## IBM-322

**Analytics for Managerial Decision Making**

**TOPIC**

Examining the Potential for Heart Attacks Using Patients' Physical and Cardiovascular Information

**Group Number :20**

|  |  |
| --- | --- |
| Harsh Kumar | 21113058 |
| Kuldeep Singh Bhadouria | 21113078 |
| Himanshu Arya | 21113062 |
| Harshvardhan Singh Mehta | 21113061 |
| Kapil Jakhar | 21113074 |

# Problem Statement

Cardiovascular diseases (CVDs) pose a significant global health challenge, imposing a substantial burden of morbidity and mortality. Early identification and precise risk assessment of individuals susceptible to CVDs are vital for effective preventive measures. This project delves into the realm of data-driven healthcare, utilizing a comprehensive dataset comprising diverse attributes pertinent to cardiovascular health.

Our dataset encompasses crucial features such as age, sex, height, blood pressure readings, cholesterol levels, smoking and drinking habits, and physical activity levels, recognized as pivotal indicators in cardiovascular risk assessment. The project's objective is to leverage machine learning techniques to construct a predictive model capable of discerning patterns within the dataset and accurately pinpointing individuals at risk of CVDs.

By harnessing the power of data analytics, we aim to contribute to ongoing efforts in preventive healthcare, furnishing clinicians and healthcare practitioners with a valuable tool to proactively manage and mitigate cardiovascular risks. This report details the analysis and findings of our investigation into detecting cardiovascular diseases using the aforementioned dataset.

# Data Analysis

We have chosen the data set from Kaggle [Cardio Vascular Disease](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset) [Dataset.](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset)

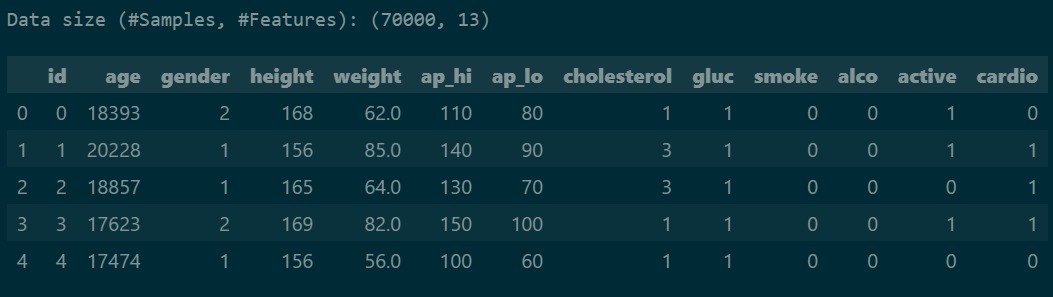
The dataset contains 70,000 rows and 13 columns .

There are 3 types of input features:

1. Objective: factual information;
2. Examination: results of medical examination;
3. Subjective: information given by the patient.

The 13 columns include patient id , age of patient (Objective Feature

) , gender (Objective Feature) , height (Objective Feature) , weight (Objective Feature) , Systolic Blood Pressue (Examination Feature) , Diastolic Blood Pressure (Examination Feature) ,cholestrol (Examina- tion Feature) , glucose (Examination Feature) , smoking and alcohol drinking status (Subjective Feature) , and physical activity (Subjective Feature) . Cardio is the target variable.

