**HEART DISEASE AND STROKE PREVENTION**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

Submitted by

**Group 6**

**Amruthendu M**

**Jagan S**

**Varsha M**

**Nitha S Kumar**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

# 

# List of Figures

* 5.1.1 Bar chart of people have heart disease and how many people suffering risk factors related to heart diseases
* 5.1.2 Pie chart of people who have heart disease and how many people suffer risk factors related to heart diseases.
* 5.1.3. Pie chart of different categories of people taken for survey.
* 5.1.4. Pie chart of different categories of people taken for the survey (Break\_out)
* 5.1.5 Histogram of Topic ( Heart diseases and Risk factors)
* 5.1.6. Bar plot of PriorityArea 1
* 5.1.7. Bar plot of PriorityArea 2
* 5.1.8. Bar plot of PriorityArea 3
* 5.1.9. Bar plot of PriorityArea 4
* 5.1.10. Bar plot of Topic
* 5.1.11. Pie plot of Topic
* 5.1.12. Pie plot of location
* 5.1.13. Count plot of Indicator
* 5.1.14 Box plot
* 5.1.15. kdeplot
* 5.1.16. Frequency graph of dataset
* 5.2.1. Count plot of Breakout Vs Category
* 5.2.2. Count plot of Breakout category Vs Category
* 5.2.3. Scatter plot of Data value and Topic
* 5.2.4. Scatter plot of priority area1 and data values
* 5.2.5. Scatter plot of priority area2 and data values
* 5.2.6. Scatter plot of priority area3 and data values
* 5.2.7. Scatter plot of category and Topic
* 5.2.8 Scatter plot between Data value and Topic
* 5.2.9 Scatter plot between Confidence limit low and Topic\
* 5.2.10 Scatter plot between Confidence limit High and Topic
* 5.2.11. Count plot of Category and Data value type
* 5.2.12. Factor plot of Topic and data value type
* 5.3.1 Heatmap of data
* 5.3.2 Pair plot of data
* 5.3.3 Pair plot of category

# List of Abbreviations

* CVD: Coronary artery disease
* HF: Heart failure
* CVDs: Cardiovascular diseases
* BRFSS : The Behavioral Risk Factor Surveillance System
* MHC : Million Hearts® Collaboration
* AHA :American Heart Association

# Table of Contents

| sl.no | Content | Page number |
| --- | --- | --- |
|  |  |  |
| 1 | LIST OF FIGURES |  |
| 2 | LIST OF TABLES |  |
| 3 | ABSTRACT | 6 |
| 4 | CHAPTER 1: PROBLEM DEFINITION           1.1 Overview           1.2 Problem statement | 8  8 |
| 5 | CHAPTER 2: INTRODUCTION | 9 |
| 6 | CHAPTER 3: LITERATURE SURVEY | 12 |
| 7 | CHAPTER 4: METHODOLOGY            4.1 Block diagram            4.2 Dataset information  4.3 Attribute Information  4.4 Dataset statistics | 19  20  20  27 |
| **6** | CHAPTER 5: EXPLORATORY DATA ANALYSIS  5.1 Univariate analysis  5.2 Bivariate analysis  5.3 Multivariate analysis | 29  46  54 |
| 7 | CHAPTER 6: DATA PREPROCESSING  6.1 Missing value handling  6.2 Outlier detection  6.3 Encoding  6.4 Feature Reduction  6.5 Feature Engineering  6.6 Standardization | 58  61  65  66  66  67 |
| 8 | CHAPTER 7: RESULT AND DISCUSSION  7.1 Modeling  7.2 Fine-tuning  7.3 Web Hosting | 71  90 |
| 9 | CHAPTER 8: CONCLUSION AND FUTURE SCOPE | 94 |
| 10 | REFERENCE |  |

# Abstract

More than 877,500 Americans die of heart disease, stroke, or other cardiovascular diseases every year. Heart disease and stroke are the first and fifth leading causes of death in the United States. A Stroke is a health condition that causes damage by tearing the blood vessels in the brain. It can also occur when there is a halt in the blood flow and other nutrients to the brain. According to the World Health Organization (WHO), stroke is the leading cause of death and disability globally. The data for the studies have been taken from the National Cardiovascular Disease Surveillance System. The data is to be analyzed with different techniques to make it clean for building a machine learning model. This survey has taken various physiological factors and the model would be able to detect the diseases depending on the data. To build the model, different types of machine learning algorithms like Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, etc. can be used. After the prediction, the best models can be chosen for future prediction.

# CHAPTER 1

# PROBLEM DEFINITION

# **1.** Problem Definition

## **1.1 Overview**

Heart disease is one of the fatal problems in the whole world, which cannot be seen with a naked eye and comes instantly when its limitations are reached. Therefore, it needs accurate diagnosis at an accurate time. The healthcare industry produces a huge amount of data every day related to patients and diseases. However, this data is not used efficiently by researchers and practitioners. Today the healthcare industry is rich in data but poor in knowledge. There are various data mining and machine learning techniques and tools available to extract effective knowledge from databases and to use this knowledge for more accurate diagnosis and decision making.

## 1.2 Problem Statement

* Study the heart disease dataset and Predicting and detecting the presence of heart disease and risk factors using machine-learning techniques.

# CHAPTER 2

# INTRODUCTION

# 2. Introduction

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pump blood to the whole body. Any disorder that can lead to disturbing the functionality of the heart is called heart disease. In all body parts the blood, oxygen is circulated by the circulatory system of the body and if the heart does not work properly then the whole human blood system will collapse. So if the heart does not function properly, then it will lead to a serious health condition, it could even lead to death.

Several types of heart disease in the world; Coronary artery disease (CAD), and heart failure (HF) are the most common heart diseases that are present. The main reason behind coronary heart disease (CAD) is blockage or narrowing down of the coronary arteries. Coronary arteries are also responsible for supplying blood to the heart. CAD is the leading cause of death. Over 26 million people are suffering from coronary heart disease (CAD) around the world, and it is increasing 2% annually due to CAD 17.5 million deaths happening globally in 2005.

In the growing world, 2% of the population around the world is suffering from CAD, and 10% of the people are older than 65 years. Approximately 2% of the annual healthcare budget is spent only to treat CAD disease. The US government spent 35 billion dollars on CAD in 2018 . According to recent statistics from the American Heart Association, coronary heart disease accounted for 13% of deaths in the USA in 2018.

      Cardiovascular disease (CVD)  also known as heart disease includes the blood and heart of the human body. Myocardial infarction (as a heart attack) is also a part of CVD. Another type of Heart Disease is called Coronary Heart Disease(CHD), in this type of disease, a substance called plaque develops in the coronary arteries. The development of plaque can block the vessel completely over the course of time.

The symptoms of the Heart Attack:

 1. Chest Pain: The most common sign of a heart attack is chest pain. It mainly happens because of the blockage of the coronary vessel of the body due to the plaque.

 2. Arms pain: The pain normally starts in the chest and moves towards the arm, mainly the left arm.

 3. Low in oxygen: Because of the plaque the level of oxygen drops in the body and causes dizziness and loss of balance.

 4. Tiredness: this causes fatigue, which means simple chores become harder to do.

5. Excessive Sweating: Another common symptom is sweating.

 6. Diabetics: In this, the patients have a heart rate of ~ 100 bpm and also occasionally have a heart rate of 130bpm.

 7. Bradycardia: In this, the patient will have a slower heartbeat of 60 bpm.

8. Cerebrovascular Disease: The patient will have a higher heart rate than normal, usually of 200 bpm and higher than this can cause a Heart attack.

Some other reasons for the occurrence of a heart disease are lifestyle habits like smoking and certain eating habits.

These conventional methods deal with medical reports of the patients. Moreover, these conventional methods are time-consuming, and it may give erroneous results because these conventional methods are performed by humans. Predicting and detection of heart disease has always been a critical and challenging task for healthcare practitioners. Hospitals and other clinics are offering expensive therapies and operations to treat heart diseases. So, predicting heart disease at the early stages will be useful to the people around the world so that they will take necessary actions before getting severe.

# CHAPTER 3

# LITERATURE SURVEY

## 

## 

# 

# 

# 3. Literature Survey

Over the years, a range of works have been done related to the heart disease prediction system using different data mining algorithms by different authors. They tried to attain efficient methods and accuracy in finding out diseases related to heart by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results

  K. Polaraju et al.,proposed Prediction of Heart Disease using the Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease. The work is performed using a training data set consisting of 3000 instances with 13 different attributes.The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of the Regression algorithm is better compared to other algorithms.

Ashwini shetty et al., recommended to develop the prediction system which will diagnose the heart disease from the patient's medical dataset. 13 risk factors of input attributes have been taken into account to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed.

Sairabi H.Mujawar et al., used k-means and naïve bayes to predict heart disease. This paper is to build a system using a historical heart database that gives diagnosis. 13 attributes have been considered for building the system. To extract knowledge from databases, data mining techniques such as clustering and classification methods can be used. 13 attributes with a total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient has heart disease or not based on the values of 13 attributes.

Gavhane, Aditi et al, proposed prediction of heart disease using machine learning. They developed an application which could predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural network had proven to be the most accurate algorithm and hence used in the proposed system.

Shinde, Rucha, et al, proposed an intelligent heart disease prediction system using k-means clustering and Naive Bayes algorithm. It was the combination of both the algorithms. It helped in predicting the heat disease using various attributes and it predicted the output. For grouping of various attributes, it used k-means algorithm and for predicting it used naive bayes algorithm.

Aala et.al., proposes a CVD prediction model using ML algorithms based on the National Health Insurance Service-Health Screening datasets. They extracted 4699 patients aged over 45 as the CVD group, diagnosed according to the international classification of diseases system (I20–I25). In addition, 4699 random subjects without CVD diagnosis were enrolled as a non-CVD group. Both groups were matched by age and gender. The extreme gradient boosting, gradient boosting, and random forest algorithms exhibited the best average prediction accuracy among all algorithms validated in this study. Their results indicate that the proposed health screening dataset-based CVD prediction model using ML algorithms is readily applicable, produces validated results and outperforms the previous CVD prediction models.

Kim et al identified that, Background Identifying people at risk of CVD is a cornerstone of preventative cardiology. Risk prediction models currently recommended by clinical guidelines are typically based on a limited number of predictors with sub-optimal performance across all patient groups. Data-driven techniques based on ML might improve the performance of risk predictions by agnostically discovering novel risk predictors and learning the complex interactions between them. Methods and findings using data on 423,604 participants without CVD at baseline in UK Biobank, and developed a ML-based model for predicting CVD risk based on 473 available variables. The ML-based model was derived using AutoPrognosis, an algorithmic tool that automatically selects and tunes ensembles of ML modeling pipelines. They compared the model with a well-established risk prediction algorithm based on conventional CVD risk factors (Framingham score), a Cox proportional hazards (PH) model based on familiar risk factors, and a Cox PH model based on all of the 473 available variables. Overall, their AutoPrognosis model improves the accuracy of CVD risk prediction in the UK Biobank population. This approach performs well in traditionally poorly served patient subgroups. Additionally, AutoPrognosis uncovered novel predictors for CVD disease that may now be tested in prospective studies. They found that the “information gain” achieved by considering more risk factors in the predictive model was significantly higher than the “modeling gain” achieved by adopting complex predictive models.

