Risk factors for Cardiovascular Heart Diseases (CHD)

Jorge Méndez-Benegassi[1], Augusto Díaz-Leante[2], Álvaro Barja[3], and Fernando Heredero[4]

[1]A20521628
[2]A20521606
[3]A20521894
[4]A20521792

May 1, 2023

Contents

Abstract

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States, specifically, one person dies every 34 seconds in the United States from cardiovascular disease (CVD). Just in 2019, CVD caused around 690,000 deaths in the country; the sad part is that a lot of these deaths could be prevented if people took steps to lower their risk.

High blood pressure, smoking, high cholesterol, or diabetes, which all increase the risk of developing CVD are very common problems among the people that live in the US. However, it is even worse for certain groups of people, like African Americans, and Hispanics/Latinos as they tend to have higher rates of CVD and worse outcomes.

It is important to note that even though there have been some improvements over the last few years in how we prevent and treat CVD, there is still a long way to go to reduce the numbers we are currently at.

The problem statement is to identify and analyze the risk factors associated with cardiovascular heart disease (CHD) and evaluate the current prevention and treatment strategies to identify areas for improvement. Besides, this project aims to develop recommendations for healthcare providers and policymakers to improve prevention and treatment strategies for CHD.

Keywords— heart, cardiovascular, disease, classification, data

1 Overview

Data mining can help the current situation of cardiovascular problems by analyzing large amounts of patient data to identify risk factors associated with the disease. This information can produce effective and valuable knowledge, which is essential for accurate clinical decision-making and risk assessment[1]. Additionally, data mining can help in many other fields of medicine by identifying patterns and trends through various sources of patient data, which can be used to improve diagnosis, predict outcomes, and develop personalized treatment plans.

The project has been structured into four different sections in which we will analyze data, process it, try different models and lastly test its performance.

We will start by preprocessing the dataset [2] to ensure that the values don't have any anomalies that may have a negative impact on our analysis. We will search for invalid types, outliers, and missing values to then transform the data accordingly. In order to do this we will use some visualization techniques such as boxplot to analyze how the data is distributed.

After this analysis, it seemed necessary to visualize the data to understand how the values were distributed across the datasets and how certain attributes would affect the objective label we were working with. This visualization was done by using pie charts and some bar graphs as they illustrated the best how the data was spread on the dataset.

The data mining stage was the most important part of the project, we have focused the analysis starting with feature selection. The analysis started by trying to reduce the dimensionality of the dataset by using Principal Component Analysis (PCA), we then performed a correlation analysis to identify highly correlated attributes and finished with significant feature analysis.

With the proper features selected, we proceed to generate three types of models to see which one performed the best, these models are: Random Forest, Binary Decision Tree, and logistic regression.

Lastly, the models were evaluated using the confusion matrix and the Receiver Operating Characteristic (ROC) curve to determine which of the models had a better performance.

2 Data preprocessing

In this section we describe the data preprocessing performed for future model training.

2.1 Missing and duplicate values

The first thing to do is to check the number of null values and duplicate values. Null values can be empty, incomplete, or corrupted values. These types of values prevent the application of certain algorithms and the extraction of statistics. Therefore, if they exist, they should be eliminated. On the other hand, duplicates simply add redundancy. However, this information can be valuable and give strength to the results, for example, when identifying patterns. Therefore, depending on the use case, different strategies can be applied. In our case, we have chosen to maintain this redundancy.

After scanning the dataset, we have concluded that there are no 0 null values and 77 duplicate rows.

2.2 Outliers

The next step is to identify the outliers. These outliers, or extreme values, are observations that are inconsistent with the rest of the samples in the dataset. For their detection, boxplots are applied to the numerical attributes with respect to the target "cardio". Thus, the following results are obtained:

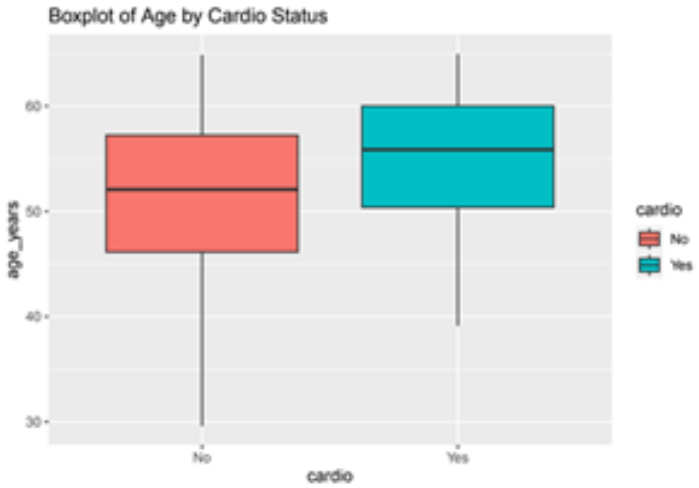

Figure 1: Boxplot of feature age.

