|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **National University of Computer and Emerging Sciences, Lahore Campus** | | | | |
| C:\Users\saif\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.Word\final design.jpg | **Course:** | **Data Mining** | **Course Code:** | **CS4059** |
| **Program:** | **BS(Data Science)** | **Semester:** | **Spring 2024** |
| **Duration:** |  | **Total Marks:** | **80** |
| **Due Date:** | **11Jun24** | **Weight** |  |
| **Section:** | **A, B & C** | **Page(s):** | **4** |
| **Exam:** | **Final Exam** | **Roll No.** |  |
| **Instruction/Notes:**   * Read the Questions carefully. Make sure you have understood the requirements/expectations of the Questions and answer accordingly. * Any form of cheating or plagiarism will result in an award of ZERO marks. * For MCQs, you must attempt them on the sheet provided and fill the MCQs on Google Classroom. * For Coding Question, you must submit them on Google Classroom renamed as “L21XXXX.ipynb” * Don’t submit the databases or any other file on Google Classroom. | | | | |

**Question #1 MCQs [40 marks]**

1. Which of the following is not a type of attribute in WEKA?

- Answer: D) Ordinal

2. What does the "Filter" option in WEKA allow you to do?

- Answer: B) Remove missing values

3. The primary algorithm behind the J48 classifier in WEKA is:

- Answer: D) Decision Tree

4. Which metric is not typically used to evaluate a classification model in WEKA?

- Answer: C) Lift

5. In a classification task, which WEKA tool can be used to assess the performance of the model using cross-validation?

- Answer: B) Experimenter

6. Which visualization tool in WEKA helps you understand the decision boundaries of a classifier?

- Answer: C) Decision Boundary Plot

7. Which clustering algorithm is commonly used in WEKA?

- Answer: C) KMeans

8. The main purpose of clustering in data mining is:

- Answer: C) Finding natural groupings in data

9. Apply the NaiveBayes classifier on the Weather dataset with 10-fold cross-validation. What is the accuracy of the model?

- Answer: B) 74%

10. Using the RandomForest classifier on the Breast Cancer dataset with default settings, what is the kappa statistic value?

- Answer: C) 0.80

11. Identify the attribute with the highest number of missing values in the Breast Cancer dataset.

- Answer: D) Node-caps

12. Apply the SMO (Support Vector Machine) classifier to the Diabetes dataset with default settings. What is the precision for the 'tested\_positive' class?

- Answer: B) 0.75

13. Using the FilteredClassifier with the Diabetes dataset, first apply the Normalize filter, then use J48. What is the accuracy of the model?

- Answer: B) 75%

14. Using the J48 classifier on the Titanic dataset, which attribute is at the root of the decision tree?

- Answer: B) Sex

15. Using the IBk (k-nearest neighbors) classifier on the Wine dataset, what is the accuracy when k=3?

- Answer: C) 95%

16. What is the value of the mean absolute error for the IBk (k-nearest neighbors) classifier with k=3 on the Wine dataset?

- Answer: B) 0.04

17. Apply the Logistic classifier on the Heart Disease dataset. What is the AUC (Area Under the ROC Curve) for the model?

- Answer: C) 0.94

18. Load the Iris dataset in WEKA and apply the k-means clustering algorithm with k=3. What is the sum of squared errors (SSE) for the clustering?

- Answer: B) 78.85

19. Load the Weather dataset in WEKA. Use the "Discretize" filter on the 'temperature' attribute. What is the number of discrete bins created by default?

- Answer: B) 10

20. After discretizing the 'temperature' attribute in the Weather dataset, apply the NaiveBayes classifier. Does the accuracy improve compared to the original dataset?

- Answer: A) Yes, by more than 5%