**OFFICIAL CURRICULUM GUIDE**

## ELEVATOR SCHOOL OF TRADE AND TECHNOLOGY

# Data Science Programme

*From Foundations to Advanced Practice — Open CDS Aligned*

**8 Modules · 2 Videos per Module · Beginner to Advanced**

Open Certified Data Scientist (Open CDS) Certification Pathway Embedded

*Submitted to the Admission Board for Screening*

**May 2026**

*Confidential — For Internal Review Only*

# Table of Contents

## Programme Overview

The Elevator School of Trade and Technology Data Science Programme is a structured, outcome-driven curriculum designed to take students from foundational programming concepts to advanced, deployment-ready data science practice. The programme is built around eight carefully sequenced modules, each delivered through two professionally produced videos of 15 minutes each.

What makes this programme distinctive is that it is the only one of its kind in this region designed to prepare students for the Open Group Certified Data Scientist (Open CDS) certification — the most rigorous, experience-based, internationally recognized data science credential available globally. Rather than teaching data science as an academic exercise, this curriculum teaches it as a professional practice, equipping students with the technical skills, professional habits, communication ability, and real-world project experience required to earn a globally credible certification.

**Programme at a Glance**

| Category | Detail | Standard |
|---|---|---|
| Total Modules | 8 Modules | Beginner to Advanced |
| Video Content | 2 videos per module | 15 minutes each |
| Text Resources | 2 text resources file per module | One per section |
| Certification Alignment | Open Group Open CDS | Global Professional Standard |
| Mid-module assessment | After module 4 | Graded with rubric |
| End-of-module Quiz | >60% before accessing next module | Pass mark: 60% |
| Group Project | 1 capstone group project | Real-world application |
| Certificate Awarded | School certificate + Open CDS pathway | Upon passing all assessments |

## Certification Pathway

Each module embeds Open CDS Critical Evaluation Criteria — the exact competencies assessed during Open CDS Milestone Badge applications and the Peer Review Board. By the time a student completes this programme, they will have documented proof of competence, a professional portfolio, and experience of working on real data problems — the three pillars of a strong Open CDS application.

### Three Routes to Open CDS Certification After This Programme

**Associate Data Scientist Credential:** Available immediately after completing the Capstone module. Students who complete a strong final project can submit for this credential.

**Level 1 Certified Data Scientist:** Achievable within 6–18 months of graduation for students who enter professional employment and continue documenting their project experience.

**Level 2 Master Certified Data Scientist:** A medium-term career target (3–5 years) for graduates who progress into lead data science roles.

## Enrolment Tiers

The programme offers three enrolment tiers, each designed for a different type of learner. All tiers access the same eight modules, the same video content, and the same assignments. The difference lies in the level of guided support, interaction, and additional access each tier provides.

**TIER 1**
### Self-Paced Track
*Open access — learn at your own speed*

#### Standard Rate

- ✓ Full access to all 8 modules
- ✓ Full access to all video content (2 videos per module)
- ✓ All module quizzes and mid-module assessment and resources included
- ✓ School certificate upon successful completion
- ✓ Independent study — no direct instructor access

**TIER 2**
### Mentorship Track
*3-month structured learning journey*

#### Professional Rate

- ✓ Everything in the Self-Paced Track
- ✓ Direct access to ask the instructor questions via a dedicated channel
- ✓ Structured pacing: one module per week for deeper understanding
- ✓ Instructor feedback on all 8 module assessment, quizzes, and structured assignments
- ✓ Progress check-in at the end of each month
- ✓ Open CDS pathway guidance documents included
- ✓ Priority enrolment for the group project
- ✓ School certificate upon successful completion

**TIER 3**
### Elite Masterclass Track
*12-month immersive programme*

#### Elite Rate

- ✓ Everything in the Mentorship Track
- ✓ Monthly closed-door Elite Masterclass session
- ✓ Each masterclass features a seasoned industry professional
- ✓ Topics: industry trends, real-world data science applications, career insights
- ✓ Small cohort for an intimate, high-value experience
- ✓ Networking access with featured industry professionals
- ✓ Priority Open CDS application support and review
- ✓ School certificate + professional recommendation letter

MODULE 01

# Foundations of Data Science and Professional Practice

*Beginner Level*

*Duration*
**2 Weeks**
2 Videos · 15 min each

## Module Overview

Module 1 lays the intellectual and professional foundation for everything that follows. Before students write a single line of code, they need to understand what data science is, why it exists, and how it creates value in the real world. This module introduces Python as the language of data science, establishes the professional habits that will carry students through to certification, and frames every subsequent technical topic within a business and professional context.