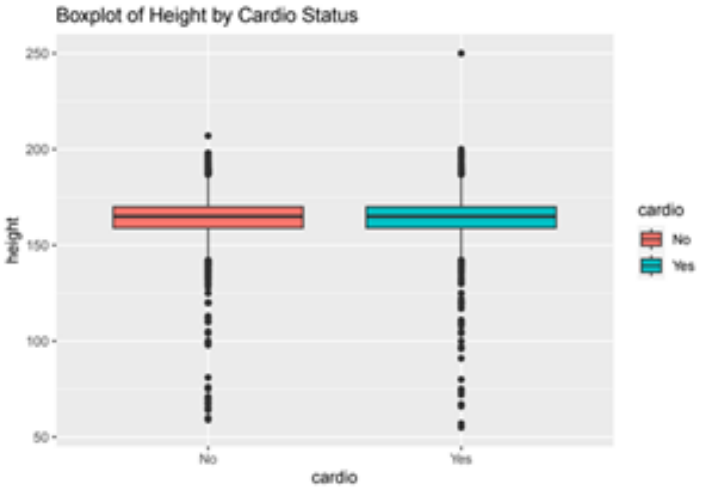

Figure 2: Boxplot of feature height.

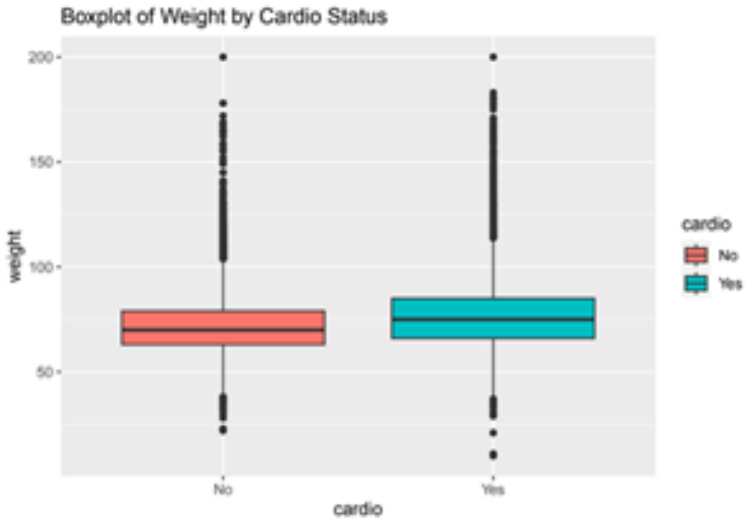

Figure 3: Boxplot of feature weight.

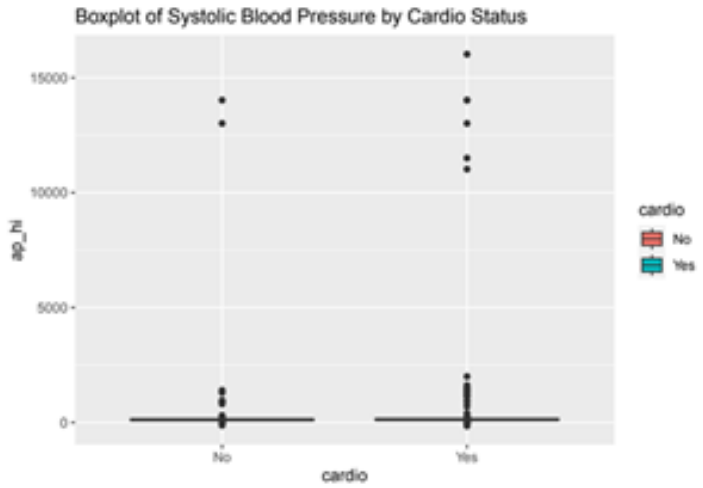

Figure 4: Boxplot of feature systolic blood pressure.

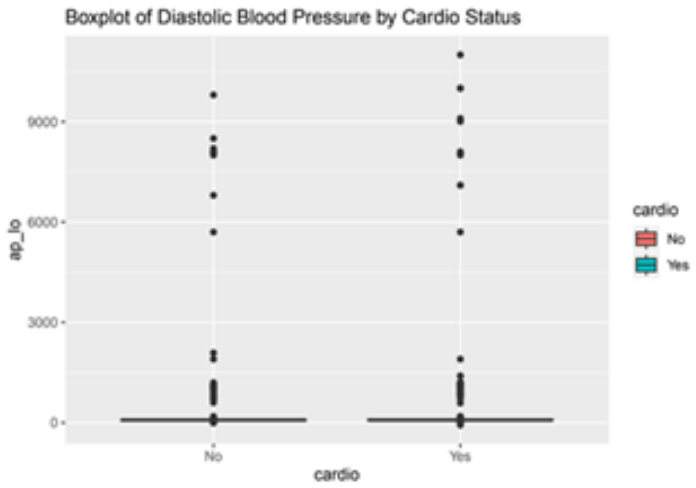

Figure 5: Boxplot of feature diastolic blood pressure.

Taking a look at the boxplots, we may consider outliers of the features height and weight to any data point that falls more than 1.5 the IQR below the first quartile or 1.5 above the third quartile. However, following this method we find height values such as 187 cm considered outliers, therefore, it is better to consider weight and height outliers to those values with no sense with respect to the age. In this way, we will first see the distribution of height and weight according to the age and extract conclusions.

