1. **Introduction to Machine Learning**

Machine Learning is making the computer learn from studying data and statistics. It can be use in different fields like Artificial Intelligence, Robotics, Generative AI, etc. It use to train or test the accuracy of machine working. This python program has designed to clean the dataset and perform machine learning algorithms to get the accurate dataset. Here, Medical dataset has taken for doing machine learning algorithms to clean the data. By importing pandas library, Medical dataset is read by using pd.read\_csv(“Medicaldataset.csv”). Firstly, perform some numerical operation like mean, median, mode, standard deviation, Variance, percentile etc. Secondly, plotted a Scatter graph, Histogram, Linear regression, etc. At last, perform some algorithms like Test and Train the dataset, build decision tree, bootstrap aggregation, K-nearest neighbour(K-nn), etc. Additionally, checked the duplicate entries and drop, check missing values and fill missing values with 0, change the column name, etc. Data stored in machine learning is not deleted it just overlap with new dataset. It can work on both old and new dataset. It also store historical data.

**Objectives**

* To understand the data structure and model.
* To generate the pattern.
* Evaluate patterns for decision making.
* Improve accuracy and Efficiency.
* For data analysis.
* Classification and Prediction of data.

1. **Methodology**

In this python program, there is a Medical dataset on which has to be performed some pre-processing steps like: -

1. Import csv file
2. Cleaning the data
3. Clustering the data
4. Perform mathematical operation like mean, median, mode, variance, etc.
5. Make Decision Tree
6. Plot histogram and Scatter graph
7. Test and train the data and check accuracy
8. Machine learning algorithms

There is excel sheet of Medical dataset which contain 1320 rows and 9 columns( Age, Gender, Heart rate, Systolic blood pressure, Diastolic blood pressure, Blood sugar, CK-MB, Troponin, Result) In this columns all the information about patient is stored. Import library like pandas, numpy, matplotlib, stats, linear\_model, StandardScaler, Decision\_Tree\_Regressor ,Kmeans, matplotlib.pyplot, datasets, tarin\_test\_split, accuracy\_score, Decision\_Tree\_Classifier, sys, matplotlib.use(‘Agg’), etc.

Csv file is stored in data variable.

**Function use to clean the data**

These all functions are use to clean the data. Data cleaning refer to removing of missing values, noisy data and inconsistent data.

data.head() is use to print first 5 entries and data.tail() is used to print last five entries. data.describe(), data.info(), data.isnull() .sum(), data.drop(), data.rename( columns = {} ), data.copy(), data.fillna(0, inplace = True), data.issnull().sum, data,dropna( inplace = True ), data.duplicate(), len(duplicate), etc.

**Arithmetical Operation**

Some of the mathematical operation are done to cluster the dataset into a group. It take out the average of every numerical data. Some methods are mean(), median(), , mode(), standard deviation( std() ), Variance( var() ), percentile( per() ), etc.

**Plotting of graph**

Plot the graph of clean data for analysis and decision making. With graph data analysis become more easy and accurate result are given. In this python code there are 3 types of graph histogram, scatter graph and clustering graph for visualization and understand the pattern of machine learning. It import matplotlib.pyplot library and use plt.hist(x, value, density = True, facecolor = ‘colour’) to plot histogram and use plt.scatter( x, y ) to plot scatter plot.

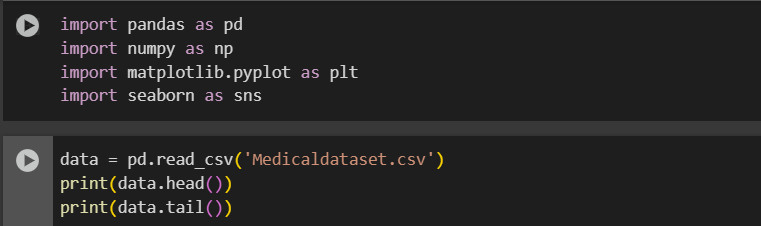
**Testing and training of machine learning**

