**PROJECT TITLE**

"Exploring the Factors and Predictive Models for Heart Disease:

A Comprehensive Analysis Using a Heart Disease Dataset"

**COURSE NAME**

Capstone Project I

**COURSE CODE**

DAB322

**SUBMITTED BY**

Vijay Lathwal (0822116)

Group: - 1

**PROBLEM STATEMENT**

*As it is rightly said, ‘Health is Wealth’. We have realized this fact in the pandemic time after witnessing the brute effects of Covid-19 on people of all age groups. Apart from this, another major contributor to the death rate is heart related diseases.*

*Located in the chest region of the body, the heart beats at around 80 times per minute. Even though it is just the size of an average human fist, it is the strongest muscle which continuously to pump blood to the body organs, even at rest.*

*Heart diseases have been known to take a major toll of people’s lives. As a layman, we may feel that the common factors for heart related diseases are cardiac arrest or blockages. But the dataset under analysis describes multiple different medical parameters associated with the heart and their typical values. We will be analyzing the relationships between them and study the implications of changes in those parameters. In this project, we will be incorporating most trending and powerful tool namely Excel, Sql, Tableau and Python.*

*Q1. What role does age play in the development and occurrence of heart diseases?*

*Q2. What is the correlation between the type of chest pain and gender?*

*Q3. How does the levels of cholesterol is changing with gender?*

*Q4. Does angina while exercising implies higher probability of chest pain?*

*Q5. What are the patterns of Thalassemia occurrence in relation to gender for the Normal, Fixed, and Reversible types of the disorder?*

*Q6. How does the slope of the peak exercise ST segment affect the ST depression?*

*Q7. What were the observed results during a fluoroscopy procedure?*

**OBJECTIVE AND SCOPE OF THE PROJECT**

*► Objective:*

*• The dataset contains the records for the patients and their medical parameters details and the target variable whether they will suffer from heart disease or not.*

*• The aim of this project is to use the given data and perform machine learning and data analysis to infer key metrics and patterns in the dataset. In addition to this, different visualizations are developed to depict meaningful relationships.*

*► Benefits:*

*• The data analysis will reveal some common and unique patterns in the dataset related to the medical parameters.*

*• Data visualizations will enhance the understanding of the effect of the high or low of these features on the chances of heart rate and give a better chance of prediction.*

**STEPS FOLLOWED**

***1) Data Extraction:*** *This step involves extracting the data from different sources relevant to the problem statement or obtaining data from the client.*

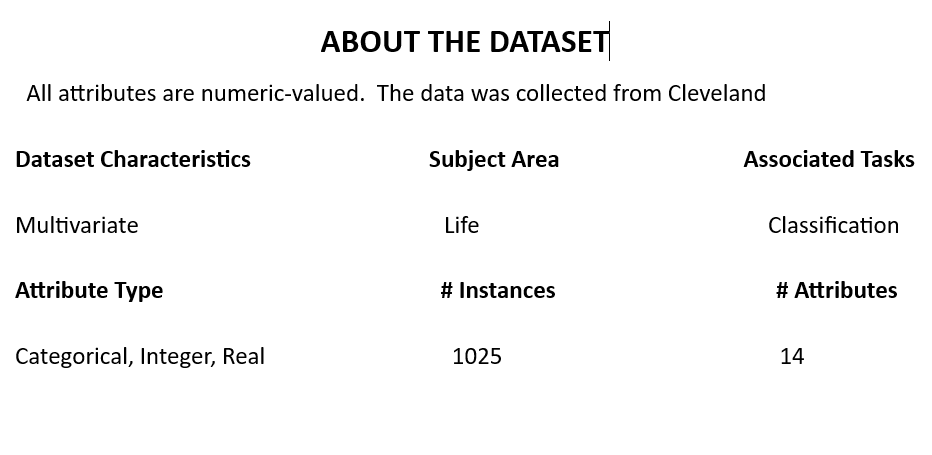
***2) Data Preprocessing:*** *Once the raw data is obtained, we need to ensure that the data is free from errors. We perform Exploratory Data Analysis followed by Data Cleaning which involves imputing missing values, removing duplicates, finding anomalies or outliers and treating them.*

***3) Data Exporting:*** *The preprocessed data is exported to a .csv file to be used for analysis.*

***4) Data Loading and Modification:*** *The preprocessed data in .csv file is loaded into SQL, added some columns and then linked it into the Tableau Desktop for analysis purpose and modified for simplicity purpose.*

***5) Data Analysis:*** *Once the data is loaded, we perform the data analysis using Tableau features and store the visualizations in Tableau worksheets.*

***6) Machine Learning:*** *Developed predictive models for heart disease diagnosis and then assessed the performance of predictive models.*



**HEART DISEASE ANALYSIS**

* We imported heart disease dataset into tableau.
* Since the dataset contained many categorical columns which store the categories in the form of integers. We converted these numbers into meaningful phrases which will be understandable to the viewer and also easy to understand the terms used in the visualizations.
* Once the data has been loaded into the Tableau Desktop software, we perform the analysis for the various medical parameters provided in the dataset and study the relationship between them.
* Link : **[Visualization](file:///C:\\Users\\amrit\\Desktop\\Project\\CapstoneProject.twb)**

**CLASSIFICATION ALGORITHMS**

Classification algorithms are machine learning algorithms that are used to classify data into different classes or categories. These algorithms learn patterns and relationships from labeled training data and then apply that knowledge to classify new, unseen data. Here are some commonly used classification algorithms:

**1.** **Logistic Regression:** Logistic Regression is a linear model that is used for binary classification problems. It estimates the probability of an instance belonging to a particular class.

