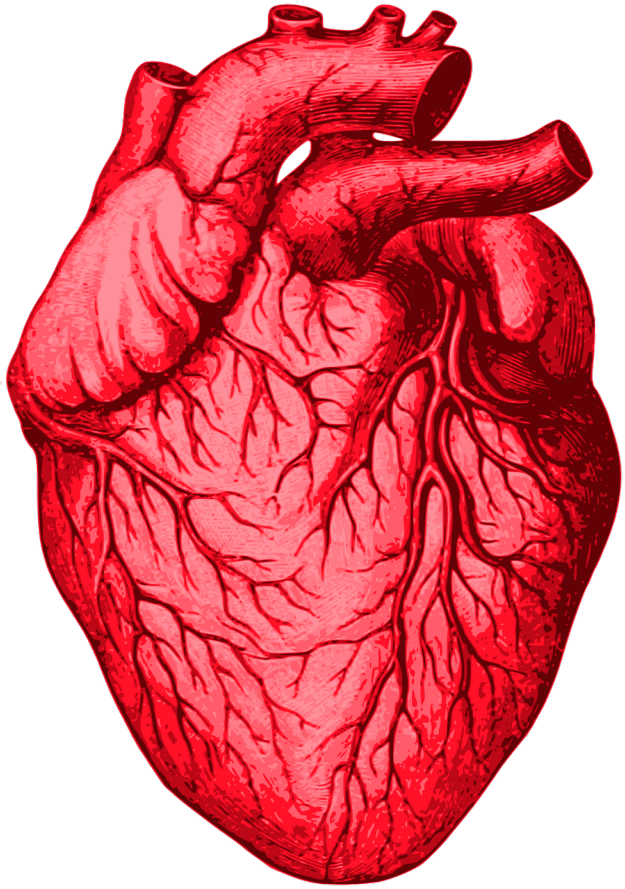
Predicting Coronary Artery disease with Random Forest, Multi-layer Perceptron and Logistic Regression Models

Raiyan Ferdous, 21101127, Faiaz Zahin, 21101090



Group- 05

I. Introduction

HEART problems are quite widespread these days. The term ”heart diseases” mostly refers to a variety of cardiac disorders. Heart disorders come in a variety of forms. Conditions involving the heart or blood vessels are collectively referred to as cardiovascular disease. It claims over 17.9 million lives annually, making it one of the world's greatest causes of death (WHO 2022). The heart could beat abnormally slowly, excessively fast, or too erratically. The following are some signs of cardiovascular disease: shortness of breath, fast or sluggish heartbeat, dizziness, fainting, chest pain or discomfort, and shortness of breath. Numerous factors, including smoking, eating unhealthily, and not exercising, contribute to it. Maintaining a healthy lifestyle can reduce your risk of CVDs AI technologies have been applied in cardiovascular medicine including precision medicine, clinical prediction, cardiac imaging analysis and intelligent robots. Research from Dawes TJW (Health IT Analytics, 2020)

Cardiovascular medicine has grown from the application of AI technologies, which include intelligent robotics, clinical prediction, precision medicine, and cardiac imaging analysis. According to research by Dawes TJW, artificial intelligence may be able to forecast when people with heart disease may pass away. AI software was used in their study to record 256 heart disease patients' blood test and cardiac magnetic resonance imaging (MRI) scan data. 93% of the time, the AI-assisted screening tool correctly identified those who were at risk of left ventricular dysfunction. In comparison, a mammography is accurate 85% of the time. (PMC, 2022)

With the knowledge provided above, we can comprehend the significance of artificial intelligence in the early detection of cardiovascular illnesses.

Our study will use specific data, such as height, weight, blood pressure, cholesterol, glucose level, smoking habit, and blood pressure, to accurately determine whether or not a patient has heart disease. Our dataset has a 50/50 ratio of damaged to healthy hearts, which allows us to get an accuracy score that is high enough.

