| Name: | Roll No: | |
|---|---|---|

# Data Mining and Machine Learning

## Quiz 1, II Semester, 2024–2025

4 February, 2025

1. Consider the following class association rules for the bank loan data set in the table below.

   **(R1)** (Own_House = false) and (Credit_rating = good) → Class = Yes

   **(R2)** (Own_House = true) and (Credit_rating = good) → Class = Yes

| ID | Age | Has_job | Own_house | Credit_rating | Class |
|---|---|---|---|---|---|
| 1 | young | false | false | fair | **No** |
| 2 | young | false | false | good | **No** |
| 3 | young | true | false | good | **Yes** |
| 4 | young | true | true | fair | **Yes** |
| 5 | young | false | false | fair | **No** |
| 6 | middle | false | false | fair | **No** |
| 7 | middle | false | false | good | **No** |
| 8 | middle | true | true | good | **Yes** |
| 9 | middle | false | true | excellent | **Yes** |
| 10 | middle | false | true | excellent | **Yes** |
| 11 | old | false | true | excellent | **Yes** |
| 12 | old | false | true | good | **Yes** |
| 13 | old | true | false | good | **Yes** |
| 14 | old | true | false | excellent | **Yes** |
| 15 | old | false | false | fair | **No** |

   Recall that support measures how often the rule is observed in the data, whereas confidence measures how much the rule is validated by the data. Which of the following is true?

   (a) Both rules have the same support but R2 has higher confidence than R1. ✔

   (b) R2 has higher support and confidence than R1.

   (c) R1 has higher support than R2 but lower confidence.

   (d) R2 has higher support than R1 but lower confidence.

   *Explanation:*

   - R1: false,good → Yes, Support = 2/15, Confidence = 2/4 = 1/2
   - R2: true,good → Yes, Support = 2/15, Confidence = 2/2 = 1

2. We build a decision tree for binary classification, without any pruning, and we discover a leaf node that is not pure — it has representatives of both classes. The most likely explanation is that.

   (a) The training data set is too small.

   (b) There are attributes contributing to the classification that are not captured in the training data. ✔

   (c) The heuristic to maximize information gain at each level has chosen a sub-optimal tree.

   (d) The training data set does not accurately represent the distribution of the entire set of samples.

   *Explanation:*

   If two items have identical attributes and different class labels, there are attributes influencing the class that are not available in the table. This issue will not be addressed by enhancing or improving the training set or changing the tree building process.

3. A new test for tuberculosis is administered to a sample of 1000 patients, 100 of whom actually have tuberculosis. The test is found to be 80% accurate—if a person has the disease, the test is positive 80% of the time and if a person does not have the disease, the test is negative 80% of the time.

   Suppose we regard the test as a classifier for tuberculosis. What is its precision and recall?

   (a) Precision 80/100, Recall 180/900

   (b) Precision 80/100, Recall 80/260

   (c) Precision 80/260, Recall 80/100 ✔

   (d) Precision 180/260, Recall 720/900

   *Explanation*

   From the given data, the confusion matrix is as follows.

| | Predict Yes | Predict No |
|---|---|---|
| Actual Yes | 80 | 20 |
| Actual No | 180 | 720 |

   - The first column gives the precision, $80/(80 + 180) = 80/260$
   - The first row gives the recall, $80/(80 + 20) = 80/100$

---