Course Outline:

Introduction to Data Science for Big Data Analytics *(Course T259)*

**What is this course about?**
Throughout this course, you will learn to derive value from vast amounts of untapped data and apply data analytics techniques for smart, data-driven decision-making. Using open source tools, this course covers the concepts necessary to move through the entire data science pipeline.

Specifically, you will learn how to harness data mining methods to answer business questions from internal and external data sources; estimate future outcomes; unearth patterns in data; pre-process unstructured data for deeper analysis; apply clustering, classification, and regression to data sets; and work with R to analyze structured and unstructured data as well as Big Data.

**Who should attend?**
This course is intended for business intelligence and database professionals, managers, data analysts, data scientists, and business analysts. Course content is also useful to those involved in forecasting and trends management.

A background in basic statistics is helpful, but not required. Programming background is helpful for understanding sample code, but not required.

**Duration:**
5 Days

**This course qualifies for:**
29 NASBA CPEs

**Chapter 1: Introduction to Data Science**
- Discuss how to turn business questions into data mining problems
- Explore diverse and wide-ranging data sources that can be used to answer business questions
- Introduce the statistical programming language, R

**Chapter 2: Exploratory Data Analysis (EDA)**
- Learn how classifiers are used to predict data categorization
- Explore how decision trees are built
- Build and apply a decision tree classifier
- Develop a random forest classifier on Big Data

**Chapter 3: Working With Unstructured Data**
- Learn how classifiers are used to predict data categorization
- Explore how decision trees are built
- Build and apply a decision tree classifier
- Develop a random forest classifier on Big Data

**Chapter 4: Categorizing Data with Classification Techniques**
• Learn how classifiers are used to predict data categorization
• Explore how decision trees are built
• Build and apply a decision tree classifier
• Develop a random forest classifier on Big Data

Chapter 5: Additional Classification Methods
• Express business problems as linear regression tasks
• Produce a linear regression model in R
• Interpret, evaluate, and improve the linear regression model
• Carry out a linear regression on Big Data

Chapter 6: Predicting Outcomes With Regression Analysis
• Express business problems as linear regression tasks
• Produce a linear regression model in R
• Interpret, evaluate, and improve the linear regression model
• Carry out a linear regression on Big Data

Chapter 7: Detecting Patterns in Data With Clustering Analysis
• Detect natural groupings in your data sets through clustering algorithms
• Explore the concept of similarity
• Use various distance measures to define similarity
• Perform top-down clustering with the K-Means algorithm
• Perform bottom-up clustering with hierarchical clustering
• Examine clustering techniques on unstructured data
• Create a MapReduce algorithm to perform clustering on Big Data

Chapter 8: Additional Techniques: Association Rules and Recommenders
• Generate association rules from transaction data
• Evaluate rules using measures of support, confidence, and lift
• Examine approaches toward improving rules
• Build recommenders from transaction data

Chapter 9: Additional Techniques: Social Network Analysis
• Analyze data that has a network structure
• Represent networks as edge lists and adjacency matrices
• Plot networks to visualize relationships
• Identify important nodes and patterns in a network

Chapter 10: Data Science and Big Data Analytics: Final Thoughts
• Examine alternative approaches to Big Data analytics
• Draw the various aspects of Big Data analytics together in a data mining life cycle
• Investigate the merging of Big Data analytics with traditional data warehouses
• Consider the ethics of Big Data analytics
• Explore the required skillsets of a data scientist

Chapter 11: Course Summary