UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Selection of predictors for peripheral arterial disease using tree-based algorithms

Author:
Margarida GONÇALVES

Supervisor:
Dr. Carles CASACUBERTA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2023

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Selection of predictors for peripheral arterial disease using tree-based algorithms

by Margarida GONÇALVES

The purpose of this thesis is to collaborate with clinicians in order to enhance knowledge of peripheral arterial disease (PAD) by leveraging machine learning techniques to select variables sharing the strongest association with PAD among a set of predictors from a recent cross-sectional medical study carried out in Barcelona (Gonçalves-Martins et al., 2021). We built several machine learning models using Random Forest, Gradient Boosting Tree, and Extreme Gradient Boost classifiers to retrieve risk factors, of which Random Forest was the most efficient. Risk factors were obtained using the Shapley Additive Explanations' (SHAP) library. Results were compared with the known outcome of the logistic regression model used in Gonçalves-Martins et al., 2021.

We were able to replicate the main results of this study, as well as to discover new nuances of the factors that play a role in the development of PAD. Consistently with the above-mentioned study, the smoking habit was found to be a strong predictor for PAD both in women and in men, whereas hypertension was found to be a strong predictor for PAD in women, whereas diabetes was found to be a strong predictor for PAD in men. Surprisingly, dyslipidemia appeared to be negatively correlated with PAD. Furthermore, cholesterol levels and blood pressure levels could be unreliable for an analysis of risk factors for PAD, due to the effect of medication. Among our findings, we discovered that REGICOR scores are most consistent when their continuous value is used, and that history of cardiovascular events is especially influential on PAD in men. In addition, abdominal perimeter proved to be more efficient in general, but especially for women, in the prediction of PAD and discernment of its risk factors for PAD than body mass index and obesity.

Acknowledgements

I am profoundly grateful to my supervisor, Carles Casacuberta, for his support and guidance throughout this project. I extend my heartfelt appreciation to the teams at Hospital Vall d'Hebron and Hospital de Sant Pau for their generosity in providing the data and valuable assistance, especially to Dr. Teresa Puig, specialist in clinical epidemiology, and Dr. Sergi Bellmunt, specialist in vascular surgery.

I would also like to thank Professor Polyxeni Gkontra and Marina Camacho for their insights and knowledge, which were essential for the development of this project.

I am indebted to my mother and sister for their unwavering support and unwavering belief in my abilities.

Finally, I extend my deepest gratitude to my friends, whose contributions were vital in completing this journey.

Contents

Chapter 1

Introduction

1.1 A brief insight on peripheral arterial disease

Peripheral arterial disease (PAD) is the main manifestation of atherosclerosis, which is a widespread condition that affects multiple artery types throughout the body, including the coronary, cerebral, and lower-extremity vessels. The causes for PAD differ from other atherosclerosis diseases, revealing the need to have clear diagnosis methods and approaches for this condition. At first glance, this reasoning is only logical, as it would be for any type of health condition. However, overall awareness regarding PAD is very limited, despite the fact that there are more than 230 million diagnosed patients (Kuijk et al., 2010). The PAD pandemic, as described in Alushi et al., 2022, counts with a higher mortality rate due to cardiovascular events (e.g. myocardial infarction and stroke) and other ischemic events (P.Marso and R.Hiatt, 2006), when compared, for example, to the known HIV pandemic that affected roughly 38.4 million patients at the end of the year 2021, according to the World Health Organization (WHO) (WHO, 2022). Studies have long shown that PAD-positive patients tend to be at higher risk to suffer from carotid artery stenosis (CAS) and cerebral infarction (CI) than the general population (Araki et al., 2012).

In Europe alone, CV diseases are responsible for more than 4.35 million deaths per year. The concern about CV diseases and their impact on quality of life should extend to PAD prevalence since individuals with PAD face a higher risk of suffering cardiovascular ischaemic events (Stehouwer et al., 2009).

PAD is originated from the buildup of plaques (composed of fat, cholesterol, and other substances) in the inside of the arteries, which causes them to dilate as much as possible to preserve the blood flow. However, when dilating is no longer viable, the blood flow shifts to nearby smaller arteries. Nevertheless, these networks of arteries are unable to support the original blood flow going through the main artery, resulting in a mismatch between the necessary blood flow for the normal functioning of muscles and the blood flow that the arteries can deliver. Quantitatively speaking, a 50% decrease in the artery's diameter causes a loss of around 75% of cross-sectional area (Zemaitis et al., 2022).

Essentially, the main hallmark of PAD is the blood restriction that occurs at the body's extremities, resulting in increased walking or running difficulty, although 60 to 70% of PAD-positive individuals do not present this symptom. This might be due to the presence of other diseases that mask these symptoms, such as arthritis, which is known to cause similar pain. Another common reason for this percentage's unexpectedly high value comes from the fact that a lot of undiagnosed PAD-positive individuals naturally avoid getting involved in physical activity that might provoke this pain. We can conclude that the rate of asymptomatic individuals who have PAD

does not depict an accurate image of reality, since there can be a high percentage of those who have PAD but are unconsciously ignoring the symptoms or misattributing them for other diseases (Levine, 2018). When the hallmark symptom is present, PAD individuals are unable to walk a couple of blocks without stopping to rest, due to the failure of the blood supply at meeting the muscles' demands, ultimately causing pain, cramping, fatigue, discomfort, or weakness (Criqui et al., 2021). Naturally, it is a disease whose risk increases substantially with age (Shu and Santulli, 2018).

In more severe and rare cases of PAD, the collateral blood flow is so restricted and unable to comply with the blood supply the lower extremities need, that symptoms remain even after having ceased any type of physical activity, allowing ischemic ulcers to develop, causing tissue loss or even gangrene, leaving around 3-4% of PAD-positive patients with no other option but to have the affected area amputated (Shu and Santulli, 2018).

PAD in the lower extremities is diagnosed through resting ankle-brachial index (ABI) values. This index is the ratio of systolic blood pressure in the ankle and the higher of the two brachial artery pressures. The toe-brachial index (TBI) is used when there is suspicions of PAD, but the ABI index is not reliable, having a value of, for example, 1.40, due to non-compressible vessels, indicating artery calcification or arterial stiffness (Shu and Santulli, 2018). Table 1.1 depicts how the various ranges of resting ABI values are categorized.

TABLE 1.1: ABI index values categories

	Ankle-Brachial Index Rest
Non-compressible	> 1.40
Normal	> 1.00 - 1.39
Borderline	0.91-0.99
Abnormal	< 0.90

1.2 Common risk factors of PAD

The risk factors for PAD include:

- **Diabetes**

  Diabetes is one of the strongest risk factors for PAD (Song et al., 2023). Individuals with diabetes are more prone to suffer from PAD, which behaves differently and more aggressively in those with Diabetes Mellitus. The extent of PAD is positively correlated with the severity of diabetes, and the disease itself progresses faster in individuals with a diabetic profile. Moreover, insulin resistance is an additional risk factor for those with diabetes since it could trigger other cardiovascular events (P.Marso and R.Hiatt, 2006), an effect that might heavily be exacerbated by the presence of PAD. In the USA, the American Diabetes Association (ADA) has even gone so far as to create guidelines for diagnosis as well as management of PAD in patients with diabetes.

- **Smoking**

  Smoking is considered the most important and preventable factor from all risk factors of PAD. Not only active smokers are at a much higher risk of developing PAD, but passive smokers generally also present a higher likelihood of having PAD (Wang et al., 2021). Not only cigarette smoking is responsible for increasing the mortality rate of coronary disease by over 70%, but when combined with PAD smoking might also very well aggravate the extent of its complications (Lakier, 1992).

- **Obesity**

  A body mass index over 30 represents an increased chance of developing PAD. However, the interesting study of Lin et al., 2022 combined multiple other studies regarding the mortality of PAD in both patients underweight and overweight and discovered that underweight patients had a higher mortality risk. This is called the *Obesity Paradox*, in which obese individuals present a higher survival rate than those who are normal-weight. Nonetheless, the presence of this paradox is not solidified within the PAD researcher community, albeit worthy of being highlighted. Generally speaking, obesity is undeniably an established risk factor for PAD. Nonetheless, PAD patients are typically sedentary due to this disease's symptoms, and high rates of obesity have naturally been reported in these subjects (Cronin et al., 2013). This raises the doubt of whether obesity is a strong risk factor for PAD or simply a major consequence.

  The abdominal perimeter, as well as Body Mass Index (BMI), are also frequently registered when diagnosing PAD.

- **High blood pressure**

  Multiple studies have repeatedly demonstrated the positive correlation between the presence of PAD and high blood pressure. The interesting study by Itoga et al., 2018 in which participated patients over 55 years old with a history of cardiovascular complications, had as a goal reanalyzing the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). Their investigation of the association of blood pressure with lower extremity PAD events showed that systolic blood pressure (SBP) $<$ 120 mm and SBP $\geq$ 160 mmHg were associated with high rates of PAD. Lu et al., 2020 presents the same findings and solidifies that diastolic blood pressure is less associated with PAD.

  Thomas Manapurathe et al., 2019 and Hsu et al., 2017 also illustrate this relation in PAD-positive patients, the former stating that SBP $\leq$ 120 mm puts them at higher risk for cardiovascular complications.

- **High cholesterol**

  Murabito et al., 2002, Krishnan et al., 2018 study and demonstrate the relationship between high-density lipoprotein cholesterol and PAD. Studies have also shown that when high in the general population, RC cholesterol - cholesterol in triglyceride-rich lipoproteins, consisting of low-density lipoproteins and intermediate-densit lipoproteins - the risk of developing PAD is five times greater than suffering from a myocardial infarction or ischemic stroke (Wadström et al., 2022). In the study performed by Araki et al., 2012, which mixed both diagnosed and undiagnosed PAD patients, high levels of total cholesterol

and LDL cholesterol were most prevalent in individuals with PAD. Interestingly enough, HDL cholesterol has been inversely associated with the risk of developing PAD (Aday et al., 2018).

Related to this effect of abnormally high cholesterol levels, Aday and Everett, 2019a, add that both abnormally high high-density lipoprotein cholesterol - also known as HDL cholesterol - and low levels of HDL cholesterol are associated with the risk of developing PAD. According to the authors, this risk is exacerbated by high triglyceride concentration. Moreover, the authors were able to find that, within the female population of their study, low levels of HDL cholesterol implied a greater risk for women of developing PAD than that of men. However, in their research, they were not able to determine any association between triglycerides and PAD.

The imbalanced relationship between levels of triglycerides, total cholesterol, HDL cholesterol, and low-density lipoprotein cholesterol is titled dyslipidemia (Manjunath et al., 2013) and constitutes a common risk factor for PAD, although Poussa et al., 2007 states that dyslipidemia is a risk factor that is not usually taken into account in vascular ambulatory clinics. They find that only 11.6% of PAD-positive patients were tested for dyslipidemia, with the probability of being tested ultimately dependent on prior vascular treatment. Aday and Everett, 2019b has also pointed out the positive correlation between dyslipidemia and PAD.

- **Increasing age and sex**

  In the extensive study of available literature, Song et al., 2019, found that the prevalence of PAD increases with age. Sampson et al., 2014 also corroborates the results of Song et al., 2019. Interestingly enough, they also discovered that PAD prevalence was lower in men than in women for the population under 75 years of age, having the reverse behavior for those over 75 years old. In the scarce studies and research on PAD, we often find both female and male subjects grouped according to age. This results in the general belief that PAD incidence is higher in men, as many studies seem to point out, and that being male constitutes a risk factor (Teodorescu, Vavra, and Kibbe, 2013). However, there are other studies that suggest otherwise (Pabon et al., 2022). Nevertheless, by combining the knowledge from Sampson et al., 2014, we can extrapolate that one of the reasons why several studies have registered a higher incidence of PAD in men might be due to the inadequate grouping of men and women of the same age. Furthermore, Pabon et al., 2022 defends that the delayed diagnosis of PAD in women might be due to how the sex chromosome complement affects lipid profiles and vascular health.

- **Family history**

  A family history that includes PAD constitutes a risk of developing other cardiovascular complications. Inversely, a family history of cardiovascular complications and/or peripheral artery is a great risk determinant of PAD. Valentine et al., 2004 has specifically found that first-degree relatives of individuals who have been diagnosed with PAD are more prone to developing cardiovascular diseases. They also focused on studying if the genetic factor was independent of the *a priori* healthy siblings' smoking habit, having detected no interaction between family history and smoking habit, with the help of multivariable logistic regression. Because PAD is a disease that typically appears

with age, family history proves itself to be a great determinant of PAD in young adults. Wassel et al., 2011 extended these conclusions to the severity of PAD.

1.3 Prevalence of PAD in general population

The prevalence of PAD differs from country to country. In the United States, PAD is quite common, albeit still strongly undiagnosed, with an overall prevalence of 4.3% in those over 40 years old, and 29% in those over 70 years old (Levine, 2018).

In Europe, however, asymptomatic PAD prevalence can reach up to 17.8%, as mentioned in the PANDORA study in Cimminiello et al., 2011. In this study, we can also find mentioned the lack of reported cases, which makes it extremely difficult to extract a sufficiently accurate description of the incidence of PAD in subjects considered not to be at risk for CV diseases.

In the exhaustive and extremely interesting research carried out by Song et al., 2019, who examine the impact of increasing age on PAD in high-income countries (HICs), and low-income and middle-income countries (LMICs), it was found that, although PAD prevalence consistently increased with age, the effect of increasing age on PAD development was higher in HICs. On the other hand, PAD prevalence for younger ages (between 40-44) was higher in LMICs.

