UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Classification in heart disease problem

Aprenentatge Automàtic 1

Grau en Ciència i Enginyeria de Dades

Cristina Teixidó i Laia Royo

June, 2024

ÍNDEX

1. INTRODUCTION

This project aims to develop a predictive model by analyzing a database of heart disease-related information. The primary objective is to assess whether a patient is likely to have a heart disease based on various medical parameters such as age, sex, cholesterol levels, and oldpeak among others.

By carrying out this project, we will be able to help medicine to decide whether a patient is at risk of heart disease and to act before the condition develops.

To achieve this, we will initially divide our dataset into training and testing sets. The training set will be used to train several machine learning algorithms like logistic regression, decision trees, and perceptron models, among others. For each model, we will compute performance metrics such as accuracy, precision, recall, and F1-score.

Afterwards, in order to select the most effective model, we will compare the performance metrics of all the trained algorithms and choose the one that exhibits the highest predictive capability. Finally, we will apply the selected model to the test dataset to generate predictions and conduct a thorough analysis of the results to evaluate the model's performance and predictive accuracy.

2. PREVIOUS WORK

To predict outcomes from a medical database, understanding its features is crucial. That's why we're examining the range of numerical attributes and the significance of various options in categorical ones.

Feature 1: AGE

The age variable represents the patient's age. In our model, we include patients older than 25. It's important to note that as age increases, blood pressure and cholesterol levels often rise, which may increase the risk of heart disease.

Feature 2: SEX

This feature helps us understand how other variables differ between males and females. For example, women usually have lower resting blood pressure when they are young but are more likely to have high blood pressure after age 60.

Feature 3: CHEST PAIN TYPE

There are different types of chest pain:

1. Typical angina: *"Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. Angina pain may even feel like indigestion."* [1]
2. Atypical angina: The patient has other types of angina like unstable angina, microvascular angina, vasospastic angina, etc.
3. Non-angina pain: The patient has another type of chest pain not related to angina.
4. Asymptomatic: The patient doesn't have chest pain.

According to the American Heart Association, types of pain 1 and 2 are usually a symptom of a heart problem.

Feature 4: RESTING BLOOD PRESSURE

Normal resting blood pressure (BP) ranges from 80 to 140 mmHg. Higher values indicate hypertension, which can lead to high cholesterol or angina, and its risk increases with age.

Feature 5: CHOLESTEROL

Cholesterol is a type of fat that circulates in our body and is essential for normal functioning. However, high levels of certain types can be harmful. Cholesterol is considered high when it exceeds 200 mg/dl.

[1] Definition of Angina (chest pain) - American Heart Association

Feature 6: FASTING BLOOD SUGAR

Normal fasting blood sugar levels are 70 to 100 mg/dL. A level over 130 mg/dL on more than two occasions indicates diabetes, which also raises cholesterol levels.

Feature 7: RESTING ELECTROCARDIOGRAPHIC RESULTS

Provides valuable information about the heart's electrical activity. The database considers:

- Normal: No problems observed.
- ST: Abnormalities linked to higher risk of heart issues.
- LVH: Thickening and enlargement of the heart's left ventricle due to high pressure.

Feature 8: HEART RATE

Heart rate, or pulse, is the number of heartbeats per minute. A normal resting heart rate is between 60 and 100 beats per minute. In some cases, it can go up to 200 beats per minute, depending on age.

Feature 9: OLDPEAK

Oldpeak measures the difference in heart depression between exercise and rest. A normal value is usually 0; other values may suggest heart disease.

Feature 10: SLOPE

The slope in heart disease refers to the pattern of ST segment[2] changes on an electrocardiogram during exercise stress testing. A normal (indicating good health) slope is typically upsloping.

[2] A part of an electrocardiogram that represents the time between ventricular depolarization (when the heart is contracted) and repolarization (when it's relaxed).

3. DATA EXPLORATION PROCESS

3.1. ABOUT THE DATASET

The chosen dataset contains 919 samples and 12 features, including the target variable. The dataset is fairly balanced, we have 508 people with a cardiovascular disease and 411 healthy people. We'll now detail the various values that the medical parameters in the database can take and whether they are categorical or numeric.

- **Age**: Numerical attribute higher than 25.
- **Sex**: Categorical attribute with values "M" (male) and "F" (female). This feature is clearly unbalanced, with nearly four times as many samples for men compared to women. This will cause biases in the model's predictions, but we can't do anything about it.
- **Chest pain type (cpt)**:. Categorical attribute that can take these values:
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic

  Category 4 is dominant, followed by similar frequencies for categories 2 and 3, with category 1 having fewer observations.
- **Resting blood pressure (restBP)**: Numerical attribute that typically falls between 80 and 140 in a healthy patient.
- **Cholesterol (chol)**: Numerical attribute of the level of cholesterol of the patient.
- **Fasting blood sugar (fastBS)**: Indicates whether the patient's fasting blood sugar is greater than 120 mg/dl, 1 if it is true or 0 if is not. Categorical attribute with 2 options. This variable is also unbalanced, with quite more patients having blood sugar levels under 120 mg/dl.
- **Resting electrocardiographic results (restecg)**: Categorical attribute that represents the results of resting electrocardiogram with 3 possible values.:
    - Normal
    - ST
    - LVH

  "Normal" category has a large number of occurrences, while "ST" and "LVH" are evenly distributed among the remaining values.

