

Database, Data Analytics and Machine Learning Foundations

Python data acquisition, SQL Server and guided analysis

FA-0778 · 4 days

Data & Analytics · Software Development

Introduction

Build a foundation in relational data, Python-based analysis and introductory machine learning. Work with prepared files, APIs and database connections, then use guided examples to clean, visualise and model data. The course combines basic programming and database skills with selected machine-learning demonstrations; it is not advanced mastery of arbitrary data sources or production infrastructure.

Learning outcomes

- Describe SDLC phases and development practices in a data project.
- Use basic Python, external libraries, files and encoding for data acquisition and preparation.
- Connect to a prepared relational database and run basic SQL queries.
- Visualise data and apply selected classification, clustering and regression methods.
- Evaluate models using separated training/test data and cross-validation.
- Recognise Linux, SSH and remote-query concepts used in the prepared environment.

Prerequisites

Basic computer, internet and file-handling skills and English-language technical reading. Use a prepared compatible environment and authorised access to course data and remote services; only permissions needed for the exercises are required.

Training outline

1. Day 1 — Python, project lifecycle and data acquisition

- SDLC phases, pre-development decisions and project-independent development practices.
- Computing/workflow design concepts illustrated through a prepared data process.
- Python syntax and basic data structures using a supported Python3 version.
- External libraries, files, encoding, extraction and conversion.
- Data acquisition through prepared files, permitted scraping/API examples and database connectivity.
- Linux, SSH and authorised remote-query basics.

2. Day 2 — SQL Server environment and SQL

- RDBMS and Microsoft SQL Server; SQL and T-SQL.
- Container and virtual-machine concepts; prepared supported Linux container rather than Ubuntu18.04 installation instructions.
- Remote access, environment security and differences between lab and production setup.
- Use SSMS or a supported client such as VS Code MSSQL; Azure Data Studio is retired.
- Databases, tables, stored procedures, functions and import/export.
- SELECT, WHERE, LIKE, ordering, INSERT, UPDATE, DELETE, IN and BETWEEN.
- Aggregates, grouping, ALTER and subqueries.

3. Day 3 — Preparation, visualisation and probability

- Data visualisation with Seaborn
- Covariance and Correlation.
- Conditional Probability.
- Bayes' Theorem.
- Data Cleaning and Normalization
- Cleaning web log data
- Normalizing numerical data
- Detecting outliers
- Feature Engineering
- Imputation Techniques for Missing Data
- Handling Unbalanced Data: Oversampling

Fit imputation, scaling and other learned preparation only on training data; oversampling belongs within the training process, not the held-out evaluation data.

4. Day 4 — Introductory machine learning

- Basic Machine Learning (using Python).
- Supervised vs. Unsupervised Learning, and Train/Test.
- Use train/test evaluation to assess polynomial-regression overfitting
- Bayesian Methods: Concepts.
- Practice with:
- Implementing a Spam Classifier with Naive Bayes.
- K-Means Clustering.
- Entropy
- Decision Trees
- Bias/Variance Tradeoff
- K-fold cross-validation to assess generalisation and guide model selection

Introduce the roles of pandas, SciPy, scikit-learn, Matplotlib, Seaborn and XGBoost in the examples. Review predictions and complete the assessment.