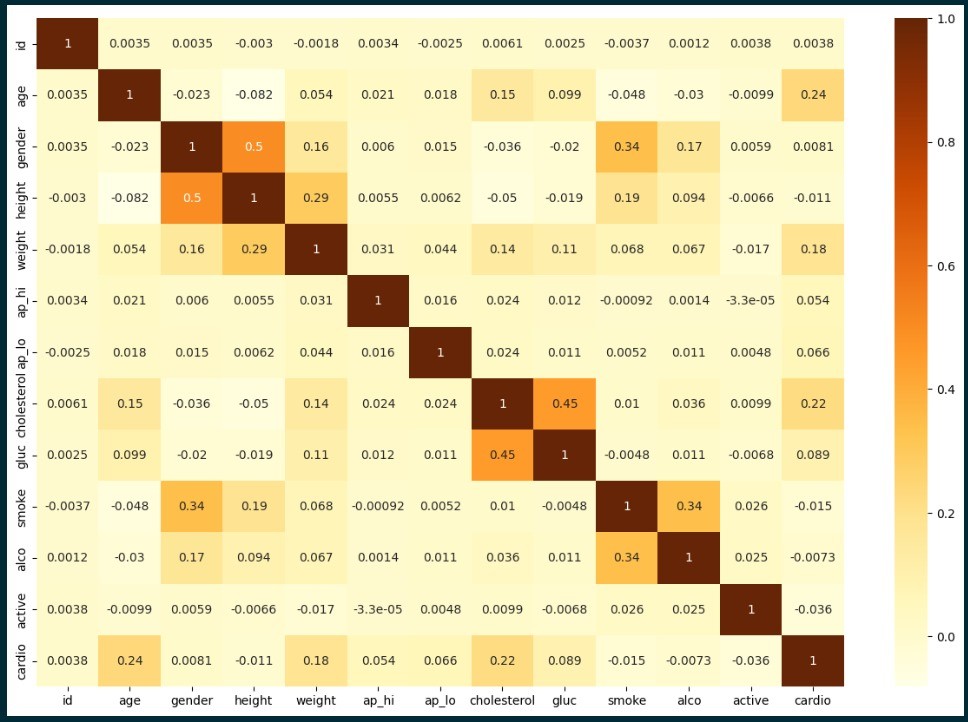


# Data Visualization

## Correlation Matrix

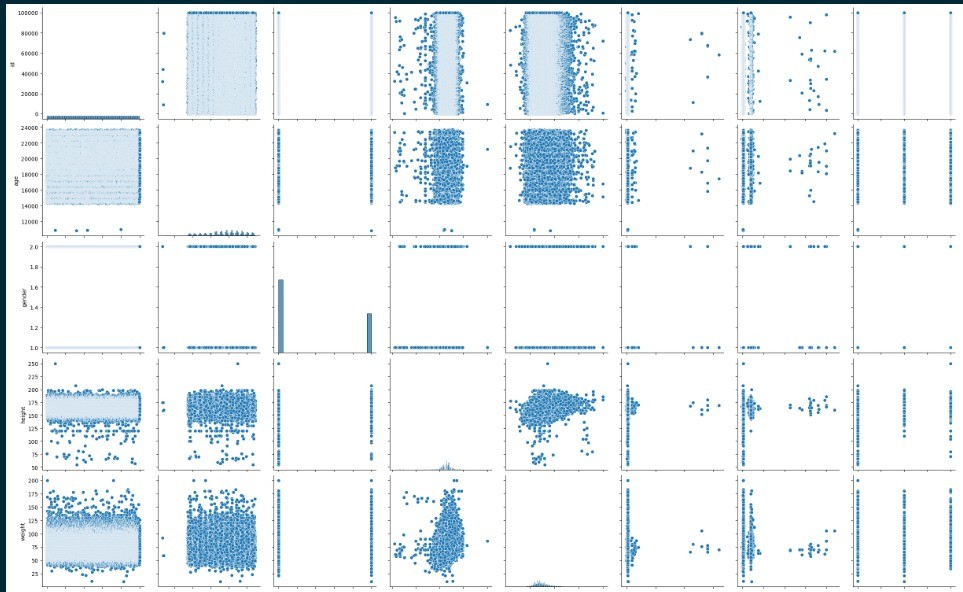
The Correlation Matrix serves as a cornerstone in uncovering complex relationships among various features in the dataset. This analytical instrument enables us to quantify and visualize both the strength and direction of associations between variables, providing a comprehensive grasp of their interconnections. Through scrutiny of the correlation matrix, we extract valuable insights regarding significant correlations among factors and discern potential multicollinearity issues..

Here a value in ith row and jth column denotes the correlation coefficient value between the ith and jth feature.



## Bi - Featurely Plot

The Bi-Feature Plot stands out as an impactful visual tool, offering a nuanced view of the connections between pairs of features in the dataset. This plotting method facilitates the examination of bivariate interactions, empowering us to detect potential patterns, trends, and outliers with clarity.



# Logistic Regression

We have used Logistic Regression in majorly two ways that are using the threshold of 0.5 for class prediction but since heart diseases are a crucial issue we have also reduced the threshold to **0.3 for reducing cases where a person had a disease but couldn’t be identified.**

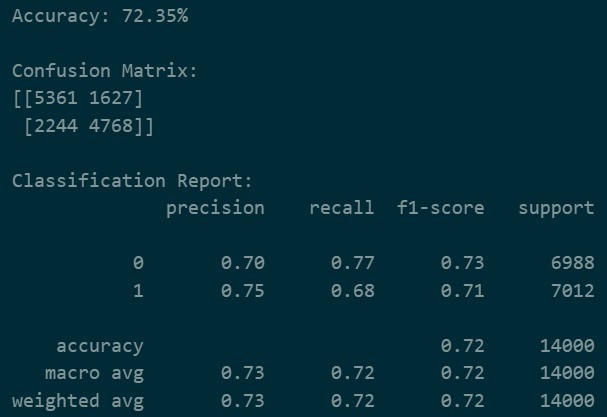
### Results for Logistic Regression with threshold 0.5

Here the model classifies a data point to be of type 1 that is having a disease if the probability of having a disease is greater than 0.5 . We get 0.7235 accuracy. Explanation of Confusion Matrix is as follows : The number of people who didn’t have the disease and were classified correctly are : 5361

The number of people who didn’t have the disease and couldn’t be clas- sified correctly are : 1627

The number of people who had the disease but couldn’t be classified correctly are : 2244

The number of people who had the disease and are classified correctly : 4768



### Results for Logistic Regression with threshold 0.3

Here the model classifies a data point to be of type 1 that is having a disease if the probability of having a disease is greater than 0.3 . We get 0.6260 accuracy.

