DATA MINING
(CSEN 3105)

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) In a picture, where 7 cats and 10 dogs are present, your dog detection algorithm has detected 9 entities, out of which only 6 are dogs and remaining are cats. What is the precision of your algorithm?
(a) 0.6 (b) 0.66
(c) 6/17 (d) 0.9.

(ii) If σ(X) represents support count of X, confidence (A ⇒ B) may be given by
(a) σ(A ∩ B) / σ(A) (b) σ(A ∪ B) / σ(A)
(c) σ(A ∩ B) / σ(B) (d) σ(A ∪ B) / σ(B)

(iii) Suppose that the minimum and maximum values for the attribute income are ₹12,000 and ₹98,000, respectively. We would like to map income to the range [0.0,1.0]. By min-max normalization, a value of ₹73,600 for income is transformed to
(a) 0.823 (b) 0.716
(c) 0.336 (d) None of (a), (b) & (c).

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

(v) The rules extracted directly from a decision tree are
(a) neither mutually exclusive nor exhaustive (b) mutually exclusive but not exhaustive
(c) exhaustive but not mutually exclusive (d) both mutually exclusive and exhaustive.

(vi) Slack variable is applicable for
(a) Decision Tree (b) Naïve Bayesian Classifier
(c) K-means clustering (d) Support Vector Machine.

(vii) Which of the following is not an example of ensemble learning algorithm?
(a) Support Vector Machine (b) Bagging
(c) Boosting (d) Random Forest.

(viii) In K-means clustering technique, K is the number of
(a) clusters (b) data points
(c) iterations (d) variables.

(ix) **Statement I:** "A non linearly-separable training set in a given feature space can always be made linearly-separable in another space."
Statement II: "Using the kernel trick, one can get non-linear decision boundaries using algorithms designed originally for linear models."
(a) Both the statements are FALSE (b) Statement I is FALSE but Statement II is TRUE.
(c) Statement I is TRUE but Statement II is FALSE (d) Both the statements are TRUE.

(x) When you find noise in data, which of the following option would you consider in k-NN classification?
(a) Value of k can be increased (b) Value of k can be decreased
(c) Noise cannot be dependent on value of k (d) None of (a), (b) & (c).

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) In a binary classification problem, the probability of one class is 0.75. Its entropy is _____.

(xiii) An attribute with possible values that have a meaningful order or ranking among them is called _____.

(xiv) If a rule R covers 6 out of 14 tuples and if it correctly classifies 4 out of them, then the accuracy of R is _____.

(xv) The total number of possible association rules, extracted from a dataset containing 3 distinct items is _____.

Group - B

2. (a) Briefly outline how to compute the dissimilarity between objects described by the following:
(i) Nominal attributes
(ii) Numeric attributes
(iii) Term-frequency vectors. *[(CSEN3105.1)(Remember/LOCQ)]*

(b) Consider the following 2-D dataset:

	A1	A2
X1	1.5	1.7
X2	2	1.9
X3	1.6	1.8
X4	1.2	1.5
X5	1.5	1.0

Given a new data point, x = (1.4, 1.6) as a query, rank the database points based on similarity with the query using Euclidean distance. *[(CSEN3105.1)(Apply/IOCQ)]*

(3 + 3 + 3) + 3 = 12

3. (a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. *[(CSEN3105.2)(Remember,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? *[(CSEN3105.2)(Apply/IOCQ)]*

6 + (5 + 1) = 12

Group – C

4. (a) Consider the following dataset:

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Predict by using Naïve Bayes classifier, whether a Red SUV from Domestic makers will be stolen or not. *[(CSEN3105.3)(Apply/IOCQ)]*

(b) Let's assume that there is one attribute in the given dataset that contains continuous values. How can you use Naïve Bayes classifier in that circumstance? Give the explanation with suitable example. *[(CSEN3105.6)(Evaluate/HOCQ)]*