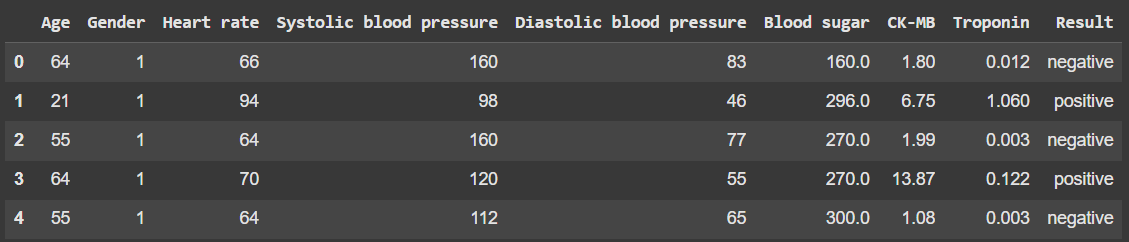
In this python code for machine learning, To get the accurate dataset and to make accurate decision. The machine should be train. Machine is train first on 80% data and test 20%. Then it will give the accurate result. Selected a column on with data should train. Then, select a range of data and plot a graph. To differentiate the distance and point. After cleaning the data and cluster it into group. Then that data may be trained and tested. Trained the data to ddo same task again and again to increase at speed and accurary of result. Test the data on new data to check that machine learn correctly from the train data and work same on test data.

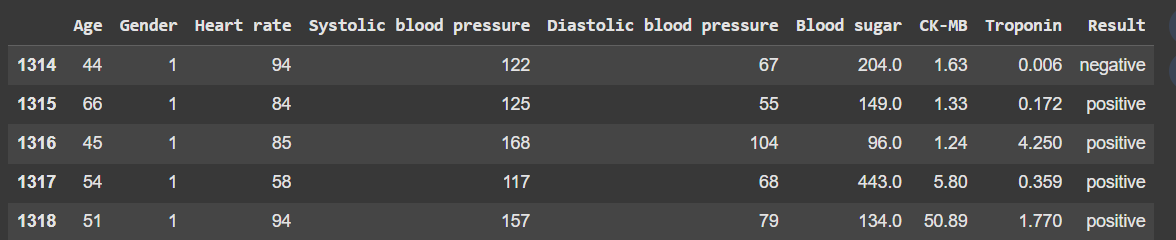
**Machine learning algorithm**

In this python program, There are 2 machine learning algorithms are used. First is clustering and second is K- nearest mean( . K-nn ). Clustering algorithms are use to check the numbers of same data group combine and make it result to perform task. K- nearest neighbors ( K-nn ) algorithm is a simple , supervised machine learning (ML) algorithm that can be used for classification or regression tasks - and is also frequently used in missing value imputation.

1. **Results**
2. **Loading dataset using pandas library and use head(), tail() function**