II. Methodology

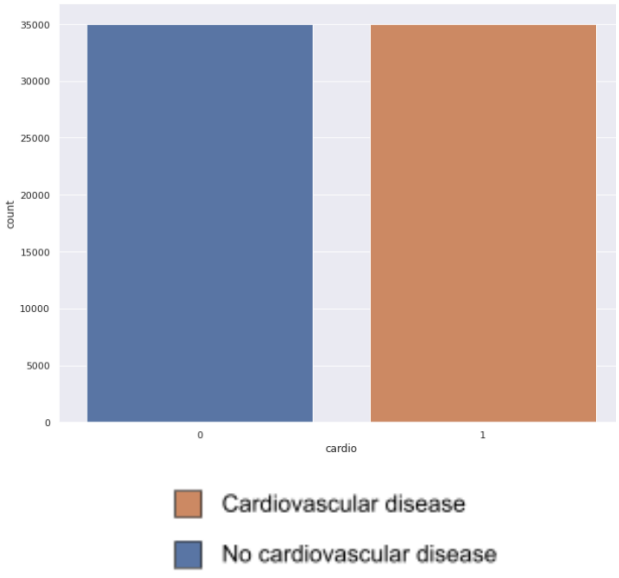
A. DATASET

Our dataset came from Kaggle. Data scientist Svetlana Ulianova from Ontario, Canada, created the dataset. With 13 features, including a unique ID, the dataset has 70,000 individual data points. These qualities are:

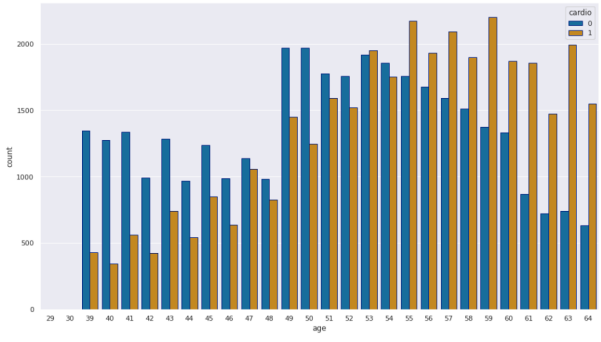
Feature	Feature type	Name in CSV file	Datatype
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective Feature	smoke	binary
Alcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular disease	Target Variable	cardio	binary

B. DATA VISUALIZATION

The data's cardio ratio is unquestionably balanced, with a nearly equal distribution of damaged and healthy hearts.



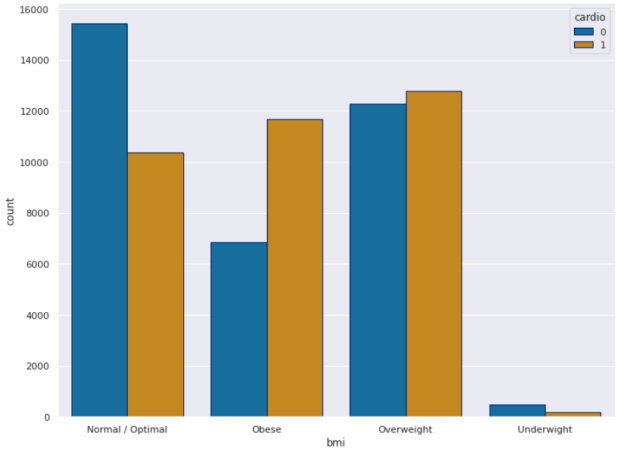
A link between an attribute and the target attribute can be seen directly in specific data attributes. like ”age.”



In According to the age factor, the orange hue in this bar chart represents the sick heart and the blue color represents the healthy heart. This suggests that the likelihood of developing CVDs increases significantly with age in comparison to earlier ages.