- **Maximum heart rate achieved (maxHR)**: The maximum heart rate achieved by the patient. It is a numerical variable.

- **Exercise induced angina (exang)**: Binary variable. It's marked as 1 if the patient had chest pain during exercise and 0 if not. Not having pain is the dominant category, making this variable unbalanced. There is an imbalance between categories, with "noexang" appearing more frequently than "exang".

- **Oldpeak**: refers to a numerical value representing the ST segment depression induced by exercise relative to rest on an electrocardiogram (ECG).

- **Slope**: The slope of the peak exercise ST segment. Categorical with 3 values:
    - 0: Upsloping
    - 1: Flat
    - 2: Downsloping

  The results show a notable imbalance, with "Flat" having the highest occurrence, followed by "Up," while "Down" has significantly fewer instances.

- **Target**: Indicates the presence of heart disease in the patient. Binary categorical variable: 0 means no presence of a heart condition and 1 otherwise.

Before separating this data in the training (75%) and testing part (25%), we have setted categorical variables as objects and we have replaced names of their categories. Then, we have saved testing observations and preprocessed training data.

3.2. PREPROCESSING

During preprocessing, we handled missing values in restBP, chol, and maxHR. Some observations in restBP and chol were set to 0, which doesn't make sense given our previous work. Similarly, maxHR had observations with values of 999, which are also unlikely. We assumed these were missing values. Although oldpeak often equals 0 (as seen in Annex 1, Figure 1.4), we didn't label these values as NaNs. Previous research suggests that a 'normal' oldpeak value for a healthy person is 0, so it makes sense that there are 271 observations where oldpeak is equal to 0.

To fill NaN values accurately, we applied the 1NN imputation method. This involves replacing missing values in a column with the nearest neighbor's value from the dataset. To utilize this method effectively, we converted categorical features into dummy variables.

Next, we tackled outliers. In the age distribution plot, we spotted an observation below 25, which falls outside the study's age range, so we removed it. Additionally, cholesterol and restBP had outliers, with cholesterol having a long tail of high values. We used the Local Outlier Factor with 30 neighbors to normalize the distribution. The histogram of cholesterol (Figure 1.6) was adjusted to have a maximum value below 400, which is very high for cholesterol, as mentioned in Previous Work.

Finally, we transformed the data using Box-Cox and scaled it to center observations around 0, achieving Gaussian distribution and consistent variance across variables. See Figure 1.6 for the results.

The restBP histogram displays an unusual pattern resembling a categorical attribute. This phenomenon, unnoticed until now as we examined column numbers individually, occurs because most values have been rounded. Despite this, we'll keep this variable as numerical due to its inherent order.

3.3. FEATURE SELECTION

Before modeling, we need to ensure that the variables are independent. To do this, we'll examine both the correlation matrix and the VIF of the variables.

In the correlation matrix (see Figure 1 in Annex 1), we observe that 5 features have high correlation values exceeding 0.6. However, these features were originally categorical and were transformed into dummy variables, so we won't remove them. Now, let's calculate the VIF to confirm that there is no multicollinearity among the variables. The results show that all VIF values are below 5, indicating no significant multicollinearity. Therefore, we can proceed without altering any variable in the dataset. None of the feature pairs exhibit high correlation except for the dummy variables.

Finally, we'll calculate the correlation between the features and the target to identify which features explain the target variable the most in linear models. Through this analysis, we discover that the slope of the peak exercise is highly correlated with the target (nearly 0.7), as well as the "chest pain type" feature in asymptomatic patients (with a correlation value of 0.541). This insight guides us in understanding which features are most important in the linear models.

4. MODELING METHODS

We're faced with a classification problem where we aim to determine whether a person, based on certain characteristics, is healthy or not. So, we will try some methods and finally choose the one with the best results:

Linear Regression, Ridge Regression and Lasso Regression: these methods are the first ones that we've tried because they are the simplest ones, but they are better for lineal problems. Ridge and Lasso are extensions that help to regulate the complexity of the models. Once we fit each model (choosing the most efficient lambda for each model) we will observe the part that they explain of the dataset (R^2).

QDA and LDA: these methods are generative classifiers. The difference between them is that LDA assumes a common covariance matrix and QDA assumes that each class has its own covariance matrix. Because of our knowledge in the dataset, we think that covariance matrices will be different and then QDA will work better. However, we will try both and study the best option.

Logistic Regression: this method is a discriminant classifier. More concretely, is an algorithm that performs binary classification. Specifically, Logistic Regression uses regular linear regression to model the 'logit' function. We will adjust it and see how it works using a confusion matrix.

