1. **Title of the project:** Heart Failure Prediction
2. **Project Brief**

Introduction: Cardiovascular diseases (CVDs) remain a leading cause of global mortality, responsible for approximately 31% of all deaths worldwide. Among CVDs, heart failure is a prevalent condition, contributing significantly to premature deaths. Early detection and management of cardiovascular diseases are crucial for individuals at risk. This project aims to leverage machine learning models to predict heart failure based on relevant patient features.

Problem Statement: Develop a classification filter using various machine learning models to predict heart failure. The dataset comprises 12 features, including patient demographics, medical history, and diagnostic indicators. The goal is to build a robust model capable of accurately classifying individuals into two categories: those with heart disease (output class 1) and those without (output class 0). Additionally, the project aims to compare the performances of different classification models to identify the most effective approach.

1. **Objective:**

The primary objective of the project is to create a classification filter for heart failure prediction. This involves implementing and evaluating various classification models on the provided dataset. The comparison of model performances will help identify the most effective model for accurate and reliable predictions.

1. **Deliverables:**

i. Project Documentation:

- Detailed project overview, objectives, and significance.

- Description of dataset attributes and their relevance to heart failure prediction.

- Explanation of the chosen machine learning models and their rationale.

- Overview of the evaluation metrics used to assess model performance.

ii. Data Preprocessing:

- Cleaning and handling missing data.

- Encoding categorical variables.

- Normalizing or scaling numerical features.

iii. Exploratory Data Analysis (EDA):

- Visualization of data distribution and correlations.

- Identification of patterns or trends within the dataset.

iv. Model Development:

- Implementation of classification models (e.g., Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, etc.).

- Hyperparameter tuning for optimal model performance.

v. Model Evaluation:

- Comparison of model performances using metrics such as accuracy, precision, recall, F1 score.

- Visual representation of model evaluation results.

vi. Conclusion and Recommendations:

- Summary of findings.

- Identification of the most effective model for heart failure prediction.

- Recommendations for further improvement or refinement.

By delivering these components, the project aims to provide a comprehensive understanding of heart failure prediction using machine learning, along with actionable insights for healthcare practitioners and researchers.

1. **Resources**

**Data set source:** Data is provided in .csv format for this project.

Also, similar data can be found in: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

**Software**: Anacoda, Jupyter notebook

Thomas Medhi

Email: thomas.medhi111@gamil.com

Phn: 9101314537

**Individual Details:**

Debajit Talukdar

Email: debajit.talukdar24@gamil.com

Phn: 6901372468

1. **Milestones**

**Define a problem:** Develop a machine learning classification model to predict heart failure based on patient attributes, aiming to assist in early detection and management of cardiovascular diseases. Evaluate and compare the performance of different classification models for effective implementation in healthcare.

**Understanding the business problem:** Addressing the global health challenge of cardiovascular diseases, this project seeks to harness machine learning to predict heart failure, facilitating timely intervention for individuals at risk. The focus is on enhancing healthcare outcomes through early detection and informed decision-making.

