DATA MINING
(AML2101)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following techniques uses mean and standard deviation scores to transform real-valued attributes?
(a) Decimal scaling (b) Min-max normalization
(c) Z-score normalization (d) Logarithmic normalization

(ii) "Determine whether an email is spam or not" is an example of
(a) Finding Frequent Item sets (b) Classification
(c) Clustering (d) Prediction

(iii) If a transaction set consist of 1000 transactions, 300 transactions contain bread, 350 transactions contain butter, 150 transactions contain both bread and butter. Then the confidence of buying bread with butter (butter ⇒ bread) is
(a) 30% (b) 42.86% (c) 50% (d) 65%

(iv) The goal in Naïve Bayes' classifier is to predict class labels using
(a) posterior probability (b) prior probability
(c) likelihood (d) evidence.

(v) The binary entropy is maximum when p(a) =
(a) 1.00 (b) 0.25 (c) 0.50 (d) 0

(vi) In an ANN model, learning constant should be
(a) small (b) constant throughout the epoch
(c) one (d) small but adaptive and remain stable to irrelevant input

(vii) Consider the following learning algorithms:
(i) Logistic Regression
(ii) Perceptron
(iii) Linear Regression
Which of the following option represents classification algorithms?
(a) Only (i) and (ii) (b) Only (i) and (iii)
(c) Only (ii) and (iii) (d) (i), (ii) and (iii).

(viii) Bagging is an ensemble technique that
(a) Combines predictions using a weighted average
(b) Trains multiple models on different subsets of the data
(c) Constructs an ensemble by iteratively updating weights
(d) Uses a committee of experts to make predictions.

(ix) Which of the following is required by K-means clustering?
(a) Defined distance metric (b) Number of clusters
(c) Initial guess as to cluster centroids (d) All of the above.

(x) DBSCAN cannot be used (with high accuracy) for datasets that are
(a) Convex (b) Uniform density
(c) Non-uniform density (d) None of the above.

Fill in the blanks with the correct word

(xi) _____ is a method in which pruning starts even before the decision tree is completely built.

(xii) A split in the construction of a decision tree is said to be homogeneous if the degree of impurity is _____.

(xiii) If the regression equation is y = 23.6 – 54.2x, then 23.6 is the _____ of the regression line.

(xiv) A lending company wants to estimate the loan amount for a customer, who has applied for a possible loan, this is an example of _____.

(xv) Random forest is an example of _____ learning algorithm.

Group - B

2. (a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
[(AML2101.2)(Understand/LOCQ)]

(b) Suppose that a hospital tested the age and % of body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Calculate the Pearson's correlation coefficient. Are these two attributes positively or negatively correlated? *[(AML2101.2)(Apply, Analyze/IOCQ)]*

5 + (6 + 1) = 12

3. (a) State the Apriori principle of a set. *[(AML2101.4)(Remember, Understand/LOCQ)]*

(b) Consider the following set of frequent 3 – itemsets (F_3):
{1,2,3}, {1,2,4}, {1,2,5}, {1,3,4}, {1,3,5}, {2,3,4}, {2,3,5}, {3,4,5}
Assume that there are only 5 items in the dataset.
Now, answer the following questions:

(i) List all candidate 4-itemsets (F_4) obtained by a candidate generation procedure using F_{k-1} x F_1 merging strategy.

(ii) List all candidate 4-itemsets (F_4) obtained by a candidate generation procedure followed in Apriori method.

(iii) List all candidate 4-itemsets (F_4) from the set obtained in part (ii) above, that survive the candidate pruning step of the Apriori algorithm.

[(AML2101.4)(Apply/IOCQ)]

2 + (4 + 4 + 2) = 12

Group - C

4. (a) Define information gain as an attribute selection measure of a decision tree.

[(AML2101.3)(Remember/LOCQ)]

(b) Consider the following data table:

Age	Income	Married	Health	Class
Young	High	No	Fair	No
Young	High	No	Good	No
Middle	High	No	Fair	Yes
Old	Medium	No	Fair	Yes
Old	Low	Yes	Fair	Yes
Old	Low	Yes	Good	No
Middle	Low	Yes	Good	Yes
Young	Medium	No	Fair	No
Young	Low	Yes	Fair	Yes
Old	Medium	Yes	Fair	Yes
Young	Medium	Yes	Good	Yes
Middle	Medium	No	Good	Yes
Middle	High	Yes	Fair	Yes
Old	Medium	No	Good	No

Using the data as shown in the above Table, predict the class label of the following record:

X = (Age= "Young", Income = "Medium", Married = "Yes", Health = "Fair") using Naive-Bayes classifier.

[(AML2101.3) (Apply/IOCQ)]

3 + 9 = 12

5. (a) Briefly describe the working principle of k-NN classification algorithm.

[(AML2101.3) (Understand/LOCQ)]

(b) Consider the following dataset:

Height (CM)	Weight (KG)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal

Apply k-NN classification algorithm on the above dataset to classify the test data (Height, Weight) = (170, 57). Assume the value of k = 3. *[(AML2101.3)(Apply/IOCQ)]*

5 + 7 = 12

Group - D

6. (a) A basic limitation of the Perceptron is that it can't implement XOR function. Explain the reason for this limitation. Show how XOR function can be implemented using Multilayer Perceptron? *[(AML2101.3)(Analyze/IOCQ)]*

   (b) Explain the significance of choosing correct value of learning constant as far as convergence is concerned. *[(AML2101.3)(Analyze/IOCQ)]*

   (c) State Hebb's hypothesis. *[(AML2101.3)(Remember/LOCQ)]*

(3 + 5) + 2 + 2 = 12

7. (a) How do you determine the best fit line for a linear regression model? *[(AML2101.3)(Understand/LOCQ)]*

   (b) Obtain linear regression equation of Y on X and estimate Y when X = 55 from the following dataset: *[(AML2101.3) (Apply/IOCQ)]*

X	40	50	38	60	65	50	35
Y	38	60	55	70	60	48	30

6 + 6= 12

Group - E

8. (a) Explain the difference between complete-link and single-link hierarchical clustering. *[(AML2101.3)(Understand/LOCQ)]*

   (b) Illustrate the strength and weakness of k-means clustering method in comparison with k-medoids clustering. Also illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme. *[(AML2101.3)(Understand/LOCQ)]*

   (c) How do you measure the validity of a cluster? Explain with example. *[(AML2101.3)(Analyze/IOCQ)]*

3 + (3 + 3) + 3 = 12

9. (a) Write the steps of DBSCAN Algorithm in detail. *[(AML2101.3)(Understand/LOCQ)]*

   (b) Describe the process of selecting the parameters Eps (radius that defines the neighbourhood of a point) and MinPts (minimum number of points in the neighbourhood of the core point) in DBSCAN method. *[(AML2101.3)(Understand/LOCQ)]*

   (c) Explain why DBSCAN does not work well for the data having varying density. *[(AML2101.3)(Analyze/IOCQ)]*

5 + 4 + 3 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	39.5	60.5	0