Decision Tree: predicts the target by learning simple decision rules. During training, it splits the data in order to learn these rules. It is very fast to train and interpretable, but it can be easily overfitted. Therefore, it will be essential to choose some hyperparameters (maximum depth, criterion, minimum sample splitting and minimum sample sheet) to avoid it.

Random Forest: this method is an ensemble of Decision Trees. We average high-variance but decorrelated individual decision trees to avoid overfitting. We will also use Out-of-Bag (OOB) error. It allows us to tune hyperparameters without having to execute costly cross-validations. It acts as a validation score.

SVM: seeks to find the most effective decision boundary, or hyperplane, that can separate data points belonging to different classes. We'll explore different kernel functions—such as polynomial, RBF, linear, or sigmoid—to determine which one is the best fit for our dataset.

Clustering: groups similar data points together based on their features, without predefined labels. It aims to identify inherent structures or patterns in data by partitioning it into subsets or clusters. This unsupervised learning method is widely used for tasks like customer segmentation, anomaly detection, and data summarization.

5. COMPARISON OF RESULTS

In this section, we'll analyze the outcomes of various models applied to our dataset to determine the most effective one. We'll use different metrics for comparison. All predictions were made using cross-validation on the train part (75% of the dataset), so the metrics were derived from these cross-validated predictions.

For regression models, we'll evaluate Linear Regression, Ridge, and Lasso based on their R-squared scores in both training and cross-validation (CV) results. We'll use the optimal lambda values previously determined in separate tests for each model

	Train R^2	CV R^2
Linear Regression (λ=0)	**0.602**	**0.547**
Ridge Regression (λ=10)	0.599	0.554
Lasso Regression (λ=0.001)	0.601	0.546

Table 1. Comparison of R^2 in different Regression models

As we have a classification problem it's seen that we obtain very bad predictions so we will directly discard these three methods.

In Logistic Regression, we have obtained better results than the previous ones. In Figure 3 (see Annex 2.2) we can see the ROC[3] curve is close to the upper left corner, and AUC[4] is near 1 indicating a good prediction.

For QDA and LDA, we can compare the two methods with this Figure 2 (see Annex 2.1). In QDA, the percentage of False Positive is 7.07% and False Negative is 7.88%. On the other hand, LDA has 7.27% False Positive and 5.86% False Negative. So, as the total percentage of False predictions in LDA is lower than in QDA, we can conclude that for our dataset **LDA** works better (on the contrary to what we initially thought).

In decision tree and random forest models, we evaluate their performance using metrics such as accuracy and class-specific scores. We compare both the default and optimized

[3] A graphical representation that illustrates the performance of a binary classification model across different threshold settings. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values. Demonstrates the trade-off between sensitivity and specificity across different threshold values.

[4] A scalar value that quantifies the overall performance of a binary classification model based on its ROC curve. It represents the area under the ROC curve.

versions of these models. The optimized models have parameters that were tuned through multiple iterations to find the best fit for our dataset. Additionally, we used a balanced random forest model to address class imbalances. The balanced random forest model achieved the best performance, with an accuracy of 0.871.

In the Support Vector Machine model we have different options: linear, non-linear (RBF), sigmoid and polynomial kernel. For each we will try different values of C and epsilon: The option with the best R2 and the lowest MSE is the SVM linear with C = 0.1.

For clustering, we used the k-means algorithm. To determine the optimal number of clusters (k), we computed several metrics for different values of k: the Calinski-Harabasz (CH) score, the Davies-Bouldin (DB) score, and the silhouette score. The best results for these metrics were achieved when k = 2, so we chose this value (Figure 5 in Annex 2.3) . To validate our conclusion, we also fitted the model using different values of k. However, the results were poor, with precision and accuracy both near 0.5 (see Figure 4 in Annex 2.3 for a representation of the results obtained)

Finally, we will compare all the previous models and choose the better. We would use accuracy[5], and F1-score[6] to compare the best model and also, we would take into account the precision[7] and recall[8] of class 1. This decision is due to the purpose of the classification: we prefer to frighten a healthy person (false positive) than letting a sick patient die (false negative), for this reason, we consider a high level of recall in the chosen model necessary.

From Table 2, we see that the two models with the highest recall are Random Forest and SVM. Although their accuracy, precision, and F1-score are slightly lower compared to some other methods, the differences are minimal. Most importantly, recall is the most critical metric for our problem, and these models achieve the highest recall, exceeding 0.9. Additionally, their performance in the other metrics is also satisfactory.

[5] measures the overall correctness of a model's predictions. It is the ratio of correctly predicted instances to the total number of instances in the dataset.

[6] is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when the classes are imbalanced.

[7] measures how much the model is predicting correctly a class with respect all the predictions of this class. We will use this metric when having false positive predictions is very harmful in our model context.

[8] measures how much the model is predicting correctly a class with respect all the real values of this class. We will use this metric when having false negative predictions is very harmful in our model context.