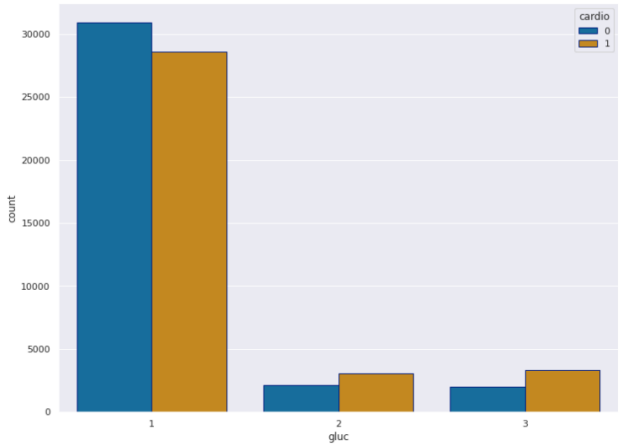
Another important factor in heart disease is BMI. The equation of BMI in equation: 1.

$$BMI = \frac{weight(kg)}{height^2(m^2)} \tag{1}$$



Although the BMI property is absent from this dataset, we can still easily find the BMI because we have the properties of height and weight. This gives us,

Fig. 1. Glucose



Normal / Optimal, Obese, Overweight, and Underweight are listed from left to right. The key finding in this instance is that an individual's risk of developing CVDs is increased if they fall into the obese or overweight category. Moreover, elevated cholesterol and glucose levels raise the risk of CVDs.

Fig. 2. Cholestorel

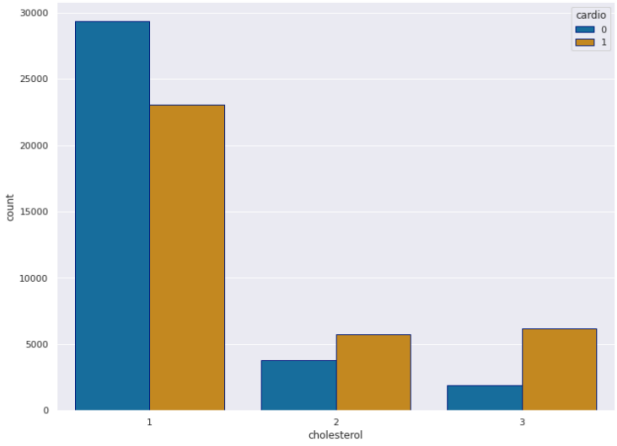


Fig. 3. Summary

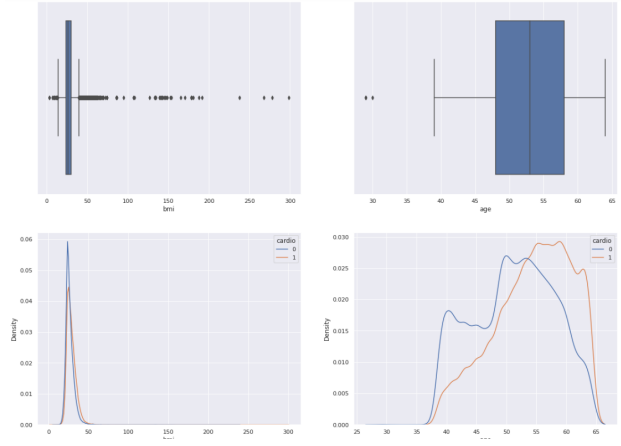
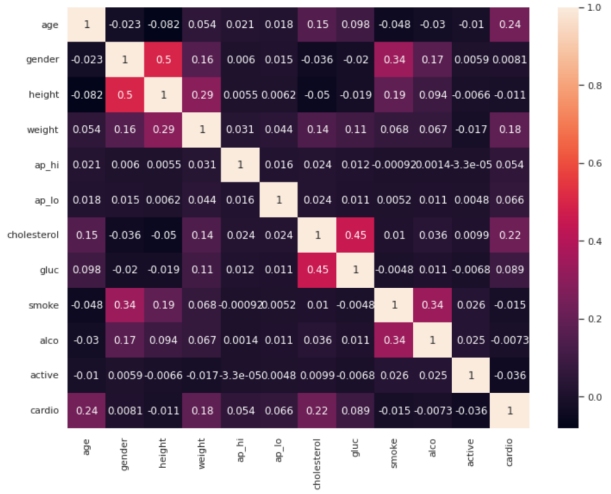


Fig. 4. The Dataset's correlation heatmap:



C. DATA PROCESSING TECHNIQUE

1) Deleting ID: An attribute called "id" is found in the dataset, but it is not required. Therefore, we may remove it to increase accuracy.

```
del data['id']
```

2) Converting age from day to year: Working with the unit days of the "age" property is not very convenient. So, by dividing by 365, we can turn it to year.

```
data['age']// = 365
```

This preprocessing can lead to slightly greater accuracy.

