CSE 422
Artificial Intelligence

BRAC
UNIVERSITY

Inspiring Excellence

“Cardiovascular disease prediction using Random Forest, K-Nearest Neighbor, Support Vector Machine, And Multi-Layer Perceptron”

CSE 422 - Section 01

Group - 13

Mehrab Hasan- 20301445
Yousuf Abdullah Harun- 20301139

Table of Contents

1 INTRODUCTION

Nowadays, heart diseases are very common. Heart diseases mainly describe a range of conditions that affect the heart. There are different types of heart diseases. Cardiovascular disease is a general term for conditions affecting the heart or blood vessels. It is one of the leading causes of death globally, taking almost 17.9 million lives each year (WHO 2022). The heart may beat too quickly, too slowly or irregularly. Cardiovascular disease symptoms can include: Chest pain or discomfort, dizziness, fainting ,fluttering in the chest, shortness of breath, racing heartbeat, slow heartbeat. It is caused due to many reasons like smoking, eating unhealthy food, and no exercise. Risk of CVDs can be lowered by leading a healthy lifestyle. If a person currently has CVDs, maintaining health might lessen the likelihood that it will worsen.

The faster a person can be diagnosed with CVDs, the better it is. A quicker diagnosis can lead to quicker access to care and quicker recovery, reducing rates of mortality, hospitalizations, and heart failure. Here, AI can come to help. An AI agent can read the symptoms and conclude to a decision based on the trained data. Before creating the AI tool, cardiologists had difficulty recognizing physical changes in the heart and defining whether they were related to disease or simply just to aging. (Health IT Analytics, 2020)

AI technologies have been applied in cardiovascular medicine including precision medicine, clinical prediction, cardiac imaging analysis and intelligent robots. Research from Dawes TJW suggests that AI can predict possible time periods of death for heart disease patients. In their research, AI software recorded the results of cardiac magnetic resonance imaging (MRI) scans and blood tests of 256 heart disease patients. The AI-assisted screening tool identified people at risk of left ventricular dysfunction 93% of the time. To put that in perspective, a mammogram is accurate 85% of the time. (PMC, 2022)

With the above information we can understand the importance of Artificial Intelligence in predicting cardiovascular diseases.

Our project will work with some values, like, height, weight, glucose level, smoking habit, blood pressure, cholesterol and from those values it will predict if the patient has heart disease or not with certain accuracy. Our dataset contains a 50-50 ratio of healthy heart and diseased heart from where we can lead to a sufficient accuracy score.

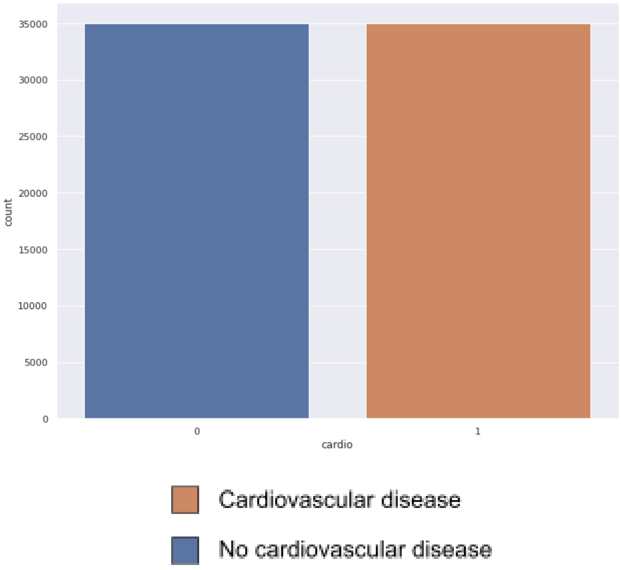
METHODOLOGY

2.1. Dataset: The dataset we used is from kaggle. The dataset was prepared by Svetlana Ulianova, a data scientist from Ontario, Canada. The dataset has 70,000 individual data with 13 attributes including unique ID. The attributes are as follows:

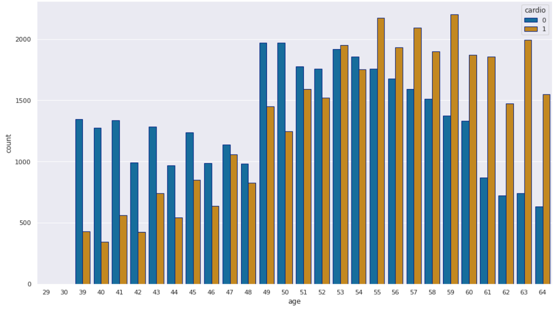
Feature	Feature type	Name in CSV file	Datatype
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination 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

2.2. Data visualization:

The cardio ratio of the data is certainly balanced, meaning it has almost a 50-50 ratio of healthy heart and diseased heart.



There are certain data attributes which directly show the relationship between that attribute and target attribute. Such as ‘age’.

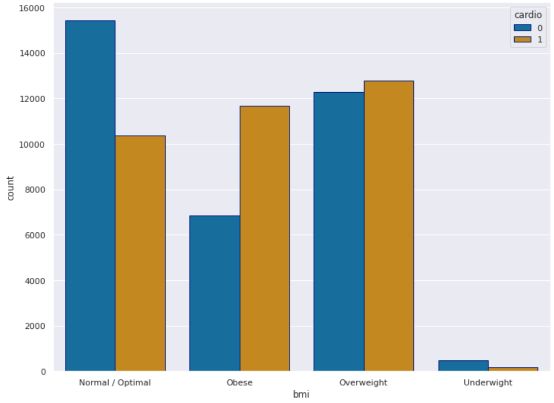


In this bar chart, the color orange shows the diseased and blue shows the healthy heart in accordance to the age factor. Which indicates that, with increasing age, the chances of having CVDs are much higher than the younger age.

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

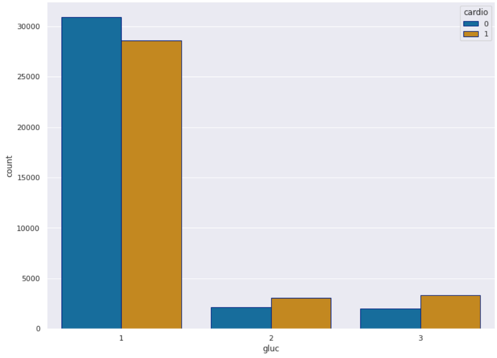
$$BMI = \frac{\text{weight (kg)}}{\text{height}^{\,2}(\,\text{m}^2)}$$

In this dataset we do not have the attribute BMI, but as we have height and weight as attributes, we can find the BMI easily. From this we get,

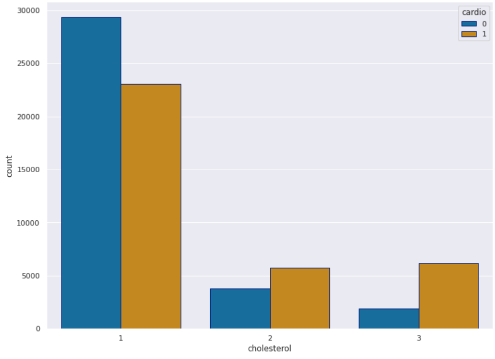


From left to right: Normal / Optimal, Obese, Overwieght, and Underweight. Here the main observation is, if a person is from obese or overweight catefory, that person has a high chance of having CVDs. Also, increasing glucose level and cholesterol level increases the chances of having CVDs.

