**Cardiovascular Disease Detection and Analysis using known Machine Learning Algorithms**

|  |  |  |
| --- | --- | --- |
|  | Student Name: Sakib-Ul-Ahsan  ID: 20-42978-1 *Dept Name: CSE* *Institute Name: AIUB* City, Country: Dhaka, Bangladesh email address: sakib.ahsan.35@gmail.com  Student Name: Md. Minhazul Bari Fahim  ID: 20-42176-1 *Dept Name: CSE* *Institute Name: AIUB* City, Country: Dhaka, Bangladesh email address: sakib.ahsan.35@gmail.com |  |
|  |  |  |

***Abstract***:

Cardiovascular diseases (CVD) represents a significant global health concern with high mortality rates. Accurate classification of CVD is crucial because any misidentification can lead to severe outcomes. Early detection and precise characterization of the disease are vital for effective treatment strategies. So, in this research there will be an attempt to analyze predictions make by known models to understand which model is the most accurate in determining CVD. Among all the models Random Forest had the highest accuracy of 72%. Further research and validation should be conducted to validate the proposed approach

# **Introduction**

The prevalence of cardiovascular diseases (CVD) presents a critical challenge to global health, contributing to substantial mortality rates. Ensuring the accurate classification of CVD is of paramount importance, as any erroneous identification can result in grave consequences. Therefore, the focus of this research revolves around the analysis of predictions generated by established models, aiming to discern the model that exhibits the highest accuracy in CVD determination.

This study seeks to address the pressing need for robust classification methodologies in the context of cardiovascular diseases. By assessing the predictive capabilities of various established models, the research aims to shed light on the most accurate model for discerning the presence and characteristics of CVD. Notably, the outcomes of these analyses reveal that among all considered models, the Random Forest model emerges as the leader, boasting an accuracy of 72% in CVD classification.

While this research provides promising insights into model performance, it also underscores the necessity for further investigations to validate and refine the proposed approach. Through rigorous validation and refinement, the research community can pave the way for more effective and reliable strategies in the early detection and characterization of cardiovascular diseases.

# **Motivation of the Project**

The motivation behind this research stems from the need for more efficient and accurate methods of CVD detection. Traditional diagnostic approaches often rely on manual interpretation and subjective judgment, which can be time-consuming and prone to human error. ML algorithms, with their ability to learn patterns and make predictions from large datasets, offer a promising solution to enhance CVD detection and analysis. By leveraging ML algorithms, we can potentially identify novel risk factors, improve risk stratification, and provide personalized treatment plans.

The general aim of this thesis is to contribute to the understanding of cardiovascular disease by developing and applying ML algorithms for CVD detection and analysis. By harnessing the power of ML, we aim to improve the accuracy, efficiency, and effectiveness of CVD diagnosis, risk assessment, and treatment planning. Our research seeks to bridge the gap between traditional diagnostic approaches and advanced computational techniques to enhance patient care and outcomes.

# **Objective of the project**

Some of the objectives of this project is described below:

1. **Comparative Model Evaluation**: This research aims to assess the comparative performance of prominent machine learning models including Random Forest, Decision Tree, KNN, Naïve Bayes and SVM.
2. **Precision of Disease Classification:** The primary goal of this study is to find the model that has the highest precision in categorizing different types of cardiovascular diseases.
3. **Early Detection Potential:** Investigating the early detection capabilities of the identified models is a significant focus of this project.
4. **Feature importance:** Insights into the feature importance hold potential for informing medical practitioners about key indicators of CVD.
5. **Validation & Optimization:** It also places emphasis on the validation and optimization of the chosen models.

# **Methodology**

A review of the pre-existing literature it is evident that current studies focus on the detection and analysis of cardiovascular diseases (CVD). In this study, a system is proposed to contribute to this particular area**.** The proposed system aims to enhance the accuracy of CVD prediction by the proper use of machine learning models and balancing dataset. By considering multiple ML algorithms, we can determine the most effective approach for CVD analysis.

A diagram of a model selection

Description automatically generated

Fig: Flowchart of this study.

## Data Collection

The dataset used in this study was collected from Kaggle, a well-known platform for data science and machine learning resources. The dataset is called “Cardiovascular Disease Dataset”.

## Data processing

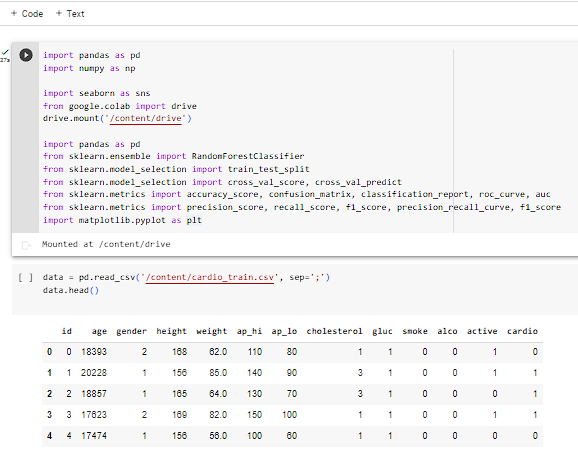
The dataset was searched for missing values at first. The there was an attempt to look for any outliers. But none was found. So, we jumped to the feature selection phase. The attribute named “id” had no relevance to whether or not the person might have CVD or not. So that particular attribute was discarded. Then the age attribute was given in days which would make the calculations for the ML models slower and more complex. So, it was converted into years by dividing with 365.

## Dataset description

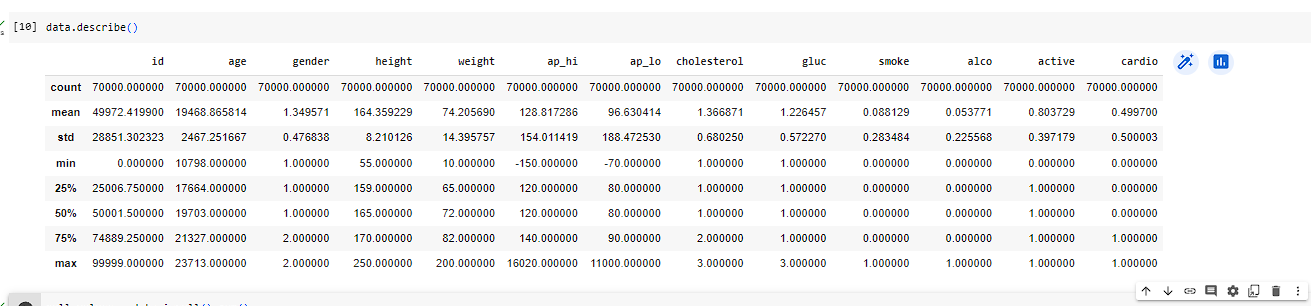
After data pre-processing the dataset comprised of 12 attributes with the target attribute too. The attributes include age, gender, height, weight, systolic blood pressure, diastolic blood pressure, glucose, smoking status, alcohol, active status, cardio. The last attribute “cardio” is the target attribute which has two values(0 and 1). 0 refers to the person not having CVD and 1 says the opposite. There are 70,000 instances and 13 columns.