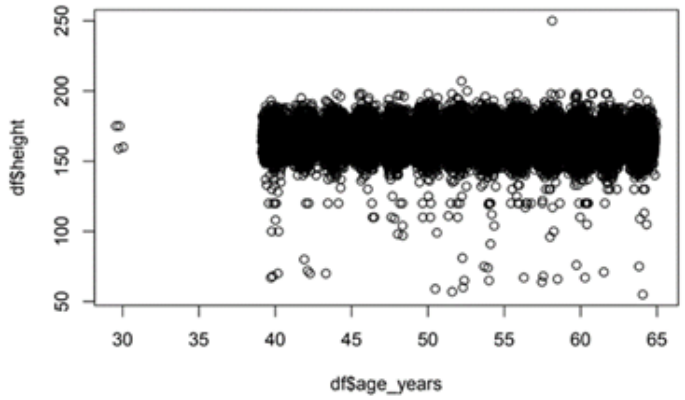

Figure 6: Scatterplot of feature height observations.

Since we have ages between 29.58 and 64.97, we will consider height outliers to any data point that falls out of the normal height values range, which is [130, 210].

We have done the same for the attribute weight, and extract the following scatter plot:

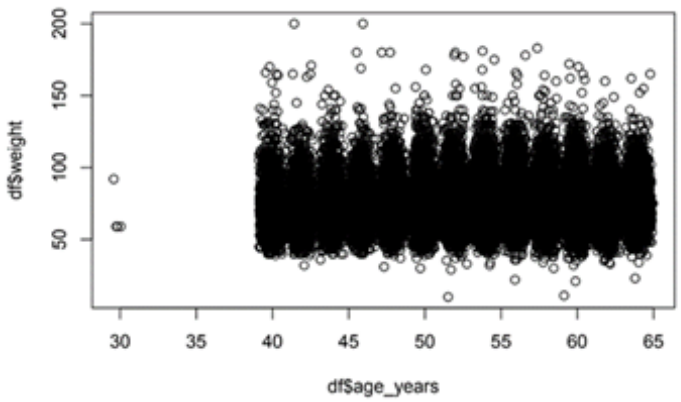

Figure 7: Scatterplot of feature weight observations.

As we can see, there are no significant outliers. There might be some high values, but it can be related to obese individuals. This is a relevant factor for cardiovascular disease prediction, thus, we won't remove any weight value.

For the feature *ap_hi* (systolic) we consider outliers to those values higher than 180 mmHg which would correspond to a stage of hypertensive crisis, requiring immediate medical attention [3]. In the same way, for the feature *ap_lo* (diastolic) we consider outliers to those values higher than 120 mmHg.

Eventually, all these outliers are removed.

2.3 Data transformation

Now our dataset contains a total of 1439 rows less. Regarding the data transformation, the attribute "age" has been modified to give the information in years instead of days. In addition, the target *cardio* has been transformed into a categorical attribute with two levels: Yes / No.

On the other hand, the attribute "gender" has been converted into a binary attribute, and cholesterol and glucose have been transformed into categorical attributes with three levels (0, 1, and 2).

Once these last two attributes have been transformed, dummy variables have been created and can be used for statistical purposes.

3 Data exploration

3.1 Statistics

In this section the important statistics of categorical features of the data used for the project are detailed. In the following table, the statistics are shown.

Feature	**Number**	**Percentage %**
Men	23880	34.83
Women	44681	65.17
Smokers	6013	8.77
Alcohol consumers	3647	5.32
Glucose level normal	58334	85.08
Glucose level above average	5028	7.33
Glucose level well above average	5199	7.58
Cholesterol level normal	51494	75.11
Cholesterol level above average	9253	13.5
Cholesterol level well above average	7814	11.4

Table 1: Feature statistics.

The presented visualizations depict the data distribution for the "Glucose level" and "Cholesterol level" features. It is evident that the cholesterol levels have a higher prevalence of values exceeding the normal

cholesterol range, compared to the glucose levels. These features hold significance in identifying cases of heart disease.

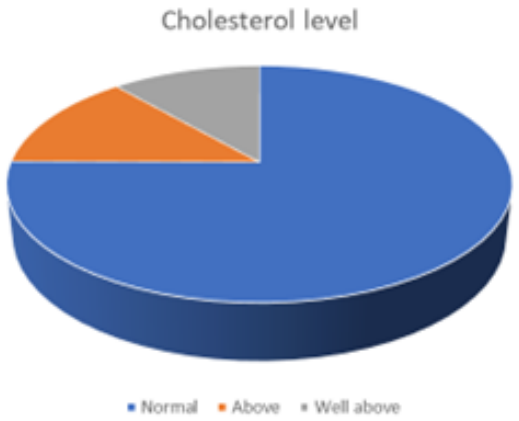

Figure 8: Cholesterol level.

Monitoring cholesterol levels is a crucial measure of heart health. Elevated levels of low-density lipoprotein (LDL) cholesterol can be particularly worrisome as it can contribute to the accumulation of plaque in the arteries. Thus, it is imperative to represent the data distribution for this feature in a visual format.

Figure 9: Glucose level.

When it comes to glucose levels, elevated levels of glucose in the bloodstream can be a contributing factor to various health complications, such as heart disease. As a result, it is crucial to represent the distribution of this particular feature visually.

In the subsequent table, you can find the statistical data for continuous features such as weight, height, and age (in years) for the entire dataset.

	Minimum	**Mean**	**Maximum**
Weight (Kg)	11.00	74.06	200.00
Height (cm)	130.00	164.44	207.00
Age	29.58	53.32	64.97

Table 2: Continuous feature statistics.