D. Train-Test split:

The data was divided into two sections for training and testing. Eighty percent of the data will be set aside for model training, and twenty percent will be set aside for model testing.

E. Used Models

1) Random Forest: The Random Forest algorithm is a nonlinear model that is produced by mixing the output of several decision trees, each one of which formed using a portion of the data. A decision tree comprises of two stages:

- It divides the predictor space into discrete rectangular sections (the split of these areas minimizes overall GINI index or entropy for tree classification and the RSS for regression tree).
- It calculates the mode, or mean, of the sample section's results in each region and utilizes that number to forecast new data.

  Using X1 to X90, that were chosen based on previous research, we compared random forest's predictive power to other non-linear models. We tried a number of designs using 500, 1000, 2000, and 3000 trees. The node size was set to the default, and the total number of variables evaluated at every split (mtry) was evaluated at five equally-spaced variables ranging from 2 to 90, as advised by Kuhn & Johnson. A different test set was utilized to assess the model's correctness, and cross-validation results were used to improve it.

2) Logistic Regression: The model chosen for this study is the logistic regression model, a well-liked machine learning method that is frequently used in real manufacturing settings in domains such as data mining, automated illness detection, and economic prediction. For instance, the study described the heart disease risk factors in depth and offered a propensity score for the condition. Since there are only two kinds of output, every one of which goes to a single category, and because logistic regression can estimate the chance that each classifying event will occur, it is a common method for classification, particularly in two-category issues. Below is an example of a logistic regression model:

$$prob(Y = 1) = \frac{e^x}{1 + e^x} \tag{2}$$

Where Y means to binary dependent variable (Y is equal to 1 if event happens; Y=0 otherwise), e stands for the foundation of natural logarithms and Z means:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{3}$$

for p predictors (j=1,2,3,. ,p)

3) Multi-layer Perceptron: In a multi-layer perceptron, the input layer neurons' sole and main job is to divide the input signal (xi) among the neurons that are in the hidden layer. Following their weighting with the intensities of the corresponding connections wji from the input layer, each neuron j in the hidden layer adds up the input signals xi and calculates its output yj as a function f of the sum, as follows:

$$y_i = f(\sum w_{ij} x_i) \tag{4}$$

Currently, f might be a simple threshold function, such as a sigmoid, or a hyperbolic tangent function. The output layer neurons choose their output according to the same rules. An overview of the steps involved in the operation of a Multi-Layer Perceptron neural networks is given below: 1) Data input is received by the input layer, which processes it to produce an expected output. 2) The anticipated outcome is subtracted from the actual result to find the error value. 3) The network then modifies the weights using a Back-Propagation approach. When changing weights, one starts with the amount of weight between the output layer nodes at the end of the layer that is hidden and moves backwards through the network. 5) The sending procedure resumes when backpropagation is finished. 6)Until the discrepancy between the expected and actual output is as little as possible, the process is repeated.

III. RESULTS

A. Train- Test Accuracy Score

1) **Random forest:**
   - Accuracy of the train data is: **98%**
   - Accuracy of the test data is: **71%**
2) **Logistic Regression:**
   - Accuracy of the train data is: **72%**

- Accuracy of the test data is: **72%**

3) **Multi-layer Perceptron:**

- Accuracy of the train data is: **72%**
- Accuracy of the test data is: **72.6%**

