DATA SCIENCE TRAINING CURRICULUM

A Structured 4-Month Program
From Foundations to Independent, Deployable Data Science Projects

16 WEEKS • 4 CORE MODULES • 1 CAPSTONE PROJECT

Prepared for Training & Upskilling Programs

Table of Contents

01 Program Overview & Learning Outcomes

02 Prerequisites & Program Structure

03 Month 1 — Foundations: Python, Statistics & Data Handling

04 Month 2 — Core Machine Learning

05 Month 3 — Advanced ML, Deep Learning & Big Data

06 Month 4 — MLOps, Specialization & Capstone Project

07 Assessment Framework & Grading

08 Tools & Technology Stack

09 Capstone Project Guidelines

10 Trainer Notes & Delivery Tips

01. Program Overview & Learning Outcomes

This curriculum is designed to take a learner from foundational programming and statistics through to independently building, evaluating, and deploying machine learning models. The program runs for **4 months (16 weeks)**, structured into four progressive modules, each ending with a graded project. The final month culminates in an end-to-end capstone project that mirrors real industry work.

By the end of this program, participants will be able to:

- Write production-quality Python code for data analysis and manipulation using NumPy and Pandas.
- Apply statistical reasoning and hypothesis testing to real-world datasets.
- Build, tune, and validate supervised and unsupervised machine learning models.
- Design and train neural networks for structured, text, and image data.
- Work with SQL databases and big-data tools for large-scale data processing.
- Deploy a machine learning model as an API and monitor it in production (MLOps basics).
- Communicate insights clearly through visualization, storytelling, and reporting.

02. Prerequisites & Program Structure

Prerequisites: Basic computer literacy and logical/mathematical aptitude. No prior programming experience is required, though familiarity with high-school level algebra and statistics is helpful.

Component	Details
Duration	4 months / 16 weeks
Weekly Commitment	10–12 hours (4 hrs live instruction + 6–8 hrs practice/assignments)
Format	Instructor-led sessions, hands-on labs, weekly assignments, and milestone projects
Modules	4 modules (1 per month), each closing with a graded project
Capstone	End-to-end project in Month 4, presented to a review panel
Certification	Awarded on completion of all 4 module projects + capstone with a passing score

MONTH 1 | Foundations — Python, Statistics & Data Handling

Weeks 1–4

Goal: Build strong programming fundamentals, statistical intuition, and comfort with data manipulation and visualization before any modeling begins.

Week	Focus Area	Key Topics	Tools	Deliverable
Week 1	Python Programming Basics	Variables, data types, control flow, functions, OOP basics, virtual environments, Jupyter/Colab setup, Git & GitHub basics	Python, Jupyter, Git	Set of 10 Python programming exercises + GitHub repo setup
Week 2	Data Handling with NumPy & Pandas	Arrays and vectorized operations; Series & DataFrames; filtering, grouping, merging, reshaping; handling missing data	NumPy, Pandas	Cleaned dataset + data-wrangling notebook
Week 3	Statistics & Probability	Descriptive statistics, probability distributions, sampling, central limit theorem, hypothesis testing, confidence intervals, correlation vs. causation	Python, SciPy	Statistical analysis report on a sample dataset
Week 4	Exploratory Data Analysis (EDA) & SQL	Data visualization (Matplotlib/Seaborn), outlier detection, univariate/bivariate analysis, SQL joins, aggregations, window functions	Matplotlib, Seaborn, SQL	Module 1 Project: Full EDA report with SQL data extraction

Month 1 Milestone Project

- Select a real-world dataset (e.g., retail sales, healthcare, or public government data).
- Clean and preprocess the data, extract subsets using SQL queries.
- Perform a complete exploratory data analysis with at least 8 visualizations.
- Present 3–5 key business insights supported by statistical evidence.

MONTH 2 | Core Machine Learning

Weeks 5–8

Goal: Understand the theory and practical implementation of core supervised and unsupervised machine learning algorithms, along with proper model evaluation techniques.

Week	Focus Area	Key Topics	Tools	Deliverable
Week 5	Supervised Learning I — Regression	Linear & polynomial regression, regularization (Ridge/Lasso), assumptions, gradient descent, bias-variance tradeoff	Scikit-learn	Regression model on a housing/price prediction dataset
Week 6	Supervised Learning II — Classification	Logistic regression, k-NN, decision trees, random forests, SVM, ensemble methods (bagging, boosting: XGBoost/LightGBM)	Scikit-learn, XGBoost	Classification model with performance comparison across 4 algorithms
Week 7	Model Evaluation & Feature Engineering	Train/test/validation split, cross-validation, confusion matrix, ROC-AUC, precision/recall, feature scaling, encoding, feature selection, hyperparameter tuning (GridSearch/Optuna)	Scikit-learn, Optuna	Tuned and validated model pipeline with documented metrics
Week 8	Unsupervised Learning & Dimensionality Reduction	K-Means, hierarchical clustering, DBSCAN, PCA, t-SNE, anomaly detection basics	Scikit-learn	Module 2 Project: Customer segmentation or fraud-pattern analysis

Month 2 Milestone Project

- Build a complete supervised learning pipeline: data prep, feature engineering, model training, tuning, evaluation.
- Compare at least 3 algorithms using cross-validation and justify the final model choice.
- Apply clustering or PCA to uncover hidden structure in a business dataset.
- Document all decisions and trade-offs in a short model report.

MONTH 3 | Advanced ML, Deep Learning & Big Data

Weeks 9–12

Goal: Move into advanced modeling territory — neural networks, deep learning for text/images/time series, and the tools needed to work with large-scale data.