## Learning Objectives

By the end of this module, students will be able to:

- Explain what data science is and articulate the difference between Artificial Intelligence, Machine Learning, and Deep Learning in plain language
- Write and execute basic Python programs using variables, data types, loops, conditionals, and functions
- Install and import the core data science libraries — NumPy, Pandas, Matplotlib, Seaborn, SciPy, Scikit-learn, TensorFlow, and PyTorch — and understand what each is for
- Frame a business problem as a data science question using a structured approach
- Define and begin maintaining a Professional Development Log as part of their Open CDS certification journey

VIDEO 1

**What is Data Science? The Professional Mindset and Python Basics**

- The data science value chain: from raw data to business decision
- Difference between Artificial Intelligence, Machine Learning, and Deep Learning — explained simply
- The CRISP-DM professional workflow: how real data science projects are structured
- Business Problem Framing: translating a business challenge into an analytical question
- Installing Python and setting up the development environment
- Python fundamentals: variables, data types, operators, conditionals, and loops
- The Professional Development Log: what it is and why it matters for Open CDS

**VIDEO 2**

**Python Essentials: Libraries, Data Structures, and Professional Identity**

- Installing essential libraries: NumPy, Pandas, Matplotlib, Seaborn, SciPy, Scikit-learn, TensorFlow, PyTorch
- NumPy arrays: creating, indexing, and performing operations
- Pandas DataFrames: creating, reading, and exploring tabular data
- Functions and modules: writing reusable, organized code
- Introduction to Jupyter Notebooks as a professional working environment
- Setting up a professional LinkedIn profile for a data science career

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

1. **Business Context and Problem Framing:** Candidates must demonstrate the ability to define a data science problem within a real business context. This module introduces the framing discipline required for every Experience Profile submission.
2. **Professional Identity and Community Engagement:** Open CDS requires evidence of professional community participation. The Professional Development Log and GitHub/LinkedIn setup started in this module begin building that evidence from day one.
3. **Data Science Lifecycle Awareness:** Understanding CRISP-DM and the professional data science workflow is foundational to writing credible Experience Profiles that demonstrate structured, professional practice.

## ASSIGNMENT — MODULE 01 | Business Problem Framing Exercise

*Students are given three real-world business scenarios (e.g. a retail company losing customers, a hospital with long wait times, a bank with rising loan defaults). For each scenario, they must produce a structured Project Brief that includes the business problem in plain language, the analytical question derived from it, the data that would be needed, and how success would be measured.*

**Deliverables:**

- ✔ Three completed Project Briefs (one per scenario), each 200–300 words
- ✔ A one-paragraph reflection on what they found challenging about thinking in business terms rather than technical terms
- ✔ First entry in their Professional Development Log documenting what they learned this week

MODULE 02

# Data Exploration, Statistics, and Storytelling

*Beginner Level*

*Duration*
**2 Weeks**
2 Videos · 15 min each

## Module Overview

Data without understanding is just noise. This module teaches students how to explore a dataset, understand what it is telling them, apply statistical thinking to draw valid conclusions, and then communicate those conclusions clearly to an audience that may not be technical. The integration of data storytelling into this module is deliberate — communication is not an afterthought; it is part of the analysis itself.

## Learning Objectives

By the end of this module, students will be able to:

- Conduct a full exploratory data analysis (EDA) on a provided dataset using Python
- Calculate and interpret descriptive statistics including mean, median, variance, standard deviation, and correlation
- Handle missing values and outliers using appropriate, defensible strategies
- Build clear, well-labelled visualizations using Matplotlib and Seaborn
- Apply hypothesis testing and interpret p-values in a business context
- Structure a data finding as a clear narrative using the SCR (Situation, Complication, Resolution) framework
- Access and evaluate public datasets from sources including Kaggle, UCI, and government repositories

VIDEO 1

### Exploring and Understanding Data: Statistics and EDA

- What is Exploratory Data Analysis and why it comes before modeling
- Descriptive statistics: mean, median, mode, variance, standard deviation — explained with real examples
- Correlation and covariance: measuring relationships between variables
- Hypothesis testing fundamentals: null hypothesis, alternative hypothesis, and p-values
- Handling missing values: drop, fill, impute — when to use which strategy
- Handling outliers: detection methods (IQR, z-score) and treatment options
- Data cleaning workflows: a step-by-step professional approach

**VIDEO 2**

### Visualization, Storytelling, and Data Sources

- Visualization with Matplotlib: line charts, bar charts, scatter plots, and histograms
- Advanced visualization with Seaborn: boxplots, heatmaps, pairplots, and violin plots
- Choosing the right chart for the right message
- Data storytelling: the SCR framework — Situation, Complication, Resolution
- Designing visualizations for non-technical audiences
- Types of datasets: structured vs unstructured, open vs proprietary
- Accessing public datasets: Kaggle, UCI Machine Learning Repository, World Bank, and government data portals
- Mini Project: Statistical insights from a consumer dataset

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

4. **Data Exploration and Investigation:** A core requirement of the Associate Data Scientist credential and every Experience Profile. Evaluators look for evidence that candidates explore data systematically before building models, understanding distributions, missing data, and relationships.
5. **Professional Communication — Visual Storytelling:** The Professional Communication Milestone Badge requires candidates to demonstrate effective communication of data findings. Visualizations designed for non-technical audiences are a primary form of this evidence.
6. **Data Quality Awareness:** Open CDS evaluators assess whether candidates identify and handle data quality issues appropriately. Students who can articulate why they made specific cleaning decisions demonstrate the professional judgment the board is looking for.

