Cx1015 : Mini-Project

Your Group may choose any ONE of the following datasets for the Mini-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 perform Regression, Classification, Clustering or Anomaly Detection, whichever you like, or a combination of techniques, on the dataset of your choice.

Dataset 1 : AirBnB Open Data from Seattle

Source : https://www.kaggle.com/airbnb/seattle

Default problem is Regression. Feel free to define your own.

Primary Data : The dataset posted by AirBnB on Kaggle.
Join the Kaggle platform to obtain the AirBnB Dataset.
This is NOT a Competition. NO participation is needed.

Dataset 2 : Modelling Earthquake Damage

Source : https://www.drivendata.org/competitions/57/nepal-earthquake/

Default problem is Classification. Feel free to define your own.

Primary Data : The competition dataset for Training.
Join the Competition to obtain the Training Dataset.
You DO NOT need to submit the final Test Prediction.
That is, NO NEED to take a part in the Competition.

Dataset 3 : Jester Joke Dataset

Source : http://eigentaste.berkeley.edu/dataset/

No default problem is defined. Define your own.

Primary Data : Dataset 1 : 4.1 million ratings
Ratings (-10.00 to +10.00) of 100 jokes by 73,421 Users
(collected between April 1999 - May 2003)

Optional Data : Dataset 3 : 2.3 million ratings
Updated version of Dataset 1 : 50 new jokes
Over 115,000 new ratings from 82,366 total users
(collected between November 2006 - Mar 2015)

Download the data as ZIP file(s) from the website.
Read the data description on the website carefully.

Dataset 4 : Aviation Accident Database

Source : https://www.kaggle.com/khsamaha/aviation-accident-database-synopses

No default problem is defined. Define your own.

> Primary Data : The dataset is from NTSB, posted on Kaggle.
> Join the Kaggle platform to obtain the NTSB Accident Data.
> This is NOT a Competition. NO participation is needed.

Dataset 5 : Epileptic Seizure Recognition

Source : https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

No default problem is defined. Define your own.

> Primary Data : The dataset is posted on the UCI ML Repository.
> Download the data from "Data Folder", in plain CSV format.
> Read the data description on the website/portal carefully.

FAQs for Mini-Project

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, evaluated through peer assessment

If you are attempting something different, especially something that you think do not fit into the grading scheme, feel free to discuss with your TA, your Lab Instructor, and the Course Coordinator.

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

Talk to your Lab Instructor and Teaching Assistant. They will be able to help you choose something interesting. It should not be something that you can solve by copy-pasting the LinearRegression or DecisionTreeClassifier codes from the regular 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 whether your problem is interesting, ask for the Lab Instructor's or the TA’s advice.

Warning : In quest for "interesting problem", please do not attempt something that you can't 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 do an impressive project.

Warning : In quest for "cool presentation", please do not try something 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 data, resizing/reshaping the data, 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 Regression, Classification, Clustering, Anomaly Detection. If you want something simple, 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, etc.

Warning : In quest for "quick impression", please do not try complex tools that takes 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 new DS/ML model for regression, classification, clustering or anomaly detection, beyond what we have already 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/extra data.

Warning : In quest for "quick impression", please do not try to learn too many new things at the same time.

What is the format for the final Presentation?

You will get 10 minutes. You will present on the Projector, in front of the whole Lab group. You may use your OWN laptop to present. Just one laptop please, as there will not be enough time to change from one laptop to another. Please bring an adapter to VGA/HDMI, and ask the Lab In-Charge for help regarding connectivity.

At least two members of your Group should deliver the presentation and demonstration. You may choose to present as a complete team too, with everyone talking about individual portions on which they contributed. The presentation is a combination of PPT/PDF/Google Slides, plus a Demonstration of your project on Jupyter Notebook/Google Colab or any other way. You DO NOT need to show raw code during the presentation.

Who will grade my Mini-Project?

The Presentations will be attended by the respective Professor in charge of your Lab Session, and your TA. The Instructors (Sourav Sen Gupta and Bo An) will NOT grade your project unless they are your Lab Instructors.

What will I finally submit for the Mini-Project?

You will have to submit your PPT/PDF slides 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 an application, a visualization tool, a dashboard or a website, you may also submit the link to that.

Note : Guidelines on Presentation and Submission of Project will be communicated closer to the date.

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

Your individual contribution to the mini-project will be judged through Peer Assessment, where your Group mates are going to judge your contribution to the project. Each one of you will be asked to judge your own individual contribution against the contribution of each of the other Group mates working together on this.

In addition, your Lab Instructor or TA may ask you questions after your presentation to judge the individual contribution, in case they feel it is non-uniform. Such a scenario will be considered on a case-to-case basis.

You MUST mention, at the end of your presentation, who did which portion of the project. The primary component of your marks (80%) will be awarded for the Group effort. Only 20% is for individual contribution.