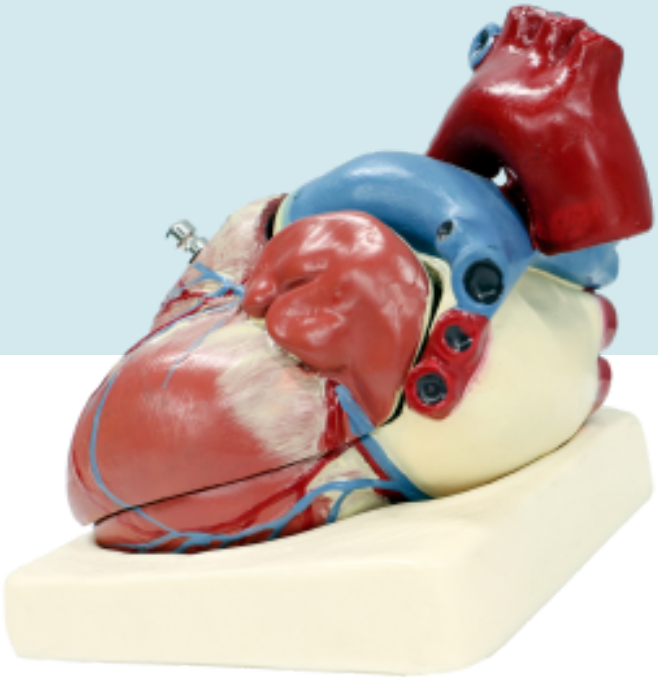
18 DECEMBER 2023

ISCHEMIA DATASET

DISTRIBUTED DATA ANALYSIS AND MINING REPORT



Group 3
Benincasa Pierfrancesco
Chiruzzi Camilla
Clot Mathilde
Seghieri Niccolò
Trentacapilli Guido

Tutor
Trasarti Roberto

CONTENTS

DATAFRAME CREATION

DATA UNDERSTANDING AND PREPARATION

CLASSIFICATION

01

ISCHEMIA DATASET DESCRIPTION

Ischemic heart disease is one of the main causes of death from cardiovascular diseases.
This dataset is made up of patients with ECGs, annotated by professionals, and diagnosed with ischemic heart disease. Each variation contained in the dataset represents an electrode placed on a different part of patient's body (arms, legs, chest) to record different angulations of heart activity. For each patient we have 12 measurements/derivations.

DEAL WITH THE DATASET

We start with 2,559 hea files and 2,559 mat files associated with each other. The hea files contains some information about the patient, such as age, sex, pathologies and information about the 12 derivations such as the starting values and the total sums. The corresponding 12 derivations will be in their entirety in a mat file.
Our goal is to create a merged dataframe from the present files, extract new attributes and, based on these, predict if the patient has or has not had an heart attack.
So, in order to create our dataframe from these files, we had to solve the problem of joining the hea files with the corresponding mat files. We therefore initially extracted the files from the main folder in two lists, depending on their format. In the mat list, we added to each element an attribute (mat file name) that will allow us to do the future join of the two lists.
We then used the parallelize function to exploit the potential of spark to distribute the data in two RDDs.
One problem we encountered was related to the maximum memory supported: in particular for the mat files, each containing 12 arrays of over 5000 elements, we had to set the number of partitions to 400.

02

HEA FILES

The next step was the creation of a function, **Get_diz**, that takes in parameter an element of the hea list (a parallelized hea file) recently created.
The purpose of this function is to create a dictionary from the information contained in a hea file. The values in this dictionary are information (essentially strings) extracted from the hea file, and the keys have been defined by us using the available documentation and the display of an example hea file.