## ASSIGNMENT — MODULE 02 | EDA and Storytelling with Data

*Students select a dataset from Kaggle or UCI and conduct a complete exploratory data analysis. They then present their findings as a business-facing data story — not a technical report, but a narrative a manager could read and act on.*

**Deliverables:**

- ✔ A Jupyter Notebook containing the full EDA with annotated code and inline commentary
- ✔ A one-page Executive Summary written in business language summarizing the three most important findings
- ✔ Three visualizations specifically designed for a non-technical audience, each with a clear title and key message
- ✔ A Professional Development Log entry noting one data community activity completed this week (a post, a comment, or a forum contribution)

MODULE 03

# Machine Learning Foundations

*Intermediate Level*

*Duration*
**3 Weeks**
2 Videos · 15 min each

## Module Overview

Module 3 is where students encounter machine learning for the first time, moving from describing data to making predictions from it. The module builds a solid conceptual foundation before introducing the mechanics, ensuring students understand not just how to run an algorithm but when to choose it, why it works, and how to evaluate whether it has actually done its job. It covers supervised learning fundamentals including regression and classification, data preprocessing, and the evaluation metrics every practitioner must master.

## Learning Objectives

By the end of this module, students will be able to:

- Explain the three types of machine learning — supervised, unsupervised, and reinforcement — and give a practical example of each
- Build and evaluate a linear regression model and interpret its output in business terms
- Build and evaluate a logistic regression model for a binary classification problem
- Split data into training and test sets correctly and understand why this separation matters
- Calculate and interpret key evaluation metrics: MAE, RMSE, Accuracy, Precision, Recall, F1-Score, and the Confusion Matrix
- Build a KNN classifier and tune its k parameter
- Handle imbalanced class distributions using SMOTE and class weighting strategies
- Standardize, normalize, and scale data appropriately as a preprocessing step

VIDEO 1

### Introduction to Machine Learning: Types, Regression, and Evaluation

- Types of machine learning: supervised, unsupervised, and reinforcement learning — with practical examples
- The machine learning workflow: data → features → model → evaluation → deployment
- Linear regression: building a model to predict a continuous outcome
- Logistic regression: building a model to classify a binary outcome
- The train/test split: why we never evaluate a model on the data it learned from
- Evaluation metrics for regression: MAE, MSE, RMSE — what they measure and when each matters

- Evaluation metrics for classification: Accuracy, Confusion Matrix, Precision, Recall, F1-Score
- Mini Project: Predicting House Prices or Loan Default

**VIDEO 2**

**Classification, Imbalanced Data, and Preprocessing**

- KNN Classifier: the intuition, the algorithm, and choosing k
- ROC Curve and AUC: measuring classifier performance across thresholds
- The Precision-Recall tradeoff: when false positives and false negatives have different costs
- Imbalanced classes: why 95% accuracy can still be a terrible model
- Handling imbalanced data: SMOTE, undersampling, and class weight adjustment
- Data standardization: when and why to standardize (StandardScaler)
- Normalization: min-max scaling and when to use it
- Activity: Build an ML system from scratch — Fraud Detection or Medical Diagnosis

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

7. **Algorithm Selection and Justification:** Open CDS Experience Profiles require candidates to explain why they chose a specific algorithm for a specific problem. This module builds the conceptual foundation for making and articulating those choices under board review.
8. **Model Testing and Validation:** A specific requirement of the Associate Data Scientist credential. Evaluators assess whether students can evaluate model performance rigorously and interpret metrics in the context of the business problem.
9. **Data Processing and Readiness:** Preprocessing decisions how and why data was scaled, how imbalanced classes were handled must be documented in Experience Profiles. This module teaches students to make those decisions consciously and defensibly.

## ASSIGNMENT — MODULE 03 | Build and Evaluate a Classification Model

*Students build a complete classification pipeline on a real dataset of their choice. The submission must demonstrate not just working code but genuine analytical thinking: why they chose their model, how they evaluated it, and what the results mean for a real decision-maker.*

**Deliverables:**

- ✔ A Jupyter Notebook with a complete classification pipeline: data loading, EDA, preprocessing, model training, and evaluation
- ✔ A written interpretation of all evaluation metrics in business language (not just numbers — what do they mean for the problem?)
- ✔ A comparison of at least two models with a clear justification for why one is preferred
- ✔ A 150-word reflection on what the most significant decision in the project was and why

MODULE 04

# Advanced Machine Learning and Feature Engineering

*Intermediate Level*

*Duration*
**3 Weeks**
2 Videos · 15 min each

## Module Overview

Module 4 takes students beyond single models into the richer territory of ensemble methods, feature engineering, and the art of building production-quality machine learning pipelines. Students learn to extract more signal from their data, build models that generalize well to unseen data, and structure their workflows in the systematic, reproducible way that professional and Open CDS-aligned practice demands.

