M.Sc. (SEM-I) Minor II EXAMINATION, 2024
MS 13: Data Warehousing and Mining

Time: 1 Hour (4 PM to 5 PM) **Date: November 8, 2024** **Maximum Marks: 20**

- **Attempt all questions. Use of Scientific Calculator is permitted.**

1. Consider the following databases:

	mc	~mc	m~c	~m~c
D1	10,000	1000	1000	100
D2	1000	100	10,000	100,000

   (i) What is Imbalance ratio? Which database out of D1 & D2 has imbalance?
   (ii) Calculate Kulczynski measure for D1 & D2.
   (iii) Do you think lift measure of D1 is in agreement with their Kulc measure? If not, then why?

2. What is generalization? How will you ensure generalization while training of a model? Consider the following outcome of a Bayesian classifier on a Test dataset of 5 tuples [0.1, 0.4, 0.8, 0.9, 1] with these as True labels [Yes, No, No, No, Yes]. Give confusion matrix for threshold cut-offs of 0.1, 0.5, and 0.9.
3. (i) What's the importance of Entropy in Data Science? Why its formula is negative?
   (ii) Calculate Entropy of datasets with these configuration [7+, 7-], [12+, 2-], and [14+,0-].
   (iii) Name two hyper-parameters of SVM classification algorithm. How do they control the algorithm?
4. (i) For a dataset of 450 samples, how many samples would be there in each fold if you are implementing 5-fold cross validation?
   (ii) Consider the following data points: `x1(1,-3), x2(6,15), x3(8, 12),x4 (2,3).` Which one is a good clustering and why? `Model 1: [(x1, x4),(x2,x3)]; Model 2: [(x1, x2), (x3, x4)]`. Which clustering performance metrics you are using and why?
   (iii) Name trainable parameters of the following clustering algorithm: DBSCAN, and explain how do they control algorithm?

M.Sc. (SEM-I) Minor II EXAMINATION, 2024
MS 13: Data Warehousing and Mining

Time: 1 Hour (4 PM to 5 PM) **Date: November 8, 2024** **Maximum Marks: 20**

- **Attempt all questions. Use of Scientific Calculator is permitted.**

1. Consider the following databases:

	mc	~mc	m~c	~m~c
D1	10,000	1000	1000	100
D2	1000	100	10,000	100,000

   (iv) What is Imbalance ratio? Which database out of D1 & D2 has imbalance?
   (v) Calculate Kulczynski measure for D1 & D2.
   (vi) Do you think lift measure of D1 is in agreement with their Kulc measure? If not, then why?

2. What is generalization? How will you ensure generalization while training of a model? Consider the following outcome of a Bayesian classifier on a Test dataset of 5 tuples [0.1, 0.4, 0.8, 0.9, 1] with these as True labels [Yes, No, No, No, Yes]. Give confusion matrix for threshold cut-offs of 0.1, 0.5, and 0.9.
3. (i) What's the importance of Entropy in Data Science? Why its formula is negative?
   (ii) Calculate Entropy of datasets with these configuration [7+, 7-], [12+, 2-], and [14+,0-].
   (iii) Name two hyper-parameters of SVM classification algorithm. How do they control the algorithm?
4. (i) For a dataset of 450 samples, how many samples would be there in each fold if you are implementing 5-fold cross validation?
   (ii) Consider the following data points: `x1(1,-3), x2(6,15), x3(8, 12),x4 (2,3).` Which one is a good clustering and why? `Model 1: [(x1, x4),(x2,x3)]; Model 2: [(x1, x2), (x3, x4)]`. Which clustering performance metrics you are using and why?
   (iii) Name trainable parameters of the following clustering algorithm: DBSCAN, and explain how do they control algorithm?