

**Data Science and Data Engineering for Architects**

**Course Number:** PYTH-246WA
**Duration:** 4 days

**Overview**

This Data Engineering and Data Science for Architects training course teaches attendees how to use applied data science, business analytics, and data engineering to gain valuable insights from data. Participants learn how to incorporate standard machine learning algorithms, Python, and libraries such as NumPy, pandas, and Matplotlib in their data analytics work.

**Prerequisites**

Participants must have some experience with Python programming and be familiar with core statistical concepts, i.e., variance, correlation, etc.

**Materials**

All Data Engineering training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Anaconda Python 3.6 or later
* Spyder IDE and Jupyter notebook (Comes with Anaconda)

**Objectives**

* Use the CoLab Jupyter Notebook Environment
* Understand data visualization in Python
* Understand NumPy
* Perform data repairing
* Understand common metrics
* Coding kNN algorithm in NumPy (Optional)
* Understand machine learning datasets in scikit-learn
* Build linear regression models
* Perform Spam detection with Random Forest, Support Vector Machines, and Logistic Regression
* Compare classification algorithms
* Understand feature engineering and EDA
* UnderstandPCA

**Outline**

* Python for Data Science
	+ Python Data Science-Centric Libraries
	+ SciPy
	+ NumPy
	+ pandas
	+ Scikit-learn
	+ Matplotlib
	+ Seaborn
	+ Python Dev Tools and REPLs
	+ IPython
	+ Jupyter Notebooks
	+ Anaconda
* Data Visualization in Python
	+ Why Do I Need Data Visualization?
	+ Data Visualization in Python
	+ Getting Started with matplotlib
	+ A Basic Plot
	+ Scatter Plots
	+ Figures
	+ Saving Figures to a File
	+ Seaborn
	+ Getting Started with seaborn
	+ Histograms and KDE
	+ Plotting Bivariate Distributions
	+ Scatter Plots in seaborn
	+ Pair plots in seaborn
	+ Heatmaps
	+ A Seaborn Scatterplot with Varying Point Sizes and Hues
* Introduction to NumPy
	+ What is NumPy?
	+ The First Take on NumPy Arrays
	+ The ndarray Data Structure
	+ Understanding Axes
	+ Indexing Elements in a NumPy Array
	+ Re-Shaping
	+ Commonly Used Array Metrics
	+ Commonly Used Aggregate Functions
	+ Sorting Arrays
	+ Vectorization
	+ Vectorization Visually
	+ Broadcasting
	+ Broadcasting Visually
	+ Filtering
	+ Array Arithmetic Operations
	+ Reductions: Finding the Sum of Elements by Axis
	+ Array Slicing
	+ 2-D Array Slicing
	+ The Linear Algebra Functions
* Introduction to Pandas
	+ What is pandas?
	+ The DataFrame Object
	+ The DataFrame's Value Proposition
	+ Creating a pandas DataFrame
	+ Getting DataFrame Metrics
	+ Accessing DataFrame Columns
	+ Accessing DataFrame Rows
	+ Accessing DataFrame Cells
	+ Deleting Rows and Columns
	+ Adding a New Column to a DataFrame
	+ Getting Descriptive Statistics of DataFrame Columns
	+ Getting Descriptive Statistics of DataFrames
	+ Sorting DataFrames
	+ Reading From CSV Files
	+ Writing to a CSV File
* Repairing and Normalizing Data
	+ Repairing and Normalizing Data
	+ Dealing with the Missing Data
	+ Sample Data Set
	+ Getting Info on Null Data
	+ Dropping a Column
	+ Interpolating Missing Data in pandas
	+ Replacing the Missing Values with the Mean Value
	+ Scaling (Normalizing) the Data
	+ Data Preprocessing with scikit-learn
	+ Scaling with the scale() Function
	+ The MinMaxScaler Object
* Defining Data Science
	+ What is Data Science?
	+ Data Science, Machine Learning, AI?