B. Confusion Matrix

Fig. 5. Random Forest

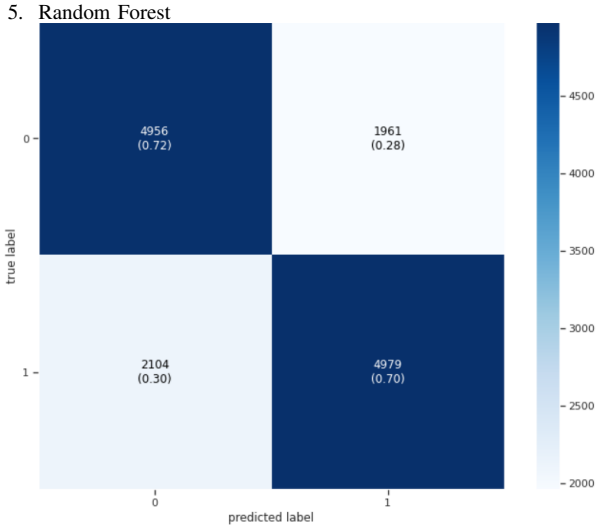


Fig. 6. Logistic Regression

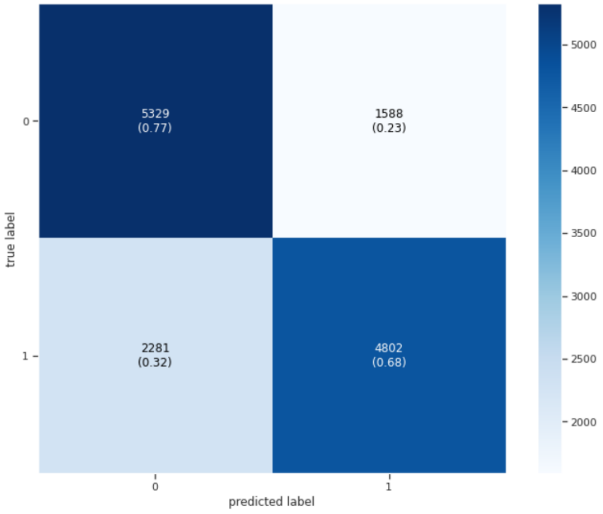
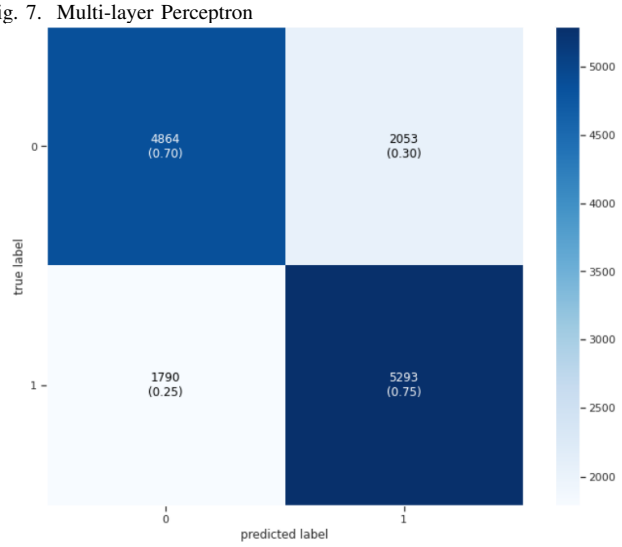


Fig. 7. Multi-layer Perceptron



C. Classification Report

Algorithm	Accuracy	Precision	Recall	F1 Score
Random Forest	0.71	0.71	0.71	0.71
Logistic Regression	0.73	0.72	0.72	0.72
Multi-layer Preceptron	0.73	0.73	0.73	0.73

IV. Conclusion

Based on the aforementioned findings, we may conclude that Random Forest Classification, with its 98% accuracy score, is the best model for training the dataset. However, with a precision of around 73%, the Multi-layer Perceptron module would be the ideal choice for testing.

V. Reference

[5.1]Scikit learn, https://scikit-learn.org/

[5.2]Random Forest Implementation in CVDs, Quantifying Health (2020).

[5.3]Logistic Regression Models in Predicting Heart Disease, January 2021, Journal of, Physics Conference Series 1769(1):012024, DOI:10.1088/1742-6596/1769/1/012024, LicenseCC BY 3.0

[5.4]Cardiovascular Disease Prediction Using KNN Algorithm, Cibhi Baskar, Data Scientist.