The importance of recall stems from our preference for having false positives over false negatives, as previously mentioned. Recall measures the ratio of correctly predicted positive instances, essentially telling us how many sick patients are correctly diagnosed by our model.

MODEL	Precision (class 1)	Recall (class 1)	Test accuracy	F1-score
Linear Regression	0.952	0.152	0.545	0.262
LDA	0.867	0.89	0.869	0.878
Logistic Regression	0.873	0.867	0.863	0.870
Decision Tree	0.899	0.878	0.883	0.888
Random Forest	0.862	0.901	0.871	0.881
Perceptron	0.827	0.779	0.796	0.802
SVM	0.874	0.920	0.887	0.896
Clustering	0.601	0.598	0.598	0.598

Table 2. Comparison of chosen models

6. FINAL MODEL AND ITS PERFORMANCE

Finally we have chosen Random Forest and SVM as final models. These models have a 0.871 and 0.887 test accuracy, respectively. However, these values are calculated with the training set. So it's the moment to test the data we had saved at the beginning. This part of the dataset is not preprocessed and has not been involved in the training of the algorithm. However, we will apply to the new data the same transformations we have applied in the training test to obtain the expected results.

By applying Random Forest, we've obtained these results:

	Predicted 0	**Predicted 1**
True 0	75	19
True 1	11	72

Accuracy: 0.831 | **Precision**: 0.798 | **Recall**: 0.856 | **F1-score**: 0.826

For SVM we have:

	Predicted 0	**Predicted 1**
True 0	73	21
True 1	11	72

Accuracy: 0.819 | **Precision**: 0.774 | **Recall**: 0.867 | **F1-score**: 0.818

As we can see, SVM has a slightly better recall, but Random Forest performs better in the other metrics. While recall is crucial for our dataset, considering the trade-off between recall and precision (as reflected in the F1-score), Random Forest shows a better balance. Therefore, we will choose Random Forest as our final model.

7. CONCLUSIONS

During preprocessing, we addressed a significant number of missing values by setting them to 0 and dealt with some outliers. We also observed that all the features in the dataset were correlated with the target and useful for prediction. We tried various models, optimized their hyperparameters, and evaluated their performance using cross-validation on the training data. After comparing all the models, we selected two finalists: Random Forest and SVM.

Next, we calculated their real accuracy using the test data. This part of the data wasn't preprocessed initially, so we transformed the appropriate variables. After making predictions with each model, we concluded that Random Forest works better for our dataset.

We have seen that the Random Forest model performs well on our dataset, correctly predicting 80% of the cases presented. Moreover, we successfully diagnosed 86% of the sick patients, which was the primary aim of the project.

In conclusion, Random Forest is our chosen model because it not only provides a high recall, correctly diagnosing 86% of sick patients, but also maintains a good balance across other metrics. This balance ensures reliable overall performance, making Random Forest the best choice for our predictive needs.

Despite these successes, we acknowledge some limitations that likely made the model less general and precise. Firstly, the data imbalance, especially in gender representation, with more male than female samples, could bias the model, making it less effective for predicting heart disease in women. Additionally, features like cholesterol and blood pressure are highly variable due to lifestyle factors and medication, affecting model accuracy. Furthermore, the model was trained and validated on a single dataset, so its generalizability to other populations or datasets is not guaranteed, as different demographics might exhibit different heart disease patterns. Lastly, preprocessing choices, such as handling missing values and outliers, can introduce bias.

8. BIBLIOGRAPHY

Riesgo cardiovascular. Fundación Española del Corazón. https://fundaciondelcorazon.com/prevencion/riesgo-cardiovascular.html

Mean fasting blood glucose. World Health Organization. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380

All about heart rate (pulse). American Heart Association. https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood

ST segment. Wikipedia. https://en.wikipedia.org/wiki/ST_segment

LeWine, H. E. (2023, June 13). What your heart rate is telling you. Harvard Health. https://www.health.harvard.edu/heart-health/what-your-heart-rate-is-telling-you

Scikit-learn and categorical features. Investigate.Ai: Data Science for Journalists. https://investigate.ai/classification/scikit-learn-and-categorical-features/

Curry, R. (2021, October 25). Simplified logistic Regression: Classification with categorical variables in python. Medium. https://medium.com/@curryrowan/simplified-logistic-regression-classification

Deshmukh, H. (2020, June 18). Heart disease UCI-diagnosis & prediction. Towards Data Science. https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction

Hottinger, M. (2020, June 26) K-Means Clustering for Analysis of Heart Disease. Medium. https://medium.com/@michellibelly/k-means-clustering-for-analysis-of-heart-disease

ANWAAR, S. (2022) [Beginner - K-means] - Heart Attack Analysis https://www.kaggle.com/code/sohaibanwaar1203/beginner-k-means-heart-attack-analysis

Arias, M., & Molina Martínez, A. (2024). Jupyter Notebooks. Aprenentatge Automàtic 1, Grau en Ciència i Enginyeria de Dades , Universitat Politècnica de Catalunya (UPC).

