

Project Assignment

Machine Learning and AI in Python

Overview:

In this project, you will explore the power of Machine Learning (ML) and Artificial Intelligence (AI) to study, predict, or discover patterns in a real-world problem of your choice. The goal is to develop, fit, and evaluate at least four different learning models to address a specific issue or trend that you are passionate about. This could be something related to your country, region, or any field in which you have significant expertise (e.g., sports, nature, finance, healthcare, etc.).

You are expected to apply and experiment with a variety of techniques, including **but not limited to**:

- K-Nearest Neighbors (KNN)
- Logistic Regression
- Linear Regression
- Linear Discriminant Analysis (LDA)
- Clustering
- Naive Bayes
- Decision Trees
- Neural Networks
- Hierarchical Clustering
- Principal Component Analysis (PCA)

You are also encouraged to explore more advanced ML techniques that we have recently covered in class, or that will be introduced soon. You may choose from the methods listed below or apply a different technique not included in this list, if it fits your project.

- NLP: embeddings
- LM: word predictions
- Text classification
- Computer Vision
- XIA

At least **four different techniques** must be used in your project.

Objectives:

1. **Problem Identification:** Identify a real-world problem, trend, or pattern you would like to study or predict. Your problem could relate to various sectors like sports

analytics, finance, environmental studies, healthcare, etc. It should ideally connect to your region or field of expertise.

2. **Model Development:** Develop at least **four machine learning models**. You are encouraged to explore different types of models (classification, regression, clustering, etc.) and justify why they are suitable for your chosen problem.
3. **Compare the results** of the different models you have applied. Where applicable, particularly if you are using different models to classify, predict, or explain the same outcome, you should evaluate and compare the performance of each model. Discuss the strengths and weaknesses of each model in terms of performance, interpretability, and suitability for the problem you are addressing.
4. **Model Training & Testing:**
 - **Fit** your models to the data.
 - **Test** the models on unseen data.
 - **Generalize** your models to ensure they perform well on data beyond the training set.
 - **Validate** the models using appropriate metrics (e.g., accuracy, precision, recall, F1-score, etc.).
 - Apply **regularization** techniques where appropriate to avoid overfitting.
5. **Data Preprocessing & Exploration:** You should handle data preprocessing, cleaning, and feature selection techniques as needed to ensure that your models are well-optimized and accurate.
6. **Visualization:** Provide clear visualizations to explain your data, model performance, and findings.
7. **Presentation:** You will be required to deliver a **5-minute presentation** in front of the class, explaining your problem, models, results, and insights. This will help you practice communicating your technical work to an audience.

Deliverables:

1. **Python Code:**
 - Submit your complete, well-documented Python code that includes all necessary comments and explanations of your methods and logic.
 - The code should be clean, readable, and well-organized.
2. **GitHub Repository** (Preferred but Optional):
 - It is **preferred** that you compile your project and host it on GitHub, including a readme file that explains the project overview, objectives, models used, and how to run the code.
 - If you choose to use GitHub, make sure the repository is public or share the link in Moodle as part of your submission.
 - However, this is **optional**, and you may submit your files directly to Moodle if preferred.
3. **Presentation:** A **5-minute presentation** where you explain:
 - The problem you are addressing.
 - The models you selected and why.
 - Key findings and results.
 - Challenges faced and solutions implemented.

Grading Criteria:

- Relevance and clarity of the problem being addressed.
- Quality and diversity of the models applied (minimum of 4 different techniques).
- Thoroughness of the model evaluation process (fitting, testing, validating, and regularizing).
- Quality of Python code and documentation.
- Effectiveness of the presentation and communication of findings.
- Use of GitHub (if applicable) to organize and share the project.

Suggestions:

- Choose a problem that is relevant to your field of interest, and try to apply your expertise in the domain to select the most appropriate techniques.
- Be creative! Think of trends or patterns that could have a meaningful impact or provide insights in your area of study.
- Regularly commit to GitHub as you progress in your project to ensure version control and track your work effectively.

This is an opportunity for you to apply what you've learned in a meaningful way, explore your own ideas, and demonstrate your skills in machine learning and AI.

Good luck!