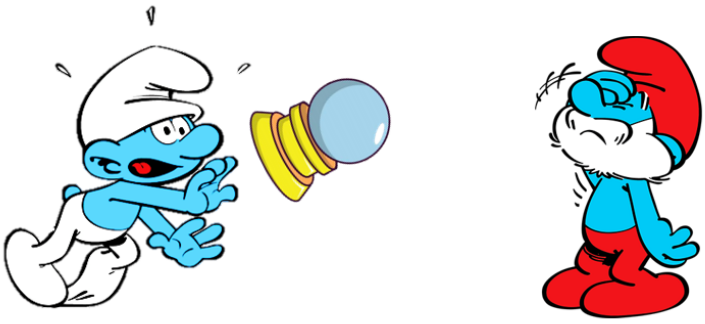
LELEC2870 - Machine Learning Project: Heart failure on the rise in the Smurf society

Academic year 2024-2025

Introduction

In the once tranquil world of Smurfs, death was an unfamiliar concept. However, significant changes in their society have cast a shadow over their formerly flawless paradise, introducing them to cardiovascular diseases they had never encountered. To aid in diagnosing and treating these newfound health issues, Papa Smurf used an extraordinary crystal ball. He manipulated this enigmatic device in collaboration with Doctor Smurf to predict the chances of heart failure for a fellow Smurf within the next 10 years of his life. While the crystal ball seemed to give accurate predictions, they did not fully understand the root causes behind these new heart problems. Regardless, calamity struck when Clumsy Smurf inadvertently shattered the magical oracle. In their quest to recover a life-saving prediction tool, they have turned to a new technology that they do not yet master: machine learning.



Fortunately for them, Doctor Smurf has meticulously gathered data on his peers. Indeed, for every Smurf he assessed, he documented information regarding his lifestyle, nutrition habits, blood test results and some other basic health indicators. Also, using the latest smurf medical imaging technology, he managed to obtain 28 by 28 pixel heart scans! Of course he also took careful note of Papa Smurf's crystal ball predictions.

Brainy Smurf immediately got to work and has recently proposed a first simple linear model for predicting heart failure risk. The results are not great, though still encouraging. He is currently working on more advanced models and documents his progress in technical reports. In the meanwhile, due to the urge of the situation, Papa smurf has turned to his most trusted human friend: you!

Your goal is to help Brainy Smurf by training a machine learning model yourself that produces predictions similar to those from the now shattered oracle. At the same time, you will try to formulate hypotheses on the causes of all these heart failures. This document provides guidelines to help you start this project, as well as a description of the data and details about the expected outcomes. Please take the time to read it carefully; a lot of little blue creatures are counting on you ! (no pressure).

Dataset

To complete your task, you will need to work with several data files. These are available on Moodle. In the `labeled_data` folder, the files `X_train.csv` and `X_test.csv` contain medical data stored as a table. Each line/row corresponds to a Smurf and each column to a measured attribute/variable/feature. A description of each variable is given bellow:

age	Age (can be well over 100 for Smurfs)
blood pressure	Systolic blood pressure (in mmHg)
calcium	Level of calcium in blood (in mmol/dL)
cholesterol	Level of LDL cholesterol ("bad cholesterol") in blood (in mg/dL)
hemoglobin	Level of hemoglobin in blood (in g/dL)
height	Height (in cm)
potassium	Level of potassium in blood (in mmol/L)
profession	Professional occupation (various)
sarsaparilla	Consumption of sarsaparilla leaves (very low - low - moderate - high - very high)
smurfberry liquor	Consumption of smurfberry liquor (very low - low - moderate - high - very high)
smurfin donuts	Consumption of smurfin donuts (very low - low - moderate - high - very high)
vitamin D	Level of vitamin B in blood (in ng/ML)
weight	Body mass of Smurf (in grams)

The risk of developping a heart failure within the next ten years is the target variable; it is stored in the `y_train.csv` and `y_test.csv` files. The indices match those of `X_train.csv` and `X_test.csv`. The last element of each line in `X_train.csv` and `X_test.csv` is the name of the image file that contains the corresponding heart scan. These images are stored in the folders `Img_train` and `Img_test`.

All these files form the labeled dataset on which you will train your models and estimate their generalization performance. You will find additional data in the folder `unlabeled_data`, for which you do not have the labels (but we do ☺). This second dataset will be used for the evaluation of your best model.

Instructions

You can work on the project individually or by groups of 2. The project is divided into 4 parts. In the first part, you will only use the tabular data and you will implement a linear model. In the second part, you are still using only the tabular data, but you will implement a nonlinear model. In the third part, you will integrate the images, and finally, in the last part, you will take a broader view and try to better understand this growing problem of heart failures in the Smurf society.

We ask you to code in python. Note that you do not have to implement everything from scratch, you may use any library you like (e.g. `scikit-learn` for traditional models and `pytorch` for deep neural architectures). In addition to the code, you will write a report to summarize your results. Here below, we detail what is required from you for each part.