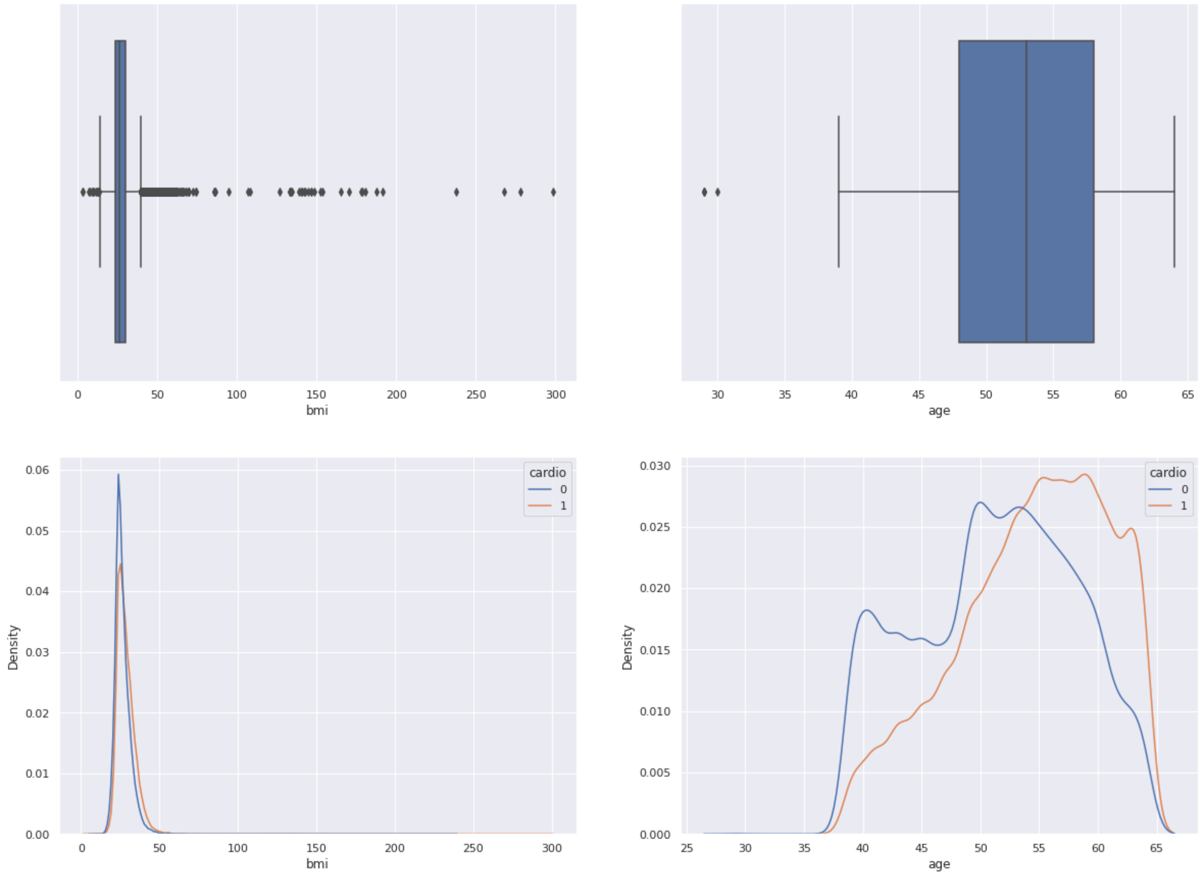
Glucose:



