# Projects for Artificial Intelligence and Machine Learning

This document describes the projects for the **Artificial Intelligence and Machine Learning** course. They are valid only for the **academic year 24/25**. Projects are mandatory, whether you take the midterm/final or the oral exam. The project will contribute to **30% of your total grade** in the course. A **further 10% will be assigned** on the day of the final/exam following theoretical and/or technical questions on the project.

## Instructions

### Choosing your project

You can **work in groups of at most 3 people**. The team's "captain" must send an email to gitaliano@luiss.it, fangeletti@luiss.it and ampanti@luiss.it. Such email must include, in CC, all the other components of the team and its subject must be **[AIML PROJECT 24/25]**. The email should contain:

> **First Project Preference**
>
> Name Surname student id of member 1 ("captain")
>
> Name Surname student id of member 2
>
> Name Surname student id of member 3
>
> **Second Project Preference**

You must send the email by **October 20th, 2024, at 23:59**. If you do not send an email by that date, you will be assigned to a project and to a team by the instructor.

### When to submit your project

You must submit your project **at least 7 days before the exam date** when you want to take the exam or register your grade. The first two exam dates are 10th December 2024, and 15th January 2025, thus you must submit either by **3rd December 2024, at 23:59**, or by **8th January 2025, at 23:59**.

### What to submit for your project

Each group must submit via mail to gitaliano@luiss.it, fangeletti@luiss.it and ampanti@luiss.it the URL of a GitHub repository. The repository's name must end with the student id of the "captain".

The repository must contain:

1. **A "README.md" file** with the following information:

- **Title and Team members**
- **[Section 1] Introduction** – Briefly describe your project
- **[Section 2] Methods** – Describe your proposed ideas (e.g., features, algorithm(s), training overview, design choices, etc.) and your environment so that:
    - A reader can understand why you made your design decisions and the reasons behind any other choice related to the project
    - A reader should be able to recreate your environment (e.g., conda list, conda envexport, etc.)
    - It may help to include a figure illustrating your ideas, e.g., a flowchart illustrating the steps in your machine learning system(s)
- **[Section 3] Experimental Design** – Describe any experiments you conducted to demonstrate/validate the target contribution(s) of your project; indicate the following for each experiment:
    - The main purpose: 1-2 sentence high-level explanation
    - Baseline(s): describe the method(s) that you used to compare your work to
    - Evaluation Metrics(s): which ones did you use and why?
- **[Section 4] Results** – Describe the following:
    - Main finding(s): report your final results and what you might conclude from your work
    - Include at least one placeholder figure and/or table for communicating your findings
    - All the figures containing results should be generated from the code.
- **[Section 5] Conclusions** – List some concluding remarks. In particular:
    - Summarize in one paragraph the take-away point from your work.
    - Include one paragraph to explain what questions may not be fully answered by your work as well as natural next steps for this direction of future work

2. **One single notebook called "main.ipynb"** with ALL the code used for the project. The notebook must have the following characteristics:
    - Text and code cells must alternate from start to finish. The text cell above must describe the contents of the code below and its output so that a reader can easily follow up on your implementation. In particular:
        - You must explain what you will do and why you chose to do so.
        - You must explain the outputs of the cell (if any) with particular attention to describing figures such that a reader already knows what he is going to see
3. An additional folder named "images" contains the figures displayed in the "README.md".

## Academic Integrity

You must write the code by yourself. The abuse of copy-paste will be taken into account during the evaluation. Any code that, for some (nonsensical) reason, is not written by yourself must be referenced (with a link to the original code). Copying the projects from other teams is also strictly forbidden. Your code will be validated by anti-plagiarism software. In the unlikely event of two projects being very similar, we will follow the Netflix Prize rules: only the first project published on GitHub will get the grade, and the other will get nothing.

# Projects proposals

## Datasets

All the datasets can be found on the Luiss Learn platform **inside the folder Datasets** in the section Project.

## Assignment

The assignment is valid for all the projects and acts as a guideline throughout the completion of the project. Depending on the task you want to accomplish, some points are applicable, some are not. It is your responsibility to understand and apply the correct algorithms/processes.