Part 1 You will find on Moodle a brief technical report (`technical_report_1.pdf`) which describes how Brainy Smurf used a simple machine learning pipeline based on linear regression to

predict heart failure risk. Your goal is to propose a better pipeline, while still using linear regression. In particular, you are expected to do three things:

- Analyze and criticize the solution proposed by Brainy Smurf
- Based on your analysis, propose your own methodology and describe it precisely. Describe your data pre-processing, your feature selection, your model selection if any (model comparison, hyper-parameter tuning, validation procedure, etc.) and other details if they are relevant.
- present your results clearly AND discuss them. Compare with the results on Brainy Smurf. Do not hesitate to use various graphs and plots.

Part 2 Same as part 1, but now you will have to use a nonlinear model (of your choice). You will refer to a second technical report which will be made accessible after the smart-week (week 7 in the university calendar). You will analyze/criticize the work of Brainy Smurf, propose and describe your method, and discuss the results. Nonlinear models typically have much more hyper-parameters than linear ones (basic linear regression has none) so you will need to focus a bit more on model selection. Also know that some nonlinear models are sensitive to uninformative features, hence the need for a good feature selection as well. The pre-processing may be different than in part 1, but needs not to be. Note that you can start to work on this part as soon as you finish part 1, even if the second technical report is not available yet.

Part 3 You will now have to integrate the heart scans to your pipeline. You will have to extract features from the images using a deep neural network and combine them to the tabular features. You will then retrain (and eventually adapt) the nonlinear model that you proposed in part 2 on the combined dataset. When you discuss your results, make sure to compare performance with and without images features. As for part 1 and 2, you will receive a technical report from Brainy Smurf for reference. Note that we will see together how to implement a deep neural network for images during the practical session in week 9 ☺.

Part 4 This last and smaller part is a bit less guided. Your goal is to formulate some hypotheses on the causes of heart failures and identify the groups of Smurfs that are the most at risk. Try to provide a striking visualization. You are free to use any tool you desire (and of course, you may also rely on results from the previous parts).

Deliverable

By the end of Week 6 (Friday, October 25), you have will have to deliver:

- **Predictions**. Once your linear model is properly trained, you are asked to produce predictions on the data from `X.csv` (in the `unlabeled_data` folder) for which we have kept secret the corresponding targets. This prediction vector should be uploaded on Moodle in a csv file named `y_pred.csv` that contains one prediction per line and no header (no quotation marks around your numbers either). Check that your format is correct by opening it with a text editor and compare it to `y_test.csv`. We will use **RMSE** as the evaluation criterion.
- **Report**. A one page report describing what you did for part 1. No worries, you will be able to modify it for the final deadline.

By the end of week 12 (Friday, December 6), you will have to deliver:

- **Predictions.** Choose your best model among part 1, 2 and 3 and use it to produce predictions on the data from `X.csv` (in the `unlabeled_data` folder).This prediction vector should be uploaded on Moodle in a csv file named `y_pred.csv`. We will use **RMSE** as the evaluation criterion. **If you transform the target at some point, do not forget to apply**

the inverse transform before estimating the generalization performance and/or before making your predictions.

- **Report**. You will produce a report documenting your technical choices and experimental results. We do not need a course on the methods you use. We are more interested in what you did and why. Be concise and go straight to the point! Follow the structure: Introduction, Part 1, Part 2, Part 3, Part 4, Conclusion. For part 1, 2 and 3, organize your text as follow: comments on technical report, description of proposed method, results and discussion. A strict **maximum of 6 pages** (one column, fontsize 11 or larger) will be observed.
- **Code**. Also on Moodle, you should submit a compressed folder containing all your python scripts (notebooks, utils.py, etc). Theses should be runnable and contain at least what you discussed in your report. There is no size limit, but these files should be structured, commented, and clear enough so that information can be easily found without deciphering everything! If you used any packages that were not used during the practical sessions, or a different version of those, don't forget to mention it in the beginning of your file.

Schedule

Below you will find the schedule for the project.

- As soon as possible: Register your group (maximum two people) on Moodle
- Friday 25/10 at 23h55: Intermediate deadline where you submit your work for part 1 as 2 seperate files (a csv file for you first predictions and a pdf for a "pre-report" on part 1)
- Thursday 7/11 at 8h30: Q/A session #1
- Thursday 21/11 at 8h30: Q/A session #2
- Friday 6/12 at 23h55: final deadline where you submit your work as 3 separate files (a csv file for your predictions, a pdf for your report, and a compressed folder for all your scripts)

Do not wait until the last minute to start, and take advantage of the Q/A sessions for asking your questions and receiving feedback. We also encourage you to discuss about the project with other groups. We do not want to see plagiarism, but we certainly value exchange of ideas and experiences. Remember to cite all your sources. Use ChatGPT wisely, don't let it fool you.

Evaluation

The project will account for half of the points in this course (10/20). Here is a rough idea of the weighting of each section: Part 1 (2/10), Part 2 (3/10), Part 3 (3/10) and Part 4 (1/10). Performance of your best model on the unlabeled data will account for the last point (1/10). This weighting may be subject to small variations. Finally you will be able to earn up to one bonus point if your linear regression submitted on the intermediate deadline performs well and granted that you respect the required file format. Also, we really insist on the quality of the report; be concise, clear, justify your choices and interpret your results! Embrace this mantra: a good project with a bad report is a bad project!