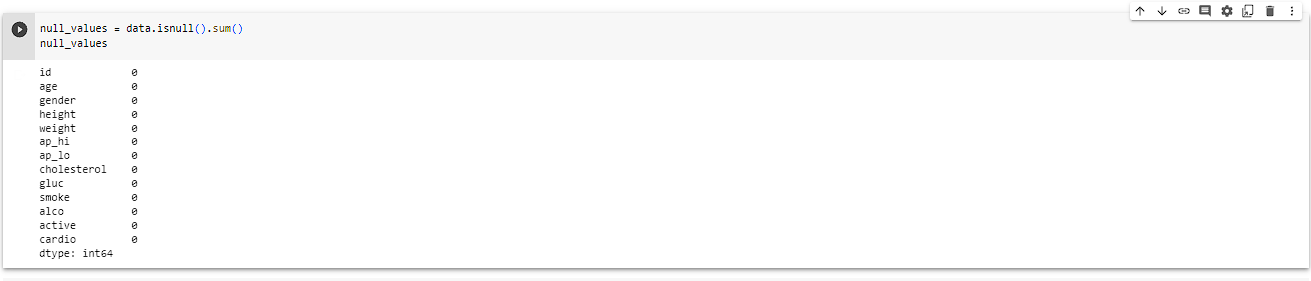
## **Machine Learning model development and evaluation**

1. Importing all the libraries and then loading the dataset and printing the headers.

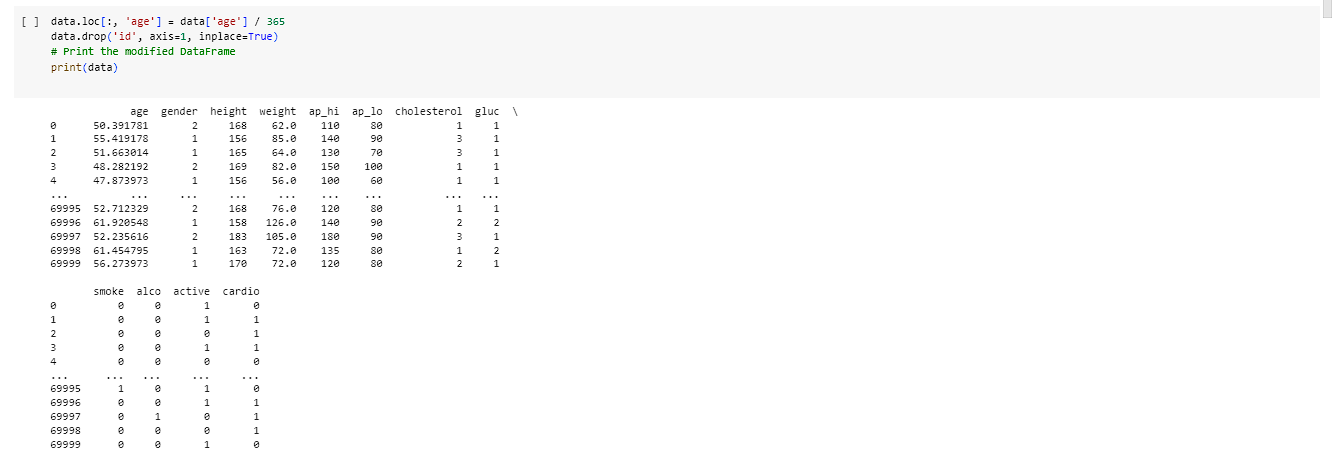


1. Dataset Description and looking for missing values

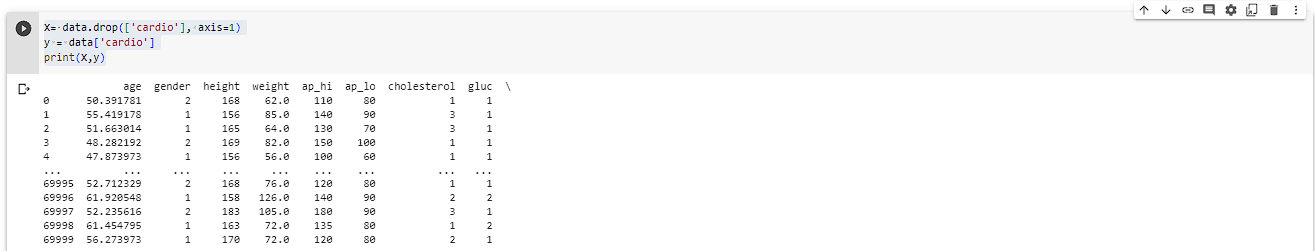




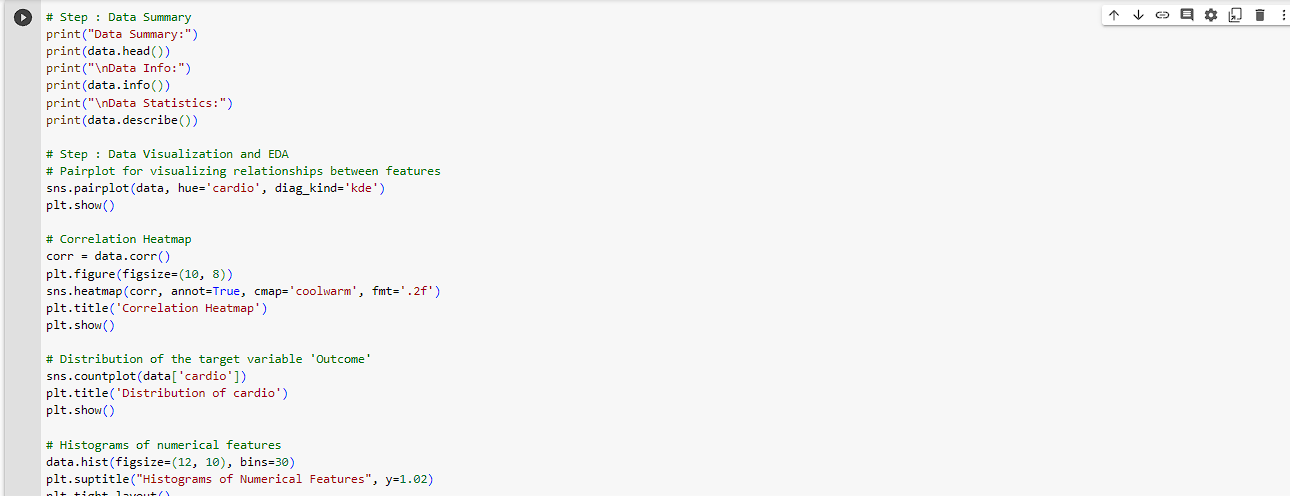
1. Data Pre-processing: Removing the “id” attribute and converting age to years(was given in days)