- Perform an Explanatory data analysis (EDA) with visualization.
- Define whether this is a regression, classification or clustering problem, explain why and choose your model design accordingly. Test at least 3 different models. First, create a validation set from the training set to analyze the behaviour with the default hyperparameters. Then use cross-validation to find the best set of hyperparameters. You must describe every hyperparameter tuned (the more, the better).
- Preprocess the dataset (remove outliers, impute missing values, encode categorical features, not necessarily in this order).
- Generate a training and test set, if needed. The test set should be used only at the end.
- Select the best model using the right metric.
- Compute the performances of the test set if applicable.
- Make a comparison of the different models
- Write the report, a.k.a. “README.md”

## Aerogel Bonding

In the development of lightweight, high-performance aerospace structures, evaluating the bonding of aerogels with carbon fiber composites is essential to ensure strong adhesion, minimize thermal expansion mismatch, and maintain the insulation properties under extreme temperatures. For high-temperature industrial machinery, bonding aerogel to steel surfaces is crucial to provide thermal protection while ensuring mechanical stability and resistance to thermal cycling in harsh environments. In the fabrication of solid-state batteries, evaluating the bonding of aerogel with lithium metal electrodes can improve thermal management, reduce weight, and enhance the overall energy efficiency and safety of the storage system. As a chief data scientist in Advance Material Corporate Ltd you need to understand whether the **aerogel bonding** is good enough for the commercial application. You must use the dataset “aerogel_bonding.csv”

### Dataset Features

- **HolidaysTaken**: Total number of holidays taken by the worker over a period of time.
- **PercentageOfCompletedTasks**: Proportion of assigned tasks successfully completed by the worker.
- **CurrentJobDuration**: Number of months the worker has been in their current job role.
- **RecentHolidaysTaken**: Number of holidays taken by the worker recently, possibly affecting work continuity.

- **RequestedProcessAmount**: Quantity of material requested for processing in the aerogel bonding process.
- **JobStatus**: Current employment status of the worker, e.g., employed, contract, or retired.
- **BondingRiskRating**: A numerical rating that reflects the risk associated with the aerogel bonding process, considering factors like material compatibility and process conditions.
- **TotalMaterialProcessed**: The cumulative amount of material processed by the worker or team in bonding operations.
- **ByproductRation**: The proportion of byproducts generated relative to the total material processed.
- **working_skills**: The skill level of the worker in performing tasks related to aerogel bonding.
- **CivilStatus**: The worker's civil status, such as married or single, which might be used for workforce analysis.
- **dependability**: A measure of how reliable the worker is in completing tasks on time and with quality.
- **MistakesLastYear**: The number of mistakes made by the worker in the previous year related to material bonding processes.
- **HighestEducationAttained**: The highest level of education completed by the worker, which may influence performance.
- **BondingSuccessful**: A binary indicator of whether the aerogel bonding processes were successful and ready for commercial usage.
- **ChurnRisk**: Risk of the worker leaving the company or being transferred to a different department or job.
- **ProcessedKilograms**: The total weight of material processed by the worker, related to aerogel bonding.
- **SkillRating**: An overall rating of the worker's skill, considering various performance metrics.
- **ProcessingTimestamp**: The date and time at which a specific bonding process or task was carried out.
- **WorkExperience**: The total work experience of the employee in months or years, particularly in material bonding processes.
- **HistoricalBehavior**: A summary of the worker’s past performance in similar tasks, possibly including metrics on punctuality, quality, and adherence to processes.
- **TotalMaterialToProcess**: The remaining amount of material yet to be processed in the current bonding project.
- **WorkHistoryDuration**: The length of the worker's overall career in months or years.
- **ApplicantAge**: The age of the worker, which may correlate with experience and performance.
- **PriorExecutionDefaults**: The number of times the worker defaulted or failed in prior bonding operations.
- **DifferentTasksCompleted**: The variety of tasks completed by the worker, indicating versatility in different bonding or material handling processes.
- **TotalChurnRisk**: Overall risk that the worker might leave their current position or not complete the project.
- **OtherCompaniesMaterialProcessed**: The quantity of material processed for other companies, indicating external work experience.
- **BondingPeriod**: Duration of time taken for each bonding process, from start to completion.
- **trustability**: A score indicating the reliability of the worker based on past performance data.
- **MonthlyExecutions**: The number of bonding operations or tasks completed by the worker monthly.

## Guilds

In the kingdom of Marendor, scholars strive to advance knowledge, and their goal is to enter the **Master Guild**, which reflects the influence of their discoveries. The council of Marendor values each scholar's contribution and carefully assesses their research to predict the potential impact their work will have on the kingdom's future progress. By analyzing various factors from the scholars' reports, the Academy aims to anticipate the guild assignment for each scholar, helping guide the kingdom's priorities in fostering groundbreaking discoveries. Use the dataset "guilds.csv".