Geographically speaking, the Western Pacific Region registered a higher incidence of PAD, as opposed to the African Region, Song et al., 2019 report. In 2015, 15 countries alone, a group in which China, India, the USA, and even Spain are included, make up more than two-thirds of the estimated global PAD cases.

1.4 PAD prevalence in Spain

In Spain, the ESTIME study (Blanes, Cairols, Marrugat, et al., 2009), which consisted of a cross-sectional study with 1324 randomly selected participants with ages in the range 55-84, confirmed a high prevalence of asymptomatic PAD, having a total prevalence of 8.03%. It was concluded that the most common profile for PAD-diagnosed subjects contained characteristics as being a male, diabetic, smoker, with a history of coronary heart disease, having higher systolic pressure, and higher triglyceride levels.

Velescu et al., 2016 claims that the incidence of PAD is slightly lower in Girona than in other Mediterranean populations, although these populations themselves already tend to have a lower prevalence of PAD. Ramos et al., 2009 registered an overall PAD presence of 4.5%, with an incidence of 5.2% in men and 3.9% in women.

However, it should be noted that PAD studies generally have very different characteristics, with the Ramos et al., 2009 study including participants from ages 35-79, and other studies exclusively considering participants of different age ranges. The nature of the participants is also varied, with some studies only focusing on already signaled patients with CV risks and other studies focusing on the general healthy population. We can conclude that apart from having little research available on PAD, they are also different in their nature, which is an added obstacle when trying to compare results and clearly define the behavior and consequences of PAD.

Chapter 2

Previous Study

The dataset used in this project stems from a single-center, population-based, cross-sectional study, part of a larger pilot screening program evaluating abdominal aorta aneurysm. The protocol was approved by the Ethics Committee of Hospital Vall d'Hebron with code PR(AG)221/2017 and by the Ethics Committee of Hospital Sant Pau with code IIBSP-AAA-2013-88. All patients signed informed consent to participate.

Their goal was to obtain the prevalence and risk factors for PAD in the selected participants, all healthcare cardholders of Barcelona Nord. It is important to note that all participants were 65 years old at the time of the study. Therefore, we will not analyze the *age* risk factor. By choosing this specific participant profile, the authors hoped to be able to define a standard profile and target individuals who are a match for an eventual screening. It is also worth mentioning that the age selected for this profiling was carefully thought out, considering the prevalence of PAD in multiple age ranges obtained by other studies conducted in the same region. Therefore, our results should be analyzed within the same scope.

2.1 Selection of participants

Participants who received a letter of invitation to participate in the screening were all women and men who were not institutionalized and were 65 years old. This amounted to a total of 2808 possible participants, which boiled down to a final number of 1173, after excluding non-eligible participants, such as those who had had amputations, and subjects who did not participate for other reasons (e.g. no response, declines, missed more than 3 scheduled appointments, others) (Gonçalves-Martins et al., 2021).

Participants were then invited to complete a questionnaire that included inquiries on: smoking habit, hypertension, dyslipidemia, diabetes, history of cerebrovascular disease, history of cardiac ischemia, chronic renal disease, and history of any aneurismatic diseases. Height, weight, waist circumference, and BMI were also retrieved, as well as the REGICOR score. Euro Qol 5 eq and the Goldberg scale (Spitzer et al., 2006) questionnaires were used to extract information on the quality of life, and anxiety and depression, respectively.

ABI was later calculated and its value was used to determine if the pathology was present. All those with ABI < 0.9 were considered PAD-positive.

2.2 Methodology

Gonçalves-Martins et al., 2021 indicate that a logistic regression model - fed by categorical and continuous variables - was used to obtain independent risk factors for PAD. The risk factors that were considered in this model were the ones previously considered relevant in a bivariate analysis that consisted of the exploration of all encoded categorical variables through Pearson's chi-square test or Fisher's exact test. Further information on the modeling of the data was not provided. Statistical significance is defined as a p-value less than 0.05, $p < 0.05$.

2.3 Results

Out of the total 1173 participants, 59.2% and 40.8% were men and women, respectively.

Out of all the participants, it was determined that PAD prevalence was 6.22% in general, with a higher percentage for men (7.91%) and 3.76% in females, having found that the difference between sexes was statistically significant. PAD incidence was also found to be higher in men with diabetes and smokers, female subjects with hypertension, and higher waist circumference, and smokers (Gonçalves-Martins et al., 2021). Furthermore, dyslipidemia was not found to be statistically significant in either sexes, although having a lower p-value for women. This behavior is also present for the variables that register past chronic renal disease, cardiac ischemia, and cerebrovascular events, with the latter having a lower p-value for men.

Interestingly enough, 48.8% of PAD-diagnosed participants were considered to be at low risk of developing a cardiovascular event, coming close to the 51.1% of healthy subjects who were considered to be of low-risk as well.

Overall, the proportions of patients at each score of the REGICOR scale are similar between non-PAD and PAD-diagnosed subjects. It is interesting to note that the REGICOR scores showed that 0% of PAD-diagnosed women are at a high risk of developing a cardiovascular event.

Gonçalves-Martins et al., 2021 have reported that they believe this sample to be a correct representation of Northern Barcelona's population.

Chapter 3

Methodology

We created several models using different classifiers: Random Forest, Extreme Gradient Boost, Gradient Boost, and for comparison purposes, Logistic Regression. The code for this project can be found in the corresponding Bitbucket repository for this project. Due to the confidentiality of the data, the Bitbucket repository can only be accessed with an invitation.

For this invitation, please contact mgoncago44@alumnes.ub.edu.com.

3.1 Training the models

One of the first concerns with this dataset was the high level of class imbalance. As mentioned before, only 6.22% of the dataset corresponds to positive participants, that is, participants ultimately diagnosed with PAD.

When building and training the models, we first did so considering the entire dataset. However, data augmentation techniques, such as SMOTE, ADASYN, SMOTENC, were ineffective. The initial metrics of the models were unsustainably low with the best F1-score, for example, reaching values up to 0.3. We could not improve the metrics despite changing the model's architecture by performing more splits, using different cross-validation techniques, or even combining data augmentation techniques.

We then tried randomly selecting samples from the dataset, although we recognize this is not the most effective nor consensual way of training a model. By selecting a random sample of 140 healthy participants, of which 70 were men and 70 were women, we were able to achieve good metrics. Having good metrics, although our main goal is not to extrapolate the models created, is just as important, as we are going to obtain the risk factors from the models created themselves.

After having seen that data augmentation techniques were not suitable for the dataset in question, we decided to downsample the majority class. However, we did so with the goal of creating a sample that was as uniform as possible. Initially, we identify the categorical variables and calculate the minimum number of samples required from each categorical to maintain as much balance as possible. By leveraging this information, we selectively sample data from each categorical feature in the dataset. For continuous variables, we binned the data using quantiles. To ensure a sample size big enough for the construction of the models and small enough for their good performance, we defined a target count of samples for each bin. Then, we extract the desired number of participants from each bin, considering the target count. The resulting sample, encompassing both categorical and continuous variables, represents a stratified subset of the original healthy dataset. From

this leveraged dataset, we perform the same process again separately for each sex, and ultimately randomly sample 70 female and male participants. Our final healthy dataset has 140 healthy participants and all 73 unhealthy participants.

Since this is not a standard way of building a sample, but rather a "tree" of binning and selecting random participants, we ran the models several times to ensure that the results were consistent and did not vary greatly according to the random sample taken. We ran all the models with 4 different random seeds for all the ablation studies discussed in 6 and obtained consistent results.

As mentioned before, although our main goal was never to predict PAD in other datasets, that is, our objective was not to extrapolate the model itself to other datasets, but only to understand what the relationship between risk factors and PAD is in the given population, since there is actually no other data to perform these tests on, bringing us back to the lack of research studies on PAD. Nevertheless, by not being able to reach decent metrics with the full dataset, we know that if the models we created are extrapolated to other datasets, the metrics obtained would not be satisfying.

In Figure 3.1, we can see the general scheme used for building the dataset.

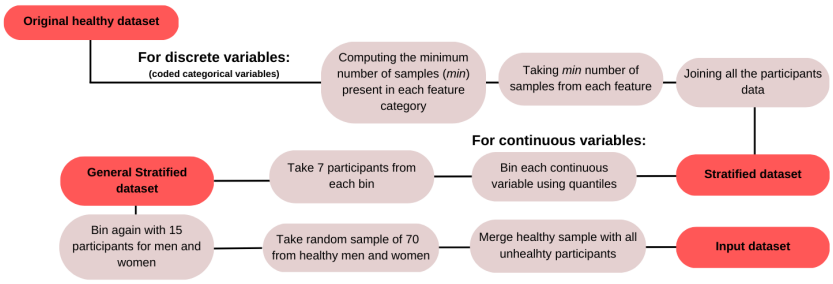

FIGURE 3.1: Scheme for building the sample dataset

For the training of the dataset itself, we tested different cross-validation techniques, such as K-fold, stratified K-fold, and repeated stratified K-fold. The latter generally gave us the best results. Inside each cross-validation, we also did a grid search for the generally considered most important parameters of each classifier. By combining the cross-validation with the grid search, we ended up doing an adapted nested cross-validation.

The imbalanced problem was generally fixed. We did slightly more repeats when the sex of the participants was specified, to ensure the existence of models with good

metrics. However, women's models were still not able to obtain nearly as good metrics as the models for men. Therefore, we applied SMOTE to women to balance their dataset. We tried to explore more innovative and recent data augmentation techniques, such as TABGAN (Ashrapov, 2020). This technique creates tabular data using a GAN. Although we saw a slight improvement in the metrics using this technique, it was not consistent. That is, the GAN always creates data that cannot be replicated. Considering our purposes, this was the main reason why we decided to stick with SMOTE and leave TABGAN behind. By using SMOTE for women, we were able to improve their models' metrics.

Nonetheless, there is the occasional case where a model trained and tested only on female participants cannot achieve an F1-score of at least 60%. For men, however, in our ablation studies, we were always able to obtain a significant amount of models with an F1-score of over 60%. This is due to the fact that we have more men with PAD than women, resulting in increased difficulty of the models performing well enough for women.

In Figure 3.2, we find a scheme of the training of the models.

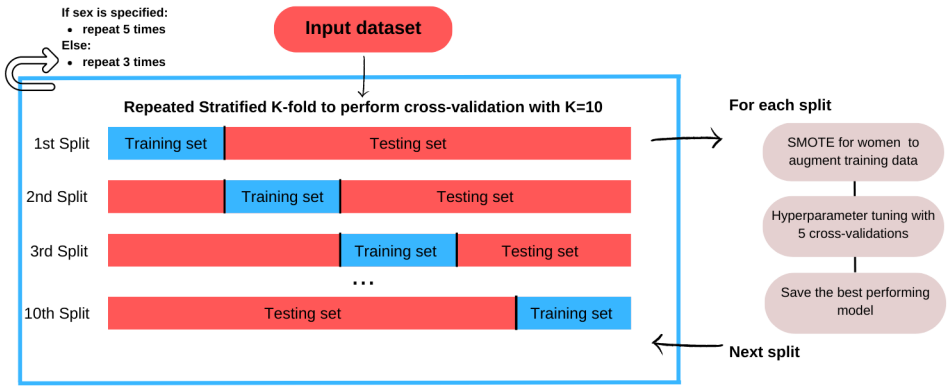

FIGURE 3.2: Scheme for building the sample dataset

3.2 SHAP for plotting risk factors importance

To extract the relationships between risk factors and PAD we used the Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017). SHAP, a method based on game theory, is currently a known explainability algorithm that allows one to explain individual predictions a model has made. Each feature is considered to be a

player, and the final prediction is the payout. To put it simply, SHAP calculates the contribution of each player, that is, each feature, to the output, which is a specific prediction.

SHAP library exists to deal with the problem of explainability and to transform the so-called "black box" into a "glass box" (Holzinger, 2018). It gives space for the use of more complex models in the medical and other fields where justifications and explanations are essential, by allowing us to access the way in which the model made the predictions and consequently letting us see the logical process behind the prediction of a specific data point.

In our study, we considered only the models that had an F1-score above 60%. We plotted the SHAP summary plot for each of these models, which gives us general information on how the values of a certain feature influence the increase or the decrease of the probability of a data point being predicted as a part of the positive class. We then carefully examined the SHAP plots generated and registered the results and conclusions of this analysis in Chapter 6.

Chapter 4

Proposal

In this project, our main goal was to replicate, and, if possible, add more insights to the conclusions achieved by Gonçalves-Martins et al., 2021 by making use of machine learning (ML) algorithms and techniques. Our research aims to address a significant gap in the knowledge of PAD and to offer valuable insights, even if it is through a small incremental advance.

4.1 Logistic Regression in the medical field

In the medical field, logistic regression (LR) models are typically preferred (Schober and Vetter, 2021). Even though consensus on what methods perform best does not exist, LR allows for simple and arguably interpretable multivariate analysis (Shipe et al., 2019).

Some of the problems that might arise with LR models derive from the fact that regression models assume the relationship between the independent and dependent variables is uniform. This means that the behavior of the impact of a risk factor in the independent variable is linear - it follows a specific direction. However, this assumption can constitute the downfall of regression algorithms in the medical field, since complex diseases cannot be modeled considering linear assumptions (Ranganathan, Pramesh, and Aggarwal, 2017).

This is why, despite having the possibility to recreate the LR analysis made by Gonçalves-Martins et al., 2021 using another approach, our interest was to introduce more complex models, and later on, apply other ML techniques to allow for these models' explanations and interpretations. Nonetheless, we still considered LR models in our study, from an ML perspective, rather than considering the standard statistic procedures.