## Learning Objectives

By the end of this module, students will be able to:

- Build and interpret Decision Tree models and explain how they make predictions in plain language
- Build Random Forest models and explain how ensemble methods improve on individual models
- Identify the most important features in a dataset and articulate why feature importance matters
- Apply cross-validation correctly and explain why a held-out test set alone is insufficient
- Recognise overfitting and apply practical strategies to prevent it
- Engineer new features from existing data to improve model performance
- Build a reusable Scikit-Learn pipeline that combines preprocessing and modeling steps
- Tune model hyperparameters using GridSearchCV and RandomizedSearchCV
- Explain the bias-variance tradeoff and its implications for model selection

VIDEO 1

### Decision Trees, Ensemble Learning, and Feature Engineering

- Decision Trees: how they work, how to build them, and how to read them
- Overfitting in decision trees and how to control it: max depth, min samples
- Random Forests: the ensemble intuition — why many imperfect trees beat one perfect one
- Feature importance: measuring and visualizing which variables drive predictions
- Cross-validation: k-fold CV, stratified CV, and when to use each

- Feature engineering: creating new variables that give the model more signal
- Feature selection: removing noise and redundancy from your input space
- Mini Project: Customer Churn Prediction

**VIDEO 2**

**Pipelines, Hyperparameter Tuning, and Data from the Web**

- Scikit-Learn Pipelines: chaining preprocessing and modeling into a reproducible workflow
- GridSearchCV: exhaustive hyperparameter search — when to use it
- RandomizedSearchCV: efficient hyperparameter search for large parameter spaces
- The Bias-Variance Tradeoff: understanding underfitting vs overfitting visually and practically
- Importing data from the internet: flat files, non-flat files, HTTP requests
- Web scraping with Beautiful Soup: parsing HTML to extract data
- Interacting with APIs: JSON responses and API authentication
- Hands-on practice: streaming and analysing data from a real API

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

10. **Methodology Depth and Justification:** Ensemble methods and hyperparameter tuning represent the kind of modeling depth Open CDS evaluators look for in Level 1 Experience Profiles. Candidates who can explain why they used a Random Forest over a single Decision Tree, and how they tuned it, demonstrate genuine professional practice.
11. **Reproducible and Structured Workflows:** Pipeline construction maps directly to the professional practice standard assessed by Open CDS. A candidate who builds reproducible, documented pipelines demonstrates the discipline expected of a certified professional.
12. **Critical Thinking About Model Limitations:** The Bias-Variance discussion prepares students for the kind of reflective reasoning the Peer Review Board assesses not just 'what did you build' but 'what are the risks and limitations of your approach?'

## ASSIGNMENT — MODULE 04 | End-to-End Machine Learning Pipeline

*Students build a complete, production-style machine learning pipeline using Scikit-Learn, incorporating preprocessing, feature engineering, model selection, cross-validation, and hyperparameter tuning. The project must include a documented decision log explaining every significant choice made.*

**Deliverables:**

- ✔ A complete Scikit-Learn pipeline notebook with inline explanations for every major step
- ✔ A hyperparameter tuning report showing the parameter grid explored and the winning configuration
- ✔ A Project Decision Log documenting at least five key decisions made during the project and the reasoning behind each
- ✔ A short business brief (250 words) describing the project as if presenting to a non-technical manager

MODULE 05

# Unsupervised Learning, Time Series, and SQL

*Intermediate–Advanced Level*

*Duration*
**3 Weeks**
2 Videos · 15 min each

## Module Overview

Module 5 expands the student's analytical toolkit into territory that does not require labeled data and into the dimension of time. Unsupervised learning opens up an entirely different class of business problems — customer segmentation, anomaly detection, pattern discovery — while time series analysis addresses the reality that much of the world's most valuable data is collected sequentially. SQL rounds out this module by giving students the data retrieval skills every professional data scientist needs.

## Learning Objectives

By the end of this module, students will be able to:

- Apply K-Means clustering to segment a dataset and interpret the resulting clusters in business terms
- Use the Elbow Method and Silhouette Score to select the appropriate number of clusters
- Apply hierarchical clustering and interpret a dendrogram
- Apply PCA to reduce a high-dimensional dataset and explain the tradeoff involved
- Decompose a time series into its trend, seasonal, and residual components
- Engineer lag features and rolling statistics for time series modeling
- Build an ARIMA model for time series forecasting and evaluate its performance
- Apply machine learning (Random Forest, XGBoost) and deep learning (LSTM) to time series problems
- Write SQL queries of intermediate complexity and integrate SQL with Python using pandas

VIDEO 1

**Unsupervised Learning, Clustering, and Dimensionality Reduction**

- What is unsupervised learning and when do you use it?
- K-Means clustering: the algorithm, the intuition, and the business applications
- Choosing k: the Elbow Method and Silhouette Score explained
- Hierarchical clustering: agglomerative approach and dendrograms
- PCA (Principal Component Analysis): reducing dimensions while preserving variance
- Interpreting PCA components in business language

- Practice: Market Segmentation using real customer data
- Time Series fundamentals: components — Trend, Seasonality, and Residuals

