

Course outline for Data Science and Predictive Analysis

FA-0488

Software Development · Data & Analytics

Introduction

This course has been designed to cater to the needs of learners who are comfortable with python and picking up a specialization of python as an additional concentrated skill. This course will also delve into the various usages of python which in real business case studies. This course aims to alleviate the unnecessary jargon and cover the parts that are indeed needed.

This course shall also enable the user to have the necessary foundation to take their training to a more advanced level should they wish to do so.

This course is designed for those who are experienced in programming with python. For absolute beginners, a Python beginners course is recommended prior to attempting this course.

Learning outcome:

1. Overview of python basics.
2. IDEs of python.
3. Understand the basic syntax structures.
4. Get updated on recent syntax changes in python 3.8 and 3.9 respectively.
5. Advanced Data handling in python.
6. Working with external libraries in python.
7. Working with files and encoding.
8. Understand different forms of data acquisition.
9. Data Extraction and conversion.
10. Learn statistical analytics of data.

11. Use Python to get the basics of Data Analytics.
12. Use simple scripting to present data.
13. Have the ability to visualize data and manipulate them using simple python.
14. Understand how data science delves into Machine Learning.
15. Machine learning and coding in python using Scikit Learn
16. Machine learning and coding in python using XGBoost
17. Using machine learning for predictions and analytics.
18. Data Acquisition and scraping.

Training outline

1. Day 1: Statistics and Probability and tools to use for Data Analytics

2. Python Revision

3. A Crash Course in matplotlib.

4. Advanced Visualization with Seaborn.

5. Covariance and Correlation.

6. Conditional Probability.

7. Bayes' Theorem.

8. Day 2: Real hands on Data Analytics

9. Basic Machine Learning (using Python).

10. Supervised vs. Unsupervised Learning, and Train/Test.

11. Using Train/Test to Prevent Overfitting a Polynomial Regression.

12. Bayesian Methods: Concepts.

13. Practice with:

14. Implementing a Spam Classifier with Naive Bayes.

15. K-Means Clustering.

16. Entropy

17. Decision Trees Assessment,

18. Day 3: Dealing with Real-World Data

19. Bias/Variance Tradeoff

20. K-Fold Cross-Validation to avoid overfitting

21. Data Cleaning and Normalization

22. Cleaning web log data

23. Normalizing numerical data

24. Detecting outliers

25. Feature Engineering

26. Imputation Techniques for Missing Data

27. Handling Unbalanced Data: Oversampling,

28. Undersampling, and SMOTE

- 29. Binning, Transforming, Encoding, Scaling, and Shuffling**
- 30. Day 4: Machine Learning on Big Data**
- 31. Introducing MLlib**
- 32. Introduction to Decision Trees in Spark**
- 33. K-Means Clustering in Spark**
- 34. TF / IDF**
- 35. Searching Wikipedia with Spark**
- 36. Using the Spark with python brief**
- 37. Day 5: E-Commerce and ML in the Real World**
- 38. Deploying Models to Real-Time Systems**
- 39. Testing Concepts**
- 40. T-Tests and P-Values**
- 41. Hands-on With T-Tests**
- 42. Determining How Long to Run an Experiment**
- 43. A/B Test Gotchas**