We then map this function to each element of the hea list and create the resulting dataframe that will includes the following columns :

- **FileName**: the identifier of the related file;
- **NDerivations**: the number of patient derivations;
- **SamplingRates**: the sampling rate;
- **DurationRecordings**: the total duration of the ECG;
- **Mat**: indicates the reference to the related mat file containing the derivations;
- **ECGRapprSignal**: 12 columns of type ECGRapprSignal1, ECGRapprSignal2 ... ECGRapprSignal12 describing the format of ECG;
- **AmplitudeUnit**: there are 12 columns of the type AmplitudeUnit1, AmplitudeUnit2 ... AmplitudeUnit12 describing the unit used for the measurement;
- **Register**: 12 columns of the type Register1, Register2... Register12 indicating a parameter;
- **OffsetPar**: 12 columns of type OffsetPar1, OffsetPar2... OffsetPar12 indicating a parameter;
- **StartingValue**: 12 columns of the type StartingValue1, StartingValue2 ... StartingValue12 indicating values for each patient derivations;
- **SumValues**: 12 columns of the type SumValues1, SumValues2 ... SumValues12, indicating the sum of the values for each patient derivations;
- **Age**: the patient's age;
- **Sex**: the sex of the patient;
- **Dx**: the pathologies, a patient may have more than one pathology indicated by identification codes;
- **Rx**: treatment;
- **Hx**: medical history;
- **Sx**: symptoms.

03

Once the dataframe was created, we therefore performed a brief analysis to see if there was any information in the newly created dataframe that was redundant. In particular, looking at the distinct values, we discovered that NDerivation, SamplingRates, ECGRapprSignal, Register, OffestPar, Rx, Hx, Sx were always constant and should therefore be eliminated.
Continuing the analysis, we observed that there were two distinct amplitude units: one corresponding to a value of 1000/mV and another of 200/mV. Since the amplitude unit is machine dependent (the 1000/mV occurs only in the China recordings) we decided to eliminate these observations from the dataframe in order to have uniformity in the recordings made, thus going from 2,559 to 2,175 rows/patients.

MAT FILES

For the creation of the dataframe for the mat files, we followed the same logic as for the mat files. We created a **Create_dict** function that associates a progressive key of the form Der1, Der2 ... Der12 to each derivation, each containing an array whose values are the derivations's data.
We then map this function to each element of the mat list and create the resulting dataframe that will includes the following columns :

- Mat: the name of the mat file;
- Der: are 12 columns of the type Der1, Der2 ... Der12, that contain the arrays in their entirety.

JOIN

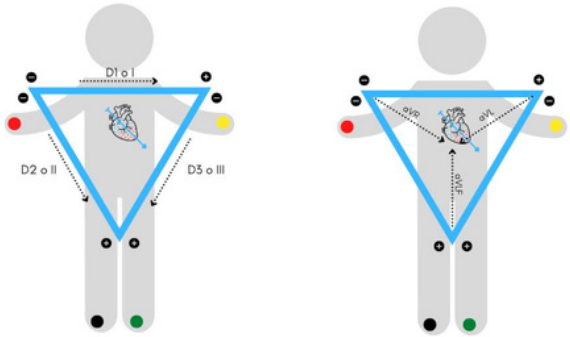
The final step is to perform a join between the two dataframes in order to obtain a single structure.
To accomplish this, we created two temporary views and executed a query using the mat file name as the join attribute.

We therefore finished this first part with the following columns in the joined dataframe:

- FileName, Age, Sex, Dx;
- StartingValue: 12 columns of the type StartingValue1, StartingValue2 ... StartingValue12;
- SumValues: 12 columns of the type SumValues1, SumValues2 ... SumValues12;
- Der: 12 columns of type Der1, Der2 ... Der12.

04

DATA UNDERSTANDING AND PREPARATION



NOTIONS ON ECG

In this part of the project, it proved very important to understand the structure of the derivations and the ECG, so as to be able to understand which data is best to work on.

Named after the theorist and father of modern electrocardiography, the Einthoven Triangle is the physiological principle on which the detection of the heart's electrical activity is based. The Einthoven Triangle is based on the imaginary arrangement of an inverted equilateral triangle on the patient's chest, the center of which coincides with the heart.

