



# Classical Machine Learning with Python

## Curriculum

1. Introduction to Data Science
    - a. What is Data Science
    - b. Applications of Data Science
    - c. Data Science Roles and Tools
    - d. Data Sources
    - e. Data Types
    - f. Data Storage and Retrieval
    - g. Data Pipelines
    - h. Data Preparation
    - i. Exploratory Data Analysis
    - j. Interactive Dashboards
  2. Introduction to Python
    - a. Python Installation and Anaconda distribution
    - b. Variables and Types – Introduction
      - i. Exercise I
      - ii. Exercise II
      - iii. Exercise III
      - iv. Exercise IV
      - v. Exercise V
    - c. Python Lists – Introduction
      - i. Exercise I
      - ii. Exercise II
      - iii. Exercise III
    - d. Sub-setting Lists – Introduction
      - i. Exercise I
      - ii. Exercise II
    - e. Manipulating Lists – Introduction
      - i. Exercise I
      - ii. Exercise II
-

- f. Functions – Introduction
    - i. Exercise I
    - ii. Exercise II
  - g. Methods – Introduction
    - i. Exercise I
    - ii. Exercise II
  - h. Package – Introduction
    - i. Exercise I
    - ii. Exercise II
3. NumPy
- a. NumPy – Introduction and Basics
    - i. Exercise – NumPy Basics
    - ii. Exercise – NumPy Side-effects
  - b. 2D – NumPy Array
    - i. Exercise
  - c. Basic Statistics
    - i. Exercise – Average and Median
    - ii. Exercise – Cricket Data
4. Matplotlib
- a. Matplotlib – Introduction
    - i. Exercise – Line Plot I
    - ii. Exercise – Line Plot II
    - iii. Exercise - Scatter plot
  - b. Matplotlib Histogram
    - i. Exercise - Building a histogram I
    - ii. Exercise - Building a histogram II
  - c. Matplotlib Customization
    - i. Exercise - Matplotlib Customization Labels
    - ii. Exercise - Matplotlib Customization Ticks
5. Dictionaries
- a. Dictionaries Part I
    - i. Exercise - Motivation for Dictionaries
    - ii. Exercise - Create Dictionaries
    - iii. Exercise - Access Dictionaries
  - b. Dictionaries Part II
    - i. Exercise – Dictionaries Manipulation I
    - ii. Exercise – Dictionaries Manipulation II

## 6. Pandas - Overview

- a. Pandas – Part I
  - i. Exercise – Dictionary to Dataframe
  - ii. Exercise – CSV to Dataframe
- b. Pandas – Part II
  - i. Exercise – Square brackets I
  - ii. Exercise – Square brackets II
  - iii. Exercise – LOC and ILOC – I
  - iv. Exercise – LOC and ILOC – II
  - v. Exercise – LOC and ILOC – III

## 7. Logical Operators

- a. Comparison Operator
  - i. Exercise – Equality
  - ii. Exercise – Greater Than and Less Than
- b. Boolean Operator
  - i. Exercise – And, OR, NOT
  - ii. Exercise – Boolean Operator with NumPy

## 8. Control Flow

- a. If, Else, Elif – Introduction
  - i. Exercise – If
  - ii. Exercise – Else
  - iii. Exercise – Elif

## 9. Filtering Pandas Dataframes

- a. Filtering Pandas Dataframes – Introduction
  - i. Exercise - Driving Right I
  - ii. Exercise - Driving Right II
  - iii. Exercise – Cars per Capita I
  - iv. Exercise – Cars per Capita II

## 10. Loops

- a. While loop – Introduction
  - i. Exercise I
  - ii. Exercise II
- b. For loop
  - i. Exercise - Loop over a list
  - ii. Exercise - Indexes and values I
  - iii. Exercise - Indexes and values II
  - iv. Exercise - Loop over list of lists
- c. Loop Data Structure - I

