

**Duration** 3 days

## Overview

Presented by **Michael J. A. Berry** co-founder of Data Miners, Inc. and co-author of Data Mining Techniques and Mastering Data Mining.

This course introduces a data mining methodology that is a superset to the SAS SEMMA methodology around which SAS Enterprise Miner is organized. The course also introduces a wide range of data mining algorithms and both theoretical knowledge and practical skills. In this class, you work through all the steps of a data mining project, beginning with problem definition and data selection, and continuing through data exploration, data transformation, sampling, portioning, modeling, and assessment.

## Learn how to

- use a data mining methodology
- build and use decision trees and neural networks for modeling and scoring
- use survival analysis and create survival curves.

**Who should attend:** Business analysts, their managers, and statisticians

## Course Contents

### Introduction to Data Mining

- what is data mining?
- directed and undirected data mining
- models
- profiling and prediction

### Data Mining Methodology

- why have a methodology?
- how data miners can inadvertently learn things that are not true
- translating business problems into data mining problems
- the importance of model stability
- finding the right input variables
- sampling to create balanced model sets
- partitioning to create training, validation, and test sets
- data preparation
- model assessment

### Data Exploration

- developing intuition about data
- data structure
- data types
- data values
- exploring distributions
- summary statistics
- histograms
- using SAS Enterprise Miner for data exploration

### Regression Models

- the null hypothesis
- statistical significance
- confidence bounds
- variance and standard deviation
- standardized values
- correlation
- linear regression
- logistic regression
- using SAS Enterprise Miner to build regression models

### Decision Trees

- decision trees as data exploration and classification tools
- decision trees for modeling and scoring
- decision trees for variable selection
- alternate representations of decision trees
- algorithms used to build decision trees
- splitting criteria
- recognizing instability and overfitting in decision tree models
- capturing interactions between variables
- using SAS Enterprise Miner to build decision trees

### Neural Networks

- origins of neural networks
- neural networks compared with regression
- algorithms used to train neural networks
- data preparation requirements for neural networks
- picking appropriate inputs for neural networks
- creating neural network models using SAS Enterprise Miner

### Working with Results and Automating Projects

- Memory-Based Reasoning
- similarity and distance
- distance metrics appropriate for different kinds of data
- the role of the training set in memory-based reasoning (MBR)
- combining the votes of several neighbors
- other K-nearest neighbor techniques
- collaborative filtering
- using the SAS Enterprise Miner MBR node

## Clustering

- more on similarity and distance
- the k-means algorithm
- divisive clustering
- agglomerative clustering
- data preparation for clustering
- interpreting clusters
- finding clusters with SAS Enterprise Miner

## Survival Analysis

- origins of survival analysis
- how business data is different from clinical data
- hazards and hazard charts
- retention curves and survival curves
- calculating survival from retention
- calculating hazards empirically
- parametric hazard models
- censoring
- competing risks
- survival-based forecasting
- using SAS code in SAS Enterprise Miner to create survival curves

## Association Rules

- market basket analysis
- association rules
- sequential pattern analysis
- using SAS Enterprise Miner to discover associations in retail data

## Link Analysis

- background on graph theory
- sphere of influence
- using link analysis to generate derived variables
- graph-coloring algorithm
- Kleinberg's algorithm

## Genetic Algorithms

- optimization techniques and problems (SAS/OR software)
- other algorithms
- linear programming problems
- genetic algorithms