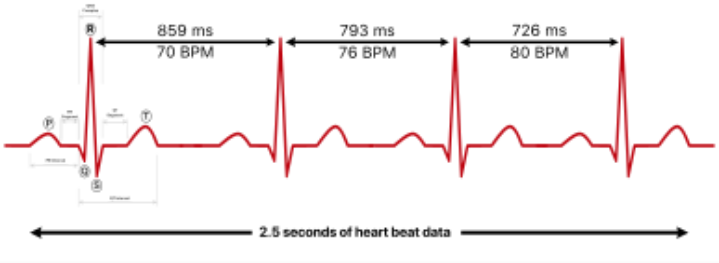
Each corner of the geometric figure is electrically coincident with a point on a specific limb that is assigned a name: **VL** (left, left) **VR** (right, right) and **VF** (foot, left foot). The remaining limb, the right foot is called neutral and does not participate in the formation of the triangle.

Placed in two body planes, the six peripheral leads observe the sagittal plane of the body, while the remaining six precordial leads observe the heart in the transverse plane.

- **I** or D1: measured between the positive electrode on the left arm and the negative electrode on the right arm;
- **II** or D2: measured between the positive electrode on the left leg and the negative electrode on the right arm;
- **III** or D3: measured between the positive electrode on the left leg and the negative electrode on the left arm;
- **aVF**: amplified lead on the left foot;
- **aVR**: amplified lead of the right arm;
- **aVL**: amplified lead of the left arm; added to the other six precordial unipolar leads (V1 to V6) allow a complete view of the heart on a three-dimensional plane and formed by many observation windows. It is as if twelve cameras were positioned around the heart ready to investigate every aspect of myocardial tissue

05

FEATURE ENGINEER



FEATURE EXTRACTION FROM ECG

For the data understanding task we start from the '.csv' dataset created in the 'data creation' phase. We start visualizing the format in which the .csv file saved the features, all the features initially are integers, except for the 12 derivations of each patient that are strings. This type of configuration will be a problem for the features analysis so we start casting into integer arrays the derivations.

The next step is the counting of Missing Values for the features. We find their presence only in the 'Age' column so we decide to fill this data with the average of the data grouped by sex (respectively 65 for the Male and 72 for the Female).

Beyond this we also do an analysis on 'DiagnosticCode' because it contains different Code (30+ distinct codes) for each patient. We code this feature into a binary feature, 1 if the corresponding user had a Heart Attack, 0 otherwise, because our goal is to predict through the data a possible Heart Attack.
Because the derivations are very hard to manage for their data structure, weight and size (5000 x 12 data for each patient), we construct new features. We start by deriving statistical measures like 'mean' and 'standard deviation' from the comparison of correlated derivations. We then obtain the following new features:

- **mean** attributes: leftleg_mean , critical_mean, leftarm_mean, rightarm_mean, total_der_mean
- **std** attributes: leftleg_std, critical_std, leftarm_std, rightarm_std, total_der_mean.

Each feature calculates statistical measurements relating to different electrode leads, more precisely :

- **leftleg** -> II, III, aVF;
- **critical** -> V2, V3, V4;
- **leftarm** -> I, aVL, V5, V6;
- **rightarm** -> aVR, V1.

To this we add three features: **tot_mean_array**, **tot_std_array** and **Bpm_mean**. tot_mean_array and tot_std_array refer to the combination of all the 12 derivations calculating the mean and the standard deviation. Bpm_mean is calculated using the library **Neurokit2** that extracts the r-peaks and calculates the intervals for each derivation.

06

DATA VISUALIZATION

ONCE WE GET ALL THE DATA CLEANED AND CODED WE START EXPLORING THE DATA THAT WE CONSIDER IMPORTANT.

