Big Data Analytics with Python

Hands-on course of 4 days - 28h
Ref.: BDA - Price 2025: 2 920 (excl. taxes)

EDUCATIONAL OBJECTIVES

At the end of the training, the trainee will be able to:

Understanding the principle of statistical modeling

Choosing regression and classification depending on data type

Evaluating an algorithm's predictive performance

Creating selections and classifications in large volumes of data to reveal trends

HANDS-ON WORK

Developing/conducting analysis in Python, with the modules pandas, NumPy, SciPy, MatPlotLib, seaborn, scikit-learn, and statsmodels.

TRAINER QUALIFICATIONS

The experts leading the training are specialists in the covered subjects. They have been approved by our instructional teams for both their professional knowledge and their teaching ability, for each course they teach. They have at least five to ten years of experience in their field and hold (or have held) decision-making positions in companies.

ASSESSMENT TERMS

The trainer evaluates each participant's academic progress throughout the training using multiple choice, scenarios, hands-on work and more.
Participants also complete a placement test before and after the course to measure the skills they've developed.

TEACHING AIDS AND TECHNICAL RESOURCES

• The main teaching aids and instructional methods used in the training are audiovisual aids, documentation and course material, hands-on application exercises and corrected exercises for practical training courses, case studies and coverage of real cases for training seminars.
• At the end of each course or seminar, ORSYS provides participants with a course evaluation questionnaire that is analysed by our instructional teams.
• A check-in sheet for each half-day of attendance is provided at the end of the training, along with a course completion certificate if the trainee attended the entire session.

TERMS AND DEADLINES

Registration must be completed 24 hours before the start of the training.

ACCESSIBILITY FOR PEOPLE WITH DISABILITIES

Do you need special accessibility accommodations? Contact Mrs. Fosse, Disability Manager, at psh-accueil@ORSYS.fr to review your request and its feasibility.

THE PROGRAMME

last updated: 07/2024

1) Introduction to modeling

- Introduction to the Python language.
- Introduction to the Jupiter Notebook software.
- Steps for building a model.
- Supervised and unsupervised algorithms.
- Choosing between regression and classification.

Hands-on work : Installing Python 3, Anaconda, and Jupiter Notebook.

2) Model evaluation procedures

- Techniques for resampling in training, validation and testing sets.
- Learning data representativeness test.
- Predictive model performance measurements.
- Confusion and cost matrix and AUC-ROC curve.

Hands-on work : Setting up data set sampling. Conducting evaluation tests on multiple provided models.

3) Supervised algorithms.

- The principle of univariate linear regression.
- Multivariate regression.
- Polynomial regression.
- Regularized regression.
- Naive Bayes.
- Logistic regression.

Hands-on work : Implementing regressions and classifications on multiple data types.

4) Unsupervised algorithms

- Hierarchical clustering.
- Non-hierarchical clustering.
- Mixed approaches.

Hands-on work : Handling unsupervised clusters in multiple datasets.

5) Component analysis

- Principal component analysis.
- Correspondence analysis.

- Multiple correspondence analysis.
- Factor analysis for mixed data.
- Hierarchical classification of principal components.

Hands-on work : Reducing the number of variables and identifying underlying factors of dimensions associated with significant variability.

6) Text data analysis

- Collecting and preprocessing text data.
- Extracting primary entities, named entities, and reference resolution.
- Grammatical tagging, syntactical analysis, semantic analysis.
- Lemmatization.
- Text vectorization.
- TF-IDF weighting.
- Word2Vec.

Hands-on work : Explore the contents of a text base using latent semantic analysis.

DATES

REMOTE CLASS
2025 : 19 août, 28 oct.