Predictive model for Coronary Disease

Scientific Programming Final Project

Blai Crespo Selma David Hidalgo Fàbregas
Miriam Iturralde Aguiló Adam Koershuis i García
Pablo Longán Gasol Aina Mas Tena Marta Meroño Rafel

Index

1 Objective of the project

The primary objective of this project is to develop a robust and accurate predictive model to assess the risk of developing coronary heart disease (CHD) over a ten-year period, based on specific clinical markers and patient demographics. Using a dataset comprising relevant health features such as blood pressure, cholesterol levels, and glucose, the study implements a complete data science pipeline involving data cleaning, exploratory analysis, and feature engineering. Machine learning algorithms are trained, optimized, and validated to identify the most effective classification approach. To ensure practical applicability, the final model is encapsulated within a functional API and containerized using Docker, providing a scalable, accessible solution for potential diagnostic support.

2 Distribution of tasks

To distribute the project workload, we followed the task structure suggested in the course guidelines equally dividing the tasks among the seven group members. Contributor 1 was responsible for the initial data cleaning and summary, managing missing values, discarding irrelevant variables, and generating the first statistical overview through histograms and boxplots. Contributor 2 focused on data aggregation and visualization, developing functions to group key categories and creating scatterplots to analyze relationships between variables such as blood pressure and BMI. Contributor 3 handled data normalization - using techniques like z-score or min-max scaling- and conducted correlation analyses to eliminate redundant features.

The predictive engine was built by Contributor 4, who implemented a Random Forest and a Logistic regression classification models to identify high-risk patients. This was followed by Contributor 5, who established the validation strategy, selected the best-performing model, and designed the initial API schema. Contributor 6 then finalized the API by developing functional endpoints, integrating the trained model, and adding technical documentation. Finally, Contributor 7 managed the deployment phase by containerizing the application with Docker and hosting the API on a cloud server. To conclude the project, the entire team collaborated on refining the final model, drafting the comprehensive technical report, and producing a demonstration video of the live deployment.

Due to the task requirements, all the tasks were completed sequentially.

3 Analysis and key results

3.1 Dataset description and preprocessing

The dataset used in this project contains clinical and demographic information related to the risk of coronary heart disease (CHD). The target variable corresponds to the presence or absence of CHD within a ten-year follow-up period. Prior to model training, the data underwent a preprocessing pipeline that primarily involved handling missing values. For managing missing values from variables such as glucose, the population was divided into diabetic and non-diabetic groups. For diabetics with missing values, the median glucose level of the diabetic group was imputed, and for non-diabetics with missing values, the median glucose level of the non-diabetic group was used. Missing BPMeds entries were imputed using the mode of the patient's hypertension group (prevalentHyp). In the BMI and Heart Rate variables, records with missing values were removed due to their low frequency relative to the 4,000+ total records. Missing data for daily cigarette consumption was imputed by assigning a value of zero to non-smokers and applying the group-specific median to current smokers. Education level was removed due to a high number of missing values and their apparent weak relationship with the target variable. Additionally, binary categorical variables were converted to 0 and 1 instead of "yes" and "no."

3.2 Exploratory data analysis

Exploratory data analysis revealed variability across several clinical indicators, particularly in cholesterol levels, systolic blood pressure, and glucose concentration. Outliers were observed in variables associated with cardiovascular risk, which were retained as they represent clinically meaningful extremes rather than measurement errors.

Correlation analysis highlighted strong relationships among certain physiological measurements. Based on these findings, representative features were selected to avoid redundancy while preserving relevant clinical information. Additionally, a moderate class imbalance was identified between CHD-positive and CHD-negative cases, which was taken into account during model evaluation.

An exploratory analysis of the dataset based on graphical representations suggests that certain factors, such as age and hypertension, are more likely to be associated with an increased risk of coronary heart disease within 10 years. Individuals with CHD tend to be older and have higher blood pressure, reflecting the proportion of hypertensive individuals and the observed systolic-diastolic ratio. Other variables, such as total cholesterol and smoking, exhibit milder or non-linear patterns of association, whereas BMI, glucose and heart rate demonstrate more subtle indications of a relationship with CHD. Together with the presence of clinically relevant outliers and correlations between variables, these findings support the careful selection of representative characteristics for multivariate risk models and the comprehensive interpretation of cardiovascular risk.

