Intro to AI – Final Project

The final project of the **Introduction to AI** course will consist of a case study on how AI can be used to process a particular dataset.

In many scenarios, companies have access to datasets (for instance, sales, logistics records, IT bug reports and support tickets, etc), but without much knowledge about how to extract relevant information from this data. Hence, automated extraction of information and patterns from data may appear as a very practical use case of AI. We propose to tackle such a scenario as the final evaluation of the class.

Project Structure

In the context of the class, you will have **3 assignments**:

1. Choose an Appropriate Dataset

The dataset must meet the following criteria:

- **Adequate Size**: It should have a sufficient number of samples to enable meaningful analysis and model training without requiring too much computational power. For beginners, a dataset with **1,000–10,000 samples** samples is usually ideal.
- **Completeness**: The dataset should be as complete as possible, with minimal missing values.
- **Easy-to-Use Format**: Clear labeling of data points, well-defined columns, and readable metadata are essential for easy understanding and processing.
- **Open License**: The dataset should be freely available for academic use.

2. Define One or Two Tasks Based on the Dataset

- Tasks must be finalized and validated by the teachers **before 22 November**.
- Validation can be done via **Discord** or **in-person** during class.

3. Evaluate How AI Can Help Solve These Tasks

Use the resources seen in class to evaluate your model.

Suggested Platforms for Open-Access Datasets

If you have difficulty finding a dataset or defining tasks, the following resources and platforms may help:

- **Kaggle**: https://www.kaggle.com/datasets
  A large repository of datasets across various domains like healthcare, finance, NLP, and more. Beginner-friendly.
  Example: Titanic Survival Prediction Dataset (binary classification with tabular data).

- **Hugging Face Datasets**: https://huggingface.co/datasets
  Particularly good for text and NLP datasets, also includes datasets for computer vision, reinforcement learning, etc.
  Example: IMDB Reviews (sentiment analysis with labeled text).

- **OpenNeuro**: https://openneuro.org/
  A great source for neuroimaging datasets, especially for students interested in exploring AI in neuroscience.

- **PhysioNet**: https://physionet.org/
  Primarily focused on physiological and clinical data, great for students interested in healthcare and bioinformatics projects.

Tools and Guidance

- You may use any AI tool, including those introduced in class (e.g., UMAP/t-SNE/PCA for visualization, neural networks, foundation models, etc.).
- You may use any AI assistant to help you (e.g. ChatGPT, Copilot, ...).
- The evaluation emphasizes **comprehension**, not the objective performance of your solution.
  - **Good Practices**: Develop a strategy, test hypotheses, and validate them. Experimenting extensively is encouraged.
  - **Avoid**: Implementations without clear motivation.

Final Project Evaluation Criteria

1. Relevance of Overall Strategy

- **"-"**: Experiments lack a clear or consistent approach.
- **"="**: Experiments follow a logical and consistent approach aligned with the given objectives.
- **"+"**: Strategy for exploring various AI techniques was particularly effective.

2. Description of Experiments and Results

- **"-"**: The presentation does not clarify which experiments were conducted.
- **"="**: The presentation is clear and detailed enough to allow for reproduction of the results (clear and comprehensive description of training hyperparameters and architecture).
- **"+"**: The exploration of AI approaches and hyperparameters is particularly precise and adapted to the chosen dataset.

3. Creativity

- **"-"**: The implementations are limited to the approaches presented in the labs.
- **"="**: The proposed implementations go beyond the labs, with original techniques, new foundation models, or custom models trained from scratch.
- **"+"**: The proposed implementations are highly original and creative (involving recent literature research, adoption of new tools and techniques, etc.)