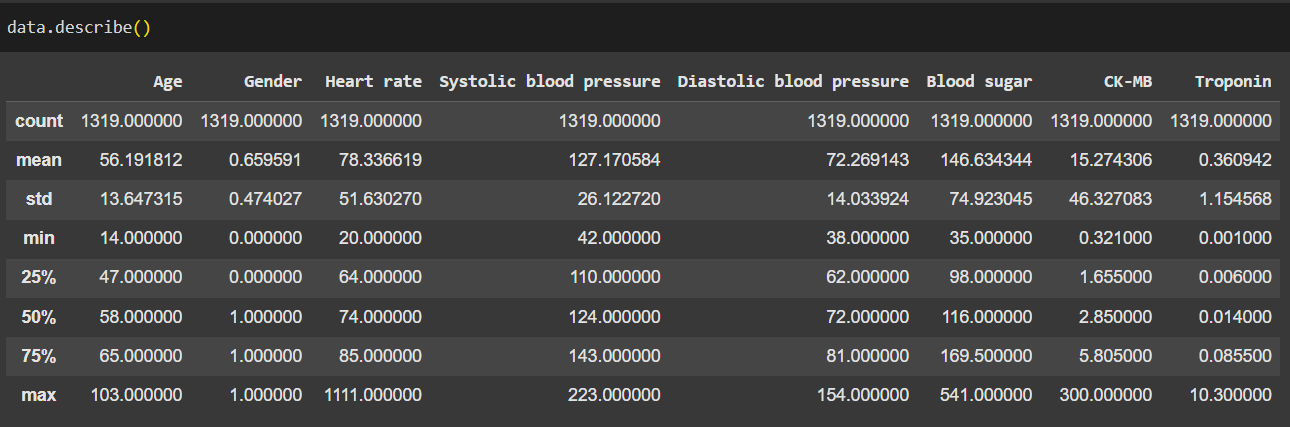


****

****

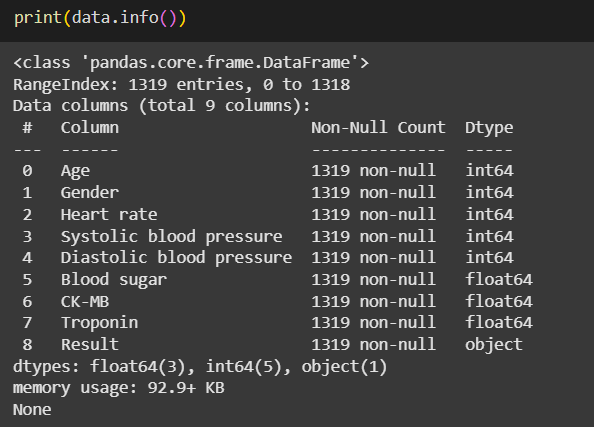
This dataset about the Medical test. The dataset contains the Age, gender, Heart rate, Systolic blood pressure, Diastolic blood pressure, Blood sugar, CK-MB, Troponin, Result. This dataset is downloaded from Kaggle website. pd.read\_csv(‘ ‘) is a method to import a csv file. data.head() is use to print first 5 entries and data.tail() is used to print last five entries.

1. **Describe method**

****

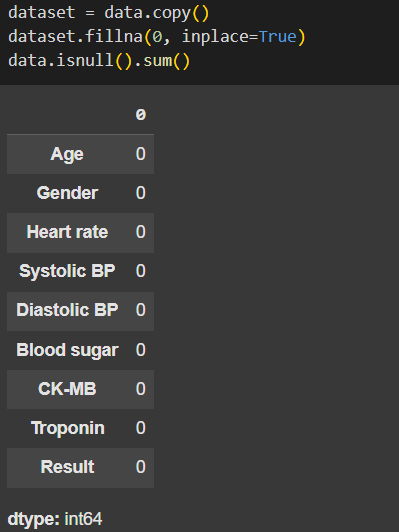
the describe() method has given the descriptive analysis of each column. But cannot assure that these findings as the data is not yet clean. The values can change after cleaning data.

1. **Info() method**

. 

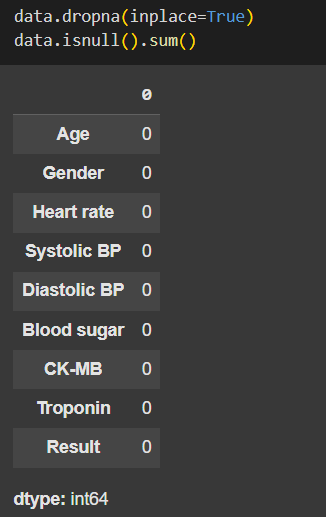
The **info()** method gives us some extra information about the dataset which includes the total number of rows and the data type of each column. There are a total of 1319 rows and 9 columns in this dataset

1. **Checking for null values**



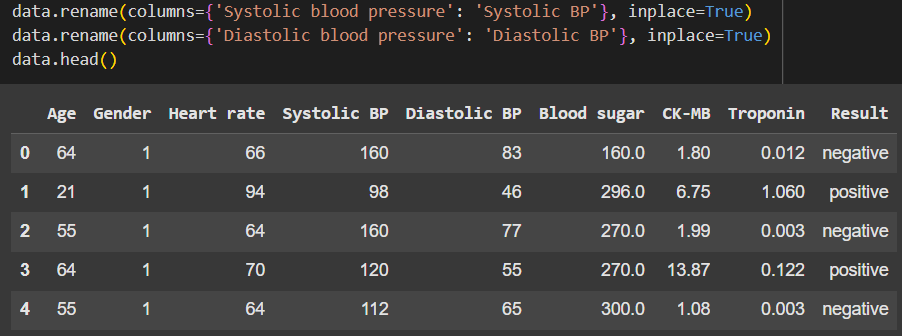
Null values or missing values occur when there is no information. Pandas provide various useful methods to check and handle null values in our dataset. For example, **isnull()** is used to check if there are null values in our dataset or not. The **fillna()** method is used to fill the null values with any other given value. Let us first create a copy of the dataset.

