CSMODEL Machine Project
Release Date: May 19, 2025 (M), 12:00 AM

Phase 1 Deadline: June 20, 2025 (F), 11:59 PM
Phase 2 Deadline: July 18, 2025 (F), 11:59 PM

---

1 Overview

This project is designed to allow students to apply the concepts and skills covered in the course in real-world data. Students will go through the process of obtaining data, preprocessing, analysis, modelling, data mining, and insight extraction.

2 Groupings

The project must be accomplished in groups of at most four (4) members. Groupings must be finalized by May 23, 2025 (F), after which no more changes can be made in the groupings. Please follow the directives of your instructor on how the groups are to be formed.

3 Phases

The project has two submission phases with the following deliverables:

- **Phase 1:** data description, target research question, preprocessing, and exploratory data analysis
- **Phase 2:** statistical inference, data mining, key insights and conclusions

4 Project Specifications

For your CSMODEL project, you will conduct a comprehensive analysis of a real-world dataset anchored on a general research question of your choice. You must produce a cohesive data narrative showcasing meaningful or interesting insights obtained from the data.

Your must ensure that you will go through all of these minimum requirement tasks in your project:

1. identify a general research question that you aim to answer in your data narrative
2. perform exploratory data analysis, covering **at least** 3 EDA questions, to get a good understanding of the data
3. conduct **at least** 3 statistical tests to establish three sound conclusions from the data
4. apply at least one of the following techniques: (1) rule mining, (2) clustering, or (3) collaborative filtering to discover meaningful insights from the data (you may also choose to apply any of the variants of the above approaches)

Before you start your project, you must plan what you intend to do with your dataset ahead to ensure that it will cover the minimum requirements listed above, while maintaining the cohesiveness of your data narrative.

5 Deliverables

You must submit the following:

1. A Jupyter Notebook containing all the data processing you did in the project. The Notebook should include Markdown cells explaining each process, and highlighting the insights and conclusions. The Notebook should be structured in a way that (1) is easy to understand, and (2) can be run sequentially to reproduce all outputs in your work.
2. A poster that communicates all key findings and insights of your work. The poster should be intuitive to understand, and intended for a general audience.

On the Phase 1 deadline, you must submit a preliminary draft of your Notebook showing your current progress. On the Phase 2 deadline, you must submit the final Notebook along with the poster.

6 Detailed Guides for Accomplishing the Project

6.1 Phase 1

The following steps must be accomplished for the first phase of the project. Note that these steps do not have to be done strictly in a sequential order.

6.1.1 Dataset Description

Each group will work on a real-world dataset. Follow your instructor's directives on what dataset you will be working on. If your instructor lets you choose your own dataset, make sure that you have their approval, and that the dataset you choose is obtained ethically and legally, in accordance with data privacy laws.

The group must familiarize themselves with their dataset. This includes not only knowing the attributes of the dataset, but also understanding how it was collected, recognizing the possible biases and implications of the sampling method, and other factors that are relevant to the potential insights that could be generated from the data. Note that you may need to look at other related sources for this part of the project.

In the Notebook, the following information should be stated in a concise manner:

- A clear, concise description of the dataset
- How the dataset was collected
- Potential implications of how the data was collected on the insights that will be generated

- Struture of the data
  - What each row and column represents (if tabular data)
  - Number of observations
  - What attributes or features are present in each observation
- Brief description of the different attributes in the dataset (if structured), or brief description of the nature of each observation in the data (if unstructured)

6.2 Data Cleaning

Ensure the integrity of the dataset by applying the relevant cleaning steps. For tabular data, this might entail addressing the following:

- Multiple representations of the same categorical value
- Incorrect datatypes
- Missing data
- Duplicate data
- Inconsistent formatting
- Outliers

For unstructured data, you must check for potentially noise or unwanted data, or apply a modelling scheme to convert them first into a tabular format.

In the Notebook, explain all the procedures applied during the data cleaning process.

6.2.1 Research Question & Exploratory Data Analysis

You must come up with a general research question that you want to explore using the dataset that you have.

For CSMODEL, please abide by the following guidelines when coming up with the research question:

- The question does not necessarily have to be groundbreaking, but it must be something that the group is genuinely interested in.
- The question should not be too specific (i.e., focused only on a few variables); it should be something that warrants an extensive exploration of the dataset. For example, instead of "Is the average sleep duration correlated with poor academic performance?", consider "What are the factors associated with poor academic performance?" If your question is too specific, you will have difficulty fulfilling the minimum requirements for the project.
- The question should be feasible to answer given the dataset/s that you have (obviously).

The identification of the research question goes hand-in-hand with the exploratory data analysis (EDA). Sometimes, the EDA can lead you to interesting research questions. On the other hand, sometimes you have an initial research question, which is then refined or confirmed by the EDA. In this project, you don't have to strictly do one before the other. You must iterate between the EDA and question formulation until you come up with a solid angle that is supported by the data.

