Course code: **WEKA**

Course title: **Data Mining with Weka (Pentaho Data Mining)**

Days: 4

Description:

Course intended for:

Training is intended for data scientists and developers, who want to create or maintain processes of data exploration with the use of Pentaho Data Mining (WEKA).

Course objective:

Course participants will gain knowledge on design, implementation, monitoring, running and tuning of DM processes, as well as will freshen up their knowledge on basic statistical terms and learn the most popular DM algorithms in details. They will be able to choose the appropriate set of tools and techniques for their real-world projects.

Course strengths:

Course curriculum includes general introduction to the subject of Data Mining and Machine Learning, as well as a comprehensive presentation of a real-world process using the Weka environment.

Requirements:

Participants are expected to have basic programming knowledge of Java.

Course parameters:

4*8 hours* (4*7 net hours) of lectures and workshops. During workshops, apart from doing simple exercises, participants will solve data exploration problems and will use and tune DM algorithms. Group size: max. 8-10 people

Course curriculum:

1. Introduction
    I. introduction to data warehouse:
        i. OLTP, OLAP, database, data warehouse, data marts
        ii. ROLAP, MOLAP, HOLAP
        iii. Normalization, aggregation, facts, dimensions

- iv. SQL, MDX, XML/A
- v. ETL
- vi. BigData, BigTable, NoSQL, non-relational data warehouses
- vii. Others

II. Pentaho BI Suite Platform

2. Data exploration
   - I. Artificial intelligence, machine learning, data exploration etc.
   - II. Basics of data mining algorithms
     - i. Algorithms
       - classification
       - clustering
       - finding patterns and association rules
       - transforming and reducing the space of attributes
     - ii. Techniques:
       - Trees and decision tables
       - linear regression
       - Bayesian networks
       - Neural networks
       - Genetic and evolutionary algorithms
     - iii. Basic statistical terms
       - Minimum, Maximum
       - Mean, Median
       - Standard deviation, Variance
       - Probability
       - Correlation
       - Distance metric
       - Statistical significance
     - iv. Others
   - III. Overview of data mining tools available on the market
3. Pentaho Data Mining (WEKA)
   - I. Architecture
   - II. Weka GUI Chooser
     - i. Explorer
     - ii. Experimenter
     - iii. Knowledge Flow
     - iv. Simple CLI
     - v. Tools: ARFF Viewer, SQL Viewer etc.
     - vi. Weka Light, Weka Server
   - III. Working with Explorer
4. Preprocessing and working with data
   - I. ARFF data format
   - II. Data preprocessing
   - III. Attribute selection
   - IV. Data filtering and types of filters in WEKA e.g. filtering, discretization, normalization etc.
   - V. Visualization

    VI. Processing of big data sets, JVM 32bit limitations
    VII. Sttream processing and incremental learning
5. Classification
    I. Classification problem definition
    II. Selecting an appropriate set of training and testing data
    III. Types of classification algorithms available in WEKA
    IV. Most popular classification algorithms in details
        i. Bayesian networks e.g. Naive Bayesian classifier
        ii. Regression e.g. linear regression
        iii. Trees and decision tables
    V. Cross-validation, overfitting
    VI. Interpretation of classification results
6. Clustering
    I. Clustering problem definition
    II. Selecting an appropriate set of training and testing data
    III. Types of clustering algorithms available in WEKA
    IV. Most popular clustering algorithms in details
        i. Centroids, e.g. k-Means
        ii. Density-based, e.g. DBSCAN
    V. Interpreting the results of clustering
7. Association rule mining
    I. Association rule mining definition
    II. Selecting an appropriate set of training and testing data
    III. Types of association rule mining algorithms available in WEKA
    IV. Most popular algorithms in details
        i. Apriori
        ii. Frequent Pattern Growth
    V. Interpreting the results
8. Transforming and reducing the attribute space
    I. Defining the problems of: attribute selection, attribute space reduction and transformation
    II. Types of algorithms for transforming the attribute space in WEKA
    III. Most popular algorithms in details
        i. Searching the attribute space, e.g. BestFirst, ExhaustiveSearch, GeneticSearch
        ii. Principal Component Analysis (PCA)
        iii. Support Vector Machines (SVM/SVMAttributeEval)
    IV. Interpreting the results
9. Other data mining algorithms and techniques available in WEKA
10. Extending WEKA
    I. Pentaho Data Mining Plug-Ins
    II. User-define DM algorithms in WEKA
11. Combining Weka with other Pentaho products
    I. Knowledge Flow Plugin and Pentaho Data Integration