Course Project – Data Mining 2024

Welcome everyone! This year, we're introducing an exciting new challenge. Instead of submitting standalone home assignments throughout the course, we'll work on a data science project you can proudly showcase in your portfolio. The focus of this project is to perform classification on a dataset obtained from https://info.data.gov.il/home/. Each student or team must select a **unique dataset!** Make sure to choose a dataset that hasn't been picked by others.

An essential aspect of this project is to conduct an **end-to-end classification pipeline**, as well as perform **unsupervised analysis** such as clustering, anomaly detection, or building a recommendation system. The project can be conducted in groups of **two students**. However, in certain scenarios, individual submission will be permitted, such as active reserves service and circumstances formally acknowledged by the Dean of Students.

Therefore, please make sure you find a fellow classmate as soon as possible.

Individual submission requires explicit approval by email.

All the code (and the data) must be available on **GitHub**. This not only helps you keep track of your work, but it also counts for **10%** of your overall grade. Your repository should include an informative **README** file, data and code in separate folders, and other relevant information such as figures and outputs. A guide for working with GitHub can be found at:

- Git Guide
- Git Version Control (Hebrew)

The first thing to do is pair up into groups of up to **two students** (you can decide to work alone or in a pair, but the grading criteria will be the same). All students should be registered on the project spreadsheet (https://tinyurl.com/mryjnmy3) by **November 24th**. Registration after this date will reduce your grade by 5 points!

Choosing a Dataset and Project Topic

To get started, visit https://info.data.gov.il/home/ and select a dataset that interests you. **Ensure that the dataset you choose is unique to avoid duplication—check the project spreadsheet to confirm that no one else has selected it.**

Your project will involve:

- **End-to-End Classification Pipeline**: This includes data preprocessing, feature engineering, model selection, training, evaluation, and interpretation.
- **Unsupervised Analysis**: Perform clustering, anomaly detection, or build a recommendation system using the same dataset.

An acceptable dataset should include more than **100 rows** with at least **10 meaningful features**

Important Note: If you choose to use another dataset instead of selecting one from the specified portal, the highest possible project grade will be **80** (i.e., **20 points will be deducted** from your final grade).

Things to Consider

- **Data Preprocessing**: While we don't want you to spend excessive time collecting raw data, inspecting and visualizing the data, experimenting with different types of preprocessing, and conducting error analysis are crucial parts of machine learning.
- **Novelty and Originality**: Your project should not merely replicate previous work. Instead, apply techniques to your chosen dataset in a novel way or improve upon existing methods.

Project Proposals (Due December 10th at 11:59 PM)

You will submit a project proposal to receive feedback. Your proposal should be a PDF document that includes the project's title and the full names of all team members and be a maximum of **one page** with a font size of **11**.

Your project proposal should include the following information:

- **Dataset Selection**: Identify the dataset you've chosen from the specified portals, including a brief description.
- **Motivation**: What problem are you tackling? Why is it significant or interesting?
- **Method**: What machine learning techniques are you planning to apply or improve upon for both the classification task and the unsupervised analysis?
- **Intended Experiments**: What experiments are you planning to run? How do you plan to evaluate your algorithms?

Grading: The project proposal is mainly intended to ensure you have decided on a project topic and to provide feedback. As long as your proposal follows the instructions and demonstrates a well-thought-out plan, you should do well. The proposal's grade is **15%** of the project's overall grade. In some cases, the project proposal may not be approved if it is too simplistic or if you didn't follow our guidelines. In these cases, the maximum grade can be up to **10 points**.

Final Write-up (Due January 9th at 11:59 PM)

We appreciate the dedication and hard work you put into your projects, and we will thoroughly review each write-up. We will post all final write-ups online for everyone to read and learn from. If you prefer to keep your write-up private, please notify us at least a week before the final submission deadline.

The final project write-up should be a maximum of **seven pages**, including appendices and figures. If your project involved collaboration or guidance from others, please acknowledge their contributions in your write-up, following the report guidelines.

Please include a section detailing each team member's contributions. If any concerns arise about your team's collaboration, please reach out via email. We may consider team contributions and evaluations when assigning project grades.

Lastly, provide a link to your **GitHub repository** containing your final project's code, data, and a requirements.txt file listing the libraries used. The final report's grade will be based on its clarity, relevance to topics covered in the Machine Learning and Advanced Machine Learning classes, the novelty of the problem, and the technical quality and significance of the work. The write-up's grade is **40%** of the overall project grade, with the presentation accounting for **10%**.

Oral Defense (TBD, around January 26th)

Each team will defend their project, demonstrating knowledge in every aspect. You must show that the code on GitHub performs exactly as presented in the final write-up. Each team member will be asked questions about the project and code and will be graded individually.

Each team member must be familiar with the entire project's code,

Each team member must have a computer with the code up and running prior to the start of the defense with access to Zoom and a microphone/phone.

Frequently Asked Questions

1. **Can we use GenAI (e.g., GPT, Claude, ...)**

While we encourage you to experiment with genAI, this project should be made by you and not by GPT. You can use GPT if you want to improve specific parts of your writing and fix typos. Anyhow, **any text or code that was written or edited by GenAI should colored red in the text or followed by comments in the code!!!!**
<u>For example</u>:

```
data=pd.read_csv('data.csv')
### CODE BY GPT###
feature_list = feature_string.split("\t")

####
dataset = dataset[feature_list]
```

2. **Should the final project use only methods taught in the class?**

We encourage you to use methods/topics/problems covered in class, but you are not restricted to them. Feel free to consult if you need clarification on any method or problem statement.

3. **Is it okay to use a dataset that is not public?**

No. You must choose a dataset from https://info.data.gov.il/home/.

4. **Can I use datasets from Kaggle or other repositories?**

No. Using datasets from Kaggle or other repositories (such as UCI) is not allowed.

5. **Can this project be combined with that of another class?**

No. The project should not be related to your final project or any other class.

6. **What are acceptable team sizes, and how does grading differ as a function of the team size?**

Teams can consist of **one or two students**. The grading criteria are the same regardless of team size. You are **not allowed** to work in groups of three or more students.

7. **Can I change groups?**

Generally not. If there is a justified reason why you can no longer be part of your team, please let us know as soon as possible. Be advised: Switching a partner 30 days before the defense is not permitted and requires explicit approval.

8. **Do I have to be on campus to submit the final report?**

No, the final report will be submitted via Moodle.

9. **What is the late-day policy for a group project?**

Each group is allocated **three late days** to divide between the different submissions (i.e., proposal and final write-up). Note that late submission applies to all group members; use these days wisely.

10. **Can we use Machine Learning libraries such as scikit-learn, or are we expected to implement them from scratch?**

You can use any library and API for the project. List all libraries used in the final report and the requirements.txt file in your repository.

11. **Are we required to use GitHub for version control?**

Yes. You must use GitHub, and we must be able to access your code once you submit your reports. After the class ends, you may choose to make your work public.

12. **What if two teams end up working on the same dataset?**

When registering on the Google spreadsheet, review the datasets your peers have chosen and ensure that no one has selected the same dataset.

13. **Will we be provided any cloud computing resource credit?**

Exceptional projects may be allowed to use the computing infrastructure in our lab. We encourage you to check out Google Colab (https://colab.research.google.com) for free GPU resources.

14. **Are we required to use Python for the project?**

Yes.

15. **Can you repeat how the grade is divided?**

- **Project Proposal** – 15%
- **Final Write-up** – 40%
- **GitHub Repository** – 10%
- **Oral Defense** – 35%