	+ The Data Science Ecosystem
	+ Tools of the Trade
	+ The Data-Related Roles
	+ Data Scientists at Work
	+ Examples of Data Science Projects
	+ The Concept of a Data Product
	+ Applied Data Science at Google
	+ Data Science and ML Terminology: Features and Observations
	+ Terminology: Labels and Ground Truth
	+ Label Examples
	+ Terminology: Continuous and Categorical Features
	+ Encoding Categorical Features using One-Hot Encoding Scheme
	+ Example of 'One-Hot' Encoding Scheme
	+ Gartner's Magic Quadrant for Data Science and Machine Learning Platforms (a Labeling Example)
	+ Machine Learning in a Nutshell
	+ Common Distance Metrics
	+ The Euclidean Distance
	+ Decision Boundary Examples (Object Classification)
	+ What is a Model?
	+ Training a Model to Make Predictions
	+ Types of Machine Learning
	+ Supervised vs Unsupervised Machine Learning
	+ Supervised Machine Learning Algorithms
	+ Unsupervised Machine Learning Algorithms
	+ Which ML Algorithm to Choose?
	+ Bias-Variance (Underfitting vs Overfitting) Trade-off
	+ Underfitting vs Overfitting (a Regression Model Example) Visually
	+ ML Model Evaluation
	+ Mean Squared Error (MSE) and Mean Absolute Error (MAE)
	+ Coefficient of Determination
	+ Confusion Matrix
	+ The Binary Classification Confusion Matrix
	+ The Typical Machine Learning Process
	+ A Better Algorithm or More Data?
	+ The Typical Data Processing Pipeline in Data Science
	+ Data Discovery Phase
	+ Data Harvesting Phase
	+ Data Cleaning/Priming/Enhancing Phase
	+ Exploratory Data Analysis and Feature Selection
	+ Exploratory Data Analysis and Feature Selection Cont'd
	+ ML Model Planning Phase
	+ Feature Engineering
	+ ML Model Building Phase
	+ Capacity Planning and Resource Provisioning
	+ Communicating the Results
	+ Production Roll-out
	+ Data Science Gotchas
* Overview of the Scikit-learn Library
	+ The scikit-learn Library
	+ The Navigational Map of ML Algorithms Supported by scikit-learn
	+ Developer Support
	+ scikit-learn Estimators, Models, and Predictors
	+ Annotated Example of the LinearRegression Estimator
	+ Annotated Example of the Support Vector Classification Estimator
	+ Data Splitting into Training and Test Datasets
	+ Data Splitting in scikit-learn
	+ Cross-Validation Technique
* Classification Algorithms (Supervised Machine Learning)
	+ Classification (Supervised ML) Use Cases
	+ Classifying with k-Nearest Neighbors
	+ k-Nearest Neighbors Algorithm Visually
	+ Decision Trees
	+ Decision Tree Terminology
	+ Decision Tree Classification in the Context of Information Theory
	+ Using Decision Trees
	+ Properties of the Decision Tree Algorithm
	+ The Simplified Decision Tree Algorithm
	+ Random Forest
	+ Properties of the Random Forest Algorithm
	+ Support Vector Machines (SVMs)
	+ SVM Classification Visually
	+ Properties of SVMs
	+ Dealing with Non-Linear Class Boundaries
	+ Logistic Regression (Logit)
	+ The Sigmoid Function
	+ Logistic Regression Classification Example
	+ Logistic Regression's Problem Domain
	+ Naive Bayes Classifier (SL)
	+ Naive Bayesian Probabilistic Model in a Nutshell
	+ Bayes Formula
	+ Document Classification with Naive Bayes
* Unsupervised Machine Learning Algorithms
	+ PCA
	+ PCA and Data Variance
	+ PCA Properties
	+ Importance of Feature Scaling Visually
	+ Unsupervised Learning Type: Clustering
	+ Clustering vs. Classification
	+ Clustering Examples
	+ k-means Clustering
	+ k-means Clustering in a Nutshell
	+ k-means Characteristics
	+ Global vs. Local Minimum Explained
* Conclusion