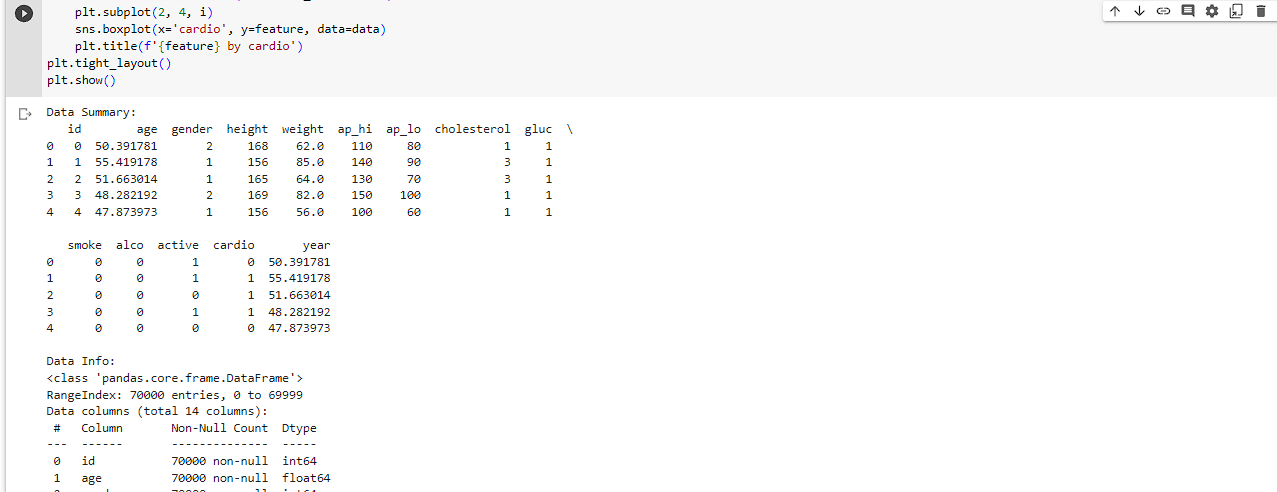


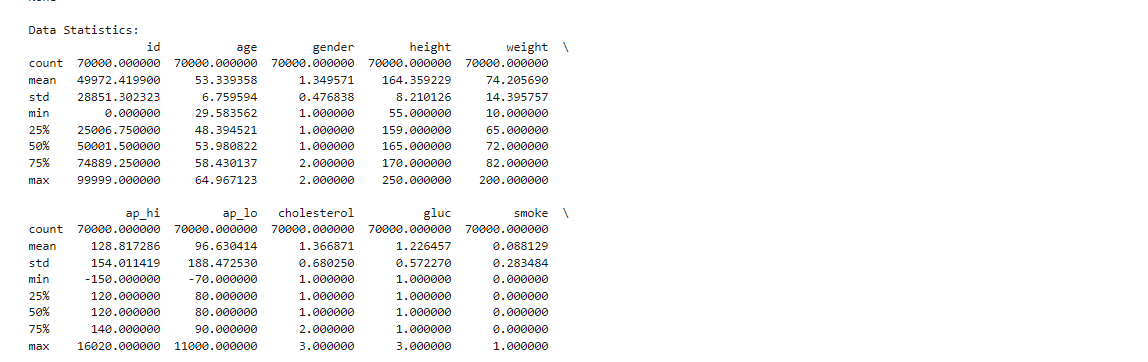
1. Separating the features and keeping it in X and the target variable in y

