National University of Computer and Emerging Sciences, Lahore Campus

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. What is the primary algorithm behind the J48 classifier in WEKA?

   A) Naive Bayes B) Decision Tree C) Neural Network D) kNearest Neighbors

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

   A) Nominal B) Ordinal C) Numeric D) Binary

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

   A) Accuracy B) Precision C) Recall D) Lift

4. In WEKA, which tool allows you to compare multiple models on the same dataset?

   A) Explorer B) Experimenter C) Knowledge Flow D) Simple CLI

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

   A) Visualize data

   B) Remove missing values

   C) Classify data

   D) Generate association rules

6. Which clustering algorithm is commonly used in WEKA?

A) kMediod B) Hierarchical Clustering

C) KMeans D) Centroidbased Clustering

7. What is the main purpose of clustering in data mining?

A) Predicting future values B) Classifying new instances

C) Finding natural groupings in data D) Visualizing data

8. Which parameter in the Apriori algorithm specifies the minimum acceptable level of confidence for rules?

A) MinSupport B) MinConfidence C) MaxSupport D) MaxConfidence

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

A) 0.60 B) 0.70 C) 0.80 D) 0.90

10. In a confusion matrix, what does a high value in the bottomright cell indicate?

A) High false negatives B) High true positives

C) High true negatives D) High false positives

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

A) Age B) Menopause C) Tumorsize D) Nodecaps

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

A) 70% B) 75% C) 80% D) 85%

13. Load the Weather dataset in WEKA. What type of data preprocessing is required if there are any missing values?

A) Normalization B) Discretization

C) Imputation D) Attribute selection

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

A) Class B) Sex C) Age D) Fare

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

A) 85% B) 90% C) 95% D) 100%

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

A) 0.02 B) 0.04 C) 0.06 D) 0.08

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

A) 0.70 B) 0.80 C) 0.94 D) 1.00

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

A) 56.67 B) 78.85 C) 102.34 D) 133.17

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?

A) 5 B) 10 C) 15 D) 20

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

A) Yes, by more than 5%

B) Yes, by less than 5%

C) No change

D) Accuracy decreases

Question #2 [40 marks]

You are provided with the "Diabetes dataset". Your task is to build a machine learning model to predict its target variable using various Data Mining techniques.

- Data Exploration and Visualization:
    - Load the dataset and explore its structure using Pandas.
    - Visualize key features to gain insights into the data.
- Data Preprocessing:
    - Handle any missing values and outliers in the dataset.
    - Perform feature scaling and transformation if necessary.
- Model Building and Evaluation:
    - Split the dataset into training and testing sets (e.g., 70% training, 30% testing).
    - Build and train a classification model using the following algorithms:

      Naïve Bayes || SVM
    - Evaluate the model's performance using metrics like accuracy, precision, recall, and F1score on the test set.
    - Visualize the confusion matrix and ROC curve for model evaluation.
    - Determine which model classification accuracy is better.