Department of Computer Science and Engineering
Green University of Bangladesh

Course Code: CSE 435 **Course Title:** Data Mining **Semester:** Fall 2025

Assignment

Instructions

- Answer all questions with necessary calculations, reasoning, and steps.
- Show intermediate values for marks.
- Use clean formatting for tables, charts, or FP-trees where applicable.

---

1. Compute the **mean**, **median**, **mode**, and **standard deviation** for the dataset:

   $$X = \{50, 60, 80, 90, 100\}$$

   Then find the z-score of each observation.
2. For the dataset

   $$Y = \{2, 5, 7, 8, 10, 15, 18, 20, 25, 60\}$$

   determine the **five-number summary**, calculate the **IQR**, and identify any **outliers**. Draw the corresponding **boxplot**.
3. Normalize the values $[25, 30, 35, 40, 50, 150]$ using:
   (a) Min–Max normalization to range [0, 1].
   (b) Z-score normalization and identify any potential **outliers** based on Z–score values.
   (c) Decimal scaling normalization.
4. The following table describes three individuals:

ID	Height (cm)	Gender	Hobbies	Income ($000)
A	170	Male	3	50
B	165	Female	4	55
C	180	Male	2	70

   Compute:
   (a) Euclidean distance between A and B (numeric attributes only).
   (b) Manhattan distance between A and C.
   (c) Simple Matching Coefficient (SMC) for gender attribute.
   (d) Mixed-type dissimilarity combining numeric and nominal data.(Dissimilarity matrix)

5. Compute the **cosine similarity** and **angle** between two document vectors:

$$D_1 = (2, 3, 0, 5), \quad D_2 = (1, 0, 4, 2)$$

Interpret the similarity value.

6. Given ages: $\{25, 30, ?, 40, 35, ?\}$, impute missing values using:
   (a) Mean substitution
   (b) Median substitution

   Compare variance before and after imputation.
7. Construct a new attribute BMI using:

$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}$$

for Weight = [60, 80, 100] and Height = [1.6, 1.8, 2.0]. Normalize the resulting BMI using Min–Max normalization.

8. Discretize the attribute "Age" = [10, 15, 18, 25, 32, 35, 38, 45, 50, 55] into three bins using:
   (a) Equal-width discretization
   (b) Equal-frequency discretization

   Compare which better preserves data distribution.
9. Two numeric attributes are given:

Sample	X	Y
1	2	4
2	3	5
3	4	8
4	5	9

Compute the **covariance** and **correlation coefficient**. Interpret the results.

10. Consider the following real-world grocery transactions collected from a local Bangladeshi marketplace:

TID	**Items Purchased**
T1	Rice, Lentil, Oil, Salt
T2	Rice, Fish, Oil, Spice
T3	Rice, Lentil, Vegetables
T4	Rice, Chicken, Oil, Salt
T5	Rice, Fish, Vegetables, Spice
T6	Rice, Lentil, Chicken, Oil
T7	Rice, Vegetables, Salt

   (a) Apply the **Apriori algorithm** step-by-step to identify all **frequent itemsets** with a minimum support threshold of **40%**. Clearly show each pass (C1 $\rightarrow$ L1, C2 $\rightarrow$ L2, etc.) with item counts.
   (b) From the frequent itemsets obtained, generate all possible **strong association rules** that satisfy a minimum confidence level of **60%**. Present each rule with its corresponding support and confidence values.

(c) Calculate the **Lift** and **Other Measures** for the rule "*Oil* $\Rightarrow$ *Rice*". Interpret the results in terms of dependency and correlation between the two items.

11. Using the same dataset, construct the **FP-tree** using the **Pattern-Growth (FP-Growth)** method. Illustrate each step as follows:
    (a) List all items in descending order of their frequency of occurrence.
    (b) Build the FP-tree structure based on the ordered frequent items.
    (c) Derive the **conditional pattern bases** and construct corresponding **conditional FP-trees** for the major items.
    (d) Enumerate all **frequent patterns** discovered using the FP-Growth approach and compare them with the Apriori results.
12. Using the same dataset, Apply the **Eclat algorithm** to find frequent itemsets with support $\geq 40\%$. Show all intermediate intersections of TID lists.
13. Given the following frequent itemsets (with support counts):

$$\{A:5, B:5, C:4, AB:4, AC:3, BC:3, ABC:3\}$$

    Identify:
    (a) All **closed** itemsets.
    (b) All **maximal** frequent itemsets.
    (c) Explain the difference between the two.
14. The following contingency table shows the co-occurrence between two grocery items, *Rice (R)* and *Oil (O)*, collected from a local supermarket dataset:

	Oil (O)	**Not Oil (¬O)**	**Total**
Rice (R)	40	10	50
Not Rice (¬R)	20	30	50
Total	60	40	100

    (a) Compute the **support** and **confidence** for the rule $R \Rightarrow O$.
    (b) Using the above contingency table, calculate the following **interestingness measures** step by step:
        i. **Confidence-Based Measures:** All-confidence, Max-confidence, Kulczynski
        ii. **Cosine Similarity**
        iii. **Jaccard Coefficient**
        iv. **Lift**
        v. **Chi-Squared Statistic** (χ^2)
    (c) Interpret the results of each measure. Discuss which metrics indicate a *positive*, *negative*, or *independent* relationship between *Rice* and *Oil*.

Submission: Submit handwritten as PDF by the due date. Include all steps and reasoning.

— End of Assignment —