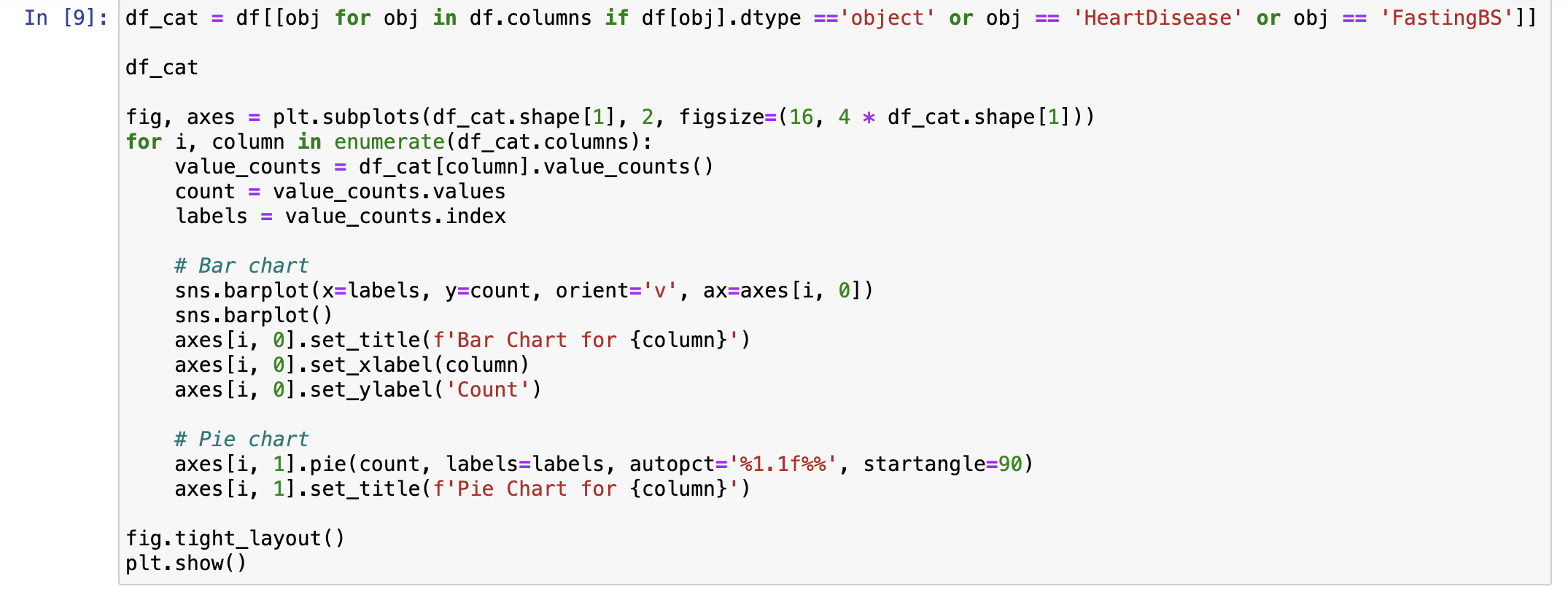
A screenshot of a computer

Description automatically generated**Get the Data:** Data is imported form the shared .csv file. CSV file is saved in the same pwd as the .ipynb file and using panda data is fetch as shown below.

**Explore, Pre-process data, EDA and Feature engineering:** Using function - head(), shape, describe(), info data is explored. Find the screenshot below