4.2 Proposed models

Our suggestion is to explore tree-based algorithms to understand how the risk factors present in our dataset influence the probability of having PAD, according to each tree-based algorithm. It is important to notice that our main goal does not include extrapolating the created models to similar datasets. Primarily because due to the high data imbalance - total PAD prevalence is 6,22% - and the erratic distribution of the variables across PAD-diagnosed and not PAD-diagnosed subjects, this extrapolation would not be successful, but also because there are no other similar datasets that we can extrapolate to, simply due to the lack of study of PAD. This was also why we did not use an external validation set when training the models. More details on the imbalance of the data can be found in Chapter 5.

4.2.1 Tree-based algorithms

Opposite to LR algorithms, tree-based algorithms do not assume a linear relationship between features and target variable(s). At its core, tree-based algorithms are combinations of multiple decision trees. A decision tree is depicted as a visual representation with a tree-like structure. At each node of the tree, a specific attribute value is tested, and the branches represent the outcomes of those tests. The leaves of the tree contain the final classification for a subject.

Random Forests (RF) consist of multiple decision trees combined through a bagging method (Hnat, Veselka, and Honek, 2022). Firstly introduced by Breiman, 2001 almost two decades ago, Random Forests, although arguably simple, are one of the trustiest and most popular ML models. They consist of an ensemble of decision trees: many decision trees are built randomly through various data points, which ultimately come together as a forest. A sample is classified by the majority of votes, that is, each individual decision tree has its say in the prediction and the majority vote wins. To prevent overfitting in relation to the predictors that are highly correlated with the target variable, RF randomly selects subsets of predictors for each data point. RF mantain its fame due to its high accuracy, simplicity, and ability to work with large datasets and large quantities of predictors.

Gradient Boosted Trees (Friedman, 2001), however, develop decision trees sequentially using gradient descent as a loss function, with the final prediction being the result of a weighted majority vote. Each tree is consequentially improving the predictive power of the former, with the ultimate prediction for a data point being that of the last decision tree in this ensemble.

XGBoost (Chen and Guestrin, 2016), is an improvement of Gradient Boosting Trees, which use regularization techniques, resulting in general better performance and lower computational time.

Multiple papers in the medical field have incorporated ML algorithms such as decision trees in their research, especially in cardiovascular-related research, such as Budholiya, Shrivastava, and Sharma, 2022, who develop a model with Extreme Gradient Boosting (XGBoost) as the base for predicting heart diseases, and Jiang et al., 2021, who used algorithms such as XGBoost and Random Forest to triage patients with higher cardiovascular disease risk.

Xie et al., 2021 used Random Forest, XGBoost, and LR to create a model for early prediction of left ventricular reverse remodeling (LVRR) - which is defined as a decrease in the volume of the ventricular chamber that has been associated with improvement in systolic and diastolic functions of the heart, resulting in a lowered probability of suffering a fatal cardiovascular event (Hnat, Veselka, and Honek, 2022) - in patients with idiopathic dilated cardiomyopathy. In their work, the tree-based models, especially XGBoost, were able to detect early LVRR, as well as differentiate correctly between LVRR patients and non-LVRR patients, reporting an AUC of 0.8205 as opposed to the LR's AUC of 0.5909.

Dinh et al., 2019 also used LR, Random Forest, and Gradient boosting to predict diabetes and cardiovascular disease, and simultaneously select the most relevant risk factors by building a weighted ensemble model. This ensemble model achieved a maxima AU-ROC of 83.9%, and the XGBoost model alone was able to have an AU-ROC of 84.4% for the pre-diabetic patients.

Outside of the cardiovascular scope, we find multiple studies that incorporate decision trees in their methodology. Yoo et al., 2020 proposed a mix between deep learning and decision trees for COVID-19 diagnosis from chest X-ray imaging. The features used in the decision trees ensemble created were extracted using a deep learning model, and predictions were made by the implemented decision tree. Esmaily et al., 2018 use decision trees, not to predict, but to select risk factors for type-2 diabetes. Further use of decision trees combined with Syntethic Minority Oversampling Technique (SMOTE) in diabetes-oriented research is found in Mirza, Mittal, and Zaman, 2018. In Ghiasi and Zendehboudi, 2021, breast cancer predictions are made using Random Forest and other tree-based models. Shaikhina et al., 2019 makes use of Decision Tree and Random Forest classifiers to detect high-risk patients in kidney transplants.

Chapter 5

Data Overview and Preliminary Analysis

5.1 Participant's profile

As mentioned in Chapter 2, the dataset arises from the previous study we are trying to replicate. It was found that PAD is higher in women than in men, however, the distribution of the risk factors themselves differs when considering the sex of the subject.

We also discovered that, in our dataset, the majority of men were active smokers, as opposed to the majority of women, who were ex-smokers. Women were also the group that had more relatives with a history of AAA, with most of them being of second degree. For men, however, despite having fewer relatives with a history of AAA, most of them were of first degree.

Men had a higher prevalence of diabetes, hypertension, and chronic renal disease than women with a slightly higher prevalence of dyslipidemia. Men also had more records of a history of CV events.

The participant's profile should be considered throughout the research, as it plays an important role in interpreting the difference in risk factors for both sexes later on.

5.1.1 Data treatment and feature selection

Before applying any machine learning algorithms, we first cleaned the data, got rid of outliers and did some exploratory analysis to better understand the distribution of the risk factors across the subjects. It should be noted that all outliers present in the dataset were healthy patients. This means that no information was lost on PAD-positive patients.

We began by excluding those variables that were not risk factors, but instead were measures taken for the diagnosis of PAD itself. We also excluded irrelevant variables such as date of the examination, center of examination, date of birth, among others of the same category. Redundant variables were also excluded.

The following variables: family history of 1st and 2nd degree of cardiovascular diseases, and aneurysm type (if existent) were also not considered for this project. These variables were excluded due to their high rate of missing values as seen in Table 5.1.

TABLE 5.1: Missing values for family history and personal history variables

	Missing values
AF grau r01	1139
AF grau 02	1139
Grau AntFam	1139
tipus AA	1157
t AA r01	1157
t AA r02	1157
t AA r03	1157
malaltia arteri	1129

The existence of these many missing values is easily justified by the fact that these subjects were randomly selected from the general population (not diagnosed with vascular diseases). This dataset arises precisely from an invitation to participate in a screening program for cardiovascular diseases, which explains the lack of records on family history regarding CV diseases.

Imputing values

Other missing values were imputed using the KNN imputer, one of the most robust and popular methods for imputing data that has been used in several research papers in the medical field (Saranya et al., 2021, Pazhoohesh, Pourmirza, and Walker, 2019, Pujianto, Wibawa, Akbar, et al., 2019) due to its ability to handle mixed datasets - with both continuous and categorical variables -, and maintain data distribution.

The problem of multicollinearity

Regarding the remaining variables that constitute risk factors, we chose to not include all of them in our models due to the problem of multicollinearity. Multicollinearity appears when a subset of predictor variables is highly correlated, which can lead to misleading results (Daoud, 2017).

We computed the Variance Inflation Factor (VIF) for each variable to determine if there was multicollinearity. The VIF measures how much the variance of a predictor is inflated. When multicollinearity is present, the standard error of the predictors' coefficients increases, leading to a higher variance of the predictor's coefficients, revealing the need to quantify this inflation.

Generally speaking, if the VIF of a variable is higher than 5, then there is evidence of mild multicollinearity. VIFs higher than 10 define the presence of high multicollinearity. No model, ML or not, is multicollinearity-proof, which is why it is important to always consider removing highly correlated variables from a dataset before applying any algorithm.

However, we still have to maintain clinical interpretability. Therefore, we cannot blindly eliminate highly correlated features without considering the impact that this might have later on in our study.

Our problematic predictors, that is, those that are highly correlated between themselves, are BMI, obesity, and abdominal perimeter. These correlations are natural, as these variables are naturally dependent on each other's values. However, it would be unwise to introduce them all at once in our models, without considering the possible optimization of subsets of these problematic variables, and also the possible misleading results that could arise from the introduction of multicollinearity in the model.

We proceeded by conducting ablation studies regarding these variables - BMI (defined as *IMC* in our dataset), abdominal perimeter, and obesity -, in order to see how that affected the performance of the models and risk factors' detection.

Most importantly, since the majority of conflicting variables are continuous, we decided to further compare the predictive power of using the mixed dataset - with continuous and categorical variables - and using only categorical variables, as this was one of the primary goals of the study of Gonçalves-Martins et al., 2021.

The results from these studies can be found in Chapter 6.

5.1.2 Cholesterol behavior on PAD-diagnosed participants

We found that PAD prevalence was higher in participants with lower levels of cholesterol. Through basic exploratory analysis the results presented in Table 5.2 were obtained.

TABLE 5.2: Total cholesterol's description

	Healthy participants	**PAD participants**
Mean	204.943	192.534
Standard deviation	38.598	41.447
Minimum registered value	89	108
Maximum registered value	312	290
25% quantile	180	159
50% quantile	205	195
75% quantile	230	216

We can state that, in this dataset, the average total cholesterol levels are higher in healthy patients than in PAD-diagnosed patients. We assume that this is going to have an effect on the behavior of the variable *Dyslipidemia* when detecting risk factors for PAD.

Moreover, although the standard deviation is slightly higher in the PAD-positive group, it is still quite similar to the standard deviation of the healhty participants' group. Nonetheless, these values indicate that the behavior of the total cholesterol in the PAD-positive group is more erratic when compared to the behavior of this same variable in the other class. This is an interesting pattern to observe, especially when we see that the range of total cholesterol levels is higher in healthy participants.

While this may seem contradictory it actually indicates that, although the PAD-positive group may have a narrower range of total cholesterol values, the individual values within that range exhibit more variability compared to the healthy participants' group. This indicates that the total cholesterol levels in the PAD-positive

group are more diverse and less concentrated around the mean, which ultimately leads to a higher standard deviation.

To confirm the behavior of the *total cholesterol* variable, we plotted its density, which can be seen in Figure 5.1.

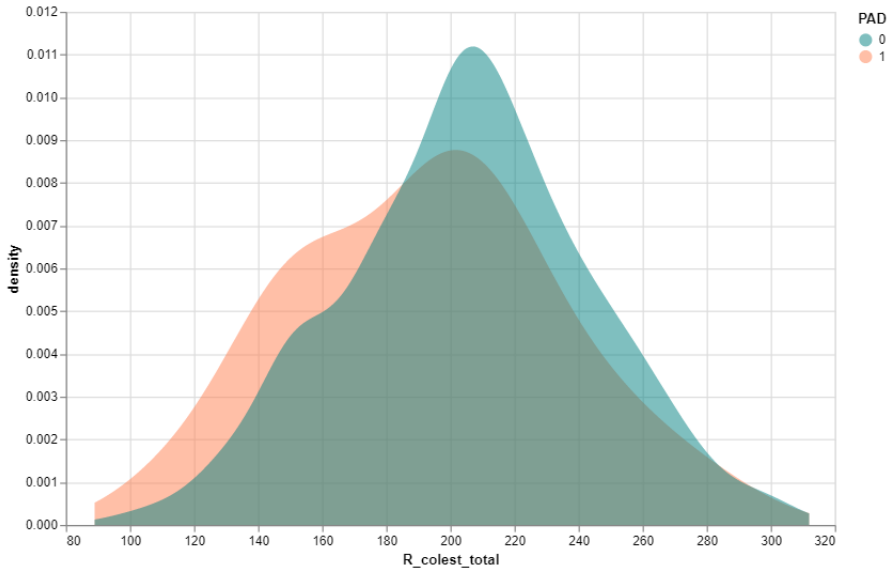

FIGURE 5.1: Density plot for total cholesterol

The density of PAD-positive participants, represented in the figure above by the color orange, is surprisingly higher for participants with the lowest total cholesterol values in the dataset. This behavior is also present when considering each sex.

On the other hand, the values for HDL cholesterol are almost identical between classes, as we can see from Table 5.3, with no apparent difference between the behavior of this variable.

TABLE 5.3: HDL cholesterol's description

	Healthy participants	**PAD participants**
Mean	54.674	52.54
Standard deviation	12.036	12.621
Minimum registered value	20	23
Maximum registered value	93	83
25% quantile	47	43
50% quantile	52	51
75% quantile	62	61

However, in the density plot (Figure 5.2) for HDL cholesterol, we clearly see that most of the participants that have lower levels of HDL cholesterol are those that have PAD, which is consistent with the known relationship between HDL cholesterol and PAD.

