Data & AI 5/Artificial Intelligence: Machine Learning lifecycle Project

Alexander Michielsen, Jan Van Sas

Overview

In this project, you will work in groups of three. Your goal is correctly implement and evaluate a Machine Learning lifecycle
Imporant: This project is not just about correctly implementing a complete lifecycle. You will be evaluated on your ability to justify your choices, analyze results, and draw conclusions. Make sure to document everything in a clear and structured manner.

Objectives

- Perform comprehensive exploratory data analysis (EDA) on a real-world dataset.
- Implement data preprocesssing and feature engineering techniques into a reproducible pipeline.
- Select and tune suitable machine learning models for the task. Compare multiple models.
- Apply sound validation techniques to assess model performance using appropriate metrics.
- Analyze results critically and reflect on the model's strengths and weaknesses.
- Justify all choices made throughout the project.
- Document the process and findings in a comprehensive notebook.
- Optionally, deploy the model using a simple endpoint.

Project Requirements

Note: These are the minimum requirements for a passing grade. The more extensive and thorough your work, the better your grade will be. Please also take into account that part of your final grade will be based on the oral examination of the project.

1. Data Understanding and Exploration (15%)

- Choose a real-world dataset and define a clear problem statement (e.g., classification, regression). The dataset should have at least 1000 instances and 10 features. **If the dataset is too clean, introduce some outliers and missing values yourself.**
- Perform extensive exploratory data analysis (EDA).
- Include visualizations and statistical summaries to understand the data distribution, relationships, identify patterns, detect anomalies,...
- Discuss the result of your EDA in detail.

2. Implementation and Data Pipeline (20%)

- Perform suitable data preprocessing and cleaning steps (e.g., handling missing values, class imbalances, ...).
- Perform feature engineering including feature transformation, selection, creation, and extraction if applicable.
- Implement a reproducible data pipeline.
- Pay special attention to data leakage and ensure that your pipeline is robust.

3. Model Selection and Justification (15%)

- Choose appropriate machine learning models for the task (e.g., classification, regression). Consider multiple models.
- Apply hyperparameter tuning techniques.
- Compare multiple models and make a **justified** choice for the final model (the justification is essential and should be completely written down in the notebook!)

4. Evaluation and Analysis (20%)

- Choose appropriate metrics to evaluate model performance.
- Apply sound validation techniques (e.g., cross-validation, train-test split).
- Visualize the results using appropriate plots.
- Interpret the results and discuss the various models' strengths and weaknesses for your problem.

5. Deployment (Optional) (5% bonus)

Note: This bonus is only available if criteria 2 to 4 are met at the 'meets' level.

- Expose the model via a simple API endpoint (e.g., using Flask or FastAPI).
- Must be demonstrable during the oral examination.

6. Oral Examination (30%)

During the exam period, you will have to defend your project. Questions will be asked about all aspects of the project, including choices made, results obtained, and potential improvements/alternatives. Each team member should be able to answer questions about the entire project especially the parts they worked on.

Deliverables

Interactive Python Notebooks

- Combine code, explanations, and results in well-documented Jupyter Notebooks.
- **All cells should be executed with the output visible and free of errors**.
- Use multiple notebooks if necessary (e.g., one for EDA, one for modeling, ...).
- Ensure that the notebooks are well-structured and easy to follow.

Additional Notes

- Each team member needs to have a complete understanding of all the work that has been delivered.
- You are free to use any tool and support that you find useful (e.g., AI tools, online resources, ...). In the case of AI tools, **keep a record** of what you used and how it helped you and include it in a separate text file in your submission. **This can be questioned during the oral examination.**
- You can go as far as you want with this project, however, keep your time management in check. Prioritize if necessary.

Submission Guidelines

- **Deadline**: *27 October 2025, 23:59*.
- **Submission Method**: Upload your Jupyter Notebooks and all relevant files to Canvas in a Zip format. Ensure that the notebooks have been run without errors and that output is visible.
- A late submission will lead to 0/20 for the project.
- **Format**:
  - Notebooks should be named clearly (e.g., `DataExploration.ipynb`).
  - All code cells should be executed in order, and outputs should be visible.

Good Luck!