**VIDEO 2**

**Time Series Modeling and SQL for Data Science**

- Lag features and rolling statistics: engineering time-aware variables
- ARIMA: understanding and applying autoregressive models for forecasting
- Machine learning for time series: Random Forest and XGBoost with lag features
- Deep learning for time series: LSTM and RNN architecture explained simply
- Sequence modeling and multi-step forecasting
- SQL Basics: SELECT, WHERE, ORDER BY, LIMIT — reading data from a database
- SQL Intermediate: INNER JOIN, LEFT JOIN, GROUP BY, HAVING, subqueries
- Integrating SQL with Python: using pandas.read_sql to query directly into a DataFrame

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

13. **Breadth of Analytical Technique:** Open CDS evaluators assess whether a candidate has exposure to a range of techniques and can select appropriately. Unsupervised learning, time series, and SQL represent three distinct competency areas that significantly strengthen an Experience Profile.
14. **Business Application of Clustering:** The Market Segmentation project directly mirrors the kind of business-relevant application Open CDS Experience Profiles require. Students practice framing the cluster results as a business recommendation, not just a technical output.
15. **Data Retrieval Competence:** SQL proficiency is assessed as a data engineering competency within Open CDS. A candidate who cannot retrieve and manipulate data from a database is considered to have a gap in their foundational practice.

## ASSIGNMENT — MODULE 05 | Segmentation Analysis and Time Series Forecast

*A two-part assignment. Part A requires students to perform a customer segmentation analysis using K-Means on a provided e-commerce dataset, producing a business-facing segment profile. Part B requires building a time series forecast for a publicly available dataset (e.g. energy consumption, stock price, weather data), using at least two different modeling approaches and comparing their performance.*

**Deliverables:**

- ✔ Part A: Clustering notebook with segment profiles written as business personas (who is this customer?)
- ✔ Part B: Time series notebook with ARIMA and at least one ML-based approach, including a forecast visualization
- ✔ A SQL query script demonstrating retrieval, joining, grouping, and filtering on a provided database
- ✔ A Professional Development Log entry and evidence of one community activity this month

MODULE 06

# Data Engineering, Version Control, and Model Deployment

*Advanced Level*

*Duration*
**2 Weeks**
2 Videos · 15 min each

## Module Overview

A data scientist who can build a model but cannot deploy it, version it, or retrieve the data it needs is only half-equipped for professional practice. Module 6 completes the production picture — teaching students how data science work is tracked, shared, and delivered in real organizations. These are the skills that bridge the gap between a notebook experiment and a working solution that a business can actually use.

## Learning Objectives

By the end of this module, students will be able to:

- Use Git for version control: initializing repositories, staging, committing, branching, and merging
- Setting up GitHub: creating a repository, writing a README, and making your first commit
- Collaborate on a shared GitHub repository and manage contributions professionally
- Identify and access appropriate data sources from open repositories and government portals
- Serialize a trained model using Pickle and Joblib for reuse and deployment
- Build a functional data visualization dashboard using Streamlit
- Build a simple model-serving API using Flask
- Deploy a trained model locally and explain the steps required for production deployment

VIDEO 1

### Data Sources, Version Control, and Git Workflows

- Types of data sources: open, private, crowdsourced, and institutional
- Public datasets: UCI, Kaggle, data.gov, World Bank Open Data — how to find and evaluate them
- Government and institutional data in Africa: relevant open data portals
- Introduction to version control: why data scientists need Git
- Git fundamentals: init, add, commit, status, log — the daily workflow
- Branching and merging: working on features without breaking the main codebase
- Collaborating on GitHub: pull requests, code review, and team workflows
- Structuring a professional data science repository: folder layout, README, and documentation standards

**VIDEO 2**

### Model Serialization and Deployment

- Why deployment matters: the gap between a notebook and a live application
- Model serialization: saving and loading models with Pickle and Joblib
- Introduction to Streamlit: building an interactive dashboard without frontend code
- Building a Streamlit app: connecting a trained model to a user interface
- Introduction to Flask: building a lightweight REST API to serve model predictions
- Deploying a model locally: end-to-end walkthrough from training to live inference
- Activity: Deploy a trained model as a working Streamlit application
- Introduction to cloud deployment concepts: what happens after local deployment (Heroku, Render, HuggingFace Spaces)

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

16. **Data Engineering and Production Readiness:** Open CDS explicitly assesses data engineering competence. The ability to source data, version code, and deploy a solution represents the full professional lifecycle — not just the modeling step — which is what certified practitioners are expected to demonstrate.
17. **Documentation and Reproducibility:** Version control is the professional standard for reproducible data science. Open CDS evaluators will probe whether a candidate's work is traceable, version-controlled, and documented in a way that a colleague could pick up and continue.
18. **Professional Portfolio Building:** The GitHub repository structure introduced in this module is the foundation of the student's professional portfolio. Every project completed from this point should follow this structure, building a body of visible, verifiable professional work.

