

**Applied Machine Learning using Python and Apache Spark**

**Course Number:** PYTH-232  
**Duration:** 3 days

**Overview**

This Applied Machine Learning using Python and Apache Spark training teaches attendees Machine Learning (ML) concepts, terminology, and usage. Students learn how to perform and scale ML tasks using Python libraries (including NumPy, Pandas, Matplotlib, and Scikit-learn) on the Apache Spark platform.

**Prerequisites**

All attendees must have familiarity with Python. Having a working knowledge of Spark is a plus, but not required.

**Materials**

All Machine Learning training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Windows, Mac, or Linux with at least 8 GB RAM
  + Most class activities will create Spark code and visualizations in a browser-based notebook environment. The class also details how to export these notebooks and how to run code outside of this environment.
* A current version of Anaconda for Python 3.x
* Related lab files that Accelebrate will provide
* Internet access

**Objectives**

* Gain a basic understanding of Machine Learning
* Understand the differences between supervised and unsupervised learning
* Understand how to use Python libraries to explore, clean, and prepare data
* Describe the role of ML and where it fits into IT strategies
* Explain the technical and business drivers that result from using Machine Learning
* Understand techniques like classification, clustering, and regression
* Discuss how to identify which techniques should be applied for a specific use case
* Understand popular machine offerings, including Amazon Machine Learning, TensorFlow, Azure Machine Learning, Google Cloud, Spark MLlib, Python, R, and more
* Install and set up Anaconda
* Use Jupyter Notebooks
* Understand the popular Machine Learning algorithms, including linear regression, decision tree, logistic regression, K-nearest neighbor, K-means clustering, and more
* Use Python libraries like NumPy, Pandas, Matplotlib and Scikit-learn
* Understand Apache Spark Processing Framework and distributed architecture
* Compare Machine learning using Python versus Apache Spark
* Use Databricks cloud with Apache Spark MLlib

**Outline**

* Introduction
  + History and background of Machine Learning
  + Compare traditional programming to Machine Learning
  + Supervised and unsupervised learning overview
* Machine Learning Patterns
  + Classification
  + Clustering
  + Regression
* Gartner Hype Cycle for Emerging Technologies
  + Machine Learning offerings in the industry
  + Install and set up Anaconda
  + Descriptive statistics
  + Jupyter Notebooks
* Essential Libraries
  + NumPy
  + Pandas
  + Matplotlib
* Exploratory Data Analysis
* Getting Data
* Feature Selection
* Essential libraries
  + Scikit-learn
* Transforming Data
* Binary Encoding
* One-Hot Encoding
* Feature Engineering
* Algorithms
  + Linear regression
  + Naive Bayes
  + Decision tree
  + Random forest
  + Logistics regression
  + Support vector machine
  + K-nearest neighbor
  + K-means clustering
* Data Modeling
* Apache Spark Overview
  + Spark libraries
* Machine Learning using Python Versus using Spark
* Databricks Cloud Community Account Setup
* Measuring Performance
  + Confusion Matrix
  + ROC Curve, Area Under Curve (AUC)
* Refining the Model
* Hyperparameter Tuning
* Grid Search
* Spark MLlib
* Conclusion and Next Steps