1. **Drop the null values**



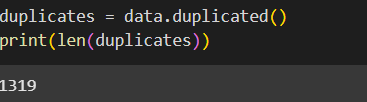
The **dropna()** method will remove the whole row which has a null value in it.

1. **Rename the column name**



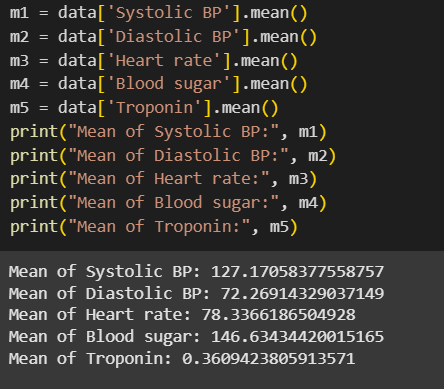
Sometimes, Column name are not clear or inappropriate. In such cases, we can use Pandas to change the name of the columns in our data frame. One of the simplest ways to change the name of the columns is to use the rename() method.

1. **Find the duplicate entries**



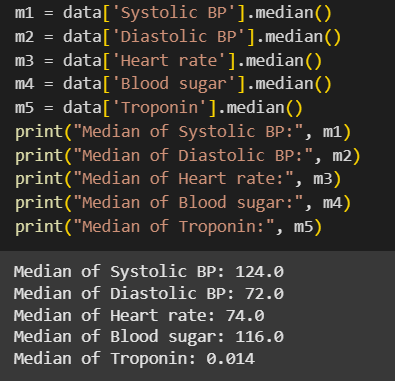
In Pandas, finding duplicate values is much simple by using **duplicated()** method. With len() method we can find how many duplicate entries are presented in our dataset. With pandas, single column or multiple column duplicate value can also be checked. And can also drop the duplicate values using dropna() method.

1. **Performing mean() operation after cleaning**



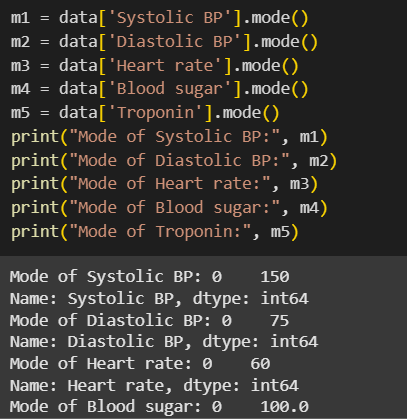
The mean value is the average value. To calculate the mean, find the sum of all values, and divide the sum by the number of values. In python, there is a predefined function to calculate mean is mean() function. It make to calculate mean in a second.

1. **Performing median() operation**



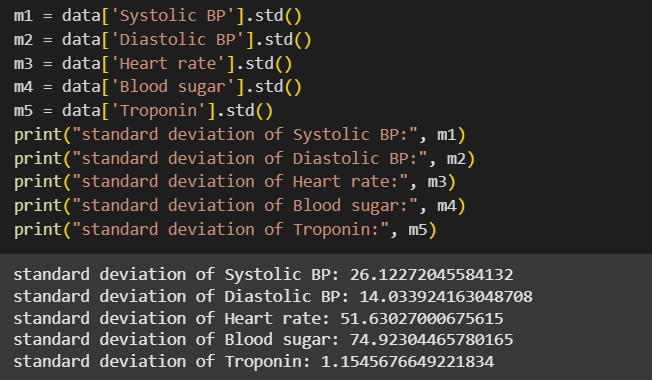
The median value is a middle value of data, after sorted all the values of data. To calculate the median, sort the data and pick middle number, if there are two number in middle,  divide the sum of those numbers by two In python, there is a predefined function to calculate median is median() function. It makes to calculate median easily in a second.

1. **Performing mode() operation**



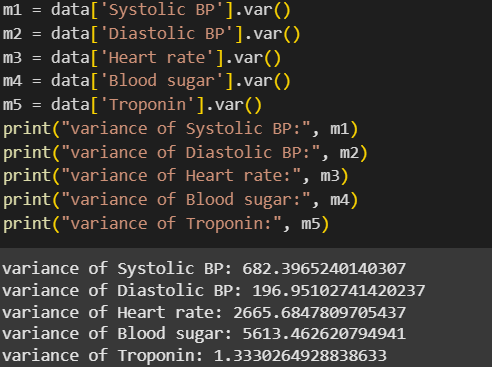
The Mode value is the value that appears the most number of times. To calculate mode, count the number of times data is repetead. In pythod, use mode() function to calculate mode and save the time. Mode() is also a predefined function python

1. **Performing standard deviation() operation**



Standard deviation is a number that describes how spread out the values are. A low standard deviation means that most of the numbers are close to the mean (average) value. A high standard deviation means that the values are spread out over a wider range. In python, To calculate standard deviation, use std() function, it is a predefined function.

