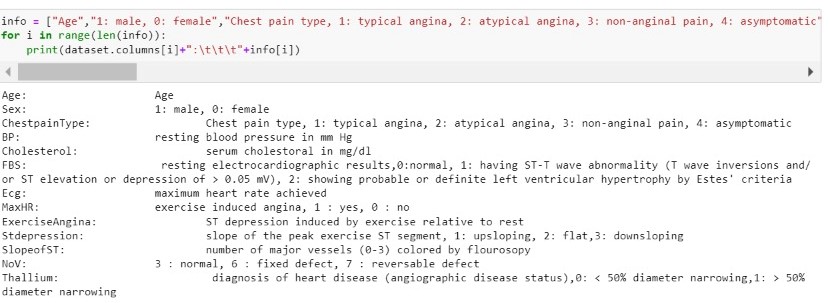
Analysing the Target variable:  




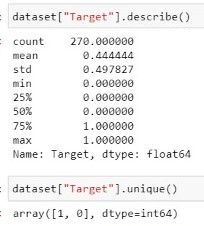




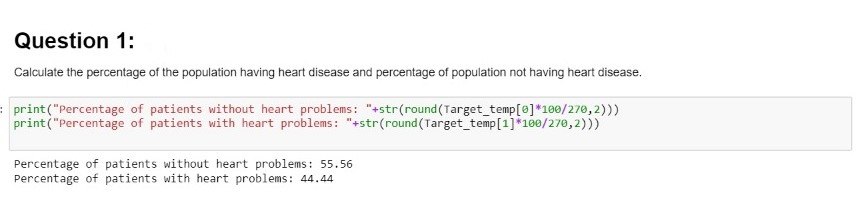
Description of the Attributes:



