

# Data Science & Machine Learning

Analyse data, apply statistics and build, evaluate and explain machine-learning models responsibly with Python

Data Analytics and Data Science · intermediate · Live online, instructor-led

This syllabus is a draft awaiting instructor review and may change.

An 80-hour live course: Python and pandas, exploratory analysis and visualisation, practical statistics, feature engineering, machine-learning algorithms with scikit-learn, model evaluation, responsible AI and a reviewed capstone.

## Length and schedule

80 hours of live teaching, on either schedule:

- Weekday Intensive (Mon–Fri): about 16 weeks before holidays
- Weekend Evening (Sat & Sun evenings): about 20 weeks before holidays

## Fees per student, whole course

- Group: Rs. 22,999
- Duo: Rs. 37,999
- One-to-One: Rs. 69,999

Launch fee — GST treatment to be confirmed.

## What you will be able to do

- Write clean, reusable Python for data work with NumPy and pandas
- Explore and visualise data to find patterns, problems and questions worth modelling
- Apply descriptive and inferential statistics, including confidence intervals and hypothesis tests
- Engineer features and build leakage-safe preprocessing pipelines
- Train, tune and compare regression, classification and clustering models with scikit-learn
- Choose evaluation metrics that match the problem and validate models honestly
- Assess fairness, privacy and explainability, and document a model with a model card
- Deliver an end-to-end capstone project and explain its results and limitations to a non-technical audience

## Before you start

- Basic programming in any language: variables, conditions, loops and functions
- School-level algebra and comfort with percentages and graphs; the statistics needed is taught in the course
- Some experience with spreadsheets or SQL is helpful but not required
- A laptop with at least 8 GB RAM (16 GB recommended) running Windows, macOS or Linux

## Tools

Python 3.13 or later, Jupyter Notebook or JupyterLab, Visual Studio Code, NumPy, pandas, matplotlib and seaborn, Plotly, SciPy and statsmodels, scikit-learn, SHAP, Streamlit, Git and GitHub, Google Colab (optional free alternative to a local setup; paid tiers are not needed)

# Syllabus

## Module 1: Python for data science

A fast Python refresher for people who already program, plus NumPy and a reproducible project setup.

1. Setup and Python core types (2 h)
2. Functions, modules and clean code (2 h)
3. NumPy and vectorised computation (2 h)
4. Python review and doubt-clearing (2 h)

## Module 2: Data wrangling with pandas

Load, clean, combine and reshape real-world data with pandas, and handle it responsibly.

5. Loading data and data provenance (2 h)
6. Selecting, filtering and transforming (2 h)
7. Cleaning messy data (2 h)
8. Grouping, merging, reshaping and time series (2 h)
9. pandas review and doubt-clearing (2 h)

## Module 3: Exploratory analysis and visualisation

Explore data systematically and communicate what you find with clear, honest charts.

10. Exploratory data analysis with matplotlib and seaborn (2 h)
11. Honest and interactive visualisation (2 h)
12. EDA workshop and doubt-clearing (2 h)

## Module 4: Statistics for data science

The probability and statistics needed to reason about uncertainty, test claims and avoid common traps.

13. Descriptive statistics and distributions (2 h)
14. Probability, sampling and the central limit theorem (2 h)
15. Confidence intervals and hypothesis testing (2 h)
16. A/B testing and statistical pitfalls (2 h)
17. Correlation, causation and linear regression (2 h)
18. Statistics review and doubt-clearing (2 h)

## Module 5: Machine learning foundations

What machine learning is, how to set up a problem correctly and the core scikit-learn workflow.

19. Framing machine-learning problems (2 h)
20. Linear regression and regularisation with scikit-learn (2 h)
21. Logistic regression and classification (2 h)
22. ML foundations review and doubt-clearing (2 h)

## Module 6: Feature engineering and pipelines

Turn raw columns into useful features and wrap preprocessing in pipelines that cannot leak.

23. Encoding, scaling and imputation (2 h)
24. Creating and selecting features (2 h)
25. Pipelines, ColumnTransformer and doubt-clearing (2 h)

## Module 7: Model evaluation and tuning

Measure models with the right metrics, validate them honestly and tune them without fooling yourself.

- 26. Evaluation metrics (2 h)
- 27. Cross-validation and hyperparameter tuning (2 h)
- 28. Error analysis, imbalanced data and doubt-clearing (2 h)

### **Module 8: Machine-learning algorithms**

How the main families of algorithms work, when to use each, and how to find structure without labels.

- 29. Decision trees (2 h)
- 30. Ensembles: random forests and gradient boosting (2 h)
- 31. Other supervised methods and an introduction to neural networks (2 h)
- 32. Unsupervised learning: clustering and PCA (2 h)
- 33. Algorithms review and doubt-clearing (2 h)

### **Module 9: Responsible AI and communicating results**

Check models for fairness and privacy risks, explain their behaviour and communicate results honestly.

- 34. Fairness, bias and privacy (2 h)
- 35. Explainability and model documentation (2 h)
- 36. Communicating and sharing results, with doubt-clearing (2 h)

### **Module 10: Capstone project and final assessment**

Plan, build and present an end-to-end data-science project, with structured reviews and the final assessment.

- 37. Capstone kickoff and data audit (2 h)
- 38. Capstone build lab and mid-point review (2 h)
- 39. Final assessment (2 h)
- 40. Capstone presentations and final review (2 h)

## **Final project**

Capstone: predict, explain and responsibly recommend. Choose one of the written briefs — for example, predicting which deliveries will arrive late for a fictional logistics company, or forecasting demand for a fictional chain of grocery stores, each with a synthetic dataset generated for the course — or propose a public dataset with a clear licence, agreed with the instructor in advance. Projects are done individually or in pairs.

**Deliverables-**

- A problem statement: who would use the model, what decision it supports, how success is measured and what harm a wrong prediction could cause.
- A data audit covering source and licence, data dictionary, quality problems found and any personal data, with how it was removed, minimised or protected.
- An exploratory analysis notebook with clear, labelled charts and the questions it raised.
- At least one statistical test or confidence interval that supports a claim made in the project.
- A scikit-learn pipeline with feature engineering, a simple baseline and at least two candidate models compared with cross-validation and a held-out test set.
- An evaluation section with metrics chosen for the problem, error analysis and a check of performance across relevant groups.
- Explanations of the model's behaviour using permutation importance or SHAP, and a model card stating intended use, limitations and risks.
- A short presentation for a non-technical audience and a GitHub repository with a README describing how to reproduce the work.

**Optional extension:** a small Streamlit app that lets a user try the model on example inputs, with its limitations shown on screen.

The capstone is presented and reviewed live in the final week, with feedback on method, honesty of evaluation and clarity of communication.

## **Certificate**

- At least 80% attendance
- All assignments submitted
- Projects passed

- Final assessment passed (60%)

A ZORK Academy course completion certificate; paying for the course does not by itself earn it.

## Next step

Book a free consultation: <https://zorkacadamy.online/consultation?course=data-science-machine-learning>

We help you prepare for work; we do not guarantee jobs, placements or salaries.