Week	Focus Area	Key Topics	Tools	Deliverable
Week 9	Neural Networks & Deep Learning Foundations	Perceptrons, activation functions, backpropagation, loss functions, optimizers, building feed-forward networks	TensorFlow/Keras or PyTorch	Feed-forward neural network on a tabular classification task
Week 10	Computer Vision & CNNs	Convolution & pooling operations, CNN architectures, transfer learning, image augmentation	TensorFlow/Keras, OpenCV	Image classification model using transfer learning
Week 11	NLP & Sequence Models	Text preprocessing, embeddings (Word2Vec/GloVe), RNN/LSTM basics, intro to Transformers and pretrained LLM APIs	NLTK/spaCy, Hugging Face	Sentiment analysis or text classification model
Week 12	Time Series & Working with Big Data	Trend/seasonality decomposition, ARIMA/Prophet, intro to distributed computing (Spark basics), working with large datasets	Prophet/ARIMA, PySpark	Module 3 Project: Forecasting model + big-data processing notebook

Month 3 Milestone Project

- Train a deep learning model for either an image or text classification problem.
- Apply transfer learning to improve performance over a baseline model.
- Build a time-series forecasting model and evaluate it against a naive baseline.
- Demonstrate basic distributed data processing on a dataset too large for a single Pandas DataFrame.

MONTH 4 | MLOps, Specialization & Capstone Project

Weeks 13–16

Goal: Learn how models move into production, choose a specialization track, and apply everything learned to a single, portfolio-ready capstone project.

Week	Focus Area	Key Topics	Tools	Deliverable
Week 13	Model Deployment & MLOps Basics	REST APIs with Flask/FastAPI, containerization with Docker, model versioning, CI/CD basics, monitoring & drift detection	FastAPI, Docker, MLflow	Deployed model API with a simple front-end demo
Week 14	Specialization Track (choose one)	Track A: Advanced NLP/LLM applications • Track B: Advanced Computer Vision • Track C: Business Analytics & Forecasting • Track D: MLOps & Data Engineering	Track-specific tools	Track-specific mini-project
Week 15	Capstone Project — Build Phase	Problem definition, data collection/cleaning, modeling, evaluation, and initial deployment of an end-to-end project of the learner's choosing	Full stack from Months 1–4	Working capstone prototype + progress checkpoint review
Week 16	Capstone Project — Finalize & Present	Final testing, documentation, dashboard/report creation, presentation rehearsal, final review panel presentation	Streamlit/Power BI/Tableau	Final capstone submission: code, report, and live presentation

Specialization Tracks (Week 14)

- Track A — NLP & LLMs: prompt engineering, fine-tuning, retrieval-augmented generation (RAG).
- Track B — Computer Vision: object detection, image segmentation, real-time inference.
- Track C — Business Analytics: A/B testing, causal inference, dashboarding for stakeholders.
- Track D — MLOps & Data Engineering: pipelines, orchestration (Airflow), scalable data infrastructure.

07. Assessment Framework & Grading

Component	Weight	Description
Weekly Assignments	20%	Hands-on exercises submitted each week, auto- or peer-reviewed
Module Projects (x4)	40%	One graded project at the end of each month (10% each)
Participation & Labs	10%	Attendance, in-class exercises, peer code reviews
Capstone Project	30%	Final end-to-end project: code quality, results, documentation, and presentation

Passing criteria: Minimum 70% overall score, with at least a passing grade on the capstone project, is required for program certification.

08. Tools & Technology Stack

Category	Tools & Technologies
Languages	Python, SQL
Data Handling	NumPy, Pandas, Polars (optional)
Visualization	Matplotlib, Seaborn, Plotly, Power BI / Tableau
Machine Learning	Scikit-learn, XGBoost, LightGBM, Optuna
Deep Learning	TensorFlow / Keras or PyTorch, Hugging Face Transformers
Big Data	PySpark, basics of cloud data warehouses (BigQuery/Snowflake)
Deployment & MLOps	FastAPI/Flask, Docker, MLflow, GitHub Actions (CI/CD basics)
Collaboration	Git & GitHub, Jupyter/Colab, Notion/Confluence for documentation

09. Capstone Project Guidelines

The capstone project is completed in Weeks 15–16 and should demonstrate the full data science lifecycle applied to a problem of the learner's choosing.

Capstone Requirements

- Clearly defined business or research problem with a measurable objective.
- Data sourced, cleaned, and documented (own dataset, public dataset, or provided by the organization).
- At least one machine learning or deep learning model, with justified algorithm choice.
- Rigorous evaluation using appropriate metrics and a documented comparison of alternatives.
- A working deployment: an API, dashboard, or interactive app demonstrating the model in action.
- A final report (5–10 pages) and a 10–15 minute presentation to a review panel.

Suggested Capstone Domains

- Customer churn prediction and retention strategy for a subscription business.
- Demand forecasting and inventory optimization for retail or e-commerce.
- Fraud or anomaly detection for financial transactions.
- Medical/health risk prediction using structured or imaging data.
- Sentiment and trend analysis from social media or customer reviews.
- Recommendation system for products, content, or services.

10. Trainer Notes & Delivery Tips

For Facilitators

- Keep a 60/40 split between live instruction and hands-on lab time within each session.
- Use real, messy datasets from Week 2 onward — avoid over-cleaned toy datasets so learners build real intuition.
- Run a short recap quiz at the start of each week to reinforce retention from the previous week.
- Pair weaker and stronger learners for lab exercises to encourage peer learning.
- Bring in a guest practitioner (if possible) before the capstone phase to discuss real-world project experience.
- Maintain a shared GitHub organization so learner projects are versioned and visible throughout the program.

This curriculum can be adapted in pace or depth depending on the cohort's starting skill level. For cohorts with prior programming experience, Week 1 can be compressed to allow extra time for the specialization track in Month 4.