**Calculation of count, Mean, Standard Deviation, Minimum value, maximum value**

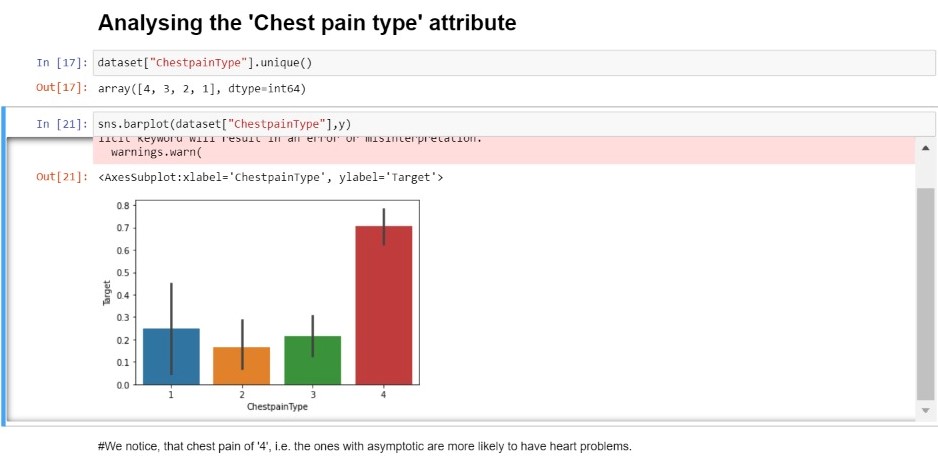


**Data Mining questions with Solutions and it’s visualization**

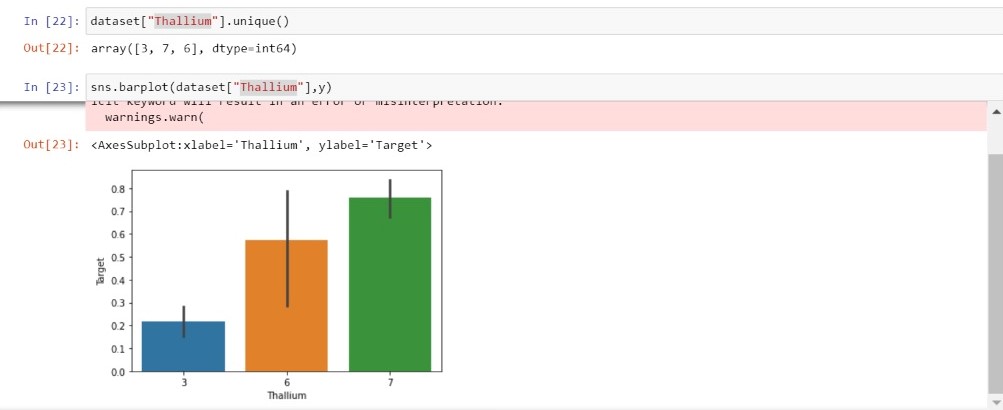


### 

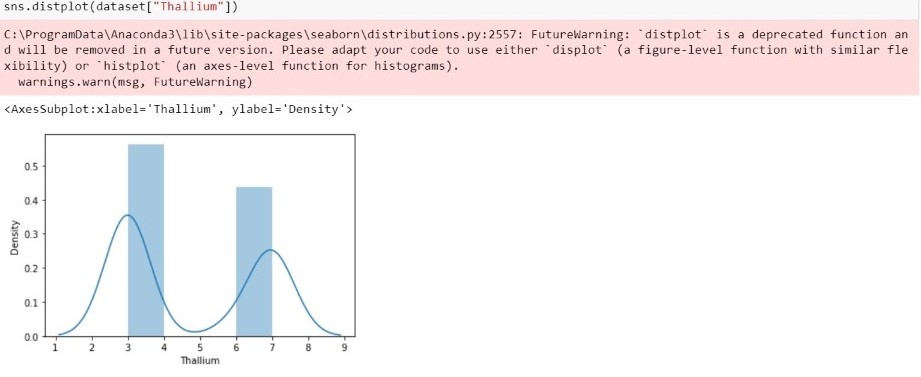
From this dataset we notice, that males are more likely to have heart problems than females. Males with sex attribute value=1 having target value = 0.55 while females with sex attribute value =0 having target value=0.24.