## Dataset Features

- **Fae_Dust_Reserve**: The subject's reserve of mystical dust indicating magical potential.
- **Physical_Stamina**: The subject's overall endurance and physical health.
- **Mystical_Index**: A numeric representation of the subject's mystical power and well-being.
- **Healer_consultation_Presence**: Indicates if the subject recently consulted a healer.
- **Elixir_veggies_consumption_Presence**: Shows if the subject consumed enchanted vegetables.
- **Mystic_Energy_Level**: Represents the level of mystical energy possessed by the subject.
- **Bolt_of_doom_Presence**: Indicates if the subject experienced a thunderstrike.
- **Age_of_Wisdom**: The subject's age, reflecting life experience.
- **High_willingness_Presence**: Shows the subject's inclination towards embracing higher magical activities.
- **Defense_spell_difficulty_Presence**: Indicates the subject's difficulty in casting defense spells.
- **Doc_availability_challenge_Presence**: Shows barriers preventing access to healers.
- **Mental_Wizardry**: Represents the subject's mental health and wizardry capacity.
- **Potion_Power_Level**: The power or effectiveness of potions used by the subject.
- **Dexterity_check_Presence**: Indicates the high dexterity.
- **Gold_Pouches_Per_Year**: The subject's annual income represented as gold pouches.
- **Wizardry_Skill**: The subject's level of proficiency in magical skills.
- **Spell_Mastering_Days**: The number of days the subject dedicated to mastering spells.
- **Level_of_Academic_Wisdom**: The highest level of knowledge achieved by the subject.
- **General_Health_Condition**: An overall assessment of the subject's health status.
- **Fruits_of_eden_consumption_Presence**: Indicates if the subject consumes fruits from Eden.
- **Knight_physical_training_Presence**: Shows if the subject did knight-like physical training.
- **Royal_family_pressure_Presence**: Indicates whether the subject is subject to pressure from the royal family.
- **Dragon_Sight_Sharpness**: A measure of the subject's visual acuity.
- **Guild_Membership**: The guild or magical faction the subject will belong.
- **Enchanted_Coin_Count**: The number of enchanted coins representing wealth.
- **Celestial_Alignment**: Represents alignment with celestial forces.
- **Knightly_Valor**: The bravery and valor displayed by the subject.
- **Heavy_elixir_consumption_Presence**: Indicates if the subject consumes heavy magical elixirs.
- **Stigmata_of_the_cursed_Presence**: Shows if the subject experienced a crisis from magical and dark powers.
- **Dragon_status_Presence**: Denotes whether the subject has the sign of the dragon.
- **Rune_Power**: Represents the power derived from magical runes by the subject.

## Euphoria

In the virtual world of **Euphoria**, a vast archipelago of islands stretches across a digital sea. Each island is unique, offering different experiences and environments to its inhabitants. The goal is to explore and understand these islands by segmenting them based on their "happiness levels" allowing similar travelers to find their perfect paradise and remain close together. Use the dataset “euphoria.csv”.

### Dataset Features

- **referral_friends**: Number of friends referred to the island.
- **water_sources**: Availability and quantity of fresh water on the island.
- **shelters**: Number of shelters or dwellings available on the island.
- **fauna_friendly**: Indicates if the island allows pets or has wildlife presence.
- **island_size**: Total physical area of the island.
- **creation_time**: The time when the island was formed or discovered.
- **region**: The geographical area or zone where the island is located.
- **happiness_metric**: The unit or scale used to measure happiness on the island.
- **features**: List of amenities or unique characteristics available on the island.
- **happiness_index**: The overall happiness level of the island's inhabitants.
- **loyalty_score**: Measures the loyalty or retention of island visitors or residents.
- **total_refunds_requested**: Number of refunds requested by visitors.
- **trade_goods**: Items or resources available for trade on the island.
- **x_coordinate**: The horizontal position of the island on the map.
- **avg_time_in_euphoria**: Average time people spend on the island.
- **y_coordinate**: The vertical position of the island on the map.
- **island_id**: Unique identifier for each island.
- **entry_fee**: Cost required to enter or visit the island.
- **nearest_city**: The closest city to the island.

## Aeropolis

In the futuristic city of **Aeropolis**, autonomous delivery drones are essential to ensure fast and efficient delivery of goods across the sprawling metropolis. Each drone's performance is evaluated based on how much cargo it can deliver per flight. However, many factors influence its performance, from weather conditions to the type of terrain it navigates. To optimize drone performance, data scientists are tasked with predicting the cargo capacity per flight based on various environmental and operational factors. Use the “aeropolis.csv” dataset.