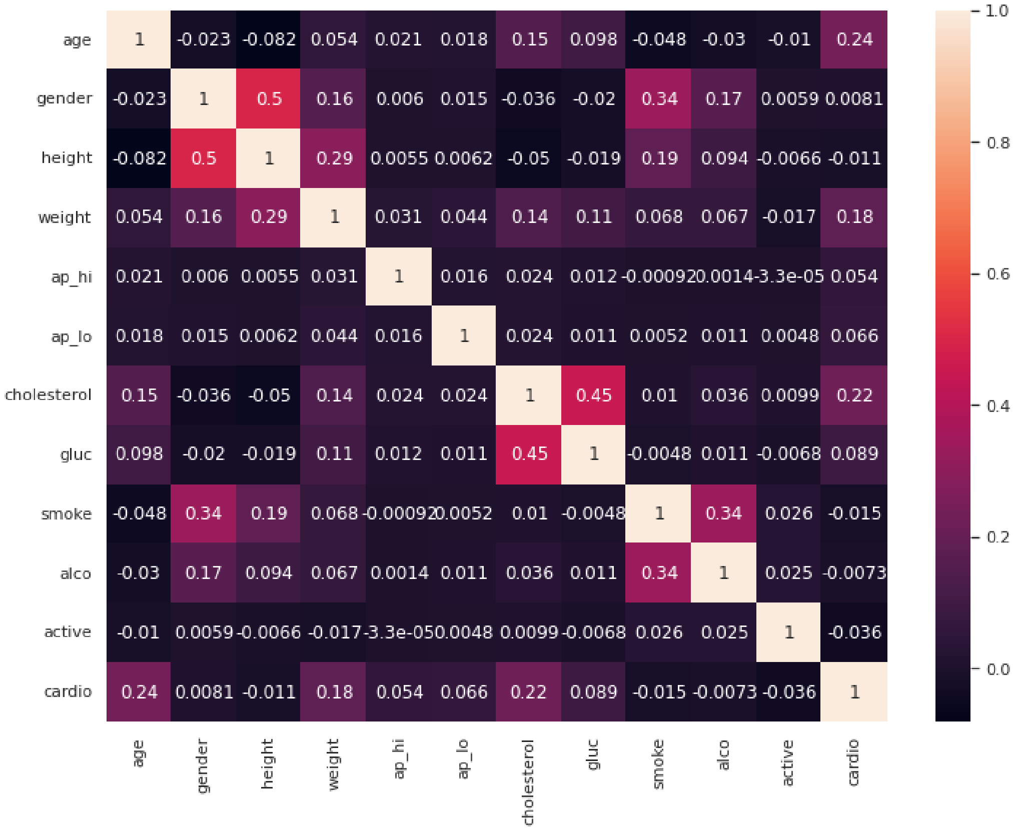
Cholesterol:



Summary:



The dataset’s correlation heatmap:



	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
age	1	-0.023	-0.082	0.054	0.021	0.018	0.15	0.098	-0.048	-0.03	-0.01	0.24
gender	-0.023	1	0.5	0.16	0.006	0.015	-0.036	-0.02	0.34	0.17	0.0059	0.0081
height	-0.082	0.5	1	0.29	0.0055	0.0062	-0.05	-0.019	0.19	0.094	-0.0066	-0.011
weight	0.054	0.16	0.29	1	0.031	0.044	0.14	0.11	0.068	0.067	-0.017	0.18
ap_hi	0.021	0.006	0.0055	0.031	1	0.016	0.024	0.012	-0.00092	0.0014	-3.3e-05	0.054
ap_lo	0.018	0.015	0.0062	0.044	0.016	1	0.024	0.011	0.0052	0.011	0.0048	0.066
cholesterol	0.15	-0.036	-0.05	0.14	0.024	0.024	1	0.45	0.01	0.036	0.0099	0.22
gluc	0.098	-0.02	-0.019	0.11	0.012	0.011	0.45	1	-0.0048	0.011	-0.0068	0.089
smoke	-0.048	0.34	0.19	0.068	-0.00092	0.0052	0.01	-0.0048	1	0.34	0.026	-0.015
alco	-0.03	0.17	0.094	0.067	0.0014	0.011	0.036	0.011	0.34	1	0.025	-0.0073
active	-0.01	0.0059	-0.0066	-0.017	-3.3e-05	0.0048	0.0099	-0.0068	0.026	0.025	1	-0.036
cardio	0.24	0.0081	-0.011	0.18	0.054	0.066	0.22	0.089	-0.015	-0.0073	-0.036	1

2.3. Data Preprocessing Techniques:

2.3.1. Deleting ID: In the dataset, there is an attribute named 'id' which is not necessary. So, we can delete it to improve the accuracy.

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

2.3.2. Converting age from day to year: The 'age' attribute has the unit *days*, which is not so convenient to work with. So, we can change it to, year by dividing by 365.

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

This preprocessing can lead to slightly greater accuracy.

2.4. Test-Train Split:

To train and test the data, we splitted the data into two parts. For training the models we are going to reserve 80% of the data and to test the models, we are going to reserve 20% of the data.

