

UGANDA INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY (UICT)
DIPLOMA IN DATA SCIENCE METHODOLOGY AND ANALYTICS
END OF FINAL EXAMINATIONS YEAR I SEMESTER II 2025/2026

COURSE NAME: DATA SCIENCE METHODOLOGY AND APPROACH

CODE: DSA1201:

DATE: TUESDAY 23.06.2026

TIME: 9AM-12PM

VENUE: LAB1, LAB2

INSTRUCTIONS:

Section A has two (2) Case Studies. Choose only One.

Section B is Compulsory

Use Python (Google Colab, Jupyter Notebook, or any Python environment).

Use the provided dataset: **Employee_dataset.csv**, Heart Diseases

Section A: Select one question

CASE STUDY 1: Predicting Student Dropout at a Ugandan Tertiary Institution (20 Marks)

The Uganda Institute of Information and Communications Technology (UICT) has observed an increasing rate of student dropout, especially among first-year students. Institutional reports show inconsistencies in data collected from systems such as LMS, ACMIS, and finance records. Some data fields have missing values (e.g., family income), duplicates, and inconsistent formats. Management wants to use data science to predict students at risk of dropping out early enough for intervention. This aligns with the data science lifecycle, which begins with problem definition and proceeds through data preparation, modeling, and deployment .

a) Problem Definition

- i. Clearly define the business problem in this case. (4 marks)
- ii. Formulate a data science question based on the problem. (4 marks)
- iii. Identify two key stakeholders. (2 mark)

b) Data Understanding & Collection

- i. Identify four types of data that would be required. (4 marks)
- ii. Suggest four possible data sources within the institution. (4 marks)
- iii. State one challenge related to data collection. (2 mark)

2. CASE STUDY 2: Crop Yield Prediction in Eastern Uganda (20 Marks)

The Ministry of Agriculture in Uganda is working with farmers in Eastern Uganda to improve crop productivity. However, crop yields have been inconsistent due to unpredictable rainfall patterns, pests, and soil variability. A dataset has been collected containing:

- Rainfall data (some missing values)
- Soil quality indicators (inconsistent formats)
- Crop yields (some outliers)
- Fertilizer usage

The Ministry wants to use data science to predict crop yields and support decision-making for farmers. This reflects real-world data science problems such as predicting agricultural outcomes using structured methodologies .

a) Using the 5WH framework :

- i. Who are the key stakeholders? (2 mark)
- ii. What is the main problem? (2 mark)
- iii. Why is this problem important? (2 mark)
- iv. Where will the data come from? (2 mark)
- v. How can data science help solve the problem? (2 mark)

b) Exploratory Data Analysis

- i. Suggest two EDA techniques that can be used. (4 marks)
- ii. Explain how EDA helps in understanding the dataset. (3 marks)
- iii. Give three example of a possible pattern that may be discovered. (3 mark)

SECTION B (All Questions in this Section are Compulsory)

Question 1: Scenario

You have been hired as a Junior Data Analyst by a company. Management wants to understand the characteristics of their employees before implementing a new Human Resource Management System.

Perform an Exploratory Data Analysis (EDA) on the provided employee dataset and answer the following tasks.

a. Loading and Understanding the Dataset (7 Marks)

Task

I. Import the required libraries:

- pandas
- seaborn
- matplotlib

II. Load the dataset into Python.

III. Display:

- First five rows of the dataset
- Number of rows and columns
- Data types of all variables

b. Output Required

A short summary describing:

- Number of records
- Number of variables
- Types of variables available

c. Univariate Analysis (8 Marks)

Task 2.1: Histogram (3 Marks)

Create a histogram showing the distribution of employee ages.

Interpretation

Write one observation about:

- Age distribution
- Concentration of employees

Task 2.2: Bar Chart (3 Marks)

Create a bar chart showing the distribution of employees by Gender.

Interpretation

Write one observation about:

- Which gender has the highest count

Task 2.3: Pie Chart (4 Marks)

Create a pie chart showing the proportion of employees in each Salary category.