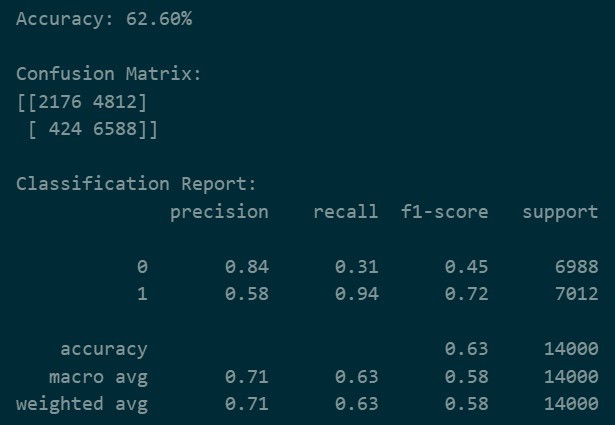
#### Explanation of Confusion Matrix

The number of people who didn’t have the disease and were classified correctly are : 2176

The number of people who didn’t have the disease and couldn’t be clas- sified correctly are : 4812

The number of people who had the disease but couldn’t be classified correctly are : 424

The number of people who had the disease and are classified correctly : 6588



# Multilayer Perceptron

#### The association between the provided features and the likelihood of Cardiac Disease may exhibit nonlinear behavior. Consequently, we employed the Multi-Layer Perceptron (MLP) model, renowned for its proficiency in capturing complex, nonlinear relationships within datasets. This approach allows for a more nuanced comprehension of the intricate interplay among various cardiovascular risk factors. The MLP's capability to discern subtle patterns renders it indispensable in our predictive modeling endeavors, aimed at unveiling concealed insights and refining the accuracy of our cardiovascular risk predictions.

#### In our implementation, we utilized an MLP model comprising 100 neurons in each hidden layer, with 50 hidden layers. This configuration yielded an accuracy of 0.7221, demonstrating the effectiveness of the model in navigating the complexity of cardiovascular risk assessment.

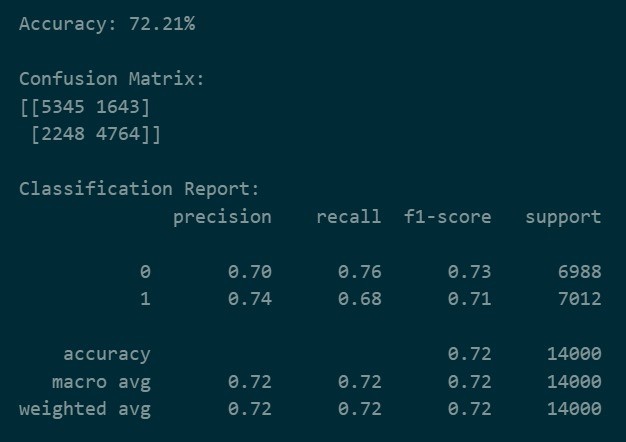
#### Explanation of Confusion Matrix

The number of people who didn’t have the disease and were classified correctly are : 5345

The number of people who didn’t have the disease and couldn’t be clas- sified correctly are : 1643

The number of people who had the disease but couldn’t be classified correctly are : 2248

The number of people who had the disease and are classified correctly : 4764

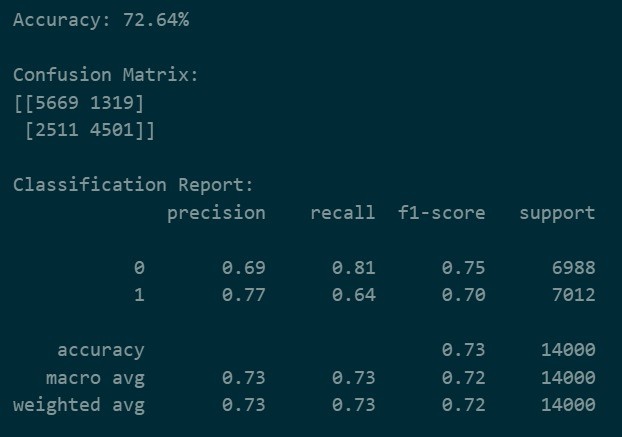


# Support Vector Machine

SVM is a supervised Machine Learning algorithm that can be used for classification or regression tasks . The basic idea behind SVM is to find a hyperplane that best separates data points of different classes in a high dimensional space. In SVM’s kernel plays a crucial role in allowing the algorithm to handle non linear relationships between features. We have used two kernels namely linear and rbf (with degree 3).

