Machine Learning Project
January 2024 Exam session

Cardiovascular diseases: a predictive analysis of the phenomenon

Team 10: Alessandro Bosi[1], Denis Bugaenco[1], Eleonora Zullo[1]

Abstract

Cardiovascular diseases represent one of the most pressing global health challenges in the 21st century, constituting the leading cause of mortality worldwide. But can machine learning contribute to the detection of this type of diseases? How effective are machine learning models in assessing an individual's likelihood of developing cardiovascular diseases?
The analysis aims to examine and identify the risk to detect a CVD using KNIME software, developing a predictive model based on a variety of clinical and non-clinical features. After an initial data cleaning, various supervised classification models were analyzed to obtain a prediction as accurate as possible regarding the probability of a correct diagnosis.

Keywords:
Machine Learning – Cardiovascular diseases – Clinical Risk Management

[1] Università degli studi di Milano Bicocca, CdLM Data Science

Contents

Introduction

Heart attacks, ischemic events and strokes are among the leading causes of death in Italy and worldwide. But what do they have in common? All three of them fall under the category of cardiovascular diseases, a pathological process affecting the heart and blood vessels, leading to the progressive narrowing of arteries.

According to a report from the American Heart Association, this type of disease affects approximately 471 million people worldwide, causing 17.6 million deaths per year. The most affected countries are the ones with low-middle income, due to a difficult access to effective and equitable healthcare, where a sedentary lifestyle and an unhealthy diet have become more prevalent [1]. The information gathered from various studies suggests that lifestyle factors, including variables such as physical activity, diet, and smoking, have a real impact on the likelihood of developing these diseases. But is it true?

In this context, the present study aims to outline an advanced classification model based on machine learning techniques to predict the probability of an individual developing cardiovascular diseases. Through an accurate analysis, our goal is to provide a diagnostic tool that assists healthcare professionals and spread awareness for cardiovascular risk.

To introduce our analysis, we decide to organize the report as follows:

1. **Data exploration**, where we present and examine the dataset used and the features that compose it.
2. **Data preprocessing**, where we prepare the dataset for the analysis considering the presence of missing values and duplicate rows, transforming and removing some wrong variables values, aggregating some variables to obtain less attributes in our data and normalizing their scales.
3. **Models and comparison**, where we show the different models used to predict the target variable, *cardio*, and the method used to select the best one.
4. **Evaluation**, where we test and evaluate the chosen model.

1. Data exploration

The dataset used to reach the main goal of the analysis is the *Cardiovascular Disease dataset*, provided by Svetlana Ulianova and available on the Kaggle platform [2]. It is a collection of patients data that can be organized in three different types of features:

- **objective features**: the factual information of the considered patient,
- **examination features**: results of medical examination on the patient,
- **subjective features**: information given by the patient.

More precisely, it consists of 70,000 records, representing 70,000 different patients, each with the following 13 features:

- *id*: the id number of the row;
- *age* (objective feature): age of the patient in days;
- *gender* (objective feature): binary variable for the gender of the patient, where 1 means female and 2 means male;
- *height* (objective feature): height of the patient in cm;
- *weight* (objective feature): weight of the patient in kg;
- *ap_hi* (examination feature): systolic blood pressure in mm hg, i.e. the pressure the blood is exerting against the artery walls when the heart contracts;
- *ap_low* (examination feature): diastolic blood pressure in mm hg, i.e. the pressure the blood is exerting against the artery walls while the heart muscle is resting between contractions;
- *cholesterol* (examination feature): cholesterol level in patient blood, where 1 means normal, 2 means above normal and 3 means well above normal;
- *gluc* (examination feature): glucose level in patient blood, where 1 means normal, 2 means above normal and 3 means well above normal;
- *smoke* (subjective feature): binary variable stating whether the patient is a smoker or not (0 means “not smoker”, 1 means “smoker”);
- *alco* (subjective feature): binary variable stating whether the patient is an alcoholic or not (0 means “not alcoholic”, 1 means “alcoholic”);
- *active* (subjective feature): binary variable stating whether the patient is involved in physical activities or not (0 means “not involved in physical activities”, 1 means “involved in physical activities”);
- *cardio* (target variable): binary variable stating whether the patient has cardiovascular disease or not (0 means “not have cardiovascular disease”, 1 means “have cardiovascular disease”).

It can be easily noticed that the variable *cardio* is the main feature of the dataset. The purpose of the analysis is in fact, as we already mentioned, predict the predisposition of a patient to detect a cardiovascular disease.

