



UGANDA INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY PROGRESSIVE ASSESSMENT EXAMINATIONS ACADEMIC YEAR 2025/2026

PROGRAMME(S): DIPLOMA IN DATA SCIENCE MANAGEMENT AND ANALYTICS (DSMA)

YEAR OF STUDY: ONE SEMESTER TWO

COURSE: APPLIED MACHINE LEARNING

COURSE CODE : DSA1203

DATE: MONDAY 29TH, JUNE 2026

TIME: 9:00 AM – 12:00 PM

DURATION: 3 HOURS

VENUE: LAB 1, LAB 2

INSTRUCTIONS:

Attempt all questions in Section A and TWO questions from Section B.

SECTION A (COMPULSORY)

Question1: A startup at the Nakawa Innovation Hub is developing a machine learning model to help SMEs in the agricultural sector predict maize yields in the Kapchorwa region. Their dataset contains historical rainfall and soil pH levels, but due to faulty sensors in remote areas, 15% of the "Soil Moisture" readings are missing.

As a data technician, explain two methods you would use to handle these missing values (e.g., Mean Imputation vs. Dropping Rows) and justify which is more appropriate for a small dataset where every observation is valuable for the model. (5 marks)

Question2: A group of students and innovators is building a fraud detection model for a local mobile money agent network. The raw data includes transaction amounts in UGX, timestamps in different formats, and a "Location" column with inconsistent naming (e.g., "K'la", "Kampala", "Kla city").

Describe the data preprocessing steps required to "clean" this dataset before it can be used to train a model. Focus specifically on feature scaling for the transaction amounts and data standardization for the location names. (5 marks)

Question 3: A project is proposed to help associations for the elderly predict the risk of malaria outbreaks based on local weather patterns. You have a dataset with features like "Average Temperature," "Stagnant Water Proximity (Yes/No)," and "Mosquito Net Usage."

Identify which variables are categorical and which are numerical. Explain how you would convert the "Yes/No" categorical data into a format that a machine learning model (like Logistic Regression) can understand. (5 marks)

Question4: The Uganda Police Force is looking to use machine learning to automatically categorize "Cybercrime Investigation" reports into types like "Phishing," "System Intrusion," or "Data Theft" based on the text descriptions in their files.

This is a Classification problem. Outline the basic steps of building this model, from splitting the data into training and testing sets to evaluating the model's performance using an accuracy score. Why is it important not to train the model on the same data used for testing? (5 marks)

Question 5:

Uganda Electricity Distribution Company (UEDCL) wants to predict monthly household electricity consumption using customer information such as household size, appliance usage, monthly income, and previous electricity consumption records. The dataset has already been cleaned and is ready for modelling.

Required:

- a) Import the dataset and identify the target and predictor variables. (2 Marks)
- b) Split the dataset into training and testing sets. (2 Marks)
- c) Normalize the numerical features where appropriate and explain why normalization is necessary. (3 Marks)
- d) Build a Linear Regression model to predict monthly electricity consumption. Show all the Python code used. (5 Marks)
- e) Evaluate the model using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. (5 Marks)
- f) Interpret the results and recommend how UEDCL can use the model for electricity demand planning. (3 Marks)

SECTION B

(Attempt any TWO questions from this section)

Question 1 (20 Marks)

UAP Old Mutual Insurance Uganda wants to analyze car sales data to better understand vehicle risk profiles. They need to predict which cars are likely to be sold quickly in the Ugandan market to set appropriate insurance premiums. The dataset contains missing values in the Price and Mileage variables.

Required:

- a) Import the dataset and display its structure and summary statistics. (2 Marks)
- b) Identify and visualize missing values in the dataset. (5 Marks)
- c) Clean the dataset by handling missing values using an appropriate technique and justify your choice. (3 Marks)
- d) Perform exploratory data analysis (EDA) to identify relationships between key variables. (5 Marks)
- e) Prepare the data by encoding categorical variables and splitting it into training and testing sets. (2 Marks)
- f) Build a classification model to predict whether a car will be sold quickly (within a specified timeframe). (4 Marks)
- g) Evaluate the model using appropriate performance metrics (e.g., Accuracy, Precision, Recall, F1-score, Confusion Matrix). (4 Marks)
- h) Identify the most important features influencing the prediction and provide business recommendations for UAP Old Mutual Insurance Uganda. (4 Marks)

Question 2

Pezesha Uganda, a fintech company offering vehicle financing, wants to develop a credit risk model to predict customers likely to default based on vehicle characteristics. The dataset requires comprehensive cleaning before it can be used for loan decision-making.

Required:

- a) Load the dataset and inspect its structure and data types. (2 Marks)
- b) Identify and resolve data quality issues, including:
 - Missing values
 - Duplicate records
 - Inconsistent data formats
 - Outliers (where appropriate). (5 Marks)
- c) Perform feature engineering or feature selection where necessary. (4 Marks)
- d) Encode categorical variables and normalize or scale numerical variables where appropriate. (4 Marks)

- e) Split the dataset into training and testing datasets. (1 Mark)
- f) Build a suitable classification model to predict customer credit default risk. (4 Marks)
- g) Evaluate the model using suitable classification metrics and interpret the results. (5 Marks)
- h) Recommend how Pezesha Uganda can use the model to improve loan approval decisions. (5 Marks)

Question 3

Mulago National Referral Hospital is using a heart disease prediction dataset to develop an early warning system. The hospital wants to identify patients at high risk of heart disease based on routine medical examination data. The dataset contains missing values that require appropriate handling.

Required:

- a) Import the dataset and display its basic information, including data types and summary statistics. (2 Marks)
- b) Examine the dataset for missing values and explain the impact of missing data on predictive modelling. (3 Marks)
- c) Handle the missing values using an appropriate technique and justify your choice. (4 Marks)
- d) Perform exploratory data analysis by generating suitable visualizations for important variables. (3 Marks)
- e) Prepare the dataset by encoding categorical variables (if any) and splitting it into training and testing datasets. (2 Marks)
- f) Build a classification model to predict heart disease risk, showing all the Python code used. (4 Marks)
- g) Evaluate the model using a confusion matrix and other appropriate performance metrics. (4 Marks)
- h) Interpret the model results and explain how they could support clinical decision-making at Mulago National Referral Hospital. (5 Marks)

END