# Risk factors for Cardiovascular Heart Diseases (CHD)

Project Team Members:

Jorge Méndez-Benegassi Márquez - A20521628 - Group leader.
Augusto Díaz-Leante Ferreiro - A20521606
Álvaro Barja Galles - A20521894
Fernando Heredero López - A20521792

## 1 Project Proposal

### 1.1 Research goal

The primary goal of this project is to identify and analyze the risk factors associated with cardiovascular heart disease (CHD) in order to better understand the disease and improve prevention and treatment strategies. Specifically, the project aims to:

1. Identify and analyze the risk factors associated with CHD, including but not limited to age, gender, height, weight, smoking, alcohol drinking, BMI and cholesterol.
2. Examine the relationship between these risk factors and the incidence and severity of CHD.
3. Evaluate the effectiveness of current prevention and treatment strategies for CHD and identify areas for improvement.
4. Develop recommendations for healthcare providers and policymakers to improve prevention and treatment strategies for CHD.

### 1.2 Specific questions

The project seeks to address the following specific questions:

1. What are the most significant risk factors associated with CHD, and how do they contribute to the incidence and severity of the disease?
2. How effective are current prevention and treatment strategies for CHD, and what gaps exist in these strategies?
3. What is the relationship between different risk factors and CHD outcomes, such as morbidity and mortality?
4. How can healthcare providers and policymakers improve prevention and treatment strategies for CHD to reduce the burden of the disease on individuals and society?

### 1.3 Proposed methodology

The proposed methodology for the analysis of the project will involve the following stages:

Data Preprocessing: The data preprocessing stage will involve cleaning, reduction, and transformation of the data. This stage will ensure that the data is in a suitable format for analysis. Cleaning will involve identifying and correcting any errors, inconsistencies, or missing values in the data. Reduction will involve selecting a subset of variables that are relevant to the analysis. Transformation will involve converting variables into a format that is suitable for analysis, such as normalizing or standardizing variables.

Data Exploration**:** The data exploration stage will involve using statistical methods and visualization techniques to gain insights into the data. Descriptive statistics, such as means, standard deviations, and correlation coefficients, will be calculated to summarize the data. Visualizations, such as scatter plots, histograms, or box plots, will be used to identify patterns and relationships in the data.

Data Mining: During the data mining stage, the most important variables that contribute to the prediction of CHD risk will be identified through feature selection. Models will then be developed to predict CHD risk based on the selected features, utilizing various classification algorithms such as logistic regression, decision trees, and random forests to determine the most accurate and efficient model. Model selection will be done using cross-validation techniques to assess the performance of the models on new data. Finally, the performance of the final model will be evaluated in terms of accuracy, sensitivity, specificity, and other relevant metrics.

Overall, the proposed methodology will enable a comprehensive analysis of the risk factors associated with CHD and contribute to the development of effective prevention and treatment strategies for the disease.

### 1.4 Metrics

A set of metrics will be used to measure the analysis results of the project, including:

Accuracy: This metric will measure the proportion of correctly classified instances by the model, which is important for determining the overall effectiveness of the model in predicting CHD risk.

Sensitivity: This metric will measure the proportion of true positive instances that were correctly identified as positive.

Specificity: This metric will measure the proportion of true negative instances that were correctly identified as negative.

Area under the ROC curve (AUC): This metric will measure the overall performance of the model in discriminating between positive and negative instances. A higher AUC indicates better model performance in identifying individuals at high risk for CHD.

Precision: This metric will measure the proportion of true positive instances out of all instances predicted as positive.

These metrics will be used to evaluate the performance of the models developed during the data mining stage and to compare the relative effectiveness of different classification algorithms and feature selection methods.

## 2. Project Outline

### 2.1 Literature review and related work- existing projects, references, papers, and relevant articles, etc.

- Academic Papers:
  - Prevalence of Conventional Risk Factors in Patients with Coronary Heart Disease. [1]
  - Risk factors for coronary heart disease [2]

## 2.2 Data sources and references

Overview:

This dataset is extracted from a Kaggle competition [3] [4]. This dataset contains detailed information on the risk factors for cardiovascular disease. It includes information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits and alcohol consumption of over 70 thousand individuals. Additionally, it outlines if the person is active or not and if he or she has any cardiovascular diseases.

