Machine Learning Project

Oral defense: May 28, 2025

Dataset

The data is taken from the KAGGLE competition website; it is the data set " Gym Members Exercise Dataset" available here: https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset.

This dataset provides a detailed look at the exercise routines, physical attributes and fitness measures of gym members. It contains 15 variables observed in 973 gym-goers:

- `Age` : age of the gym member.
- `Gender` : Gender of gym member (qualitative with two options: male or female).
- `Weight..kg.` : Member's weight in kilograms.
- `Height..m.` : Member's height in meters.
- `Max_BPM` : Maximum heart rate (beats per minute) during workout sessions.
- `Avg_BPM` : Average heart rate during workout sessions.
- `Resting_BPM` : Heart rate at rest before workout.
- `Session_Duration..hours.` : Duration of each workout session in hours.
- `Calories_Burned` : Total calories burned during each session.
- `Workout_Type` : Type of workout performed (qualitative with 4 modalities: cardio, weight training, yoga, HIIT).
- `Fat_Percentage` (Pourcentage de graisse) : Body fat percentage of the member.
- `Water_Intake..liters.` : Daily water intake during workouts.
- `Workout_Frequency..days.week.` : Number of training sessions per week (qualitative to 4 modalities: 2 to 5).
- `Experience_Level` : Level of experience (qualitative with 3 modalities: from 1 for beginner to 3 for expert).
- `BMI` : Body Mass Index, calculated from height and weight.

In this project, we first want to predict the variable `Calories_Burned` from all the other variables, and then predict the variable `Experience_Level` from all the other variables (including `Calories_Burned`).

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. Don't 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 calories burned

We now consider the problem of predicting the variable `Calories_Burned` 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), and carefully adjust the hyperparameters of each model (by 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?

Experience level prediction

Repeat the previous steps to predict the variable `Experience_Level` from all the other variables.

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 25 at 11:59pm**, a zip file containing the two compiled Jupyter notebooks (R and Python),
- **on May 27 at 6:30 pm**, the slides of the presentation **in pdf format**.

Oral presentation on May 28, 2025: 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 just 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.).