The provided statistics offer a comprehensive overview of the dataset used in the model. Additionally, it is worth noting that weight and age are relevant factors in terms of heart diseases. Therefore, these features should be taken into consideration when examining the data and identifying potential cases of heart disease.

3.2 Advanced statistics

In this section, the percentage of individuals with cardio problems was calculated based on their gender. The results indicate that 49.77% of men and 49.% of women had cardio problems. The similarity of these results suggests that gender is not a dependent factor in the development of heart disease.

The subsequent statistics outline the percentage of heart disease cases based on glucose and cholesterol levels.

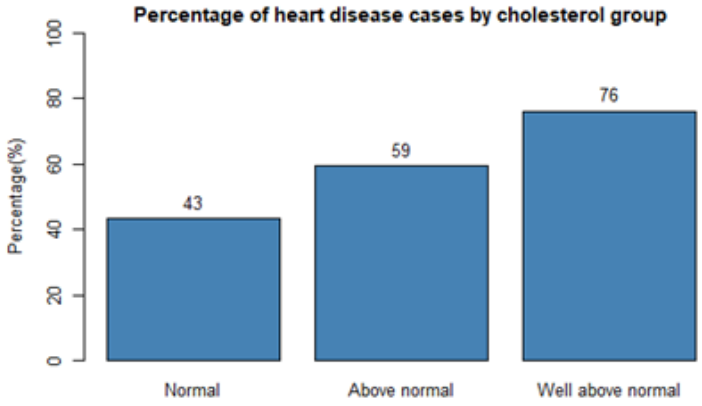

Figure 10: Percentage of heart disease cases by cholesterol group.

The displayed plot indicates a noticeable increase in the probability of developing heart disease in relation to cholesterol level. Specifically, the increment in the probability of heart disease is 16% when the cholesterol level is above the normal range, in comparison to individuals with normal cholesterol levels. The risk of heart disease is even higher (17% and 33% compared to normal) for individuals with well above normal cholesterol levels. It is important to note that these levels were classified based on the following LDL cholesterol ranges:

- Normal level of cholesterol: less than 200 mg/dL.
- Above the normal level of cholesterol: 200 mg/dL - 239 mg/dL.
- Well above the normal level of cholesterol: above 240 mg/dL.

The subsequent plot represents the percentage of heart disease cases based on the glucose level of individuals.

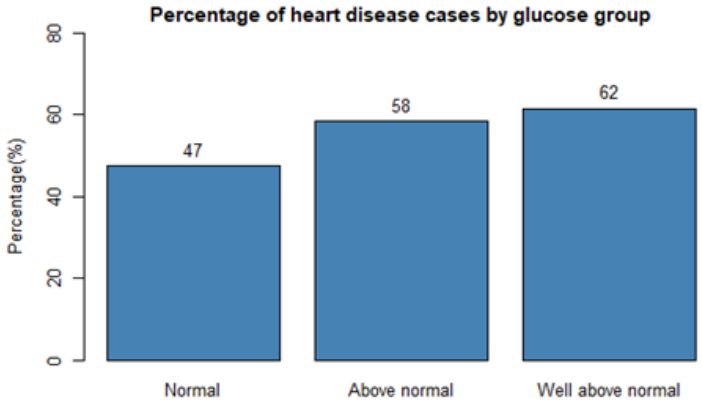

Figure 11: Percentage of heart disease cases by glucose group.

The provided plot shows an increase in the probability of developing heart disease in individuals with glucose levels above the normal range. The increment in risk is 11% for individuals with above-normal glucose levels compared to those with normal levels. However, the increment of risk is only 4% for individuals with well above normal glucose levels in comparison to those with above normal levels. Nonetheless, the risk of

heart disease is substantially higher (15%) for individuals with well above normal glucose levels compared to those with normal levels.

It should be noted that these levels were classified based on the following glucose ranges:

- Normal glucose level: less than 100 mg/dL.
- Above normal glucose level: 110 - 125 mg/dL.
- Well above normal glucose level: above 126 mg/dL.

4 Data mining

After completing the data transformation and data exploration stages, the next phase to concentrate on is data mining.

In this phase, we will dive into the data and employ various techniques to uncover valuable insights, patterns, and trends that can help us answer our research questions. The data mining phase will enable us to apply algorithms and models to the data to identify hidden patterns and relationships that would be difficult to detect through manual analysis.

The selection of the model is a critical decision, and it can have a significant impact on the accuracy and effectiveness of the results.

4.1 Feature selection

The initial step of the data mining phase is feature selection, where we select the most relevant attributes from the data that will be used in the analysis. This step is essential because not all features contribute equally to the performance of the model. Selecting the right set of features can significantly improve the model's accuracy, while using irrelevant or redundant features can lead to overfitting or poor model performance. Feature selection can also help to reduce the complexity and computational burden of the model, making it easier to interpret and implement.

First of all, we tried reducing the dimensionality of the dataset with a Principal Components Analysis (PCA). PCA works by finding the linear combinations of the original features that account for the most variance in the data, resulting in a new set of features that capture the most important information in the dataset.