Heidenreich et al identified that, To prepare for future cardiovascular care needs, the American Heart Association developed methodology to project future costs of care for hypertension, coronary heart disease, heart failure, stroke, and all other CVD from 2010 to 2030. This methodology avoided double counting of costs for patients with multiple cardiovascular conditions. By 2030, 40.5% of the US population is projected to have some form of CVD. Between 2010 and 2030, real (2008$) total direct medical costs of CVD are projected to triple, from $273 billion to $818 billion. Real indirect costs (due to lost productivity) for all CVD are estimated to increase from $172 billion in 2010 to $276 billion in 2030, an increase of 61%. These findings indicate CVD prevalence and costs are projected to increase substantially. Effective prevention strategies are needed if we are to limit the growing burden of CVD

Van et al .(2020) Concluded that, With increasing age, associations between traditional risk factors (TRFs) and cardiovascular disease (CVD) shift. It is unknown which mid-life risk factors remain relevant predictors for CVD in older people.They included 12 studies, comprising 11 unique cohorts. TRF were evaluated in 2 to 11 cohorts, and retained in 0–70% of the cohorts: age (70%), diabetes (64%), male sex (57%), systolic blood pressure (SBP) (50%), smoking (36%), high-density lipoprotein cholesterol (HDL) (33%), left ventricular hypertrophy (LVH) (33%), total cholesterol (22%), diastolic blood pressure (20%), antihypertensive medication use (AHM) (20%), body mass index (BMI) (0%), hypertension (0%), low-density lipoprotein cholesterol (0%). In studies with low to moderate risk of bias, systolic blood pressure (SBP) (80%), smoking (80%) and HDL cholesterol (60%) were more often retained. Model performance was moderate with C-statistics ranging from 0.61 to 0.77. Compared to middle-aged adults, in people aged 60+ different risk factors predict CVD and current prediction models perform only moderate at best. According to most studies, age, sex and diabetes seem valuable predictors of CVD in old-age. SBP, HDL cholesterol and smoking may also have predictive value. Other blood pressure and cholesterol related variables, BMI, and LVH seem of very limited or no additional value. Without competing risk analysis, predictors are overestimated.

Wang et al ., (2014) detected that, Considering that early detection, diagnosis and screening of hypertension plays a significant role in the prevention and reduction of the onset of cardiovascular diseases as well as the improvement of quality of life, it is of great value to figure out hypertension-related risk factors and further establish a model for the prediction of hypertension with the identified risk factors. Thus, in this paper, they put forward an integrated logistic regression analysis and Artificial Neural Networks (ANNs) model for the selection of risk factors and the prediction of chronic diseases by taking a case study of hypertension. First, a binary logistic regression model was applied on an experimental dataset collected from BRFSS to select factors statistically significant to hypertension in terms of the pre-defined p-value. Then, a Multi-Layer Perception (MLP) neural network model with Back Propagation (BP) algorithm was constructed and trained for the prediction of hypertension with the selected risk factors as inputs to ANNs. Experimental results showed that their proposed approach achieved more than 72% prediction accuracy acceptable in the diagnosis of hypertension and that the Area Under the receiver-operator Curve (AUC) was more than 0.77. The results indicate that integration of logistic regression and artificial neural networks provides us with an effective method in the selection of risk factors and the prediction of hypertension, as well as a general approach for the prediction of other chronic diseases.

Khan, Sundas Naqeeb,et al,(2017) proposed comparative analysis for heart disease prediction. It aimed to draw a comparison among different algorithms used to predict heart disease. The results so obtained would help in developing an understanding of the recent methodologies used for heart disease prediction models. The analysis results of significant data mining techniques that could be used in developing accurate and efficient prediction model which would help doctors in reducing the number of deaths caused by heart disease

# 

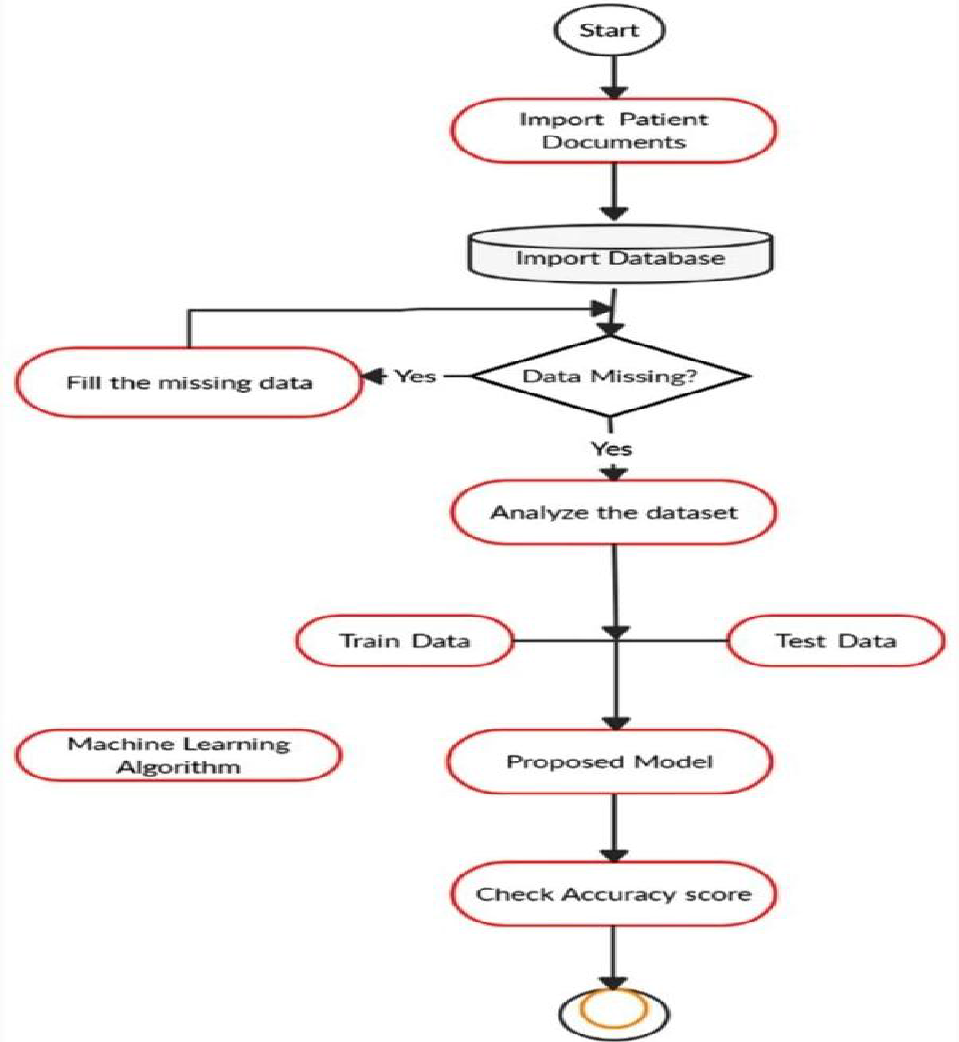
# CHAPTER 4

# METHODOLOGY

# 4. Methodology

Complete machine learning methodology is mentioned below, that is how input dataset is loaded into a program and how it will be processed is mentioned clearly in text format and also in block diagram format.

## 4.1 Block diagram

****

## 4.2 Dataset information

The dataset is publicly available on Kaggle website. This is one of the datasets provided by the National Cardiovascular Disease Surveillance System. The system is designed to integrate multiple indicators from many data sources to provide a comprehensive picture of the public health burden of CVDs (Cardiovascular Diseases) and associated risk factors in the United States. The data are organized by location (national, regional, state, and selected sites) and indicator, and they include CVDs (e.g., heart failure) and risk factors (e.g., hypertension). This dataset gives information related to heart disease. The dataset includes over 85800 records and 29 attributes.

Huge amounts of incomplete and noisy data are present in real-life data. Noise and missing values are common when cleaning collected data. The data must be cleaned of noise and missing values must be filled in order to obtain an accurate and efficient result. Transformation is the process of changing the format of data from one type to another in order to make it more understandable.

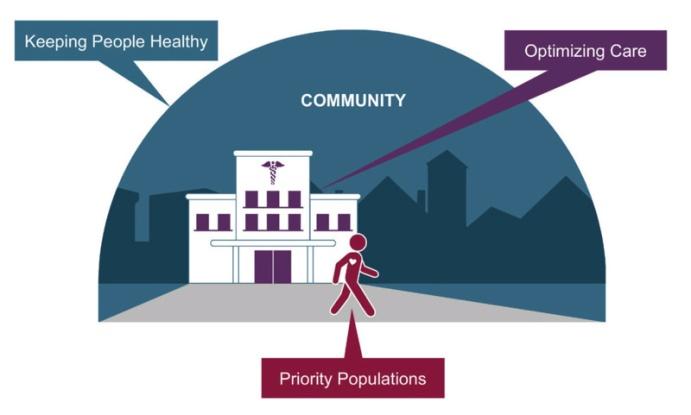
## 4.3 Attribute Information:

1. **Year**(2011-2015)
2. **LocationAbbr** : Location abbreviation
3. **LocationDesc** :Location description
4. **DataSource**: Abbreviation of data source BRFSS

**The Behavioral Risk Factor Surveillance System (BRFSS)** is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

1. **PriorityArea1**-Priority Area (**Million Hearts® or None**)

Million Hearts® is a national initiative to prevent 1 million heart attacks and strokes within 5 years. It focuses on implementing a small set of evidence-based priorities and targets that can improve cardiovascular health for all. The Million Hearts® Collaboration to Prevent Heart Disease and Stroke (MHC) was established in 2015 to bring together national, state and local partners to implement Million Hearts® strategies. MHC organizations work together to disseminate evidence-based cardiovascular disease prevention strategies and resources, promote the use of consistent cardiovascular health messaging, and provide opportunities for sharing best practices and evidence-based approaches by state and local partners. **American Heart Association commitment to Million Hearts®** The American Heart Association (AHA) supports Million Hearts® by integrating strategies in its work with healthcare providers and health systems, leveraging the reach of communication channels including specific meetings, as well as identifying solutions for policy initiatives. AHA, in partnership with the National Forum on Heart Disease and Stroke Prevention guides the efforts of the MHC.



:

1. 20% reduction in sodium intake
2. 20% reduction in tobacco use
3. 20% reduction in physical inactivity
4. 80% performance on the ABCS Clinical Quality Measures
5. 70% participation in cardiac rehab among eligible patients
6. **PriorityArea2**-Priority Area (**ABCS or None**)

Million Hearts® has the potential to help improve the health of millions of Americans by improving the ABCS:

1. Aspirin Therapy When Appropriate
2. Blood Pressure Control
3. Cholesterol Management
4. Smoking Cessation
5. **PriorityArea3**-Priority Area (**Healthy People 2020 or None**)

**Healthy People 2020** are the federal government’s prevention agenda for building a healthier nation. It is a statement of national health objectives designed to identify the most significant preventable threats to health and to establish national goals to reduce these threats. The vision of Healthy People 2020 is to have a society in which all people live long, healthy lives. The overarching goals of Healthy People 2020 are to: attain high-quality, longer lives free of preventable disease, disability, injury, and premature death; achieve health equity, eliminate disparities, and improve the health of all groups; create social and physical environments that promote good health for all; and promote quality of life, healthy development, and healthy behaviors across all life stages.

1. **PriorityArea4**-Priority Area (**AHA 2020 Goals: Cardiovascular Health Metrics or None**)
2. **Category**-Category description(**Cardiovascular diseases and Risk factors**)

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. Strokes can be caused by bleeding from a blood vessel in the brain or from blood clots

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These “intermediate risk factors” can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications. Cessation of tobacco use, reduction of salt in the diet, eating more fruit and vegetables, regular physical activity and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease. Health policies that create conducive environments for making healthy choices affordable and available are essential for motivating people to adopt and sustain healthy behaviors.

In addition, drug treatment of hypertension, diabetes and high blood lipids are necessary to reduce cardiovascular risk and prevent heart attacks and strokes among people with these conditions.

.

1. **Topic**-Topic description
   1. Acute Myocardial Infarction (Heart Attack): A heart attack (medically known as a myocardial infarction) is a deadly medical emergency where your heart muscle begins to die because it isn’t getting enough blood flow. This is usually caused by a blockage in the arteries that supply blood to your heart. If blood flow isn’t restored quickly, a heart attack can cause permanent heart damage and death.
   2. Cholesterol Abnormalities: Lipids, or fats that circulate in the blood, provide energy, produce hormones, and help with many other important functions. However, abnormal levels of certain lipids (either too high or too low) can lead to cardiovascular disease, heart attack, or stroke.
   3. Hypertension: High blood pressure (hypertension) is a common condition in which the long-term force of the blood against your artery walls is high enough that it may eventually cause health problems, such as heart disease.
   4. Obesity: Obesity is a complex disease involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. It's a medical problem that increases the risk of other diseases and health problems, such as heart disease, diabetes, high blood pressure and certain cancers.
   5. Coronary Heart Disease: Coronary heart disease is a type of heart disease that develops when the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. It is the leading cause of death
   6. Diabetes: Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. [Insulin](https://www.niddk.nih.gov/Dictionary/I/insulin), a [hormone](https://www.niddk.nih.gov/Dictionary/H/hormone) made by the [pancreas](https://www.niddk.nih.gov/Dictionary/P/pancreas), helps glucose from food get into your cells to be used for energy. Sometimes your body doesn’t make enough or any insulin or doesn’t use insulin well. Glucose then stays in your blood and doesn’t reach your cells. Over time, having too much glucose in your blood can cause [health problems](https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems).
   7. Major Cardiovascular Disease
   8. Nutrition
   9. Physical Inactivity
   10. Smoking
   11. Stroke: A stroke occurs when a blood vessel in the brain ruptures and bleeds, or when there’s a blockage in the blood supply to the brain. The rupture or blockage prevents blood and oxygen from reaching the brain’s tissues.