2.5 Training Models:

2.5.1. Random Forest:

Random forest is a non-linear model that consists of aggregating the results of an ensemble of decision trees, each created using a subsample of the data. A decision tree works in a 2-step process:

→ It divides the predictor space into separate rectangular regions (these regions are split in a way so that they minimize the RSS for regression trees, and the GINI index or entropy for classification trees).

→ It calculates the mean (or mode) of the outcome values for the portion of the sample in each region which is then used to predict new data.

We compared the accuracy of random forest to other non-linear models in predicting Y using X^1 to X^{90} chosen based on previous studies. We tried different models with 500, 1000, 2000, and 3000 trees, the number of variables tried at each split (mtry) was evaluated at 5 evenly-spaced values ranging from 2 to 90 as suggested by Kuhn & Johnson, and the node-size was set to default.

The cross-validation performance was used to tune the model, and a separate test set was used to evaluate its accuracy.

2.5.2. Logistic Regression:

This research chooses the logistic regression model, a widely used model in machine learning, which is frequently used in the real manufacturing setting in domains like data mining, automatic disease detection, and economic prediction. For instance, the risk factors for heart disease were described in this study, along with a prediction of the likelihood that the disease will occur. Since there are only two types of output, each of which represents a single category, logistic regression is widely used for classification, primarily two-category problems, and it can predict the likelihood that each classification event will occur.

Logistic regression model is shown below:

$$\mathrm{prob}(Y = 1) = \frac{e^{z}}{1 + e^{z}}$$

Where Y refers 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:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p$$

with constant β_0, coefficients β_j and predictors X_j, for p predictors (j=1,2,3,.....,p)

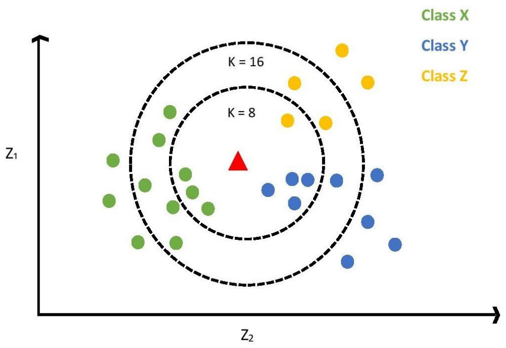
2.5.3. Multi-layer Perceptron:

In Multi-layer Perceptron the lone and primary task of the neurons in the input layer is the division of the input signal x_i among neurons in the hidden layer. Every neuron j in the hidden layer adds up its input signals x_i once it weights them with the strengths of the respective connections w_{ji} from the input layer and determines its output y_j as a function f of the sum, given as,

$$y_i = f(\sum w_{ji} x_i)$$

At this instant it is possible for f to be a simple threshold function such as a sigmoid, or a hyperbolic tangent function. The output of neurons in the output layer is determined in an identical fashion. The working of Multi-Layer Perceptron Neural Network is summarized in steps as mentioned below: 1) Input data is provided to the input layer for processing, which produces a predicted output. 2) The predicted output is subtracted from the actual output and the error value is calculated. 3) The network then uses a Back-Propagation algorithm which adjusts the weights. 4) For weights adjusting it starts from weights between output layer nodes and last hidden layer nodes and works backwards through the network. 5) When back propagation is finished, the forwarding process starts again. 6) The process is repeated until the error between predicted and actual output is minimized.

2.5.4. K-Nearest Neighbor:



When we add a new point to a dataset using the KNN algorithm we can predict which class the new point belongs to. In order to start the prediction, the very first thing we need to do is select the value of K. According to fig 1, points with green color belong to class X, points with blue color belong to class Y and yellow color points belong to class Z. When K=8, we need to select 8 neighbor points that have the least distance to the new point which is represented by the triangle. As demonstrated in fig 1 when K=8 new points are close to one yellow point, three green points, and four blue points. Since we have a majority of blue points, in this case, we can say that for K=8 the new point belongs to class Y.

Moving on ahead if K=16, we have to look for 16 different points which are closest to the new points. After calculating the distance, it is found that when K=16 new points are closer to three yellow points, five blue points, and eight green points. Therefore, we can say that when K=16 the new point belongs to class X.

In order to find the best K value, we can use a cross-validation technique to test several values of K.