9. ANNEX

9.1. Annex 1: about the dataset

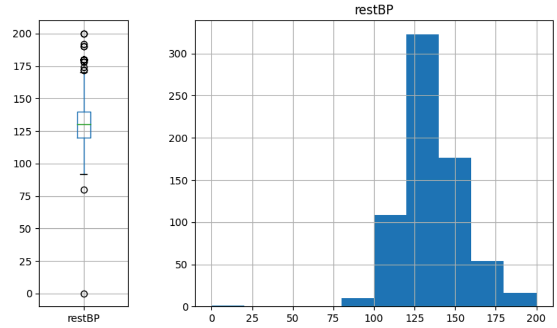

Figure 1.1 Boxplot and Histogram of "restBP" column

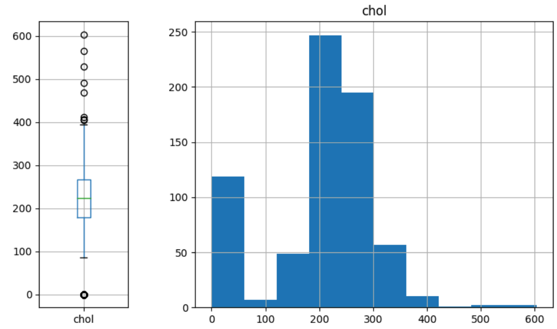

Figure 1.2 Boxplot and Histogram of "chol" column

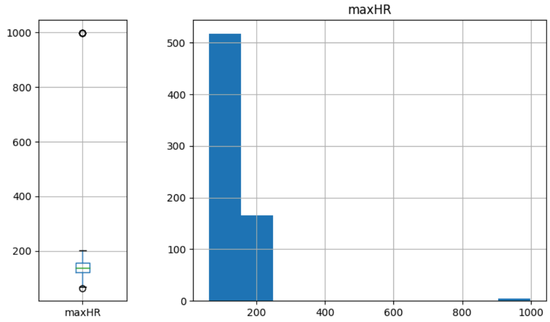

