Machine Learning Project Description

1. Introduction

The objective of this assignment is to create a ML solution using a (open publicly available) dataset and one or more of the ML techniques introduced in the course following a Data Science approach. Project examples are predicting the effects of a specific action in each domain or optimizing an existing solution from a given domain. To realize this project, concrete requirements and guidelines are further provided.

Requirements and guidelines

2. Description

Data Science practical approach

In this project you will build a ML solution using a (open publicly available) dataset from a field that captures your interest or is reach in data. Once the aim of this project is established, and the data is collected (from the Internet), you will start working on the implementation and its corresponding report that should contain max. 2500 words following the requirements and guidelines provided in the sections below.
The project is done at individual level.

Implementation and a max.2500 words report

Reading: read the course materials corresponding to the technique(s) selected and data analysis process.

Communication: communicate your plans or questions with the teaching team using the Discussion forum in Brightspace and by e-mail.

Time: the expected time needed for this assignment is about 20 hours.

Deadline on 31.01.2025

Deadline: submit this assignment by 31.01.2025.

Project Grade is 80% project execution + 20% report

Grading: this project represents 40% from the final grade associated with this course. The grade for the project is calculated as follows: 80% project execution + 20% report.

3. Deliverables

The deliverables of this assignment are as follows:

- The report of max. 2500 words in .PDF format structured as described in Section 4 of this document.
- The source code of the corresponding implementation using Python in (a set of) Jupyter Notebooks script(s).
- The dataset used in a .CSV or a similar format together with its source (e.g., link).

4. Requirements

Below each of the components of this assignment are described.

4.1.Dataset

Dataset information

Provide the following information to describe the dataset:

- Dataset source e.g., link.
- Dataset name and corresponding domain.

4.2.Report

Report structure

The report should be structured as follows:

- **Introduction:** provide a short introduction into the context of the topic and the topic selected integrating the following elements:
  - Name of student and student ID.
  - Topic selected.
  - Relation between the topic selected and the aim of the course.
  - Limitations of earlier research and own contribution.
- **Goal:** discuss the objective of this project and split it into multiple smaller objectives that will be tackled in this project.
  - Main goal and main research question to be addressed.
  - Smaller objectives and report outline.
- **Data analysis:** describe the dataset used considering entities such as features and target, and discuss processing and cleaning aspects like duplicates and outliers:
  - Dataset description.
  - Data pre-processing / cleaning and analysis of outliers.
  - Data visualization of general properties and interpretation.
- **Methodology and Implementation:** describe the followed methodological approach and discuss important implementation choices providing concrete examples.
  - Research Methodology discussion.
  - Design elaboration.
  - Implementation choices discussion.
- **Evaluation and Results:** discuss the evaluation method and criteria used, and further interpret the results obtained in the context of the topic considering both technical and ethical aspects i.e., positioning in relation to existing socio-ethical norms and values.
  - Evaluation mechanism discussion and evaluation criteria selected.
  - Results description using different data visualization methods e.g., data plots.

- Interpretation of the results obtained and positioning in the field considering both relevant technical and ethical aspects.

➢ **Conclusions and Discussion:** summarize the approach followed in this project and draw conclusions based on the results obtained. Furthermore, discuss the meaning and findings of this project in the context of the course.

- Summarize the context and main research question.
- Discuss the findings of this project in a broader context and in comparison, with past related work.
- Conclusions based on the approach followed of possible ethical and theoretical aspects of this project.
- Discuss limitations and future (research) directions based on the findings obtained.

4.3.Source code

Source code information

The solution should be implemented in Python using Jupyter Notebooks. You can download the platform and find useful documentation using the following links:

- https://jupyter.org/install
- https://docs.anaconda.com/ae-notebooks/4.3.1/user-guide/basic-tasks/apps/jupyter/

5. Supplementary Resources

Additional useful resources

In the Open Datasets section of the course in Brightspace you can find a set of publicly available open datasets that could be used for implementing this assignment. If you would like to use a dataset that is not in this list, please contact the teaching team to let them know about the dataset that you would like to use instead.