2.5.5. SVC:

SVC is a supervised learning classification model. SVM first maps all the training data into a vectorized formation in multidimensional planes and then uses an algorithm to propose a hyperplane that maximizes the aggregate distance between the classified vectors in the space. The real capability comes into play in SVM where it can perform both linear and nonlinear classification using kernel. The default kernel in sklearn sklearn.svm.SVC is ‘rbf’ (radial basis function) which we are using to train our svc model. It uses the following formula:

$$K(x^i, x^j) = \phi(x^i)^T \phi(x^j) = e^{(\gamma||x^i - x^j||^2)}$$

Here, $\gamma > 0$

Results

3.1: Train - Test Accuracy Score:

3.1.1: Randomforest:

Accuracy for the train data is: **98%**

Accuracy for the test data is: **71%**

3.1.2: Logistic Regression:

Accuracy for the train data is: **72%**

Accuracy for the test data is: **72%**

3.1.3: Multi-layer Perceptron:

Accuracy for the train data is: **72%**

Accuracy for the test data is:**72.6%**

3.1.4: K-Nearest Neighbor:

Accuracy for the train data is: **75%**

Accuracy for the test data is: **71%**

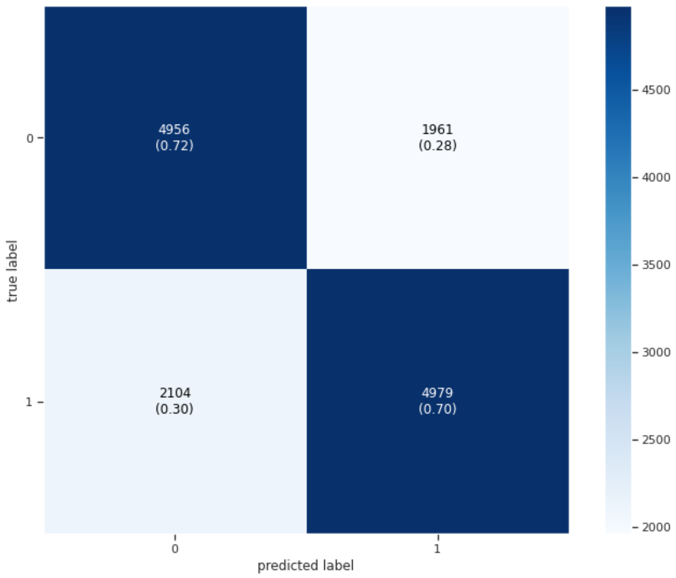
3.1.5: SVC:

Accuracy for the train data is: **72%**

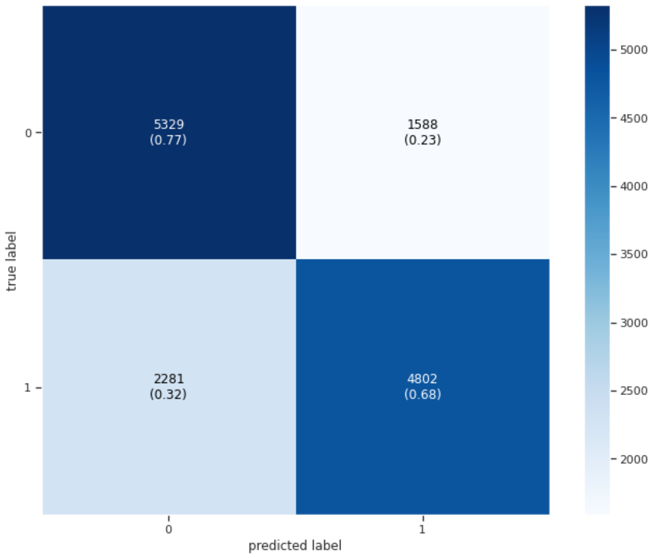
Accuracy for the test data is: **72%**

3.2: Confusion matrix:

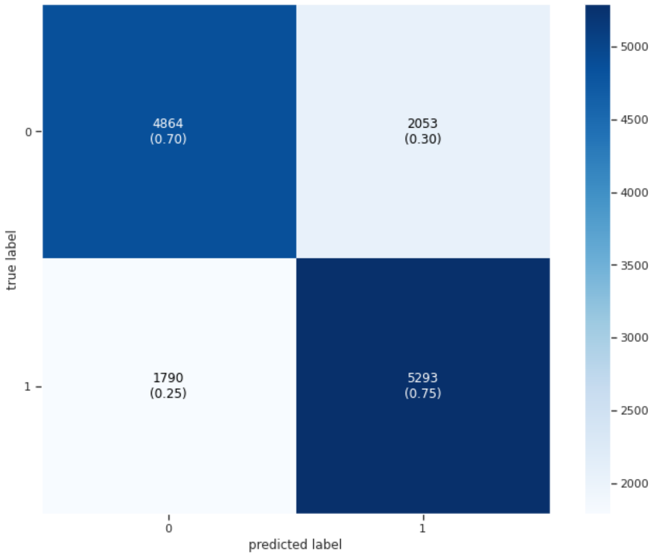
3.1.1: Randomforest:



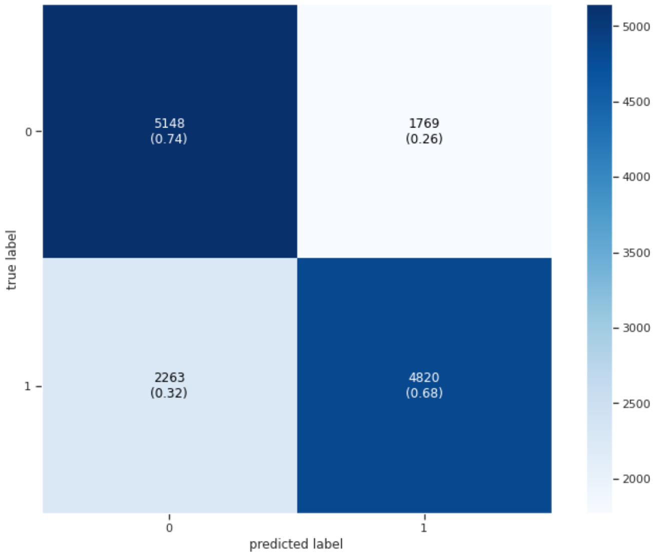
3.1.2: Logistic Regression:



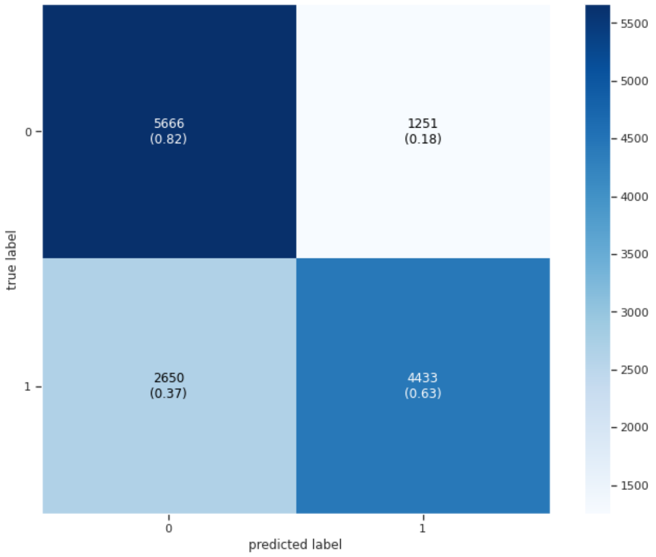
3.1.3: Multi-layer Perceptron:



3.1.4: K-Nearest Neighbor



3.1.5: SVC:



3.3. 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 Perceptron	0.73	0.73	0.73	0.73
K-Nearest Neighbor	0.71	0.71	0.71	0.71
SVC	0.73	0.72	0.72	0.72

4 Conclusion

From all the above results, we can say that, for training the dataset, the best model would be Random Forest Classification with the accuracy score of 98%. On the other hand, for testing, the best module would be Multi-layer Perceptron with an accuracy of around 73%.

5 References

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.