Sanjay Ghodawat University, Kolhapur 2019-20

Established as State Private University under Govt. of Maharashtra. Act No XL, 2017 EXM/P/09/01

Year and Program: 2019-20 **School of Technology** **Department of CSE**

Course Code: CST313.3 **Course Title: Data Mining and Warehousing** **Semester – V**

Day and Date: Tuesday 26/11/19 **End Semester Examination (ESE)** **Time:** 3hrs 10.30am to 1.30pm **Max Marks: 100**

Instructions:
1) All questions are compulsory.
2) Assume suitable data wherever necessary.
3) Figures to the right indicate full marks.

Q.1	**Attempt the Following**	Marks	Bloom's Level	CO
CO2	a) Calculate Correlation Coefficient value for finding out relation between variable X and Y based on following dataset. What does the resulting value mean?	07	L_3	CO1

X	Y
1	10
2	20
3	30
4	40
5	50

OR

CO1	a) Calculate covariance value for given dataset and discuss data integration techniques used in preprocessing.	07	L_3	CO1

Aptitude Score	Statistics Score
95	85
85	95
80	70
70	65
60	70

	b) Compare between OLAP and OLTP.	08	L_4	CO1

OR

CO2	b) Differentiate between Star schema, snowflake schema and fact constellation schema with example of each.	08	L_4	CO3

ESE

Q.2 **Attempt the Following**

a) Elaborate on various data mining functions. 07 L_2 CO1

OR

a) What is assumption in naïve Bayes classification? Classify following record 'X' using Naïve Bayes classifier. Use given dataset. 07 L_4 CO2

X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

age	income	student	credit_rating	Class (buys_computer)
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

b) Elaborate on major clustering approaches. 08 L_2 CO1

OR

b) "Group following data points into **three** clusters using k-means clustering. Each point denotes the x and y coordinates of a location. Use Euclidean distance measure. Use A1, B1, and C1 as the initial clusters. 08 L_3 CO2

A1(2,10) A2(2,5) A3(8,4) B1(5,8) B2(7,5) B3(6,4) C1(1,2) C2(4,9)

Q.3 **Solve any Two**

a) List and explain data preprocessing techniques. 08 L_2 CO1

b) Elaborate on data warehouse design. 08 L_2 CO3

c) Find out the best binary split on attribute '**CarType**' using **Gini** index for following data set: 08 L_3 CO2

ESE

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

d) Differentiate between two types of hierarchical clustering and find out clusters at all levels using AGNES for given dissimilarity matrix. 08 L_4 CO2

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Q.4 **Solve any Two**

a) Find out all possible frequent item sets using Apriori algorithm for given transaction dataset, given min_sup = 2. 09 L_3 CO2

ESE

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

b) Discuss various methods used for finding frequent item sets in transaction datasets. 09 L_3 CO2

c) Discuss different types of traversal patterns used under Web usage mining. Find out sequential pattern from following dataset of sessions containing web pages: {A, B, C} 09 L_3 CO2

D = {S1={U1, <A,B,C>}, S2={U2, <A,C>}, S3={U1,<C,B>}

Note: A sequence is large if it is contained in at least one customer's sequence.

Q.5 **Solve any Two**

a) Elaborate on various types of crawlers. Discuss focused crawler in detail. 09 L_2 CO1

b) Discuss PageRank and HITS algorithm used for web structure mining. 09 L_3 CO2

c) Find out frequent item-sets from list of Sessions given below without considering ordering of Web page accesses: 09 L_3 CO3

D = {S1={U1, <A,B,C>}, S2={U2, <A,C>}, S3={U1,<B,C,E>, S4={U3,<A,C,D,C,E>}}, Min_sup_count=1.

Q.6 **Solve any Three**

a) Find frequent item sets from following transactions using 'FP-growth' 06 L_3 CO2

ESE

technique. Min_sup_count =2.

Transaction Id	Items Bought
10	A,B,C
20	A,C
30	A,D
40	B,E,F

b)	Find frequent 1-item item-sets and frequent 2-item item-sets from transactions in Q. 6 a) using Vertical data partitioning method where min_sup_count =1.	06	L_3	CO2
c)	Elaborate on support, confidence and frequent items sets in mining transaction data.	06	L_2	CO1
d)	Discuss issues and applications of Web usage mining.	06	L_3	CO1

*****

ESE