**2. Decision Trees:** Decision Trees create a tree-like model where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label. Decision Trees can handle both binary and multi-class classification.

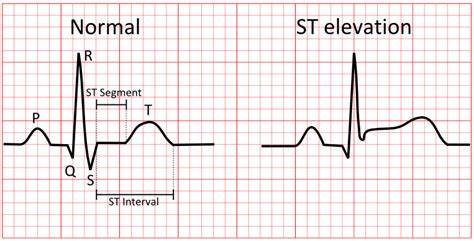
**3. Support Vector Machines (SVM)**: SVM is a powerful algorithm that finds a hyperplane in a high-dimensional space to separate the classes. It can handle both linear and non-linear classification tasks using kernel functions.

**4. K-Nearest Neighbors (KNN):** KNN is an algorithm that classifies instances based on the majority vote of their k nearest neighbors in the feature space. It does not make any assumptions about the underlying data distribution.

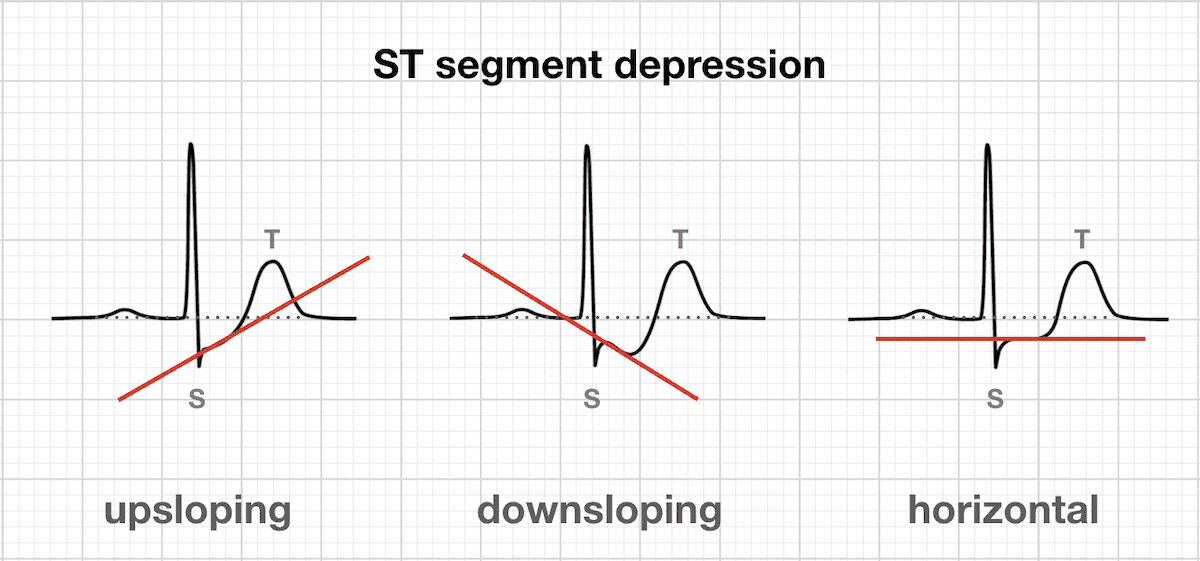
**5. Random Forest:** Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It's an ensemble learning method that combines multiple decision trees to make more accurate predictions.

**ATTRIBUTES**

* **age**: - in years
* **sex:** - (1 = male; 0 = female)
* **chest pain type**: -
* Value 0: typical angina (physical exertion or emotional stress)
* Value 1: atypical angina (related to reduced blood flow to the heart
* Value 2: non-anginal pain (muscle strains, acid reflux (heartburn), inflammation of the chest wall, anxiety, or other non-cardiac conditions)
* Value 3: asymptomatic (there are no noticeable symptoms or pain present)
* **resting blood pressure** (in mm Hg on admission to the hospital)
* **serum cholesterol** (in mg/dl)
* **fasting blood sugar > 120 mg/dl** (1 = true; 0 = false)
* **resting electrocardiographic results: -**
  + Value 0: normal
  + Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)



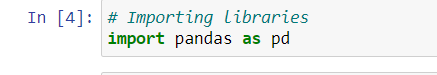
* + Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
* **maximum heart rate achieved**
* **exercise induced angina** (1 = yes; 0 = no)
* **old peak** = ST depression induced by exercise relative to rest
* **the slope of the peak exercise ST segment**
* Value 1: upsloping
* Value 2: flat
* Value 3: down sloping