ECG DERIVATION

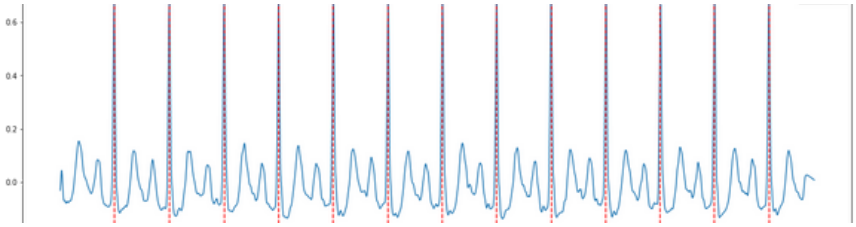
Below a representation of all the derivation using and the first derivation (I) of patient HR00102

R-R PEAKS:

In electrocardiography (ECG or EKG), an R peak is a prominent, upward deflection on the ECG waveform that corresponds to the depolarization (the process of electrical activation and contraction of the heart muscle) of the ventricles of the heart. The R peak is an important feature of the ECG signal and is used to measure various aspects of cardiac activity.

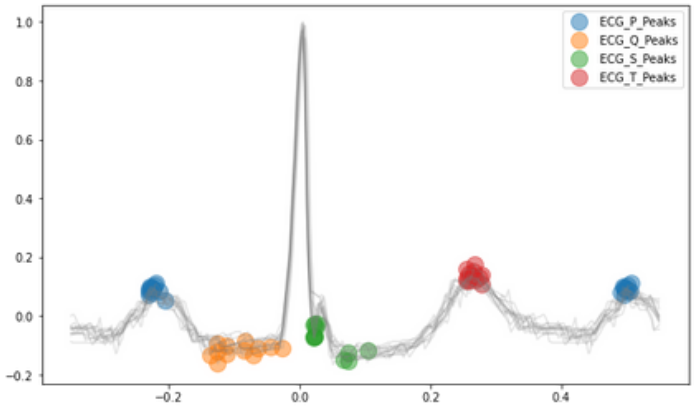
07

Data concerning R Peaks and measures are essential for assessing the electrical activity of the heart and diagnosing various cardiac conditions. Clinicians use ECG analysis to detect arrhythmias, evaluate heart rate variability, and identify abnormalities in the heart's conduction system.
It's important to note that the specific measures and their interpretation may vary based on the clinical context, and healthcare professionals are trained to interpret ECG signals in the context of a patient's overall health and medical history, for this reason this data can characterize the tuning of the parameter in the future model.



PEAKS PQST

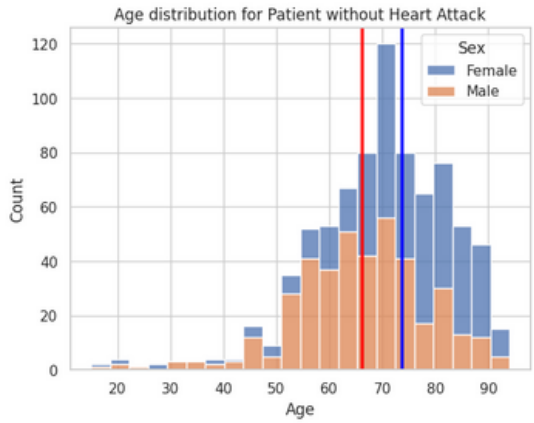
Beyond these information, there are many different 'waves' that are very significant and crucial in the analysis of an ECG like the **P Wave** (Atrial Depolarization: where the cardiac cycle begin), **T Wave** (Ventricular Repolarization: after depolarization comes repolarization, represents the electrical recovery phase of the ventricles), **QRS** (Ventricular Depolarization: comprises the Q, R, and S waves, the electrical activity associated with the contraction of the ventricles. The R wave is the most prominent upward deflection.) Below the different peaks extract from the patient.