1. **Indicator**-Indicator description
   1. Prevalence of acute myocardial infarction (heart attack) among US adults (18+); BRFSS
   2. Prevalence of cholesterol screening in the past 5 years among US adults (20+); BRFSS
   3. Prevalence of consuming fruits and vegetables less than 5 times per day among US adults (18+); BRFSS
   4. Prevalence of coronary heart disease among US adults (18+); BRFSS
   5. Prevalence of current smoking among US adults (18+); BRFSS
   6. Prevalence of diabetes among US adults (18+); BRFSS
   7. Prevalence of healthy weight among US adults (20+); BRFSS
   8. Prevalence of high total cholesterol among US adults (20+); BRFSS
   9. Prevalence of hypertension among US adults (18+); BRFSS
   10. Prevalence of hypertension medication use among US adults (18+) with hypertension; BRFSS
   11. Prevalence of major cardiovascular disease among US adults (18+); BRFSS
   12. Prevalence of obesity among US adults (20+); BRFSS
   13. Prevalence of physical inactivity among US adults (18+); BRFSS
   14. Prevalence of post-hospitalization rehabilitation among heart attack patients, US adults (18+); BRFSS
   15. Prevalence of stroke among US adults (18+); BRFSS
2. **Data\_Value\_Type**: Data Value Type (mean, rate, percentage)
   1. Age standardized
   2. Crude
3. **Data\_Value\_Unit**: Data Value Unit (%, rate per 100,000, etc.)
4. **Data\_Value**: Data value (point estimate)
5. **Data\_Value\_Alt**: Equal to data value, but formatting is numeric
6. **Data\_Value\_Footnote\_Symbol** :Symbol that would be used to flag footnotes
7. **Data\_Value\_Footnote**:Footnote description
8. **Confidence\_Limit\_Low**-95% confidence interval lower bound
9. **Confidence\_Limit\_High**-95% confidence interval upper bound
10. **Break\_Out\_Category**-Break out category description
    1. **Age**
    2. **Race**
    3. **Gender**
    4. **Overall**
11. **Break\_Out**-Break out group description
    1. Non-Hispanic Black
    2. Female
    3. Other
    4. Male
    5. Non-Hispanic Asian
    6. Overall
    7. Non-Hispanic White
    8. Hispanic
    9. 65+
    10. 75+
    11. 25-44
    12. 45-64
    13. 35+
    14. 18-24
    15. 20-24
12. **CategoryId**:Category lookup value
13. **TopicId**:Topic lookup value
14. **IndicatorID**: Indicator lookup value
15. **Data\_Value\_TypeID:**Data value type lookup value
16. **BreakOutCategoryId**:Break out category lookup value
17. **BreakOutId**:Breakout group lookup value
18. **LocationID:**Location lookup value
19. **GeoLocation**
    1. **Dataset statistics**

| Number of variables | 29 |
| --- | --- |
| Number of observations | 85800 |
| Missing cells | 204816 |
| Missing cells (%) | 8.2% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 130.1 MiB |
| Average record size in memory | 1.6 KiB |

Variable types

| Categorical | 24 |
| --- | --- |
| Numeric | 5 |

# 

# CHAPTER 5

# EXPLORATORY DATA ANALYSIS

# 5. Exploratory Data Analysis

By definition, exploratory data analysis is an approach to analyzing data to summarize their main characteristics, often with visual methods. In other words, we perform analysis on data that we collected, to find important metrics/features by using some nice and pretty visualizations.

Exploratory Data Analysis is majorly performed using the following methods:

* **Univariate analysis**: - provides summary statistics for each field in the raw data set (or) summary only on one variable. *Ex*:- CDF,PDF,Box plot, Violin plot.
* **Bivariate analysis**: - is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them. *Ex*:-Box plot, Violin plot.
* **Multivariate analysis:**- is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

## 5.1 Univariate analysis

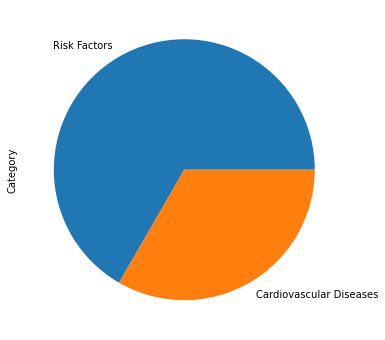
Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

## 5.1.1 Bar chart of people who have heart disease and how many people suffering risk factors related to heart diseases



* It seems that about 28600 people are diseased and 57200 people have the risk factors related to cardiovascular diseases.
* From this plot it is clear that Risk Factors related to heart disease are high in U.S people as compared to people having cardiovascular diseases.

## 5.1.2 Pie chart of people who have heart disease and how many people suffer risk factors related to heart diseases.



* From these graphs it is clear that Risk factors are high as compared to other cardiovascular diseases.

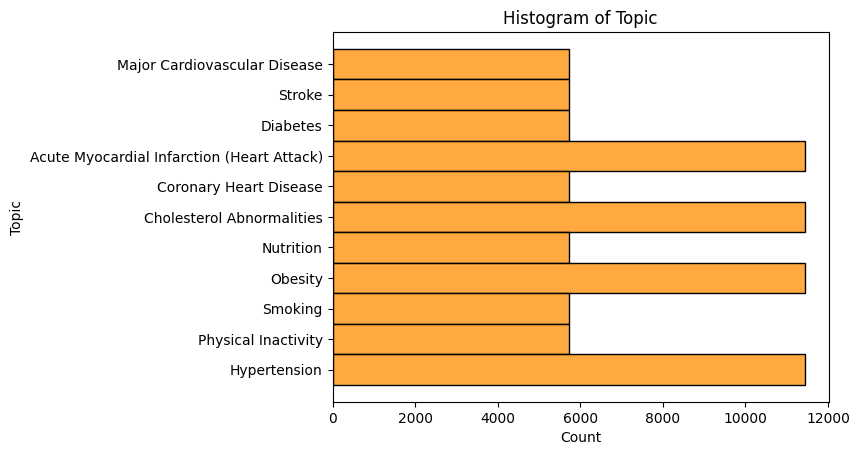
## **5.1.3. Pie chart of different categories of people taken for survey.** **3.png**

* Break out category is divided into Age, Gender, Race and Overall and which is again divided into the different categories.
* From this classification of Race is higher.
* Overall category is less as compared to others

## 5.1.4. Pie chart of different categories of people taken for survey (Break\_out)4.png

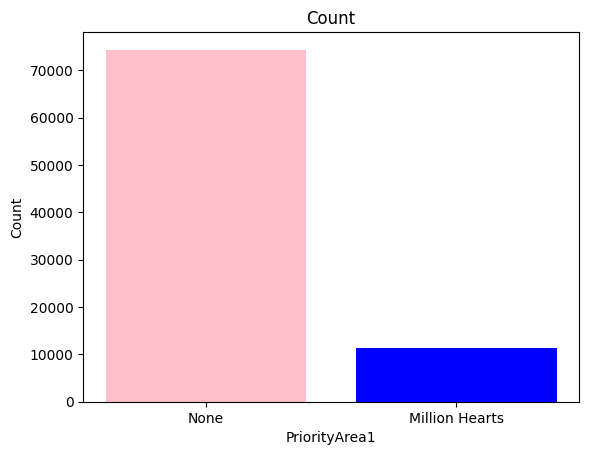
* Break out is divided into Non-Hispanic Black, Female, Other, Male, Non-Hispanic Asian, Overall, Non-Hispanic White ,Hispanic ,65+ ,75+ ,25-44 ,45-64,35,18-24 and 20-24.
* From this Non-Hispanic Black, Female, Other, Male, Non-Hispanic Asian, Overall, Non-Hispanic White, Hispanic having more number of data and all data are equal in numbers.
* 20-24 age group data is very less as compared to others.

## **5.1.5 Histogram of Topic ( Heart diseases and Risk factors***)*



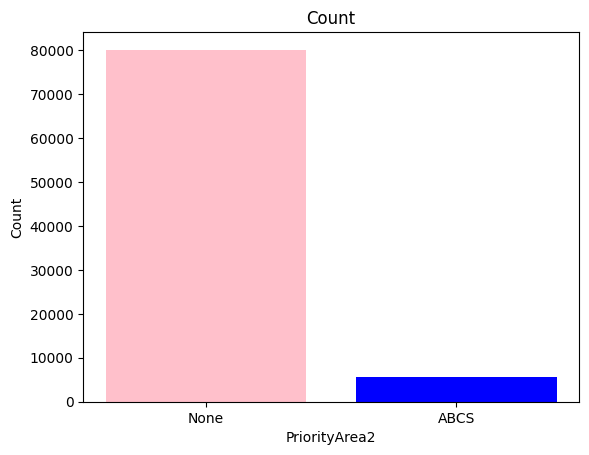
* This graph shows the health conditions of US residents
* From this graph, it is clear that Acute Myocardial Infarction (Heart Attack), Cholesterol Abnormalities, Hypertension, Obesity are very high as compared to other risk factors.

## **5.1.6. Bar plot of PriorityArea 1**

****

* Million hearts and none are two unique values in PriorityArea 1.
* 11440 people included in million hearts for improving the nation's health and preventing stroke.
* Others may be included in other government initiatives or not.

## **5.1.7. Bar plot of Priority Area 2**

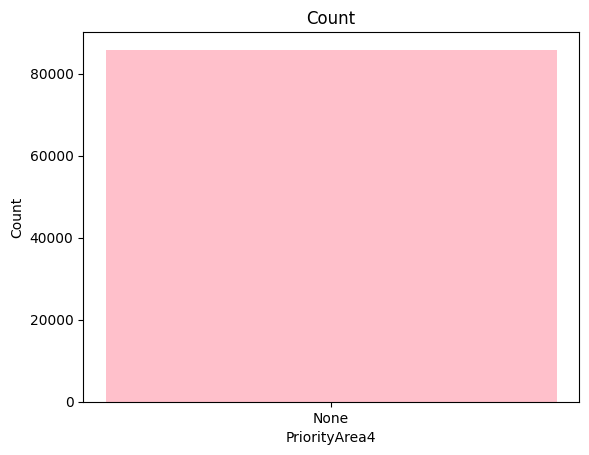
****

* ABCS and None are two unique values in PriorityArea 1.
* Only 5720 people included in ABCS initiative
* Others may be included in other government initiatives or not.

## **5.1.8 Count plot of Priority area 3**https://lh3.googleusercontent.com/42KBQELi6uyY-umHhRM-ezHGsHQ-2_iHruHK-iosJbwIUYSqAuVQQQOMuRK45IZwY19hZUyE9FHSVmoav-5Zx-9lKV0WKp230OWLJo0odp5kPAaiw1fMJ98PA2OT_g

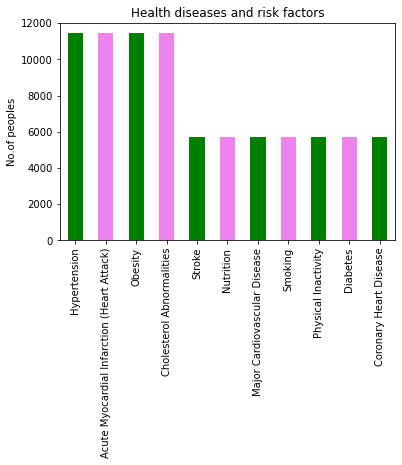
* Healthy People 2020 and None are two unique values in PriorityArea3.
* 45760 people included in Healthy people 2020 having vision "a society in which all people live long ,healthy lives"
* None value is 40040.
* There is a balance between the data and more people included in this initiative as compared to other priority areas.

## 5.1.9   Count plot of Priority area 4

****

* Only one unique value is present in priority area 4 that is none.