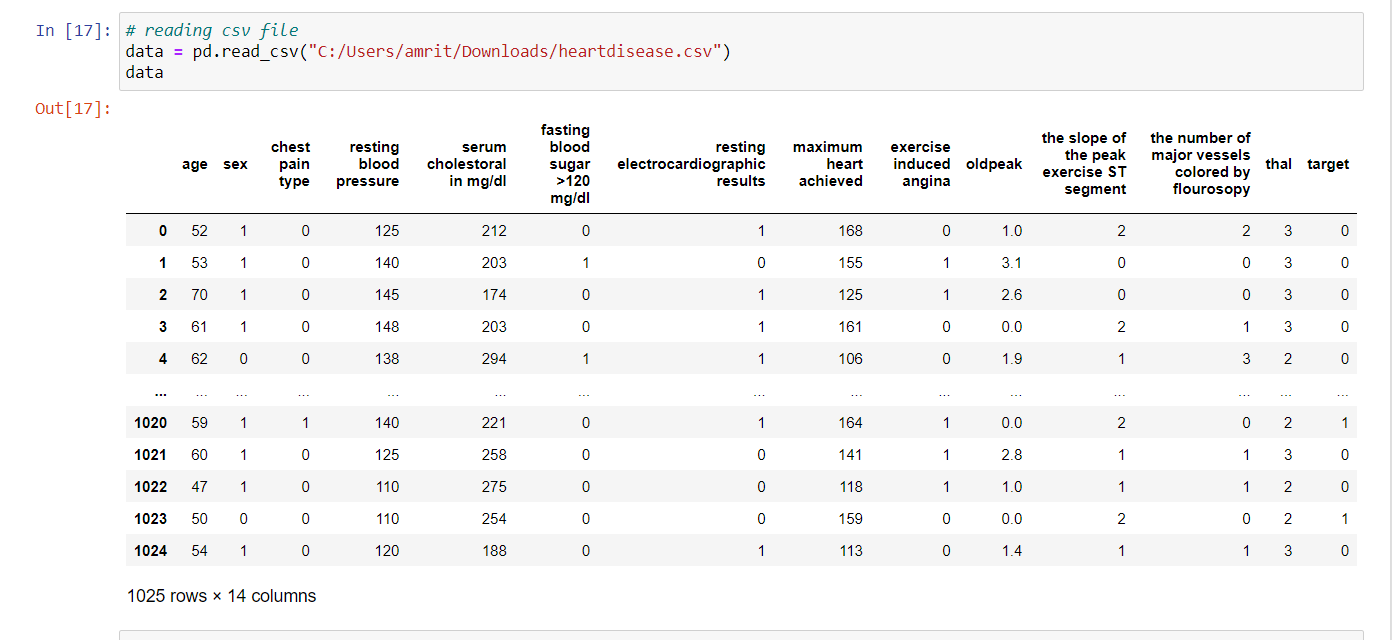
* **# of major vessels (0-4) colored by fluoroscopy (**It refers to the # of major coronary arteries that are visualized and show blood flow during a fluoroscopy procedure)
* **Thalassemia:**
* 0 = normal (there are no significant defects in blood flow to the heart)
* 1 = fixed defect (It suggests an area of the heart where the blood flow is permanently reduced or blocked, often due to scar tissue from a previous heart attack or other heart conditions)
* 2 = reversable defect (It suggests an area of the heart where blood flow is temporarily reduced during stress or exercise but improves when the stress is relieved)

**STEPS TAKEN**

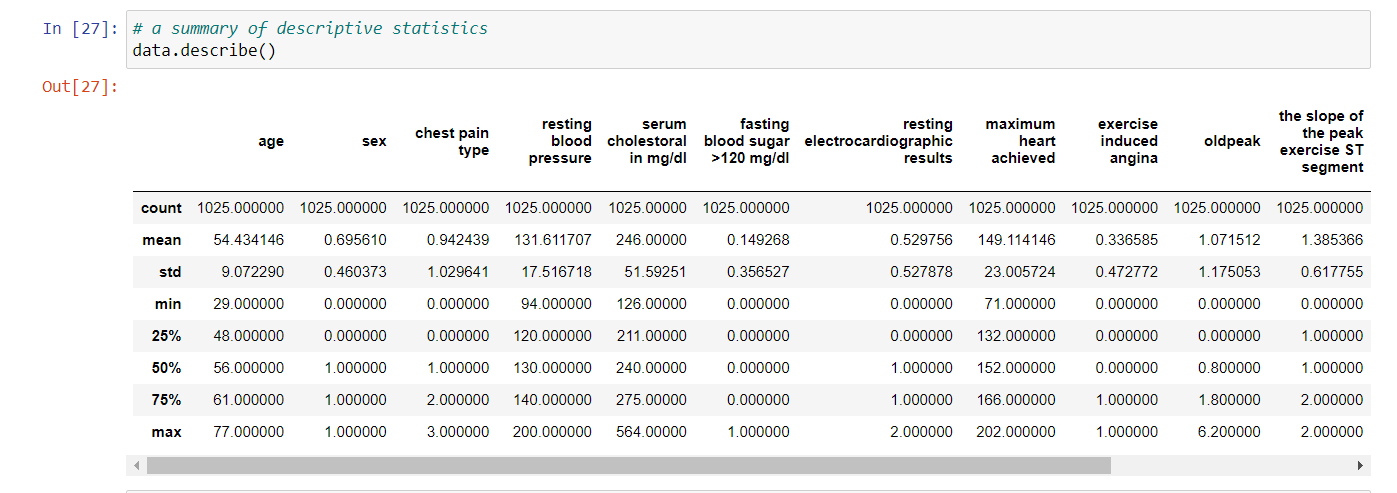
1. **Importing libraries: -** “import pandas as pd” imports the pandas library and assigns it the alias pd. This allows you to use pandas functions and classes in your code.