1. **Performing Variance () operation**



Variance is another number that indicates how spread out the values are.

To calculate the variance you have to do as follows:

1. Find the mean:

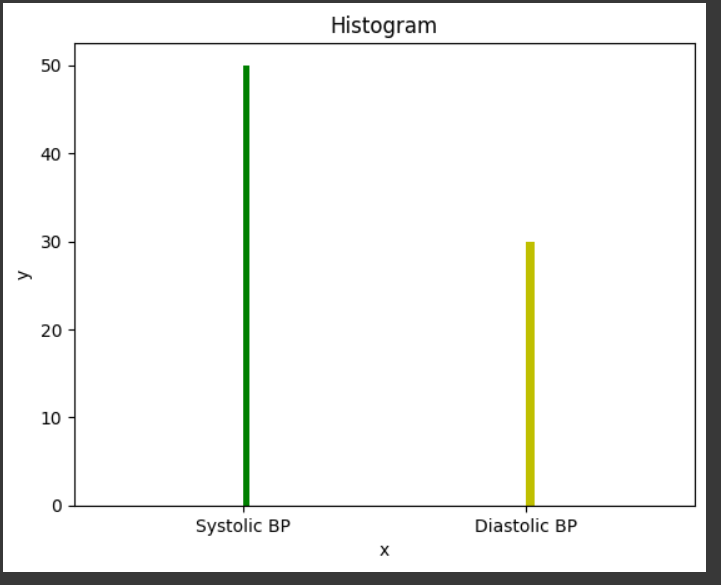
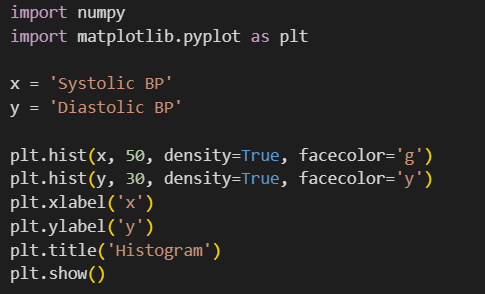
2. For each value: find the difference from the mean:

3. For each difference: find the square value:

4. The variance is the average number of these squared differences:

This process is time taking and confusing. Even can’t assure answer is correct or not luckily , python has var() function to calculate variance

1. **Plot a histogram**

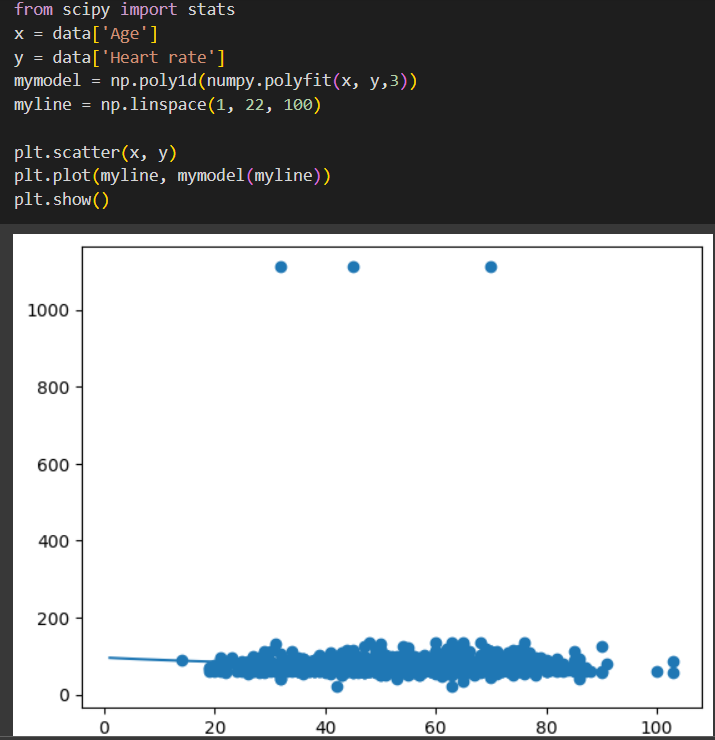


To visualize the data set properly, pandas prefer to draw graph like, A histogram with the data collected. The Python module use Matplotlib.pyplot library to draw a histogram or any graph. As shown in above figure. To plot a histogram, there is keyword hist, to draw histogram give a argument plt.hist(x). The x axis represent Systolic BP and the y axis represent the Diastolic BP.