## 5.1.10. Bar plot of Topic

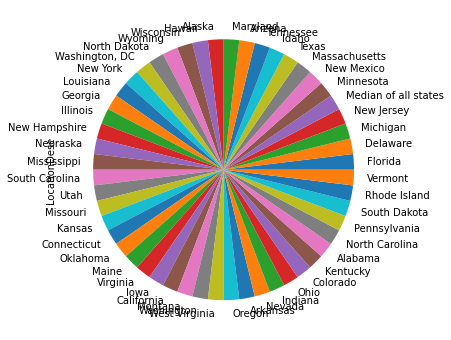


## 5.1.11. Pie plot of Topic

##### **6.png**

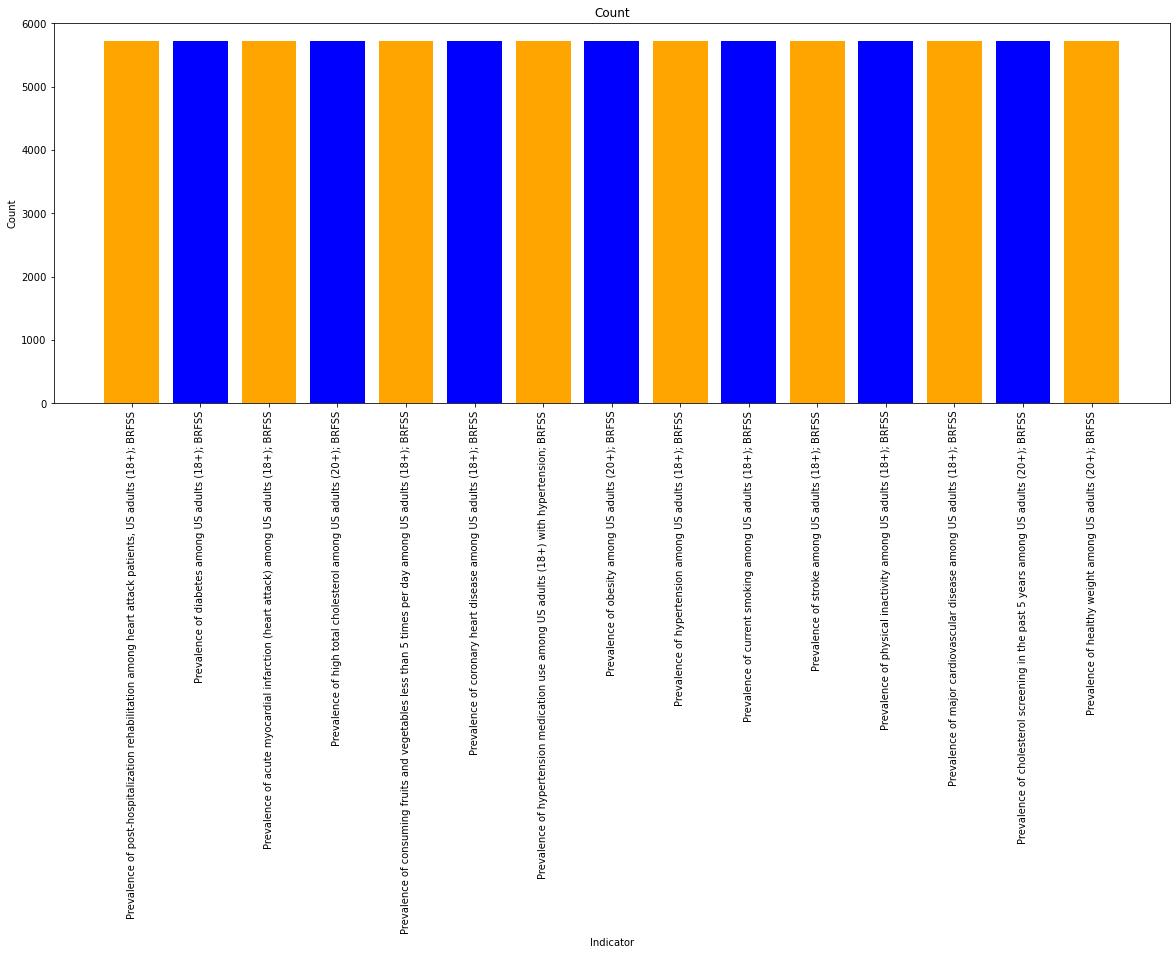
* From this graph, it is clear that Acute Myocardial Infarction (Heart Attack), Cholesterol Abnormalities, Hypertension, Obesity are very high as compared to others.
* Risk Factors related to heart disease are high in U.S people as compared to people having cardiovascular diseases.

## 5.1.12. Pie plot of location

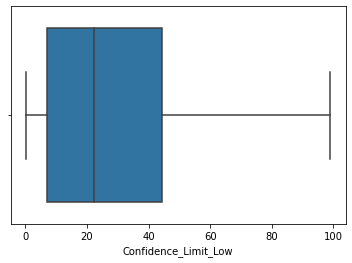
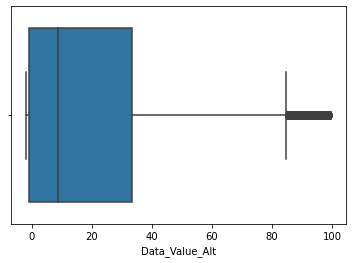


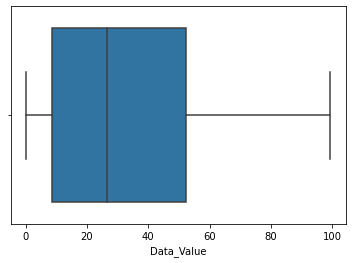
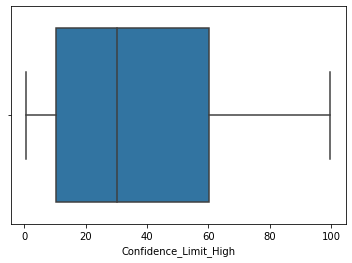
* From this plot it is clear that an equal number of data was collected each location of U.S

## 5.1.13. Count plot of Indicator



## 5.1.14 Box plot





* From this plot it is clear that there are outliers present in the 'Data\_Value\_Alt.

##### **5.1.15. kdeplot77.png**

* Most of the data of 'Data\_Value\_Alt', 'Data\_Value', 'Confidence\_Limit\_Low', 'Confidence\_Limit\_High' in between 0 to 40

## 5.1.16. Frequency graph of the dataset

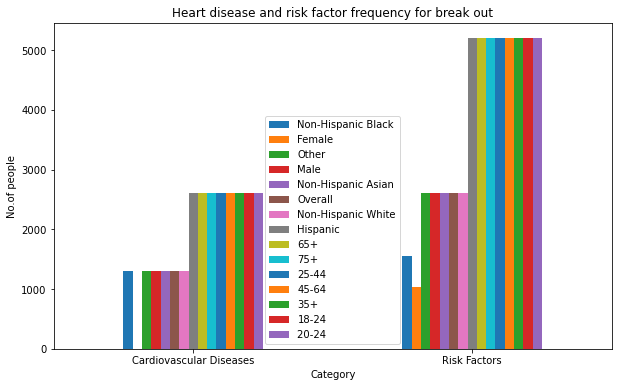


# 5.2 Bivariate analysis

## 5.2.1. Count plot of Breakout Vs CategoryA.png

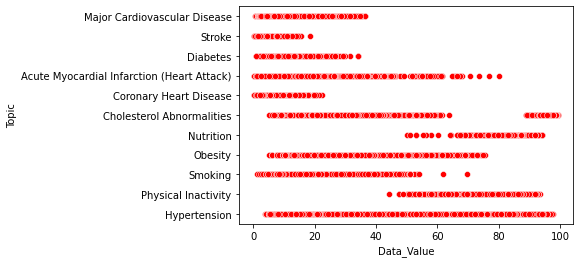
* Based on gender, 5000 people are having cardiovascular disease and 1000 people have Risk factors related to CVDs.
* Based on Age, 7500 people are having cardiovascular disease and 15000 people having Risk factors related to CVDs.
* Based on Race, 13500 people are having cardiovascular disease and 25000 people having Risk factors related to CVDs.
* From this plot it is clear that, On the basis of Break\_Out\_Category Risk factors are high in people.

## 5.2.2. Count plot of Breakout category Vs Category



* High risk factor id for Non Hispanic Black
* Cardiovascular disease is high for females
* Males, Non Hispanic Asian, Non Hispanic white and others having high risk factors.
* "65+","75+","25-44","45-64","35+","18-24" and "20-24'' are at risk factors.

## 5.2.3. Scatter plot of Data value and Topic



* Hypertension, Physical inactivity, Obesity, Cholesterol Abnormalities have high data values.
* This means those are the main conditions present in U.S residents.

## 5.2.4. Scatter plot of priority area1 and data valuesDE.png

# 5.2.5. Scatter plot of priority area2 and data valuesE.png

## 5.2.6. Scatter plot of priority area3 and data values

****

* 2.1.2.5. scatter plot of priority area2 and data values, 2.1.2.4. scatter plot of priority area1 and data value, 2.1.2.6. scatter plot of priority area3 and data values gives an idea about how data values are distributed with priority areas
* In the last plot, there is a balance between Healthy people 2020 and none values.

## 5.2.7. Scatter plot of category and Topic

****

## 

## 5.2.8 Scatter plot between Data value and Topic



* From this Graph it is clear that Hypertension has a higher number of Data Values and less is for stroke.
* Hypertension, Physical inactivity, Obesity, Cholesterol Abnormalities have high data values. This means those are the main conditions which are present in U.S residents.

## 5.2.9 Scatter plot between Confidence limit low and Topic

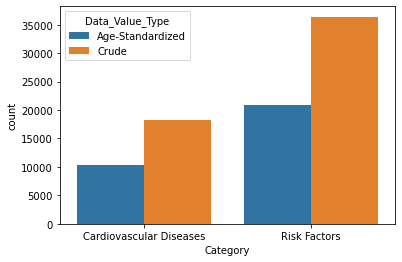


* This graph visualizes the relationship between confidence limit low and Topic.

## 5.2.10 Scatter plot between Confidence limit High and Topic



## 5.2.11. Count plot of Category and Data value type



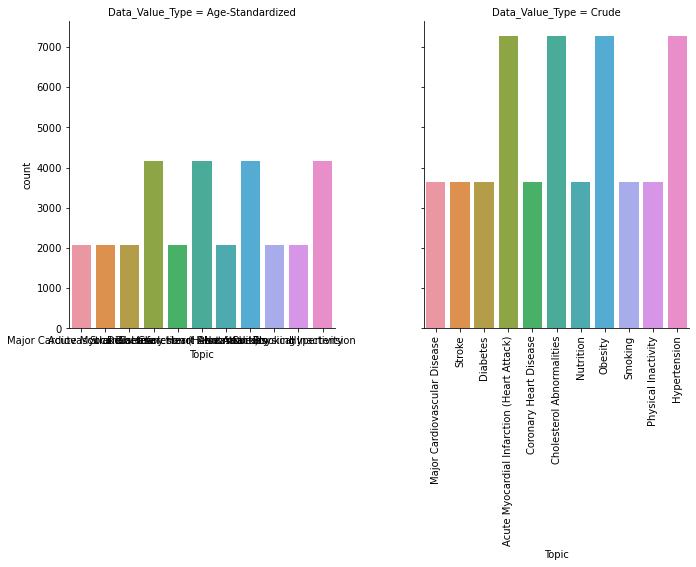
- In both the cases Crude data is very high as compared to Age standardized data value type.

- In the cardiovascular diseases category, 10040 people belong to Age standardized and 20800 for crude.

- In Risk factors category, 36400 people belong to Age standardized and 18200 for crude.

- In Both the categories, more people belong to the crude data type.

