School of Computational Sciences and Engineering
B.Tech. CSE 5th Semester
2025-26
Data Warehousing & Data Mining (DCS-501)

Assignment-2

Q1. "Data pre-processing is vital step before data mining". Justify the statement and elaborate the different pre-processing steps. (CO2)

Q2. How does Attribute-Oriented Induction help in summarizing large datasets using concept hierarchies, and how can it be applied to real-world data analysis? Provide examples to support your explanation. (CO2)

Q3. How are statistical measures applied to extract meaningful insights from large databases in real-world scenarios? (CO2)

Q4. Elaborate the concept of data integration, transformation and loading. (CO2)

Q5. Find the frequent itemsets and generate association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%) (CO2)

Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

Q6. Compare Apriori algorithm with F-P Growth algorithm. (CO2)

Q7. Use these methods to *normalize* the following group of data: (CO2)

   200, 300, 400, 600, 1000

(a)   min-max normalization by setting *min* = 0 and *max* = 1
(b)   z-score normalization
(c)   z-score normalization using the mean absolute deviation instead of standard deviation
(d)   normalization by decimal scaling

Q8. Use a flowchart to summarize the following procedures for *attribute subset selection*:

(a)   stepwise forward selection
(b)   stepwise backward elimination
(c)   a combination of forward selection and backward elimination

Q11. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem. (CO2)

Q12. Write the steps for computing Principal Component Analysis. (CO2)