DS3030: Data Analytics Lab
Assignment 10

Date: Oct 12, 2025

Timing: 2:00 to 4:45 PM Max marks: 14

Instructions

- Part 1 results for WEKA should be demonstrated in the lab itself
- Submit one .ipynb file named as
  [**student name**]_**assignment[number]**_**part[number].ipynb**
  each for Parts 2,3, containing all questions (text blocks) and solutions (code blocks)
- Write **justifications/comments** as required.

---

1 Part I - Association Rule Mining (WEKA)

Objective: Get familiarized with the data mining tool WEKA
Total Marks: 4
Dataset: "*supermarket.arff*"

1. (0.5 points) Load the dataset & remove the "*total*" attribute
2. (1.0 points) Perform Association Rule mining using the **Apriori** algorithm. Choose the appropriate parameters
   - "*metricType*" (with its minimum value)
   - "*minSupport*" (with its upper and lower bounds)
   - "*numRules*"
3. (1.0 points) Perform Association Rule mining using the **FP Growth** algorithm. Choose the appropriate parameters
   - "*metricType*" (with its minimum value)
   - "*minSupport*" (with its upper and lower bounds)
   - "*numRulesToFind*"
4. (0.5 points) In both cases, for any of the top 5 association rules, give your inference for the measures "*Support*", "*Confidence*", "*Lift*"

Part II - Association Rules: Metrics Exploration (Python)

Total Marks: 4

Task Description: For this lab, you can refer to Lab 9 code for the basics of association rule mining. In Lab 10, we will extend that work by exploring two different approaches for generating rules: Filtered Approach and Global Approach.

Note: Write the code for each step in a separate cell in your Jupyter Notebook.

Data Cleaning: Before generating the baskets, you need to perform data cleaning on Online_Retail.csv [1]

(1) Remove null values in the Description column;

(2) Remove cancelled transactions where InvoiceNo starts with 'C';

(3) Remove transactions with Quantity less than 1.;

(4) Clean the Description column: remove extra spaces and convert all text to lowercase;

(5) Remove duplicate rows based on InvoiceNo and Description columns.

1. **TASK 1: Frequent Rule on Filtered Transactions (Item Size = 4)** [1.5]

   (1) After performing the data cleaning steps, create a transactional basket (Invoice × Items)(using crosstab).

   (2) Filter the basket to include only transactions that contain exactly 4 items.

   (3) Apply the fpgrowth algorithm to find frequent itemsets with a minimum support of 0.01.

   (4) Generate association rules from these frequent itemsets using a minimum confidence of 0.1

   (5) Calculate the union size for each rule (number of items in antecedent + consequent) and retain only those rules whose union size equals 4

   (6) Identify and print the best rule sorted by confidence, support, and lift

2. **TASK 2: Rule Mining on All Transactions** [1.5]

   (1) After performing the same data cleaning, consider all transactions regardless of the number of items.

(2) Apply fpgrowth to find frequent itemsets with a minimum support of 0.01

(3) Generate association rules from these frequent itemsets using a minimum confidence of 0.1

(4) Identify and print the best rule sorted by confidence, support, and lift

2 Part III - Clustering Analysis with K-Means (Python)

Total Marks: 6

Description: Apply the K-Means clustering algorithm to analyze two real-world datasets and determine optimal cluster configurations using evaluation metrics.

Datasets:

1. **countries_continents.csv:** Global geographic clustering dataset
   Columns:

   - Country Name: Name of the country
   - Latitude: Geographic latitude coordinate
   - Longitude: Geographic longitude coordinate
   - Continent: The continent the country belongs to

   Use Case: Discover natural geographic groupings of countries based on their locations. This simulates a scenario where we want to cluster countries without prior knowledge of continents.

2. **market_segmentation_data.csv:** Customer behavior analysis dataset
   Columns:

   - Customer Satisfaction (CSAT): Self-reported satisfaction score [1-10]
     - 1 = Very dissatisfied
     - 10 = Very satisfied
   - Brand Loyalty Score: Calculated from churn rate, retention, CLV [-2.5, 2.5]
     - Negative values = Low loyalty
     - Positive values = High loyalty

   Use Case: Segment customers into behavioral groups to develop targeted marketing strategies. The goal is to identify patterns like "fans" (high satisfaction, high loyalty) vs "alienated" customers (low satisfaction, low loyalty).

1. (3 points) **Task 1: Country Clustering Analysis**

    1. **Part A: Visual Exploration**
    Using the countries dataset:
    (a) Load the dataset.
    (b) Implement K-Means clustering with k = 3, 5, and 7 clusters
    (c) Convert continent names to numerical values for clustering
    (d) Generate scatter plots showing:
        - Data points colored by cluster assignment
        - Cluster centroids marked clearly

    Deliverable: Three plots (one for each k value) with a brief observation (2-3 sentences) about how cluster formations change.

    2. **Part B: Optimal Cluster Determination**
    (a) Implement K-Means clustering for k ranging from 2 to 15
    (b) For each k, calculate:
        - WCSS (Within-Cluster Sum of Squares)
        - Silhouette Score
    (c) Create a visualization showing both metrics across all k values
    (d) Identify the optimal number of clusters based on:
        - Elbow point in WCSS curve
        - Maximum Silhouette Score

    Deliverable: One plot showing WCSS and Silhouette Score trends; Report the optimal number of clusters with justification (2-3 sentences)

2. (3 points) **Task 2: Market Segmentation Analysis**
Scaled Clustering using the market dataset

    1. Load the customer behavior dataset
    2. Scale the data appropriately (use StandardScaler or similar)
    3. Apply K-Means clustering with k = 2, 4, and 6 clusters
    4. Generate scatter plots for each k value showing:
        - Customer segments (clusters)
        - Cluster centroids

    Deliverable: Three plots with observations about customer segmentation patterns (2-3 sentences).