**CORPORATE TRAINING COURSES**

**Title:** Introduction to Data Science

**Duration:** 1 Day

**Pre-requisite:** Basic Programming Knowledge

**Description**

This Introduction to Data Science course will expose you to real-world applications of data science and why it’s become such an integral part of business and academia. We will discuss the data science process and the various tools used to analyze data sets, and used to apply machine learning to defined data problems. We will begin by installing Anaconda, which provides the Python and R programming languages, in addition to the tools that we’ll use to explore and apply data science to real-world use cases. We will perform data exploration, analysis, modeling and visualization in order to learn the data science process, and to understand the importance of data science in countless industries.

**Learning Objectives**

* Install Anaconda on personal computer.
* Understand the Data Science Field.
* Become familiar with Descriptive and Inferential Statistics and statistical analysis.
* Learn the primary toolkit for data science in Python including NumPy, Pandas, Matplotlib and Scikit-learn.
* Learn how to perform exploratory data analysis.
* Learn the importance of data cleaning.
* Utilize common Machine Learning algorithms such as Linear and Logistic Regression.
* Learn how to evaluate models and choose the most effective one.
* Understand how to interpret a Confusion Matrix
* Understand the uses of the AUC-ROC curve in model evaluation.
* Solidify understanding by completing hands-on exercises and milestones.
* Create 2 data science projects.
* Understand the big picture and the importance of data science in business, industry, and technology.

**Course Outline**

* Course Introduction
* Installing Anaconda
* Overview of Data Science
* The Difference Between Business Analytics (BI), Data Analytics and Data Science
* The Field of Data Science
* The Data Science Process
  + Define the Problem
  + Get the Data
  + Explore the Data
  + Clean the Data
  + Model the Data
  + Communicate the Findings
* Descriptive Statistics Fundamentals
* Central Tendency
  + Mean
  + Median
  + Mode
* Spread of the Data
  + Variance
  + Standard Deviation
  + Range
* Relative Standing
  + Percentile
  + Quartile
  + Inter-quartile Range
* Inferential Statistics Fundamentals
  + Normal Distribution
  + Central Limit Theorem
  + Standard Error
  + Confidence Intervals
  + Other Distributions
  + Samples
  + Hypothesis Testing
* Milestone 1: Perform statistical analysis on a given data set.
* Essential Python Data Science Libraries
  + Numpy
  + Pandas
  + Matplotlib
  + Scikit-learn
  + Statsmodels
* Data Exploration
  + Describe
  + Merging
  + Grouping
  + Evaluating Features
* Data Visualization
* Line Chart
* Scatterplot
* Pairplot
* Histogram
* Density Plot
* Bar Chart
* Boxplot
* Customizing Charts
* Milestone 2: Perform Exploratory Data Analysis
* Data Cleaning
  + Dropping Rows
  + Imputing Missing Values
  + Feature Evaluating
* Feature Engineering
* Data Transformation
  + One-Hot Encoding
  + Standardization
  + Normalization
* Test/Train Split
* Model Training
* Machine Learning
* Linear Regression
* Logistic Regression
* Support Vector Machine
* Decision Tree
* K-Means
* Clustering
* Model Refining
  + Grid Search
  + K-Fold Cross Validation
  + Nested Cross Validation
  + Learning Curve
* Model Evaluation
  + Accuracy
  + Underfitting
  + Overfitting
  + Confusion Matrix
  + Sensitivity
  + Specificity
  + F1 Score
  + AUC ROC
* Milestone 3: Apply machine learning algorithms, select and refine the best model.
* Conclusion: Data Science in the real world, next steps.

**Structured Activity/Exercises/Case Studies**

Milestone Project 1: Perform a statistical analysis on a given data set.

Milestone Project 2: Perform Exploratory Data Analysis

Milestone Project 3: Apply machine learning algorithms, select and refine the best model.

**Training material provided:** Yes (Digital format)