Therefore, *cardio*, which expresses the presence or the absence of the CVD in the initial dataset based on the characteristics of the considered patient, can be exploited to predict the probability of cardiovascular diseases in the machine learning models.

It is important to underline that *cardio* is suitable for this purpose since it is well balanced: the patients without cardiovascular disease, where *cardio* is equal to 0, are in fact the 50.02% of the whole dataset, while the patients with cardiovascular disease, where *cardio* is equal to 1, are 49.98%.

2. Data preprocessing

After the data exploration process, the analysis proceeds with the cleaning and the preparation of the dataset.

In order to make the dataset more suitable for the subsequent processes, the following techniques have been implemented.

2.1 Missing and duplicate values

The first step of the preprocessing is to verify the presence of missing values in the dataset, to decide how to deal with them. In the analysis, it has been decided to remove all the rows in which a missing value exists, as the dataset is composed by 70,000 records. However, no missing values are detected in the data.

Subsequently, a similar analysis has been done for the duplicate rows. The duplicate rows have been selected and then removed since they do not contribute to the prediction analysis and they may also alter it.

2.2 Variable conversion

After inspecting the data format of the variables, the variable *age* results to be expressed in days.

To make the information more comprehensible, the variable has been transformed in years dividing each value by 365.25, considering the presence of leap years, and rounding half down the result. The obtained values are collected in the new variable *age_y*.

2.3 Processing ap_hi and ap_low variables

To fulfill a proper work on the attributes related to the systolic blood pressure (*ap_hi*) and to the diastolic blood pressure (*ap_low*), it is important to consider some verifiable medical information [3]:

- diastolic and systolic blood pressure can only register positive values,
- since the diastolic blood pressure represents the pressure of blood when the heart muscle is resting and since the systolic blood pressure represents the pressure of blood when the heart muscle is contracting, we should not have values of diastolic blood pressure which are higher than the systolic one.

Therefore, this step of the data cleaning simply consists in the removal of all the records containing a negative value for the *ap_hi* or for the *ap_low* variables and in the removal of all the records where the diastolic blood pressure has a higher value than the systolic blood pressure.

2.4 Variable creation

Successively, we decided to combine the variables *weight* and *height* into a new variable *BMI*. The *BMI* – Body Mass Index – can be in fact simply calculated as the ratio between the weight and the square of the height.

The purpose of this operation was in fact to verify the possibility to get additional insights from our updated data and, if necessary, to remove *weight* and *height* in order to reduce the dimensionality of the initial dataset. We opted to keep all the variables until the feature selection phase to select with more precision which variables influence effectively *cardio*.

2.5 Outliers

In our analytical journey, we employed the boxplot and the Interquartile Range (IQR) to identify and eventually eliminate outliers from our dataset.

The boxplot, with its visual representation of the data's central tendency and dispersion, provides in fact a comprehensive overview of the distribution to detect anomalous data points. Subsequently, by calculating the IQR — a measure of statistical dispersion between the first and third quartiles — we established a robust criterion for identifying outliers.

Armed with this information, we systematically removed data points lying beyond the acceptable range determined by the IQR. After various analysis, we decided to use 2 as IQR multiplier, denoted as *k*, to scale the range. We noted that the chosen *k* is the ideal compromise to remove unrealistic values, while preserving data integrity and representing real-world variability.

To better clarify the process for the removal of the outliers, we give an example showing what has been done with the variable *weight*. The boxplot providing the initial distribution of the variable is shown in the figure 1.

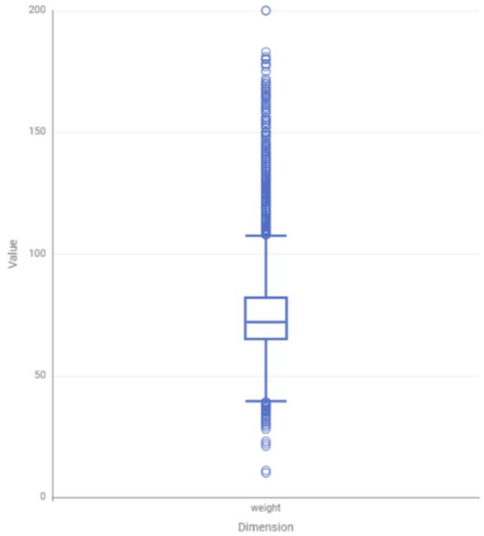


FIGURE 1 - INITIAL BOXPLOT FOR *WEIGHT*

It can be observed that there is a strong and evident presence of outliers for the variable *weight*. However, if we focus on the values of each outlier, it can be easily noticed that, even if not so common, there could exist patients with a weight ranging from, for example, 110 kg to 120 kg or from 45 kg to 35 kg. On the other hand, a patient with a weight of 11 kg, the least outlier, or with a weight of 200 kg, the greatest outlier, cannot exist in this type of dataset – where we have patients with an age ranging from 30 to 65.