Application:

The dataset presented here offers a valuable opportunity for researchers to utilize state-of-the-art machine learning methods to investigate the possible links between risk factors and cardiovascular disease. Such exploration can ultimately enhance comprehension of this critical health concern and facilitate the development of more effective preventive measures.

Feature description:

Details of the dataset from Kaggle, are explained below:

| Field Name | Type | Description |
|---|---|---|
| Age | Integer | Age of the individual (days) |
| Gender | Binary | Gender of participant |
| Height | Integer | Height measured in centimeters |
| Weight | Integer | Weight measured in kilograms |
| Ap_hi | Integer | Systolic blood pressure reading taken from patient |
| Ap_lo | Integer | Diastolic blood pressure reading |
| Cholesterol | Integer | Cholesterol level of the individual |
| gluc | Integer | Glucose level of the individual |
| smoke | Binary | Smoking status of the individual |
| alco | Binary | Alcohol consumption status of the individual |
| active | Binary | Physical activity level of the individual |
| cardio | Binary | Presence or absence of cardiovascular disease |

Notes:

The cholesterol feature is categorized in three levels. Being the first level the lowest and the three level corresponds the category that covers the individuals with a high cholesterol. The glucose level is categorized in the same manner that cholesterol. Smoke, alco, and active are Boolean values which states if the individual smokes or not, drinks alcohol or not, or if it is an active person.

Regarding considered an active or non-active individual is the most subjective feature in the dataset. For that reason, it will be necessary to work on it carefully to consider or in other case, remove it from the dataset. Finally, the cardio feature which shows if the individual suffers a cardiovascular disease (1) or not (0).

## 2.3 Data processing and pipeline

Data processing:

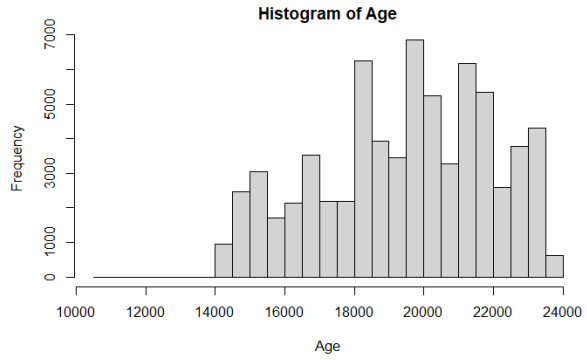
- Importation of data from Kaggle data source.
- Checking of data to remove duplicates and outliers.
- Split the dataset into training and testing sets.
- Possibility to create new variables for the dataset.

Pipeline:

- Scaling numerical features using standardization or normalization
- Feature selection using correlation analysis.

## 2.4 Data stylized facts

Regarding distributional analysis, this example shows an histogram which represent the age distribution of the individuals. This can help us to identify the age groups that may be more susceptible to cardiovascular disease.




Besides, a clustering technique in blood pressure feature could be interesting to identify the blood pressure patterns that are associated with increased risk of cardiovascular disease.

## 2.5 Model selection

The goal of the model is to predict whether an individual has a cardiovascular disease based on their demographic characteristics, lifestyle habits, and biological markers. It would be used several classification algorithms such as logistic regression, decision trees. We will compare different models evaluating their performance.

## 2.6 Software packages

Software packages, applications, libraries, and associated tools

Libraries/Packages:

ggplot2, dplyr, tidyr, caret, randomForest, ROCR, e1071

Software:

RStudio, R

# 3. References

[1] Khot, U.N. (2003) “Prevalence of conventional risk factors in patients with coronary heart disease,” JAMA, 290(7), p. 898. Available at: https://doi.org/10.1001/jama.290.7.898.

[2] Roeters van Lennep, J. (2002) “Risk factors for coronary heart disease: Implications of gender,” Cardiovascular Research, 53(3), pp. 538–549. Available at: https://doi.org/10.1016/s0008-6363(01)00388-1.

[3] Devastator, T. (2023) Risk factors for cardiovascular heart disease, Kaggle. Available at: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas.

[4] Kuzak Dempsy's datasets (2021) data.world. Available at: https://data.world/kudem