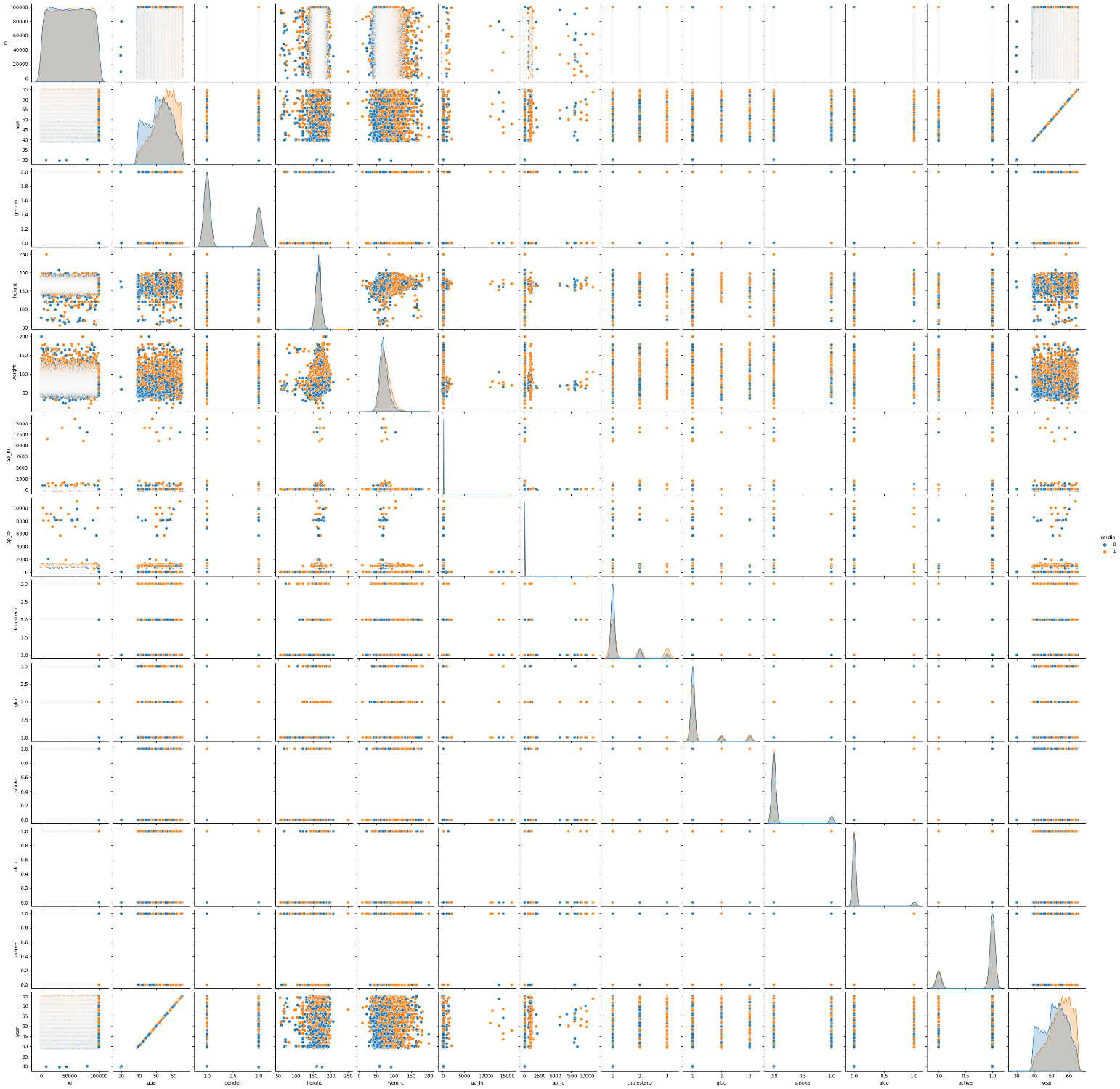


1. Creating summary of the data:

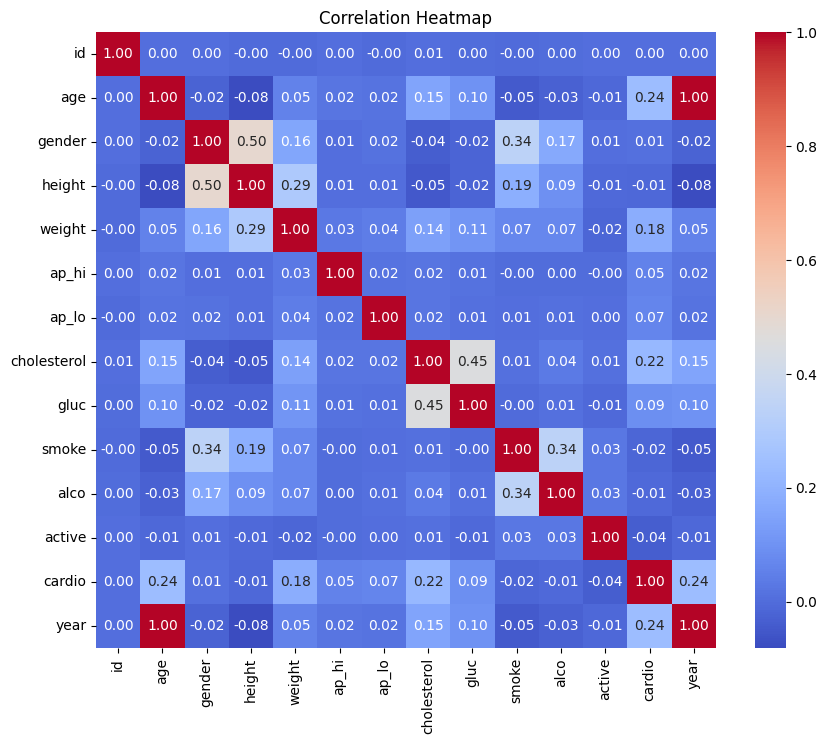


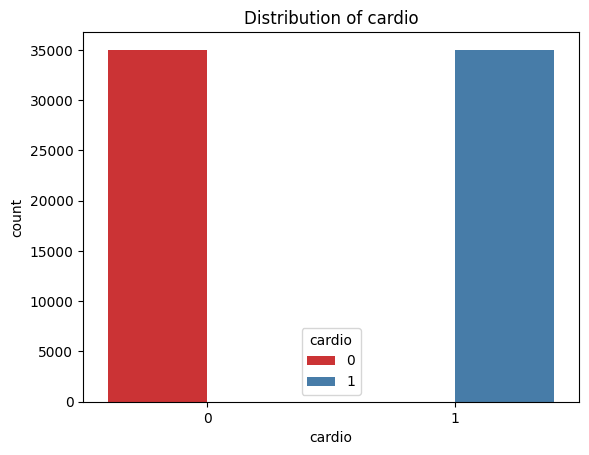


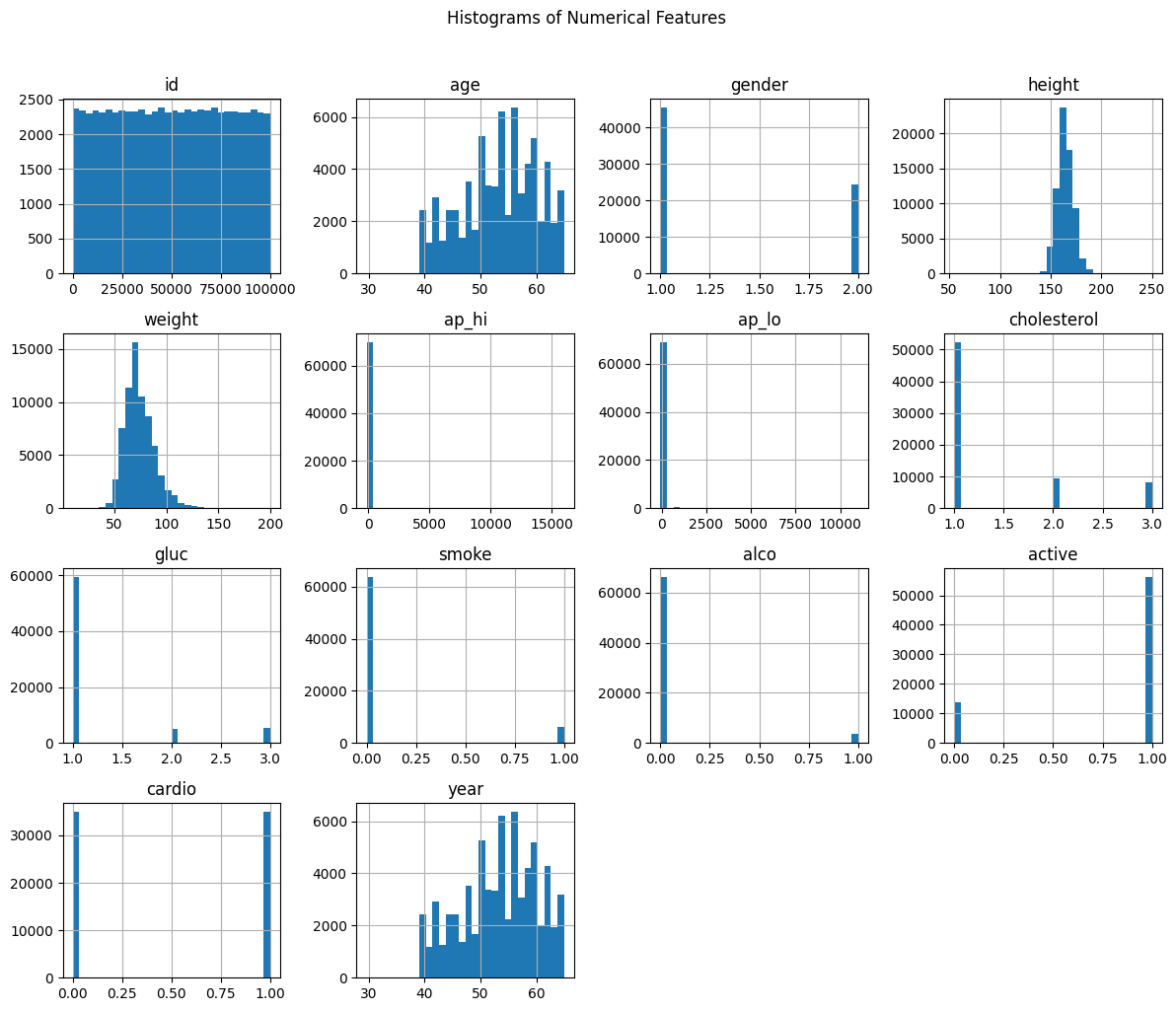


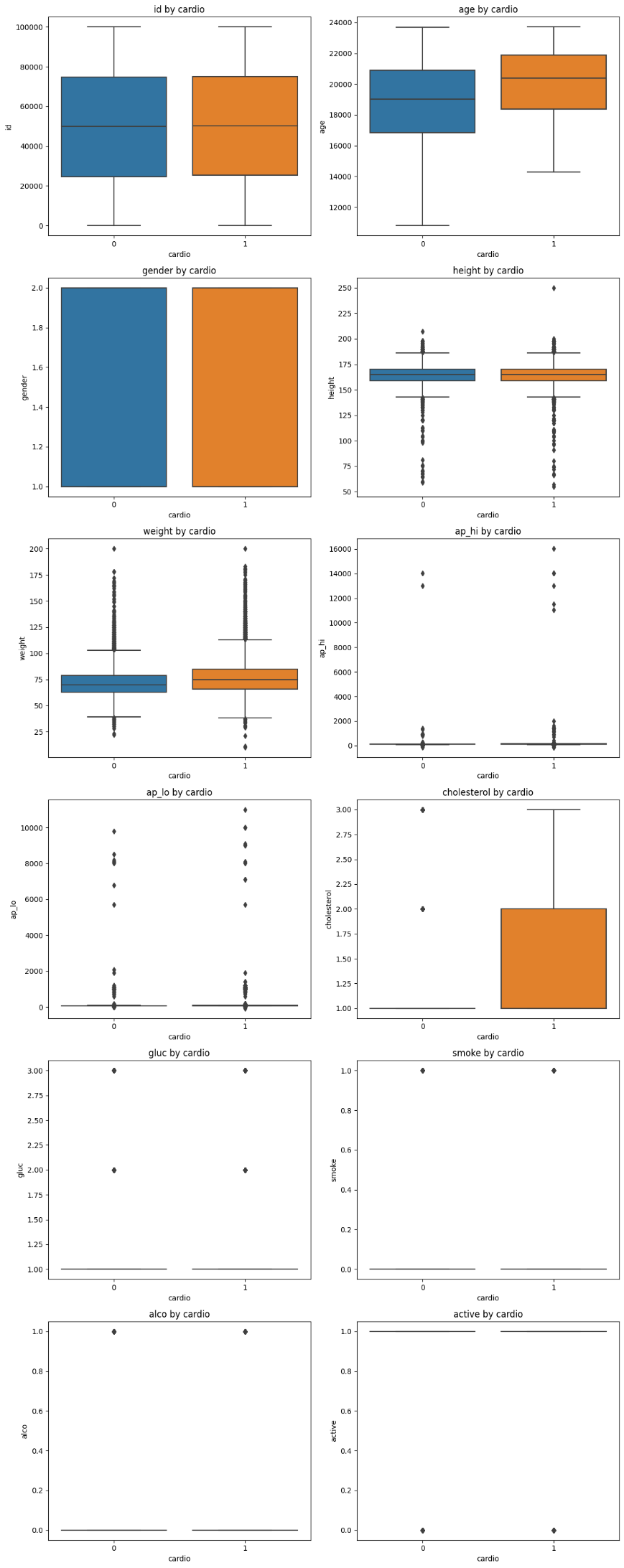


Pairplot