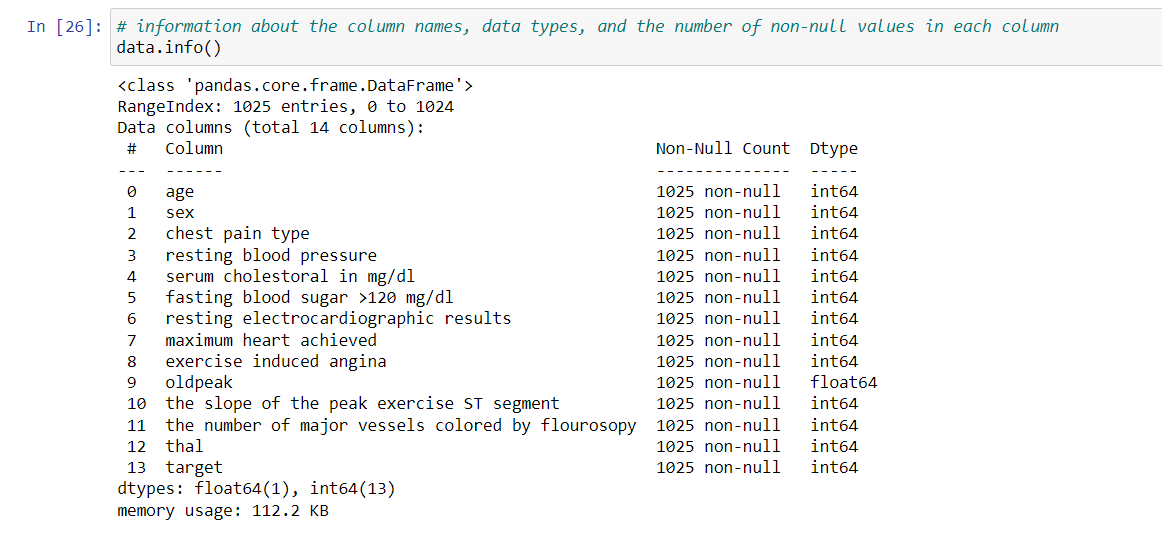
**2. Importing a dataset: -** The given code reads the CSV file located at the specified file path and assigns the resulting data to the Data Frame data.



**3. Summary: -** The data.describe() function provides descriptive statistics for each numerical column in the Data Frame data. It calculates various statistical measures that summarize the distribution, central tendency, and spread of the data.