Figure 1.3 Boxplot and Histogram of "maxHR"column

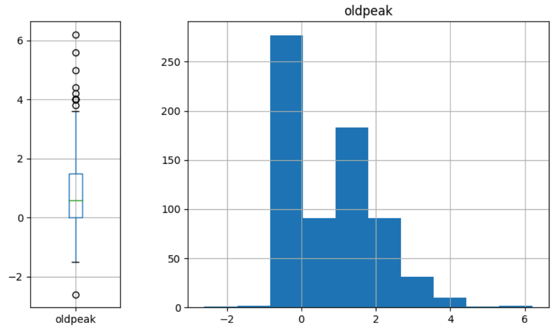

Figure 1.4 Boxplot and Histogram of "oldpeak" column

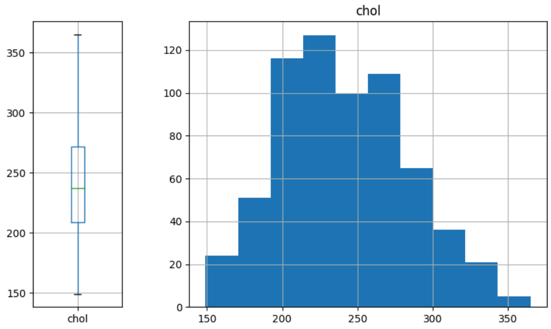

Figure 1.5 Boxplot and Histogram of "chol" column

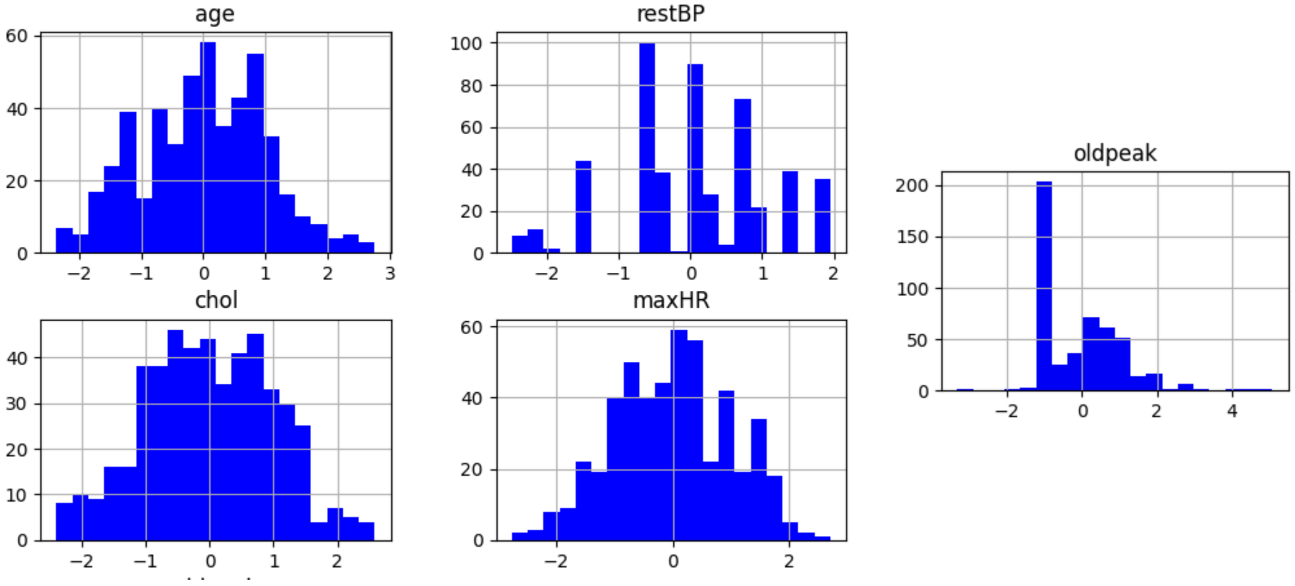

Figure 1.6. Histograms of transformed numerical variables

