# Introduction to Machine Learning With Apache Spark – 3 days

This course teaches doing Machine Learning with the popular Apache Spark framework.

This course is intended for data scientists and software engineers.

We assume no previous knowledge of Machine Learning – We teach popular Machine Learning algorithms from scratch.

**Please note the following:**

* This is an introductory to intermediate course.
* In-depth coverage of Math / Stats is behind Machine Learning is beyond the scope of this course.
* This course is taught using Spark & Python.
* Working knowledge of Spark is recommended for this course

# Objectives:

* Learn popular machine learning algorithms
* Practice the application of these algorithms using Apache Spark

# What you will learn

* ML Concepts
* Regressions
  + Linear Regression
  + Logistic Regressions
* Classifications
  + Naïve Bayes
  + SVM
  + Decision Trees
  + Random Forest
* Clustering (K-Means)
* Principal Component Analysis (PCA)
* Recommendations

# Duration

3 days

# Audience

Data analysts, Software Engineers

# Skill Level

Beginner to Intermediate

# Prerequisites

* Good programming background
* Working knowledge of Spark is essential for this course
* familiarity with Python would be a plus, but not required
* No machine learning knowledge is required

# Lab environment:

Working Spark environment in the cloud will be provided for students.

Students would run all the labs in the provided cloud environment.

# What to Bring

* A modern laptop
* The class would need open access to Internet.   
  Highly restrictive firewalls/VPNs might prevent the participant from connecting to the cloud environment.
* SSH client : Putty for Windows, Mac/Linux already has ssh clients installed

# Detailed Course Outline:

**Section 1: Machine Learning (ML) Overview**

* Machine Learning landscape
* Machine Learning applications
* Understanding ML algorithms & models (supervised and unsupervised)

**Section 2: ML in Python and Spark**

* Spark ML Overview
* Introduction to Jupyter notebooks
* Lab: Working with Jupyter + Python + Spark
* Lab: Spark ML utilities

**Section 3: Machine Learning Concepts**

* Statistics Primer
* Covariance, Correlation, Covariance Matrix
* Errors, Residuals
* Overfitting / Underfitting
* Cross validation, bootstrapping
* Confusion Matrix
* ROC curve, Area Under Curve (AUC)
* Lab: stats

**Section 4: Feature Engineering (FE)**

* Preparing data for ML
* Extracting features, enhancing data
* Data cleanup
* Visualizing Data
* Lab: data cleanup
* Lab: visualizing data

**Section 5: Linear regression**

* Simple Linear Regression
* Multiple Linear Regression
* Running LR
* Evaluating LR model performance
* Labs
* Use case : House price estimates

**Section 6: Logistic Regression**

* Understanding Logistic Regression
* Calculating Logistic Regression
* Evaluating model performance
* Labs
* Use case: credit card application, college admissions

**Section 7: Classification: SVM (Supervised Vector Machines)**

* SVM concepts and theory
* SVM with kernel
* Labs
* Use case: Customer churn data

**Section 8: Classification: Decision Trees & Random Forests**

* Classification and Regression Trees (CART) introduction
* Decision Tree concepts
* Pruning trees
* Gini index
* Bias Variance Tradeoff
* Random Forest concepts
* Random Forests features and examples
* Labs
* Use case: predicting loan defaults, estimating election contributions

**Section 9: Classification: Naive Bayes**

* Naïve Bayes theory
* Example walkthrough
* limitations
* Lab
* Use case: spam filtering

**Section 10: Clustering (K-Means)**

* Theory behind K-Means
* Running K-Means algorithm
* Estimating the performance
* Lab
* Use case: grouping uber trips data, grouping shopping data

**Section 11: Principal Component Analysis (PCA)**

* Understanding PCA concepts
* PCA applications
* Running a PCA algorithm
* Evaluating results
* Lab
* Use case: analyzing wine quality data

**Section 12: Recommendation (Collaborative filtering)**

* Recommender systems overview
* Collaborative Filtering concepts
* Lab
* Use case: movie recommendations, music recommendations

**Section 13: Final workshop (time permitting)**

Students will analyze a couple of datasets and run ML algorithms.

This is done as a group exercise. Each group will present their findings to the class.