### Dataset Features

- **Cargo_Capacity_kg**: The amount of cargo the drone can carry in kilograms per flight.
- **Air_Temperature_Celsius**: The air temperature during the drone's flight, measured in Celsius.
- **Weather_Status**: The current weather conditions during the drone's operation.
- **Package_Type**: The type of package or cargo the drone is delivering.
- **Vertical_Landing**: Indicates if the drone uses vertical landing capability.
- **Equipment_Cost**: The cost of the drone equipment.
- **Market_Region**: The geographical market where the drone operates.
- **Flight_Duration_Minutes**: The total duration of the flight, measured in minutes.
- **Terrain_Type**: The type of terrain over which the drone flies (e.g., urban, rural).
- **Water_Usage_liters**: The amount of water used by the drone during the flight, in liters.

- **Flight_Hours**: The cumulative flight hours logged by the drone.
- **Delivery_Time_Minutes**: The total time taken for the drone to complete the delivery.
- **Cleaning_Liquid_Usage_liters**: The amount of cleaning liquid used by the drone for maintenance.
- **Climate_Zone**: The climate zone in which the drone is operating.
- **Quantum_Battery**: Indicates if the drone uses a quantum-powered battery.
- **Flight_Zone**: The operational zone where the drone conducts its flights.
- **Autopilot_Quality_Index**: A rating of the autopilot system's performance.
- **Vertical_Max_Speed**: The maximum vertical speed the drone can achieve.
- **Wind_Speed_kmph**: The wind speed during the drone's flight, measured in kilometers per hour.
- **Route_Optimization_Per_Second**: The number of route optimizations performed per second by the drone's navigation system.

## Alien Galaxy

In the future, where interstellar exploration is common, different alien species have begun colonizing planets across the galaxy. However, each species prefers different planetary environments, and their success on each planet is determined by various factors. The goal is to **group together** these colonized planets to understand which types of planets attract certain alien species, optimize colonization strategies, and promote peaceful coexistence among different species. Use the dataset “alien_galaxy.csv”.

### Dataset Features

**Peace_Treaty_Accords**: Number of formal peace agreements made between alien species.
**Technological_Advancements**: The level of technological innovations achieved by the planet's inhabitants.
**Ammonia_Concentration**: The concentration of ammonia in the planet’s atmosphere or environment.
**Precious_Metal_Trade_Tons**: The total trade volume of precious metals conducted by the planet, measured in tons.
**Food_Production_Tons**: The amount of food produced on the planet, measured in tons.
**Trade_Agreements_Signed**: The total number of interplanetary trade agreements signed by the planet's inhabitants.
**Last_Contact_Days**: The number of days since the last interstellar communication or visit.
**Discovery_Date**: The date when the planet was first discovered by alien explorers.
**Mineral_Extraction_Tons**: The quantity of minerals extracted from the planet, measured in tons.
**Galactic_Visits**: The number of visits made by interstellar travelers to the planet.
**Sulfur_Concentration**: The level of sulfur present in the planet’s atmosphere or soil.
**Exploration_Missions**: The number of exploration missions launched to study or survey the planet.
**Biological_Research_Units**: The amount of biological research conducted on the planet, measured in units.
**Offspring_Colonies**: The number of colonies established by the dominant species as offspring expansions.
**Cultural_Exchange_Programs**: The number of cultural exchange initiatives with other planets or species.
**Military_Engagements**: The number of military conflicts or engagements that have occurred on the planet.
**Inhabitants_Disputes**: The number of disputes or conflicts among the planet’s inhabitants.

**Resource_Mining_Operations**: The total number of resource extraction operations active on the planet.
**Resource_Allocation_Credits**: The amount of credits allocated to managing and distributing resources.
**Young_Colonies**: Newly established colonies or outposts of the dominant species.
**HeavyMetals_Concentration**: The concentration of heavy metals in the planet's environment.
**Terraforming_Initiatives**: The number of projects initiated to alter the planet's environment to support life.
**Planet_ID**: A unique identifier assigned to the planet.
**Liquid_Energy_Consumption_Terawatts**: The planet's total consumption of liquid energy resources, measured in terawatts.
**Alien_Population_Count**: The total population of aliens inhabiting the planet.
**CO2_Concentration**: The concentration of carbon dioxide in the planet's atmosphere.
**Dominant_Species_Social_Structure**: The social organization or hierarchy of the planet's dominant species.
**Hydrogen_Concentration**: The concentration of hydrogen in the planet's atmosphere or environment.
**Colonization_Year**: The year the planet was first colonized by alien species.
**Species_Expansion_Response**: The dominant species' response or activity related to expanding their influence or territory.
**Galactic_Trade_Revenue**: The revenue generated from intergalactic trade conducted by the planet.
**Alien_Civilization_Level**: The level of development and sophistication of the alien civilization inhabiting the planet.
**Interstellar_Contact_Cost**: The total cost associated with maintaining communication and relations with other planets.
**Interplanetary_Communications**: The number of communications or signals exchanged between planets.