## **5.2.12. Factor plot of Topic and data value type**



## **5.3 Multivariate Data Analysis**

##### **5.3.1 Heatmap of data4444.png**

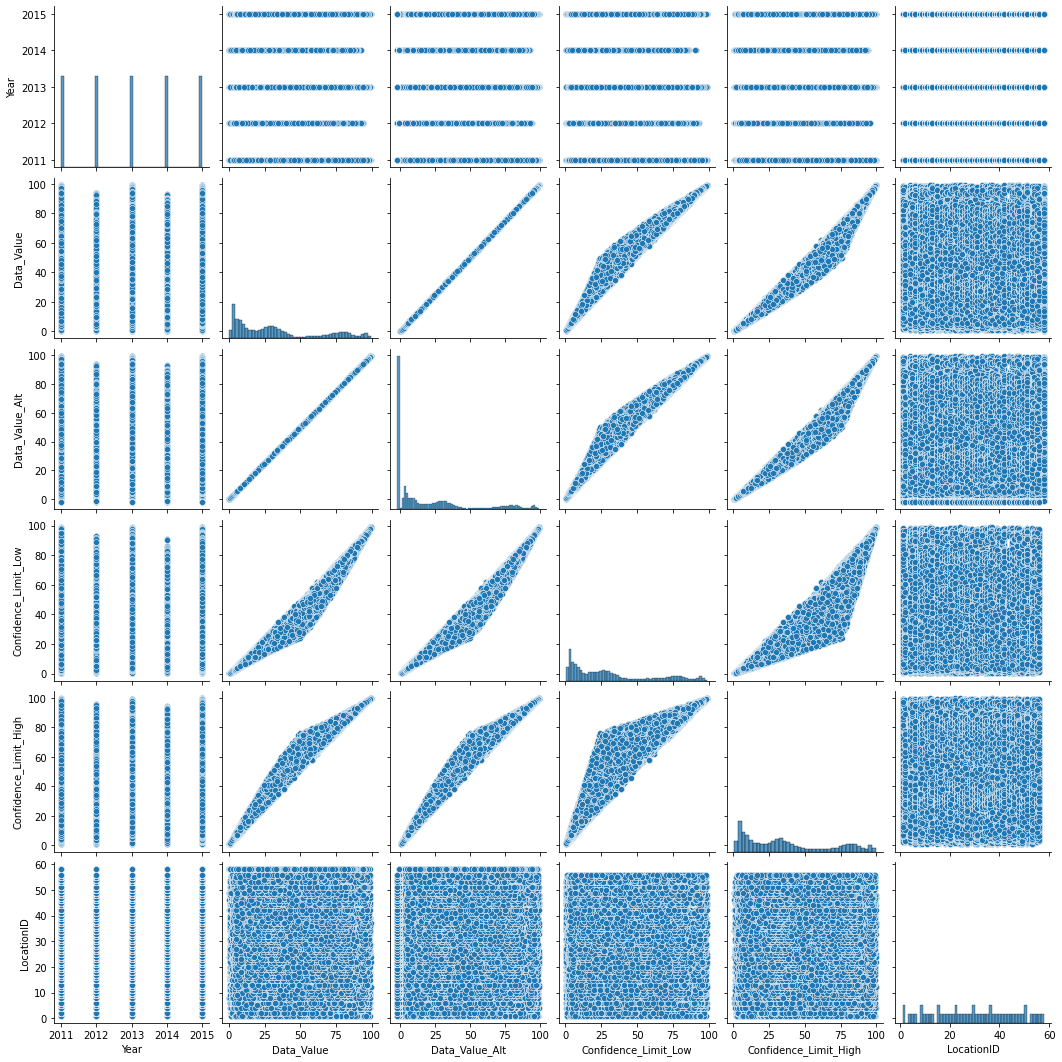
### The correlation between Data\_Value and Data\_Value\_Alt are 1.

### Confidence\_Limit\_Low and Data\_Value are highly correlated.

### Confidence\_Limit\_Low and Confidence\_Limit\_High are highly correlated.

### Correlations between Year and other features are very low.

## 5.3.2 Pair plot of data



## 5.3.3 **Pair plot of category***I.png*

# 

# CHAPTER 6

# DATA PREPROCESSING

## 6.1 Missing value Handling

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. In the dataset, blank shows the missing values. In Pandas, usually, missing values are represented by NaN. It stands for Not a Number. There can be multiple reasons why certain values are missing from the data. Reasons for the missing data from the dataset affect the approach of handling missing data. So it’s necessary to understand why the data could be missing.

Some of the reasons are listed below:

* Past data might get corrupted due to improper maintenance.
* Observations are not recorded for certain fields due to some reasons. There might be a failure in recording the values due to human error.
* The user has not provided the values intentionally.

It is important to handle the missing values appropriately.

* Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.
* You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly.
* Missing data can lead to a lack of precision in the statistical analysis.

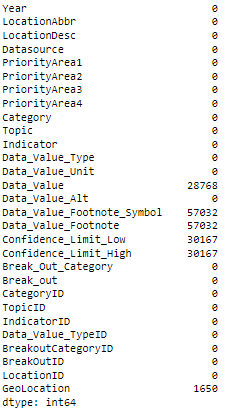
## 

## 

## 

## Checking for missing values

The first step in handling missing values is to look at the data carefully and find out all the missing values. The following code shows the total number of missing values in each column. It also shows the total number of missing values in the entire data set



From the above output, we can see that there are 6 columns – Data\_Value , Data\_Value\_Footnote\_Symbol, Data\_Value\_Footnote, Confidence\_Limit\_Low, Confidence\_Limit\_High, and Geolocation having missing values.



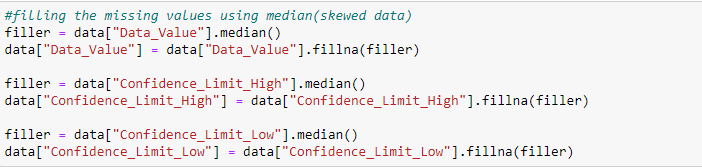


There are 204816 missing values in total. From this, 28768 Missing values in Data\_value. 57032 missing values for the Data value footnote symbol and Data footnote. 30167 missing values for confidence limit low and confidence limit high respectively. 1650 missing values for Geolocation.

#### Imputing the Missing Values:

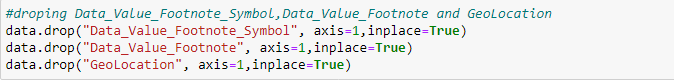
One of the techniques is mean imputation in which the missing values are replaced with the mean value of the entire feature column. In the case of fields like Data\_Value,Confidence\_Limit\_High and Confidence\_Limit\_Low the data is skewed. In such cases, it may not be a good idea to use mean imputation for replacing the missing values. imputing missing data with mean values can only be done with numerical data.

Another technique is median imputation in which the missing values are replaced with the median value of the entire feature column. When the data is skewed, it is good to consider using the median value for replacing the missing values. imputing missing data with median value can only be done with numerical data. Here the data is right skewed. So the median is used for filling missing values.



#### Dropping the missing values

See that this contains columns like Data\_Value\_Footnote\_Symbol,Data\_Value\_Footnote and GeoLocation We won’t be working with all the columns in the dataset, so deleting these columns.



Rechecking for missing values after filling and dropping null values



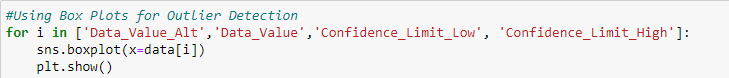
## 6.2 Outlier detection

An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining.

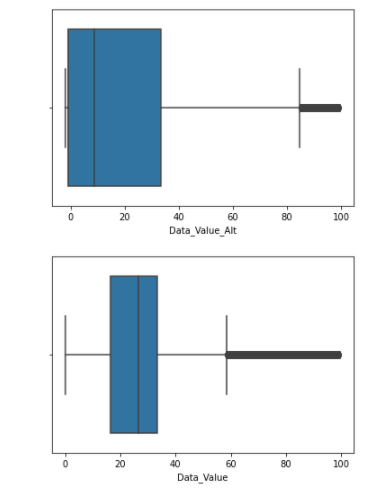
## Using Box Plots for Outlier Detection:

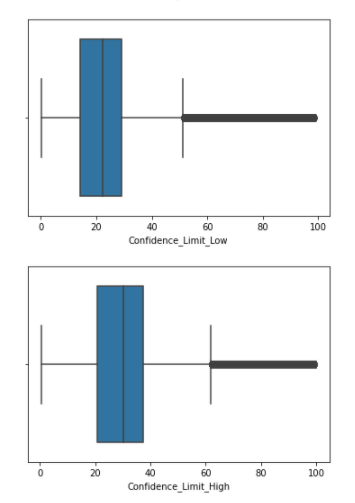
## Outliers can be detected using visualization, implementing mathematical formulas on the dataset, or using the statistical approach.

## input:



output:



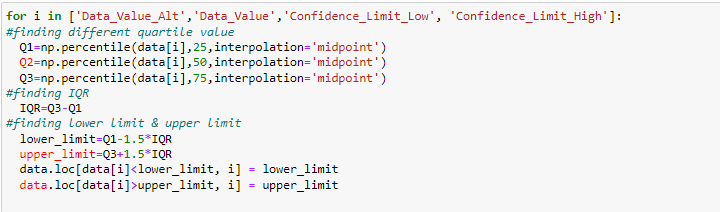


The dots in the box plots correspond to extreme outlier values. Data\_Value\_Alt', 'Data\_Value', 'Confidence\_Limit\_Low', and 'Confidence\_Limit\_High having outliers. Here we using IQR method for removing outliers

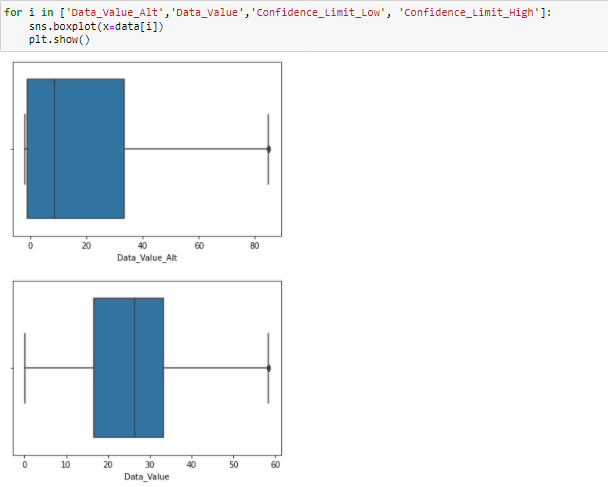
***Inter quartile range (IQR) method:*** Each dataset can be divided into quartiles. The first quartile point indicates that 25% of the data points are below that value whereas the second quartile is considered as the median point of the dataset. The inter quartile method finds the outliers on numerical datasets by following the procedure below

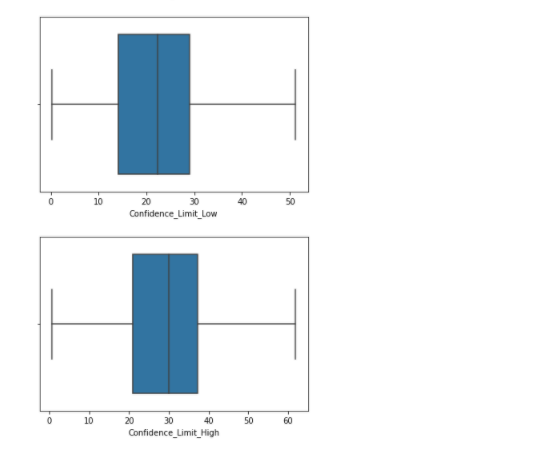
* Find the first quartile, Q1.
* Find the third quartile, Q3.
* Calculate the IQR. IQR= Q3-Q1.
* Define the normal data range with a lower limit as Q1–1.5\*IQR and upper limit as Q3+1.5\*IQR.
* Any data point outside this range is considered an outlier and should be removed for further analysis.

It has the minimum and maximum point defined as Q1–1.5\*IQR and Q3+1.5\*IQR respectively. Any point outside this range is outlier.

****

Similar boxplots are generated after the outliers are removed.



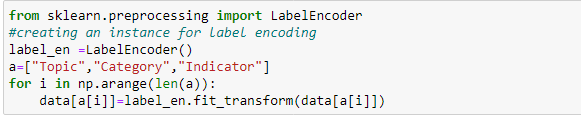


## **6.3 Encoding**

## **6.3.1 Label encoding**

In label encoding in Python, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4).

The sklearn library of Python offers users pre-defined functions in order to work with Label Encoding on the dataset. We have used the fit\_transform() method to apply the functionality of the label encoder pointed by the object to the data variable.

****

## **6.3.2 one hot encoding**

One-hot encoding is used to convert categorical variables into a format that can be readily used by machine learning algorithms.