## ASSIGNMENT — MODULE 06 | Deploy a Data Science Application

*Students take a model they have already built in a previous module and deploy it as a working Streamlit application. The application must be user-friendly, clearly documented, and version-controlled on a public GitHub repository with a professional README.*

**Deliverables:**

- ✔ A deployed Streamlit application that accepts user input and returns a model prediction
- ✔ The complete project on GitHub with a professional README (including project description, how to run it, and key findings)
- ✔ A version-controlled commit history showing at least five meaningful commits (not just one 'final' commit)
- ✔ A 200-word reflection on what a real user or client would need from this application that it does not yet have — and how you would address that

MODULE 07

# Ethics, Explainability, and Responsible AI Practice

*Advanced Level*

*Duration*
**2 Weeks**
2 Videos · 15 min each

## Module Overview

Data science does not happen in a vacuum. The models and systems that data scientists build affect real people — their access to credit, their medical diagnoses, their job applications, their interactions with government services. Module 7 confronts this reality directly, equipping students with the technical tools to interpret and explain their models and the professional ethics framework to practice responsibly. This is one of the most distinctive modules in the curriculum and one of the most directly relevant to Open CDS certification.

## Learning Objectives

By the end of this module, students will be able to:

- Apply SHAP and LIME to generate interpretable explanations for complex model predictions
- Visualize feature importance and explain model behavior to both technical and non-technical audiences
- Identify at least five distinct sources of bias that can enter a machine learning system
- Apply fairness metrics using the Fairlearn library to audit a model for differential impact
- Explain the key provisions of the Nigeria Data Protection Act (NDPA) and GDPR and their implications for data science practice
- Write a professional Ethics Impact Assessment for a data science project
- Articulate the data scientist's professional ethical obligations in a workplace context

VIDEO 1

**Model Explainability: SHAP, LIME, and Interpretable AI**

- Why interpretability matters: when a 95% accurate model is not good enough
- The difference between global and local model interpretability
- Feature importance visualization: which variables drive the model overall?
- SHAP (SHapley Additive exPlanations): the theory explained simply, with hands-on practice
- LIME (Local Interpretable Model-agnostic Explanations): explaining individual predictions
- Communicating model explanations to a non-technical stakeholder
- When black-box models are acceptable and when they are not

- Regulatory context: why explainability is increasingly a legal requirement

VIDEO 2

**Advanced Ethics, Fairness, and Professional Responsibility**

- Data science as a profession with responsibilities: the analogy to medicine and law
- Types of bias: historical, representation, measurement, aggregation, and deployment bias
- Real-world case studies: algorithms that caused measurable harm
- Fairness metrics: demographic parity, equalized odds, calibration — and their inherent tradeoffs
- Hands-on: auditing a model for bias using the Fairlearn Python library
- Privacy and consent: NDPA, GDPR, and what they mean for how you collect and use data
- Writing an Ethics Impact Assessment: structure, language, and what to include
- How to raise ethical concerns professionally in an organization

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

19. **Ethical Practice as a Standalone Competency:** Open CDS treats ethical practice as a first-class professional requirement, not a footnote. The Peer Review Board will ask candidates directly about their ethical responsibilities and how they handled ethical considerations in their documented projects.
20. **Model Interpretability and Stakeholder Trust:** The ability to explain a model's predictions — not just report its accuracy — is assessed in Experience Profiles. Evaluators look for evidence that candidates considered how their outputs would be used, by whom, and with what potential for harm.
21. **Regulatory and Legal Awareness:** Professional certification requires awareness of the legal and regulatory environment in which data science practice operates. Knowledge of NDPA and GDPR signals the kind of responsible professionalism the Open CDS standard upholds.

## ASSIGNMENT — MODULE 07 | Explainability and Ethics Impact Assessment

*A two-part assignment. Part A applies SHAP and LIME to a model the student has already built, producing interpretable explanations and a visualization suite suitable for a non-technical stakeholder. Part B requires the student to write a full Ethics Impact Assessment for the same project.*

**Deliverables:**

- ✔ Part A: A notebook with SHAP and LIME explanations including a summary visualization and a written paragraph interpreting the results in business language
- ✔ Part B: A two-page Ethics Impact Assessment covering: potential harms, affected populations, fairness risks, privacy considerations, and proposed mitigation steps
- ✔ A 150-word reflection: 'What would you do differently in this project if you were required to explain every prediction to the person it affected?'

**MODULE 08**

# Capstone, Real Client Experience, and Certification Readiness

*Advanced Level*

*Duration*
**4 Weeks**
2 Videos · 15 min each

## Module Overview

Module 8 is the culmination of the entire programme. It brings together every technical skill, professional habit, and analytical discipline developed across the previous seven modules and applies them in a real-world, client-facing context. Students complete a live project for a genuine external client, build their complete professional portfolio, and prepare directly for Open CDS certification. This is the module that transforms a student into a practitioner.