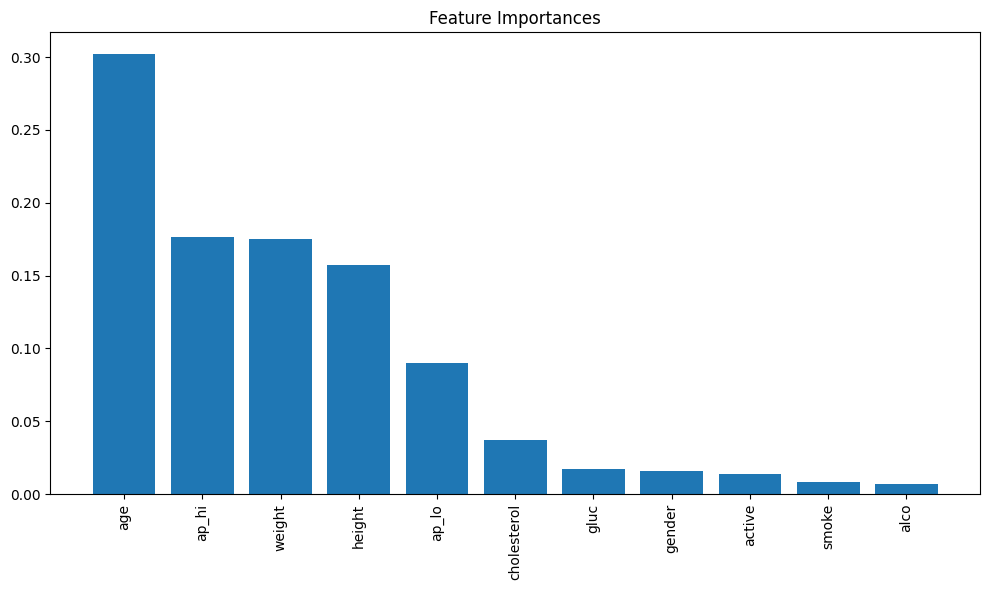




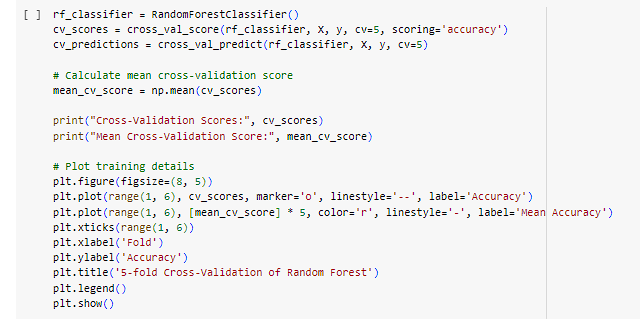


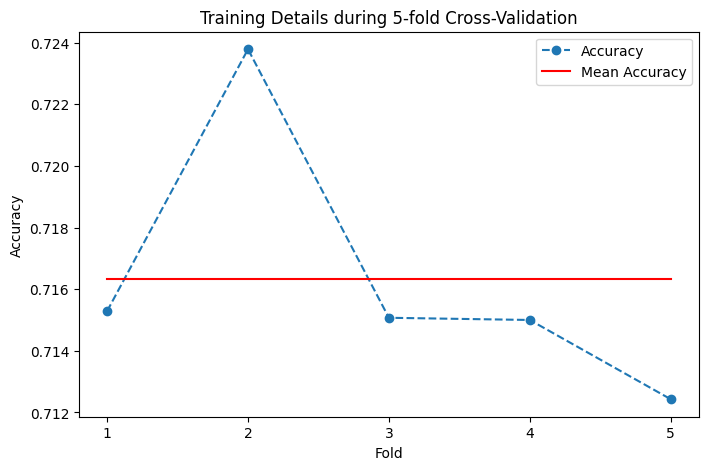


1. **Feature importance**

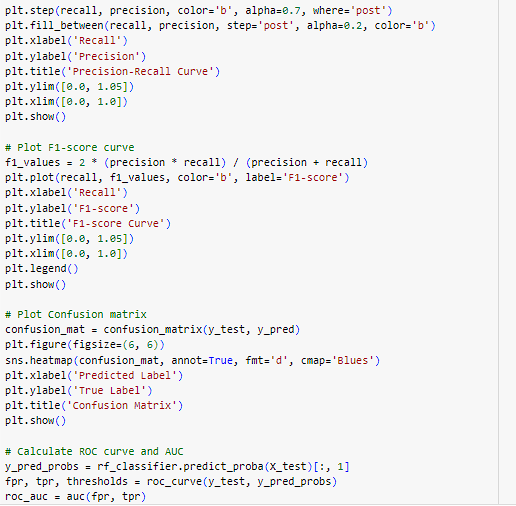


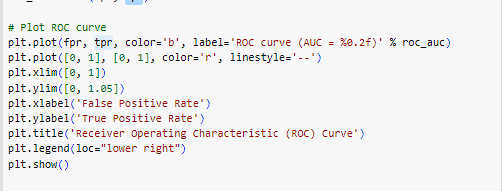
