SolidQ Central & Eastern Europe
Invalidenstraße 7/10
1030 Wien
Austria

solidq.com
+43 (720) 880 092 office
+43 (1) 253 033 300 30 fax
ceinfo@solidq.com

# DATA SCIENCE ALGORYTHMS IN SSAS, EXCEL, R, AND AZURE ML

## *Statistics, data mining and machine learning explained*

## Delivery format

**Instructor-led training** in class, with maximum number of attendees **12**, 24 training hours spread in **3 days**.

## Author and Instructor

**Dejan Sarka**, MCT and SQL Server MVP, is an independent trainer and consultant that focuses on development of database & business intelligence applications. Besides projects, he spends about half of the time on training and mentoring. He is the founder of the Slovenian SQL Server and .NET Users Group. Dejan Sarka is the main author or coauthor of thirteen books about databases and SQL Server. Dejan Sarka also developed many courses and seminars for Microsoft, SolidQ and PluralInsight.

### Language

Deliver possible in English, Slovenian, Serbian, Croatian; material in English.

## Summary

Don't just use statistics, data mining, and machine learning without understanding how it works. Get the insights in the most popular algorithms.

## Abstract

Advanced data analysis techniques are gaining popularity. With modern statistics / data mining / machine learning engines, products and packages, like SQL Server Analysis Services (SSAS), Excel, R, and Azure ML, data mining has become a black box. It is possible to use data mining without knowing how it works. However, not knowing how the algorithms work might lead to many problems, including using the wrong algorithm for a task, misinterpretation of the results, and more. This course explains how the most popular data mining algorithms work, when to use which algorithm, and advantages and drawbacks of each algorithm as well. Demonstrations and labs show the algorithms usage in SQL Server Analysis Services, Excel using the SSAS algorithms, R language and SQL Server R Services, Azure ML native algorithms, and using the R algorithms in Azure ML. The attendees also learn how to evaluate different predictive and unsupervised models.

Algorithms explained include Naïve Bayes, Decision Trees, Neural Networks, Logistic Regression, Perceptron Model, Linear Regression, Regression Trees, Ordinal

Regression, Poisson Regression, Principal Component Analysis, Support Vector Machines, Hierarchical Clustering, K-Means Clustering, Expectation-Maximization Clustering, Association Rules, Sequence Clustering, Auto-Regressive Trees with Cross-Prediction (ARTXP), Auto-Regressive Integrated Moving Average (ARIMA), and Time Series.

The course also includes the explanation of the introductory statistics, including descriptive statistics, correlations and linear associations. Even the information theory is touched briefly. All of these methods are useful for gathering understanding of the data used for later analysis and advanced data profiling. Mining unstructured data, specifically texts, is covered in the course as well. Finally, a practical real life example, namely anomaly detection, concludes the course.

## Class Focus

The focus of the training is the theoretical concepts of advanced analytics. The importance for the attendees to fully understand how the algorithms work, how to correctly use them, how to prepare the data, and how to interpret the results is the first training goal. The software part is used just for showing the concepts and enriching the concept with examples. It helps a lot in understanding how to work with data, how to prepare useful derived variables, or to smooth values of a variable appropriately, or to discretize them correctly, etc. Attendees can and should be able to use different tools in the future.

## Prerequisites

Attendees should have basic understanding of data analysis, relational data models; knowledge in statistics and mathematics is a very desired to get the maximum results of this training.

## Course format

This class contains about 65% theory and demos explained by the Trainer. About 35% of the time attendees perform practical exercise. After every module the group discuss the results of the lab to make sure that the concept and the practical scenarios are well understood.

## Technical Prerequisites

Every attendee should work on a dedicated virtual machine with the following software installed:

- Windows Server 2012 R2
- SQL Server 2016 Database Engine
- SQL Server 2016 Analysis Services in Multidimensional and Data Mining mode

- SQL Server R Services
- SQL Server Integration Services
- SQL Server Data Tools
- SQL Server Management Studio
- RStudio or (and) R Tools for Visual Studio
- AdventureWorks and AdventureWorksDW demo databases
- Microsoft Excel 2013 or 2016
- Microsoft Excel Office Data Mining Add-ins and Azure ML add-in
- Microsoft Excel Power Map and Power View add-ins enabled
- Power BI Desktop
- Azure ML free account created
- All the necessary Lab files copied in the virtual machine

## Course Material

Every attendee gets a .PDF printout of all slides and detailed lab instructions. In addition, attendees are welcome to copy the demo and lab solutions for further reference.

## Knowledge Assessment

To evaluate the knowledge of the attendees we developed 60 different questions. The questions can be split into two halves to assess the knowledge before and after the training.

## Modules

1. Introduction to data mining, machine learning, and statistics
2. Introducing advanced analytics in SSAS, Excel, Azure ML and R
    a. Lab: Getting familiar with the tools
3. Statistics for data profiling and understanding
    a. Lab: Data profiling and introductory statistics
4. Data preparation
    a. Lab: Using SSIS to split the data into training and test set and checking the split with Decision Trees
5. Classification and prediction algorithms
    a. Lab: Using the Naïve Bayes, Decision Trees, Logistic Regression, and Neural Network algorithms, and evaluating predictive models
6. Estimation algorithms
    a. Lab: Using the Linear Regression and Regression Trees algorithms
7. Unsupervised algorithms
    a. Lab: Using the Association Rules, Clustering, and Sequence Clustering Algorithms
8. Forecasting algorithms
    a. Lab: Using the Time Series, ARIMA and ARTXP algorithms
9. Personal analysis of geographical and temporal data

    a. Lab: Using Excel with Power Map and Power View, and Power BI Desktop
10. Advanced personal analytics
    a. Lab: Using Excel for data mining, using R in Power BI Desktop
11. Analyzing texts with SSIS, Transact-SQL, SSAS, R, and Azure ML
    a. Lab: Text mining
12. Task example: anomaly detection