Adventist University of Central Africa
P.O. Box 2461 Kigali, Rwanda | www.auca.ac.rw | info@auca.ac.rw

Faculty of Information Technology

Department: IT

FINAL- EXAM CAPSTONE PROJECT Academic year 2024-2025, SEM III

Course Code and Name: INSY 8413| Introduction to Big Data Analytics

Assistant Lecture: Eric Maniraguha **Exam Duration**
Time: **Date:** Saturday, July 26, 2025
Group: A, B & E **Total max:** /40

Tools: Python (22 marks), Power BI (14 marks), Innovation (4 marks)

Instructions: *This exam is based on a project with two parts:* ***Python (Data Analysis)*** *and* ***Power BI (Visualization)****. Each student must explain and justify their individual work and answer related technical questions.*

- Each student must:
  - ✓ Choose a real-world **problem in one or more sectors** (e.g., Health, Agriculture, Education, etc.)
  - ✓ Find or use an existing **public dataset**
  - ✓ Conduct **data preprocessing, exploratory analysis, and basic modeling** in Python
  - ✓ Create **a Power BI dashboard** that explains findings clearly
  - ✓ Present your work in a **structured GitHub repository and PowerPoint file**
- *Failure to explain your work (code or dashboard) may lead to mark deductions.*

PART 1: PROBLEM DEFINITION & PLANNING

You are required to define and plan your capstone project using the prompts below:

I. Sector Selection

Indicate which sector(s) your project will focus on: - Fill the share link here
☐ Agriculture ☐ Health ☐ Education ☐ Environment ☐ Energy ☐ Government ☐ Retail ☐ Finance ☐ Transportation ☐ Other: _____

II. Problem Statement

Clearly define a specific real-world problem you aim to address using Big Data Analytics.
(Example: "Can we detect patterns of student dropouts based on socio-economic data?")

III. 3. Dataset Identification

Complete the following details about your selected dataset:

- **Dataset Title**:

- **Source Link**:
- **Number of Rows and Columns**:
- **Data Structure**:
  ☐ Structured (CSV, Excel) ☐ Unstructured (Text, Images)
- **Data Status**:
  ☐ Clean ☐ Requires Preprocessing

PART 2: PYTHON ANALYTICS TASKS

In this section, you are required to use Python (in Jupyter Notebook or similar) to:

1. **Clean the Dataset**
   - Handle missing values, inconsistent formats, and outliers
   - Apply necessary data transformations (e.g., encoding, scaling)
2. **Conduct Exploratory Data Analysis (EDA)**
   - Generate descriptive statistics
   - Visualize distributions and relationships among variables
3. **Apply a Machine Learning or Clustering Model**
   - Choose a suitable model (classification, regression, or clustering)
   - Train the model on your dataset
4. **Evaluate the Model**
   - Use appropriate evaluation metrics (accuracy, precision, RMSE, silhouette score, etc.)
5. **Structure Your Code Properly**
   - Use modular functions
   - Include markdown explanations and comments for clarity and reproducibility
6. **Incorporate Innovation**
   - Add any custom function, ensemble technique, or creative model approach that enhances your project

PART 3: POWER BI DASHBOARD TASKS

Using Power BI, design a dashboard to present your data insights:

1. **Communicate the Problem & Insights Clearly**
   - Include context and summaries drawn from your data analysis
2. **Incorporate Interactivity**
   - Use slicers, filters, and drill-down options for user interaction
3. **Use Appropriate Visuals**
   - Match chart types (bar, line, scatter, etc.) with your data goals
4. **Ensure Design Clarity**
   - Apply consistent color themes, clear labels, and tidy layouts
5. **Add Innovative Features**

- Include advanced features like DAX formulas, AI visuals, custom tooltips, bookmarks, or R visuals

PART 4: SUBMISSION & COMMUNICATION

Make sure your final project includes:

1. **GitHub Repository**
   - Well-structured folders
   - README file with overview, instructions, and screenshots
   - All code files and Power BI .pbix file
2. **PowerPoint Presentation**
   - Slide deck summarizing:
     - Project Introduction
     - Methodology
     - Results
     - Recommendations
     - Future Work

PART 5: COMPLEXITY & CREATIVITY (Optional)

You are encouraged (but not required) to:

- Use multiple or large-scale datasets
- Introduce a novel methodology or data science approach

TIMELINE CHECK

☐ Have you submitted all components by the deadline?