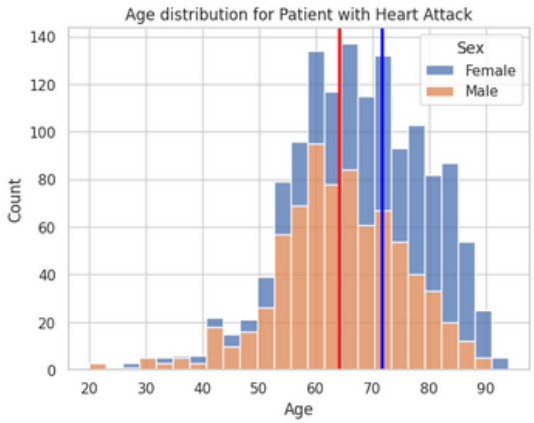


08

AGE DISTRIBUTION

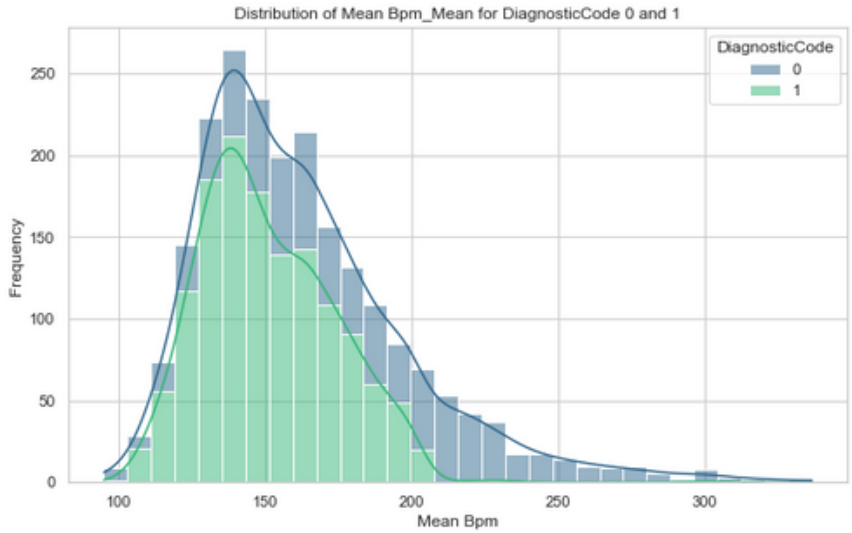
It is always is important to contextualize who is the subject of analysis and his medical history, which is why we delve deeper into the data of the patient we're analyzing.
We are therefore interested in the distribution of age conditioned by sex for patients who have had a heart attack and for those who have not.





AVERAGE BPM

Finally, the last plot, we check the average BPM for each DiagnosticCode. We can notate that the values are more similar in the central part of the distribution, but the DiagnosticCode equal to 0 presents a heavier right tail, so the patients who have not had a Heart Attack reach higher BPM.