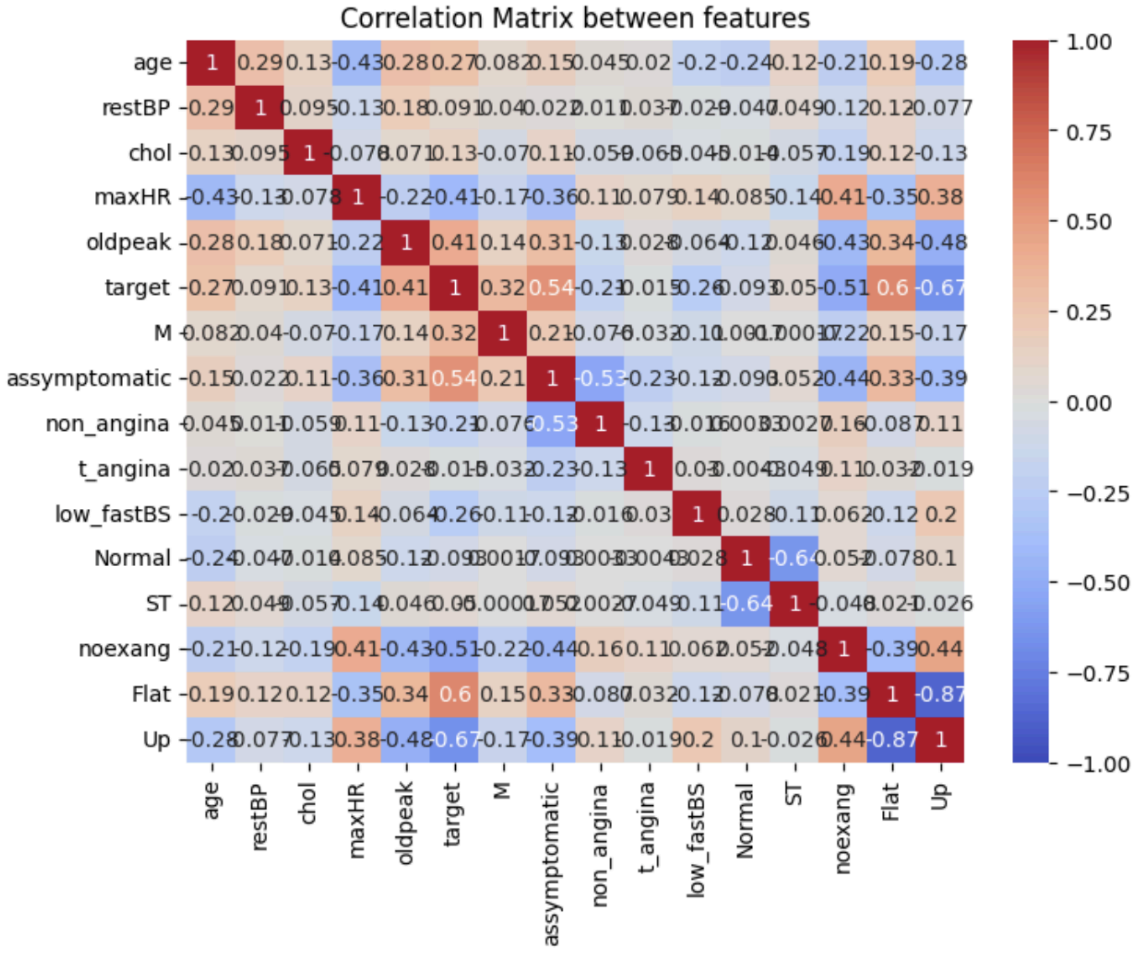

Figure 1.7. Correlation matrix between features

9.2. Annex 2: models

9.2.1. Annex 2.1: QDA and LDA

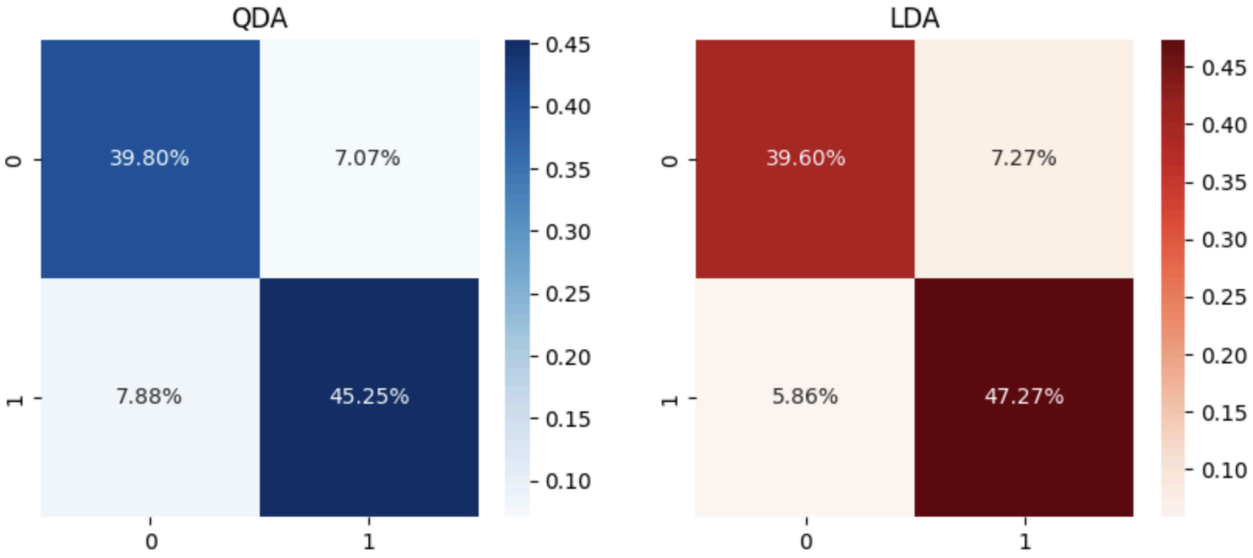

Figure 2. Confusion matrices for QDA and LDA

9.2.2. Annex 2.2: Logistic regression

