RapidMiner & Data Science: Foundations

Course Overview

RapidMiner & Data Science: Foundations is a two-day course focusing on data mining and predictive analytics with RapidMiner Studio. Over the course of two days, students will explore a simplified business use case and build a strong analytical model while becoming familiar with the graphical interface and the main product features and functionality. They will also be introduced to basic concepts in modern data science and several popular machine learning algorithms for predictive modeling.

The course is structured in a mentoring fashion where the entire group performs tasks alongside the instructor as members of a data science team. After successfully completing this course, participants will have a solid understanding of how RapidMiner Studio functions. Participants will be able to import data into RapidMiner from common files types. They will be able to prepare data using common ETL transformations for data mining. They will learn how to create and validate predictive models, and evaluate them using a variety of common model performance criteria.

Practical exercises during the course prepare students to take the knowledge gained and apply it to their own complex data challenges. The class exercises and labs are hands-on, so students will internalize the topics covered, which will provide a jumpstart to the real world application of these techniques.

Prerequisites & Target Audience

This class is aimed at Analysts and Data Scientists. It assumes a basic knowledge of computer programming principles and higher mathematics (through calculus), but does not require prior knowledge of RapidMiner software or any academic preparation in applied statistics or data science.

Course Objectives

After the training, students will have the ability to use RapidMiner to:

- Perform all common data preparations and transformations for data mining
- Build strong analytical predictive models based on best-practice validation approaches
- Evaluate model quality with respect to several different performance criteria
- Deploy analytical predictive models

Course Outline

- Overview
  - Introduction to the RapidMiner ecosystem
  - Business Scenario
  - Analytics Taxonomy & Hierarchy
  - CRISP-DM & Data Mining in the Enterprise
- Getting Started with RapidMiner Studio
  - User Interface
  - Creating and Managing RapidMiner Repositories
  - Operators and Processes
  - Storing Data, Processes, and Result Sets
- EDA: Exploratory Data Analysis
  - Loading Data
  - Quick Summary Statistics
  - Visualizing Data & Basic Charting
- Data Preparation
  - Basic Data ETL (Extract, Transform, and Load)
  - Data Types & Transformations of Value Types
  - Handling Missing Values
  - Handling Attribute Roles
  - Filtering Examples and Attributes
  - Normalization and Standardization
- Building Better Processes
  - Organizing, Renaming, & Relative Paths
  - Sub-Processes
  - Building Blocks
  - Breakpoints
- Predictive Modeling Algorithms
  - k-Nearest Neighbor
  - Naïve Bayes
  - Linear Regression
  - Decision Trees & Rules
- Model Construction and Evaluation
  - Machine Learning Theory: Bias, Variance, Overfitting & Underfitting
  - Splitting Data
  - Split and Cross Validation
  - Evaluation Methods & Performance Criteria
  - Optimization and Parameter Tuning
  - Applying Models