- i. Exercise - Loop over Dictionary
  - ii. Exercise - Loop over NumPy array
- d. Loop Data Structures - II
  - i. Exercise - Loop over Dataframe I
  - ii. Exercise - Loop over Dataframe II
  - iii. Exercise - Add column I
  - iv. Exercise - Add column II
- e. User-defined Functions
  - i. Exercise - Write a simple function
  - ii. Exercise - Single-parameter functions
  - iii. Exercise - Functions that return single values
- f. Multiple parameters and return values
  - i. Exercise - Functions with multiple parameters
  - ii. Exercise - A brief introduction to tuples
  - iii. Exercise - Functions that return multiple values
- g. Bringing it together
  - i. Exercise I
  - ii. Exercise II

## 11.Pandas

- a. Introduction to Pandas
- b. Transforming Data – Introduction to Dataframes
  - i. Exercise - Inspecting a Dataframe
  - ii. Exercise - Parts of a Dataframe
- c. Sorting and Sub-setting
  - i. Exercise - Sorting rows
  - ii. Exercise – Sub-setting columns
  - iii. Exercise – Sub-setting rows and categorical variables
- d. New columns
  - i. Exercise - Adding new columns
  - ii. Exercise - Combo Attack
- e. Summary Statistics
  - i. Exercise - Mean and Median
  - ii. Exercise - Summarizing Dates
  - iii. Exercise - Efficient summaries
  - iv. Exercise - Cumulative statistics
- f. Counting
  - i. Exercise - Dropping duplicates
  - ii. Exercise - Counting categorical variables
- g. Grouped summary statistics
  - i. Exercise - What percent of sales occurred at each store type

- ii. Exercise - Calculation with groupby ()
  - iii. Exercise - Multiple grouped summaries
- h. Pivot Table
  - i. Exercise - Pivoting on one variable
  - ii. Exercise - Fill in missing values and sum values with pivot tables
- i. Explicit indexes
  - i. Exercise - Setting & removing indexes
  - ii. Exercise - Sub-setting with. loc[]
  - iii. Exercise - Setting multi-level indexes
  - iv. Exercise - Sorting by index values
- j. Slicing and Sub-setting with. loc and. iloc
  - i. Exercise - Slicing index values
  - ii. Exercise - Slicing in both directions
  - iii. Exercise - Slicing time series
  - iv. Exercise – Sub-setting by row, column number
- k. Working with pivot tables
  - i. Exercise - Pivot temperature by city and year
  - ii. Exercise - Calculating on a pivot table
- l. Visualizing your Data
- m. Exercise - Which avocado size is most popular
- n. Exercise - Changes in sales over time
- o. Exercise - Avocado supply and demand
- p. Exercise - Price of conventional vs. organic avocados
- q. Missing values
- r. Exercise - Finding missing values
- s. Exercise - Removing missing values
- t. Exercise - Replacing missing values

## 12. Merging Dataframes Using Pandas

- a. Preparing Data – Reading Multiple Data Files
  - i. Exercise - Reading Dataframes from multiple files
  - ii. Exercise - Reading Dataframes from multiple files in a loop
  - iii. Exercise - Combining Dataframes from multiple data files
- b. Re-indexing Dataframes
  - i. Exercise - Sorting Dataframe with the Index & columns
  - ii. Exercise – Re-indexing Dataframe from a list