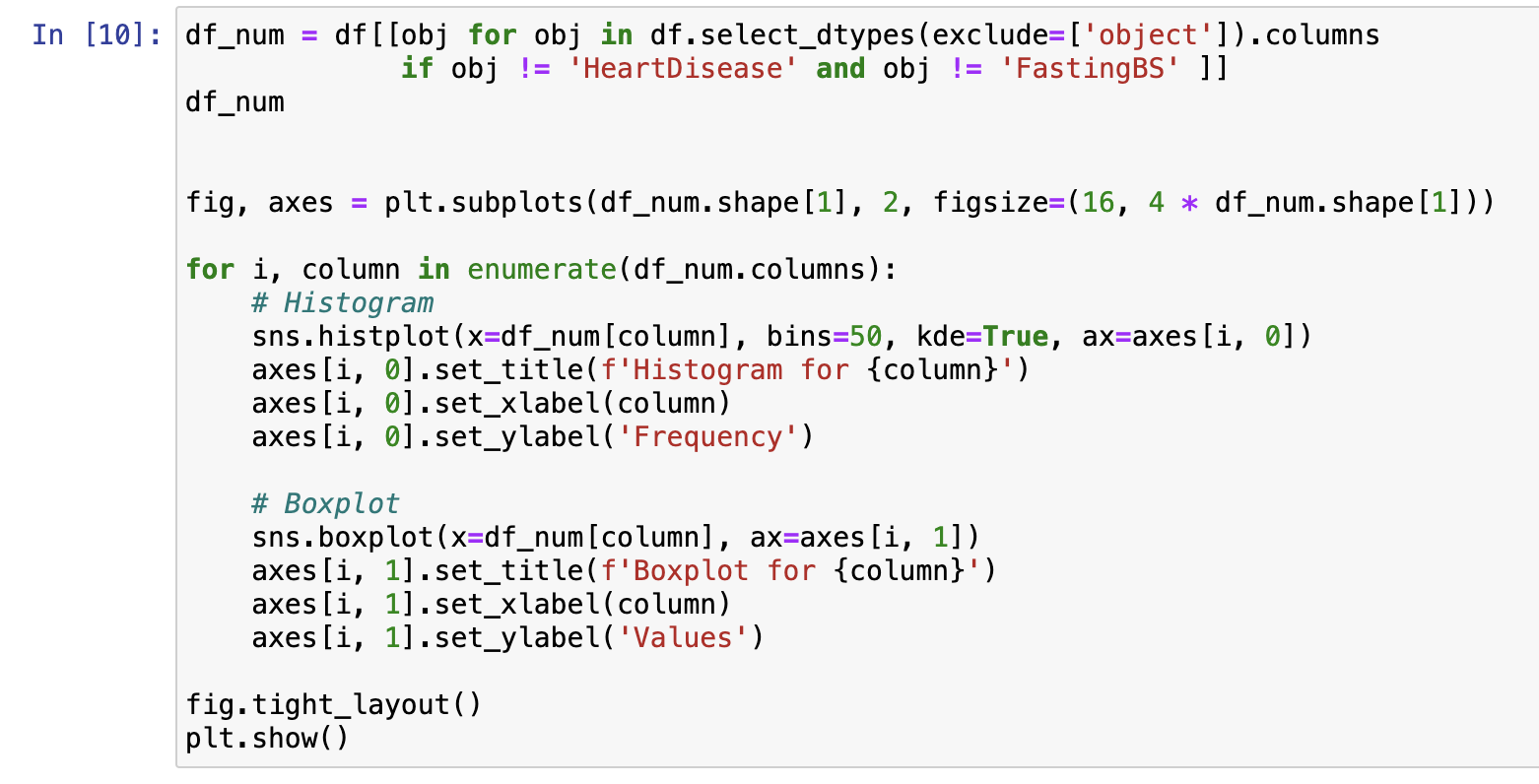
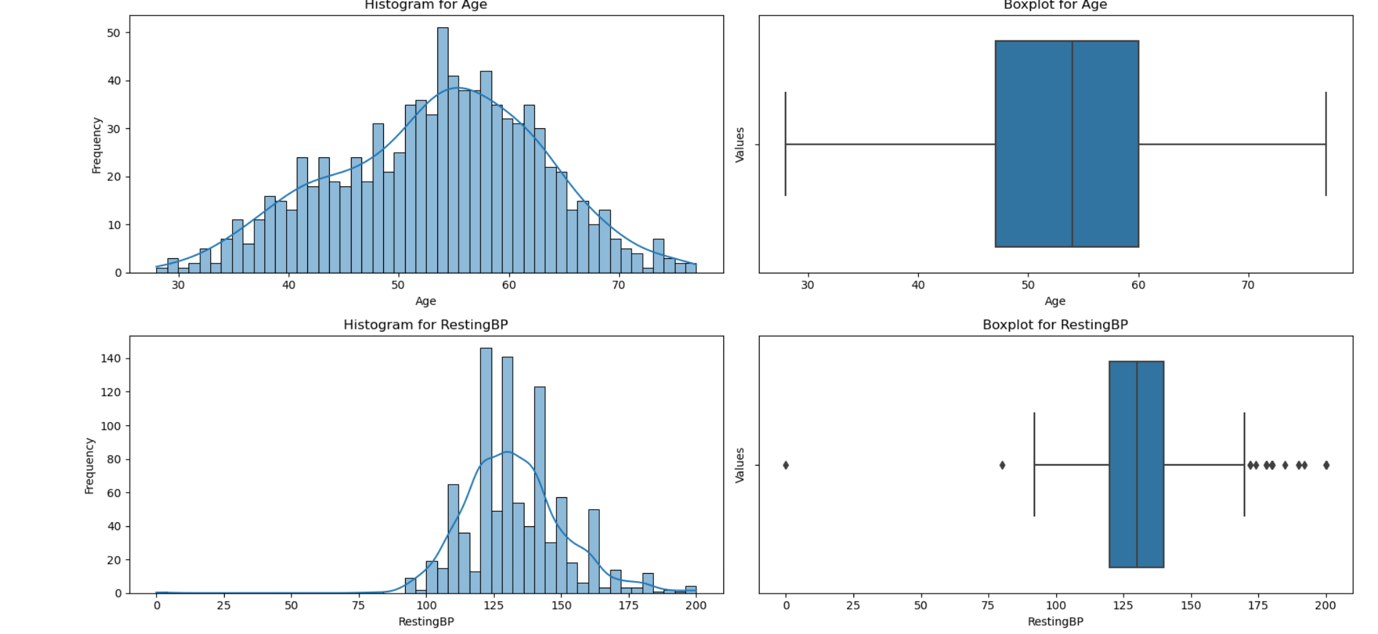
A screenshot of a computer

