SASTRA DEEMED UNIVERSITY

(A University under section 3 of the UGC Act, 1956)

End Semester Examinations

Nov 2025

Course Code: **INT317**

Course: **DATA MINING AND ANALYTICS**

QP No. **:S1144-7**

Duration: **3** hours

Max. Marks:**100**

PART – A

Answer any Four questions **4 x 20 = 80 Marks**

1. a) A hospital has collected patient health records with over 200 attributes and 2 million entries. The data includes many redundant attributes (e.g., age in years and date of birth) and very detailed logs (e.g., second-by-second heart rate readings). As a data scientist, which data reduction techniques would you apply during preprocessing to make the dataset more manageable without losing critical medical information? (10)

   b) Consider the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Using the data for age, answer the following:

      i) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0]. (3)

      ii) Use z-Score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. (2)

      iii) Use normalization by decimal scaling to transform the value 35 for age. (2)

      iv) Apply binning by means method to smooth these data, using a bin depth of 3. (3)

2. A database has 9 transactions. Let min sup = 2 and min conf=70%.

TID	Items
1	{I1, I2, I5}
2	{I2, I4}
3	{I2, I3}
4	{I1, I2, I4}
5	{I1, I3}
6	{I2, I3}
7	{I1, I3}
8	{I1, I2, I3, I5}
9	{I1, I2, I3}

a) Find all frequent item sets using Apriori algorithm. (15)

b) List all the strong association rules. (5)

3. a) Create the dissimilarity matrix between the items using a simple distance measure based on the various forms of data. The ordinal traits are ranked as follows: Perfect - 1, Good - 2, and Poor - 3. (10)

Object1	Attribute 1 (nominal)	Attribute 2 (ordinal)	Attribute3 (Numerical)
1	A1	Perfect	23
2	A2	Good	45
3	A1	Poor	78
4	A3	Perfect	31

b) Describe multiple logistic regression and how the forward and backward methods are used for variable selection in multiple logistic regression. Specify the statistical criteria used in the selection procedure. (10)

4. a) Consider the given training data and apply Naïve Bayes algorithm to test the data, {Age<=30, Income=Medium, Student=yes, Credit Rating=fair} and predict the Buy Computer is yes or no. (10)

Age	Income	Student	Credit Rating	Buy 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) A university wants to predict whether a student will Pass or Fail an exam based on their Hours of Study and Number of Absences. The following dataset is available: (10)

Student	Hours of Study	Absences	Result
S1	2	7	Fail
S2	3	6	Fail
S3	4	4	Fail
S4	5	2	Pass
S5	6	2	Pass
S6	7	1	Pass
S7	8	0	Pass
S8	9	1	Pass
S9	4	5	Fail
S10	6	3	Pass

Now, a new student S11 studies for 5 hours and has 3 absences. Using K = 3 (3-Nearest Neighbors) and Euclidean distance, classify whether S11 will Pass or Fail.

5. a) A market trader sells ball-point pens on his stall. He sells the pens for a different fixed price, x pence, in each of six weeks. He notes the number of pens, y, that he sells in each of these six weeks. The results shown in the following table.

x	10	15	20	25	30	35
y	68	60	55	48	38	32

i) Calculate the least square regression line y on x. (3)

ii) Predict the number of pens when he sells for 45. (3)

iii) Calculate the coefficient of determination R^2. (4)

b) Compare linear and logistic regression. Derive the equation for sigmoid function in logistic regression. (10)

6. a) The annual salaries (in thousands of dollars) of 8 men in middle management at a given company are: 55.5, 64.8, 68.2, 70.2, 52.4, 56.8, 60.6, 72.5 while those for 6 women are: 56.2, 48.8, 58.4, 50.9, 60.2, 54.5. Let X and Y denote the salaries of the men and women respectively. Assuming normal distribution and equal standard deviation, test the null hypothesis $\mu_x = \mu_y$ against the alternative hypothesis $\mu_x > \mu_y$ at 5 percent level of significance and the critical value at 5% significance is 1.78. (10)

b) Explain the following:
i) Semiparametric regression model. (5)
ii) Non-Parametric regression methods. (5)

PART – B

Answer the following **1 x 20 = 20 Marks**

7. a) Find the Root node of the decision tree for the following Dataset. (15)

ID	Age	Income	Credit History	Loan Approval
1	Young	Low	Bad	No
2	Young	Medium	Bad	No
3	Young	Medium	Good	Yes
4	Young	High	Good	Yes
5	Middle-aged	Low	Bad	No
6	Middle-aged	Medium	Good	Yes
7	Middle-aged	High	Bad	Yes
8	Middle-aged	High	Good	Yes
9	Senior	Low	Good	No
10	Senior	Medium	Bad	No
11	Senior	Medium	Good	Yes
12	Senior	High	Good	Yes

b) Describe the Yule-Walker equations and explain how they are used for parameter estimation in autoregressive (AR) models of time series analysis. (5)

*****