- c. Arithmetic with Series & Dataframes
  - i. Exercise - Broadcasting in arithmetic formulas
  - ii. Exercise - Computing percentage growth of GDP
  - iii. Exercise - Converting currency of stocks
- d. Concatenating Data - Appending and concatenating Series
  - i. Exercise - Appending Pandas Series
  - ii. Exercise - Concatenating Pandas Series along row axis
- e. Appending and concatenating Dataframes
  - i. Exercise - Concatenating Pandas Dataframes along column axis
  - ii. Exercise - Reading multiple files to build a Dataframe
- f. Concatenation, keys, and Multi-indexes
  - i. Exercise - Concatenating vertically to get Multi-indexed rows
  - ii. Exercise - Slicing Multi-indexed Dataframes
  - iii. Exercise - Concatenating Dataframes from a dictionary
- g. Outer and Inner Joins
  - i. Exercise - Concatenating Dataframes with inner join
  - ii. Exercise - Resampling & concatenating Dataframes with inner join
- h. Merging Dataframes – Overview
  - i. Exercise - Merging on a specific column
  - ii. Exercise - Merging on columns with non-matching labels
  - iii. Exercise - Merging on multiple columns
- i. Joining Dataframes – Overview
  - i. Exercise - Left & right merging on multiple columns
  - ii. Exercise - Merging Dataframes with outer join
- j. Ordered merges – Overview
  - i. Exercise - Using merge\_ordered()

### 13. Introduction to Data Visualisation Using Matplotlib

- a. Introduction to data visualization with Matplotlib
  - i. Exercise - Using the matplotlib.pyplot interface
  - ii. Exercise - Adding data to an Axes object
- b. Customizing your plots
  - i. Exercise - Customizing data appearance
  - ii. Exercise - Customizing axis labels and adding titles
- c. Small multiples
  - i. Exercise - Creating small multiples with plt.subplots

- ii. Exercise - Small multiples with shared y axis
  - d. Plotting time-series data
    - i. Exercise - Read data with a time index
    - ii. Exercise - Plot time-series data
    - iii. Exercise - Using a time index to zoom in
  - e. Plotting time-series with different variables
    - i. Exercise - Plotting two variables
    - ii. Exercise - Defining a function that plots time-series data
    - iii. Exercise - Using a plotting function
  - f. Annotating time-series data
    - i. Exercise - Annotating a plot of time-series data
    - ii. Exercise - Plotting time-series: putting it all together
  - g. Quantitative comparisons: bar-charts
    - i. Exercise - Bar chart
    - ii. Exercise - Stacked bar chart
  - h. Quantitative comparisons: histograms
    - i. Exercise - Creating histograms
    - ii. Exercise - "Step" histogram
  - i. Statistical plotting
    - i. Exercise - Adding error-bars to a bar chart
    - ii. Exercise - Adding error-bars to a plot
    - iii. Exercise - Creating boxplots
  - j. Quantitative comparisons: scatter plots
    - i. Exercise - Simple scatter plot
    - ii. Exercise - Encoding time by colour
  - k. Preparing your figures to share with others
    - i. Exercise - Switching between styles
  - l. Saving your visualizations
    - i. Exercise - Saving a file several times
  - m. Automating figures from data
    - i. Exercise - Unique values of a column
- 14. Data Visualization with Seaborn
  - a. Introduction to Seaborn
    - i. Exercise I - Making a scatter plot with lists
    - ii. Exercise II - Making a count plot with a list
  - b. Using pandas with Seaborn
    - i. Exercise I - "Tidy" vs. "untidy" data
    - ii. Exercise II - Making a count plot with a Dataframe
  - c. Adding a third variable with hue
    - i. Exercise I - Hue and scatter plots
    - ii. Exercise II - Hue and count plots
  - d. Visualizing Two Quantitative Variables - Introduction to relational plots and subplots

