ngee ann polytechnic

School of InfoComm Technology

Machine Learning

Specialist Diploma in Data Analytics (SDDA)

TFIB

ASSIGNMENT 2

(40% of Machine Learning Module)

Timeline:

Presentation: 16th Oct 2025 (Thurs)

Report & Codes Submission: 17th Oct 2025 (Fri), 23:59

Student Name:	
Student Number:	

Penalty for late submission:

10% of the marks will be deducted every day after the deadline.

NO submission will be accepted after 24th Oct 2025, 23:59.

1 Problem Statement

1. OBJECTIVES

In this assignment we will utilize the **Cluster analysis** and **Association rules** to discovers the similarities between observations and hidden patterns in data using Python.

You will utilise related Python Libraries (through Jupyter Notebook platform) to do this. You are also required to create visualisations to help users explore the data. You can use Python Libraries (e.g. WordCloud, Matplotlib) or other suitable visualisation tools to do this. Finally, you are required to summarise and interpret the results.

2. DATASETS

2.1. BANKS LOAN (CLUSTER ANALYSIS)

The dataset contains 1000 customers' information from a bank. Each customer has a set of features including Age, Debt Ratio, Monthly Income etc. Please refer to the dataset (**LoanDefault.csv**) for more details.

Feature	Description	Type
Age	Age of borrower in years	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer
SeriousDelinquency	Delinquency of 90 days past due or worse	1=yes, 0=no

2.2. TV SHOWS (ASSOCIATION RULES)

We have a set of data containing 9690 record of TV show watched by different customers. We are interested to find correlations between the different TV show, so that we can develop the ability to predict and perhaps recommend the next TV show that a customer will be interested in watching. Please refer to the dataset (**TVShows.csv**) for more details.

3. SUGGESTED TASKS

You are suggested to tackle this problem in the below steps.

3.1. BANKS LOAN (CLUSTER ANALYSIS)

Step 1 – Clustering and visualisation of numerical data

- Perform simple data exploration to familiarise yourself with the dataset. You may perform descriptive analysis in Python using info(), describe(), histograms, boxplots etc. Are there any missing values or outliers. How did you handle these?

- Perform simple data manipulation on numerical data to prepare the data for clustering modeling. Do you need to scale the data or not? Did you ignore any numerical variables? Why?
- Build clustering model using numerical data only.
    - Build using Hierarchical clustering technique
    - For the models:
        - Present your choice for the optimal number of clusters. How did you arrive at this number?
        - Evaluate the models using proper metrics, e.g. Silhouette Scores etc.
        - Analyse the formed clusters using proper visualisation tools (e.g. Matplotlib etc.)

Step 2 - Summarise and Interpret

- Summarise your findings and provide names for clusters.
- Provide an interpretation of each cluster.

3.2. TV SHOWS (ASSOCIATION RULES)

Step 1: Load and Visualise Data

- Perform simple data exploration to familiarise yourself with the dataset. You may perform descriptive analysis in Python using info(), describe(), barchart etc. Are there any empty cells? How did you handle these?
- Download the dataset (TV Shows.csv) from POLITEMall. You are encouraged to utilize visualization approaches to familiarize yourself with the datasets.
- Perform simple data manipulation to prepare the data for association rule mining modeling. Do you need to transform the data or not? Why?

Step 2: Preprocess the Data

- Prepare the data in transaction format.

Step 3: Generate Frequent Itemsets using Apriori Algorithm

- Short-list frequently occurring item sets.

Step 4: Generate the Association Rules

- Generate relevant association rules from item sets.
- Compare and interpret the impacts on association rules by setting up different support thresholds.
- What recommendations would you make?

4. SUGGESTED REPORT FORMAT & CONTENT GUIDELINES

Based on the above, write an **INDIVIDUAL** report with the following sections (see Table below). Sample content description is provided for each section. You are free to include other relevant information you deem necessary in the sections. You are strongly advised to use screenshots to capture details of work done.

*(Note: For a page with 1 inch margins, 11 point Calibri font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Table of Contents	NA
2.	Summary/Overview	500 words
3.	Banks Loan (Cluster analysis) • Data exploration and manipulation on numerical data. Build Hierarchical Clustering model using numerical data • Summarise and interpret the clusters	Min: 1000 words Max: 2000 words
4.	TV Shows (Association rules) • Data exploration and manipulation on numerical data • Preprocess the Data • Generate Frequent Itemsets using Apriori Algorithm • Generate the Association Rules and make recommendations	Min: 1000 words Max: 2000 words
5.	Conclusion • Summarize your work on these two problems	Min: 250 words Max: 500 words
6.	Reflection • Suggest possible further improvement(s) to the current solution. • With reference to the module learning objectives stated, reflect on the skills learnt and the skills you could have learnt better.	Min: 500 words Max: 1000 words

5. PRESENTATION AND DEMONSTRATION

Each student will be given 10 minutes for presentation. Students will present their work remotely through MS Teams in Week17. Your tutor will provide more information later regarding your presentation slot.

6. DELIVERABLES

Presentation and demonstration

- Each student will be given 10 mins on **16th Oct 2025.**
- The presentation will be done remotely through MS Teams.
- Students to present their findings using the Jupyter notebook**. Powerpoint is optional.**
- Your tutor will provide more information later regarding your presentation slot

Assignment report

- Submit the **softcopy** of the report via Turnitin in **POLITEMall**. Deadline for softcopy submission is **17th Oct 2025, 2359 hours**.
- Submit the Jupyter Notebook file (ML_Assignment_2_<name>.ipynb) in POLITEMall. This is the completed Jupyter note book used for your presentation on 16th Oct 2025. Deadline for softcopy submission is **17th Oct 2025, 2359 hours**.

Note: DO NOT PLAGIARIZE (please refer to Ngee Ann Polytechnic Plagiarism Policy webpage for more information)

7. GRADING CRITERIA

	Grading Criteria	Component Weightage
Presentation	a) Quality of work b) Flow of presentation based on content guidelines (see section 4) c) Quality of presentation slides d) Presentation and articulation skills	**50%**
Final Report	a) Quality of work b) Completeness of report based on suggested report sections and content guidelines (see section 1.4) c) Clarity of report, use of proper visual aids and use of proper grammar d) Quality of recommendations for further improvements	**50%**