Description automatically generated

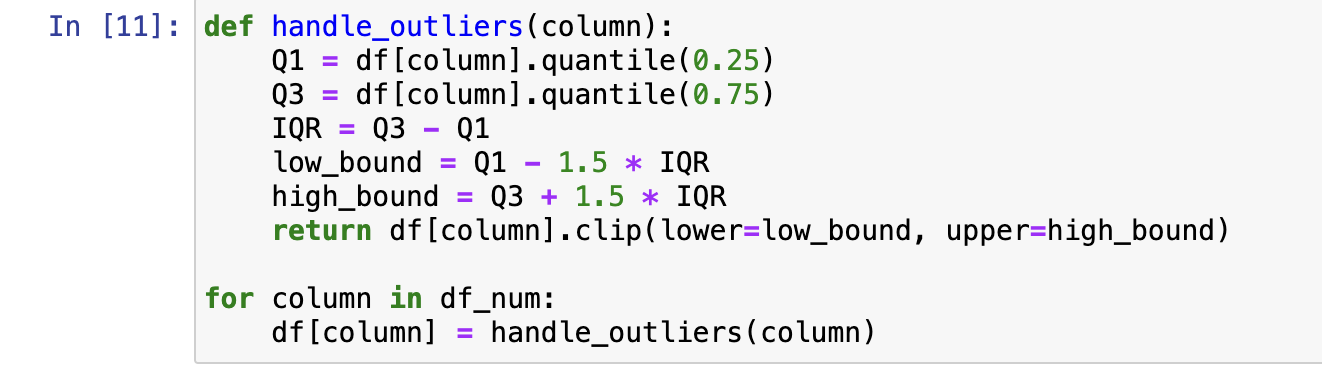
Also, for categorical data fields bar and pie chart have been plotted to check the count by each filed.

A screenshot of a graph

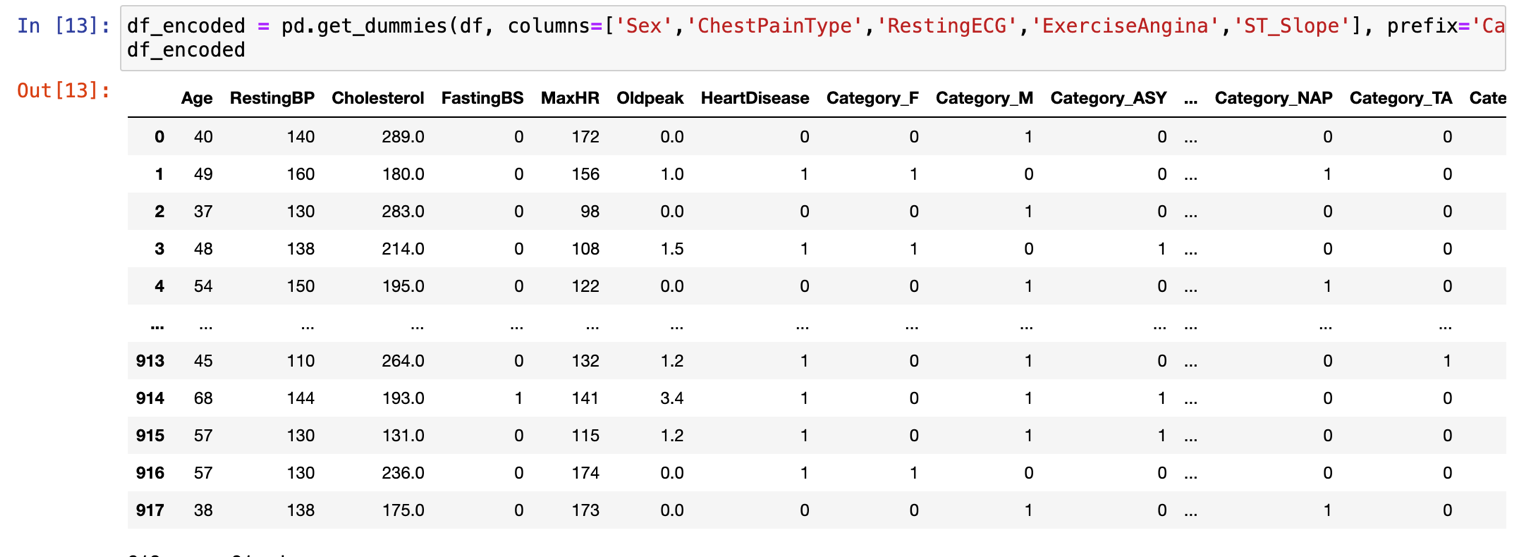
Description automatically generated

For numerical data fields, distribution and box plot have been plotted to check distribution and 5 point summary of each data field.

