IE0005 Mini Project

Students are encouraged to form team consisting of 4 students. By week 8, those who do not have a team will be randomly assigned by your Tutor to a team or those team with less than 4 students. Your team may choose any ONE of the following datasets for the Mini-Project. You may also find your OWN dataset but get your Tutor's approval first before you embark on the project!

The exact Data Science problem that you define on the dataset may be different for every group, even if the dataset is the same. You may want to do Regression, Classification, Clustering or Anomaly Detection, whichever you like, on the dataset of your choice.

- What's Cooking?
  https://www.kaggle.com/c/whats-cooking/

- Cardiovascular Disease Prediction
  https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

- Fake Job Posting Prediction
  https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction

- Supermarket Store Branches Sales Analysis
  https://www.kaggle.com/surajjha101/stores-area-and-sales-data

Start talking to your teammates as soon as possible, and decide a problem on the datasets. You may seek help from your Tutor regarding any aspect of the Mini Project.

FAQs

What is the Grading Scheme for the Mini-Project?

- 10% for coming up with an interesting problem based on the dataset
- 10% for exploratory data analysis / visualization to understand the data
- 10% for preparing the dataset to suit your specific problem definition
- 20% for the use of data science / machine learning to solve the problem
- 10% for learning something new and/or doing something beyond the course
- 20% for the presentation of your project, teamwork, and overall impression
- 20% for your individual contribution, tested through Q&A after presentation and based on the regularity of your tutorial attendance record

What is an "interesting problem" based on the dataset?

It should not be something that you can solve by copy-pasting the Linear Regression or Decision Tree Classifier codes from our course material. There should be something beyond that, for which you will have to learn something new, or apply some new technique. If you are unsure if your problem is interesting, ask for your Tutor's advice -- they will know much better.

Warning: In the quest for "interesting problem", please do not attempt something that you cannot finish in time.

How much of Visualization should be presented?

It's worth only 10%, so do not spend the bulk of time on cool visualization tools. Do standard exploration of the data, and standard statistical visualizations, as done during the course, just to understand your data well enough. You DO NOT need to produce data dashboards and cool web interfaces to complete a standard DS/ML project.

Warning: In the quest for "cool presentation", please do not try visualization tools that takes too much time to learn.

What do you mean by "preparing the dataset"?

The dataset given to you may not be in the proper format to solve the problem you targeted. Preparing means cleaning the dataset, resizing/reshaping the dataset, removing outliers (if necessary), balancing imbalanced classes (if necessary), grouping the rows/columns as necessary, etc. This is an important part of any DS/ML project.

How much of DS / ML tools should I use for the project?

This is one of the main parts of your project. You may use any tool and technique that you have seen during the course, for example, Regression, Classification, Clustering, or Anomaly Detection. If you want something simple, just stick to Scikit-Learn as your DS/ML toolbox. You may also choose to use new models that you have not seen in the course, like Random Forest for regression, Naive Bayes for classification, DBSCAN for clustering and anomalies, etc.

Warning: In the quest for "quick impression", please do not try complex tools that may take too much time to learn.

What do you mean by "learning something new" beyond the course?

The goal for the mini-project is to make you learn something new. Try to use a new DS/ML model for regression, classification, clustering or anomaly detection, beyond what we have covered in the course. That's the quickest way to prove you learned something new. You may also want to explore a new visualization tool (like Plotly), or a new technique for data preparation (like resampling), or explore additional data (to add to the given datasets).

Warning: In the quest for "quick impression", please do not try too many new things,
which may take too much time.

What is the format for the Presentation?

You will get 8 minutes followed by a 5 minutes Q&A. You may choose to present as a team (everyone talking about individual portions), or a single person leading the presentation. You MUST mention who did which portion of the project. The presentation is a combination of a PPT (slide) presentation, plus a Demonstration of your project on Jupyter Notebook or any other way. You DO NOT need to show raw code during the presentation. You will anyway submit all the code later.

How should we structure the Presentation?

This is just a guideline -- feel free to choose your own style, depending on the work you did.

- First 3 minutes of your presentation
    - You MUST mention which dataset you worked on, and what EXACTLY was your objective.

    - You may then want to BRIEFLY present your Exploratory Data Analysis and observations.
    - Why type of ML project (regression/classification/clustering/anomaly) have you done?
    - You may now briefly clarify why and how the ML problem(s) aim(s) to solve your objective.
- Next 3 minutes of your presentation
    - Is the data already clean/structured for your ML problem? If not, how did you prepare the data?
    - How did you apply the ML technique to solve your problem? Which models did you use? Why?
    - Did you only use tools and techniques learned within the course? What else did you learn/try?
- Last 2 minutes of your presentation
    - What is the outcome of your project? Did you meet your initial objective? Anything interesting?
    - You MUST conclude by clearly mentioning who in your team worked on which portion of work.

Everything mentioned above should be presented within the 8 minutes. You will be timed, so practice it! The Tutor will ask questions after your presentation is over -- this will be for another 5 minutes, per team.

You may choose to present as a team (everyone talking about individual portions), or a single person leading the presentation. You MUST mention who did which portion of the project. The presentation may be a combination of a PPT (slide) presentation, plus a Demonstration of your project on Jupyter Notebook or any other way. Slides are highly recommended, as it structures your flow. You may present from Jupyter/Dashboard too, but be organized.

What will I finally submit for the Mini Project?

You will have to submit your PPT (slide) used for the Presentation, all your codes (Jupyter Notebooks, Python codes, Visualization codes, etc.), with reference to the resources you used during the project. If you build a cool visualization tool or a website, or something similar, you may also submit the link for accessing it, as applicable.

Submission Deadline: 14 April 2023, 23:59. Any ONE member from the team should submit through your NTULearn IE0005 **Lab** site (not the main site). Do not submit in duplicate, please. Late submissions will not be considered.

Submission Format: a single ZIP file containing:

- Gather all your codes (.ipynb notebooks, .py files, etc.) along with the data you used in a single folder.
- Your presentation slides (PPT or PDF).

How is my "individual contribution" judged for the project?

Peer ranking (to be completed by week12) and your Tutor will ask every member of your team questions about the project. This will be done during the 5-min Q&A after your final presentation. You MUST also mention who did which portion of the project. The regularity of your tutorial attendance will also be taken into consideration.