Interpretation

Write one observation about:

- The salary category with the largest percentage

d. Bivariate Analysis (6 Marks)

Task 3.1: Department vs Years of Experience (4 Marks)

Create a bar plot showing:

- Department on the X-axis
- Average Years_of_Experience on the Y-axis

Interpretation

Write one observation about:

- Which department has the most experienced employees

Task 3.2: Age vs Years of Experience (6 Marks)

Create a scatter plot showing:

- Years_of_Experience on the X-axis
- Age on the Y-axis

Interpretation

Write one observation about:

- Whether there appears to be a positive relationship between age and experience

e. Multivariate Analysis (Heatmap) (9 Marks)

Task

Generate a correlation heatmap for all numerical variables in the dataset.

Interpretation

Answer the following:

- a) Which two variables have the strongest positive correlation?
- b) Are there any variables with weak or negative correlation?
- c) What does a correlation value close to +1 mean?

Question 2

Dataset: Heart Disease Dataset (heart_disease.csv)

SCENARIO

A regional referral hospital has recorded patient medical information including age, gender, blood pressure, cholesterol levels, heart rate, chest pain type, and other health indicators.

The hospital management wants to predict whether a patient is likely to suffer from heart disease and use the results to support early intervention and clinical decision-making.

As a Data Scientist, you are required to apply the complete Data Science Methodology and Approach using the provided Heart Disease Dataset.

(a) Under definition State the following:

- 1. The business problem. (2 marks)
- 2. The data science problem. (2 marks)

(b) State TWO objectives of the project. (2 Marks)

(c) State Two benefits the hospital expects from this project. (2 Marks)

DATA UNDERSTANDING AND PREPARATION

Using the Heart Disease Dataset attempt the following Required Tasks

(d) Import Libraries (2 Marks)

Import appropriate libraries for:

- I. Data manipulation
- II. Visualization
- III. Machine Learning

(e) Load Dataset (1 Marks)

Load the Heart Disease dataset into a Pandas DataFrame.

(f) Explore Dataset (4 Marks)

Display:

1. First five records
2. Last five records
3. Dataset shape
4. Dataset information

(g) Data Cleaning (4 Marks)

Perform the following:

1. Check for missing values.
2. Handle missing values appropriately.
3. Check for duplicate records.
4. Remove duplicates where necessary.

(h) Generate descriptive statistics for all numerical variables and Interpret your findings. (2 Marks)

(i) Using Visual Analysis Create and interpret: (8 Marks)

1. Histogram of Age
2. Bar chart of Heart Disease Cases
3. Pie Chart showing Gender Distribution
4. Correlation Heatmap
5. Scatter Plot of Age vs Cholesterol

(j) Identify and explain at least FIVE insights from the dataset. (5 Marks)

(k) Feature Selection Select appropriate independent variables. (2 Marks)

(l) Split the dataset into: (2 Marks)

- 80% Training Data
- 20% Testing Data

(m) Build a classification model using ONE of the following: (6 Marks)

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier

(n) Train the model using the training data. (3 Marks)

(p) Generate predictions using the test dataset. (3 Marks)

(q) Generate and interpret the Confusion Matrix (2 marks)

(r) Generate and interpret Classification Report (4 Marks)

- I. Accuracy
- II. Precision
- III. Recall
- IV. F1 Score

(s) Discuss whether the model is suitable for deployment in a healthcare environment. (4 Marks)

Prepare a management report containing:

Problem Statement, Methodology Used, Key Findings, Model Performance, Recommendations, Conclusion

(t).Create a simple deployment solution for the trained model.

You may use either:

Option A: Streamlit Application (4 marks)

The application should:

1. Load the trained model.
2. Allow a user to enter:
 - Age
 - Cholesterol
 - Blood Pressure
 - Heart Rate
 - Other relevant features
3. Predict whether the patient is likely to have heart disease.
4. Display the prediction.

END