Machine Learning Project

Oral defense: May 27 or 29, 2026

Dataset

The data is taken from the KAGGLE competition website; it is the data set "Cardivascular Disease Risk Prediction Dataset" available here: https://www.kaggle.com/datasets/bertnardomariouskono/cardiovascular-disease-risk-prediction-dataset.

This dataset contains 15,000 synthetic patient medical records specifically designed to predict the risk of cardiovascular disease. Although synthetic, the data is generated using medical heuristics to ensure realistic correlations between variables, such as the relationship between age, BMI, and blood pressure. The dataset includes 19 variables for 15,000 patients:

- `Patient_ID` : Unique identifier for each patient
- `Age` : Age of the patient
- `Gender` : Gender (qualitative with two modalities)
- `Height_cm` : Patient's height in centimeters
- `Weight_kg` : Patient's weight in kilograms
- `BMI` : Body Mass Index (kg/m2)
- `Systolic_BP` : Systolic Blood Pressure (mmHg)
- `Diastolic_BP` : Diastolic Blood Pressure (mmHg)
- `Cholesterol_Total` : Total serum cholesterol (mg/dL)
- `Cholesterol_LDL` : Low-Density Lipoprotein / "Bad" cholesterol (mg/dL)
- `Cholesterol_HDL` : High-Density Lipoprotein / "Good" cholesterol (mg/dL)
- `Fasting_Blood_Sugar` : Blood glucose level after fasting (mg/dL)
- `Smoking_Status` : 0: Non-smoker, 1: Smoker (qualitative with two modalities)
- `Alcohol_Consumption` : 0: None, 1: Moderate, 2: Heavy. qualitative with three modalities)
- `Physical_Activity_Level` : Activity level scale (0: Sedentary to 3: High) (qualitative with four modalities)
- `Family_History` : Family history of heart disease (0: No, 1: Yes). (qualitative with two modalities)
- `Stress_Level` : Self-reported stress level (Scale 1 - 10)
- `Sleep_Hours` : Average hours of sleep per night.
- `Heart_Disease_Risk` : 0: Low Risk, 1: High Risk (qualitative with two modalities).

In this project, we first want to predict the variable `Heart_Disease_Risk` from all other variables, and then predict the variable `Cholesterol_LDL` from all other variables (except `Heart_Disease_Risk`).

Questions

Exploratory data analysis (choice of R or Python)

The first step is to explore the different variables, an essential preliminary to the analysis. Below are a few basic questions. You can complete the analysis according to your own ideas.

1. Start by checking the nature of the different variables and their encoding. Do not forget to convert all categorical variables.
2. Start your exploration with a unidimensional descriptive analysis of the data. Do you think transformations of quantitative variables are relevant?
3. Continue with a two-dimensional descriptive analysis. Use visualization techniques such as scatterplot, correlation graphs, parallel boxplots, mosaicplot... What variables seem to be linked?
4. Perform a principal component analysis of quantitative explanatory variables and interpret the results. Visualize any dependencies between the variables to be predicted and the explanatory variables.

Modelization (R and Python languages)

Before you start this part, make sure you perform the same variable transformations in both languages.

Prediction of Heart's Disease Risk

We consider the problem of predicting the variable `Heart_Disease_Risk` from the other variables from a machine learning point of view, i.e. focusing on model performance. The aim is to determine the best performance we can expect, and which models achieve it. Here are a few questions to guide you.

1. Divide the dataset into a training sample and a test sample. Take a percentage of 20% for the test sample.Why is this step necessary when we're focusing on algorithm performance?
2. Compare the performance of a linear model (possibly generalized) with/without variable selection, with-/without penalization, SVR/SVM, optimal tree, random forest, boosting, and neural networks. Justify your choices (e.g. kernel for SVR/SVM), identify the hyperparameters for each model and adjust them carefully (using cross-validation).
3. Compare the different optimized models on your test sample. Which models perform best? How accurate are they? Which models should be retained if an interpretability constraint is added?
4. Interpretation and feedback on data analysis: are your results consistent with the exploratory data analysis, for example in terms of the importance of variables?

Prédiction of the variable `Cholesterol_LDL`

Repeat the previous steps to predict the variable `Cholesterol_LDL` from all the other variables (except the variable `Heart_Disease_Risk`).

Methods and assessment

You will complete the project in groups of 4 students. Assessment will be based on an oral presentation and two Jupyter notebooks (one in R and one in Python).

Assignment: As a deliverable, each group will place **at the latest** on Moodle :

- **on May 22 at 6:30pm**, a zip file containing the two compiled Jupyter notebooks (R and Python),
- **on May 22 at 6:30 pm**, the slides of the presentation **in pdf format**.

Oral presentation on May 27 or 19, 2026: 20 minutes for the presentation, followed by 5 to 10 minutes of questions. The presentation should include an introduction presenting the data and all the transformations you have performed, a brief description of the algorithms used (making it clear which hyperparameters you have optimized and how), an interpretation of the results, and a conclusion. Questions may relate to your code (so remember to open your notebooks and, if possible, compile them before the presentation).

Evaluation criteria: The evaluation will take into account the quality of the oral presentation (clarity, argumentation, interpretation of results, etc.), the coherence of the study, the quality of the presentation of the notebooks (don't forget to comment on your code), and the interpretation of the results (graphs, etc.). Generative AI can only be used to improve work that you have already done without using it (and not to do the work for you). This applies to all aspects of the work: coding, writing comments, and creating slides. In order to verify that this instruction is followed, some questions will focus on your code: any misunderstanding of your own code will be heavily penalized.