In order to avoid a biased estimation of the principal components, the attributes are scaled beforehand, resulting in a standardized dataset with a mean of zero and a standard deviation of one. Then, after performing the PCA we obtained the principal components as well as the Cumulative and Not Cumulative Variance Explained plots:

	PC1	PC2	PC3	PC4	PC5	PC6
gender	-0.019299906	0.473987426	-0.27232249	0.09631200	-0.02667212	0.139010902
height	-0.008313125	0.452123153	-0.25767221	0.10327731	-0.16364497	0.388377253
weight	-0.208072014	0.312117457	0.00937784	0.01294855	-0.20425707	0.297780121
ap_hi	-0.301505336	0.261302543	0.40042680	-0.05575893	-0.06883631	-0.146045102
ap_lo	-0.277072503	0.270911771	0.38063866	-0.05158977	-0.08701425	-0.140348517
cholesterol0	0.436452568	0.138177402	0.12704409	0.16430721	-0.43129494	-0.258775525
cholesterol1	-0.223636512	-0.049132497	-0.12971191	-0.60205015	0.31347237	0.240668656
cholesterol2	-0.353415461	-0.135187580	-0.03339775	0.42376810	0.24980390	0.093339504
gluc0	0.404191924	0.208838657	0.31891947	-0.04924837	0.38955724	0.179599823
gluc1	-0.232567429	-0.087285541	-0.25073273	-0.41432889	-0.48438142	-0.244031543
gluc2	-0.314917429	-0.195088370	-0.18227486	0.47428296	-0.04725087	-0.001385995
smoke	-0.023796115	0.347060006	-0.29551528	0.02488778	0.25748102	-0.363322220
alco	-0.041175448	0.252268260	-0.23415598	-0.01053848	0.30318274	-0.507200204
active	0.008364329	0.003475734	-0.02718571	0.01814434	0.15298576	-0.266297804
age_years	-0.179744868	-0.009106179	0.25159509	0.06959742	0.03719892	-0.118325657
cardio	-0.268163424	0.136277782	0.34066293	-0.01171695	0.01335319	-0.023680960

	PC7	PC8	PC9	PC10	PC11	PC12
gender	-0.0052840696	-0.23644580	0.38877450	-1.573604e-02	0.03497942	0.207683182
height	-0.1011192616	-0.11300065	0.03464625	-9.052982e-05	0.14407334	0.192572901
weight	-0.1223240546	0.11660103	-0.70754630	7.605484e-02	-0.14287711	-0.315036773
ap_hi	-0.0515070780	0.20190275	0.18879341	1.605974e-01	-0.04070135	0.040684047
ap_lo	-0.0699273131	0.28314429	0.21624585	2.708140e-01	-0.08965168	0.106210102
cholesterol0	0.0208537950	-0.01975713	-0.02174136	2.794332e-02	0.13144013	-0.112869821
cholesterol1	-0.0196630190	-0.02361846	0.08604059	1.864648e-01	0.23682493	-0.155909924
cholesterol2	-0.0072333158	0.05227837	-0.06292972	-2.385135e-01	-0.43348844	0.321218374
gluc0	0.0024779791	0.00129546	-0.06910258	-4.047637e-02	-0.12163427	0.067980437
gluc1	0.0001519272	-0.09958471	-0.03050711	-1.993166e-01	-0.24943400	0.198840115
gluc2	-0.0034843210	0.09632201	0.12303566	2.507463e-01	0.40931633	-0.287290328
smoke	0.1551752017	-0.01566676	0.15960845	-1.060765e-01	-0.37300025	-0.598357950
alco	0.1943416274	0.14524787	-0.42706930	8.994933e-02	0.34988812	0.398342053
active	-0.9290320028	-0.18078624	-0.03397925	-6.082266e-02	0.05630838	-0.025081708
age_years	0.1724598210	-0.83240942	-0.15105912	3.551837e-01	-0.07583028	0.003437931
cardio	0.1079721615	-0.16426685	-0.01054029	-7.409187e-01	0.41426176	-0.152730743

	PC13	PC14	PC15	PC16
gender	-0.647738503	1.560963e-02	-1.683094e-14	-6.077011e-15
height	0.676397517	-3.589059e-02	-2.278390e-15	-1.399502e-15
weight	-0.271497664	1.735123e-02	7.252809e-16	-1.165793e-15
ap_hi	0.024274962	-7.365758e-01	9.332224e-16	2.843779e-16
ap_lo	0.077965418	6.666795e-01	-1.347902e-16	-8.466046e-16
cholesterol0	-0.014468275	-1.721350e-03	2.383557e-01	6.365363e-01
cholesterol1	-0.006649889	7.850097e-03	1.883504e-01	5.029956e-01
cholesterol2	0.026836891	-6.098393e-03	1.751732e-01	4.678056e-01
gluc0	-0.003936558	-2.285053e-03	-6.481548e-01	2.427063e-01
gluc1	0.027605162	3.525905e-05	-4.742868e-01	1.776002e-01
gluc2	-0.021886417	3.040363e-03	-4.816351e-01	1.803518e-01
smoke	0.185225210	1.460231e-02	1.990153e-17	-1.395129e-16
alco	-0.005354042	-9.053549e-03	1.129383e-16	8.321385e-17
active	0.002847230	9.196033e-03	1.018479e-16	-1.273763e-16
age_years	0.077158239	2.725148e-02	-8.277571e-17	2.203695e-16
cardio	0.003673905	9.961404e-02	1.723512e-16	1.590878e-16

Figure 12: Principal Components information.