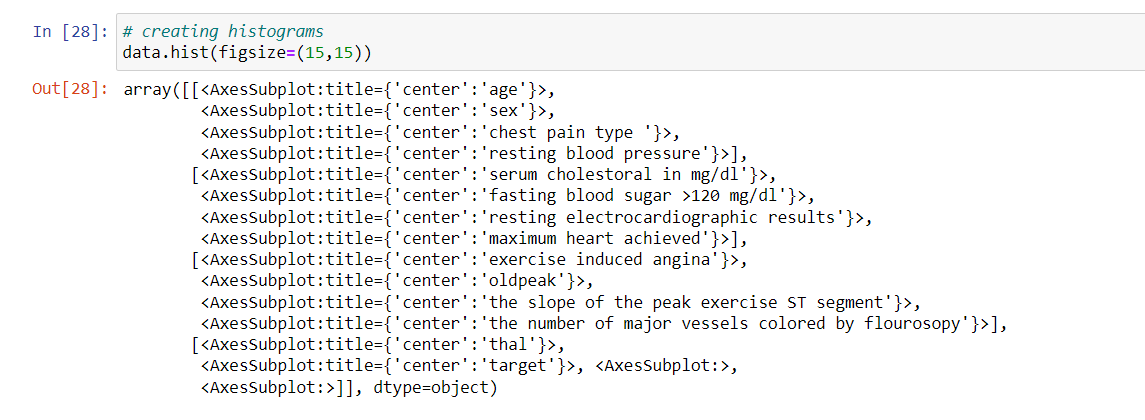


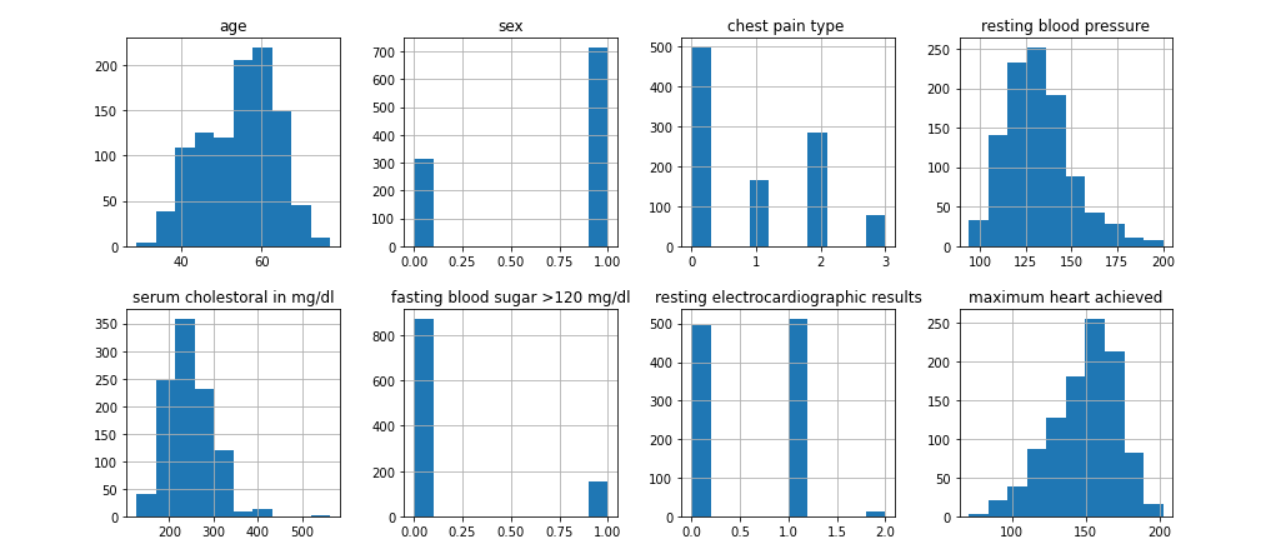
**4. Gathering information: -** The data.info() function provides a concise summary of the Data Frame df, including information about the column names, data types, and the number of non-null values in each column.

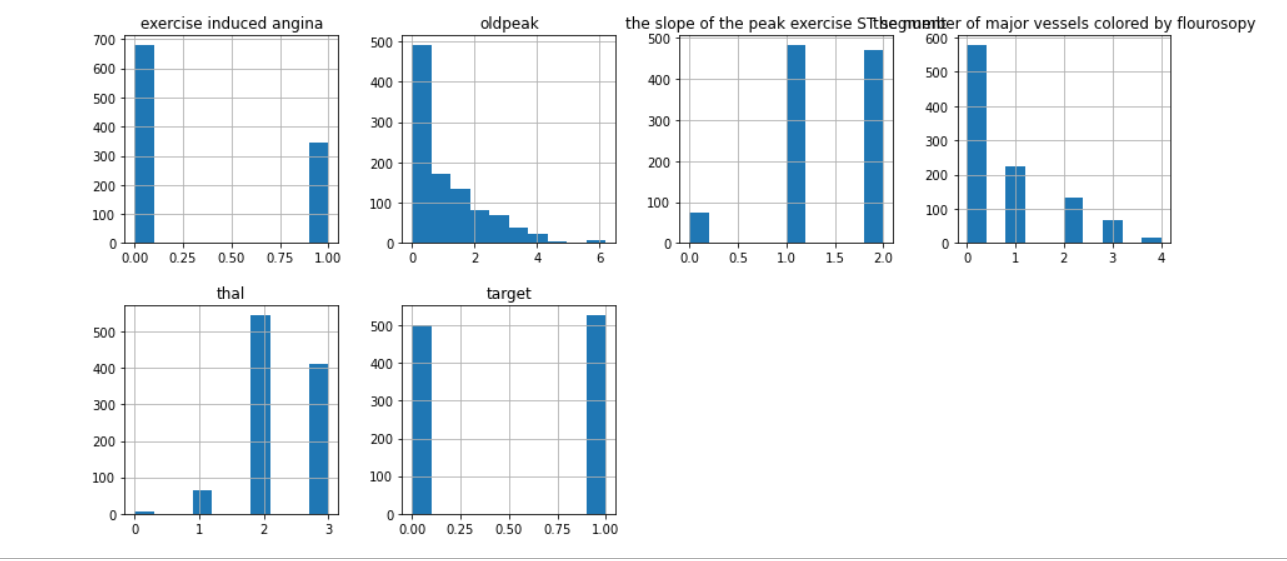


**5. Plotting histogram: -** By calling data.hist(figsize=(15,15)), you can visualize the distribution of values in each column of the DataFrame data. Histograms are a useful tool to understand the underlying distribution of numerical data, including information about central tendency, spread, and potential outliers. They provide insights into the shape of the data and can help identify patterns or anomalies in the dataset.