The basic idea of one-hot encoding is to create new variables that take on values 0 and 1 to represent the original categorical values. import the OneHotEncoder() function from the sklearn library and use it to perform one-hot encoding.



Here we encoded 'Data\_Value\_Type' and 'Break\_Out\_Category'.

## **6.4 Feature Reduction**

Reducing the number of input variables in training data. It is desirable to have simple models that generalize well, and in turn, input data with few input variables. In many instances, some columns are not relevant to your analysis. We can use Pandas drop() function to drop multiple columns from a dataframe. Pandas drop() is versatile and it can be used to [drop rows of a dataframe](https://cmdlinetips.com/2018/04/how-to-drop-one-or-more-columns-in-pandas-dataframe/) as well.

To use Pandas drop() function to drop columns, we provide the multiple columns that need to be dropped as a list. In addition, we also need to specify the axis=1 argument to tell the drop() function that we are dropping columns. With axis=0 drop() function drops rows of a dataframe.

To drop columns without creating a new data frame we specify “inplace=True”.here we drop "LocationAbbr",'CategoryID', 'TopicID','IndicatorID', 'Data\_Value\_TypeID', 'BreakoutCategoryID', 'BreakOutID','LocationID','Data\_Value\_Unit' and 'Datasource'. Now we have 20 columns and 85800 rows.

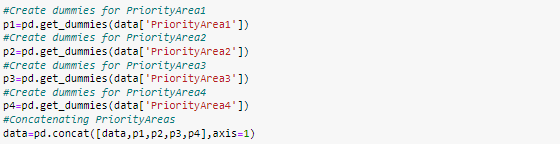
## 

## 6. 5. Feature Engineering

Feature engineering is about creating new input features from your existing ones. It is the most important art in machine learning which creates the huge difference between a good model and a bad model.

Here, we created a new column named “PriorityArea” by combining all other four priority areas

First we created dummies for each priority area and concatenated all priority areas using the code:

****

then created individual columns for each priory area such as “Million Hearts” , “ABCS” and “Healthy People 2020”.

****

created a new column named “PriorityArea” by combining all other individual columns.

****

After this, we label encoded to the newly created columns and drop other unnecessary columns.

****

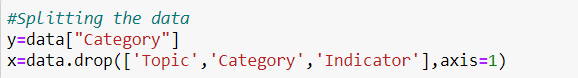
## 6.6 Standardization

Standardization is used on the data values that are normally distributed. Further, by applying standardization, we tend to make the mean of the dataset as 0 and the standard deviation equivalent to 1.

That is, by standardizing the values, we get the following statistics of the data distribution

* mean = 0
* standard deviation = 1

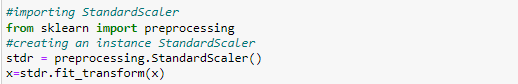
Step 1: split the dataset into dependent and independent variables.



step 2: imported the necessary library required. We import sklearn library to use the StandardScaler function

step 3: Create an instance StandardScaler().

step 4: Apply the function into the dataset using the fit\_transform( ) function

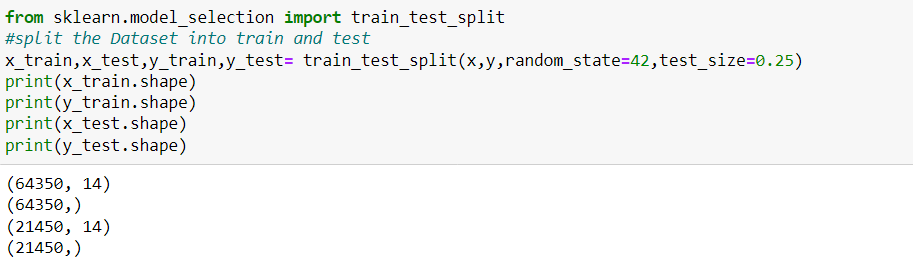


output :

## 

## Splitting our Dataset

Here we used the ‘train\_test\_split’ to split the data in 75:25 ratio i.e. 75% of the data will be used for training the model while 25% will be used for testing the model that is built out of it.



# CHAPTER 7

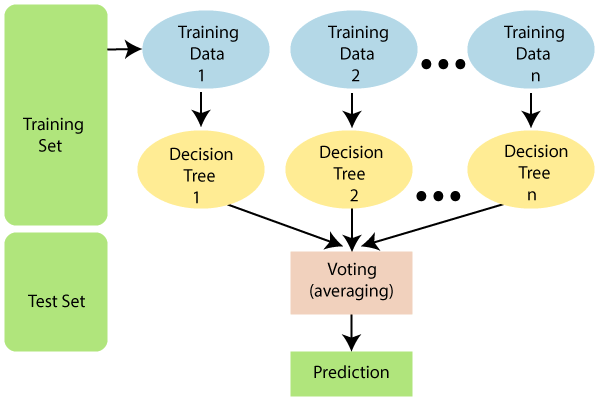
# RESULT AND DISCUSSION

# 

# 7. 1. MODELING

## 1. Random forest classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.



As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest works in two-phase first is to create the random forest by combining N decision trees, and the second is to make predictions for each tree created in the first phase.

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets)

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5**:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

### Python Implementation of Random Forest Algorithm:

### Fitting the Random Forest algorithm to the training set:

To fit it, we will import the RandomForestClassifier class from the sklearn.ensemble library and create a rf object, which will help to fit the model to random forest4.

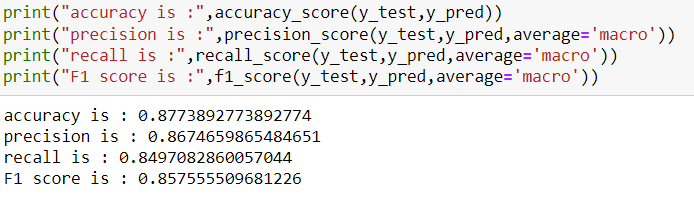
The code is given below:



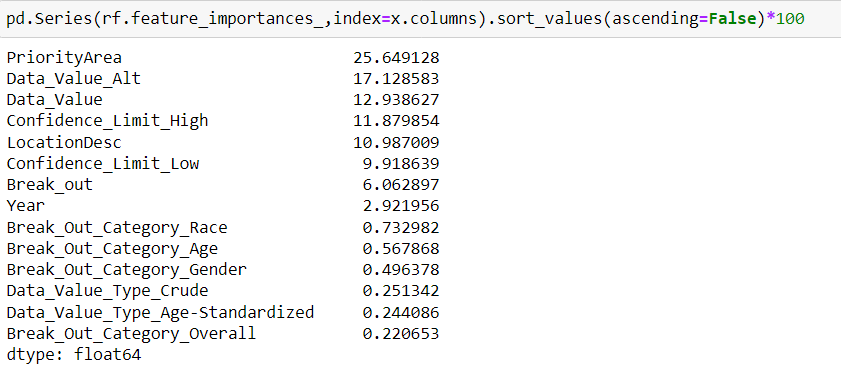
Since our model is fitted to the training set, we can now predict the test result. For prediction, we will create a new prediction vector y\_pred.



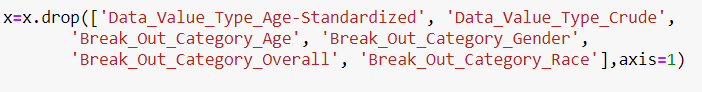
After training, check the accuracy using actual and predicted values.



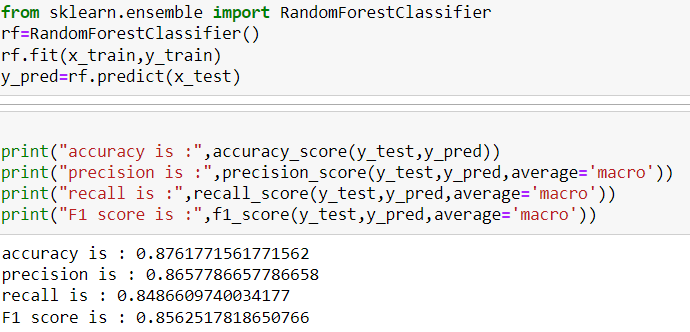
Finding important features or selecting features in the heart dataset.



Here, you can remove the ‘Data\_Value\_Type\_Age-Standardized', 'Data\_Value\_Type\_Crude','Break\_Out\_Category\_Age','Break\_Out\_Category\_Gender','Break\_Out\_Category\_Overall' and 'Break\_Out\_Category\_Race' . because it has very low importance as compared to others.



After this, generate a model on the selected training set features, perform predictions on the selected test set features, and compare actual and predicted values.



After removing the least important features, there is no change in inaccuracy.

## 2. Logistic regression classifier

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. Its basic fundamental concepts are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.

The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:



Where, y is a dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:



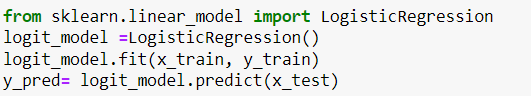
Apply Sigmoid function on linear regression:



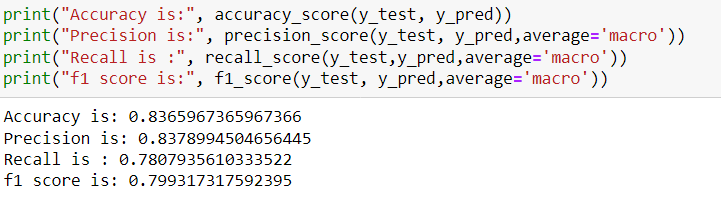
### Python Implementation of Logistic regression classifier:

First, import the Logistic Regression module and create a Logistic Regression classifier object using the LogisticRegression() function.

Then, fit your model on the train set using fit() and perform prediction on the test set.

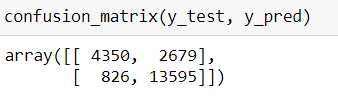


After training, check the accuracy using actual and predicted values.



### Model Evaluation using Confusion Matrix:

A confusion matrix is a table that is used to evaluate the performance of a classification model. The fundamental of a confusion matrix is the number of correct and incorrect predictions summed up class-wise.



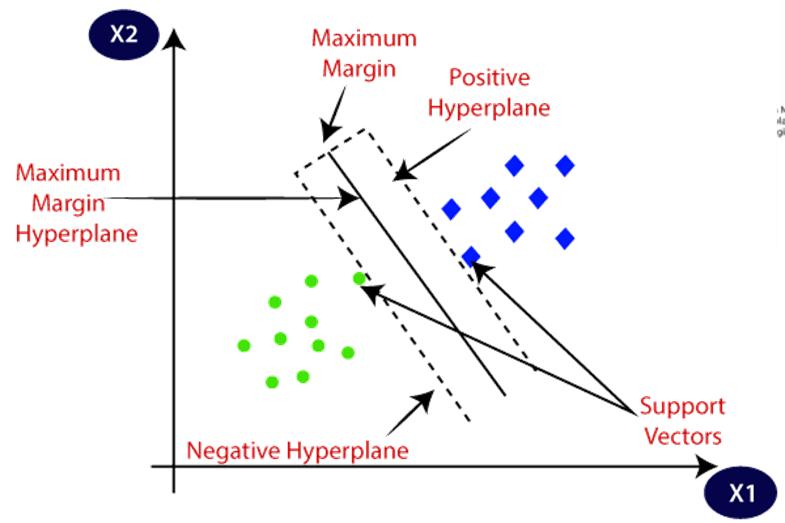
Here, you can see the confusion matrix in the form of the array object. The dimension of this matrix is 2\*2 because this model is binary classification. You have two classes 0 and 1. Diagonal values represent accurate predictions, while non-diagonal elements are inaccurate predictions. In the output, 4350 and 13595 are actual predictions, and 826 and 2679 are incorrect predictions.

## 3. SVM (Support vector machine)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as a Support Vector Machine.



Here we have used three types of SVM.

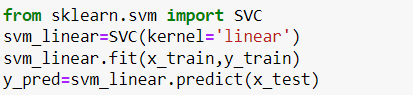
## 3.1 Linear SVM

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and the classifier is used as Linear SVM classifier.

