BUSINESS INTELLIGENCE AND ANALYTICS

ASSIGNMENT WEEK 6:

Total marks = 18 Marks
12 Qns * 1 marks = 12 marks
3 Qns * 2 marks = 6 marks

1. Which of the following statements accurately describes a characteristic of classifiers in machine learning? (1 Mark)

a) They are designed to generate continuous predictions based on input features.
b) They aim to identify the most prominent features that influence a target variable.
c) They categorize objects into distinct and mutually exclusive groups based on their characteristics.
d) They analyse the relationships between variables and estimate their strength and direction.

Ans: c) They categorize objects into distinct and mutually exclusive groups based on their characteristics.

2. Which classification technique is primarily statistics and probability-based? (1 Mark)

a) Decision trees
b) Bayes' Classifiers
c) Support Vector Machines (SVM)
d) Artificial Neural Networks (ANN)

Answer: B) Bayes' Classifiers

3. Which are the two measures used in ROC curves to visualize the performance of classifiers? (1 Mark)

a) Sensitivity and specificity
b) Precision and recall
c) Accuracy and error rate
d) Sensitivity and precision

Answer: A) Sensitivity and specificity

4. Which metric measures the ratio of correctly predicted positive observations to the total predicted positives? (1 Mark)

a) Accuracy
b) Sensitivity
c) Specificity
d) Precision

Answer: D) Precision

5. Imagine you're building a spam filter that classifies emails as spam or not spam. After testing your model, you get the following results:

- True Positives (TP): 100 emails correctly classified as spam

- False Positives (FP): 5 emails incorrectly classified as spam
- False Negatives (FN): 10 emails correctly classified as not spam but are actually spam

What is the recall of your spam filter? (2 Marks)

a) 0.812
b) 0.525
c) 0.909
d) 0.455

Ans: d) 0.909

6. Which technique primarily uses a set of if-else decision rules to categorize data? (1 Mark)

a) Decision trees
b) Artificial Neural Networks (ANN)
c) Support Vector Machines (SVM)
d) Genetic algorithms

Answer: A) Decision trees

7. How does the test data variation contribute to the errors in predicting Y values? (1 Mark)

a) It adds to the reducible error due to inaccurate estimation of f.
b) It causes the irreducible error due to randomness.
c) It directly affects the learning techniques.
d) It minimizes the error through cross-validation.

Answer: B) It causes the irreducible error due to randomness.

8. What does classifier accuracy represent in classification tasks? (1 Mark)
a) The percentage of test set tuples correctly classified by the classifier.
b) The similarity between training and test sets.
c) The number of rules generated by the classifier.
d) The predictive mapping function's complexity.

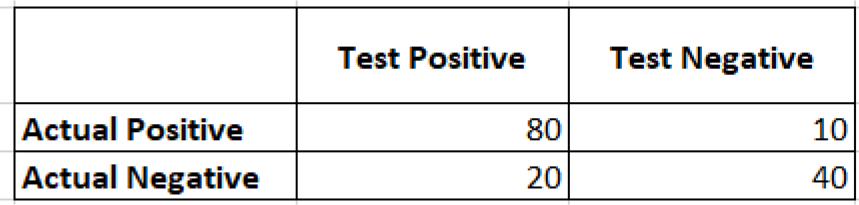
Answer: A) The percentage of test set tuples correctly classified by the classifier.

9. In classification, what does the term "reducible error" primarily refer to? (1 Mark)

a) Error due to randomness
b) Error caused by the classifier model
c) Error that can be minimized by better learning techniques
d) Error that cannot be reduced

Answer: C) Error that can be minimized by better learning techniques

10. In a medical study evaluating a diagnostic test for a certain disease, 150 patients were tested. Of these, 90 patients were diagnosed with the disease, while 60 patients did not have the disease. The model predictions are as follows:



	Test Positive	Test Negative
Actual Positive	80	10
Actual Negative	20	40

Choose the correct option that represents the error rate of the diagnostic test based on the provided classification outcomes. (2 Marks)

a) 0.25
b) 0.2
c) 0.15
d) 0.18

Ans: B) 0.2

11. Overfitting occurs when a classifier incorporates anomalies of the training data that are not present in the general dataset. (True/False) (1 Mark)

Answer: True

12. In unsupervised learning, for every observation i = 1,..., n, we observe a vector of measurements xi but no associated response yi.(True/False) (1 Mark)

Answer: True

13. What is the lift obtained by a marketing team if, without data mining, they achieve a 15% response rate by randomly selecting 20% of potential customers, while with predictive analytics, they target 20% of likely customers and achieve a response rate of 25%? (2 Marks)
a) 2.5
b) 1.67
c) 3.25
d) 6.67

Answer: B) 1.67

14. Choose the correct answer:
    1. K-nearest neighbours or KNN is an unsupervised classification algorithm
    2. K-means Clustering is a supervised classification algorithm.
a) 1 and 2 are correct
b) Only 1 is correct
c) Only 2 is correct
d) Both are wrong

Ans: d) Both are wrong

15. Which of the following is NOT a commonly used classification technique? (1 Mark)
    a) Decision trees
    b) Logistic regression

c) K-nearest neighbours (KNN)
d) Principal component analysis (PCA)

Answer: D. Principal component analysis (PCA)