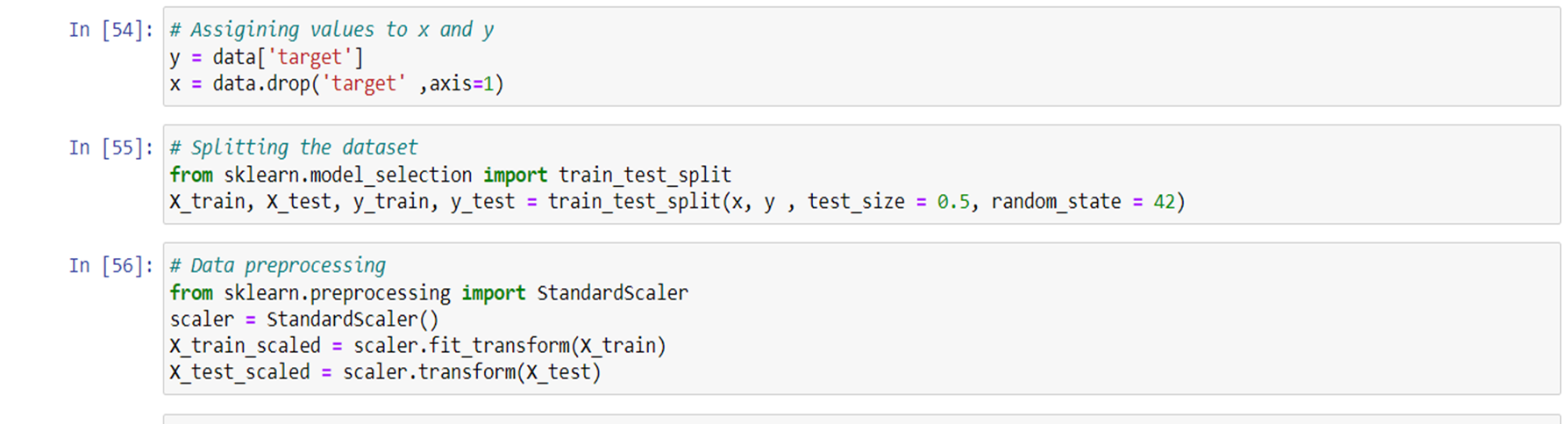
* data.hist() is a function provided by pandas that generates histograms for each numerical column in the Data Frame data. It creates a separate histogram for each column and displays them together.
* The figsize = (15,15) argument specifies the size of the figure that will be created to display the histograms. In this case, the figure size is set to 15 inches by 15 inches.



****

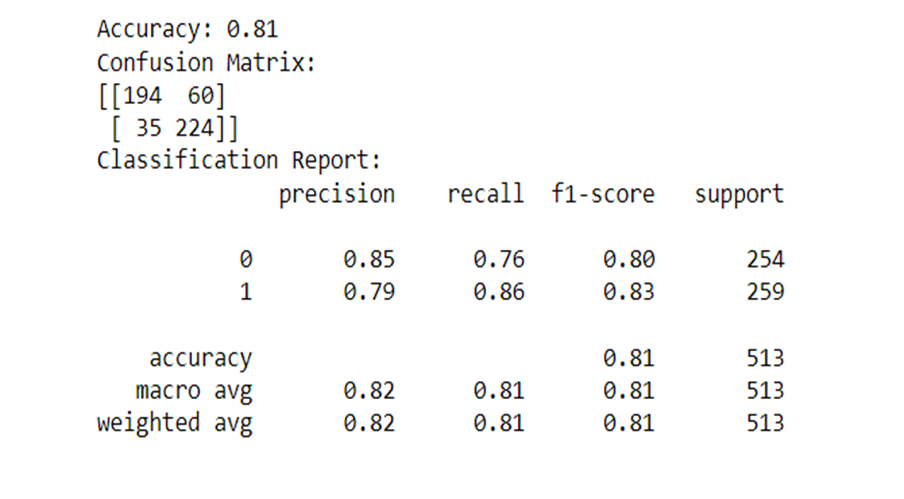


***6. Classification algorithms****: - before applying algorithms, we will preprocess the data . Data preprocessing is a crucial step in machine learning and it is very important for the accuracy of the model. Data contains noise, missing values, it is incomplete and sometimes it is in an unusable format which cannot be directly used for machine learning models. These techniques allow us to transform the raw data into a clean and usable data set and make the data more meaningful by rescaling, standardizing, binarizing and so on.*

