School of Computer Sciences

CPC351/CPM351 Principles of Data Analytics

Academic Session: Semester 1, 2024/2025

Project – Data Analytics

I. Discovery

You are provided with seven datasets, as outlined in Table 1. Analyze the datasets and frame each as an analytics problem to be addressed. Additionally, develop hypotheses to test. Ensure that your problem statements and objectives are clearly defined.

Table 1: Datasets.

No	Name of Dataset	Description	Dataset Size
1	01_Building_Energy_Performance_Data_2020.csv	This dataset contains a listing of building energy performance data from Singapore in 2020. URL to download the dataset: eLearn@USM. Dataset and variables description: https://data.gov.sg/datasets?topics=housing&page=2&coverage=&formats=CSV|XLSX&resultId=de86d8a219d0936dbb321ade068a381da	Rows: 565 Columns: 23
2	02_Purchase_Card_Fiscal_Year_2014.csv	This dataset contains information on purchases made through the purchase card programs administered by the state of Oklahoma and higher education institutions. URL to download the dataset: eLearn@USM. Dataset and variables description: https://catalog.data.gov/dataset/purchase-card-pcard-fiscal-year-2014	Rows: 442458 Columns: 11
3	03_Open_Checkbook_FY2021.csv	This dataset contains expenses with vendors/contractors for the first three quarters of fiscal year 2021 (July 2020 through March 2021) for the city of Baltimore, USA. URL to download the dataset: eLearn@USM. Dataset and variables description: https://catalog.data.gov/dataset/open-checkbook-fy2021-dataset-02dd0	Rows: 99296 Columns: 13

No	Name of Dataset	Description	Dataset Size
4	04_City_Expenditures.csv	This dataset contains the 2002-03 to 2021-22 city financial transactions or expenditures for a city in the state of California, USA. URL to download the dataset: eLearn@USM. Dataset and variables description: https://catalog.data.gov/dataset/city-expenditures-86ecf	Rows: 1048575 Columns: 13
5	05_Social_Housing_Register_30_june_2024	This dataset contains details of applications for social housing in Australia, as of 30 June, including type of assistance required, program type, application date and level of assessed housing need URL to download the dataset: eLearn@USM. Dataset and variables description: https://data.gov.au/dataset/ds-qld-963c1f5e-3819-4b02-af0c-8695739ca4cf/details?q=housing%20assistance	Rows: 25222 Columns: 24
6	06_Weather.csv	This dataset contains micro-climate sensors readings at set intervals throughout the day. The sensors are located at various locations in the City of Canning, Western Australia. URL to download the dataset: eLearn@USM. Dataset and variables description: https://data.gov.au/dataset/ds-wa-9b869e8b-e4e0-4574-b211-a7907c708bec/details?q=weather	Rows: 12945 Columns: 7
7	Fuel_Prices_Jan_24.csv	This dataset contains fuel Prices from Queensland, Australia service stations. URL to download the dataset: eLearn@USM. Dataset and variables description: https://data.gov.au/dataset/ds-qld-c59ba00b-8d2b-4a61-896c-889e0b926d22/details?q=transaction	Rows: 39831 Columns: 12

II. Data Preparation

Conduct exploratory data analysis (EDA) and preprocess the data. Based on the selected dataset(s) and the defined problem(s), data preprocessing steps may be required. These could include tasks such as converting variables to the appropriate data type, handling missing values, or removing irrelevant variables, etc.

III. Model Planning and Development

Based on the project goals you have defined; you need to select two machine learning models to apply to the data. Choose any two from the following options: clustering, classification, regression, or association rules analysis.

You can either:

- Apply two different types of models (e.g., clustering and classification) to a single dataset to address a problem, or
- Use one type of algorithm (e.g., classification) on one dataset and a different algorithm (e.g., association rules analysis) on another dataset to solve separate problems.

If the dataset has a large number of attributes (columns), consider using feature selection techniques to reduce the dimensionality.

IV. Submission

This is a group project (a group of four members). The grouping setting will be inherited from Assignment 01 and Assignment 02.

You are required to submit a zip/rar package which consists of the following items to the eLearn@USM:

- R script (in .R format).
- A project report of not more than 10 pages (in pdf format). Only the sample output screen shots and relevant explanation/write-up/description are expected. Also, a cover page which contains your details must be included in your assignment report (not counted as a page limit).

The zip/rar package must be named according to the following notation: CPC351_CPM351_[GroupNumber]_PROJ. For example, for Group03, they must name the zip/rar package as CPC351_CPM351_Group03_ PROJ.

One of the group members is required to submit the zip/rar package. Kindly communicate with your group members before the submission to avoid any miscommunication.

The submission deadline is 19 January 2025 (Sunday), 23:59 p.m. Failure to submit the assignment will be a disadvantage to you.

You will need to make a presentation based on your project submission. Further information about the presentation will be announced via eLearn@USM.

Reference: Kindly state any source of reference in your assignment script should you refer to various sources to complete this assignment.

IMPORTANT: Students who copied or plagiarized other's work or let their work be copied or plagiarized will be given an F grade. The student may be barred from sitting for final exam and reported to the university's disciplinary board.

V. Grading Rubric

This project will be graded according to the project and presentation grading rubrics as shown in Table 2 and Table 3 respectively.