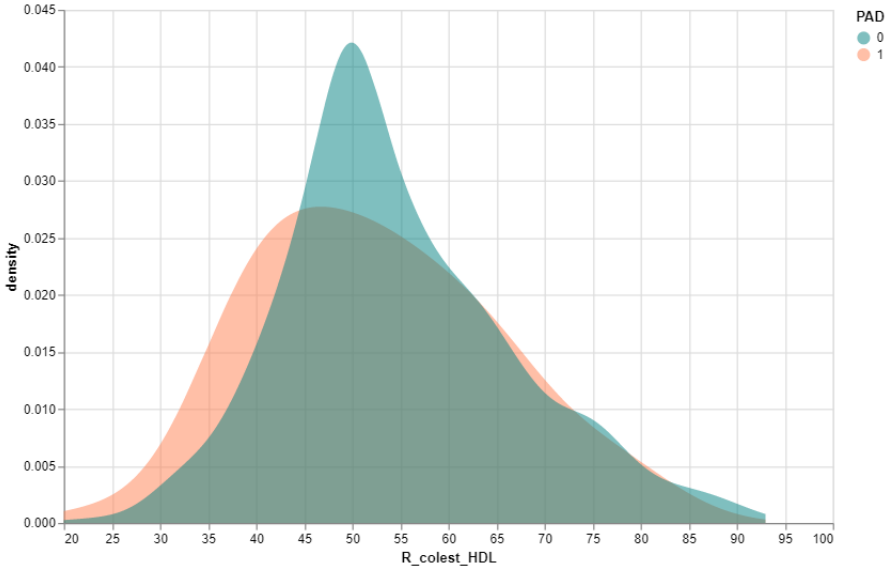

FIGURE 5.2: Density plot for HDL cholesterol

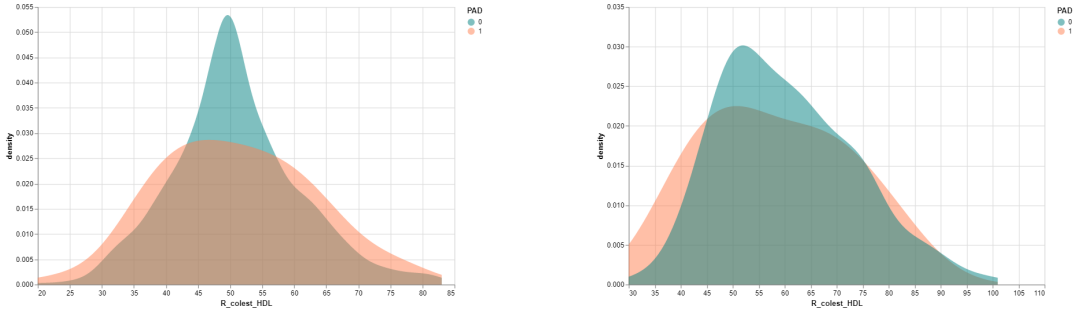

FIGURE 5.3: Density plot for HDL cholesterol for men (left) and woman (right)

Surprisingly, when checking if the behavior of this variable was consistent in both sexes, we found that HDL cholesterol behaves differently depending on the sex of the participant (Figure 5.3). In men, the density of the HDL cholesterol in PAD-positive participants is almost evenly distributed throughout all the registered values. However, in women, the presence of lower HDL cholesterol is stronger in the positive class. Ultimately, we expect this behavior to be reflected in the ML models and explanation algorithms as well as on the behavior of the variable *Dyslipidemia*.

5.1.3 Systolic and diastolic blood pressure

In the dataset, there are also continuous variables that contain the values for systolic and diastolic blood pressure. Furthermore, there is also a dichotomic variable that takes the value 1 if the participant is hypertensive or if the participant is being medicated for hypertension and 0 otherwise. In Figure 5.4 we find the density plots obtained for both systolic and diastolic blood pressure for the general population.

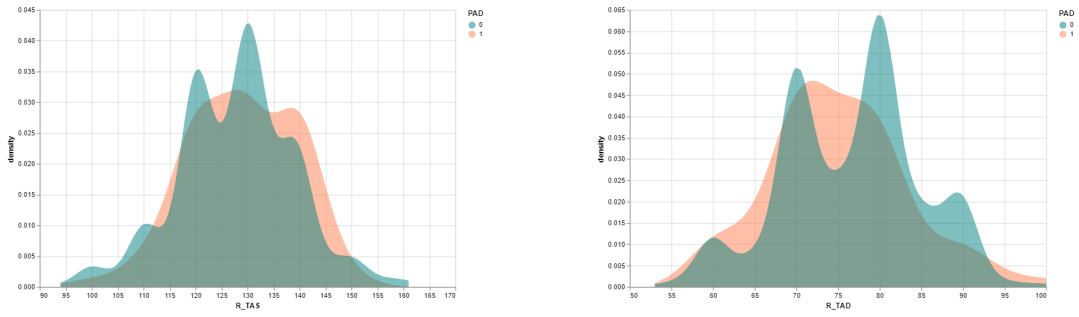

FIGURE 5.4: Density plot for systolic blood pressure (left) and diastolic blood pressure (right)

The distribution of these two variables for healthy participants, that is, participants who were not diagnosed with PAD, seems to be bimodal. We have verified this bimodal pattern also exclusively for women and men. The team of experts we consulted, mentioned that the cause for this apparent binomial distribution can be rooted in the use of different mechanisms to measure blood pressure.

In addition, it is important to take into consideration that some participants are being medicated, revealing the need to create a variable that encompasses this information. Therefore, we are expecting the dichotomic variable for hypertension to be far more informative and robust than the continuous blood pressure values when it comes to detecting risk factors for PAD.

5.1.4 Dyslipidemia's presence on PAD-positive participants

From our exploratory data analysis, PAD incidence was roughly the same between participants who had dyslipidemia and participants who did not have lipid imbalances. A percentage of 4.17% of healthy patients had dyslipidemia, as opposed to 3.74% of PAD-positive patients who had dyslipidemia. In female participants, the ratio slightly decreases, with only 1.47% of healthy females having dyslipidemia against 2.09% of PAD-positive females, who also had dyslipidemia. In male participants, 3.15% of the healthy group had dyslipidemia, and almost the same proportion (3.07%) of the PAD group males had lipid imbalances.

Dyslipidemia is strongly related to cholesterol levels, so due to the anomalies in the density of the *total cholesterol* and *HDL cholesterol* variables we are expecting abnormal behavior of the *dyslipidemia* variable as well. This could be due to the fact that some participants were already being medicated for the presence of lipid imbalance. Contrary to the variable *hypertension*, the variable *dyslipidemia* does not take into consideration whether a participant is being medicated for lipid imbalance.

5.1.5 REGICOR score

The REGICOR score (Salas-Salvadó et al., 2017) measures the risk at which a subject has to suffer from a CV event in the following 10 years. This score takes into consideration age, sex, history of smoking, diabetes, blood pressure, as well as total cholesterol and HDL cholesterol levels. Low scores, indicating a low risk of suffering a CV event in the next 10 years, are considered to be those below 5%. Subjects

with high risks of suffering a CV event would present a REGICOR score with values over 10%.

Out of all 1173 participants, 99 had a high REGICOR score, and only 10 of these were PAD-diagnosed participants. Although we saw in the general population that those who received a higher REGICOR score, were typically hypertensive and/or diabetic, we discovered that most of the 10 PAD-diagnosed participants who were given a high risk of suffering a CV event in the next 10 years had dyslipidemia, were not diabetic, did not have hypertension, and were non-smokers.

A number of 29 out of the 73 PAD-positive participants were given a low risk of suffering a CV event in the next 10 years, and 39 out of 73 PAD positive-participants were given a medium risk. Men were typically given higher risk scores than women, although most men were given a medium risk score, as is the case with women.

The weak relationship between REGICOR score and PAD, and even personal history of CV events indicates that this score should perhaps include more parameters. The possibility of including the presence of previous CV events in the calculation of this risk score should be studied.

5.1.6 Incidence of diabetes and historical variables in women

Before the analysis of the risk factors for PAD in women, it is important to notice that we are expecting inconclusive behavior from the variable diabetes. This is because only 4 out of the 18 female participants diagnosed with PAD had diabetes, therefore we intuitively predict that the models that we are going to develop do not model correctly the exact correlation between the presence of diabetes and PAD in women, simply because we do not have enough data.

The same happens for the majority of variables that register the presence of cardiovascular diseases. No women diagnosed with PAD had chronic renal disease, for instance. Out of all the 18 diagnosed women, only 1 of them had had a history of cardiac ischemia. We were able to find out that this participant was an ex-smoker, had diabetes, hypertension, and dyslipidemia, was in pain, and was dependent on others to do their daily tasks. All the while only having a medium REGICOR score, bringing us back to the apparent inefficiency of the REGICOR score for this particular dataset. We were able to also see that the only woman who had suffered from an aneurysm had dyslipidemia, was in pain, had mobility issues, was diabetic, was a smoker, and was hypertensive. She was also given a medium risk of suffering a CV event in the next 10 years. We also verified that these patients did not have any relevant data imputed.

5.1.7 Incidence of diabetes and historical variables in men

As opposed to female participants, 20 male PAD-positive participants had diabetes, out of a total of 65 male PAD-positive participants. Regarding historic variables, men had slightly more previous CV events than women, although not enough for us to be able to extract clear relationships between these historic variables and PAD. This lack of data is expected to be reflected in an inconclusive behavior of the historic variables later on in this project.

Chapter 6

Results and Discussion

6.1 Ablation studies

We performed several ablation studies. These include running the previously explained models with the most concerning variables mentioned in 5 - that we know to introduce multicollinearity into the model -, a comparison of predictive power between categorical and continuous variables, as well as a small study on the REGICOR score variable. Table 6.1 contains all the combinations of problematic variables studied.

TABLE 6.1: Combinations of problematic variables studied

Combinations studied
Q perimetre abd
Q perimetre abd + Obesitat
Q perimetre abd + Obesitat + IMC
Obesitat
Obesitat + IMC
IMC

6.1.1 Abdominal perimeter

The impact of this variable on the models' prediction was coherent throughout our ablation studies and was consistently at the top of the list of the most important variables. Nonetheless, its impact was more evidently linear for women, than for men.

Apart from the linear relationship between PAD and abdominal perimeter being more evident in the female group, we also concluded that it is considered more important for women than men. When combining abdominal perimeter, BMI and obesity, the behavior of the relationship between the abdominal perimeter and PAD remains untouched, as well as its importance. In addition, all models consider the abdominal perimeter to be of higher importance for women when detecting PAD.

6.1.2 BMI or *IMC*

The impact of BMI on the model's predictions of PAD varies greatly from women to men. In women, we generally see a positive relationship between high BMI and an increase in the probability of being predicted as part of the positive class. Nevertheless, the models seem to have difficulty detecting the precise relationship between PAD and BMI both in men and women. In men, this difficulty is exacerbated. The

models seem to indicate that BMI is strongly influencing the final prediction of PAD, but cannot pinpoint why, or, at least, cannot display a consistent behavior in the relationship each of these models learns between BMI and PAD, suggesting that for this population the BMI's relationship with PAD is quite complex.

In A.13, we can observe the ambiguous behavior of BMI both in men and in women - although more present in men - that is displayed throughout the SHAP plots we built for each model and random sample.

To further study the behavior of this variable, we introduced the variable obesity into the model, which did not cause significant changes concerning the models' capacity to shape the relationship between BMI and PAD. On the other hand, introducing the variable abdominal perimeter to the model also did not change the above description of the BMI behavior in relation to PAD. Similarly, introducing all these three variables - including obesity - together does not provide additional information.

6.1.3 Obesity

We concluded that PAD-positive subjects do not necessarily follow the pre-conceived and arguably intuitive idea that subjects classified as obese or overweight have a higher incidence of PAD.

The profile of PAD participants when it comes to obesity varies depending on the sex of the participant. For women, we verify that, in the overall set of patients, obesity is negatively correlated with the probability of our models predicting a female subject as part of the positive class. This means that the models learned that being overweight, or even obese in some cases, in this population, does not constitute a risk factor for PAD, although clinically speaking, this should always be a variable considered. The information we obtain from this analysis is that women might not follow the expected pattern for PAD regarding obesity.

On the other hand, in men, we recurrently find that obesity is positively correlated with the probability of our models predicting a male subject as part of the positive class. The vast majority of SHAP plots indicate that, in general, men follow the average profile of a PAD-positive subject.

When we introduced both obesity and BMI into the model, the multicollinearity did not provoke any change in the behavior of the obesity variable. However, the influence of obesity in the final prediction of the models significantly decreased for men, indicating that BMI might have more "weight" at the time of prediction, although, as mentioned before, the models cannot shape the relationship between BMI and PAD in a simplistic way. Still, obesity was still generally considered one of the most important features, noting that in women, the obesity factor should be considered carefully at the moment of diagnosis. That is, the reason for discarding an eventual PAD screening should not exclusively be based on whether the subject is obese or not, especially for dubious cases. For women, we seem to detect the so-called *obesity paradox* mentioned in Chapter 1.

Instead of introducing BMI to the model, if we introduce the abdominal perimeter, the behavior of obesity for men becomes less consistent, which is a common consequence of introducing multicollinearity into a model. In parallel, introducing all three variables into the model, the variable obesity loses relative importance when

compared to the abdominal perimeter and BMI. Nevertheless, it is important to notice that the shift in obesity's relationship with PAD is only present in male subjects. In female participants, obesity is still negatively correlated with an increase in the probability of any model classifying a female subject as part of the positive class.

Since the variable *obesity* was not able to provide us with coherent information about its relationship with PAD, we decided to introduce the true values of the weight for each participant in the models. We suspect that this odd behavior of this variable is intrinsically related to its nature. This variable categorizes the participants as obese, normal weight, or underweight not taking into consideration the weight itself. This can provoke the loss of more subtle and precise information when it comes to modeling the relationship between obesity and PAD. When we introduced the true weight component into the models we were able to see that this variable was weakly positively correlated with PAD, confirming the conclusion that weight on its own is not an efficient variable to predict PAD.

6.2 Discussion

Apart from the conclusions we were able to derive from the ablation studies described above, we also detected some interesting patterns. We were able to achieve the same results found in Gonçalves-Martins et al., 2021, and also detect anomalies in the relationship of some variables with PAD. These anomalies are found in the cholesterol levels, both total and HDL, which consequently produce anomalies in the relationship between dyslipidemia and PAD.

It should be noted that the variable *sex* was found to be very relevant.

6.2.1 Tree-based models vs. Logistic Regression

When comparing the interpretability of logistic regression (LR) models and tree-based models, it becomes evident that tree-based models allow for the extraction of much more complex and richer information from SHAP (SHapley Additive exPlanations) plots. Unlike LR, which assumes a linear relationship between feature values and the probability of being predicted as a positive instance, tree-based algorithms offer a more flexible framework that captures non-linearities and reveals subtle nuances in the data.