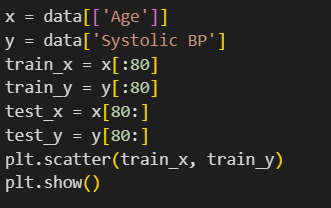
1. **Plot a scatter**

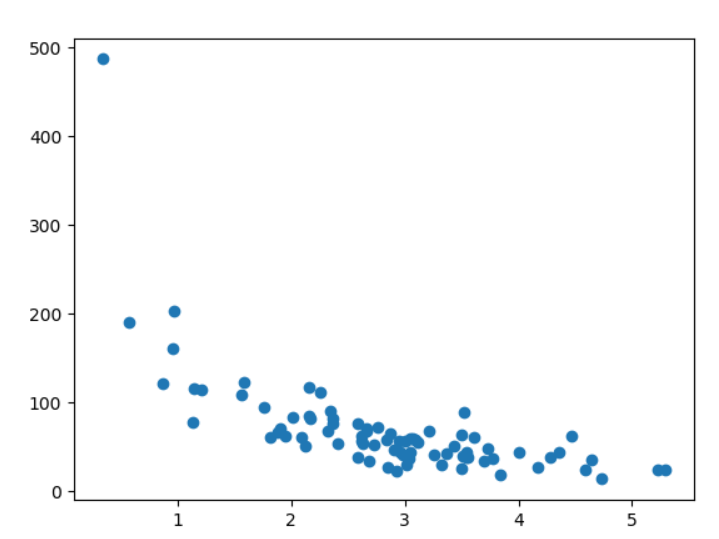
A scatter plot is a diagram where each value in the data set is represented by a dot. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis. The x axis represent Age and the y axis represent the heart rate.

In Machine Learning the data sets can contain thousands-, or even millions, of values. So firstly data should be clean to make scatter plot accurate.

****

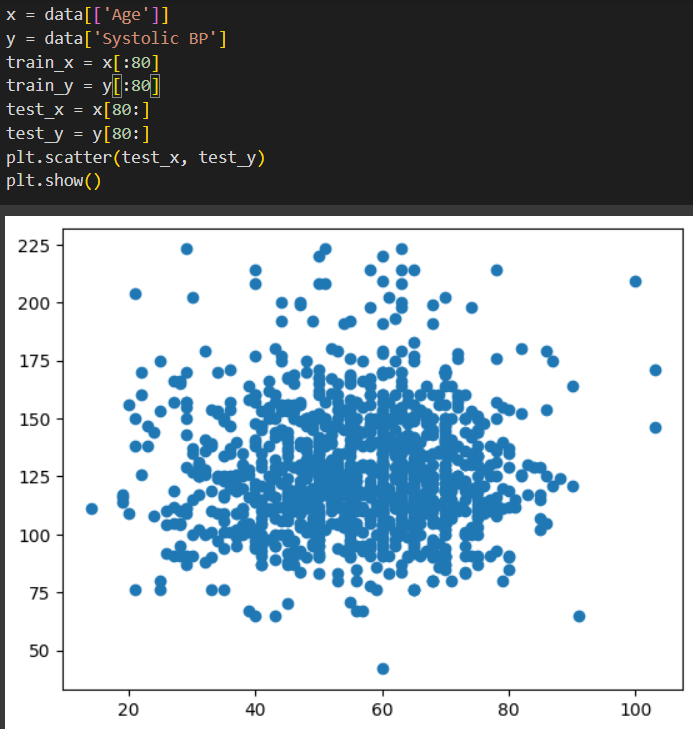
1. **Train data**

****

****

Train data is a method to measure the accuracy of dataset or model. It is called Train because it split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. Train data is used for training set or create the model with help to work on robotics, machine learning , generative AI, etc.