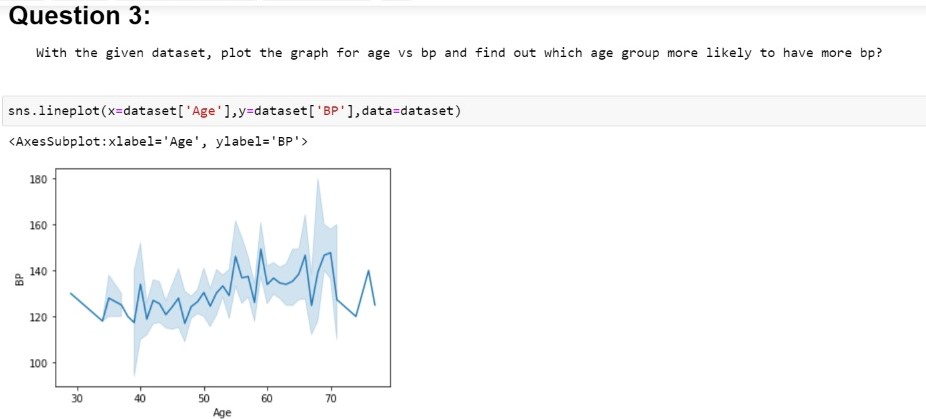


Visualization of Thallium Attribute with Target

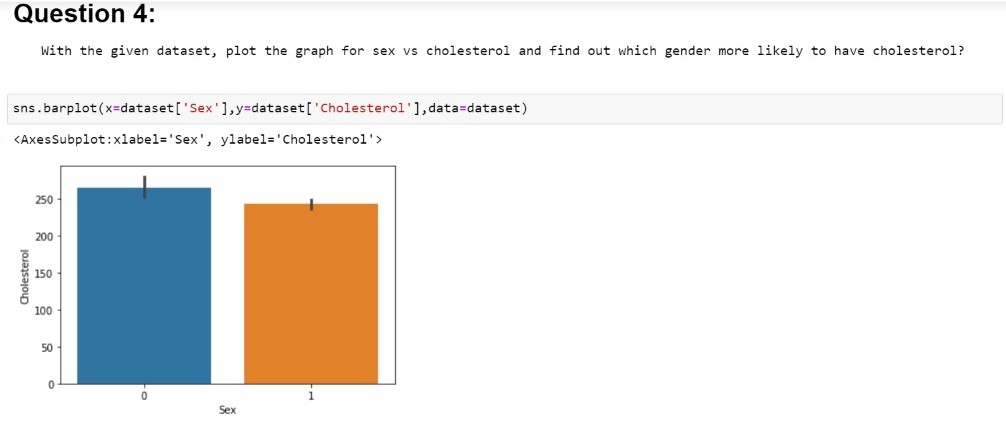


Density Distribution of Thallium

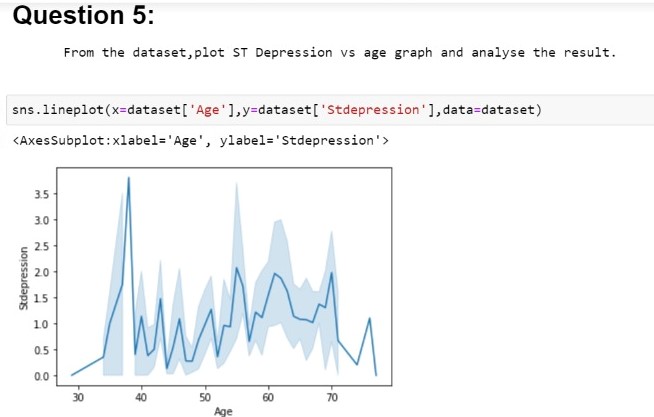




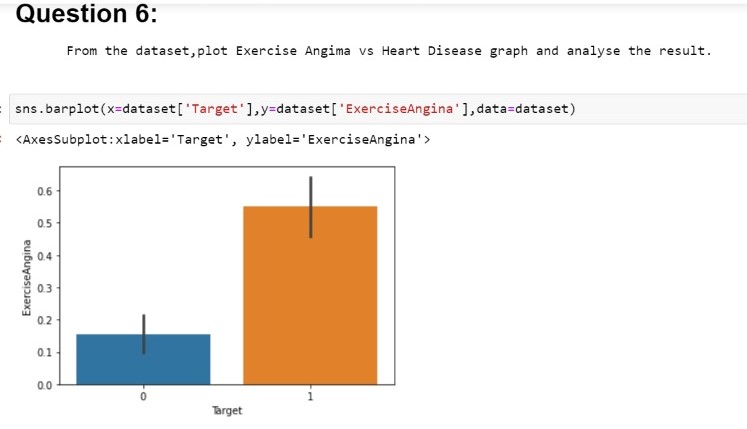
From the above graph we can observe that bp increases at the age of 50-60



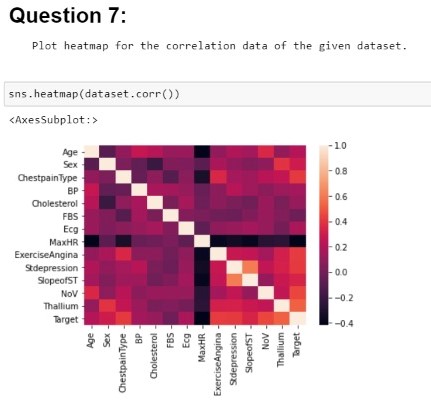
From the above graph we can observe that females have more cholesterol than male population.



Observing the above graph we can see that depression mostly increases bw the age group of 30-40.



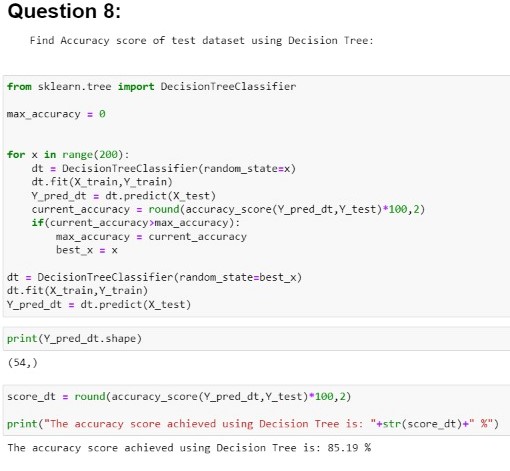
Observing the above graph we can see that person with high exercise angina has more chances of heart disease.



A heatmap contains values representing various shades of the same colour for each value to be plotted. Usually the darker shades of the chart represent higher values than the lighter shade