### Results for Support Vector Machine with linear kernel:

We have got an accuracy of 0.7264.



The explanation for confusion matrix is as follows :

The number of people who didn’t have the disease and were classified correctly are : 5669

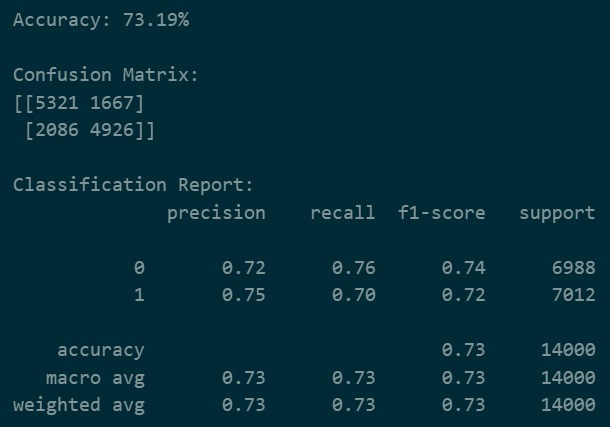
The number of people who didn’t have the disease and couldn’t be clas- sified correctly are : 1319

The number of people who had the disease but couldn’t be classified correctly are : 2511

The number of people who had the disease and are classified correctly: 4501

### Results for Support Vector Machine with rbf kernel

We have got an accuracy of 0.7319.



The explanation for confusion matrix is as follows :

The number of people who didn’t have the disease and were classified correctly are : 5321

The number of people who didn’t have the disease and couldn’t be clas- sified correctly are : 1667

The number of people who had the disease but couldn’t be classified correctly are : 2086

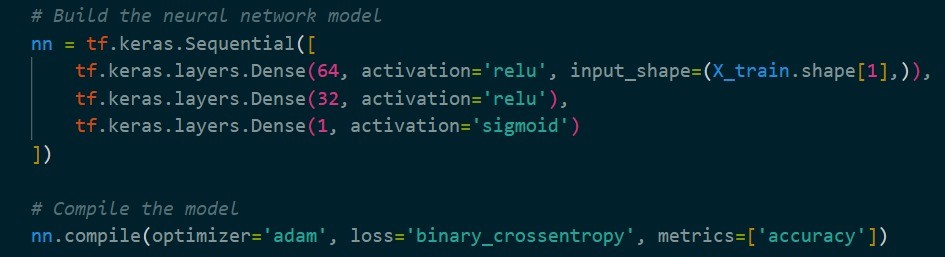
The number of people who had the disease and are classified correctly : 4926

# Neural Network

While Multilayer Perceptron (MLP) typically lacks the flexibility to adjust the internal neural network structure, our approach grants us complete control over the activation function for each layer. This means that each layer can employ a different activation function, enabling us to tailor our neural network for optimal predictions based on the dataset. To achieve this fine-grained control over the layers, we leverage the TensorFlow library, a powerful tool for building and customizing neural networks.

In our implementation, we compile the model using the Adam Optimizer, a widely used optimization algorithm. Adam optimizes Stochastic Gradient Descent by dynamically adjusting learning rates based on individual weights, thereby enhancing the efficiency of the training process and improving the model's predictive performance.

#### Building the neural network



The first layer contains 64 nodes and uses ReLu activation function, second layer contains 32 nodes and uses ReLu activation function, the third layer contains 1 node and uses sigmoid activation function.

#### Prediction Results for Neural Network

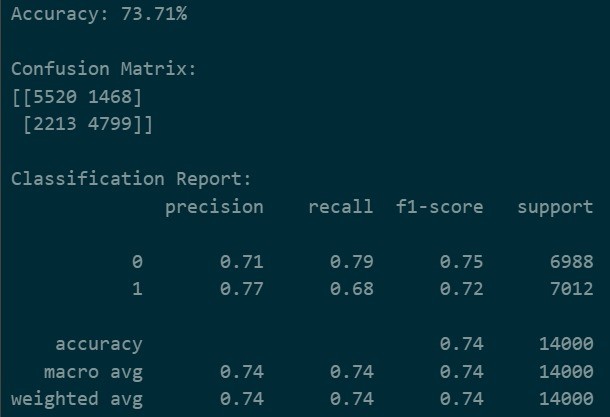
We have got an accuracy 0.7371 . The explanation for confusion matrix is as follows :

The number of people who didn’t have the disease and were classified correctly are : 5520

The number of people who didn’t have the disease and couldn’t be clas- sified correctly are : 1468

The number of people who had the disease but couldn’t be classified correctly are : 2213

The number of people who had the disease and are classified correctly : 4799



**References**

<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>