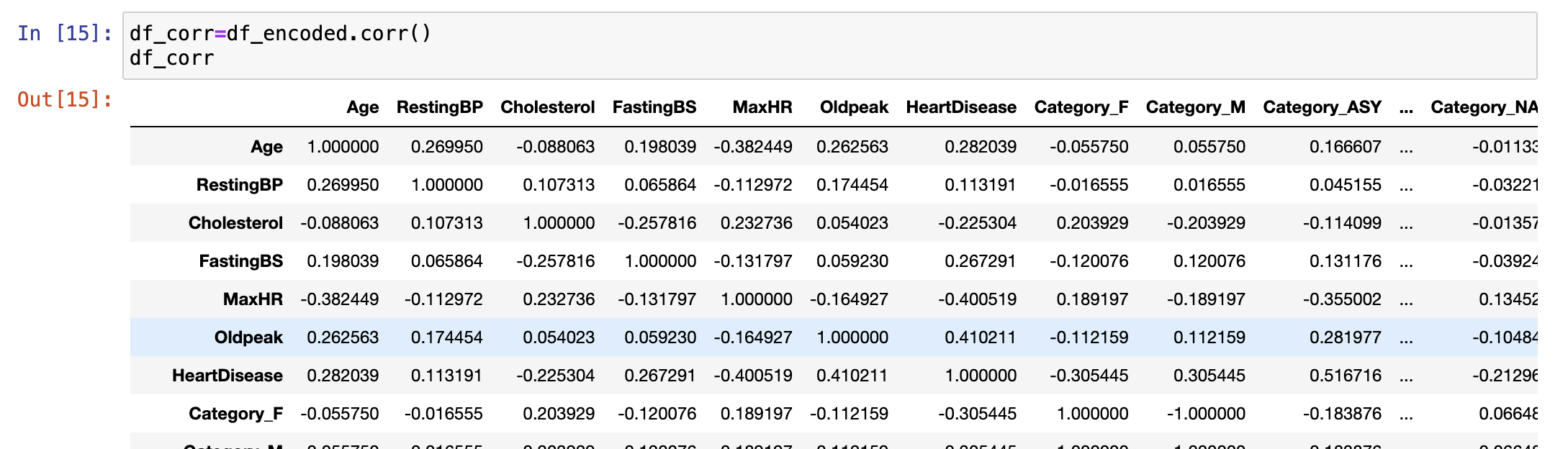
From box plot we have noticed that there are outliers in some fields. So, outlier treatment is done as shown below to improve the models.



Featuring engineering is done for categorical data using one hot encoding method, as shown below.



Correlation for all the data fields are checked as shown below. Similarly, pair plot is created which gives us a overview of the data scatter and inter relation among the variables .



