CSMODEL Term 1, AY 2021 – 2022

Project 2 Specifications – Data Mining

Groupings: 3 members in a group
Deadline: January 28, 2022 (Friday) 11:59 PM
Percentage: 20% (part of the 40% for Projects)

Deliverables:

`Zip` file containing:

- Jupyter Notebook file – `ipynb` file
- Other Python 3 files (if necessary) – `py` files
- Dataset files – `csv` files

Submission guidelines: Submit the `zip` file to AnimoSpace
Filename format: `CSMODEL-Project2-<Section>-Group<#>.zip`

SPECIFICATIONS

You are tasked to go through the process of selecting a dataset, analyzing data, modelling data, and answering research questions.

The project is to be submitted as a Jupyter Notebook and, optionally, some Python 3 source files. The notebook should be a self-explanatory document containing a report of the entire process undertaken to come up with the generated insights from the raw dataset. It should contain markup cells explaining the processes undertaken in the project, as well as code cells showing all the code that was performed. Please make sure that the codes could be successfully run sequentially to replicate the processes done in the project. The notebook should have four distinct sections, namely: Dataset Representation, Exploratory Data Analysis, Data Mining, and Insights and Conclusions.

Dataset Representation

Each group should select their own dataset to analyze from the list of 15 datasets available for this project. Only 1 group per section is allowed to work on a specific dataset from the list. Thus, the instructor will set up a sign-up sheet for groups to reserve a dataset.

In this section of the notebook, you must fulfill the following:

- Load the dataset into a `DataFrame`. Perform necessary operations to properly load your dataset into a `DataFrame`.
- State the name of the dataset. Describe the structure of the dataset file. How many observations are there in the dataset? How many variables are there in the dataset?

- Discuss the observations in the dataset file. What does each observation represent? Since the dataset description is not provided, the group can presume the entity represented by each observation. For example:
    - For an association rule mining dataset, the group may presume that an observation represents a customer transaction from a store, a list of words in a document, and other similar examples.
    - For a clustering dataset, the group may presume that an observation represents a song from a group of songs, an image from a group of images, and other similar examples.
    - For a collaborative filtering dataset, the group may presume that an observation represents a movie being rated by people, a book being rated by readers, and other similar examples.
- Discuss the variables in the dataset file. Since the description of the dataset is not provided, the group can presume the entity represented by each variable. For example:
    - For an association rule mining dataset, the group may presume that a variable represents the presence of a certain item in a customer transaction from a store, the presence of a word in a document, and other similar examples.
    - For a clustering dataset, the group may presume that a variable represents a certain characteristic or feature of a song (i.e., value representing the tempo, rhythm, pitch, and others), a certain characteristic or feature of an image (i.e., amount of black, amount of white, and others), and other similar examples.
    - For a collaborative filtering dataset, the group may presume that a variable represents a rating of a specific person to a movie, a rating of a specific reader to a book, and other similar examples.

Exploratory Data Analysis

Perform exploratory data analysis comprehensively to gain a good understanding of your dataset. The exploratory data analysis should guide you in formulating the research questions of the project.

In this section of the notebook, you must fulfill the following:

- Identify 2 interesting exploratory data analysis questions. Properly state the questions in the notebook.
- Answer the EDA questions using <u>both</u>:
    - Numerical Summaries – measures of central tendency, measures of dispersion, and correlation
    - Visualization – Appropriate visualization should be used. Each visualization should be accompanied by a brief explanation.

- To emphasize, both numerical summary and visualization should be present to answer each question. The whole process should be supported with verbose textual descriptions of your procedures and findings.

Data Mining

Identify the correct data mining technique to apply to your chosen dataset. The technique that you will apply should be appropriate for the dataset. Apply the data mining technique with the provided hyperparameters and answer the provided questions.