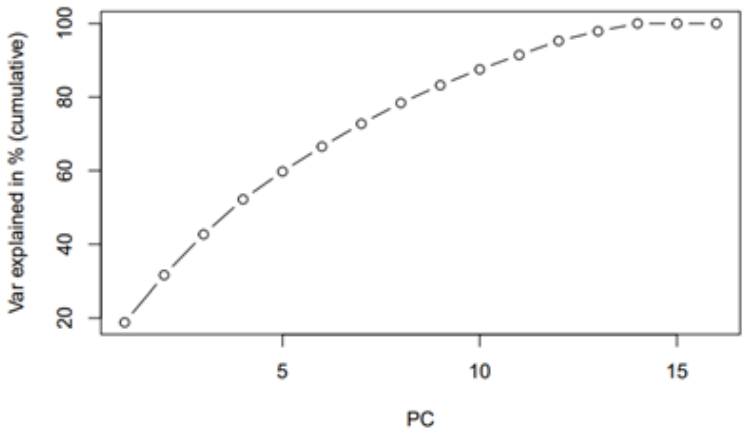

Figure 13: Percentage of variance explained by Principal Components.

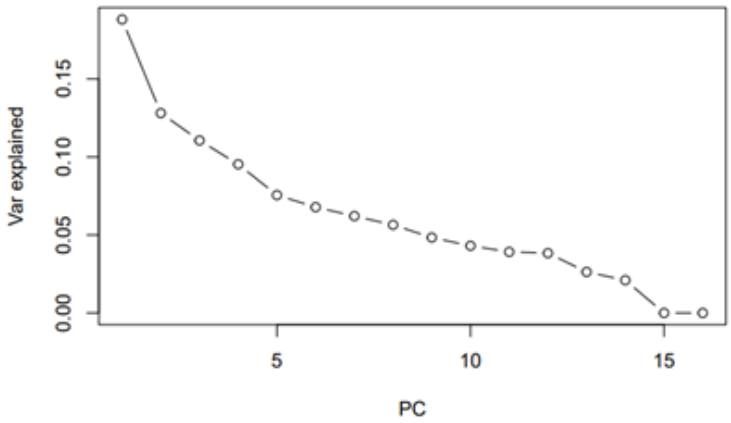

Figure 14: Variance explained by Principal Components.

By looking at the Cumulative Explained Variance plot, we can see that the first 10 PCA's are needed to explain around 85% of the variance. Then, since the first few principal components do not account for a significant amount of variance in the data, we decide to discard using PCA as a dimensionality reduction technique.

We then decide to continue with the feature selection, but this time performing a correlation analysis. It is important to identify and remove highly correlated attributes before conducting any analysis, since they can introduce redundancy and distort the results.

In order to identify highly correlated attributes in our dataset, we performed a correlation matrix analysis. We carefully evaluated the matrix and found that there were no values with an absolute value high enough to suggest that any of the features were highly correlated and needed to be removed.

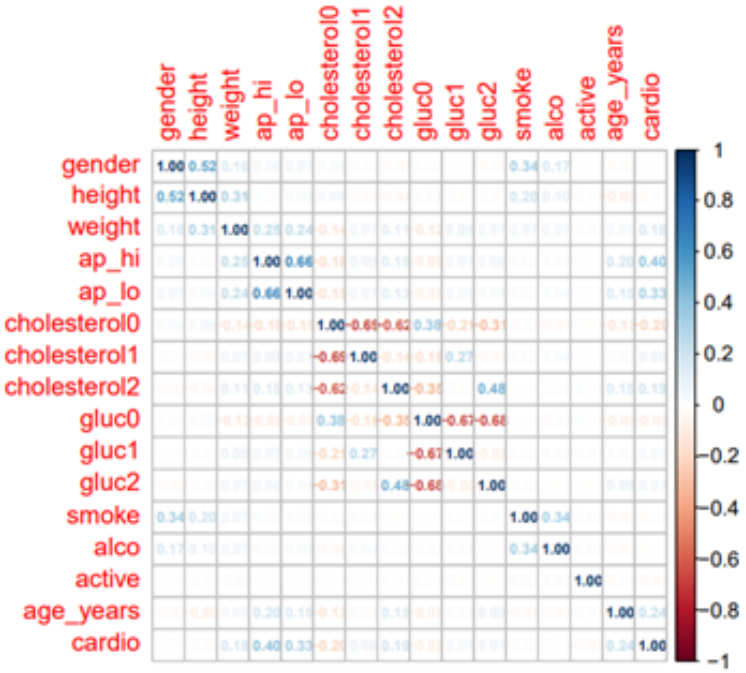

Figure 15: Correlation matrix.

The last technique we employed in our feature selection process was to consider only the statistically significant features. We did this because statistically significant features are the ones that have the most impact on the outcome variable, and by including only these features, we can improve the accuracy and reliability of our analysis.

