Introduction

This three-day course gives an introduction to the basics of Data Analysis with the KNIME® environment. The course is for students already familiar with KNIME® but who have no previous experience with data analysis. Topics covered by the course include starting out with KNIME®, data collection, data mining, processing, visualisation, and how to programme the most important data models using node workflows.

Who is this course for?

This course is for scientists who want to learn all they need to know about data analysis with the KNIME® environment.

You will learn how to . . .

•Collect your data from different sources (CSV, SDF, web), mine your data and understand its nature, decorrelate, filter out non-useful data and associated noise.

•Choose the right model for your data, test it and visualise your results in an open, powerful and versatile environment.

To follow this course you should . . .

•Already have some experience using the KNIME® environment, or in data flow programming with Pipeline Pilot®, Spotfire® or LabVIEW® to an equivalent level to that obtained in the PIKAÏROS course « An Introduction to Data Mining with KNIME » with L1 KNIME certification.

•Understand basic computer English (the KNIME® interface is in English).

Terms, access time and adaptation to people with disabilities

This course is entirely based on the KNIME educational software with the implementation of workflows as a learning method.

A continuous evaluation by MCQ is carried out on the exercises proposed during the training.

This course is offered only face-to-face in intra-company at the client's premises.

Possibility of adapting the time and mode of training to people with disabilities.

Consider 1 month waiting time to access the training from the order.

Full programme

First day

Presentation of the Knime analysis environment

Introduction to Data Analysis

- What is Data Analysis?
- What is the relationship between Data Analysis and other scientific disciplines: Data mining, Machine learning, Pattern recognition, Artificial intelligence, Statistics, Big Data
- The main steps of Data Analysis : Defining the need, data collection, data mining, processing and modeling, visualising and communicating the results

Problems solved with the help of Data Analysis: Regression, Inference, Classification, Partitioning, Time Series Prediction, . . .

The hidden nature of data

- What makes a good data set: the importance of data dimensions, sample volume, the number of variables and their nature
- Data processing and filtering: Selecting variables, correlation, the data validity domain, the race for dimensionality
- Why and how to split up a learning set : data sampling, cross-validation techniques

Exercises : Data Collection, processing and filtering in the KNIME® environment

- Reading data
- Sampling data and generating learning, cross validation and testing sets
- Correlation assessment of variables
- Implementing the selection of variables

Second day

The nature of Data Models:

- **Classical models**
  - Linear model
  - Bayes model
  - Traditional data analysis:
    - Principal component analysis
    - Fuzzy logic
    - Decision trees and hierarchical clustering

Exercises: Practising data modeling

- Putting into practice the methods presented in the course using real examples
  - Linear classifier
  - Principal component analysis
  - Naive Bayes model
  - Fuzzy logic classifier
  - Decision trees
  - A practical example of Overfitting
  - The effects of the race for dimensionality
  - Writing and Reading Data Models in KNIME®

Third day

The nature of Data Models:

- **Machine learning methods**
  - What is Machine Learning?
    - Concept of Learning and Cost Function
    - Degrees of freedom of a model
    - Overfitting and regularisation methods
  - Supervised learning:
    - Kernel methods (RBF and SVM)
    - Random forests
    - Ensembles of classifiers
  - Unsupervised learning:
    - k-means clustering
    - Expectation-Maximisation
    - Hierarchical Clustering Methods

Exercises: Practising data modeling

- Putting into practice the methods presented in the course using real examples:
  - Kernel methods (RBF and SVM)
  - Bagging, Boosting, Random forests, Ensembles of classifiers
  - K-Means, Expectation-Maximisation, Hierarchical Clustering