The SHAP plots derived from tree-based models provide a deeper understanding of the intricate relationships between features and the predicted outcomes. These plots showcase the variable importance and the impact of different feature values on the final prediction, highlighting the intricate interplay among the predictors. By examining the SHAP values for tree-based models, we can discern how specific feature values contribute to the overall prediction and uncover hidden patterns and interactions.

In A, we can find an example of a SHAP plot built using Logistic Regression.

Although the metrics for all models, both tree-based and non-tree-based, are good and consistent, we find that the difference in the quality of the algorithms does not necessarily lie in the values of the metrics, but in the interpretations and relationships the algorithms create and *believe* exist within the dataset. Therefore, to choose the best algorithm we need to leverage carefully the quality of the metrics as

well as the quality of the SHAP plots, that is, the interpretations and relationships the models create and follow in the prediction process.

For the above-mentioned reasons, choosing the best algorithm proved to be a rather complex task, as we needed to both analyze the interpretability of the SHAP plots and the quality of the metrics. Ultimately, we concluded that the Random Forest algorithm was more efficient in the creation of relationships between risk factors and PAD.

Despite all tree-based models having good metrics, we found that the SHAP plots built for Random Forest models were more interpretable than the SHAP plots for the remaining tree-based algorithms. Although this does not necessarily correlate with better metrics, such as accuracy, recall, and precision, among others, we believe this finding suggests that Random Forest algorithms might be more clinically interpretable than others.

6.2.2 Ablation studies' conclusion

With our ablation studies, we were able to detect that the abdominal perimeter was, in this case, the most efficient variable from the group of variables studied in the ablation studies described above. We were also able to detect the differences in the importance of abdominal perimeter and BMI between women and men. Women who present higher abdominal perimeter and/or BMI in this population, are more likely to be predicted as PAD-positive participants.

6.2.3 Diabetes

As predicted before, the variable that encodes diabetes for women either had no importance or inconclusive importance due to the lack of women with diabetes. For men, however, due to more information regarding the profile of diabetic participants, the model was able to clearly define a relationship between diabetes and PAD, revealing that men with diabetes are much more prone to be predicted as a part of the positive class.

6.2.4 Well-being related variables

Due to several studies that have related mental health problems with PAD (Thomas et al., 2020), we decided to study the effects of the dichotomic variable that expresses whether a participant has mental health problems or not at the moment of prediction.

This variable that encodes the mental health state had a very opposite relationship with PAD depending on the sex, which was unexpected. Although we verified that the participants who had mental health disturbances had extremely similar profiles in men and in women, we see that the models typically assign a positive correlation between mental health disturbances and PAD for men and a negative correlation for women. The only major difference we found between these participants' profiles was in the smoking habit and in the REGICOR scores: all men who reported mental health problems were either ex-smokers or smokers, while women were evenly distributed between non-smokers, ex-smokers, and smokers; regarding REGICOR scores, 1 man had been given a high REGICOR score, as opposed to 0 women who were given a high REGICOR score.

We believe this unexpected behavior of the mental health variable is related to its construction. This variable is the result of a combination of multiple other variables on mental health. This combination of variables might hide underlying information that would otherwise be relevant to our analysis. Since we had the possibility to explore similar variables to this one, we included in our study another dichotomic variable that expresses whether a participant has either anxiety or depression. The behavior of this newly added variable was the same both for men and women and demonstrated to be quite relevant in the prediction of PAD, a conclusion that is in accordance with several existing studies relating to mental health and PAD. Nonetheless, the team of experts we consulted did not point to anxiety and depression as concerning risk factors, pointing out that these might even be consequences of PAD.

Furthermore, we studied a variable that registers the level of physical activity and we found that it displays the same behavior as the mental health variable. We believe this to be due to the lack of data belonging to this variable's categories since most of the PAD-positive participants do not have mobility problems - revealing, yet again, that PAD is, more often than not, asymptomatic -, essentially culminating in SHAP plots that either don't associate mobility issues with PAD or that do so negatively.

6.2.5 Cholesterol levels

We commonly found that low levels of total cholesterol, high HDL levels, and the inexistence of dyslipidemia in men were generally associated with PAD. For women, both high levels of total cholesterol and the inexistence of dyslipidemia also increased the probability of being predicted as a PAD-positive patient. These findings were expected, given the previous exploratory analysis described in Chapter 3. Furthermore, this abnormal behavior of the cholesterol values, which is ultimately reflected in the relationship the models build between PAD and dyslipidemia, has also been witnessed in other research (Aday and Everett, 2019b; Chyou and Eaker, 2000; Brotons, Moral, and Vicuña, 2022; Ravnskov, Diamond, Hama, et al., 2016).

We then introduced a new variable into the model, hoping to extract some clear pattern from the relationship between total cholesterol and HDL cholesterol, and PAD. The new variable is the cholesterol ratio, which consists of dividing the total cholesterol by HDL cholesterol. In turn, it was clear that this newly added variable was very relevant for the detection of PAD. We were able to conclude that this ratio is negatively correlated with PAD, which further corroborates the inverse relationship of cholesterol and PAD. Participants whose total cholesterol was composed of a high portion of HDL cholesterol were more likely to be predicted as PAD-positive participants. This means that participants with a normal (high) rate of HDL cholesterol and low (normal) values of cholesterol had a higher probability of having PAD.

6.2.6 Blood pressure

As expected, we also detected erratic behavior from the variables related to blood pressure values. These behaviors from blood pressure and cholesterol are not concerning, since these were the behaviors expected due to the possible presence of medication. As mentioned before, the presence of medication has a significant effect on the behavior of these continuous variables. Without the additional information on medication, detecting patterns in these continuous variables proves to be a large obstacle.

6.2.7 Risk scores for CV events

In our dataset, we have multiple variables that are related to the risk of suffering a CV event in the future. Two of these correspond to the REGICOR score, one of them being the encoded values. Therefore, REGICOR score of under 4% corresponds to a value of 1; a REGICOR score between 5% and 9% corresponds to a value of 2; lastly, a REGICOR score of above 10% corresponds to a value of 3. The remaining variables were considered to be redundant since they also measure the risk of suffering a CV event, although in a dichotomic way. We discarded them after having confirmed that they were not informative.

We initially discarded the REGICOR variables from our studies since we already have the features that make up this score and because our goal was to examine how these individual features might play a role in the development of PAD. Nonetheless, we were curious to see how the models shaped the relationship between the risk scores and PAD.

When fed to the models, the continuous REGICOR score was more efficient in predicting PAD than the encoded REGICOR score.

This might be due to the high number of participants whose REGICOR score is on the "bounds" of the previously defined intervals, which would ultimately translate to a poor SHAP plot. Therefore, we believe it would be clinically more efficient to introduce the real values of the REGICOR score instead of the encoded ones, to properly investigate the relationship between REGICOR score and PAD. This constitutes an alternative to explicitly labeling subjects, as such a procedure could lead to a loss of information regarding the complexity of the relationship between the risk score and PAD. This is a common problem when deciding to encode continuous variables in a limited number of bins.

6.2.8 Dichotomic vs. Continuous variables

We tested running the models using only dichotomic variables, as opposed to using a mix of dichotomic and continuous variables to test their predictive and explanatory power alone. However, depriving the models of the continuous variables generally decreased the models' predictive power. Although our goal is not to extrapolate these models to other datasets, it is still desirable we achieve good metrics for the models to obtain reasonable SHAP plots. It seems that using continuous variables, specifically the abdominal perimeter, allows the model to capture more nuances of the relationships between the features and the target variable.

Nevertheless, as mentioned in Chapter 5, continuous variables such as blood pressure and cholesterol that are known to not be informative, should be left out. Therefore, we recommend using a selected mix of continuous and dichotomic variables in these types of research, a mix that we recommend includes variables such as abdominal perimeter and/or BMI. This selection of variables should be made based on previous knowledge from the population being studied. If previous knowledge is non-existent, or small, we recommend first analyzing all of the variables available, and later on, performing ablation studies by selecting only a sample of risk factors to study. As always, exploratory data analysis should be carried out, as this analysis typically already provides us with a considerable amount of information that, perhaps, will allow us to diminish the dimensionality of the dataset, ultimately reducing the complexity and length of the above-mentioned ablation studies.

6.3 Limitations

The limitations of this work go beyond the field of machine learning. The results of this project have once again shown that there is no standard profile for a subject with PAD and that this profile might vary greatly geographically and demographically, which leads us to the conclusion that these results cannot precisely be extrapolated to the general population.

The ultimate goal would be to create models that can be extrapolated, that is, models that we can successfully use to predict whether or not a subject has PAD. Nonetheless, it would be highly unlikely that we would be able to create a one size fits all model. As mentioned above, the profile for PAD-positive subjects varies depending, for example, on the country we consider. Therefore, at best, we could only confidently extrapolate this model to the geographic region the training data was extracted from to get accurate results. To do this, we would need to have more information on PAD incidence in the Catalan population than what we have access to today to train an unsupervised model that would later be applied to the population of Catalonia, more specifically, Barcelona. In this project, the lack of existing data on PAD population proved to be a sizeable obstacle. The models created were done on a very small database, and the testing set almost always had fewer participants than what we would have preferred, which led us to invest more computational time in the training of more models.

6.4 Model bias

As mentioned before, due to the lack of female PAD participants, we were not able to obtain as many well-performing models for women as we did for men, which led us to explore the differences in metrics between the trained models for women and for men. We performed a fairness analysis by comparing the metrics obtained - accuracy, recall, precision, F1-score, AUC-ROC - for the models that were used to construct the SHAP plots and analysis. The code created took into consideration the models created with all the random samples taken from the original dataset.

We had by far more models for men than for women (more than twice), to which we applied the ANOVA test to determine whether the difference in metrics of these models was statistically significant, having had a negative result. This was consistent for all the models created with the ablation studies.

Chapter 7

Conclusions

We were able to corroborate the main conclusions from Gonçalves-Martins et al., 2021. With our project, smoking habit was classified as an important risk factor for PAD, both for men and women. Diabetes was only found to be a strong predictor for men, whereas hypertension was only found to be a strong predictor for women.

Despite the small number of PAD cases in our dataset, we do believe that we have detected key characteristics of PAD-positive subjects in northern Barcelona, such as an intriguing inverse relationship between total cholesterol and PAD, and the importance of having a variable that registers whether a participant is taking medication to either regulate total cholesterol levels, hypertension, or others. Moreover, we believe that it would be interesting to further explore the mental health's impact on the development of PAD, as this is a topic that is recent within the scope of the study of this disease.

We also believe that having additional variables related to medication intake would greatly increase the value of this analysis and possibly add new insights allowing us to better understand the modeled relationships between risk factors and PAD. We were especially intrigued by the relationship between PAD-positive subjects and total cholesterol, as well as blood pressure, and were left wondering if these findings are the result of the influence of medication.

In addition to acknowledging the overall lack of data on PAD, it is essential to highlight the scarcity of epidemiological studies specifically focused on vascular pathologies. This dearth of research in the field of vascular diseases stands in stark contrast to the vast number of publications dedicated to cardiac pathologies.

It is crucial to emphasize the urgent need for robust data on PAD. The scarcity of information on this disease, particularly regarding PAD incidence in the general population, is a significant challenge for clinicians to better understand this disease. With this project, we aim to address this knowledge gap. Our work seeks to contribute to the extremely limited existing body of literature and provide valuable insights into PAD to inform better clinical practices and improve patient outcomes, even if in the slightest form.

Appendix A

Plots and metrics

A.1 Average metrics and model bias

We obtained the average metrics for each ablation study, as well as the average difference between the metrics obtained for women's and men's models. The average metrics do not vary greatly across ablation studies, nor across different classifiers.

	Avg metrics for women	Avg metrics for men	Avg metrics Standard deviation for women	Avg metrics Standard deviation for men	Average metrics Difference	Average Standard Deviation
accuracy	0.882	0.746	0.043	0.076	0.136	-0.033
AUC-ROC	0.840	0.741	0.092	0.071	0.099	0.021
precision	0.802	0.745	0.221	0.141	0.057	0.080
recall	0.762	0.706	0.256	0.131	0.056	0.125
f1 score	0.714	0.709	0.087	0.078	0.006	0.009

FIGURE A.1: Example of average metrics

A.2 Sex specified blood pressure density plots

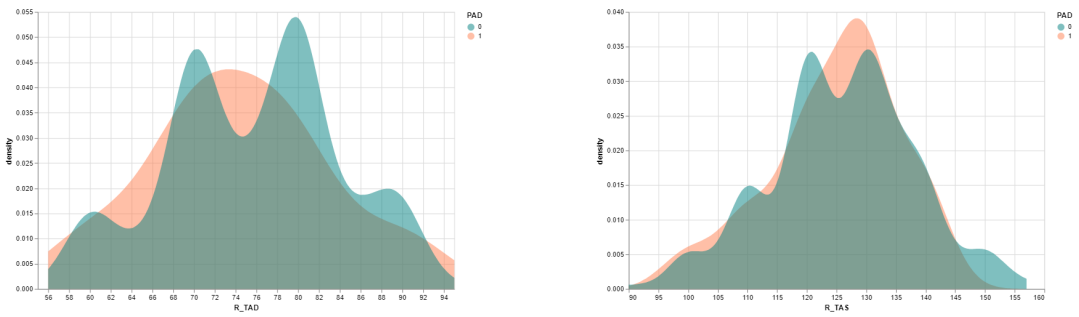

FIGURE A.2: Density plots for diastolic (left) and systolic (right) blood pressure for women

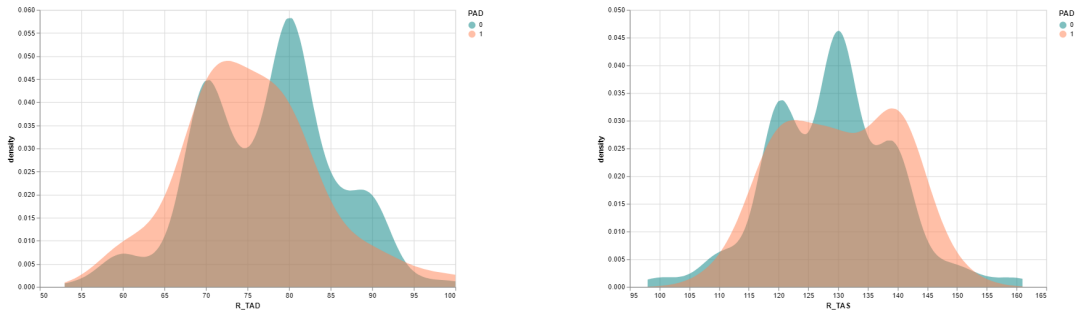

FIGURE A.3: Density plots for diastolic (left) and systolic (right) blood pressure for men

A.3 SHAP plots

A.3.1 General conclusions

In figure A.4 we find an example of a SHAP plot for both women and men.