To identify the statistically significant features, we performed hypothesis testing, which allows us to evaluate the significance of each feature based on its p-value. Based on the results of our testing, we found that the

Gender attribute was the only non-significant attribute, with a p-value higher than 0.05.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.6534870	0.0075450	86.612	< 2e-16	***
gender	-0.0009582	0.0043451	-0.221	0.825	
height	-0.0085488	0.0020783	-4.113	3.90e-05	***
weight	0.0342596	0.0018710	18.311	< 2e-16	***
ap_hi	0.1362601	0.0022887	59.535	< 2e-16	***
ap_lo	0.0477322	0.0022569	21.150	< 2e-16	***
cholesterol0	-0.2089635	0.0062606	-33.378	< 2e-16	***
cholesterol1	-0.1238404	0.0074709	-16.576	< 2e-16	***
cholesterol2	NA	NA	NA	NA	
gluc0	0.0583522	0.0073123	7.980	1.49e-15	***
gluc1	0.0607421	0.0095079	6.389	1.68e-10	***
gluc2	NA	NA	NA	NA	
smoke	-0.0259750	0.0066651	-3.897	9.74e-05	***
alco	-0.0402084	0.0080269	-5.009	5.48e-07	***
active	-0.0448806	0.0042515	-10.556	< 2e-16	***
age_years	0.0726252	0.0017476	41.556	< 2e-16	***

Figure 16: Testing result.

We then moved forward with the updated dataset, which now did not include the Gender attribute.

4.2 Model training

Since the target variable is a binary variable (classification problem) we decided to use three models: binary decision tree, random forest, and logistic regression. These models are well-suited for binary classification problems and can effectively capture the relationships between the input variables and the target variable. They also provide interpretability and flexibility, allowing us to gain insights into the underlying patterns.

Before training these models, we first used the createDataPartition function to split the data into training and testing sets. In this case, we did an 80-20 split, in order to balance between having enough data to train the model effectively and having enough data to test its performance accurately.

Before training our models, it is important to check whether the data set is balanced, meaning that the number of positive and negative instances in the target variable is roughly equal. In this case, we can see that it is a balanced dataset, since there are 27,813 observations with a negative target value (no heart disease) and 27,036 observations with a positive target value (heart disease).

4.2.1 Binary decision tree

The first technique we used was a binary decision tree. It is a useful approach for classification problems and it results in a tree-like structure where each internal node corresponds to a decision based on one of the input features and each leaf node corresponds to a class label.

This is the resulting tree:

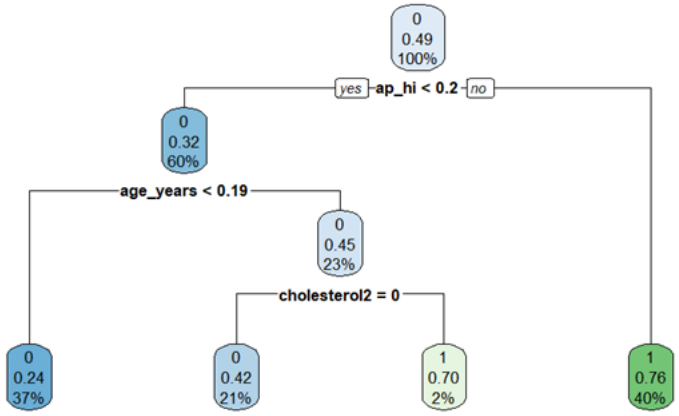

Figure 17: Binary decision tree.

4.2.2 Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees to improve the accuracy and stability of the model. It works by creating a large number of decision trees, each trained on a random subset of the features and a random subset of the data. The trees are then combined to make a final prediction, with each tree having an equal vote. The results will be shown in the Model Fitting section.

4.2.3 Logistic regression

Logistic regression is a statistical method used for binary classification problems. It estimates the probability of an event occurring based on a given dataset of independent variables, also known as features or predictors.

In our case, we have used logistic regression to predict the probability of heart disease presence in our dataset. If the resulting probability is higher than 0.5, the target variable is positive, indicating the presence of heart disease, and if not, it is negative, indicating the absence of heart disease.

4.3 Model fitting

Now that we have trained the Binary Decision Tree, Random Forest, and Logistic Regression models with the training dataset, the next step is to evaluate their performance with the test dataset. We fit the models to the test dataset to predict the target variable, based on the input features. By doing this, we can assess how well the models generalize to new data and make accurate predictions. The obtained accuracies are:

- **Binary decision tree**: 0.7247
- **Random Forest**: 0.7368
- **Logistic regression**: 0.7285

Based solely on the accuracy results, the Random Forest model performed the best with an accuracy of 0.7363, followed closely by the Logistic Regression model with an accuracy of 0.7285. The Binary Decision Tree model had the lowest accuracy with 0.7247.

However, accuracy is not the only metric to consider when selecting a model. It is also important to consider other metrics. The choice of the model depends on the specific problem, the data, and the trade-offs between different performance metrics. Therefore, a more in-depth analysis and comparison of the models based on different metrics and considerations is performed before selecting a final model.

5 Model evaluation

A deeper analysis of the results of each model is necessary to select the best model.

When analyzing the performance of the model, one of the essential tools to use is the confusion matrix, which provides valuable information, such as true positives, true negatives, false positives, and false negatives, that can be used to calculate various metrics, including accuracy, precision, recall, and F1 score.

Sensitivity and specificity are two important metrics used to evaluate the performance of a classification model, particularly in the medical field. Sensitivity refers to the proportion of true positives, or the percentage of actual positive cases that are correctly identified by the model. On the other hand, specificity measures the proportion of true negatives, or the percentage of actual negative cases that are correctly identified by the model. In medical diagnosis, false negatives can be life-threatening, while false positives can lead to unnecessary treatment or procedures, making sensitivity and specificity of utmost importance. A balance between sensitivity and specificity is required to ensure that the model is both sensitive to the target condition and specific to the exclusion of other conditions

Prediction/Reference	**0**	**1**
0	5519	2355
1	1420	4418

Table 3: Binary decision tree confusion matrix.

