KOENIG SOLUTIONS · COURSE OUTLINE

Data Pre-Processing: Cleaning, Scaling & PCA

Learn to clean, transform and reduce data through scaling, feature selection and PCA for machine learning workflows.

3 Days	10	41
24 HOURS	MODULES	TOPICS

COURSE OVERVIEW

This course is designed to provide practical examples and hands-on exercises to reinforce the concepts. Additionally, encourage students to work on a mini-project or real datasets to gain practical experience in data preprocessing.

WHO THIS COURSE IS FOR

- **Data Analysts**
- **Data Scientists**
- **Machine Learning Engineers**
- **Data Engineers**
- **Business Intelligence Analysts**

LEARNING OBJECTIVES

- Explain the role of data preprocessing in the data science workflow
- Identify methods for collecting data from web sources and APIs
- Apply techniques for handling missing values and outliers
- Compare one-hot encoding and label encoding for categorical data
- Analyse data distributions using Matplotlib and Seaborn
- Apply feature selection methods including PCA and t-SNE
- Plan data splits using cross-validation and stratified sampling
- Describe preprocessing for time series, text, image and audio data

COURSE SYLLABUS

10 Modules, 41 Topics

01

Introduction to Data Preprocessing

TOPICS

- What is Data Preprocessing?
- Why Data Preprocessing is Essential in Machine Learning
- The Data Science Workflow
- Common Challenges in Real-World Data

02

Data Collection and Acquisition

TOPICS

Data Sources and Formats · Web Scraping and APIs · Handling Structured and Unstructured Data · Data Privacy and Ethics

03

Data Cleaning

TOPICS

- Identifying Missing Data
- **Handling Missing Values:** Imputation Techniques
- Outlier Detection and Handling
- Data Duplicates and Noise Removal

04

Data Transformation

TOPICS

- Scaling and Normalisation
- **Encoding Categorical Data:** One-Hot Encoding vs. Label Encoding
- **Feature Engineering:** Creating New Features
- Date and Time Handling

05

Data Visualization and Exploration

TOPICS

- Data Visualization Libraries (Matplotlib, Seaborn)
- Univariate and Bivariate Analysis
- Correlation Analysis
- Data Distribution and Transformation

06

Feature Selection and Dimensionality Reduction

TOPICS

- Importance of Feature Selection
- Common Feature Selection Techniques
- Principal Component Analysis (PCA)
- t-Distributed Stochastic Neighbour Embedding (t-SNE)

07

Data Splitting and Sampling Techniques

TOPICS

- Training, Validation, and Test Sets
- Cross-Validation
- Stratified Sampling
- Oversampling and Under-sampling for Imbalanced Data

08

Handling Time Series Data

TOPICS

Understanding Time Series Data · Resampling and Interpolation · Time Series Feature Engineering · Seasonal Decomposition

09

Text Data Preprocessing

TOPICS

- Tokenization and Stopword Removal
- Text Vectorization (TF-IDF, Word Embeddings)
- Handling Text Data Challenges (Spelling Errors, Noise)
- Sentiment Analysis

10

Handling Image and Audio Data

TOPICS

Image Preprocessing (Resizing, Normalisation) · Image Augmentation · Audio Data Preprocessing · Spectrogram and MFCC Features · Challenges in Image and Audio Data

WHAT IS INCLUDED

Everything You Need to Succeed

- Koenig course material for every module
- Delivery by certified instructors and subject matter experts
- Hands-on exercises throughout the course
- Course completion certificate from Koenig

ABOUT KOENIG

Who We Are

Established in 1993, Koenig Solutions is a globally recognised training provider with 35+ technology partnerships, 5,000+ training programmes and a network of certified instructors and subject matter experts, supporting individuals and enterprise teams worldwide.

FLEXIBLE DELIVERY

Learn the Way That Suits You

Classroom Training | Live Online Classes (ILO) | Fly-Me-A-Trainer (FMAT) | Flexi (Self-Paced) | 1-on-1 Training

Ready to Get Started?

Talk to a learning advisor about schedules, group training and customised delivery.

Visit koenig-solutions.com | info@koenig-solutions.com