

RapidMiner & DataScience: Foundations

Course Overview

RapidMiner & DataScience: 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.

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 prepare data, create and validate predictive models, and will be ready to extend their knowledge to advanced topics such as *RapidMiner & DataScience: Advanced* and *RapidMiner Server: Deployment and Web Apps*.

Practical exercises during the course prepare students to take the knowledge gained and apply it to their own respective data mining problems, solving them quickly and easily. Since the class labs are hands-on and performed on the participants' personal laptops, students will take actual classwork home with them, which will provide a jumpstart to data mining in the real world.

Target Audience

Analysts, Advanced Analysts and Data Scientists

Prerequisites

Basic knowledge of computer programs and mathematics

Course Objectives

After the training, students will have the ability to:

- Perform all common data preparations
 - Build strong analytical predictive models
 - Evaluate model quality with respect to different performance criteria
 - Deploy analytical predictive models

Course Outline

- **Overview**
 - ◇ 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
- **Additional Workshops**
 - ◇ Outlier Detection
 - ◇ Random Forests
 - ◇ Ensemble Modeling