Projects for Artificial Intelligence and Machine Learning

This document describes the projects for Artificial Intelligence and Machine Learning. They are valid only for the academic year 2023/24. **Projects are mandatory**, whether you take Test1/Test2 or the oral exam. The project will contribute to 30% of your total grade in the class. 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 and fangeletti@luiss.it with the subject [AI/ML PROJECT 23/24] and, in CC, the components of the team. 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 25 2023 at 23:59 (CET, Central European Time). If you do not send an email by the deadline, 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. For instance, if the first two exam dates are **December 12 2023** and January **9 2024**, then you must submit your project by **December 5 2023** or by **January 2 2024**, respectively.

What to submit for your project

Each group must submit via mail to gitaliano@luiss.it and fangeletti@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. 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.

Datasets

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

Projects proposals

Project 1: Trains

As part of your duties as senior data scientist for the famous ThomasTrain company, you are assigned to understand the satisfaction of the customers even without a direct evaluation. To accomplish this task, the company provided you with the "trains_dataset.csv". Understanding the customers' satisfaction will help the marketing team to effectively target users with promotions and making the retention higher.

Dataset features

- Satisfied: whether the customer is satisfied
- Onboard General Rating: rating from 0 to 5 about the service on board
- Work or Leisure: was the travelling for work or leisure
- Baggage Handling Rating: rating from 0 to 5 about the handling of the baggage
- Age: the age of the customer
- Cleanliness Rating: rating from 0 to 5 about the cleanliness of the train
- Ticket Class: the class of the ticket
- Loyalty: is part of a loyalty program?
- Food'n'Drink Rating: rating from 0 to 5 about the food and bevarages on board
- Gender: whether male or female
- Online Booking Rating: rating from 0 to 5 about the online booking experience
- Ticket ID: unique ID assigned to the travel ticket
- Onboard Service Rating: rating from 0 to 5 about the service onboard
- Legroom Service Rating: rating from 0 to 5 about the space for the legs
- Arrival Delay in Minutes: the delay on the arrival of the train
- Departure Delay in Minutes: the delay on the departure of the train
- Checkin Rating: rating from 0 to 5 about the checkin experience
- Onboard Entertainment Rating: rating from 0 to 5 about the onboard entertainment experience
- Distance: the distance of the specific travel
- Boarding Rating: rating from 0 to 5 about the boarding
- Onboard WiFi Rating: rating from 0 to 5 about the WiFi service
- Date and Time: of the travel
- Seat Comfort Rating: rating from 0 to 5 about the comfort of the seating
- Track Location Rating: rating from 0 to 5 about the track where the train has been boarded
- Departure Arrival Time Rating: rating from 0 to 5 about the timing of the travel

Assignment

- Perform an Explanatory data analysis (EDA) with visualization
- Generate a training and test set. The test set should be used only at the end
- Preprocess the dataset (remove outliers, impute missing values, encode categorical features with one hot encoding, not necessarily in this order). Your goal is to estimate whether a customer is satisfied
- 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)
- Select the best architecture using the right metric
- Finally, compute the performances of the test set.

Project 2: ShopEasy

Imagine a platform named ShopEasy, a leading e-commerce site that sells a variety of products, from books and gadgets to furniture and fashion. Over the years, they have amassed a vast amount of user data. This data is a gold mine of insights waiting to be discovered. ShopEasy aims to provide personalized user experiences, special promotions, and improved services. But to do this effectively, they first need to understand the buying habits and behaviors of their customers. By applying segmentation to this dataset, ShopEasy aims to uncover these hidden patterns and provide an enhanced, personalized shopping experience for its users.

Dataset features

- personId: Unique identifier for each user on the platform
- accountTotal: Total amount spent by the user on ShopEasy since their registration
- frequencyIndex: Reflects how frequently the user shops, with 1 being very frequent and values less than 1 being less frequent
- itemCosts: Total costs of items purchased by the user
- singleItemCosts: Costs of items that the user bought in a single purchase without opting for installments
- multipleItemCosts: Costs of items that the user decided to buy in installments
- emergencyFunds: Amount that the user decided to keep as a backup in their ShopEasy wallet for faster checkout or emergency purchases
- itemBuyFrequency: Frequency with which the user makes purchases
- singleItemBuyFrequency: How often the user makes single purchases without opting for installments
- multipleItemBuyFrequency: How often the user opts for installment-based purchases
- emergencyUseFrequency: How frequently the user taps into their emergency funds
- emergencyCount: Number of times the user has used their emergency funds
- itemCount: Total number of individual items purchased by the user