3.3 Data Normalization and Feature Selection

Following the data cleaning and aggregation phase, we performed data normalization and feature selection to prepare the dataset for model training. First, we applied Min-Max scaling to all continuous numeric variables. We chose Min-Max scaling over standardization (z-score) because our cleaned dataset contains minimal outliers after the previous preprocessing steps, and Min-Max scaling preserves the original distribution shape while ensuring all features contribute equally to distance-based algorithms.This normalization step is critical because variables in our dataset operate on vastly different scales. Next, we conducted correlation analysis within physiologically related variable groups (“Blood Pressure”: systolic and diastolic BP; “Metabolic”: total cholesterol, glucose, and BMI; “Lifestyle”: cigarettes per day, BMI, and age) to identify and remove redundant features. When two variables exhibited high correlation ($|r| > 0.75$), we retained the variable with lower average correlation to other features in the group, as it provides more unique information. By removing redundant variables, we reduce model complexity, which improves computational efficiency, enhances model interpretability, and reduces the risk of overfitting during the classification phase. The output of this phase is a cleaned, normalized dataset with non-redundant features, optimally prepared for the subsequent modeling and classification tasks.

3.4 Feature engineering

To prepare the data for model training, it was necessary to aggregate the dataset so that each subject was represented by a single observation. This was achieved by grouping the data according to categorical variables -namely sex, age, current smoker status, use of blood pressure medication (BPMeds), history of stroke (prevalentStroke), presence of hypertension (prevalentHyp), diabetes status- and the outcome variable TenYearCHD. Although age is a numerical variable, it was treated as categorical in this context, as it represents an intrinsic characteristic of each subject. In contrast, the number of cigarettes smoked per day (cigsPerDay) was retained as a numerical variable. These transformations enabled the model to learn from consolidated and noise-reduced representations of patient data, while also preventing the overrepresentation of specific subject profiles in the presence of class imbalance.

3.5 Model selection and validation

To predict CHD risk, two classification models were implemented: Logistic Regression and Random Forest. These models were selected for their interpretability and strong performance in binary classification tasks. For Logistic Regression, model performance was evaluated by tuning the regularization parameters, while for the Random Forest classifier, the number of trees was explored. Model selection was based on accuracy and sensitivity, complemented by additional evaluation metrics including the confusion matrix, precision, specificity, and F1-score. The optimal configurations were chosen according to the combination of these metrics, resulting in robust and well-balanced predictive performance.

Model validation was performed using the aggregated dataset produced in the previous step of the pipeline. At this stage, the task was formulated as a binary classification problem, with the variable `TenYearCHD` indicating whether an individual developed coronary heart disease within a ten-year period.

The dataset was split into **training (80%) and test (20%)** subsets using stratified sampling to preserve the proportion of positive and negative cases. Model selection was conducted on the training data using 5-fold stratified cross-validation, allowing model performance to be evaluated across multiple data splits while accounting for class imbalance.

Two classification models were evaluated: Logistic Regression, used as a linear baseline, and Random Forest, selected for its ability to capture non-linear relationships between aggregated clinical features. Both models were trained using class weighting to reduce bias toward the majority class. Performance was assessed using ROC-AUC, recall, and accuracy, with ROC-AUC used as the primary criterion for model comparison.

Cross-validation results showed that the Random Forest model achieved higher ROC-AUC values than Logistic Regression. The selected model was subsequently evaluated on the independent test set, where it reached a ROC-AUC of approximately 0.71. Based on these results, **Random Forest was selected as the final model** and saved for integration into the project's API.

3.6 Key results

Among the evaluated models, the selected classifier achieved the best balance between predictive accuracy and generalization performance. The model demonstrated strong recall for CHD-positive cases, which is particularly important in a clinical risk assessment context where false negatives carry significant consequences. These results informed the selection of the final model for deployment within the API.

3.7 Implications for deployment

To make the Heart Disease Prediction easy to access and reliable, the project used a container based deployment approach. The first step to do this is creating a Dockerfile, which will set up a lightweight environment using python:3.10-slim. This Dockerfile handled the installation of the required dependencies, the definition of the project structure (api, src, models), and the start of the FastAPI application using Uvicorn. This Docker based approach ensures consistent API behavior across different environments.

In the final step, the API was deployed on Render. The GitHub repository was connected to Render, allowing automatic builds and deployments whenever the code is updated. Render builds the Docker image and makes the API available through a secure HTTPS endpoint, providing a stable and scalable way to run the API without manual server setup.

4 Conclusions

By leveraging a comprehensive data science pipeline, ranging from meticulous clinical data preprocessing to the optimization of machine learning algorithms, this project developed a reliable predictive tool for cardiovascular risk assessment. The final result is a functional, containerized API that allows users to interact with a high-performance Random Forest model from any environment. By inputting specific patient markers, the system provides an immediate classification of high-risk or low-risk status, demonstrating how integrated AI solutions can be transformed into accessible, scalable, and practical tools for clinical decision support.

5 Link to the API

https://heart-disease-api-fpqb.onrender.com/docs#

6 Link to the GitHub Repository

https://github.com/koershuis/SP_Final_Project_E