## Learning Objectives

By the end of this module, students will be able to:

- Manage a complete, end-to-end data science engagement from client brief to final delivery
- Navigate the realities of working with genuinely messy, incomplete, real-world data
- Communicate progress and findings to a real client audience throughout a project
- Write a compelling Open CDS Experience Profile narrative for a completed project
- Compile a complete Professional Development Portfolio meeting the standard required for badge applications
- Articulate their technical decisions, ethical considerations, and professional growth in a peer-review-style discussion
- Submit a final capstone project that qualifies as evidence for the Associate Data Scientist credential

**VIDEO 1**

### Real Client Engagement, Reflective Practice, and Portfolio Building

- The professional client engagement lifecycle: from briefing to delivery
- What real-world data actually looks like — and how to manage the difference from textbook datasets
- The Project Decision Log: documenting choices, alternatives, and reasoning as you work
- Writing a compelling Open CDS Experience Profile: structure, language, and what evaluators look for
- Building your Professional Portfolio: GitHub, LinkedIn, and the written Experience Profile

- The Professional Development Portfolio: consolidating your community contributions into a badge-ready submission
- Reflective writing for professionals: turning experience into evidence

VIDEO 2

### Open CDS Certification Preparation and Final Capstone Briefing

- The Open CDS certification pathway: Associate credential, Level 1, Level 2, and Level 3
- Understanding Milestone Badges: Professional Communication, Professional Development, and Experience Profile
- The Peer Review Board: what it is, what to expect, and how to prepare
- Mock board review: common questions, how to answer them, and how to handle challenges
- The Associate Data Scientist credential: what your capstone project needs to qualify
- Capstone project briefing: requirements, client context, and submission format
- Career pathways after this programme: roles, salary benchmarks, and the road to Level 1 certification
- Final programme reflection: what you have built, what you know, and where you go next

## ★ OPEN CDS CRITICAL EVALUATION CRITERIA

*Competencies assessed toward Open Certified Data Scientist certification*

22. **Experience Profile — Full Cycle:** The capstone project is explicitly designed to produce a complete, submission-ready Open CDS Experience Profile. Students who complete this module with a strong project and a well-written profile can submit for the Associate Data Scientist credential immediately upon graduation.
23. **Peer Review Board Readiness:** The mock board exercise in this module directly simulates the Open CDS Peer Review Board experience. Students practice articulating their work under professional scrutiny — the exact competency the board is designed to assess.
24. **Professional Development Milestone Badge:** The Professional Portfolio compiled in this module — including the Development Log, community contributions, LinkedIn profile, and GitHub portfolio — constitutes the primary evidence for a Professional Development Milestone Badge application.
25. **Professional Communication Milestone Badge:** The capstone presentation to the real client, and the written deliverables produced for them, represent the kind of evidence required for a Professional Communication Milestone Badge application.

## ASSIGNMENT — MODULE 08 | Final Capstone Project

*The capstone project is a complete, client-facing data science engagement conducted over four weeks. Students work in pairs or small teams on a real business problem provided by a genuine client (arranged by the school). The final submission is designed to serve simultaneously as the student's Open CDS Associate credential application.*

**Deliverables:**

- ✔ A complete Jupyter Notebook repository on GitHub: data pipeline, EDA, modeling, evaluation, and deployment
- ✔ A formal Client Report (maximum 8 pages) presenting findings and recommendations in business language
- ✔ A live demonstration or recorded presentation of the solution to the client
- ✔ A completed Open CDS Experience Profile document (using The Open Group's official template)
- ✔ A completed Professional Development Portfolio: Development Log, community contribution evidence, LinkedIn profile, and GitHub portfolio
- ✔ An individual reflection (400 words): what was the hardest decision you made in this project, and why did you make it?

# Final Examination

The final examination is taken after all eight modules and their assignments have been completed. It is a comprehensive assessment that tests theoretical understanding, applied reasoning, and professional judgment across the full scope of the programme.

**Examination Structure**

| Section | Format | Marks |
|---|---|---|
| Section A: Conceptual Foundations | 25 multiple-choice questions covering all 8 modules | 25 marks |
| Section B: Analytical Reasoning | 5 short-answer questions (100–150 words each) on methodology and decision-making | 25 marks |
| Section C: Case Analysis | One extended case study: students analyse a provided dataset scenario and recommend an approach | 30 marks |
| Section D: Ethics and Professional Practice | Two scenario-based questions on ethical decision-making and professional conduct | 20 marks |
| TOTAL | | 100 marks |

**Passing Score:** 60% (60/100). Students who score below 60% may retake the exam once after a mandatory 2-week review period.

**Duration:** 3 hours. The exam is open-note for Sections A and B but requires individual, unaided responses for Sections C and D.

# Group Project

The Group Project runs in parallel with the final weeks of Module 8 and is completed as a team exercise. It is distinct from the individual Capstone Project and is designed to assess collaboration, communication, role specialization, and the ability to deliver a coherent team output — all of which are assessed in the Open CDS Peer Review Board.

## Project Brief

Teams of 3–4 students are assigned a real or realistic data science problem relevant to an African industry context — examples include agricultural yield prediction, healthcare facility utilization, fintech fraud detection, or mobile network churn. Teams are responsible for the complete project lifecycle from brief to delivery.