- i. Exercise I - Creating subplots with col and row
    - ii. Exercise II - Creating two-factor subplots
  - e. Customizing scatter plots
    - i. Exercise I - Changing the size of scatter plot points
    - ii. Exercise II - Changing the style of scatter plot points
  - f. Introduction to line plots
    - i. Exercise I- Interpreting line plots
    - ii. Exercise II - Visualizing standard deviation with line plots
    - iii. Exercise III - Plotting subgroups in line plots
  - g. Visualizing a Categorical and a Quantitative Variable - Count plots and bar plots
    - i. Exercise I - Count plots
    - ii. Exercise II - Customizing bar plots
  - h. Box plots
    - i. Exercise I - Create and interpret a box plot
    - ii. Exercise II - Omitting outliers
    - iii. Exercise III - Adjusting the whiskers
    - iv. Point plots
    - v. Exercise I - Customizing point plots
    - vi. Exercise II - Point plots with subgroups
  - i. Customizing Seaborn Plots - Changing plot style and color
    - i. Exercise I - Changing style and palette
    - ii. Exercise II - Changing the scale
    - iii. Exercise III - Using a custom palette
  - j. Adding titles and labels: Part 1
    - i. Exercise I - FacetGrids vs. Axes Sub-plots
    - ii. Exercise II - Adding a title to a FacetGrid object
  - k. Adding titles and labels: Part 2
    - i. Exercise I - Adding a title and axis labels
    - ii. Exercise II - Rotating x-tick labels
  - l. Putting it all together
    - i. Exercise I - Box plot with subgroups
    - ii. Exercise II - Bar plot with subgroups and subplots
- 15.Descriptive Statistics Using Python
  - a. Interpreting Data Using Descriptive Statistics with Python – Introduction
  - b. Measure of Central Tendency – Mean, Median, Mode
    - i. Exercise – Mean, Median, Mode
  - c. Measures of Dispersion & Understanding Variance
    - i. Exercise - Computing IQR, Variance and Standard Deviation
  - d. Gaussian Normal Distribution



- e. Confidence Intervals
- f. Skewness and kurtosis
  - i. Exercise - Skewness and kurtosis
- g. Covariance and Correlation
  - i. Exercise - Covariance and Correlation
- h. Summary

## 16. Inferential Statistics Using Python

- a. Module Overview
- b. Prerequisites and Course Outline
- c. Descriptive Statistics to Summarize Data
- d. Hypothesis Testing - Introduction
  - i. Lady Tasting Tea
  - ii. The Power, Alpha and p-value of a Statistical Test
  - iii. t-test and z-test
  - iv. One Sample Location t-test
  - v. z-test
  - vi. Other types of t-test
  - vii. Demo - Preparing Data for Hypothesis Testing
  - viii. Demo Performing the t-test
  - ix. Demo Performing the Paired Difference t-test
- e. ANOVA – Analysis of Variance - Introduction
  - i. Demo – One-way ANOVA
- f. Two-way ANOVA - Introduction
  - i. Demo - Two-way ANOVA
- g. Pearson's Chi-Square Test - Introduction
  - i. Demo Chi-square Test
- h. Module Summary

## 17. Introduction to Importing Data in Python (0.5 hours)

- a. Chapter 1: Introduction and flat files
- b. Feature Engineering for Machine Learning in Python
- c. Creating Features
- d. Creating Features
  - i. Getting to know your data
  - ii. Selecting specific data types
- e. Dealing with categorical features
  - i. One-hot encoding and dummy variables
  - ii. Dealing with uncommon categories
- f. Numeric variables
  - i. Binarizing columns

- ii. Binning values
- g. Dealing with Messy Data
- h. Why do missing values exist?
  - i. How sparse is my data?
  - ii. Finding the missing values
- i. Dealing with missing values (I)
  - i. Listwise deletion
  - ii. Replacing missing values with constants
- j. Dealing with missing values (II)
  - i. Filling continuous missing values
- k. Dealing with other data issues
  - i. Dealing with stray characters (I)
  - ii. Dealing with stray characters (II)
  - iii. Method chaining
- l. Conforming to Statistical Assumptions

#### 18.Data distributions

- a. What does your data look like? (I)
- b. What does your data look like? (II)

#### 19.Scaling and transformations

- a. Normalization
- b. Standardization
- c. Log transformation

#### 20.Removing outliers

- a. Percentage based outlier removal
- b. Statistical outlier removal

#### 21.Scaling and transforming new data

- a. Train and testing transformations (I)
- b. Train and testing transformations (II)