For Association Rule Mining:

- Use the `rule_miner.py` file from our exercises. Make sure that your code is working properly. Set `support_t` to 10 and the `confidence_t` to 0.6.
- Perform association rule mining.
- Answer the questions:
    - Using the provided support threshold and confidence threshold, what is/are the association rules that we derived from the dataset?
    - What is/are the confidence of the derived association rules? (Limit to 2 decimal values)

For Clustering:

- State the number of observations per group before clustering.
- Use the `kmeans.py` file from our exercises. Make sure that your code is working properly. Set the `k`, `start_var`, `end_var`, `num_observations`, and `data` to their appropriate values according to the dataset.
- Perform clustering with maximum iterations set to 300.
- Answer the question: After clustering, how many observations of each class are included per cluster?

For Collaborative Filtering:

- Use the `collaborative_filtering.py` file from our exercises. Make sure that your code is working properly. Set `k` to 5.
- Perform collaborative filtering.
- Answer the question: Give the top 5 items that are most similar to the item at index 0.

Insights and Conclusions

Clearly state your answers from the data to answer each provided question. Make sure that all conclusions are backed up with proper data mining procedures.

WORKING WITH GROUPMATES

For this project, you are encouraged to work in groups of at most 3 members. Make sure that each member of the group has approximately the same amount of contribution for the project. Problems with groupmates must be discussed internally within the group, and if needed, with the lecturer.

DELIVERABLES

Submit a `zip` file containing the source code files via AnimoSpace. All exploratory data analysis, data modelling, and core algorithms should be performed using Python 3 code and integrated into the Jupyter Notebook. Other code that you used for the project other than those in the Notebook should also be included in the submission of the project.

HONESTY POLICY AND INTELLECTUAL PROPERTY RIGHTS

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 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 from scratch by yourselves, i.e., in cooperation with your groupmates.

According to the handbook (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."

RUBRIC FOR GRADING

Criteria	Ratings			Points
Description of Variables / Observations / Structure of the Data	**COMPLETE** **10 pts** A description of the variables, observations, and/or structure of the data is provided. The group properly presumed the entities represented by an observation and a variable in the dataset.	**INCOMPLETE** **4 pts** A description of variables, observations, and/or structure is present but is missing for some aspects of the dataset.	**NO MARKS** **0 pt** No overview or description of the data is provided.	10 pts
Exploratory Data Analysis 1	**COMPLETE** **5 pts** The first exploratory data analysis question is sufficiently answered, and both the appropriate numerical summaries and visualizations are presented.	**INCOMPLETE** **2 pts** The first exploratory data analysis question is not sufficiently answered, or the appropriate numerical summaries or visualizations is not presented.	**NO MARKS** **0 pt** There is no analysis done for the first exploratory data analysis question.	5 pts

Exploratory Data Analysis 2	**COMPLETE** **5 pts** The second exploratory data analysis question is sufficiently answered, and both the appropriate numerical summaries and visualizations are presented.	**INCOMPLETE** **2 pts** The second exploratory data analysis question is not sufficiently answered, or the appropriate numerical summaries or visualizations is not presented.	**NO MARKS** **0 pt** There is no analysis done for the second exploratory data analysis question.	5 pts
Data Mining - Identification	**COMPLETE** **10 pts** The data modelling technique is appropriate for the chosen dataset.		**NO MARKS** **0 pt** The data modelling technique used is not appropriate for the chosen dataset.	10 pts
Data Mining - Code	**COMPLETE** **10 pts** The data modelling technique is implemented properly and correctly. All hyperparameters used are correct.	**INCOMPLETE** **4 pts** The data modelling technique is not implemented properly and correctly for all cases; or some of the hyperparameters used are incorrect.	**NO MARKS** **0 pt** The data modelling technique is not implemented at all; or all hyperparameters used are incorrect; or the data modelling technique used is not appropriate for the chosen dataset.	10 pts

Insights and Conclusions	**COMPLETE** **10 pts** The insights and conclusions to the provided question are stated clearly and backed up with correct data mining procedures.	**INCOMPLETE** **4 pts** The insights and conclusions to the provided question are stated but not clearly enough, or some data mining procedures are not implemented correctly.	**NO MARKS** **0 pt** No insights or conclusions are presented for the provided question; or the data modelling technique used is not appropriate for the chosen dataset.	10 pts
			Total points:	50