**

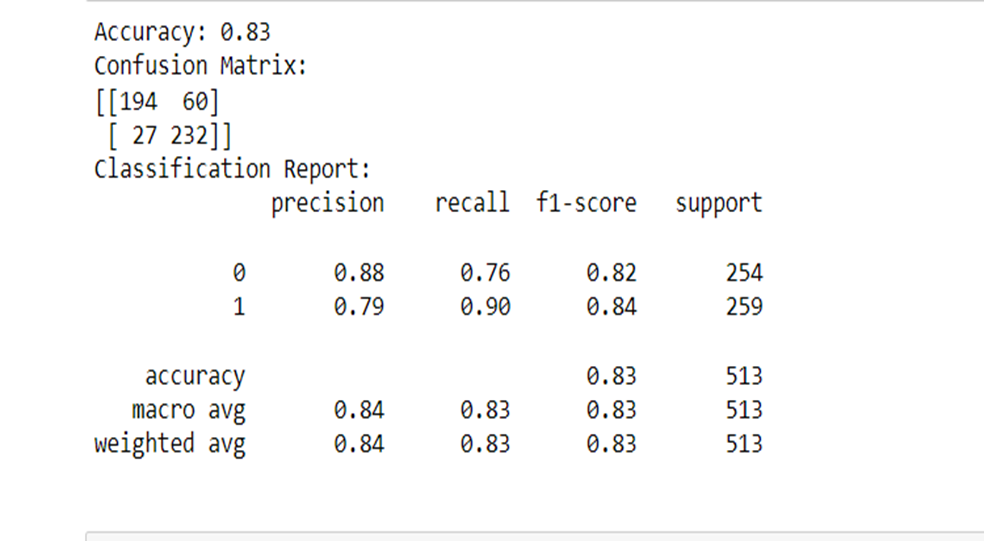
**LOGISTIC REGRESSION**





**OPTIMIZATION**

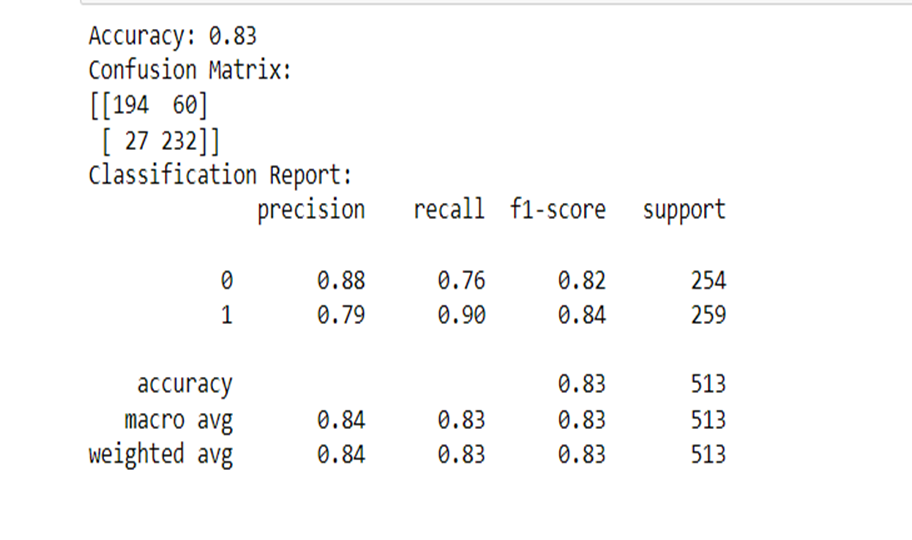
****



**Decision tree**

A screenshot of a computer code

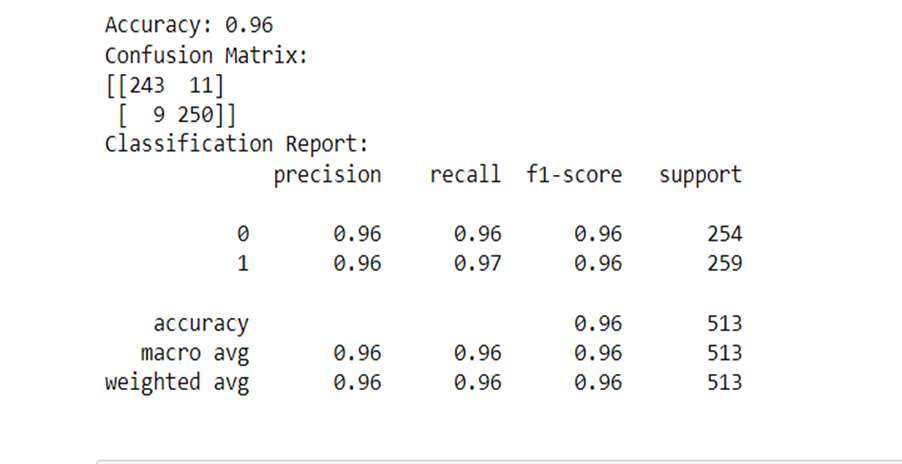
Description automatically generated



**RANDOM FOREST**

A screenshot of a computer program

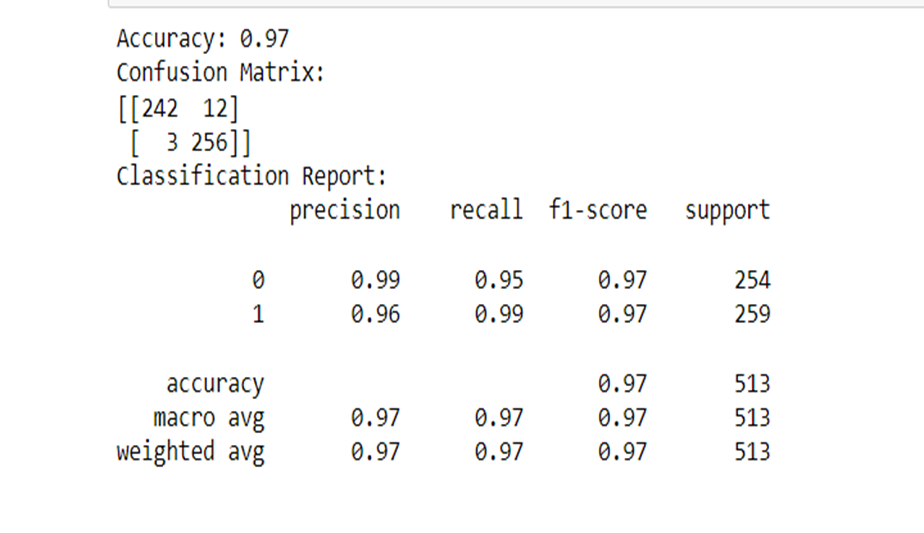
Description automatically generated



**GRADIENT BOOSTING**

A screenshot of a computer program

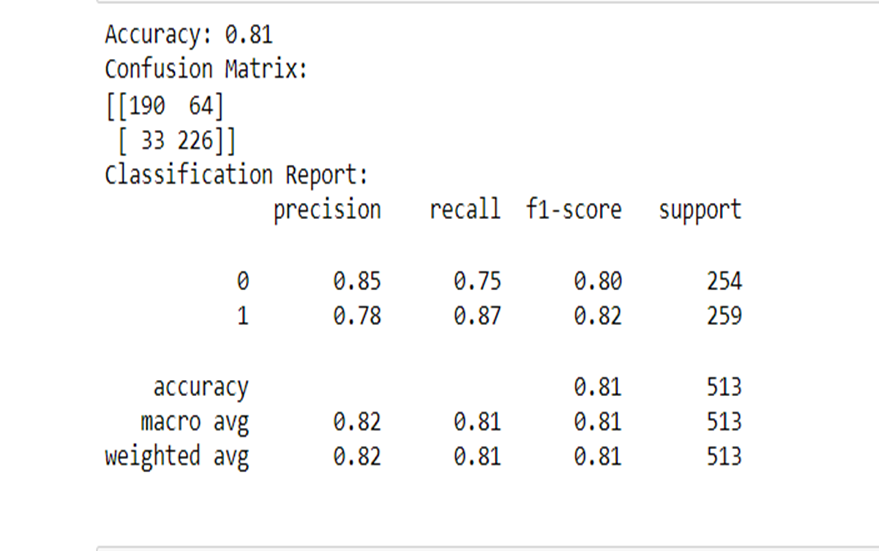