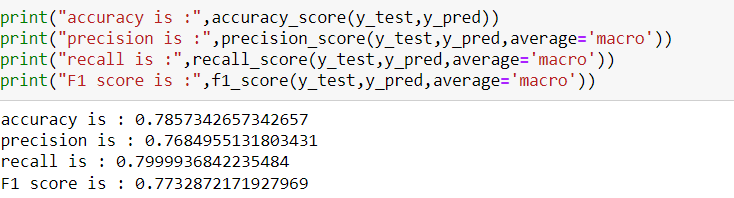
### Python Implementation of Linear SVM:

The training set will be fitted to the SVM classifier. To create the SVM classifier, import the SVCclass from the Sklearn.svm library.

In the code, kernel = 'linear', as here creating SVM for linearly separable data. And then fitted the classifier to the training dataset (x\_train, y\_train).



After training, we checked the accuracy using actual and predicted values.



## 3.2 Polynomial SVM

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

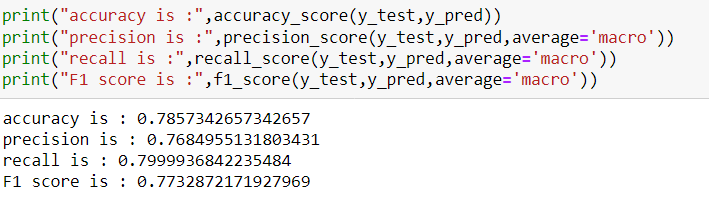
### Python Implementation of Polynomial SVM:

The training set will be fitted to the SVM classifier. To create the SVM classifier, import the SVCclass from the Sklearn.svm library.

In the code, kernel = 'poly', as here creating SVM for linearly separable data. And then fitted the classifier to the training dataset (x\_train, y\_train).

## 

After training, checked the accuracy using actual and predicted values.



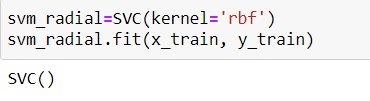
## 3.3 Radial SVM

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

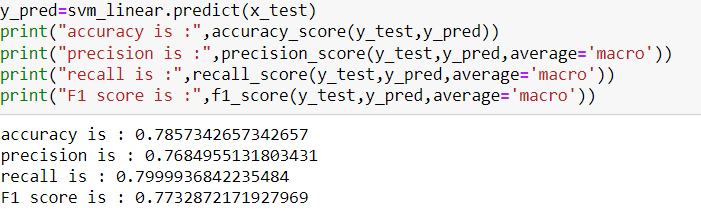
### **Python Implementation of** **Radial SVM:**

The training set will be fitted to the SVM classifier. To create the SVM classifier, import the SVCclass from the Sklearn.svm library.

In the code, kernel = 'rbf', as here creating SVM for linearly separable data. And then fitted the classifier to the training dataset (x\_train, y\_train).



After training, checked the accuracy using actual and predicted values.



## 

## 4. Naïve Bayes Classifier Algorithm

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and is used for solving classification problems.
* It is mainly used in text classification that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
* Some popular examples of the Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

## 

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

* **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
* **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

## Bayes' Theorem:

* Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
* The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

## Types of Naïve Bayes Model:

There are three types of Naive Bayes Model, which are given below:

* **Gaussian**: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
* **Multinomial**: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.  
  The classifier uses the frequency of words for the predictors.
* **Bernoulli**: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

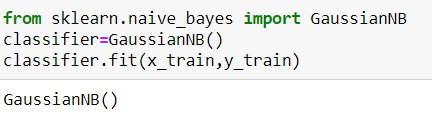
Here we have used two of the models such as Gaussian and Bernoulli

## 4.1 Gaussian Naive Bayes

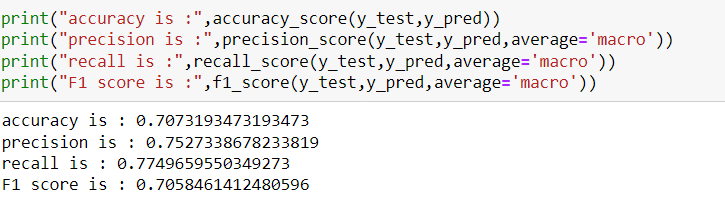
### Python Implementation of Gaussian Naive Bayes:

First, import the Gaussian NB module and create a classifier object using the GaussianNB() function. Then, fit your model on the train set using fit() and perform prediction on the test set.

# 



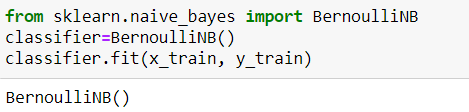
After training, check the accuracy using actual and predicted values.



## 4**.2 Bernoulli Naive Bayes**

### Python Implementation of Bernoulli Naive Bayes:

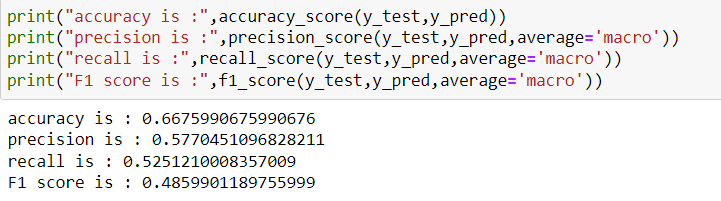
First, import the Bernoulli NB module and create a classifier object using the BernoulliNB() function. Then, fit your model on the train set using fit() and perform prediction on the test set.



Since our model is fitted to the training set, we can now predict the test result. For prediction, we will create a new prediction vector y\_pred

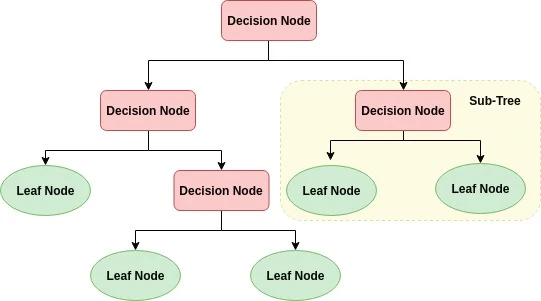


After training, check the accuracy using actual and predicted values.



## 5**. Decision Tree Classifier**

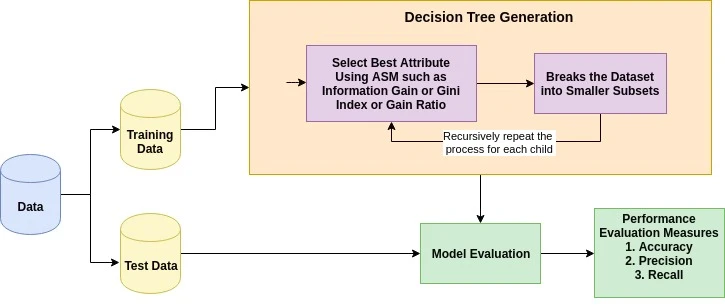
A decision tree is a flowchart-like tree structure where an internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in a recursive manner called recursive partitioning. This flowchart-like structure helps you in decision-making. It's visualized like a flowchart diagram which easily mimics human-level thinking. That is why decision trees are easy to understand and interpret.



Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures(ASM) to split the records.
2. Make that attribute a decision node and break the dataset into smaller subsets.
3. Start tree building by repeating this process recursively for each child until one of the conditions will match:
   * All the tuples belong to the same attribute value.
   * There are no more remaining attributes.
   * There are no more instances.

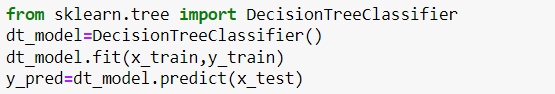


### Python Implementation of Random Forest Algorithm:

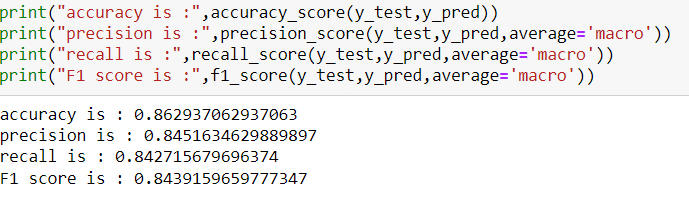
### Fitting the Decision tree algorithm to the training set:

To fit it, we will import the Decision Tree Classifier class from the sklearn.ensemble library and create a dt\_model object , which will help to fit the decision tree model.

The code is given below:

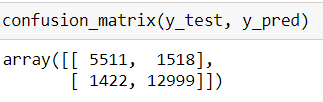


After training, check the accuracy using actual and predicted values.



### Model Evaluation using Confusion Matrix:

A confusion matrix is a table that is used to evaluate the performance of a classification model. The fundamental of a confusion matrix is the number of correct and incorrect predictions summed up class-wise.

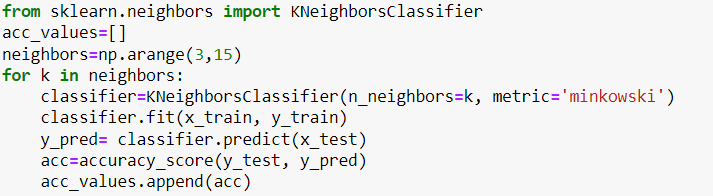


Here, you can see the confusion matrix in the form of the array object. The dimension of this matrix is 2\*2 because this model is binary classification. You have two classes 0 and 1. Diagonal values represent accurate predictions, while non-diagonal elements are inaccurate predictions. In the output, 5511 and 12999 are actual predictions, and 1422 and 1518 are incorrect predictions.

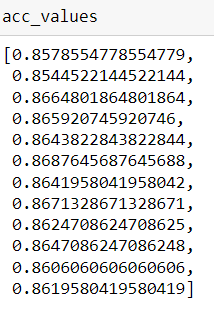
## 6. KNN

KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure is determined from the dataset. This will be very helpful in practice where most of the real-world datasets do not follow mathematical theoretical assumptions. A lazy algorithm means it does not need any training data points for model generation. All training data is used in the testing phase. This makes training faster and the testing phase slower and costlier.

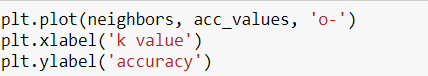
First, import the KNeighborsClassifier module and create a KNN classifier object by passing the argument number of neighbors in the KNeighborsClassifier() function. Then, fit our model on the train set using fit() and perform prediction on the test set using y\_pred

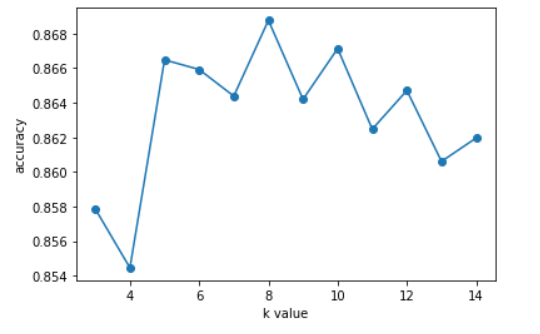


One of the challenges in a k-NN algorithm is finding the best 'k' i.e. the number of neighbors to be used in the majority vote while deciding the class. Generally, it is advisable to test the accuracy of your model for different values of k and then select the best one from them.

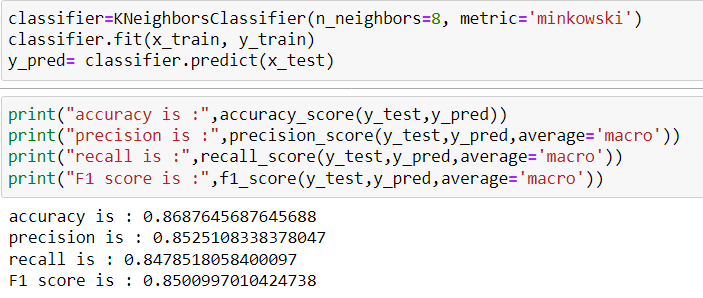


plotting a graph between k value and accuracy





by looking at the above graph, it looks like when k value =8, the model performs the best. So, selecting k=8 and re-running the training once again.



It looks like the model was able to classify 86.87% of the testing data correctly.

## 

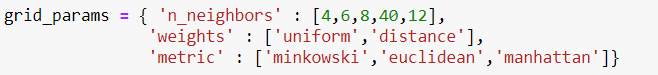
## 7.2. Fine-tuning

We will use the Exhaustive Grid Search technique for hyperparameter optimization. An exhaustive grid search is a good way to determine the best hyperparameter values to use, but it can quickly become time-consuming with every additional parameter value and cross-validation that you add.