Table 2 consists of four main components (total = 100%, scaled to 20% of your overall grade):

1. Problem framing and objective identification (15%): Frame and explain the problem statements, objectives, and initial hypothesis.
2. Data preparation (25%): Describe and implement exploratory data analysis which includes (data cleaning, data pre-processing, data visualization.
3. Model planning and development (50%): Justify, explain, and implement the machine learning models. This section covers the explanation of the results and insights.
4. Problem and pitfalls (10%): Discuss the mistakes that have been made and the knowledge & experience gained throughout the project implementation.

Table 3 consists of five main components (total = 50%, scaled to 5% of your overall grade):

1. Clear delivery of ideas (10%)
2. Confident delivery of ideas (10%)
3. Effective and articulate delivery of ideas (10%)
4. Understand and respond to questions (10%)
5. Organization (10%)

Table 2: Project grading rubric (scaled to 20% of your overall grade).

	Very Weak (1 – 2 points)	Weak (3 – 4 points)	Fair (5 – 6 points)	Good (7 – 8 points)	Very Good (9 – 10 points)
Problem framing and objective identification (15%)	Not able to frame a problem and objectives.	Able to frame a problem and objectives with minimal clarity.	Able to frame a problem and objectives with satisfactory clarity.	Able to frame a problem and objectives with good clarity.	Able to frame a problem and objectives with excellent clarity.
Data preparation (25%)	Not able to explain and perform exploratory data analysis. Not able to explain and generate visuals to understand the data. Not able to explain and perform the relevant data pre-processing to facilitate the machine learning tasks.	Able to explain and perform exploratory data analysis (with minimal clarity/correctness). Able to explain and generate visuals to understand the data (with minimal clarity/correctness). Able to explain and perform the relevant data pre-processing to facilitate the machine learning tasks (with minimal clarity/correctness).	Able to explain and perform exploratory data analysis (with satisfactory clarity/correctness). Able to explain and generate visuals to understand the data (with satisfactory clarity/correctness). Able to explain and perform the relevant data pre-processing to facilitate the machine learning tasks (with satisfactory clarity/correctness).	Able to explain and perform exploratory data analysis (with good clarity/correctness). Able to explain and generate visuals to understand the data (with good clarity/correctness). Able to explain and perform the relevant data pre-processing to facilitate the machine learning tasks (with good clarity/correctness).	Able to explain and perform exploratory data analysis (with excellent clarity/correctness). Able to explain and generate visuals to understand the data (with excellent clarity/correctness). Able to explain and perform the relevant data pre-processing to facilitate the machine learning tasks (with excellent clarity/correctness).
Model planning and development (50%)	Not able to apply any new idea or knowledge to a given problem. The algorithm implementation is not correct and not comprehensive. Not able to explain the diagnostics and insights of the models.	Limited ability to apply new ideas or knowledge. The algorithm implementation is minimally correct. Able to explain the diagnostics and insights of the models with minimal clarity.	Able to apply new ideas or knowledge to a given problem. The algorithm implementation is partially correct. Able to explain the diagnostics and insights of the models with satisfactory clarity.	Able to apply new ideas or knowledge to a given problem. The algorithm implementation is correct and comprehensive. Able to explain the diagnostics and insights of the models with good clarity.	Able to apply new ideas or knowledge to a given problem and able to propose alternative applications. The implementation based on the alternative applications is correct and comprehensive. Able to explain the diagnostics and insights of the models with excellent clarity.
Problems and Pitfalls (10%)	Not able to perform reflection.	Able to deliver a reflection report with minimal clarity.	Able to deliver a reflection report with satisfactory clarity.	Able to deliver a reflection report with good clarity.	Able to deliver a reflection report with excellent clarity.

Table 3: Presentation grading rubric (scaled to 5% of your overall grade).

	Very Weak (1 – 2 points)	Weak (3 – 4 points)	Fair (5 – 6 points)	Good (7 – 8 points)	Very Good (9 – 10 points)
Clear delivery of ideas (10%)	Not able to deliver ideas clearly and require major improvements.	Able to deliver ideas and require further improvements.	Able to deliver ideas fairly and require minor improvements.	Able to deliver ideas clearly.	Able to deliver ideas with great clarity.
Confident delivery of ideas (10%)	Not able to deliver ideas confidently.	Able to deliver ideas with limited confidence and require further improvements.	Able to deliver ideas fairly and require minor improvements.	Able to deliver ideas confidently.	Able to deliver ideas with great confidence.
Effective and articulate delivery of ideas (10%)	Not able to deliver ideas effectively.	Able to deliver ideas with limited effect and require further improvements.	Able to deliver ideas fairly and require minor improvements.	Able to deliver ideas effectively and articulately.	Ability to deliver ideas with great effect and articulate.
Understand and respond to questions (10%)	Not able to understand and respond to a question.	Able to understand and answer questions but not able to accurately answer the question.	Able to understand and answer questions satisfactorily.	Able to respond to questions well.	Able to fully understand and respond to questions very well.
Organization (10%)	Information is not arranged and unstructured.	Information is arranged in a confused way.	Information is articulated clearly but it is difficult to follow the presentation.	Information is articulated clearly but the flow is somewhat hampered.	Information is articulated clearly and is organized in a structured way with logical flow between parts.

~~END OF PROJECT~~