For this reason, we decide to do not eliminate all the outliers – all the data points located outside the whiskers represented by the IQR – but to maintain the ones located immediately after the whiskers. This has been possible simply choosing the IQR multiplier, *k*, equal to 2. The result of the operation on the variable *weight* is reported in the figure 2.

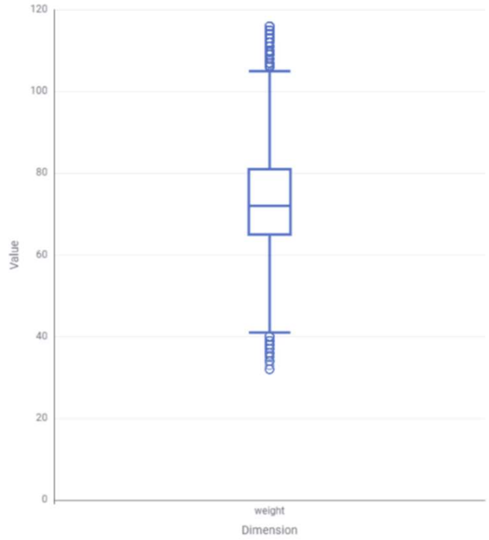


FIGURE 2 – FINAL BOXPLOT FOR *WEIGHT*

This approach not only enhanced the reliability of our analysis but also ensured that our conclusions were based on a more accurate representation of the data.

2.6 Normalization

By employing normalization techniques, we systematically transformed the attributes of the dataset to a standardized range, between 0 and 1.

This process is particularly useful when dealing with features that exhibited disparate scales or ranges. Normalization ensures that each feature contributes proportionally to our analysis, preventing any unexpected influence from variables with larger numeric values.

At the end of the preprocessing phase the resulting dataset, initially composed by 70,000 records, has been reduced to 66,8116 records.

2.7 Feature selection

Finally, to optimize the performance and the efficiency of our machine learning models, we recognized the importance of identifying and retaining only the most informative and relevant attributes of the dataset.

This is possible thanks to the feature selection process that, through the correlation matrix, can underline which attributes linearly affect the target variable of interest.

Feature selection aims in fact to enhance the model's predictive accuracy, to reduce computational complexity and to mitigate the risk of overfitting.

The correlation matrix obtained from our data, which shows the linear relationships between the different variables, is the one in the figure 3.

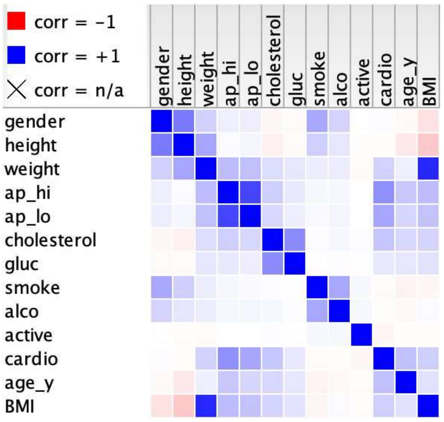


FIGURE 3 - CORRELATION MATRIX

As the matrix shows, among all the attributes, only five of them linearly affect the variable *cardio* and, therefore, are truly relevant to develop a proper predictive model: *ap_hi*, *ap_lo*, *cholesterol*, *age_y* and *BMI*.

3. Models and comparison

The data used for the realization of the predictive models are the results of a strategic partitioning of our original dataset into distinct training and test sets using the stratified sampling technique based on the target variable *cardio*. More precisely, this division, commonly known as the 70-30 split, involves allocating 70% of the data for training the machine learning models and reserving the remaining 30% for testing its predictive capabilities.

In fact, the training set serves as the foundation upon which the model honed its parameters and learns the underlying patterns, while the test set functions as an independent benchmark, evaluating the model's ability to make accurate predictions on new and previously unseen instances.

Therefore, using only 70% of the data and exploiting the attributes selected in the feature selection process, we decided to implement three different types of classification models to predict the target variable *cardio*: the Random Forest, the Logistic Regression and the Naive Bayes model.

3.1 Random Forest

Random Forest is an ensemble learning technique that operates by constructing a multitude of decision trees. The "forest" built by the model is a collection of decision trees usually trained with the bagging method; this technique creates multiple subsets of the training dataset by randomly sampling with replacement.