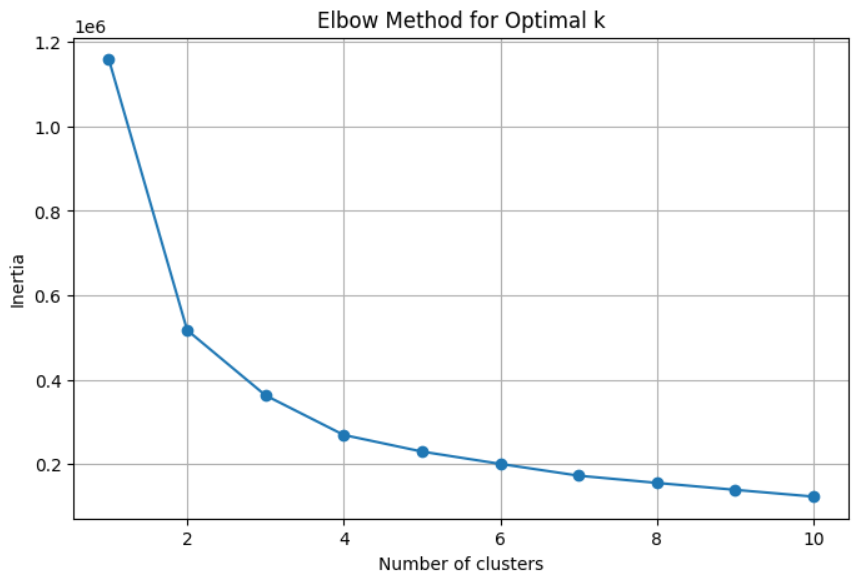
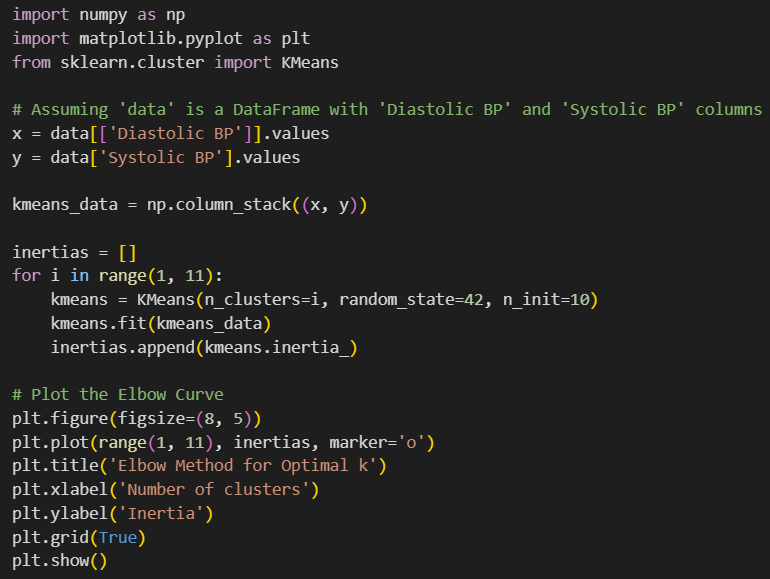
1. **Test the data**

****

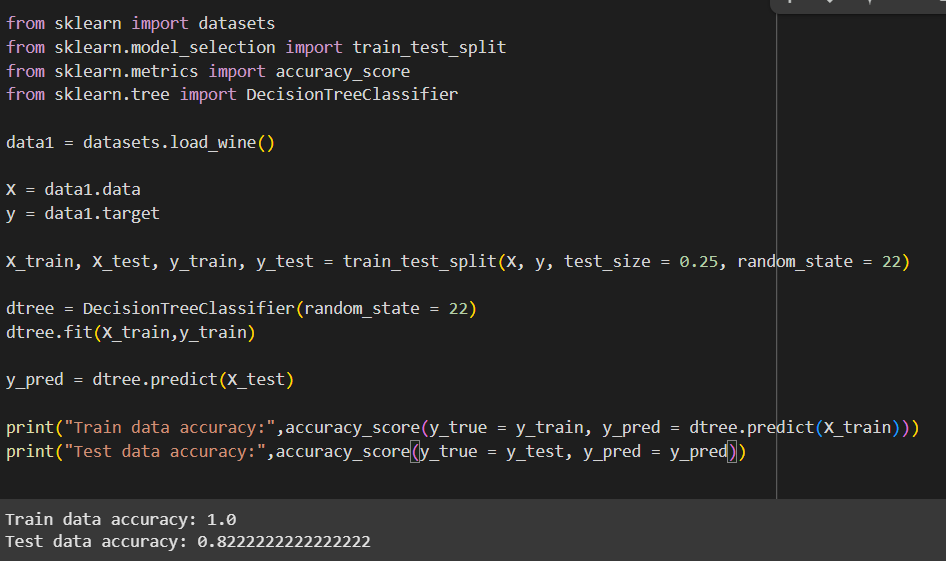
Test data is a method to measure the accuracy of dataset or model. It is called Test because it split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. Test data is used for testing set or test the accuracy of the model with help to work on robotics, machine learning , generative AI, etc.