- **Sensitivity**: 0.7954
- **Specificity**: 0.6523

Prediction/Reference	0	1
0	5523	2193
1	1416	4580

Table 4: Random Forest confusion matrix.

- **Sensitivity**: 0.7959
- **Specificity**: 0.6762

Prediction/Reference	0	1
0	5544	2228
1	1495	4545

Table 5: Logistic regression confusion matrix.

- **Sensitivity**: 0.7846
- **Specificity**: 0.6710

With these values, we can calculate more metrics:

	Prediction/Reference	0	1
Accuracy	0.725	0.737	0.7285
Precision	0.701	0.716	0.71
Recall	0.795	0.796	0.785
F1 Score	0.745	0.754	0.745

Table 6: Confusion matrices metrics.

Based on the evaluation metrics, it appears that the Random Forest model is the best-performing model among the Binary Decision Tree, Random Forest, and Logistic Regression models. In addition to having the highest accuracy, precision, and recall, the Random Forest model also has the best sensitivity and specificity values, which are really important in the medical field. However, it is important to note that the differences in performance metrics between the models are relatively small, and the decision is not entirely clear.

An additional evaluation of the model's performance can be performed using the Receiver Operating Characteristic (ROC) curve. The ROC curve is a graphical representation of the trade-off between the true positive rate (TPR) and the false positive rate (FPR) as the discrimination threshold is varied. A higher AUC value indicates better discrimination performance of the model. By using the ROC curve, we can better visualize the performance of the model and get a more comprehensive understanding. Therefore, in addition to the evaluation metrics calculated earlier, the ROC curve and AUC should also be considered when selecting the final model.

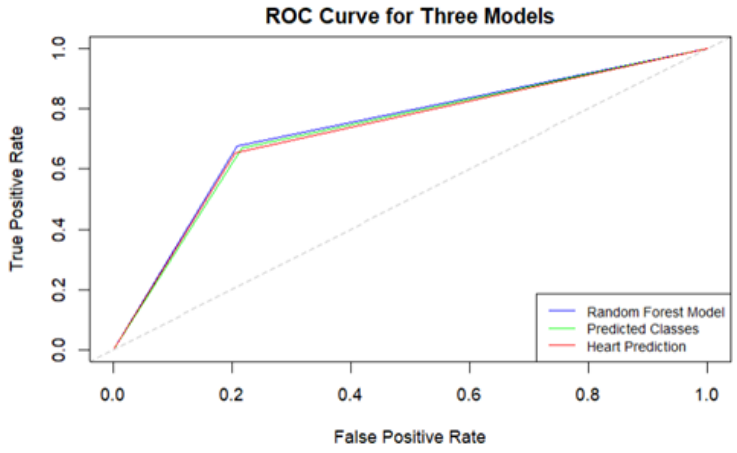

Figure 18: ROC curve.

According to the ROC curve, the Random Forest model is the best performing model among the three models. The Random Forest model's ROC curve is closest to the top-left corner of the plot, indicating that it has the highest True Positive Rate and the lowest False Positive Rate for a range of threshold values. This suggests that the Random Forest model has a better ability to distinguish between the positive and negative classes compared to the other models.

Therefore, based on both the evaluation metrics and the ROC curve, the **Random Forest model is the selected model** for this binary classification task.

Finally, it is good to say that since the selected model for this binary classification task is the Random Forest model, which is a combination of multiple decision trees, it is not necessary to perform cross-validation techniques. This is because the Random Forest model already performs internal cross-validation during the training process, where each decision tree is built on a randomly selected subset of the training data and tested on the remaining data.

6 Conclusions

This project aimed to investigate the significance of various features in predicting the occurrence of heart disease. The study identified cholesterol, glucose level, and age as crucial variables, while gender was found to be irrelevant. Consequently, gender was excluded from the training and testing datasets.

To predict the binary outcome variable, three models, namely, binary decision tree, random forest, and logistic regression, were studied. Upon evaluating each model's performance, it was concluded that the random forest model produced the best metrics. The evaluation process involved a comprehensive analysis of various metrics.

This project opens up new avenues for future research. Firstly, it would be beneficial to employ a different dataset with distinct relevant features such as family history or stress. Secondly, exploring other ensemble methods like gradient boosting could be useful. Lastly, alternative model selection methods such as support vector machines, naïve Bayes, or k-nearest neighbors could potentially improve the model's performance.

References

[1] Roeters van Lennep, Jeanine E and Westerveld, H.Tineke and Erkelens, D.Willem and van der Wall, Ernst E, «Risk factors for coronary heart disease: implications of gender,» *Cardiovascular Research*.

[2] Kuzak Dempsy, «Kaggle,» [Online]. Available: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas.

[3] Khot, Umesh N. and Khot, Monica B. and Bajzer, Christopher T. and Sapp, Shelly K. and Ohman, E. Magnus and Brener, Sorin J. and Ellis, Stephen G. and Lincoff, A. Michael and Topol, Eric J., «Prevalence of Conventional Risk Factors in Patients With Coronary Heart Disease,» *JAMA*.

A Source code

For the visualization of the code, visit the following link: https://github.com/fherederolopez/CSP-571