#### 22.Dealing with Text Data

#### 23.Encoding text

- a. Cleaning up your text
- b. High level text features

#### 24.Word counts

- a. Counting words (I)
- b. Counting words (II)
- c. Limiting your features
- d. Text to DataFrame

#### 25.Term frequency-inverse document frequency

- a. Tf-idf

- b. Inspecting Tf-idf values
  - c. Transforming unseen data
- 26.N-grams
  - a. Using longer n-grams
  - b. Finding the most common words
  
- 27.Exploratory Data Analysis in Python (3-4 hours)
  - a. Chapter 1: Read, clean, and validate
  - b. Chapter 2: Distributions
  - c. Chapter 3: Relationships
  - d. Chapter 4: Multivariate Thinking
  
- 28.Machine Learning for Everyone
  - a. What is machine learning?
  - b. Machine learning concepts
  - c. Machine learning workflow
  
- 29.Machine Learning Models
  - a. Supervised learning
  - b. Evaluating performance
  - c. Improving performance
  
- 30.Deep Learning
  - a. Deep learning
  - b. Computer vision
  - c. Natural Language Processing
  - d. Limits of machine learning
  
- 31.Supervised Learning with scikit-learn
  - a. Introduction
    - i. Supervised learning
    - ii. Exploratory data analysis
  - b. Classification
    - i. The classification challenges
    - ii. k-Nearest Neighbors: Fit
    - iii. k-Nearest Neighbors: Predict
  - c. Measuring model performance
    - i. The digits recognition dataset
    - ii. Train/Test Split + Fit/Predict/Accuracy
    - iii. Overfitting and underfitting

- iv. Regression
- d. Introduction to regression
  - i. Importing data for supervised learning
- e. The basics of linear regression
  - i. Fit & predict for regression
  - ii. Train/test split for regression
- f. Cross-validation
  - i. 5-fold cross-validation
  - ii. K-Fold CV comparison
- g. Regularized regression
  - i. Regularization I: Lasso
  - ii. Regularization II: Ridge
  - iii. Fine-tuning your model
- h. How good is your model?
  - i. Metrics for classification
- i. Logistic regression and the ROC curve
  - i. Building a logistic regression model
  - ii. Plotting an ROC curve
- j. Area under the ROC curve
  - i. AUC computation
- k. Hyperparameter tuning
  - i. Hyperparameter tuning with GridSearchCV
  - ii. Hyperparameter tuning with RandomizedSearchCV
- l. Hold-out set for final evaluation
  - i. Hold-out set in practice I: Classification
  - ii. Hold-out set in practice II: Regression
- m. Pre-processing and pipelines

### 32. Dimensionality Reduction and Unsupervised Learning

- a. How to do PCA Dimensionality Reduction
- b. Implement PCA in python
- c. How to do TSNE Dimensionality Reduction
- d. Implement TSNE on a Kaggle dataset
- e. What is K-Means Clustering
- f. Solve a "Kaggle Competition" using k-Means
- g. What is hierarchical Clustering
- h. Implement hierarchical clustering on a Kaggle dataset

### 33. Machine Learning with Tree-Based Models in Python

- a. What is Decision Tree Classification?

- b. Implement Decision Tree Classification, using a Kaggle dataset
- c. What is Decision Tree Regression?
- d. Implement Decision Tree Regression, using a Kaggle dataset
- e. What is Random Forest?
- f. Implement Random Forest Regression, using a Kaggle dataset
- g. Implement Random Forest Classification, using a Kaggle dataset

#### 34. Extreme Gradient Boosting with XGBoost

- a. Classification with XGBoost
- b. Regression with XGBoost
- c. Fine-tuning your XGBoost model
- d. Using XGBoost in pipelines

#### 35. Machine Learning for Time Series Data in Python

- a. Time Series and Machine Learning Primer
- b. Time Series as Inputs to a Model
- c. Predicting Time Series Data
- d. Validating and Inspecting Time Series Models