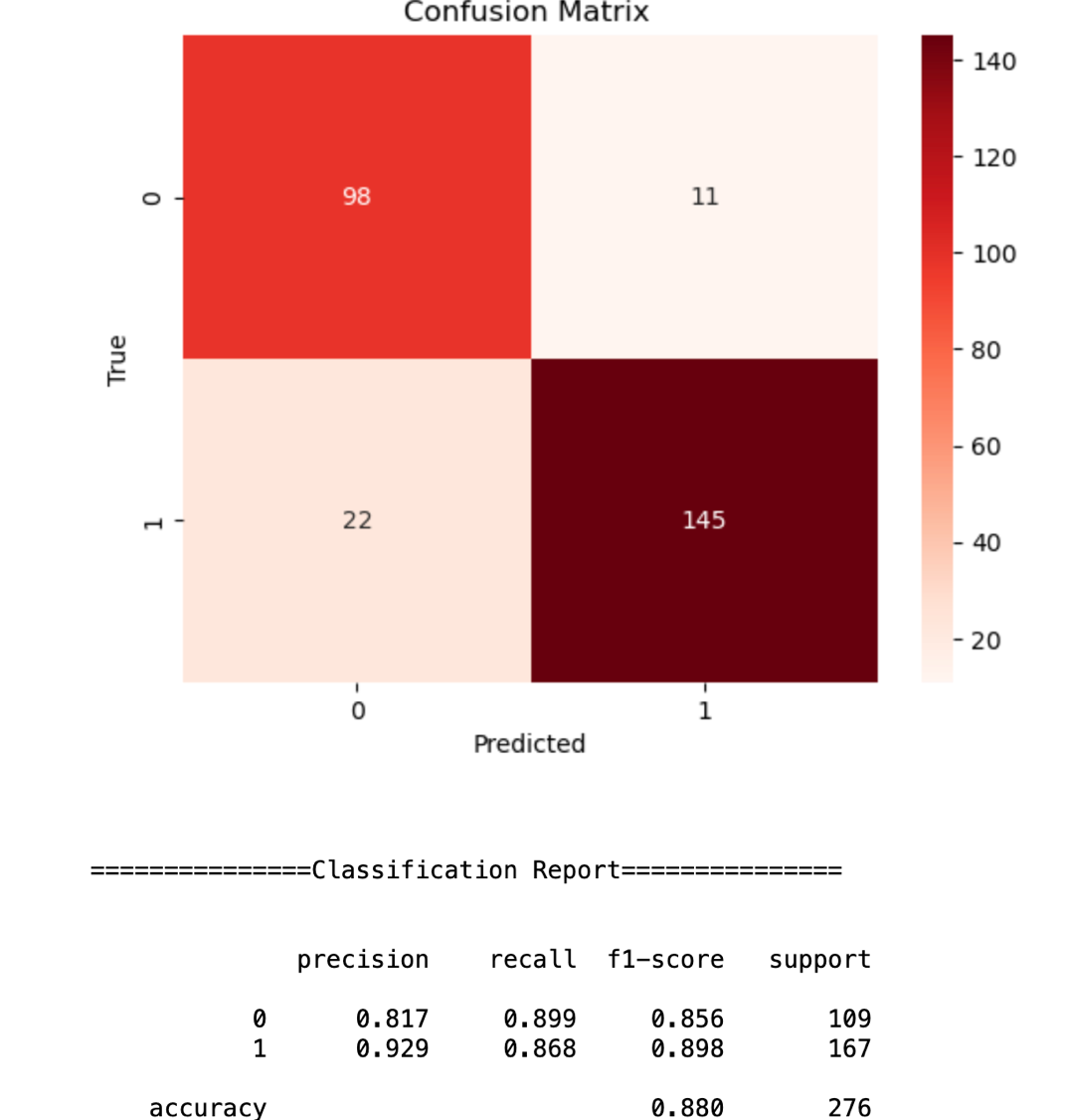
**Create Model and Model Evaluation:**

**Logistic Regression:**

1. Using logistic regression model is built and train data is fitted. Accuracy is checked.
2. Hyperparameter tuning is done using GridSearchCV. Accuracy is checked

Confusion matric and classification report is checked for this model.

A screenshot of a computer code

Description automatically generated

**Decision Tree:**

Using DecisionTreeClassifier mode is created and fitted and accuracy is checked.

To improve accuracy hyper parameter tuning is done by cost\_complexity\_pruning\_path

