

Cortana Intelligence Suite Workshop – AML001: K-Means Clustering And the Team Data Science Process

# Training Overview

Welcome to the Cortana Intelligence Suite workshop delivered by your Microsoft Data Science team. In this session, we’ll examine a real-world question from an organization. We’ll use the Microsoft Team Data Science Process (TDSP) and identify our source data, document the data sources, import the data into the AML environment, cleaning and the data and testing algorithms to get the correct training mechanism, train the model and evaluate the best path forward. We’ll then publish our model and use it in an application. It’s an end-to-end session, using Machine Learning in a practical application.

This course is designed to take approximately one day, depending on what is covered and how many of the labs are done in-class. All materials are provided regardless of the length of the course.

**Note:**

If you’re getting these materials without classroom instruction, simply walk through the videos and other materials on the delivery site step-by-step, making sure you read and follow the links in the handouts for each topic.

# Audience

# Technical professionals (Data Scientists, Database professionals, Analysts, BI Professionals) who are familiar with building solutions but not familiar with the entire CIS Platform of products

# Prerequisites

There are a few things you need prior to coming to class:

* A subscription to Microsoft Azure (this may be provided through your company or as part of your invitation – you **must** have this enabled *prior to class* – you will be using Azure throughout the course, for all labs, work and exercises)
  + You can use your MSDN subscription – <https://azure.microsoft.com/en-us/pricing/member-offers/msdn-benefits/>
  + Your employer may provide Azure resources to you, but make sure you check to see if you can deploy assets and that they know you’ll be using their subscription in the class.
  + Optionally, you may receive instructions in your class invitation.
* We’ll be using the Data Science Virtual Machine in Azure for the course. It has all the tools you will need to work with the materials. Make sure you’re able to use the Remote Desktop Protocol (RDP) from your system to be able to work through the labs.
* If you would also like to work with some of the tools locally (you still need an Azure subscription for this class), you can optionally obtain:
  + A laptop that you can install software on
  + Visual Studio installed – the Community Edition (free) is acceptable – Version 2015 preferable (<https://www.visualstudio.com/en-us/products/visual-studio-community-vs.aspx>)
  + Azure SDK and Command-line Tools installed (<https://azure.microsoft.com/en-us/downloads/> )
  + Azure Storage Explorer (<http://go.microsoft.com/fwlink/?linkid=698844&clcid=0x409>)
* It’s also a good idea to have a general level of predictive and classification Statistics, and a basic understanding of Machine Learning

# Syllabus

## Course Session

Each Training Module guides you through a logical progression with hands-on tasks in do-verb form. Each day is broken up into 1-4 hour Modules, where you will learn and perform labs as a group and individually.

***NOTE:*** *The workbooks you receive as part of the classes contain many resources to lead you through the course, and provide a rich set of references that you can use to learn much more about these topics. If the links do not resolve properly, type the link address in manually in your web browser. If the links have changed or been removed, simply enter the title of the link in a web search engine to find the new location or a corollary reference.*

We’ll cover the following concepts and technologies in this course, and you’ll have the following skills when you complete all the labs and class activities:

**Learning Objectives**

1. Understand the Team Data Science Process, CIS Platform components, Tools installation and overview
2. Understand Data sourcing, Feature selection techniques, Data cataloging, Data Ingestion, Data Exploration
3. Understand Data selection, including Features, Dimension reduction, Data processing, Data transformation and augmentation
4. Determine Algorithm selection and application, Parameter selection and adjustment
5. Show Business validation of report and results, Model testing and cross-validation
6. Deploy the solution using Data Destinations, API's, or Queries and Reports

**Technologies covered**

1. The Data Science Process, Azure Portal, ADC Interface, Visual Studio Interface, Power BI Interface
2. Azure Machine Learning Interface, Azure PowerShell, Azure Storage Explorer
3. Azure Data Catalog, Azure Storage, Techniques for discovery
4. Azure Machine Learning, Microsoft R Server overview
5. Azure Machine Learning, Microsoft R Server overview, Business Validation, Azure Storage
6. Azure Data Storage, Azure Machine Learning API,
7. Cortana Intelligence Process, Cortana Intelligence Suite Platform

**Live Deliveries: Follow-up Q&A with MS architects and trainers**