Description automatically generated



**SUPPORT VECTOR MACHINE**

A screenshot of a computer program

Description automatically generated



**OPTIMIZATION**

A screenshot of a computer program

Description automatically generated

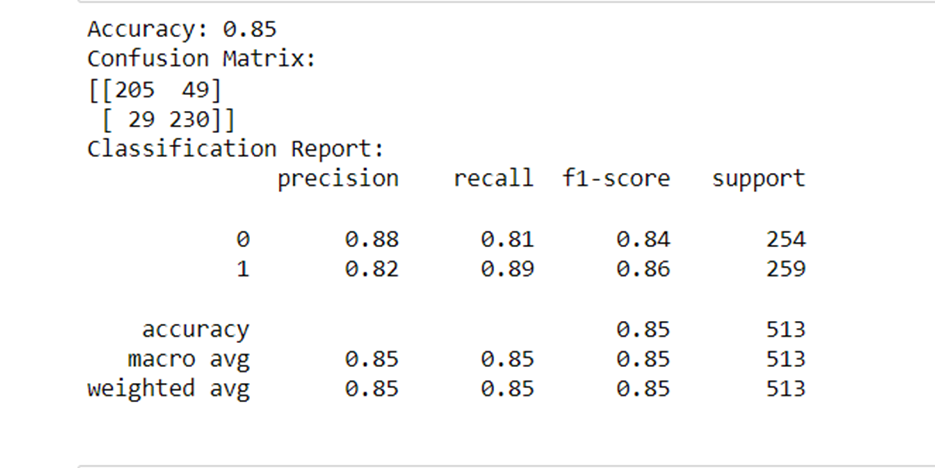
A screenshot of a computer

Description automatically generated

**K NEAREST NEIGHBORS**

A screenshot of a computer program

Description automatically generated



**EXPORTING RESULTS TO EXCEL FILE**

A screenshot of a computer program

Description automatically generated