QP Code: SN23BTECH209 Reg. No | | | | | | | | | | AY 23

Gandhi Institute of Engineering and Technology University, Odisha, Gunupur (GIET University)

B. Tech (Fifth Semester - Regular) Examinations, November 2025

23BCSPC35002 - Data Mining and Data Warehousing

(CSE)

Time: 3 hrs Maximum: 70 Marks

Answer ALL questions

(The figures in the right hand margin indicate marks)

PART – A **(2 x 5 = 10 Marks)**

Q.1. Answer ***ALL*** questions	CO #	Blooms Level
a. Why we need to normalize data?	CO1	K1
b. Differentiate nominal, ordinal and binary attribute.	CO1	K2
c. Why pruning is necessary in decision tree induction?	CO3	K1
d. What are the limitations of k-means clustering algorithm?	CO4	K1
e. Write a short note on Web Scraping.	CO5	K1

PART – B **(10 x 5 = 50 Marks)**

<u>Answer ***ALL*** the questions</u>

		Marks	CO #	Blooms Level
2. a.	Explain the data preprocessing techniques in detail.	5	CO2	K1
b.	Find the cosine similarity between the given two term frequency vectors: X=[3,2,0,5,0,0,0,2,0,0] , Y=[1,0,0,0,0,0,0,1,0,2]	5	CO1	K3

(OR)

c. A survey is conducted to see if there is any relationship between gender and preference for a product. — 5 — CO1 — K3

Gender	Like	Dislike	Total
Male	40	20	60
Female	30	10	40
Total	70	30	100

Perform a Chi-Square Test to check if product preference depends on gender (use 5% level of significance).

		Marks	CO #	Blooms Level
d.	The following are the ages (in years) of employees in a company: 21, 23, 25, 28, 29, 30, 33, 35, 36, 38, 40, 42, 45. Construct a box plot and identify any outliers.	5	CO1	K3
3.a.	Differentiate between OLAP and OLTP systems with suitable examples and tabular comparison.	5	CO2	K2
b.	Explain the different types of attributes in data mining with example.	5	CO2	K2

(OR)

		Marks	CO #	Blooms Level
c.	Draw a neat diagram of data warehouse architecture.	5	CO1	K1
d.	Use these methods to normalize the following group of data: 200, 300, 400, 600, 1000	5	CO2	K2

i. min-max normalization by setting min 0 and max 1
ii. z-score normalization
iii. normalization by decimal scaling.

4.a. Consider the Data set D. Given the minimum support2, apply apriori algorithm on this dataset. — 5 — CO3 — K3

Transaction ID	100	200	300	400
Items	A,C,D	B,C,E	A,B,C,E	B,E

b. Calculate Entropy for the dataset, also determine information gain for the attribute Outlook — 5 — CO3 — K3

Outlook	Temp	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes

(OR)

c. Write short notes on Support Vector Machine. 5 CO3 K1

d. The relationship between study hours (X) and marks scored (Y) of 10 students is: 5 CO3 K3

X	25	30	28	29	32	24	36	28	27	21
Y	18	20	21	16	14	13	22	15	19	12

i. Determine the regression line $Y = b_0 + b_1 X$ using the method of least squares.

ii. Estimate the marks for a student who studies 35 hours.

5.a. Suppose that the data mining task is to cluster the following eight points (with (*x*, *y*) representing location) into three clusters: 5 CO4 K3

*A*1(2, 10), *A*2(2, 5), *A*3(8, 4), *B*1(5, 8), *B*2(7, 5), *B*3(6, 4), *C*1(1, 2), *C*2(4, 9): The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the centre of each cluster, respectively.

Use the *k-means* algorithm to show *only*

i. The three cluster centres after the first-round execution

The final three clusters

b. Perform Agglomerative Hierarchical Clustering for the following points using single linkage**:** 5 CO4 K3

Label	**A**	**B**	**C**	**D**	**E**
X	**1**	**2**	**5**	**6**	**3**
Y	**1**	**1**	**4**	**5**	**2**

i. Show distance matrix,

ii. Steps of merging, and

iii. Draw the dendrogram**.**

(OR)

c. Both *k-means* and *k-medoids* algorithms can perform effective Clustering. Illustrate the strength and weakness of *k-means* in comparison with the *k-medoids* algorithm. 5 CO4 K3

d. A store records the purchase patterns of 10,000 customers regarding two products: Computer Games and Video Rentals. The data is summarized in the following table: 5 CO1 K3

	Video: Yes	**Video: No**	**Total**
Game: Yes	4000	3500	7500
Game: No	2000	500	2500
Total	6000	4000	10000

Using correlation analysis, check game -> video pattern is interesting or not. Consider a minimum support of 30% and minimum confidence of 60%.

6.a. Define text mining. Explain the main steps involved in text pre-processing. 5 CO5 K1

b. Convert the above text inti vector form using TF-IDF. 5 CO5 K3

Document	**Text**
D1	"Data mining techniques"
D2	"Data preprocessing and mining"
D3	"Big data analytics"

(OR)

c. Differentiate between Web Content Mining and Web Structure Mining 4 CO3 K2

d. Describe the real-world applications of data mining in 6 CO1 K1

i. Cyber Security ii. Business Intelligence

--- End of Paper ---