

**B.E. COMPUTER SCIENCE AND ENGINEERING FOURTH YEAR FIRST
SEMESTER - 2025**

DATA MINING(HONS.)

Time : Three hours

Full Marks : 100

Answer any **FIVE** Questions

1. (a) Positively Correlated Data vs. Negatively Correlated Data
 (b) K-Means vs. K-Medoids
 (c) Diana vs. Agnes
 (d) Data Mining vs. Data Exploration
 (b) Support vs. Confidence
 [CO1,CO2,CO3,CO4,CO5] [5x4]

2. (a) Why need measuring the Dispersion of Data ? What is Box plot Analysis, Histogram Analysis? What is Quantile Plot, Quantile-Quantile (Q-Q) Plot, Scatter plot?
 (b) What is Cosine Similarity? Find the Cosine Similarity between the documents(Doc)? (d1&d2, d1&d3, d1&d4, d2&d3, d2&d4, d3&d4)
[CO2]

[10+10]

Doc	Play	Ton	Goal	Cricket	Football	Tennis	Chess	Win	Loss
d1	10	5	4	8	10	1	0	3	0
d2	9	1	6	3	5	0	1	1	0
d3	6	3	1	2	1	6	0	0	1
d4	7	0	1	3	1	0	2	1	0

[Turn over

3. (a) What do you mean by Data Quality? What are the Data Quality parameter?
 (b) Suppose two stocks A and B have the following values in one week:
 (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
 Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
 (c) What is Data Reduction? Data Reduction Strategies are? Describe Dimensionality reduction by Principal Components Analysis (PCA)
 [CO3]

[(3+3)+4+(2+3+5)]

4. (a) Why need Statistical description of data ? Role of measuring the central tendency of data
 (b) Find the correlated group(s) from the data given below in tabular form :

	Play Chess	Not Play Chess	Sum(row)
Like science fiction	250(90)	200(360)	450
Not Like science fiction	50(210)	1000(840)	1050
Sum (Col.)	300	1200	1500

- (c) Suppose a hospital tested the age for the patients are 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70 110 then mean, median and standard deviation of the patients.
 [CO2,CO3][(2+5)+7+6]

5. (a) Discuss supervised learning vs. unsupervised learning. Elaborate the role(s) of Entropy, Information Gain and Gain Ratio in attribute selection
 (b)

Tid	Items
10	A,C,D
20	B, C, E
30	A,B,C,E
40	B, E

Find frequent item set (s) (with minimum support is 2 using Apriori algorithm)
 Discuss the demerits of Apriori algorithm.
 FP-growth algorithm in mining frequent pattern is better (Yes /No)? Give reason.
 [CO4][(2+4)+(5+2+7)]

6. (a) What is a cluster ?What is cluster analysis ? Explain with example(s) Clustering for Data Understanding and Applications.
(b) Utility of Clustering? Indicate the Major Clustering Approaches?
Discuss the advantages and disadvantages of partitioning based clustering approach. [CO5]

[(1+2+4)+(2+4+7)]

7. Write short notes on (any four) : [CO1,CO2,CO3,CO4,CO5] [5x4]
(a) Neural Network
(b) Evaluation of Clustering
(c) Correlation Analysis of Nominal Data
(d) Outlier Detection Methods
(e) Computational Complexity of Frequent Item set Mining
(f) DBSCAN