The whole process must cover **at least** three (3) exploratory data analysis questions. Explicitly state the questions in the Notebook. Having more than 3 questions is acceptable, especially if 3 questions are not enough to get a sufficient understanding of your data.

Answer each EDA question using appropriate numerical summaries (measures of central tendency, measures of dispersion, correlation) and visualizations (histograms, distribution plots, box plots)

6.3 Phase 2

The following steps must be accomplished for the first phase of the project. Note that these steps do not have to be done strictly in a sequential order.

6.3.1 Data Mining

Your project must involve the use of at least one (1) of the following data mining techniques: rule mining, clustering, or collaborative filtering. Variants of the above approaches, even if not discussed in class, are also welcome provided that you are able to explain them on a theoretical level. You are also welcome to apply other computational techniques beyond the scope of CSMODEL, such as machine learning, provided that you have already satisfied the minimum data mining requirement.

In the Notebook, keep in mind the following reminders when presenting the data mining sections:

- If needed, perform preprocessing techniques to transform the data to the appropriate representation before performing data mining. This may include binning, log transformation, conversion to one-hot encoding, normalization, standardization, interpolation, truncation, and feature engineering.
- Some algorithms require the values to be scaled. Make sure to consider this before data mining.
- You are encouraged to use the code that you have from your assignments **in a thoughtful manner**.

6.3.2 Statistical Inference

Your project must involve the use of statistical inference techniques to confirm (or disprove) the conclusions generated. At least three (3) statistical tests must be performed. These tests can be incorporated whereever appropriate in the data narrative. For example, it can be used to test the significance of findings from exploratory data analysis questions, or it can be used to test the validity of the conclusions generated from the data mining techniques.

In the Notebook, keep in mind the following reminders when presenting the statistical inference sections:

- Use statistical inference methods discussed in class.
- Properly state both hypotheses (your hypothesis and the null hypothesis)
- **Make sure to show the necessary assumptions and requirements of the statistical tests used, and justify why it is apprpropriate to use that test for your data.** For example, some statistical tests might require that data is more or less normally distributed.
- Discuss any preprocessing steps you did before computing for the p-value.
- Explicitly mention the important values of the tests such as the p-value and the significance level used.

6.3.3 Insights and Conclusions

At the end of the Notebook, you must include a summary of the key findings and insights that were generated from the project. Make sure that all conclusions are backed up with evidence and statiscal tests from the empirical data.

Keep in mind that you **must produce a cohesive data narrative** about your dataset. That is, the process that you do in the project, including the data mining and statistical inference requirements, must tie together into a meaningful story centered around your general research question. The insights and conclusions at the end of the Notebook must provide a satisfying resolution to the narrative you are telling.

7 Project Presentations

Be prepared to present your project during the Phase 1 and Phase 2 deadlines. Please follow the directives of your instructors regarding presentation.

8 Academic Honesty Policy

Honesty policy applies. Please take note that you are not allowed to borrow and/or copy-and- paste – in full or in part – any existing related program code or solutions from the internet or other sources (such as printed materials like books, or source codes by other people that are not online). You should develop your own codes and solutions from scratch by yourselves. Shorter snippets of code, intended to perform smaller sub-tasks, may be copy-and-pasted, but one must be prepared to explain how they work.

The student handbook states that (Sec. 5.2.4.2):

"Faculty members have the right to demand the presentation of a student's ID, to give a grade of 0.0, and to deny admission to class of any student caught cheating under Sec. 5.3.1.1 to Sec. 5.3.1.1.6. The student should immediately be informed of his/her grade and barred from further attending his/her classes."

The student handbook also states that (Sec. 10.3):

"A student caught cheating, as defined in Sec. 5.3.1.1., shall be penalized with a grade of 0.0 in the requirement or in the course, at the discretion of the faculty member, without prejudice to an administrative sanction. In cases of alleged cheating, the faculty member should report the incident to the Student Discipline Formation Office (SDFO)."

9 Generative AI Use Policy

Following the generative AI use policy stated in the syllabus (allowed in certain contexts), generative AI tools may be used to assist the group in completing the project. However, AI tools should only be used to generate code **snippets** and should not be used to generate the whole code. More specifically, majority of the code (at least 80%) should still be written by the students. Non-code blocks, i.e., text blocks, should be entirely written by the students without the help of any AI tool.

Any generative AI use must be accompanied by the following:

- Critical evaluation and understanding of the outputs produced by generative AI before they are incorporated into the project (if you can't explain it, or it doesn't make sense, don't include it).
- Accountability on the resulting output (if the output is wrong, you are the one responsible for it, not AI)
- Disclosure of generative AI use, following the format below.

When disclosing the use of generative AI tools, include it in a separate section at the end of the Notebook, following the format below:

```
Statement: During the preparation of this work the author(s) used [NAME TOOL/SERVICE]
for the following purposes:

[enumerate a description of all uses of generative AI]

After using this tool/service, the author(s) reviewed and edited the content as needed and take(s)
full responsibility for the content of the publication.
```

Failing to abide by the above three points when using generative AI will result shall be treated as academic dishonesty and will have grave consequences.