8 + 4 = 12

5. (a) Define Gini Index. *[(CO1,CO3)(Remember/LOCQ)]*

(b) Consider the training examples shown in the following table for a binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	Male	Family	Small	C0
2	Female	Luxury	Large	C0
3	Male	Sports	Extra Large	C0
4	Female	Sports	Small	C0
5	Male	Sports	Extra Large	C0
6	Male	Luxury	Extra Large	C1
7	Male	Family	Large	C1
8	Male	Luxury	Extra Large	C1
9	Female	Sports	Large	C1
10	Female	Sports	Large	C1

Evaluate the following:
(i) The Gini Index for the overall collection of training examples
(ii) The Gini Index for the Gender attribute

(iii) The expected information needed to classify a tuple in the given dataset
(iv) The expected information needed to classify a tuple in the given dataset, if the tuples are partitioned according to Car Type
[(CO3,CO6)(Evaluate/HOCQ)]

(c) Discuss briefly about Overfitting in decision tree. *[(CO1,CO3)(Understand/LOCQ)]*

2 + 6 + 4 = 12

Group - D

6. (a) What do you mean by support and confidence of any association rule? When do you refer a rule as a strong rule?
[(CSEN105.4)(Remember,Understand/LOCQ)]

(b) Consider the following dataset of transactions with each letter representing an item:

Transaction ID	Items
T1	{E, K, M, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, M}
T4	{C, K, M, U, Y}
T5	{C, E, I, K, O}

Construct the FP-tree for the above dataset and find all frequent item-sets using FP-growth approach considering the minimum support count as 3. *[(CSEN3105.4)(Apply/IOCQ)]*

(2 + 1) + (5 + 4) = 12

7. (a) Using the transaction data shown in the following table, find out the frequent itemsets using Apriori algorithm. Assume minimum support count = 3

Transaction ID	Lists of Items
1	A, B, D
2	A, B, C
3	B, F
4	A, D
5	B, C
6	A, B, D, E
7	A, B, D, F
8	A, C, E
9	A, B, F
10	A, C, E, F

[(CO4,CO6)(Apply/IOCQ)]

(b) Write a short note on Ensemble Classifiers. *[(CO5)(Understand/LOCQ)]*

7 + 5 = 12

Group - E

8. Perform K-means clustering (using Euclidean distance as distance function) on 2-dimensional data points as given in the following table. Assume the initial centroids as (2, 10), (5, 8) and (1,2). Show the centroids and clusters, in all the iterations (maximum 3 iterations).

Points	X coordinate	Y coordinate
P1	2	10
P2	2	5
P3	8	4
P4	5	8
P5	7	5
P6	6	4
P7	1	2
P8	4	9

[(CO3,CO6)(Apply/IOCQ)]

12

9. (a) Define Core Point, Border Point and Noise Point in the perspective of DBSCAN clustering algorithm. *[(CO3)(Remember/LOCQ)]*
(b) Explain why DBSCAN does not work well for data having varying density. *[(CO3)(Understand/LOCQ)]*

(c) Consider the data points provided in the table below. Perform hierarchical clustering, considering complete link method (MAX distance) to generate a cover. Show the dendogram with merging distance on y-axis.

Points	X coordinate	Y coordinate
P1	1	9
P2	2	10
P3	3	4
P4	3	12
P5	4	9
P6	5	6
P7	6	11
P8	7	4
P9	7	6
P10	8	1
P11	10	3
P12	11	2

[(CO3,CO6)(Apply/IOCQ)]

3 + 2 + 7 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	35.41	54.16	10.41

Course Outcome (CO):

After the completion of the course students will be able to

CSEN3105.1 Remember different terminologies in respect of data mining techniques.
CSEN3105.2 Understand and apply the various data pre-processing methods as and when required.
CSEN3105.3 Understand and apply different classification, clustering algorithms to solve various real-life problems.
CSEN3105.4 Analyse various methods for mining the frequent patterns in different real-life situations.
CSEN3105.5 Apply several ensemble techniques, like bagging, boosting, random forests etc. as and when required.
CSEN3105.6 Evaluate various data mining techniques to solve real-world problems.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*