Each subset is used to train a separate decision tree and allow us to break down information into multiple variables to arrive at a singular best decision to a classification problem.

For each decision tree, only a random subset of features is considered at each split. This introduces diversity among the trees and helps prevent overfitting to specific features.

3.2 Logistic Regression

Logistic Regression is a binary classification algorithm that models the probability of an instance to belong to a particular class. Hence, despite its name, Logistic Regression is used for classification rather than regression.

More precisely, the output of the model is interpreted as the probability that the given instance belongs to the positive class: if the output is close to 1, it is predicted to belong to the positive class; if the output is close to 0, it is predicted to belong to the negative class.

In the specific case of our analysis, the output can be read in the column *prediction(),* where we can see the probability for each patient to detect a cardiovascular disease or not.

3.3 Naive Bayes

A third classification method that we decided to use is the Naive Bayes algorithm: it is based on conditional probability, the likelihood that an event occurs, considering that a second event has already occurred, and on Bayes' theorem, whose formula is now provided:

$$P(C|X) = \frac{P(X|C) \cdot P(X)}{P(X)}$$

In this case, we are considering the probability that a target variable C is predicted correctly given a set of features X. The Naive assumption is that the features are independent given the classes. That being said, we can rewrite the theorem as follows:

$$\frac{P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C) \cdot P(C)}{P(X)}$$

where $x_1, x_2, \ldots, x_n$ are the features of the instance to be classified.

3.4 Cross Validation

To select the best model for our purposes between the three just presented, we exploited the Cross Validation technique.

We decided to use the Cross Validation as it is a fundamental technique in machine learning to assess the performance and the generalizability of a model.

In fact, by dividing the data into multiple subsets and iteratively training the model on different combinations of training and test sets, Cross Validation provides a more robust estimate of the model's accuracy. In essence, it serves as a valuable tool for model selection, for evaluating the stability of the parameters and for ensuring the reliability of the models even when different data are used as input.

Moreover, we decided to use the Cross Validation rather than the Holdout technique as the first one is less sensitive to the randomness in the data split compared to Holdout validation, fundamental aspect considering the dimension of the dataset.

In particular, we decided to use the 10-fold cross validation: the dataset was split in ten subsets and the models were trained ten times, using, at each iteration, nine parts as training set and one part as test set and providing a specific level of accuracy for each iteration. The output of this process is, indeed, a collection of ten different values of the accuracy for the three different models.

Successively, we firstly computed a mean for the ten values of the accuracy for each model, to have an initial preview of our model's performance, and then, we decided to make a more precise performance evaluation exploiting the boxplot.

We joined in fact the ten accuracies for the three models to realize a boxplot for each model and then, we compared the three obtained boxplots, shown in the figure 4.

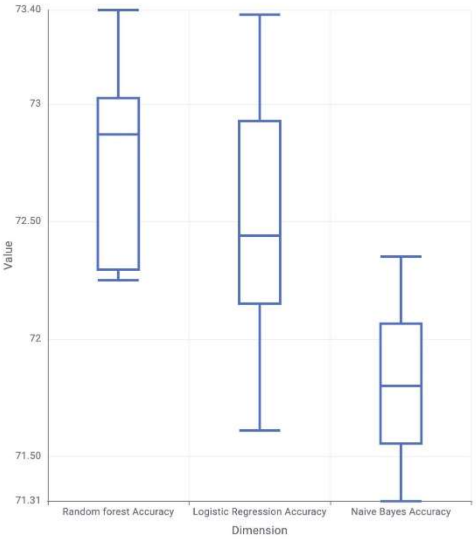


Figure 4 - Boxplots for the classification models

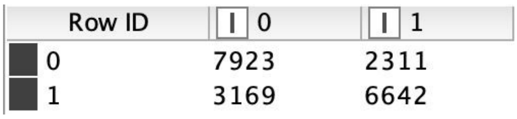
It can be easily noticed that Random Forest model registers the highest value for the accuracy, with a mean of 72.778.

More specifically, considering the ten iterations during which the model has been trained, the maximum value of the accuracy recorded is 73.4 and the minimum one is 72.25. Therefore, this model has been selected as the best one to achieve the initial task. To confirm our choice, we evaluate deeper the model and the result is shown in the following paragraph.

4. Results and evaluation

After the selection of the Random Forest as the best model to predict and classify the variable *cardio*, we decided to test it, using the 30% of the data from the partitioning, and we evaluated more precisely its performance.

Firstly, as the accuracy has already been analyzed, we focus on the confusion matrix for the model, shown in the figure 5.