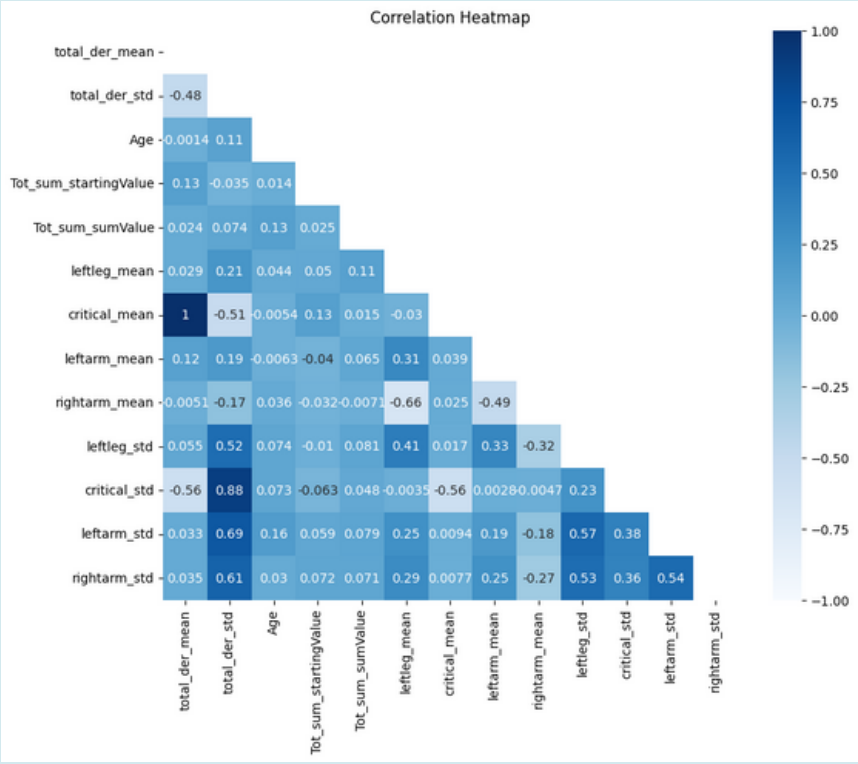
09

CORRELATION MATRIX

From the correlation matrix of the main variables of interest, we can observe how **total_der_mean**, the extracted feature representing the mean of values from all derivations combined (from I to V6), and **critical_mean**, representing instead the mean of values from the critically important ECGs for detecting heart attacks, are highly correlated (1).

For this reason, the decision was made to drop one of the two variables, specifically total_der_mean, to avoid redundancy in the information.

It is also noteworthy that, understandably, groups of variables representing the mean, as well as groups representing standard deviations, generally exhibit a high correlation, at least when compared to the correlation indices of other variables.



10

CLASSIFICATION

DATA PREPARATION

Regarding the classification phase, as mentioned earlier, the target variable is the presence or absence of a heart attack.

The two chosen classification models are:

- **random forest**
- **neural network**

Before proceeding with the classification, data preparation was conducted. Specifically, we utilized StringIndexer on the 'Sex' variable, creating a fitting VectorAssembler (while, of course, removing the target variable DiagnosticCode), and scaled all the features.

The dataset was then split into training (80%) and test (20%).
However, we soon realized how imbalanced it was, particularly with the training set comprising 638 patients with class label 0 and 1122 with class label 1.

As a solution, we decided to create a balanced training set using random oversampling with replacement, resulting in a perfectly balanced dataset.

RANDOM FOREST

We focused on optimizing the hyperparameters of our RF classification model and subsequently training the model with the identified optimal settings.
To systematically explore and select the best combination of hyperparameters, we employed a grid search technique.

Specifically, we constructed a parameter grid (ParamGrid) consisting of candidate values for the number of trees (numTrees) and maximum depth of each tree (maxDepth). The chosen values were [80] for numTrees and [6, 8, 12] for maxDepth.

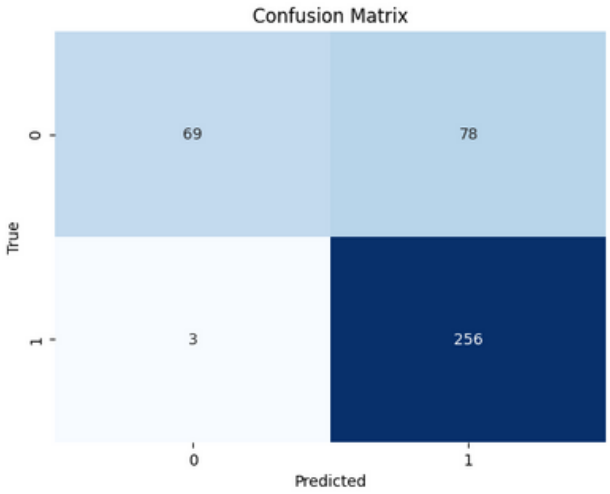
To ensure the robustness of our model evaluation, we implemented a five-fold cross-validation strategy.

