# Pre-Project

*Machine Learning Course*

Major: Cloud Computing and Cybersecurity (Group CCC1)

**Supervised By:**

Nédra MELOULLI
Christophe RODRIGUES

**Written By:**

Alex QUILIS VILA
Carolina ROMERO DELGADO
Livia GAMERO HAMMERER

Date Last Edited: November 8, 2024

# Contents

# Chapter 1

# Introduction and State-of-the-Art

## 1.1 Introduction

Heart disease remains one of the leading causes of death worldwide. Early diagnosis is critical to preventing fatalities and improving patient outcomes. The challenge is to develop a machine learning model that can predict heart disease based on patient data, helping healthcare professionals identify at-risk individuals and provide timely interventions. This solution can generate value by reducing the strain on healthcare systems through early prevention and personalized treatment, while improving overall patient care.

## 1.2 State-of-the-Art

Current techniques in heart disease prediction involve using machine learning models like Logistic Regression, Decision Trees, and more advanced methods like Support Vector Machines and Neural Networks. These models utilize medical data to predict outcomes based on features like age, blood pressure, cholesterol levels, and other clinical variables. A key challenge in these models is managing imbalanced data, where patients without heart disease outnumber those with it. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) are often used to address this issue.

# Chapter 2

# Data Description and Sources

## 2.1 Dataset Overview and Key Variables

The dataset consists of medical records from individuals, containing 14 variables related to heart health. The target variable (num) indicates whether a person has heart disease, while the features describe patient demographics and clinical measurements, such as age, cholesterol levels, resting blood pressure, and chest pain type. Notably, two features (ca, thal) have missing values that need to be addressed through imputation or data engineering.

- **Age**: Age of the patient in years.
- **Sex**: Gender of the patient.
- **cp (Chest pain type)**: Type of chest pain experienced by the patient.
- **chol (Serum cholesterol)**: Cholesterol level in mg/dL.
- **fbs (Fasting blood sugar)**: Whether fasting blood sugar ¿ 120 mg/dL.
- **thalach (Max heart rate achieved)**: Maximum heart rate during exercise.
- **exang (Exercise-induced angina)**: Whether angina occurred during exercise.
- **oldpeak**: ST depression induced by exercise relative to rest.
- **ca**: Number of major vessels colored by fluoroscopy (with missing values).
- **thal**: Thalassemia test results (with missing values).

- **num**: Target variable, indicating the presence of heart disease.

## 2.2 Data Source

Link to data source [1]: data.

# Chapter 3

# Project Objectives and Development Plan

## 3.1 Primary Business Objective

The main objective of this project is to develop a machine learning model that can accurately predict heart disease in patients based on their medical records. This model aims to support healthcare providers in identifying high-risk patients for further screening and intervention.

## 3.2 Model Development Workflow

1. **Data Cleaning**
   - Task: Handle missing values in ca and thal.
   - Outcome: A clean and complete dataset, ready for model training.
2. **Feature Engineering and Selection**
   - Task: Scale numerical features (e.g., age, trestbps) and encode categorical features (e.g., sex, cp).
   - Outcome: Feature set optimized for machine learning algorithms.
3. **Model Building**
   - Task: Implement and train models like Logistic Regression, Random Forest, and SVM.
   - Outcome: A trained predictive model for heart disease.
4. **Model Evaluation and Tuning**

- Task: Evaluate models using accuracy, precision, recall, and AUC-ROC.
- Outcome: Best-performing model selected for deployment.

## 3.3 Work Plan

1. **Step 1: Pre-project (08/11)**: Choose a relevant subject with guidance from the major head and tutors and define business objectives and scope.
2. **Stage 1: Implementation of Standard Solutions (15/11)**: Analyze and preprocess data, implement solutions using basic algorithms, define learning/testing plan, and critically analyze results.
3. **Stage 2: Improving Solutions (21/11)**: Implement advanced algorithms, update learning/testing plan, critique results, and explore ensemble learning.
4. **Stage 3: Further Improvements (12/12)**: Select the best algorithm, justify choice, and evaluate and compare results.

# Chapter 4

# Conclusions

This project aims to address the critical challenge of early heart disease detection by leveraging machine learning models trained on medical data. By accurately predicting heart disease risk, the solution can provide healthcare professionals with a valuable tool to offer preventive care and tailored treatments. The project will follow a structured approach, from data preparation to model evaluation, ensuring a robust and reliable outcome.

# Bibliography

[1] Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.