## Team Roles

Each team member takes on a defined role that they are responsible for. Roles rotate on subsequent assignments to ensure balanced exposure:

- Data Engineer — responsible for data sourcing, cleaning, and pipeline construction
- Modelling Lead — responsible for algorithm selection, training, and evaluation
- Communication Lead — responsible for the final report, visualizations, and presentation
- Ethics and Quality Lead — responsible for the Ethics Impact Assessment, fairness audit, and project documentation

## Deliverables

- A shared GitHub repository with a clear commit history showing all team members' contributions
- A 10-minute team presentation delivered to a panel (instructor and at least one invited guest)
- A formal written report (maximum 10 pages) addressing the brief, methodology, findings, and recommendations
- An Ethics Impact Assessment for the project
- An individual peer evaluation (each team member rates their teammates' contributions)

# Scoring Guide and Certificate Award Criteria

The following scoring guide governs the assessment of all components across the programme. It is designed to be transparent, consistent, and directly linked to the competencies required for Open CDS certification.

## Final Examination — Scoring Guide

| Score | Grade and Interpretation |
|---|---|
| 90–100 | Distinction — Exceptional understanding across all domains. Demonstrates the depth of knowledge expected of a certified professional. |
| 75–89 | Merit — Strong performance with clear mastery of core concepts and solid professional reasoning. |
| 60–74 | Pass — Satisfactory understanding of programme content. Eligible for school certificate. May benefit from further study before Open CDS application. |
| Below 60 | Retest Required — Students must complete a 2-week review and retake. A score below 60 on the retake requires a structured remediation plan. |

## Group Project — Scoring Rubric (100 marks)

| Criterion | Marks Available | Description |
|---|---|---|
| Project Delivery and Technical Quality | 30 marks | The solution addresses the brief, the technical work is sound, and the pipeline is reproducible |
| Presentation and Communication | 25 marks | The team presents clearly, handles questions confidently, and communicates findings in business language |
| Collaboration and Role Execution | 20 marks | Each team member demonstrates ownership of their role; the work reflects genuine collaboration |
| Ethics and Documentation | 15 marks | The Ethics Impact Assessment is thorough; the project is well-documented and version-controlled |
| Peer Evaluation | 10 marks | Scores from team members' peer evaluations, assessing individual contribution to the team |

## Certificate Award Criteria

### Requirements for School Certificate

**Completion Requirement:** All 8 module assessments and quizzes passed (minimum 60% per assessment).

**Final Exam:** Minimum passing score of 60%.

**Group Project:** Participated in and contributed to the group project (minimum 50% score).

**Professional Portfolio:** Submitted a complete Professional Development Portfolio including a Development Log with at least six monthly entries.

**Overall Programme Score:** Calculated as: Module Assignments (50%) + Final Exam (30%) + Group Project (20%).

| Overall Programme Score | Certificate Level | Open CDS Pathway Status |
|---|---|---|
| 85–100% | Distinction Certificate | Strongly recommended to submit Associate Data Scientist credential immediately |
| 70–84% | Merit Certificate | Recommended to submit Associate Data Scientist credential with instructor review |
| 60–69% | Pass Certificate | Eligible for Associate credential after strengthening capstone documentation |
| 50–59% | Completion Certificate | Programme completed. Additional project experience recommended before certification application |
| Below 50% | No Certificate (Incomplete) | Structured remediation plan required; specific modules to be retaken |

# Appendix: Curriculum Map — Open CDS Competency Coverage

The table below shows how each module maps to the Open Group CDS Conformance Requirements, the Milestone Badge it supports, and the Associate Data Scientist credential criteria it addresses.

| Module | Open CDS Competency Area | Milestone Badge Supported |
|---|---|---|
| Module 1: Foundations and Professional Practice | Business Problem Framing; Professional Identity; CRISP-DM Lifecycle | Experience Profile (Problem Definition) + Professional Development (log initiated) |
| Module 2: Data Exploration and Storytelling | Data Exploration; Professional Communication; Data Quality | Professional Communication Badge + Associate Credential (Data Exploration) |
| Module 3: Machine Learning Foundations | Algorithm Selection; Model Evaluation; Data Preprocessing | Experience Profile (Methodology Depth) + Associate Credential (Testing & Validation) |
| Module 4: Advanced ML and Feature Engineering | Reproducible Workflows; Modeling Depth; Critical Reflection | Experience Profile (Technical Rigor) + Professional Communication |
| Module 5: Unsupervised Learning, Time Series, SQL | Breadth of Technique; Data Engineering; Business Application | Experience Profile (Advanced Techniques) + Associate Credential (Data Processing) |
| Module 6: Deployment and Version Control | Data Engineering; Professional Tools; Reproducibility | Experience Profile (Production Readiness) + Professional Development (GitHub portfolio) |
| Module 7: Ethics and Explainability | Ethical Practice; Interpretability; Legal Awareness | All Experience Profiles (Ethics section) + Professional Communication |
| Module 8: Capstone and Certification Readiness | Full Certification Cycle; Client Engagement; Portfolio | All Badges + Associate Data Scientist Credential |