Introduction to Machine Learning

End-Term Exam

The goal of this project is to use the knowledge you've gained in the course and apply it to a dataset of your choice. It's designed to be flexible, allowing you to enhance your technical skills and problem-solving abilities.

I've observed that many students tend to focus on memorization instead of truly understanding the material. However, university is about developing critical thinking and enable individuals to become independent citizens, who can contribute meaningfully to their society. I hope this project will help you grow in that direction.

This document was built as a guideline and to give you some ideas, but it should not be strictly followed as a list of tasks.

General information :

- You should work in group of two or three students.
- The defense will be held during the last week (21st - 26th of April).
- The defense planning will be uploaded in teams.
- The defense should last 15 to 20 minutes maximum.
- If you didn't defend before the 27th April, your grade will automatically be set to zero (its the deadline for the second assessment).
- Deadline for submitting the project in teams : 21st of April 23h59.

Before starting : Chose a dataset and clearly identify the purpose of your study. What do you want to predict, and why ?
Here are three website you can find different datasets (please feel free to use other source) :

- UC Irvine
- Kaggle
- data gouv fr (its in french but they have interesting datasets and you might find a way to translate the website)

Some restrictions on the data :

- Your dataset should be tabular
- Don't pick visual dataset
- Don't pick time series dataset
- Don't pick a dataset we used during the semester
- If you have some doubts about your dataset, please contact me to validate it.

Data-analysis and processing : Proceed to some data-analysis. I want you to pay more attention to this part than you did so far. Your analysis should :

- provide insights into the distribution of variables, correlations, and relationships between features. If possible, translate the different correlations with words. Also wonder and try to see (its not always possible) the nature of these correlations (cf mid term exam).
- help you to find a good method to transform/scale your data (if necessary).
- detects data quality issues, such as missing values, duplicates, and outliers, which can significantly impact model performance if not addressed.
- give you some feature engineering idea : could you create new features from existing ones ?

Model evaluation and interpretation : Experiment different machine learning models and evaluate them. Document these iterations in the notebook to explain your reasoning and results.

- ML projects frequently require experimenting with different models, feature sets, and hyper-parameters to find the optimal solution.
- The process of trial and error is essential because the dataset and problem specifics often dictate which combinations work best. Its normal to have a non-efficient or insightful model at first. What truly matters is not the initial performance but understanding why the model did not achieve the desired results.
- Analyze the shortcomings, whether they are from insufficient data, poor feature selection, inappropriate hyper-parameters, or the choice of algorithm itself. Use these insights to iteratively refine your approach.
- Use appropriate metrics depending on the type of problem.
- Provide an interpretation of the result, i.e. do not limit yourself to just reporting numbers.
- Try to explain what these metrics imply in the context of your project.
- Analyze feature importance (if relevant) to understand which features contributed the most to the model's predictions.
- Finally, select one model with a set of feature that seem to solve your problem. Highlight its limitations, and/or point out how it could be improved.

Deliverable : a jupyter notebook (and your defense).

- I'll pay more attention to the quality of your code and the structure of your notebook than before.
- Your code should be modular and well-structured, with clear function or module definitions.
- Your notebook should be readable as a book, telling the story of your experiments.

Your mark will be given according to :

- your ability to solve a problem
- your ability to question your own work
- the understanding of the ML tools you used
- the quality of your code
- the quality of your explanations
- The top three teams will be awarded 20, 19, and 18 points, respectively. The remaining teams will be graded with a maximum of 17 points.