- maxSpendLimit: The maximum amount the user can spend in a single purchase, set by ShopEasy based on user's buying behavior and loyalty
- monthlyPaid: Total amount paid by the user every month
- leastAmountPaid: The least amount paid by the user in a single transaction
- paymentCompletionRate: Percentage of purchases where the user has paid the full amount
- accountLifespan: Duration for which the user has been registered on ShopEasy
- location: User's city or region
- accountType: The type of account held by the user. Regular for most users, Premium for those who have subscribed to ShopEasy premium services, and Student for users who have registered with a student ID
- webUsage: A metric (0-100) indicating the frequency with which the user shops on ShopEasy via web browsers. A higher number indicates more frequent web usage

Assignment

- Perform an Explanatory data analysis (EDA) with visualization using the entire dataset
- Preprocess the dataset (remove duplicates, encode categorical features with one hot encoding, not necessarily in this order)
- Define whether this is a regression, classification or clustering problem, explain why and choose your model design accordingly. Test at least 2 different models
- Identify the proper number of segments, and evaluate different options
- Describe the properties of the segments you have identified
- Describe the properties of the customers belonging to each segment

Project 3: MedCare Wellness Research Center

Welcome to MedCare Wellness Research Center! As you know, our primary goal is to better understand the health and wellness of elderly populations across different communities. In the past few years, we've been conducting extensive surveys across various senior living centers and communities to get insights into their health patterns, habits, and lifestyle choices.

Recently, we've gathered a comprehensive dataset from our nationwide surveys. This dataset contains records of thousands of elderly individuals, capturing a wide variety of health metrics and lifestyle attributes. Our ultimate objective is to proactively predict and prevent health issues in our aging population, improving their quality of life.

Dataset features

- Walking Difficulty: Indicates if the individual has difficulty walking (Y for Yes, N for No)
- Torsades de Pointes: A specific type of abnormal heart rhythm (Y for Yes, N for No)
- Skin Cancer: Indicates if the individual has/had skin cancer (Y for Yes, N for No)
- Hours of Sleep: Average number of hours the individual sleeps per night
- How do you Feel: The individual's self-reported feeling about their health (Poor, Fair, Good, Very good, Excellent)
- Asthma Status: Indicates if the individual has asthma (Y for Yes, N for No)

- Do you Exercise: Indicates if the individual exercises regularly (Y for Yes, N for No)
- Gender: Gender of the individual (M for Male, F for Female)
- Kidney Disease: Indicates if the individual has kidney disease (Y for Yes, N for No)
- Is Smoking: Indicates if the individual smokes (Y for Yes, N for No)
- Ethnicity: The ethnic group the individual identifies with (e.g., White, Black, etc.)
- Diabetes: Indicates if the individual has diabetes (Y for Yes, N for No)
- How many Drinks per Week: Number of alcoholic beverages the individual consumes per week
- Age Group: Age bracket the individual falls under (e.g., 65-69, 70-74, etc.)
- Mental Health: Score indicating the mental health status (higher is better)
- Body Mass Index (BMI): A measure that uses weight and height to estimate body fat
- Physical Health: Score indicating the physical health status (higher is better)
- History of Stroke: Indicates if the individual has had a stroke (Y for Yes, N for No)
- Patient ID: A unique identifier for each individual

Assignment

- Perform an Explanatory data analysis (EDA) with visualization.
- Generate a training and test set. The test set should be used only at the end. Your goal is to estimate an individual's "Physical Health" feature
- Preprocess the dataset (remove outliers, encode categorical features with one hot encoding, not necessarily in this order)
- 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)
- Select the best architecture using the right metric
- Compute the performances of the test set
- Explain your findings

Project 4: SafeComm Digital Security Solutions

Welcome to SafeComm Digital Security Solutions! In the modern digital age, people across the globe communicate largely through text messages. SMSs have become an integral part of our daily lives. However, with this ease of communication, there comes a dark side: SMS-based fraud. Unsuspecting individuals often receive malicious or scam texts intending to deceive or cause harm.

SafeComm has recently partnered with a major telecom provider that has shared anonymized SMS data. This dataset comprises a mix of regular day-to-day messages and some potentially fraudulent ones. The objective is to design a mechanism that identifies and flags these fraudulent messages automatically. This way, we can warn users or even prevent these messages from being delivered altogether.

Dataset features

- Fraudulent: Binary indicator if the SMS is fraudulent (1 for Yes, 0 for No)
- SMS Text: The content of the SMS
- ID: A unique identifier for each SMS
- Date and Time: Timestamp indicating when the SMS was sent

Assignment

- Perform an Explanatory data analysis (EDA) with visualization using the entire dataset..
- Preprocess the dataset (impute missing values, encode categorical features with one-hot encoding). Your goal is to estimate whether an SMS is fraudulent
- 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)
- Select the best architecture using the right metric
- Compute the performances of the test set
- Explain your results