MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY

P.O. Box 972-60200 – Meru-Kenya
Tel: +254(0) 799 529 958, +254(0) 799 529 959, + 254 (0) 712 524 293,
Website: info@must.ac.ke Email: info@must.ac.ke

University Examinations 2024/2025

THIRD YEAR FIRST SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF COMPUTER TECHNOLOGY

CDS 3302: LARGE SCALE DATA ANALYSIS

DATE: DECEMBER 2024 **TIME: 2 HOURS**

INSTRUCTIONS: *Answer question **ONE** (Compulsory) and any other **TWO** questions*

QUESTION ONE (30 MARKS)

a) Explain the concept of overfitting in machine learning and provide two strategies to prevent it in WEKA. (5 Marks)

b) Discuss the role of cross-validation in model evaluation, and describe how you would implement it in WEKA. (5 Marks)

c) Scenario: You are given a highly imbalanced dataset. Outline how you approach training a classification model in WEKA to handle the imbalance? (5 Marks)

d) Explain how the confusion matrix is used to evaluate the performance of classifiers. (5 Marks)

e) Discuss how boosting can improve the accuracy of a classifier. Provide an example using WEKA. (5 Marks)

f) Scenario: You are training a Support Vector Machine model on a dataset using WEKA. Explain the key parameters you would tune to improve the model's performance. (5 Marks)

QUESTION TWO (20 MARKS)

a) Describe the process of clustering customer data to identify market segments. Use a specific clustering method available in WEKA. (6 Marks)

b) Explain the concept of "distance metrics" in clustering and how it influences the clustering results. (7 Marks)

c) Scenario: You are given sales data from a retail chain. Explain how K-Means clustering can be used to segment the data. (7 Marks)

QUESTION THREE (20 MARKS)

a) Discuss the advantages and limitations of using decision trees for classification tasks in WEKA. (8 Marks)

b) Scenario: You are tasked with building a model to classify loan applications as approved or denied. Describe the steps you would take to pre-process the data and build a decision tree model using WEKA. (6 Marks)

c) Explain how Random Forest differs from a single decision tree and describe its advantages. (6 Marks)

QUESTION FOUR (20 MARKS)

a) Scenario: You are analyzing social media usage data. Explain how you would use WEKA to apply association rule mining and extract insights from this data. (6 Marks)

b) Discuss the factors that determine the choice of minimum support and confidence when generating association rules. (7 Marks)

c) Provide an example of a rule generated from a market basket analysis and explain its business implications. (7 Marks)

QUESTION FIVE (20 ARKS)

a) Describe the role of feature engineering in building a robust machine learning model. How can this be done in WEKA? (6 Marks)

b) Scenario: You have a dataset with numerical and categorical features. Explain the pre-processing steps you would take in WEKA before applying a machine learning algorithm. (6 Marks)

c) Discuss the importance of scaling and normalization in machine learning. How can these techniques be applied in WEKA? (8 Marks)