1. Accuracies of different ML models
2. Random Forest:

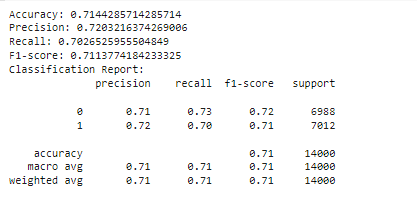


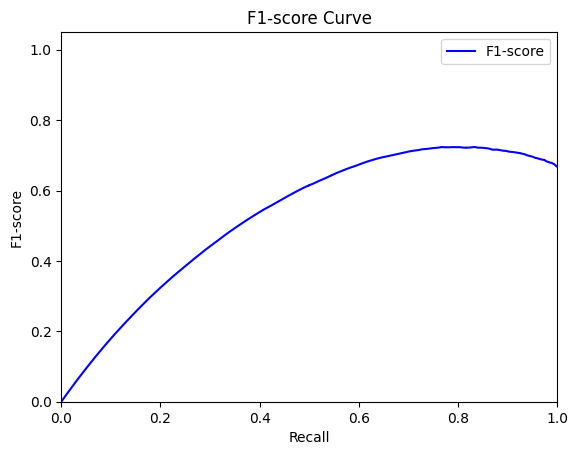
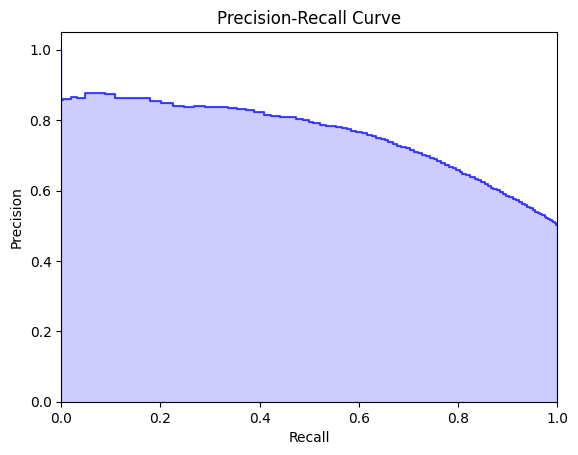


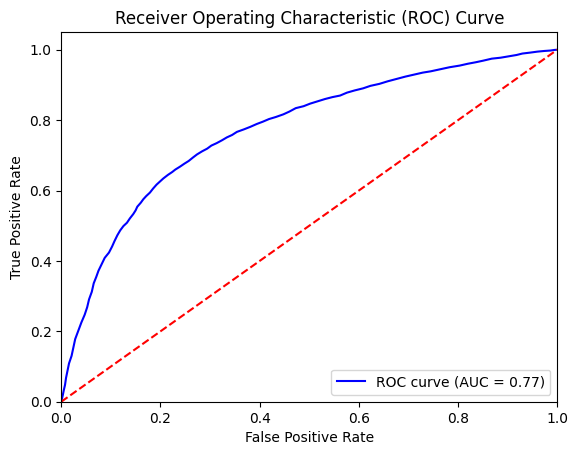
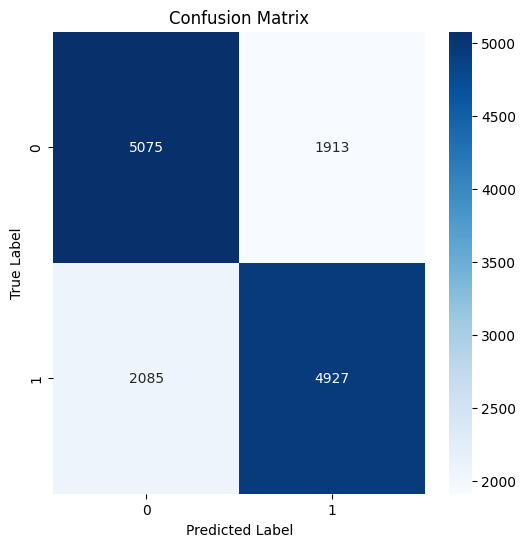