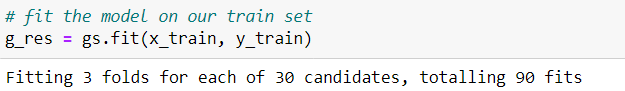


We will use three hyperparameters- n-neighbors, weights, and metric.

1. n\_neighbors: Decide the best k based on the values we have computed earlier.
2. weights: Check whether adding weights to the data points is beneficial to the model or not. 'uniform' assigns no weight, while 'distance' weighs points by the inverse of their distances meaning nearer points will have more weight than the farther points.
3. metric: The distance metric to be used will calculate the similarity.





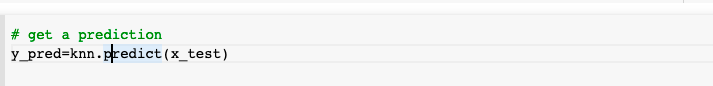


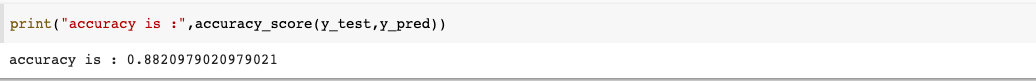
Once the model is fit, we can find the optimal parameter of K and the best score obtained through GridSearchCV and we got the best score is 0.8688578088578088.



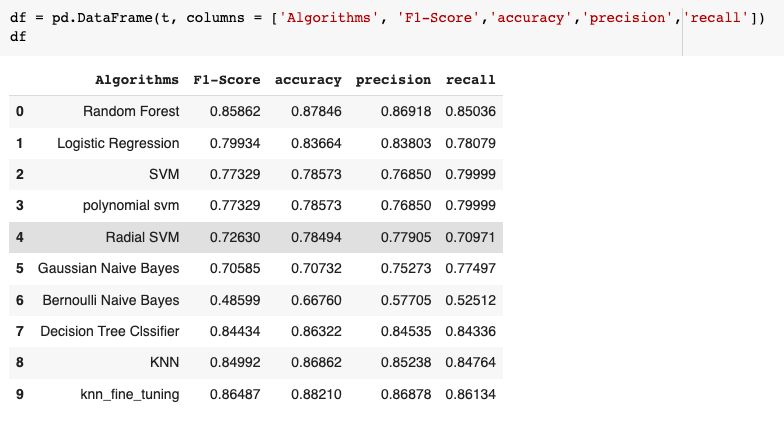
Since now we have the best hyperparameter of 'metric': 'manhattan', 'n\_neighbors': 12, 'weights': 'uniform', this can be used to fit a KNN model and check the prediction values.



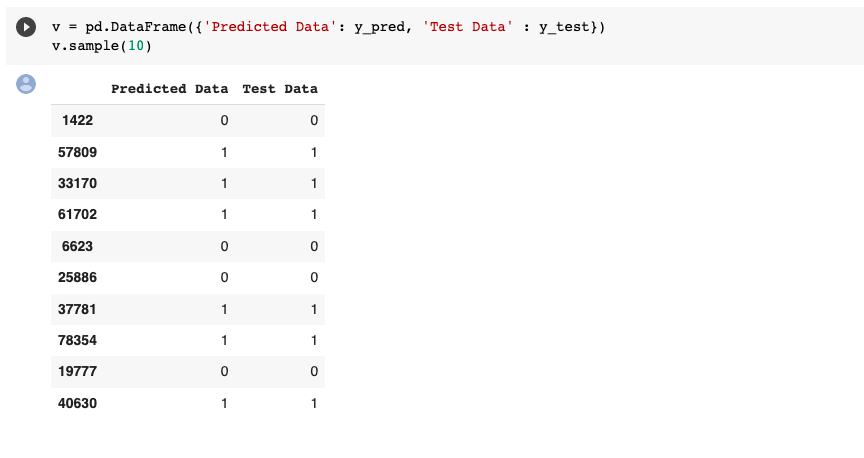


By altering the parameter values as a part of the fine-tuning method, the model’s accuracy is improved. The previous accuracy was 86% and now it is raised to 88%.

The given table is the comparison of different algorithms and their respective F1-score, accuracy, precision, and recall.



The given table is the test data vs predicted data.



After passing through different models, the accuracy of KNN is high as compared to others. Therefore, KNN is chosen as the best model for predicting the heart disease dataset.

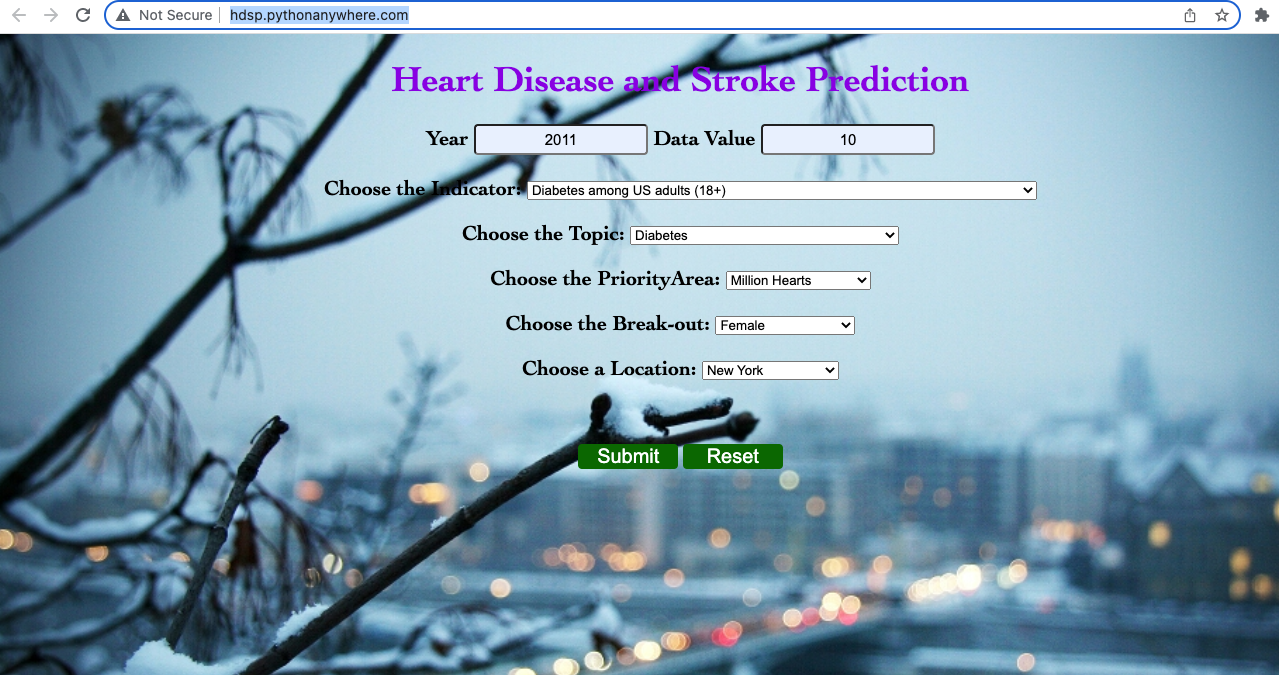
## 

## 7.3 Web Deployment

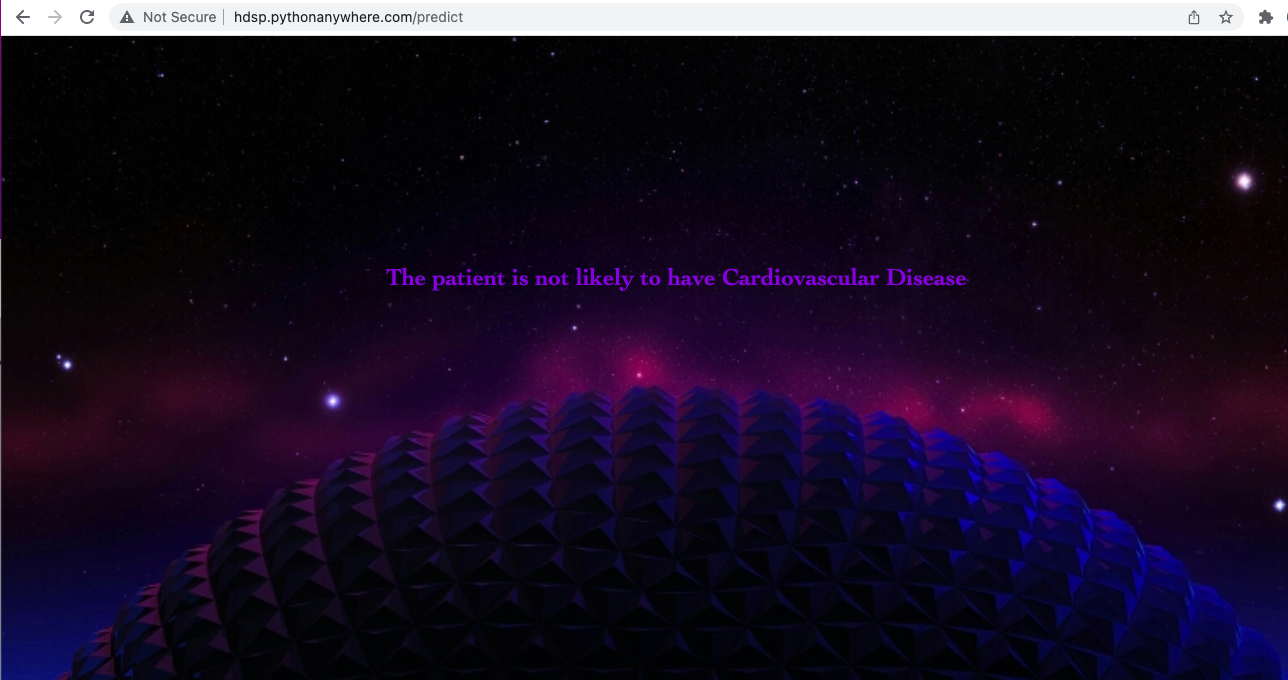
Link for the Heart Disease and Stroke Prediction is,

<http://hdsp.pythonanywhere.com/>

Home Page



Result Page



# 

# 

# 

# 

# 

# CHAPTER 8

# CONCLUSION AND FUTURE SCOPE

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Conclusion and Future Scope

Heart is one of the essential and vital organs of the human body and prediction about heart diseases is also an important concern for human beings so that the accuracy for algorithms is one of the parameters for analysis of the performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that is used for training and testing purposes. Based on the data, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart-related diseases. Each of the above-mentioned algorithms has performed extremely well in some cases but poorly in some other cases. When we perform the analysis of algorithms on the basis of dataset and confusion matrix, we find KNN is the best one. For the Future Scope, a more machine learning approach will be used for the best analysis of heart diseases and for earlier prediction of diseases so that the rate of death cases can be minimized by the awareness about the diseases.

# References

1. *K. Polaraju, D. Durga Prasad, “Prediction of Heart Disease using Multiple Linear Regression Model”, International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.*
2. *Ashwini Shetty A, Chandra Naik, “Different Data Mining Approaches for Predicting Heart Disease”, International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277- 281..*
3. *Sairabi H.Mujawar, P.R.Devale, “Prediction of Heart Disease using Modified K-means and by using Naïve Bayes”, International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273*
4. *Gavhane, Aditi, et al. "Prediction of heart disease using machine learning." 2018 second international conference on electronics, communication and aerospace technology (ICECA). IEEE, 2018.*
5. *Shinde, Rucha, et al. "An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm." International Journal of Computer Science and Information Technologies 6.1 (2015): 637-639*.
6. *Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & Van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PloS one, 14(5), e0213653.*
7. *Wang, A., An, N., Xia, Y., Li, L., & Chen, G. (2014, September). A logistic regression and artificial neural network-based approach for chronic disease prediction: a case study of hypertension. In 2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) (pp. 45-52). IEEE*
8. *Van Bussel, E. F., et al. "Predictive value of traditional risk factors for cardiovascular disease in older people: a systematic review." Preventive medicine 132 (2020): 105986.*
9. *Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., ... & Woo, Y. J. (2011). Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation, 123(8), 933-944.*
10. *Khan, Sundas Naqeeb, et al. "Comparative analysis for heart disease prediction." JOIV: International Journal on Informatics Visualization 1.4-2 (2017): 227-231.*

*.*