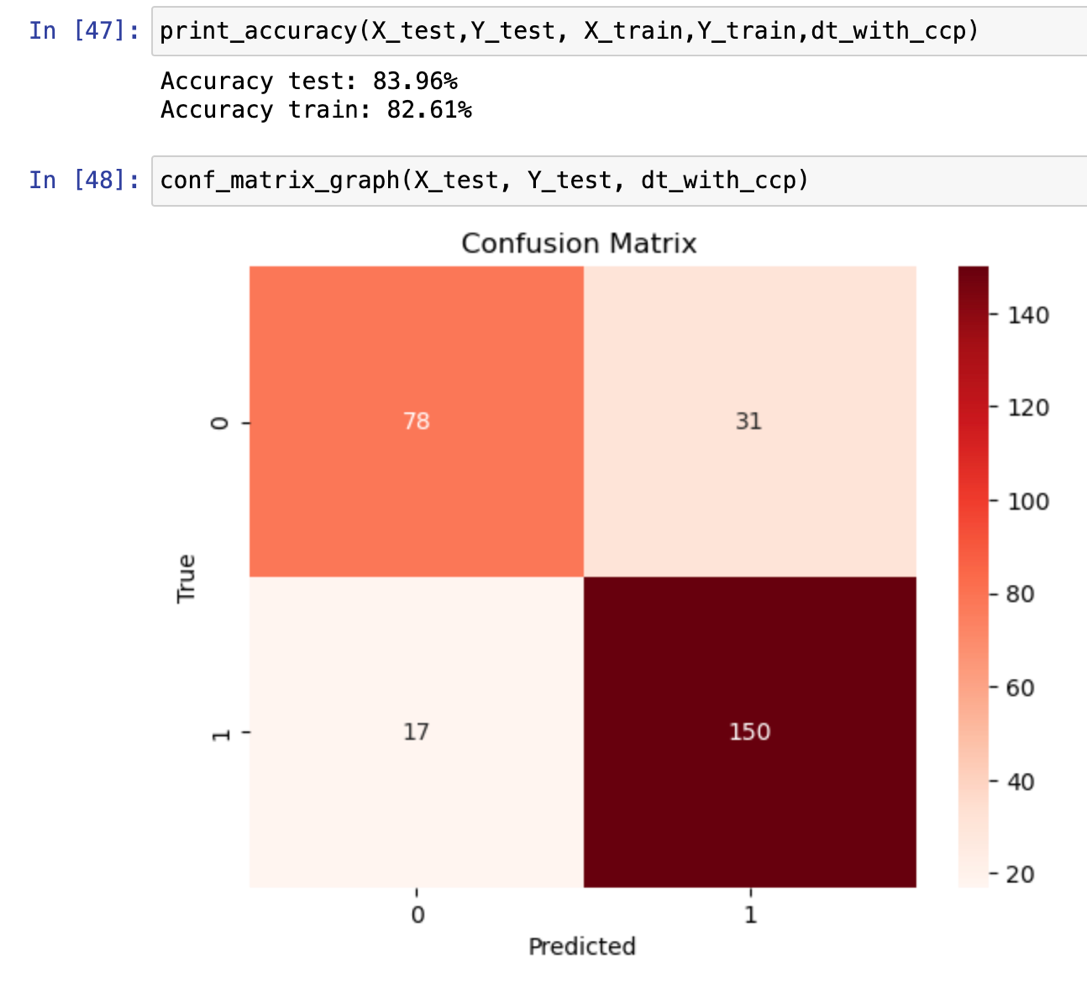
A screenshot of a computer code

Description automatically generated

A screen shot of a graph

Description automatically generated

Accuracy and confusion matrix checked.

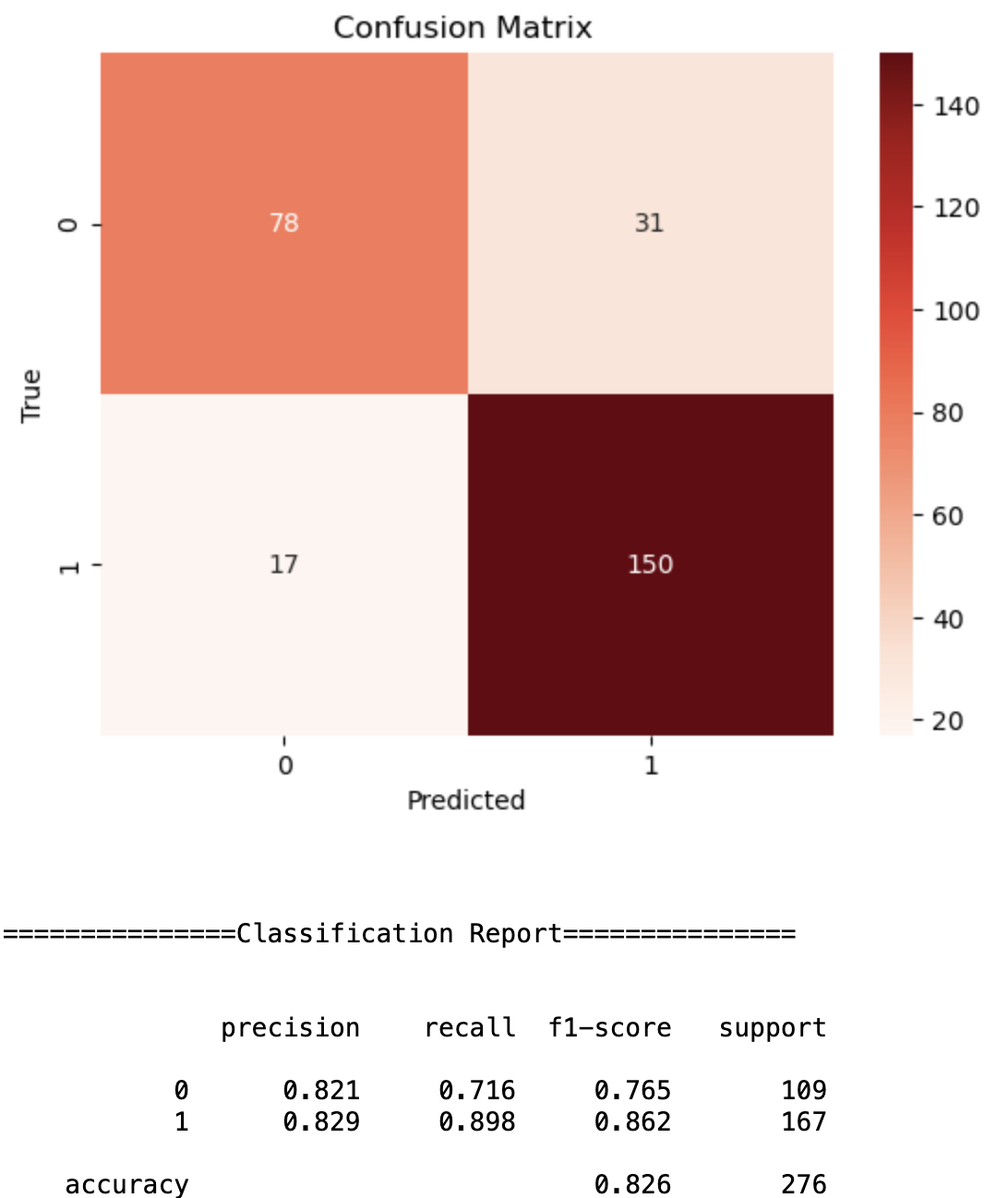


# RandomForest

Using. RandomForestClassifier mode is created and fitted and accuracy is checked.