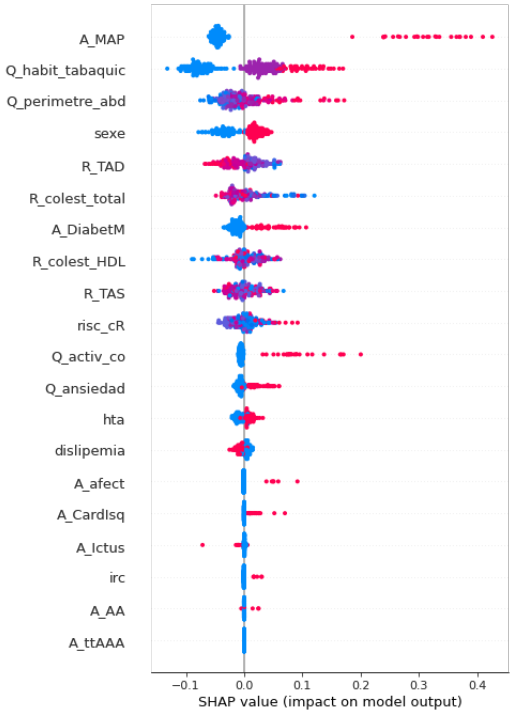

FIGURE A.4: General summary plot

Figures A.5 and A.6 show the initial SHAP plots from which we extracted an analysis of the main differences in risk factors between women and men.

Figures A.7 and A.8 include the created variable *total HDL ratio*. For these SHAP plots, we included this variable and excluded both the variables that had the values of total cholesterol and HDL cholesterol because we had previously discarded them due to the lack of added information.

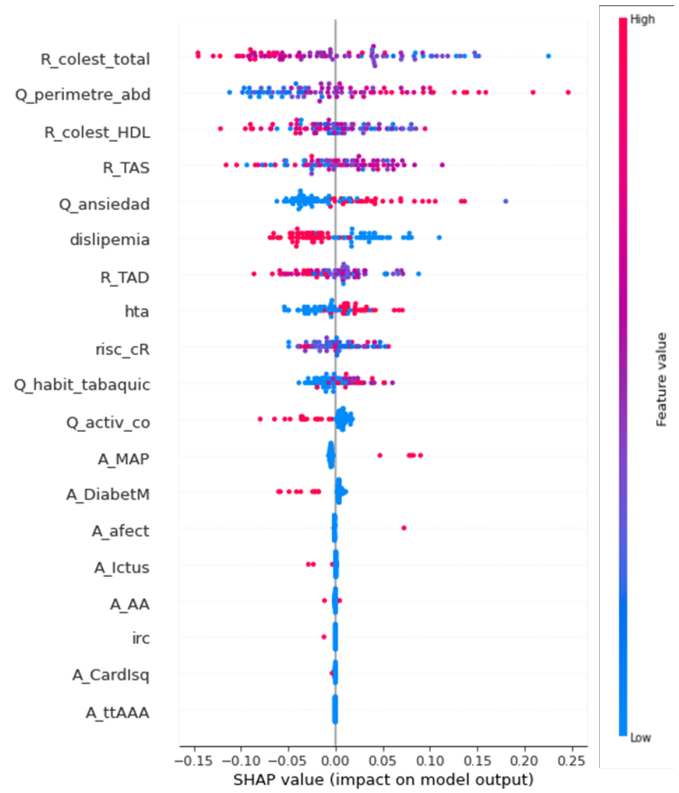

FIGURE A.5: Initial summary plots subject to women

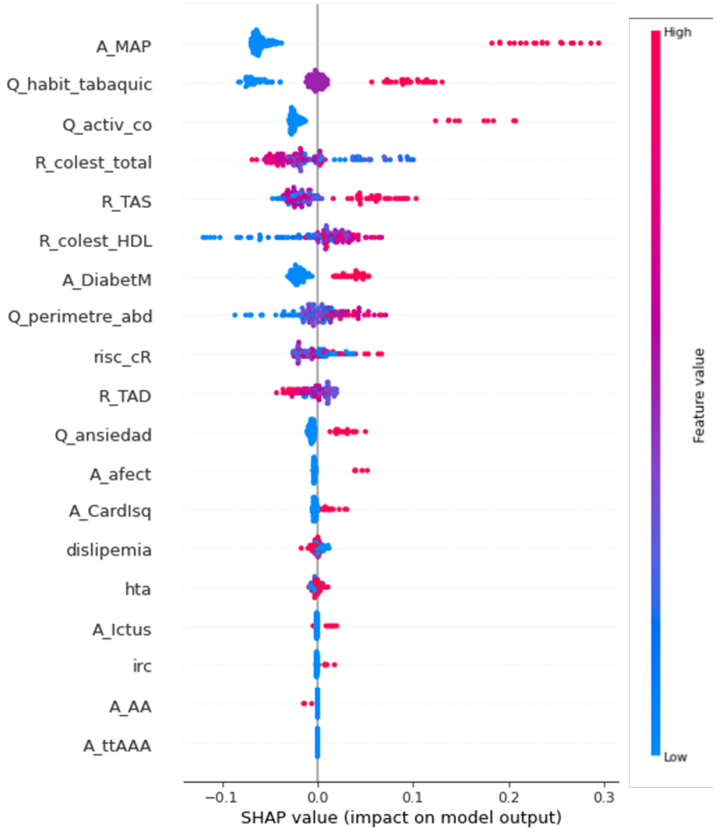

FIGURE A.6: Initial summary plots subject to men

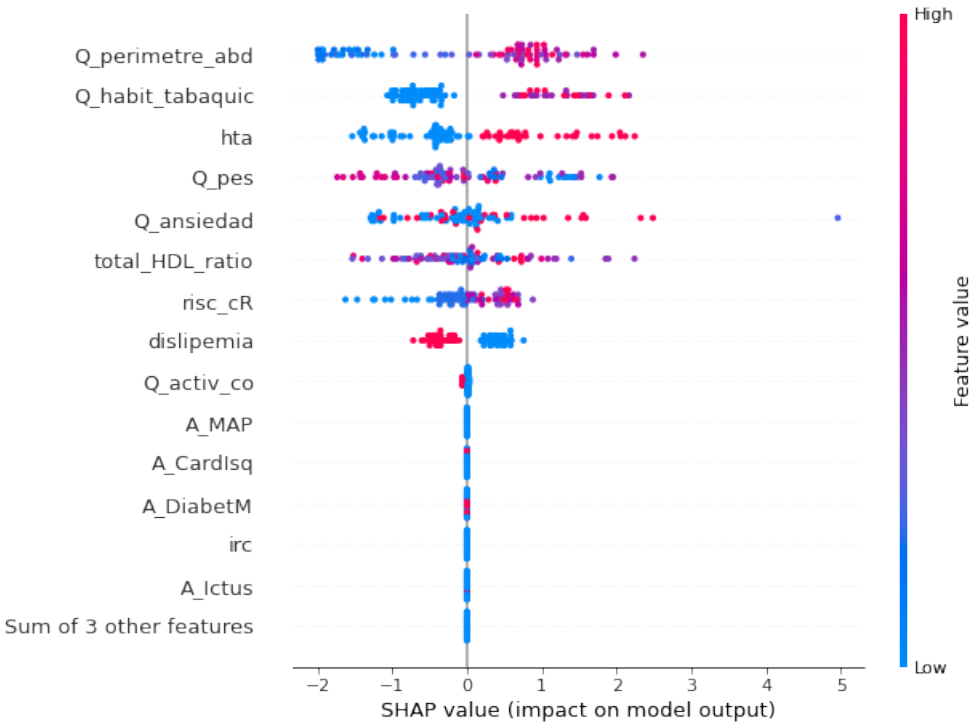

FIGURE A.7: SHAP plot including newly created variable *total HDL ratio* for women

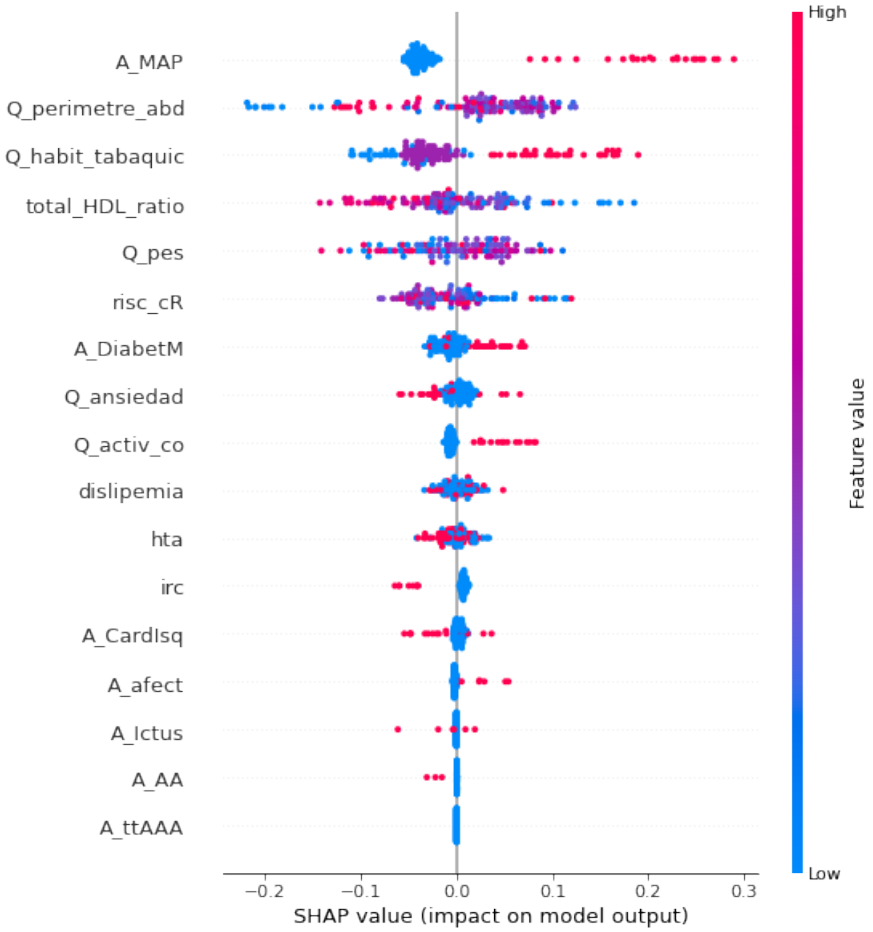

FIGURE A.8: SHAP plot including newly created variable total *total HDL ratio* for men

A.3.2 Ablation studies

In figures A.9 and A.10, we can see the difference between the impact of the abdominal perimeter in the prediction of PAD in women and men.

In Figures A.11 and A.12, we can observe the ambiguous behavior of BMI both in men and in women - although more present in men - that is displayed throughout the SHAP plots we build for each model and random sample.

In Figures A.13 and A.14, we see how the relationship modeled between obesity and PAD varies according to sex.

In Figures A.15 and A.16, we can see the relationship between the risk scores and PAD and the previously mentioned anxiety factor more present in women than in men.

In Figure A.17, we can see an example of how a SHAP summary plot built using Logistic Regression displays the relationships between the risk factors and PAD.

