School of Computer Sciences

CPC351/CPM351 Principles of Data Analytics

Academic Session: Semester 1, 2021/2022

Project – Data Analytics

I. Discovery

You are given three datasets as shown in Table 1. Study the dataset and frame the problem as an analytics problem to be solved. You also need to formulate hypotheses to test. Clearly define your problem statements and the objectives.

Table 1: Datasets.

No	Name of Dataset	Description	Dataset Size
1	MY_Covid-19_Death.csv	This dataset contains information of 31428 patients who died because of Covid-19 in Malaysia. URL to download the dataset: eLearn@USM Dataset and variables description: https://github.com/MoH-Malaysia/covid19-public/tree/main/epidemic/linelist	31428 rows 15 columns
2	Road_Traffic_Accidents.csv	This dataset contains 1907 accidents happened in UK. URL to download the dataset: eLearn@USM Dataset and variables description: https://data.europa.eu/data/datasets/road-traffic-accidents/?locale=en. Please download the "Guidance" to see the descriptions of variables.	1907 rows 14 columns
3	2018 data -WHO values_SO2.xlsx	This dataset gives a snapshot on the air pollutant (SO_2) concentrations in Europe for the year 2018. URL to download the dataset: eLearn@USM Dataset and variables description: https://www.eea.europa.eu/data-and-maps/data/air-pollutant-concentrations-at-station	1679 rows 14 columns

No	Name of Dataset	Description	Dataset Size
4	H&M-Sales-2018-2019.xlsx	This dataset contains records for the Hennes & Mauritz AB (H&M) products sale for 2018 and 2019. H&M is a Swedish multinational clothing-retail company known for its fast-fashion clothing for men, women, teenagers, and children. URL to download the dataset: eLearn@USM Dataset and variables description: https://data-flair.training/blogs/download-hm-sales-2018-data/ https://data-flair.training/blogs/download-hm-sales-2019-data/	249 rows 15 columns

II. Data Preparation

Perform exploratory data analysis and pre-process the data. Depending on the dataset(s) that have been chosen and problem(s) that have been defined, you may have to perform some data pre-processing, e.g., perform conversion to ensure the variable is in the desired type, treating missing values, or remove irrelevant variables etc.

III. Model Planning and Development

Depending on the goal of the project that you have defined, you are required to identify two models to apply to the data. Kindly choose two out of the following four machine learning models (i.e. clustering, classification, regression, or association rules analysis).

You may use two different types of model (e.g., clustering + classification) on one dataset to solve a problem **OR** you can apply one algorithm (e.g., classification) on one dataset and one different algorithm (e.g., association rules analysis) on another dataset to solve different problems.

If there are large number of attributes (columns), you can perform feature selection to reduce the number of attributes.

IV. Submission

This is a group project (a group of three members). Follow the assignment group formation.

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 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 project report.

The zip/rar package must be named according to the following notation: CPC351_CPM351_[Matric]_PROJ. For example, for a group of three students with matric number of 112211, 112222, and 112233 respectively, they must name the zip/rar package as CPC351_CPM351_112211_112222_112233_PROJ.

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

The submission deadline 31 January 2022 (Monday), 23:59 p.m. Failure to submit the project will be a disadvantage to you.

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

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 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 visulization.
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 done 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 idea 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 idea 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 idea 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 idea 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 clearly 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 confidently 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 effectively 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 confused way.	Information articulated clearly but it is difficult to follow the presentation.	Information articulated clearly but the flow is somewhat hampered.	Information articulated clearly and is organized in a structured way with logical flow between parts.

~~END OF PROJECT~~