ADVANCED THINKING (Optional Prompt)

Can you suggest a machine learning or advanced data visualization technique that fits your problem? How would you implement it using Python or Power BI?

PART 6: ACADEMIC INTEGRITY & SUBMISSION GUIDELINES

All students must adhere to the following rules. This capstone project is an academic assessment that must reflect your **original work and understanding**.

Academic Integrity Is Mandatory

- You are required to **submit your project scope (title, sector, problem statement, and dataset) before grading day** to allow for verification and early review.

- **Plagiarism will not be tolerated.**
  Any project found with **significant similarity in implementation, structure, or code**—whether due to copying from another student, or online source—**will be considered plagiarized.**
- This plagiarism assessment may occur **before or after grading**.
  → If plagiarism is confirmed, involved students will receive **low marks or a score of zero**, depending on the severity of the violation.
- **→ Do not copy code from online platforms (e.g., GitHub, blogs, AI tools) without understanding or adapting it.**
  → Do not use someone else's work or allow your work to be used by others.
- Do not share your code, report, dashboard, or GitHub repository with anyone outside.

Responsible Use of AI Tools

- AI tools (e.g., ChatGPT, GitHub Copilot, etc.) may support your work, but:
  - Use them **ethically and responsibly**.
  - Ensure you fully understand any AI-generated code or content you include.
  - You must be able to **explain any part of your project** if asked.

Project Submission Instructions

- Submit your:
  - **GitHub repository link**
  - **PowerPoint presentation (PDF or PPTX)**
- Send your submission to: **eric.maniraguha@auca.ac.rw**

Deadline: Before the grading session begins. Late submissions may be penalized unless prior approval is given.

ADDITIONAL RESOURCE – DATASET SOURCES

As part of your capstone project, you are expected to find and use a reliable dataset relevant to your selected problem and sector. Below is a list of recommended open data platforms where you can search for datasets:

Dataset Sources

Website Name	Link	Types of Data Available
Kaggle	https://www.kaggle.com/datasets	Health, agriculture, education, environment, e-commerce, transport
Data.gov	https://data.gov	US government data across all sectors
WHO Data	https://www.who.int/data	Health statistics, disease tracking, global indicators
World Bank Open Data	https://data.worldbank.org	Global education, health, economy, development
Google Dataset Search	https://datasetsearch.research.google.com	General dataset search across domains
FAOStat	https://www.fao.org/faostat/	Agriculture, food, land use, environment
UNESCO Data	http://data.uis.unesco.org	Education, science, culture
Rwanda Data Portal (NISR)	https://www.statistics.gov.rw	Rwanda national statistics: population, economy, health, education
Rwanda Open Data	https://opendata.rw	Local government, agriculture, finance, education
Cybersecurity Datasets (CIC)	https://www.unb.ca/cic/datasets/	Network attacks, malware, DDoS, phishing, cybersecurity logs
UCI Machine Learning Repo	https://archive.ics.uci.edu/ml/index.php	Classic datasets (includes some cybersecurity and e-commerce)
Amazon Review Data	https://nijianmo.github.io/amazon/index.html	Product reviews and metadata for Amazon items
Uber Movement Data	https://movement.uber.com	Traffic and mobility data from Uber in major cities
AWS Open Data Registry	https://registry.opendata.aws	Public datasets including satellite, genomics, transportation, etc.
Microsoft Research Open Data	https://msropendata.com/	Cybersecurity, NLP, computer vision, healthcare

Etc.

Note: Choose a dataset that:

- Matches your problem statement and sector.
- Is large enough to allow for meaningful analysis.
- Is either structured (CSV, Excel, SQL) or unstructured (text, image) as required.
- Requires some preprocessing, so you can demonstrate data cleaning and preparation skills.

Final Encouragement

"Whatever you do, work at it with all your heart, as working for the Lord, not for human masters."
— Colossians 3:23 (NIV)

Dear Students,
As you work on your capstone project, remember that excellence is not just about grades—it's about integrity, growth, and purpose. Do your best, stay honest, and trust the process. You are capable of great things.
May your work be guided by wisdom, and may your efforts bear meaningful results.
Blessings and success,

Eric Maniraguha
Assistant Lecturer | Faculty of Information Technology @2025