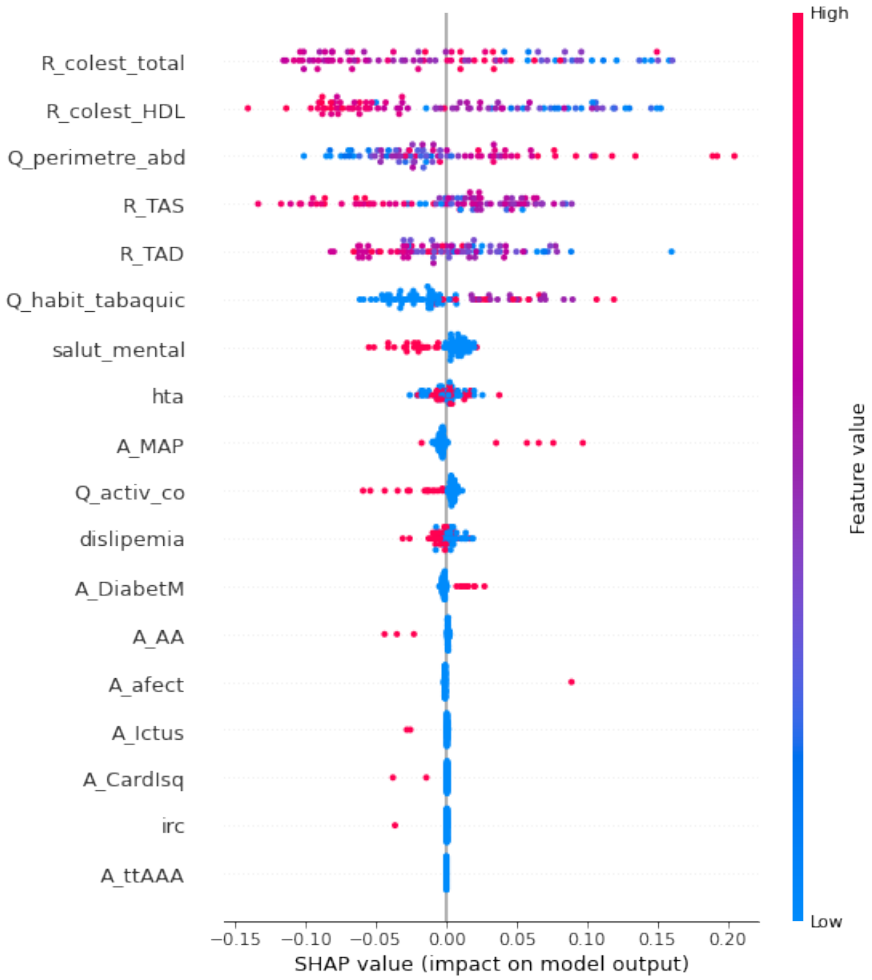

FIGURE A.9: Summary plots subject to women, considering abdominal perimeter

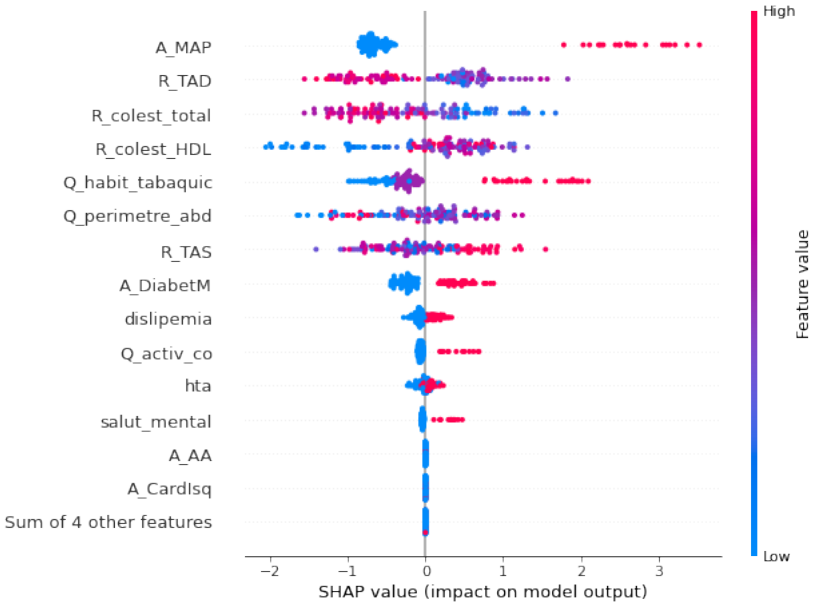

FIGURE A.10: Summary plots subject men, considering abdominal perimeter

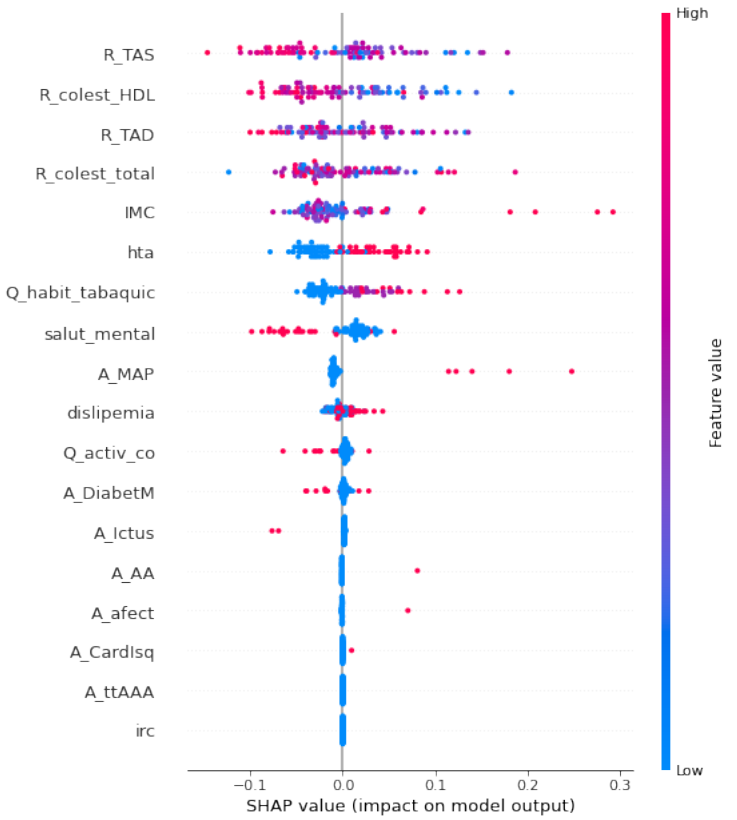

FIGURE A.11: Summary plots subject to women, considering BMI

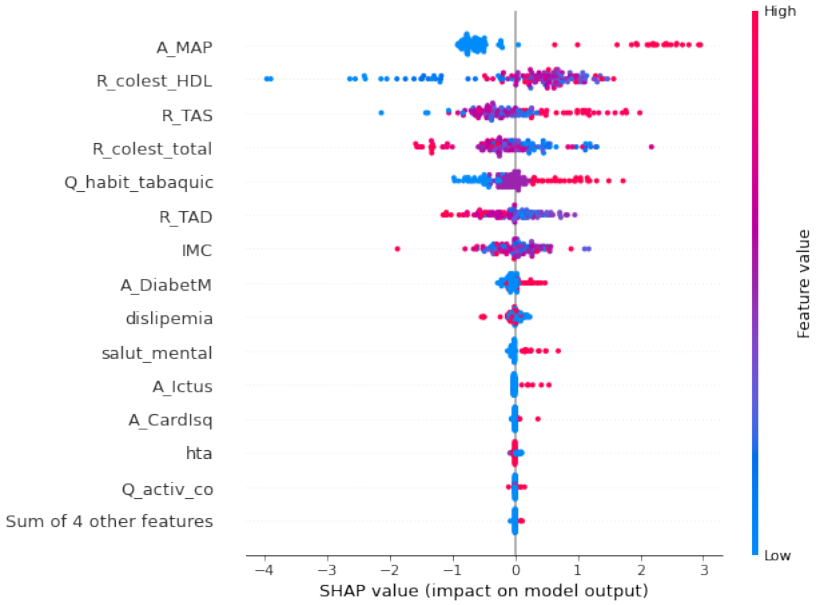

FIGURE A.12: Summary plots subject men, considering BMI

FIGURE A.13: Summary plots subject to women, considering obesity

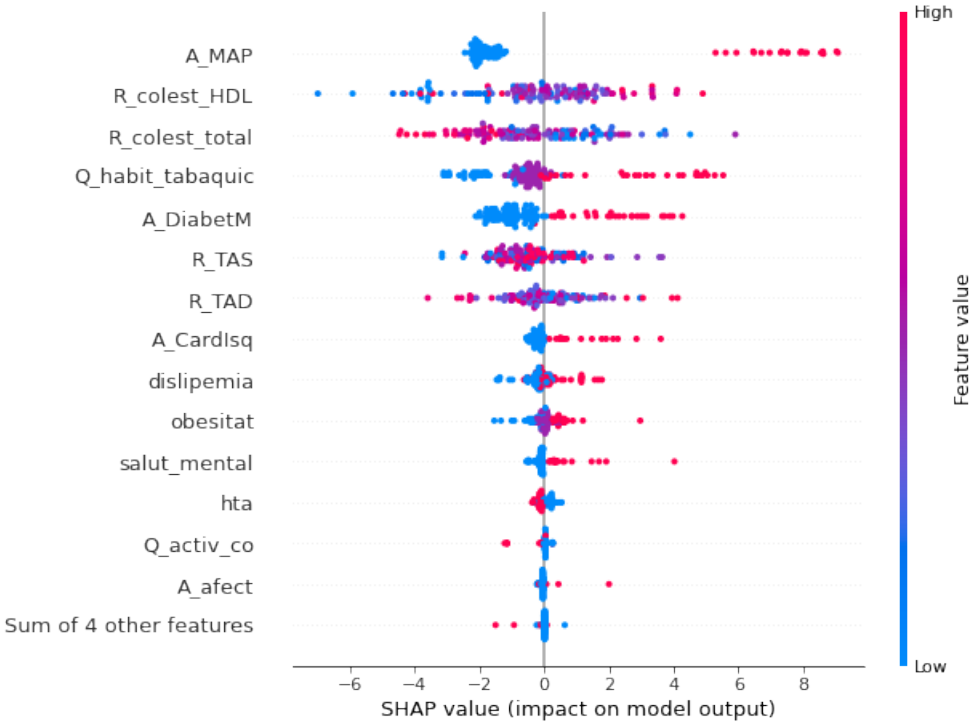

FIGURE A.14: Summary plots subject men, considering obesity

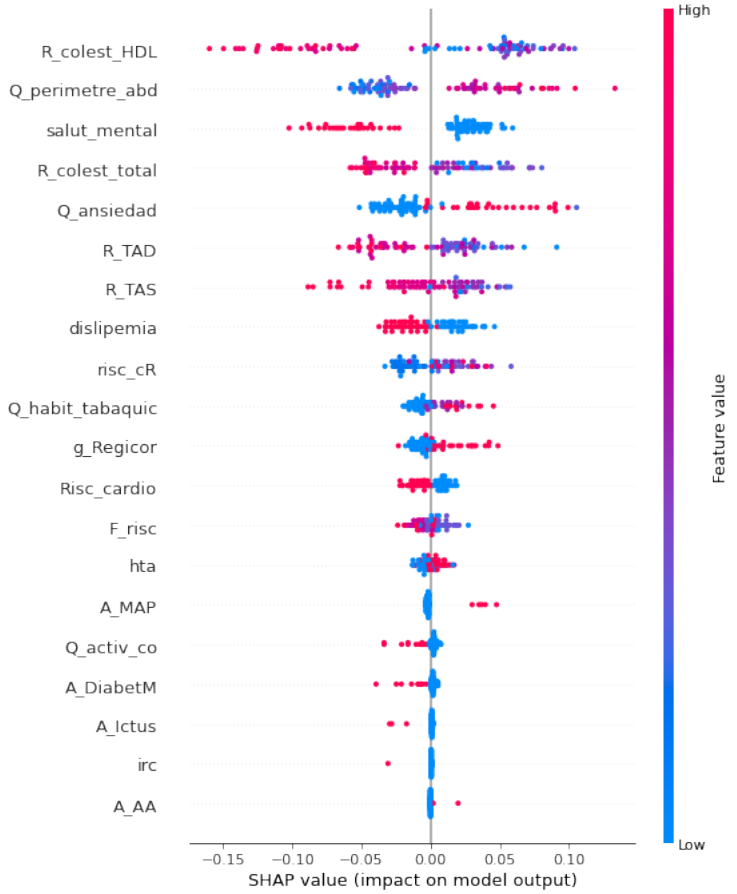

FIGURE A.15: Summary plots subject to women, considering REGICOR

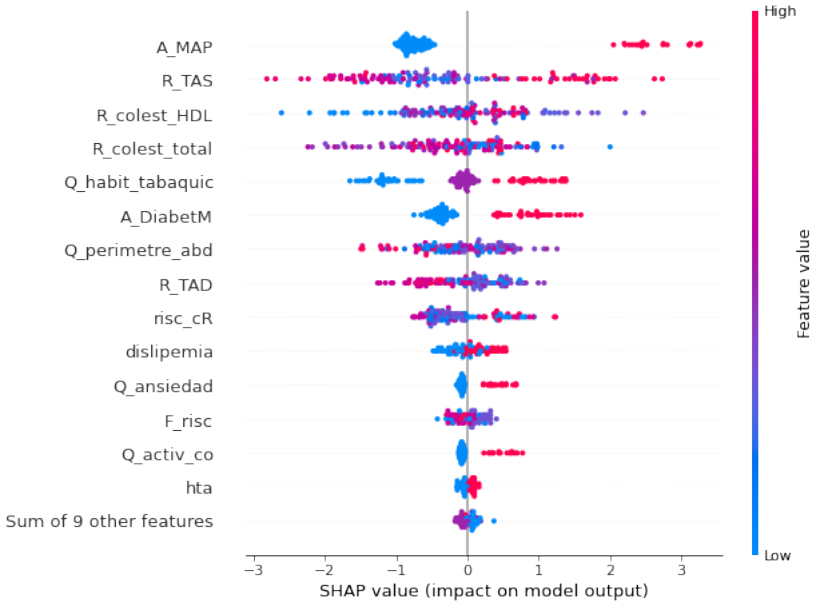

FIGURE A.16: Summary plots subject to men, considering REGICOR

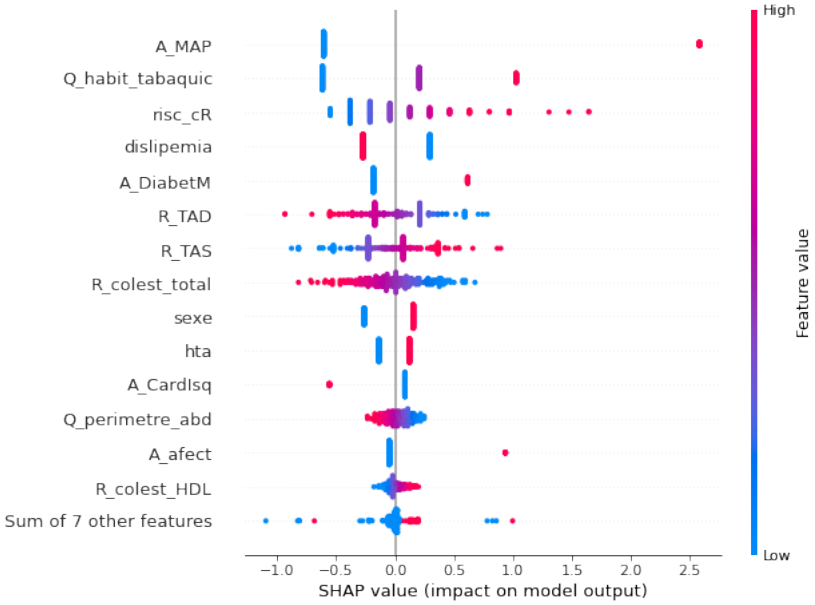

FIGURE A.17: Example of summary plot built using Logistic Regression algorithm

Bibliography

Aday, Aaron W and Brendan M Everett (2019a). "Dyslipidemia profiles in patients with peripheral artery disease". In: *Current cardiology reports* 21, pp. 1–9.