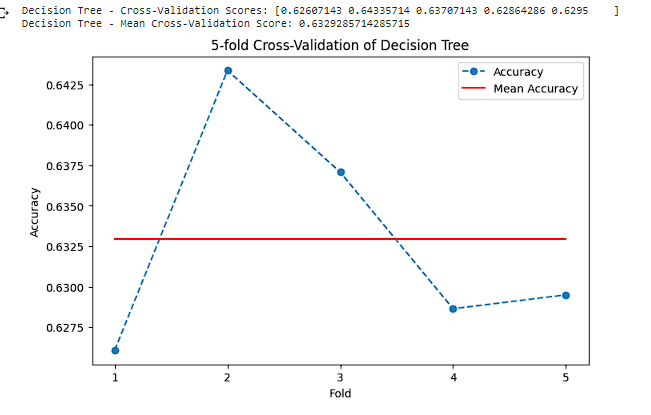


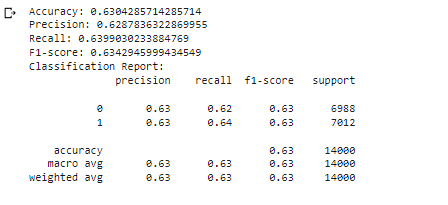


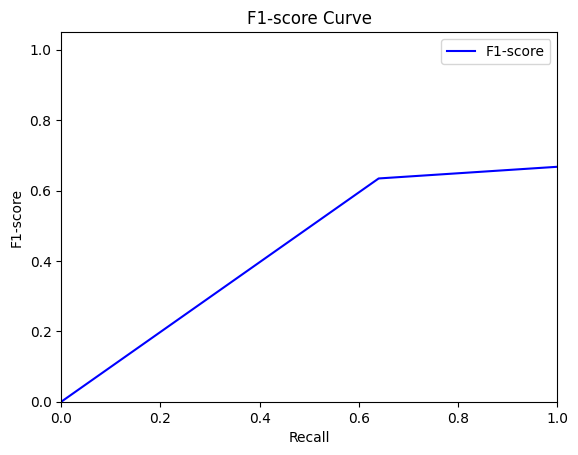
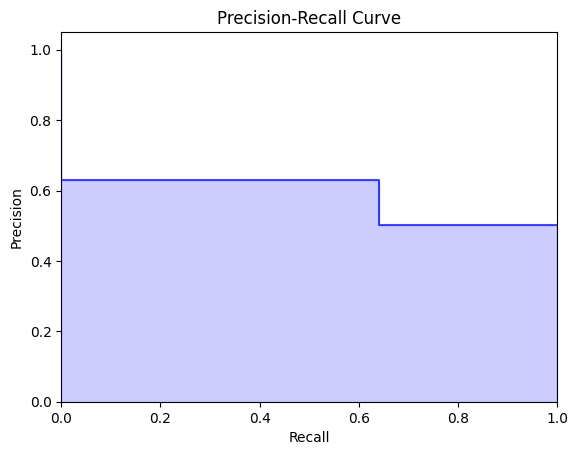


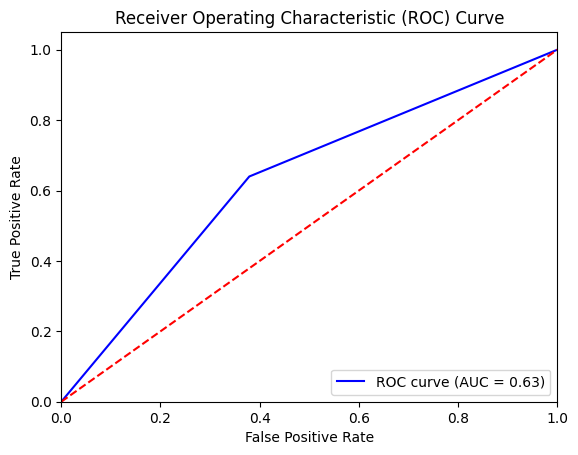
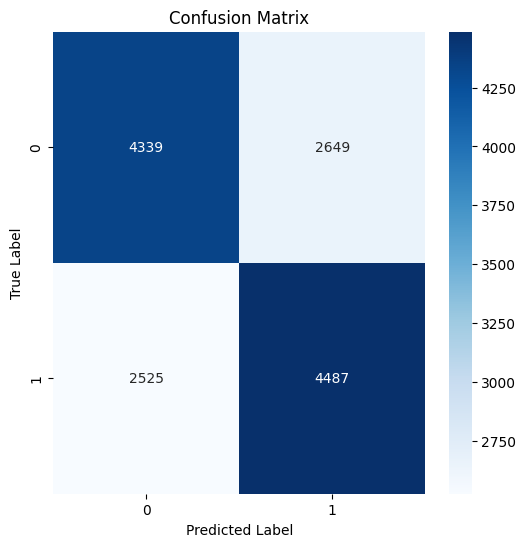


1. **Decision Tree**

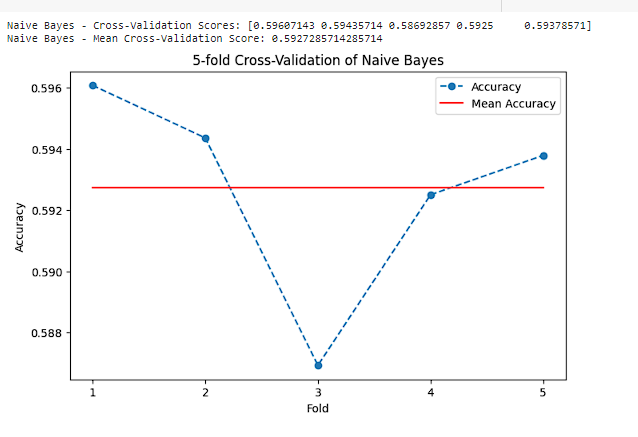


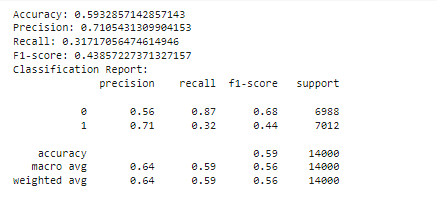


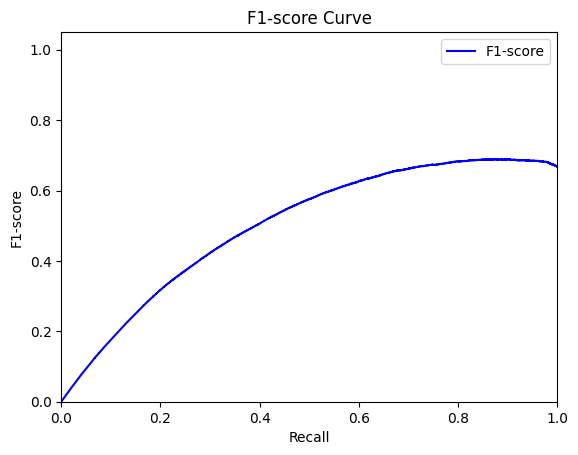
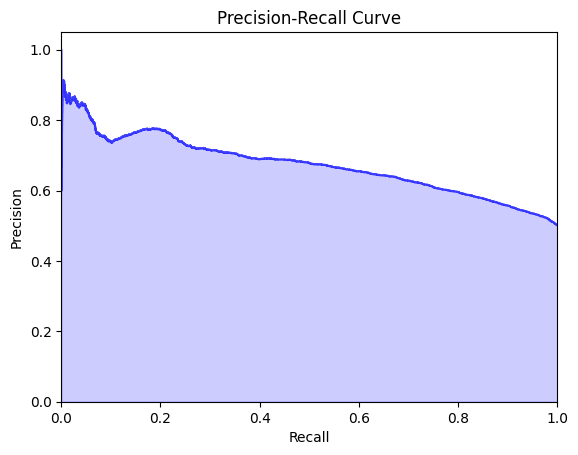


