CONCEPT

Python for Data Analysis and Machine Learning: pandas, Visualisation and scikit-learn

Python for Data Analysis and Machine Learning: pandas, Visualisation and scikit-learn

Technical depth: Practitioner · **Practical mode:** Lab

Introduction

Many organisations hold more operational data than their analysts can process in spreadsheets, so questions about customers, costs and service demand wait for specialist teams or are answered by manual workarounds that nobody can repeat. This Core Concept course builds hands-on Python capability for analysts, from cleaning and exploring data with pandas to training and validating first machine learning models with scikit-learn. Every session runs in Jupyter on realistic datasets, and participants leave with a Reproducible Analysis and Baseline Model Notebook built on a problem from their own work.

Course Objectives

- Set up a reproducible Python analysis environment with Jupyter, virtual environments and version control
- Clean, reshape and join tabular data with pandas and assess its fitness for use against recognised data quality measures
- Explore and visualise data with Matplotlib and seaborn to test business hypotheses and communicate findings
- Frame a business question as a supervised or unsupervised learning problem following the CRISP-DM process
- Train, validate and tune regression, classification and clustering models with scikit-learn and select suitable metrics
- Package an analysis as a documented notebook and model card that colleagues can rerun and review

Target Audience

- Data and business analysts moving from spreadsheets to programmatic analysis
- Finance, HR and operations analysts responsible for recurring reporting and ad hoc investigations
- Business intelligence developers extending dashboards with statistical and predictive analysis
- Researchers and statisticians in public service and policy units working with administrative data
- Engineers and quality specialists analysing process, sensor and inspection data
- Junior data scientists consolidating their foundations in Python and machine learning

Course Outline

Day 1: Python Foundations and the Analytics Workflow

- CRISP-DM Phases Mapped to Everyday Analytical Work
- Jupyter Notebook, Virtual Environment and Package Management Setup
- Python Data Types, Control Flow and Functions for Analysis
- NumPy Arrays and Vectorised Computation
- Importing CSV, Excel, SQL and API Data into pandas

Day 2: Data Structures, Wrangling and Quality Standards

- pandas Series and DataFrame Indexing and Selection
- Missing Values, Duplicates and Type Conversion in pandas
- Reshaping and Joining with groupby, merge, pivot_table and melt
- Data Quality Profiling Against ISO/IEC 5259-2 Measures
- ML System Components and Terminology per ISO/IEC 23053 and ISO/IEC 22989

Day 3: Exploratory Analysis and First Models

- Descriptive Statistics and Distribution Analysis in pandas
- Matplotlib and seaborn Chart Selection for Analytical Questions
- Correlation Analysis and Hypothesis Testing with SciPy
- Regression, Classification and k-Means Clustering with scikit-learn Pipelines
- Evaluation Metrics: RMSE, Precision, Recall, ROC AUC and Silhouette Score

Day 4: Model Validation, Pitfalls and Optimisation

- k-Fold Cross-Validation and Data Leakage Prevention
- Overfitting Diagnosis with Learning Curves and Regularisation
- Class Imbalance Handling with Stratification and Class Weights
- Hyperparameter Tuning with GridSearchCV
- Permutation Feature Importance for Model Explanation

Day 5: Lab Work and the Reproducible Analysis Notebook

- Customer Churn Lab: Classification Model Build and Interpretation
- Public Service Demand Lab: Clustering and Segment Profiling
- Model Card Drafting for a Baseline Model
- Reproducible Analysis and Baseline Model Notebook Build
- Peer Code Review Against PEP 8 Conventions and Notebook Handover

Skills You Will Gain

- Python Programming for Analytics
- Data Wrangling
- Data Quality Assessment
- Exploratory Data Analysis
- Statistical Hypothesis Testing
- Supervised Learning
- Model Validation
- Reproducible Analysis

Why Attend This Course

- Return to work with a Reproducible Analysis and Baseline Model Notebook built on your own data problem and reviewed by peers
- Replace repetitive spreadsheet steps with scripted pipelines that can be rerun whenever new data arrives
- Recognise when a model result is too good to be true, from leakage to overfitting, before it reaches decision makers
- Work through the same lab datasets as analysts from other sectors and compare approaches to common data problems

Conclusion

Python has become the working language of analytics because it lets one analyst move from raw extracts to tested models in a single, repeatable workflow. This course builds that workflow step by step: environment and data handling, quality profiling against recognised measures, exploratory analysis, first models and the validation discipline that keeps them honest. The final day turns the material into a Reproducible Analysis and Baseline Model Notebook that participants take back to work and extend as new data and questions arrive.