| U.S.N. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|

# B.M.S. College of Engineering, Bengaluru-560019

**Autonomous Institute Affiliated to VTU**

## September / October 2024 Supplementary Examinations

| | |
|---|---|
| **Programme: B.E.** | **Semester: V** |
| **Branch: Information Science and Engineering** | **Duration: 3 hrs.** |
| **Course Code: 22IS5PCDMG** | **Max Marks: 100** |
| **Course: Data Mining** | |

**Instructions**: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

**Important Note:** Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

| | | **UNIT - I** | ***CO*** | ***PO*** | **Marks** |
|---|---|---|---|---|---|
| 1 | a) | For the following vectors, x and y, calculate the indicated similarity or distance measures.<br>i. x = (0,1,0,1), y = (1,0,1,0) cosine, correlation, Euclidean<br>ii. x = (1,1,0,1,0,1), y = (1,1,1,0,0,1) cosine, correlation, Jaccard | *CO1* | *PO1* | **08** |
| | b) | Data preprocessing is considered as a diverse field encompassing various strategies and techniques. Provide detailed description for the following techniques:<br>i. Feature subset selection<br>ii. Discretization and Binerization<br>iii. Sampling | *CO2* | *PO2* | **12** |
| | | **UNIT - II** | | | |
| 2 | a) | Provide the procedural steps for Hunt's algorithm. Apply Hunt's algorithm for the following set of records. | *CO2* | *PO2* | **10** |

| ID | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

b) Consider the following data set for a binary class problem. CO2 PO2 **10**

| **A** | **B** | **Class Label** |
|---|---|---|
| T | F | + |
| T | T | + |
| T | T | + |
| T | F | - |
| T | T | + |
| F | F | - |
| F | F | - |
| F | F | - |
| F | T | - |
| F | F | - |

a) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

**OR**

3 a) Comprehend on rule ordering schemes with respect to rule-by-rule basis and class-by-class basis. CO2 PO2 **06**

b) Consider the following set of records: CO2 PO2 **14**

| Owns Home | Married | Gender | Employee | Credit Rating | Risk Class |
|---|---|---|---|---|---|
| Yes | Yes | Male | Yes | A | B |
| No | No | Female | Yes | A | A |
| Yes | Yes | Female | Yes | B | C |
| Yes | No | Male | No | B | B |
| No | Yes | Female | Yes | B | C |
| No | No | Female | Yes | B | A |
| No | No | Male | No | B | B |
| Yes | No | Female | Yes | A | A |
| No | Yes | Female | Yes | A | C |
| Yes | Yes | Female | Yes | A | C |

Using Entropy and classification error, Find the best split for each of the attributes.

**UNIT - III**

4. a) Elucidate Naïve Bayes classification algorithm. CO2 PO2 **08**

b) Apply Apriori algorithm for the following set of transactions to find frequent item set with minimum support as 30% CO2 PO2 **12**

| Transaction ID | Items Bought |
|---|---|
| 1 | {a,b,d,e} |
| 2 | {b,c,d} |
| 3 | {a,b,d,e} |
| 4 | {a,c,d,e} |
| 5 | {b,c,d,e} |
| 6 | {b,d,e} |
| 7 | {c,d} |
| 8 | {a,b,c} |
| 9 | {a,d,e} |
| 10 | {b,d} |

Identify the not considered itemset, frequent itemset and infrequent itemset from the set of candidate itemset.

**UNIT – IV**

5 a) Generate Frequent itemset using FP-tree representation for the following market basket analysis data given below ( Minimum support: 20% ): *CO2* *PO2* **10**

| **TID** | **Items** |
|---|---|
| 1 | {a,b} |
| 2 | {b,c,d} |
| 3 | {a,c,d,e} |
| 4 | {a,d,e} |
| 5 | {a,b,c} |
| 6 | {a,b,c,d} |
| 7 | {a} |
| 8 | {a,b,c} |
| 9 | {a,b,d} |
| 10 | {b,c,e} |

b) Elucidate the following *CO2* *PO2* **10**

i) Procedure for rule generation in Apriori algorithm.
ii) Maximal frequent itemset
iii) Closed frequent itemset

**OR**

6 For the following dataset, determine frequent itemset using FP growth algorithm with minimum support as 20%. {10M} *CO3* *PO3* **20**

Generate the rule set with minimum confidence as 70% {10M}

| T/D | Items |
|---|---|
| 1 | {a,b,e} |
| 2 | {b,d} |
| 3 | {b,c} |
| 4 | {a,b,d} |
| 5 | {a,c} |
| 6 | {b,c} |
| 7 | {a,c} |
| 8 | {a,b,c,e} |
| 9 | {a,b,c} |

**UNIT - V**

7 a) Use the Euclidean distance matrix given below to perform single and complete link hierarchical clustering. Represent the results using Dendrogram. *CO2* *PO2* **10**

| | p1 | p2 | p3 | p4 | p5 | p6 |
|---|---|---|---|---|---|---|
| p1 | 0.00 | | | | | |
| p2 | 0.24 | 0.00 | | | | |
| p3 | 0.22 | 0.15 | 0.00 | | | |
| p4 | 0.37 | 0.20 | 0.15 | 0.00 | | |
| p5 | 0.34 | 0.14 | 0.28 | 0.29 | 0.00 | |
| p6 | 0.23 | 0.25 | 0.11 | 0.22 | 0.39 | 0.00 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  | b) | Demonstrate your understanding of the DBSCAN algorithm by providing a detailed explanation of its fundamental concepts and the procedural steps that constitute its implementation. | *CO1* | *PO1* | **05** |
|  |  | c) | Apply your understanding of the basic fuzzy c-means algorithm by providing a comprehensive explanation of its key components and the step-by-step process involved in its execution. | *CO2* | *PO2* | **05** |

******