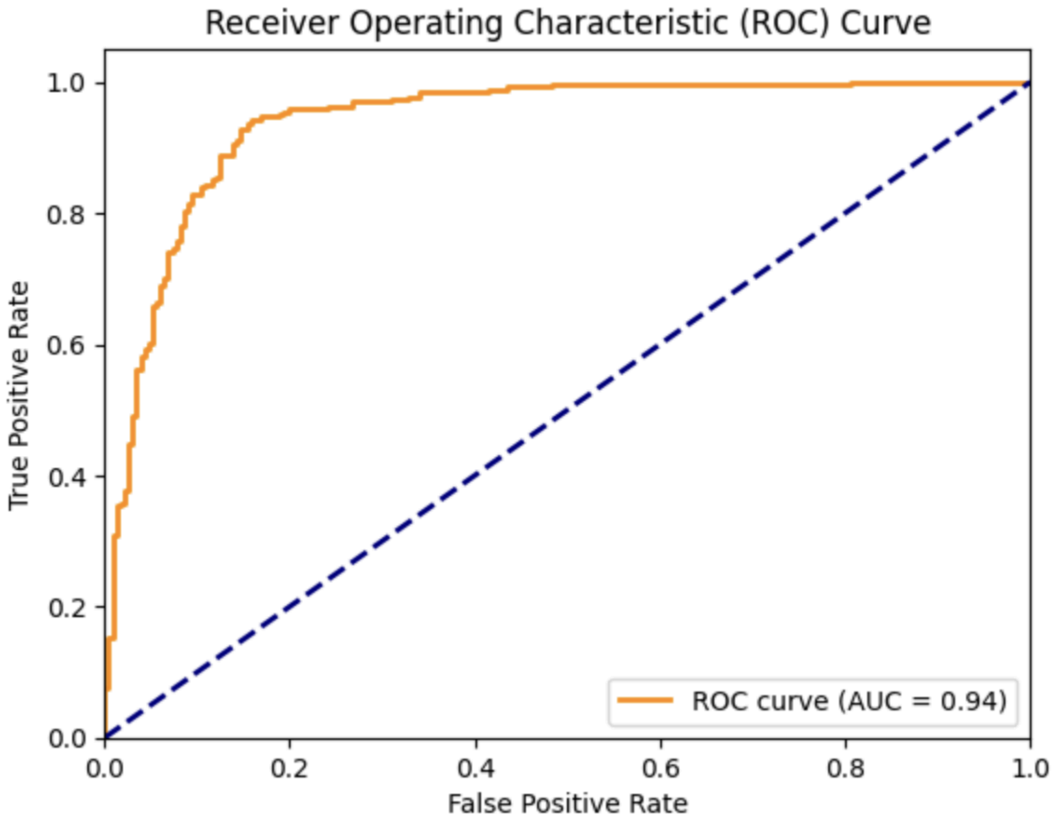

Figure 3. ROC curve

9.2.3. Annex 2.3: Clustering

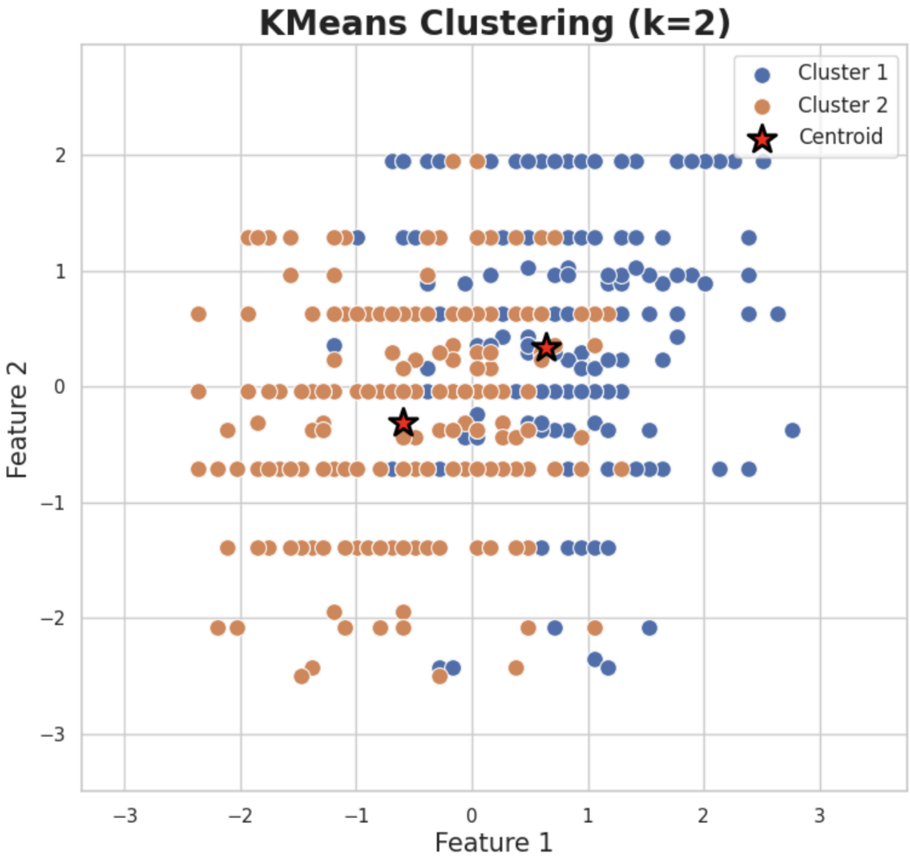

Figure 4. Visualization result of KMeans with k =2

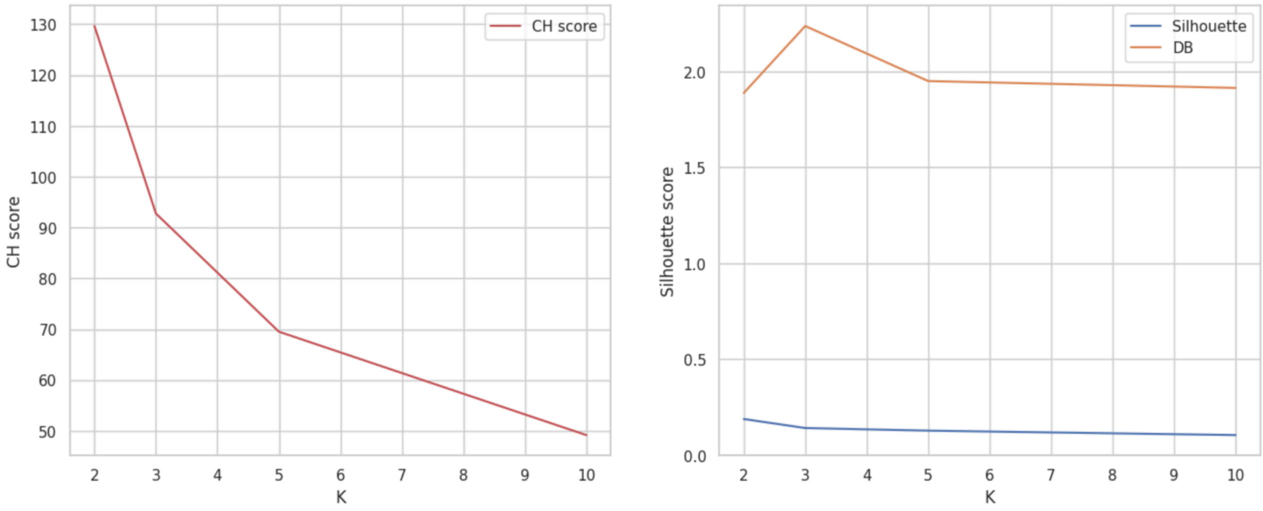

Figure 5. Evolution of CH score and Silhouette score between k = 2 and k = 10