11

RF
IMBALANCED DATA

Best parameter configuration:
(NumTrees: **80**, MaxDepth: **6**)

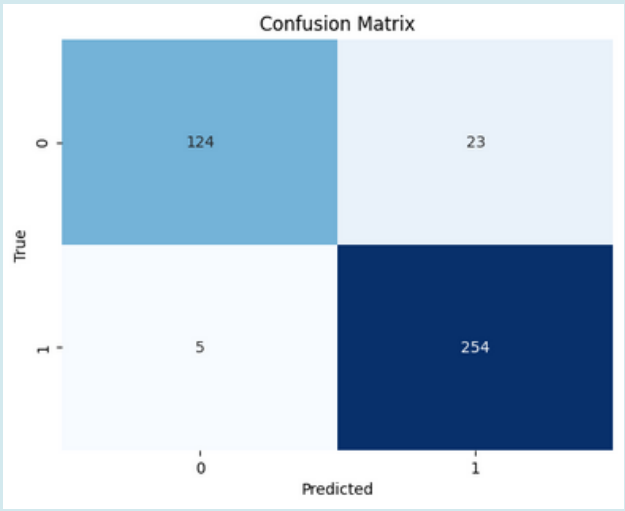
Results:
AuC: **0.87**
Accuracy: **0.80**
Precision: **0.77**
Recall: **0.99**



RF
BALANCED DATA

Best parameter configuration:
(NumTrees: **80**, MaxDepth: **12**)