**ACCURACY CALCULATION:**

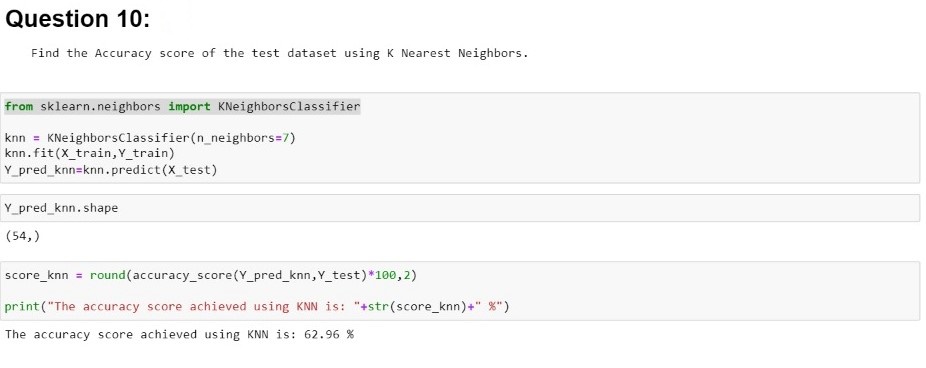
**Decision Tree:**



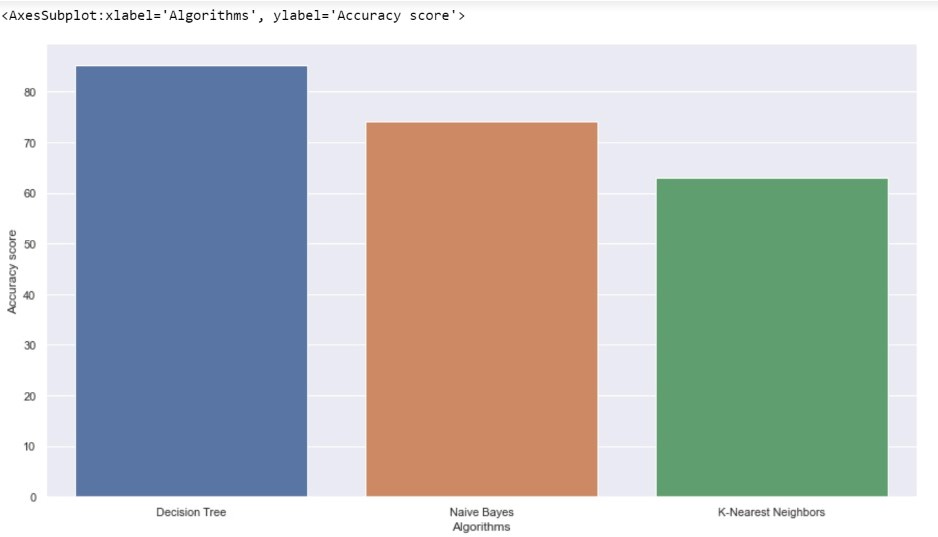
**Naïve Bayes:**



**K Nearest Neighbors:**







We can observe from the graph that Decision Tree gives more Accuracy result compared that of Naive Bayes and KNN Algorithms.

**5. Conclusion & Future Direction:**

**Conclusion:**

# Our Project is all about predicting heart disease using Data Mining Techniques. There are an immense diversity of algorithms and among them, if the individual analysis is to be considered Decision Tree have better performance. When compared the 3 Algorithms i.e.; KNN, Naïve Bayes, Decision Tree The outcome reveals that these data mining techniques can predict heart disease early with an accuracy of 62.96%, 74.07%, 85.19% respectively.

# Future Work

# The similar prediction systems can be built for various other chronic or fatal diseases like Cancer, Diabetes, etc with the help of recent technologies like machine learning, fuzzy logics, image processing and many others. Also, new algorithms can be proposed to achieve more accuracy and reliability.

# The Big Data Technology like Hadoop can be used to store huge chunks of data of all the users worldwide and to manage the data or reports of the user; technologies like Cloud Computing can be made use of

# 6. Reflection Portfolio:

# From this Project we have learnt how to implement Data Pre-processing and various Data Mining techniques, and Models .

# We learnt various classification tasks like KNN algorithm, decision tree Algorithm, and Naïve Bayes Algorithm and its practical application.

# 7.References:

<https://www.kaggle.com/>

<https://www.kaggle.com/rishidamarla/heart-disease-prediction>

<https://www.tutorialspoint.com/data_mining/index.htm>

<https://ieeexplore.ieee.org/abstract/document/9331158?casa_token=K-dfqnAGevcAAAAA:nw-VhQg8qo0wY4kfwo1kcs5fJfpB6lKmGK4O0_1Xv9HIi0DJ9IUfxdIt4NdrmCHBh6HJUaD_OMWd>

<https://cprimestudios.com/blog/how-does-data-mining-help-healthcare#data_mining_in_healthcare>

<https://ieeexplore.ieee.org/document/7154917>