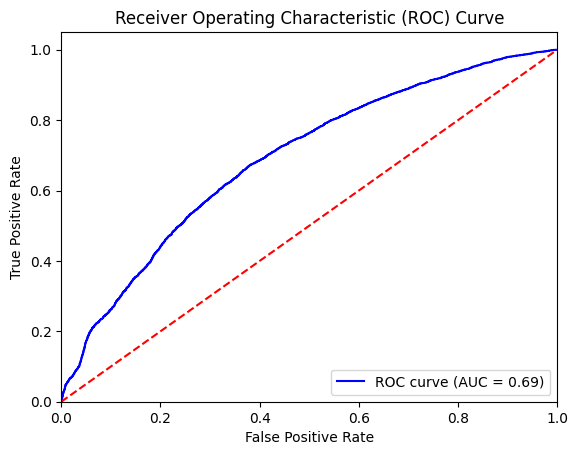
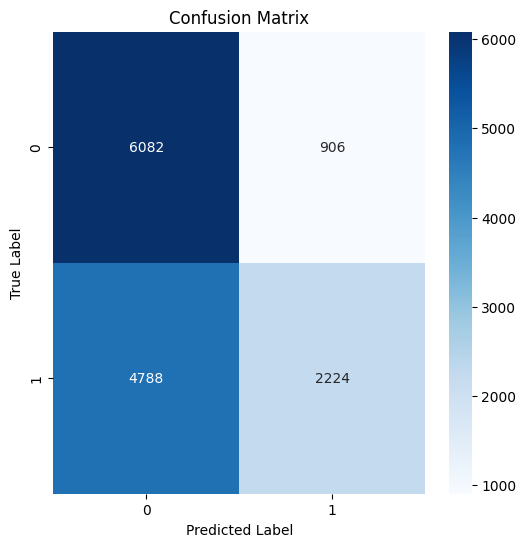


1. Naïve Bayes:

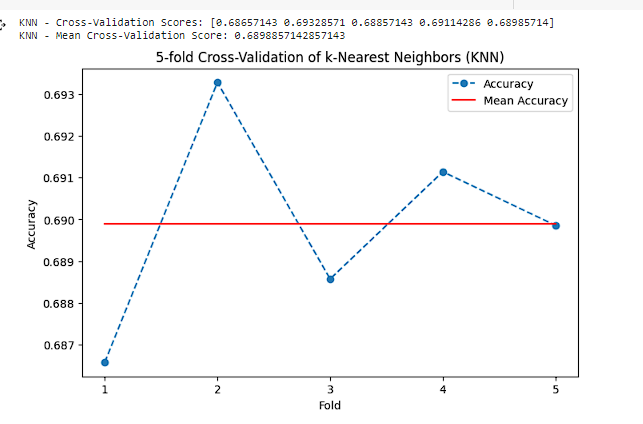


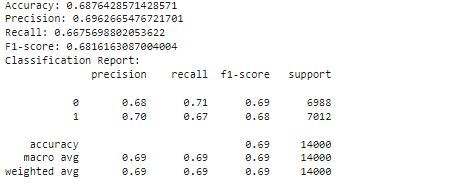


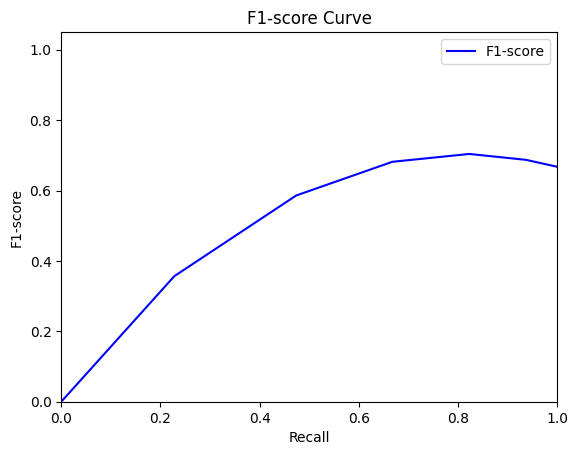
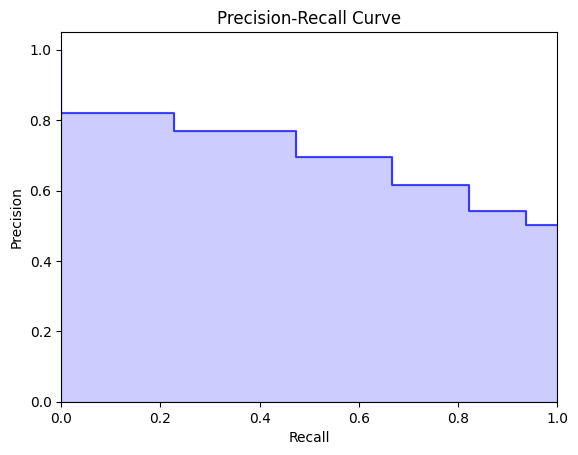
\

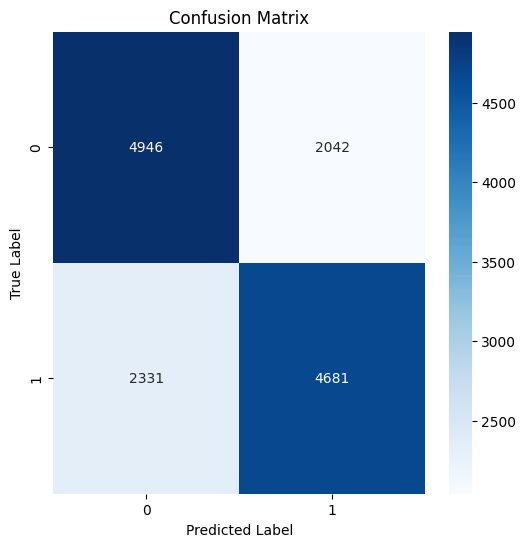
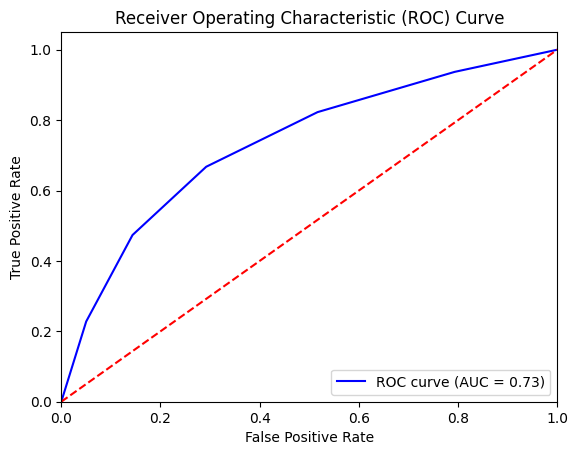


**KNN:**









# Results