Results:
AuC: **0.97**
Accuracy: **0.93**
Precision: **0.92**
Recall: **0.98**



12

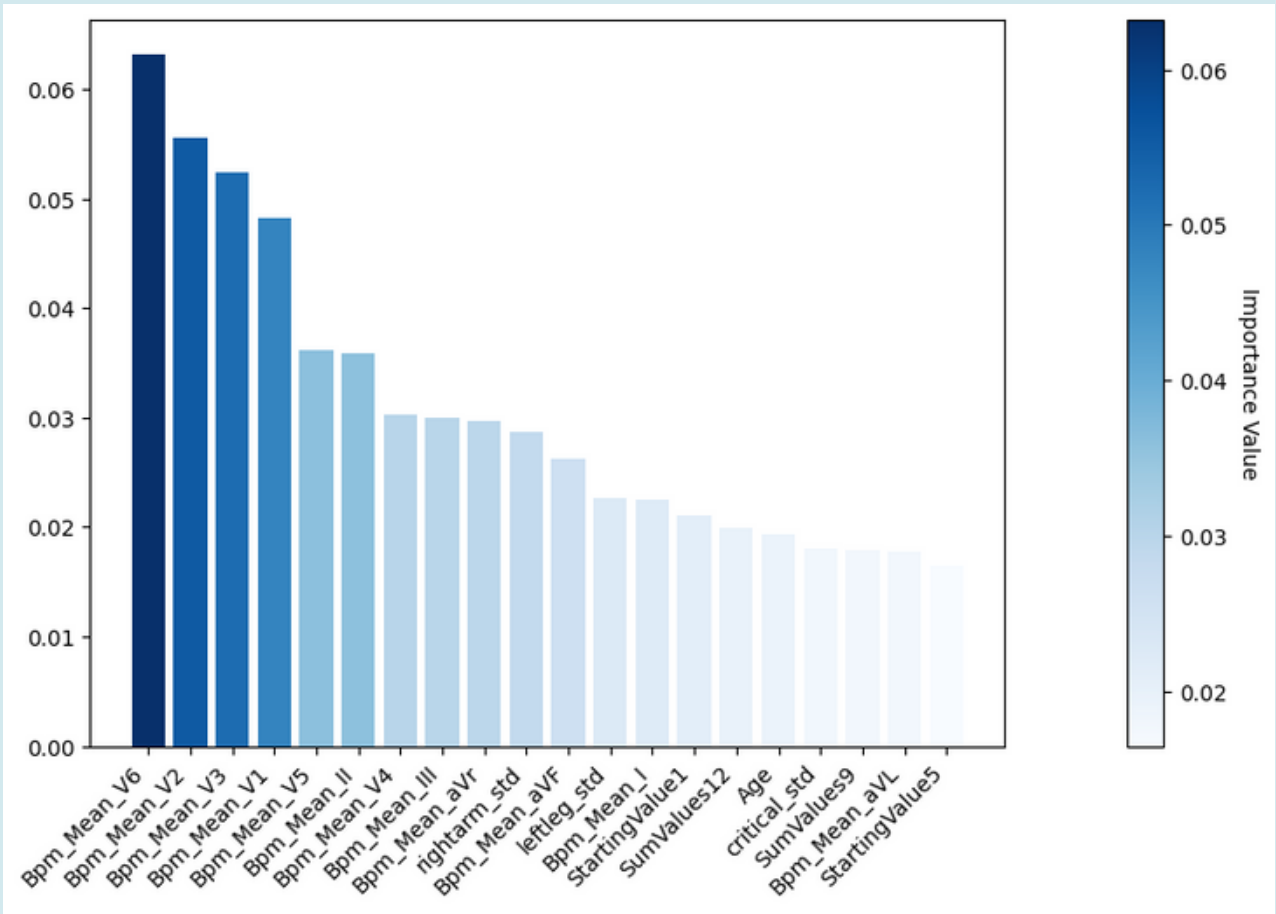
RF

FEATURE IMPORTANCE

By looking at the plot below, representing a graphical representation of feature importance, we can discern the key variables that strongly influence the model's decision-making process.

Notably:

- **Mean BPM** Variables: V6 stands out as the most influential, followed by V2 and V3. These findings underscore the significance of mean heart rate in predicting heart attacks.
- **Age**: emerges as a noteworthy predictor that recognizes age as a key factor in cardiovascular risk.
- **Standard Deviation of Critical District**: The variability in ECG signals within the critical district proves to be a meaningful contributor to the model's decision-making.



13

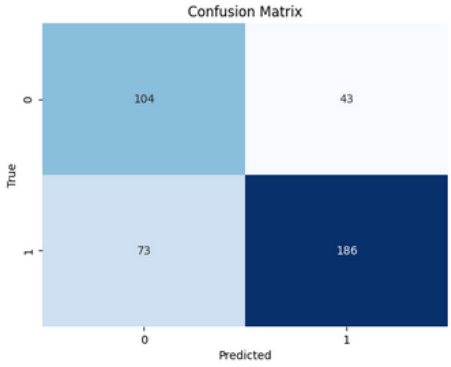
NEURAL NETWORK

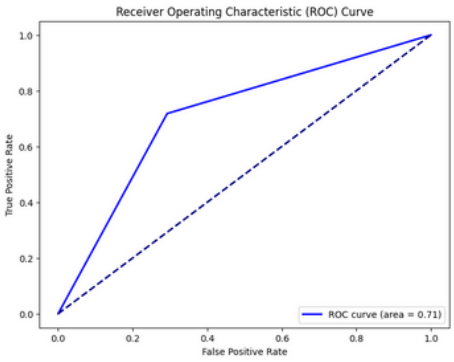
For the neural network model, a Multilayer Perceptron classifier was employed. The model's architecture, determined by the layers parameter, was configured for optimal performance. A grid search explored various layer configurations, and a three-fold cross-validation strategy ensured model generalizability.

The best parameter configuration is the following one:

- Block Size: 128
- Features Column: scaled
- Label Column: DiagnosticCode
- Maximum Iterations: 100
- Prediction Column: prediction
- Probability Column: probability
- Raw Prediction Column: rawPrediction
- Seed: -221787752041816854
- Solver: l-bfgs
- Step Size: 0.03
- Tolerance: 1e-06
- Layers: [49, 20, 10, 5, 2]

Results:
Accuracy: **0.71**
Auc: **0.71**





RESULT DISCUSSION

In the imbalanced setting, RF showed an impressive AuC, excelling in recall but at the expense of precision. The rebalancing process led to a significant boost, resulting in an AuC of 0.97, enhanced precision (0.92), and sustained high recall (0.98).
The Neural Network, demonstrated competitive but marginally inferior performance. The architectural choices might not have fully captured the complexity of the data.
The Random Forest, especially on a balanced dataset, outperformed the Neural Network in accuracy and discriminatory power. The enhanced performance of RF could be attributed to its ability to handle imbalanced data effectively and the refined understanding gained from the balanced dataset.
The ensemble approach employed by Random Forest proves well-suited for the heart attack prediction task.