Row ID	I 0	I 1
■ 0	7923	2311
■ 1	3169	6642

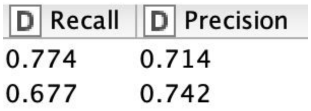
Figure 5 - Confusion matrix for Random Forest

The confusion matrix indicates that:

- 7923 values have been predicted as negative and they are effectively negative (TN),
- 6642 values have been predicted as positive and they are effectively positive (TP),
- 3169 values have been predicted as negative, but they are positive (FN),
- 2311 values have been predicted as positive, but they are negative (FP).

In other words, the chosen model has correctly predicted 14,565 observations over 20,045 – our test set.

Secondly, we analyzed the recall and the precision.



D Recall	D Precision
0.774	0.714
0.677	0.742

Figura 6 - Recall and Precision for Random Forest

In a binary classification model, we can compute two different values each for the precision and for the recall, referring to both the positive and negative classes. In this way, we obtain a more detailed vision of the model performance.

More precisely,

- 0.774 represents the negative recall, namely the ratio between TN and TN+FP,
- 0.677 represents the positive recall, whereby the ratio between TP and TP+FN,
- 0.714 represents the negative precision, the ratio between FP and FP+TN,
- 0.742 represents the positive precision, the ratio between TP and TP+FP.

Our analysis aims to find, among the patients, those who could suffer from a cardiovascular disease.

Therefore, the most important values to focus on are the positive recall, 0.677, and the positive precision, 0.742.

The precision measures in fact the proportion of the true positive predictions made by the model among all instances predicted as positive, that is how accurate the presence of cardiovascular disease prediction is. On the other hand, the recall measures the proportion of true positive predictions among all actual positive instances.

Finally, to properly evaluate our model we chose the Receiver Operating Characteristic (ROC) curve, a graphical tool used to assess the performance of a binary classification model.

The curve illustrates the relationship between the positive records correctly predicted by the model, also known as True positive rate (TPR), and the wrongly predicted positive records, also known as false positive rate (FPR). The curve is compared with a diagonal, which represent the prediction of a random model. The bigger the area delimited by the two curves, the better the performance of the model will be.

The ROC curve designed for our model is shown in the figure 7.

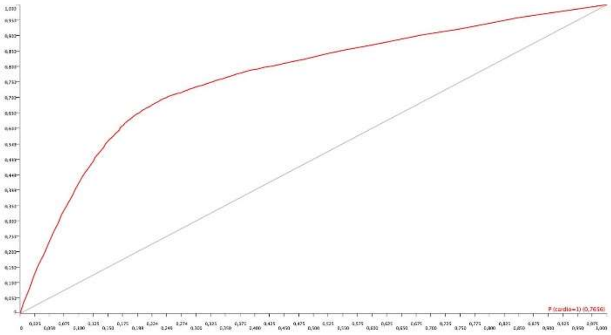


FIGURE 7 - ROC CURVE FOR RANDOM FOREST

The area under the curve, AUC, is an effective way to summarize the overall diagnostic accuracy of the test. It takes values from 0 to 1, where a value of 0 indicates a perfectly inaccurate test and a value of 1 reflects a perfectly accurate test.

In general, an area of 0.5 suggests a scenario where the test is not able to distinguish between individuals with the CVD condition, *cardio* = 1, and those without it, *cardio* = 0.

Since the area under the ROC curve is equal to 0.7656, we can declare that the chosen model gives an acceptable result. The AUC value indicates, indeed, that the cardiovascular disease test has moderate discriminatory power, offering some useful information for making predictions but there is still room for improvement.

Conclusion

In conclusion, our machine learning project has shown promising results in predicting the presence of the disease. The model achieved good performance in distinguishing individuals with cardiovascular disease from those without. However, it's crucial to acknowledge that the algorithm is not perfect, and there is still room for improvement.

While our model demonstrated effectiveness, we recognize the importance of exercising caution in its deployment. Given the critical nature of cardiovascular health and the potential impact on individuals' lives, we must prioritize safety and accuracy.

The model's current performance is promising but not flawless, emphasizing the need for ongoing refinement and enhancement. Therefore, healthcare practitioners may need to consider other factors in conjunction with the test results for a comprehensive assessment of cardiovascular health.

By using the model as a support tool, practitioners can benefit from its predictive capabilities while exercising their judgment and experience to make well-informed decisions.

References

[1] https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2]https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[3] https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

National Library of Medicine, National Center for Biotechnology Information:
https://www.ncbi.nlm.nih.gov/books/NBK535419/

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar; *Introduction to Data Mining (Second Edition);* 2018

Kevin P. Murphy, *Machine Learning A Probabilistic Perspective*, 2012