— (2019b). "Dyslipidemia profiles in patients with peripheral artery disease". In: *Current cardiology reports* 21, pp. 1–9.

Aday, Aaron W et al. (2018). "Lipoprotein particle profiles, standard lipids, and peripheral artery disease incidence: prospective data from the Women's Health Study". In: *Circulation* 138.21, pp. 2330–2341.

Alushi, Kastriot et al. (Oct. 2022). "Distribution of Mobile Health Applications amongst Patients with Symptomatic Peripheral Arterial Disease in Germany: A Cross-Sectional Survey Study". In: *Journal of Clinical Medicine* 11.3. URL: https://www.mdpi.com/2077-0383/11/3/498.

Araki, Yoshihiro et al. (2012). "Prevalence and risk factors for cerebral infarction and carotid artery stenosis in peripheral arterial disease". In: *Atherosclerosis* 223.2, pp. 473–477.

Ashrapov, Insaf (2020). *Tabular GANs for uneven distribution*. arXiv: 2010.00638 [cs.LG].

Blanes, JI, MA Cairols, J Marrugat, et al. (2009). "Prevalence of peripheral artery disease and its associated risk factors in Spain: The ESTIME Study". In: *Int Angiol* 28.1, pp. 20–5.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Brotons, Carlos, Irene Moral, and Johanna Vicuña (2022). "La complejidad del papel del colesterol unido a HDL". In: *Revista Española de Cardiología*.

Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma (2022). "An optimized XGBoost based diagnostic system for effective prediction of heart disease". In: *Journal of King Saud University-Computer and Information Sciences* 34.7, pp. 4514–4523.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chyou, Po-Huang and ELAINE D Eaker (2000). "Serum cholesterol concentrations and all-cause mortality in older people." In: *Age and Ageing* 29.1, pp. 69–74.

Cimminiello, Claudio et al. (2011). "The PANDORA study: peripheral arterial disease in patients with non-high cardiovascular risk". In: *Internal and emergency medicine* 6, pp. 509–519.

Criqui, Michael H. et al. (July 2021). "Lower Extremity Peripheral Artery Disease: Contemporary Epidemiology, Management Gaps, and Future Directions: A Scientific Statement From the American Heart Association". In: *American Heart Association* 28613496. URL: https://doi.org/10.1161/CIR.0000000000001005.

Cronin, Oliver et al. (2013). "The association of obesity with cardiovascular events in patients with peripheral artery disease". In: *Atherosclerosis* 228.2, pp. 316–323.

Daoud, Jamal I (2017). "Multicollinearity and regression analysis". In: *Journal of Physics: Conference Series*. Vol. 949. 1. IOP Publishing, p. 012009.

Dinh, An et al. (2019). "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning". In: *BMC medical informatics and decision making* 19.1, pp. 1–15.

Esmaily, Habibollah et al. (2018). "A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes". In: *Journal of research in health sciences* 18.2, p. 412.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

Ghiasi, Mohammad M and Sohrab Zendehboudi (2021). "Application of decision tree-based ensemble learning in the classification of breast cancer". In: *Computers in biology and medicine* 128, p. 104089.

Gonçalves-Martins, G. et al. (Oct. 2021). "Prevalence of Peripheral Arterial Disease and Associated Vascular Risk Factors in 65-Years-Old People of Northern Barcelona". In: *Journal of Clinical Medicine* 10.4467. URL: https://doi.org/10.3390/jcm10194467.

Hnat, Tomas, Josef Veselka, and Jakub Honek (2022). "Left ventricular reverse remodelling and its predictors in non-ischaemic cardiomyopathy". In: *ESC Heart Failure* 9.4, pp. 2070–2083.

Holzinger, Andreas (2018). "From machine learning to explainable AI". In: *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, pp. 55–66.

Hsu, Bang-Gee et al. (2017). "High serum resistin levels are associated with peripheral artery disease in the hypertensive patients". In: *BMC cardiovascular disorders* 17.1, pp. 1–7.

Itoga, Nathan K et al. (2018). "Association of blood pressure measurements with peripheral artery disease events: reanalysis of the ALLHAT data". In: *Circulation* 138.17, pp. 1805–1814.

Jiang, Huilin et al. (2021). "Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease". In: *International Journal of Medical Informatics* 145, p. 104326.

Krishnan, Mangalath Narayanan et al. (2018). "Prevalence of peripheral artery disease and risk factors in the elderly: A community based cross-sectional study from northern Kerala, India". In: *Indian Heart Journal* 70.6, pp. 808–815.

Kuijk, Jan-Peter van et al. (Apr. 2010). "Long-term prognosis of patients with peripheral arterial disease with or without polyvascular atherosclerotic disease". In: *European Hearth Journal* 31.8. URL: https://doi.org/10.1093/eurheartj/ehp553.

Lakier, Jeffrey B (1992). "Smoking and cardiovascular disease". In: *The American journal of medicine* 93.1, S8–S12.

Levine, Shel D. (Mar. 2018). "Peripheral Arterial Disease: A Case Report From the Henry Ford Hospital". In: *Journal of Clinical Exercise Physiology* 7, pp. 15–21. URL: https://doi.org/10.31189/2165-6193-7.1.15.

Lin, Donna Shu-Han et al. (2022). "Mortality risk in patients with underweight or obesity with peripheral artery disease: A meta-analysis including 5,735,578 individuals". In: *International Journal of Obesity* 46.8, pp. 1425–1434.

Lu, Yifei et al. (2020). "2017 ACC/AHA blood pressure classification and incident peripheral artery disease: The Atherosclerosis Risk in Communities (ARIC) Study". In: *European journal of preventive cardiology* 27.1, pp. 51–59.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Manjunath, CN et al. (2013). "Atherogenic dyslipidemia". In: *Indian journal of endocrinology and metabolism* 17.6, p. 969.

Mirza, Shuja, Sonu Mittal, and Majid Zaman (2018). "Decision support predictive model for prognosis of diabetes using SMOTE and decision tree". In: *International Journal of Applied Engineering Research* 13.11, pp. 9277–9282.

Murabito, Joanne M et al. (2002). "Prevalence and clinical correlates of peripheral arterial disease in the Framingham Offspring Study". In: *American heart journal* 143.6, pp. 961–965.

Pabon, Maria et al. (2022). "Sex differences in peripheral artery disease". In: *Circulation Research* 130.4, pp. 496–511.

Pazhoohesh, Mehdi, Zoya Pourmirza, and Sara Walker (2019). "A comparison of methods for missing data treatment in building sensor data". In: *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*. IEEE, pp. 255–259.

P.Marso, Steven and William R.Hiatt (Mar. 2006). "Peripheral Arterial Disease in Patients With Diabetes". In: *Journal of Clinical Exercise Physiology* 47.5. URL: https://doi.org/10.1016/j.jacc.2005.09.065.

Poussa, Heikki et al. (2007). "Diagnosis and treatment of dyslipidemia are neglected in patients with peripheral artery disease". In: *Scandinavian Cardiovascular Journal* 41.3, pp. 138–141.

Pujianto, Utomo, Aji Prasetya Wibawa, Muhammad Iqbal Akbar, et al. (2019). "K-nearest neighbor (k-NN) based missing data imputation". In: *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, pp. 83–88.

Ramos, R et al. (2009). "Prevalence of symptomatic and asymptomatic peripheral arterial disease and the value of the ankle-brachial index to stratify cardiovascular risk". In: *European Journal of Vascular and Endovascular Surgery* 38.3, pp. 305–311.

Ranganathan, Priya, CS Pramesh, and Rakesh Aggarwal (2017). "Common pitfalls in statistical analysis: logistic regression". In: *Perspectives in clinical research* 8.3, p. 148.

Ravnskov, Uffe, David M. Diamond, Rokuro Hama, et al. (2016). "Lack of an association or an inverse association between low-density-lipoprotein cholesterol and mortality in the elderly: a systematic review". In: *BMJ Open* 6, e010401. DOI: 10.1136/bmjopen-2015-010401.

Salas-Salvadó, J et al. (2017). "Prediction of cardiovascular disease by the framingham-REGICOR equation in the high-risk PREDIMED cohort: Impact of the mediterranean diet across different risk strata". In.

Sampson, Uchechukwu KA et al. (2014). "Global and regional burden of death and disability from peripheral artery disease: 21 world regions, 1990 to 2010". In: *Global heart* 9.1, pp. 145–158.

Saranya, N et al. (2021). "Diagnosing chronic kidney disease using KNN algorithm". In: *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. Vol. 1. IEEE, pp. 2038–2041.

Schober, Patrick and Thomas R Vetter (2021). "Logistic regression in medical research". In: *Anesthesia and analgesia* 132.2, p. 365.

Shaikhina, Torgyn et al. (2019). "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation". In: *Biomedical Signal Processing and Control* 52, pp. 456–462.

Shipe, Maren E et al. (2019). "Developing prediction models for clinical use using logistic regression: an overview". In: *Journal of thoracic disease* 11.Suppl 4, S574.

Shu, J. and G. Santulli (Aug. 2018). "Update on peripheral artery disease: Epidemiology and evidence-based facts". In: *Atherosclerosis* 275, pp. 379–381. URL: https://doi.org/10.1016/j.atherosclerosis.2018.05.033.

Song, Peige et al. (2019). "Global, regional, and national prevalence and risk factors for peripheral artery disease in 2015: an updated systematic review and analysis". In: *The Lancet Global Health* 7.8, e1020–e1030.

Song, Yi et al. (2023). "Remnant cholesterol is independently asssociated with an increased risk of peripheral artery disease in type 2 diabetic patients". In: *Frontiers in Endocrinology* 14, p. 351.

Spitzer, Robert L et al. (2006). "A brief measure for assessing generalized anxiety disorder: the GAD-7". In: *Archives of internal medicine* 166.10, pp. 1092–1097.

Stehouwer, Coen DA et al. (2009). "Peripheral arterial disease: a growing problem for the internist". In: *European Journal of Internal Medicine* 20.2, pp. 132–138.

Teodorescu, Victoria J, Ashley K Vavra, and Melina R Kibbe (2013). "Peripheral arterial disease in women". In: *Journal of vascular surgery* 57.4, 18S–26S.

Thomas, Merrill et al. (2020). "Mental health concerns in patients with symptomatic peripheral artery disease: Insights from the PORTRAIT registry". In: *Journal of psychosomatic research* 131, p. 109963.

Thomas Manapurathe, Diana et al. (2019). "Cohort study examining the association between blood pressure and cardiovascular events in patients with peripheral artery disease". In: *Journal of the American Heart Association* 8.6, e010748.

Valentine, R James et al. (2004). "Family history is a major determinant of subclinical peripheral arterial disease in young adults". In: *Journal of vascular surgery* 39.2, pp. 351–356.

Velescu, A et al. (2016). "Peripheral arterial disease incidence and associated risk factors in a Mediterranean population-based cohort. The REGICOR Study". In: *European Journal of Vascular and Endovascular Surgery* 51.5, pp. 696–705.

Wadström, Benjamin Nilsson et al. (2022). "Elevated remnant cholesterol increases the risk of peripheral artery disease, myocardial infarction, and ischaemic stroke: a cohort-based study". In: *European heart journal* 43.34, pp. 3258–3269.

Wang, W. et al. (Sept. 2021). "Smoking and the Pathophysiology of Peripheral Artery Disease". In: *Frontiers in cardiovascular medicine* 8.704106. URL: https://doi.org/10.3389/fcvm.2021.704106.

Wassel, Christina L et al. (2011). "Family history of peripheral artery disease is associated with prevalence and severity of peripheral artery disease: the San Diego population study". In: *Journal of the American College of Cardiology* 58.13, pp. 1386–1392.

WHO (July 2022). "Summary of the Global HIV epidemic, 2021". In: URL: https://www.who.int/data/gho/data/themes/hiv-aids#:~:text=Globally%2C%2038.4%20million%20%5B33.9%E2%80%93,at%20the%20end%20of%202021.

Xie, Xiangkun et al. (2021). "Early prediction of left ventricular reverse remodeling in first-diagnosed idiopathic dilated cardiomyopathy: a comparison of linear model, random forest, and extreme gradient boosting". In: *Frontiers in Cardiovascular Medicine* 8, p. 684004.

Yoo, Seung Hoon et al. (2020). "Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging". In: *Frontiers in medicine* 7, p. 427.

Zemaitis, M. R. et al. (July 2022). "Peripheral Arterial Disease". In: *National Library of Medicine* 28613496. URL: https://www.ncbi.nlm.nih.gov/books/NBK430745/.