First Assignment

Comparative Study of Classification Algorithms

Course: Machine Learning
Instructor: *Federico Fusco*

Due Date: *21/11/2025*

Objective

The goal of this project is to develop an understanding of foundational supervised learning algorithms by implementing, analyzing, and comparing multiple models on a real-world dataset.

Assignment Description

Part 1: Data Selection

Choose a dataset from the UCI Machine Learning Repository or Kaggle, and decide whether you want to solve a classification or regression task. The dataset chosen should contain at least 1,000 data points.

Part 2: Data Preprocessing

Perform appropriate preprocessing steps:

- Handle missing or noisy data.
- Normalize or standardize features.
- Split data into training, validation, and test sets.
- **Optional**: apply dimensionality reduction (e.g., PCA) or feature selection.

Part 3: Model Implementation and Training

Train the following models (and set the hyper-parameters via cross-validation):

1. **Naïve Bayes** (Gaussian or Multinomial) - if you have chosen a classification task
2. **Linear Regression** (linear and/or feature-based) - if you have chosen a regression task
3. **Logistic Regression** - if you have chosen a binary classification task
4. **Softmax Regression** - if you have chosen a multi-class classification task
5. **Decision Tree** - if you have chosen a classification task
6. **Random Forest** - if you have chosen a classification task
7. **SVM** (linear and kernel-based) - if you have chosen a binary classification task

Part 4: Evaluation

Evaluate each model using the following metrics, when applicable:

- Accuracy, Precision, Recall, F1 Score

- Confusion Matrix
- ROC and AUC (binary and multiclass)
- Training vs. validation performance
- **Optional**: Computational cost and training time

Part 5: Comparative Analysis

Include a 1–3 page analysis discussing:

- Which models performed best and why.
- How model assumptions influence performance.
- Observations on overfitting trade-off.
- Visualizations: learning curves, decision boundaries, feature importance.

Deliverables

1. **Technical Report (PDF) prepared in LaTex**:
   - Introduction and motivation
   - Dataset description
   - Methodology and models
   - Results and analysis
   - Conclusion
2. **Google Colab Notebook** with clean, commented code, to reproduce all the experiments presented in the technical report.

Suggested Tools

There is no need to implement everything from scratch, you can use the following tools:

- Python: `NumPy`, `Pandas`, `Matplotlib`, `scikit-learn`, . . .
- Jupyter Notebook or Google Colab
- LLMs: they are great to guide you through the project, **but** (there is a big **but**) you should not use them black-box without understanding what they are doing and why. Most importantly, they will not be allowed at your written exam, so make sure to use this assignment as an opportunity to learn by doing.

Evaluation Criteria

This is an open ended project. There is no wrong way to set it up, as long as you perform the tasks required and submit it on time. So feel free to be imaginative and creative, play with the models that we have studied in class and test them on real life datasets!

Submission Details

Hard Deadline: 23:59 (CET), 21st of November 2025.
Submission Form: `https://forms.gle/gy1piVfCJMwbjsJaA`
Team Formation: You can decide to submit the project as a team of at most *two* participants