Using GridSearchCV, K-foldCV and cross validation, RandomizedSearchCv hyper parameter tuning is done.

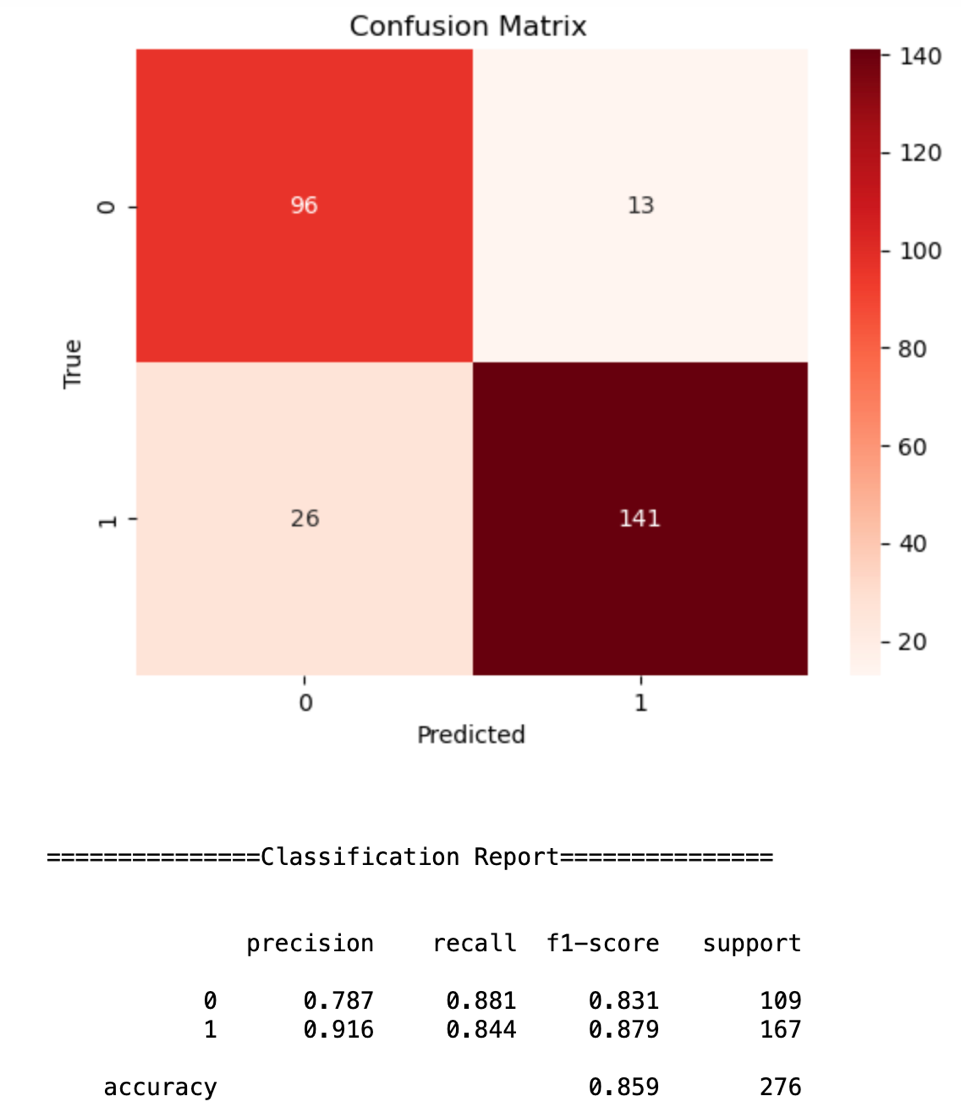
Accuracy and confusion matrix checked.



# KNN

# Using KNeighborsClassifier model is built, fitted and accuracy is checked.

Performing grid search with cross-validation best value of K is found and model is created. Accuracy and confusion matrix is checked for this model.



# SVM

# Using ‘rbf' kernel, model is built, fitted and accuracy is checked.

# 

# XGBoost

# Using XGBClassifier model is build, fitted and accuracy is checked.

# 

# Naive Bayes

# Using GaussianNB model is built, fitted and accuracy and confusion matrix is checked.

# A screenshot of a computer screen Description automatically generated