1. **Machine learning algorithm (Clustering with k nearest mean)**

K-means is an unsupervised learning method for clustering data points. The algorithm iteratively divides data points into K clusters by minimizing the variance in each cluster. In this 3 libraries are imported numpy, matplotlib.pyplot and Kmean subpart sk.learn.cluster. In this graph, it has plotted a Elbow curve graph with scatter to represent the group in dots. It is easy to supervise the data cluster with curve and dot which represent the number of cluster. On the x axis it shows number of cluster where on y axis it shows inertia with the use of K mean.



1. **Bootstrap aggregation**

****

Machine Learning - Bootstrap Aggregation (bagging). In this python module such libararies are used that was dataset, train\_test\_split, accuracy\_score, DecisionTreeClassifier.  Decision Trees, can be used to overfitting on the training set which can lead to wrong predictions on new data. So bagging is used to overcome this problem. It ensembling method that attempts to resolve overfitting for classification or regression problems. Bagging aims to improve the accuracy and performance of machine learning algorithms

1. **Discussion**

The first process is to load the data from csv file and It provides an understanding of the data structure of the dataset and identifying issues such as missing data. Data cleaning was done to remove currupt data, remove missing values, noisy data and inconsistent data containing duplicate values. Methods use to clean the data are data.head() is use to print first 5 entries and data.tail() is used to print last five entries. data.describe() is use to get descriptive data, data.info() give extra information, data.isnull() .sum() is use to check missing values, data.drop() is use to drop missing or duplicate values, data.rename( columns = {} ) is use to change column name with is not proper , data.copy(), data.fillna(0, inplace = True) is used to fill 0 at missing values place, data.issnull().sum to sum the missing value, data.dropna( inplace = True ) drop the missing value, data.duplicate() to check duplicate values, len(duplicate) count number of duplicate entries. The selection of these methods is based on the type and amount of missing data as well as the purpose of the analysis.

Descriptive statistics represent are mean(), median(), , mode(), standard deviation( std() ), Variance( var() ), percentile( per() ), etc. Graphical representations such as histograms and scatter plots are useful for visually exploring data distributions and relationships prior.

Application of linear regression and polynomial models on the 'Heart rate' and 'Age' relationship enables modeling and capturing the trend between the variables. Scaling data using StandardScaler comes in handy for scale-sensitive algorithms like most machine learning algorithms.

Data splitting into training and test is common practice for checking the performance of predictive models on new data. Decision Tree Regressor is a non-linear model that can model relationships. The K-Means cluster analysis, specifically the application of the Elbow Method, is one common method to determine the number of clusters in unsupervised learning. Lastly, the K-Nearest Neighbors classifier is an easy-to-use yet powerful algorithm for proximity-based classification to known points.

1. **Conclusion**

In this machine learning, importing a csv file and work on it by python programming language. It ensure to perform preprocessing steps on the dataset for better analysisnand make decision for business user. Firstly, data cleaning was done to remove currupt data, remove missing values, noisy data and inconsistent data containing duplicate values. By After cleaning the data. Some pre defined functions are perform to get the numerical data like mean(), median(), , mode(), standard deviation( ), Variance( ), percentile( ), etc. After getting the data Visualization was done such as histogram, scatter plot are generated. linear regression was doe on distributed data. Then after all this steps when pattern is evaluated, Train and test the data module. 80% of training data and 20% of testing the data. At the end Train and test data are used in machine learning algorithms to create a clustering, bootstrap aggregation and calculate the accuracy of data.