Starjite Institute of Technology

Comprehensive Course Outline for a Data Science and Machine Learning Course using Python

Compiled By:

Engr. John Stanley

Email Address:

johnstanley669@gmail.com

Module 1: Introduction to Data Science and Python

Overview of Data Science

The Data Science Process (Data Collection, Cleaning, Analysis, Visualization, and Interpretation)

Real-world Applications of Data Science

Overview of Machine Learning

Setting Up Python Environment

- Installing Anaconda and Jupyter Notebook
- Python IDEs (VS Code, PyCharm)
- Introduction to Google Colab for online coding

Python for Data Science

- Overview of Python Basics (Syntax, Data Types, Variables)
- Control Flow: Conditionals and Loops
- Functions and Modules in Python

Module 2: Data Handling and Manipulation using Python

- Introduction to NumPy
- Arrays and Matrix Operations
- Broadcasting and Element-Wise Operations
- Working with Missing Data
- Data Manipulation with Pandas
- Pandas DataFrames and Series
- Importing and Exporting Data (CSV, Excel, SQL)
- Data Cleaning and Preprocessing
- Handling Missing Values, Duplicates, and Outliers
- Grouping, Aggregating, and Pivot Tables

Advanced Pandas

- Merging, Joining, and Concatenation of DataFrames
- Working with Time Series Data
- String Manipulation in Pandas

Module 3: Data Visualization

- Introduction to Data Visualization
- Importance of Data Visualization in Data Science
- Data Storytelling and Insights
- Matplotlib and Seaborn for Visualization
- Plotting Line Graphs, Bar Charts, Histograms, and Scatter Plots
- Heatmaps, Pairplots, and Customizing Plots

Advanced Visualizations

- Plotly for Interactive Visualizations
- Dashboard Creation with Plotly Dash
- Data Visualization Best Practices

Module 4: Introduction to Statistics and Probability

Descriptive Statistics in Python

- Central Tendency: Mean, Median, Mode
- Variability: Standard Deviation, Variance, Range
- Correlation and Covariance

Probability Basics

- Probability Theory and Distributions
- Conditional Probability and Bayes' Theorem

Inferential Statistics

- Hypothesis Testing (T-tests, Chi-Square Tests, ANOVA)
- Confidence Intervals and P-Values

Module 5: Machine Learning Fundamentals with Scikit-Learn

Introduction to Machine Learning

- Types of Machine Learning: Supervised, Unsupervised, and Reinforcement Learning
- Steps in Building a Machine Learning Model
- Introduction to Scikit-Learn

Supervised Learning

- Linear and Polynomial Regression
- Classification Algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees
- Random Forest and Gradient Boosting Algorithms
- Support Vector Machines (SVM)

Unsupervised Learning

- K-Means Clustering
- Hierarchical Clustering
- Principal Component Analysis (PCA)
- Anomaly Detection

Module 6: Deep Learning with TensorFlow and Keras

Introduction to Neural Networks

- Biological Inspiration and Neural Network Basics
- Perceptron, Activation Functions (ReLU, Sigmoid, Tanh)

Building Neural Networks with Keras

Introduction to Keras

- Creating and Training a Neural Network
- Model Evaluation and Optimization (Loss Function, Optimizers)

Convolutional Neural Networks (CNNs)

- CNN Architecture (Convolution, Pooling, Fully Connected Layers)
- Image Classification using CNNs

Recurrent Neural Networks (RNNs)

- Understanding Sequential Data
- Introduction to RNNs and LSTM (Long Short-Term Memory) Networks

Module 7: Model Evaluation and Hyperparameter Tuning

Model Evaluation Metrics

- Classification Metrics: Confusion Matrix, Accuracy, Precision, Recall, F1-Score, ROC-AUC
- Regression Metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2
- Cross-Validation and Train-Test Split

Hyperparameter Tuning

- Grid Search and Random Search
- Overfitting vs. Underfitting
- Regularization Techniques (L1, L2 Regularization)

Model Saving and Loading

- Saving and Loading Models with Scikit-Learn and Keras

Module 8: Time-Series Analysis and Forecasting

Introduction to Time-Series Data

- Components of Time-Series Data: Trend, Seasonality, Noise
- Stationarity and Differencing

Time-Series Models

- Moving Averages and Exponential Smoothing
- Autoregressive Integrated Moving Average (ARIMA)
- Seasonal ARIMA (SARIMA)

Evaluation of Time-Series Models

- Time-Series Cross-Validation
- Metrics: RMSE, MAE, MAPE

Module 9: Natural Language Processing (NLP) with Python

Introduction to NLP

- Text Preprocessing: Tokenization, Stemming, Lemmatization
- Removing Stop Words, Handling Punctuation
- Bag of Words (BoW) and TF-IDF

Advanced NLP Techniques

- Word Embeddings: Word2Vec, GloVe
- Sentiment Analysis with Python
- Topic Modeling (LDA)

Text Classification

- Spam Detection
- Named Entity Recognition (NER)

- Document Classification

Module 10: Big Data and Scalable Machine Learning

Introduction to Big Data

- Challenges in Handling Big Data
- Hadoop and Spark Overview

Machine Learning with Apache Spark (PySpark)

- Introduction to PySpark
- Working with Spark DataFrames
- Machine Learning with Spark MLlib

Module 11: Model Deployment with Python

Deploying Machine Learning Models

- Model Deployment Strategies: Local, Cloud, Edge
- Using Flask/Django to Deploy Models as Web Applications

Containerization with Docker

- Introduction to Docker for Data Science
- Deploying Models with Docker Containers

Introduction to MLOps

- Continuous Integration and Continuous Deployment (CI/CD) for ML
- Model Monitoring and Retraining

Module 12: Capstone Project

End-to-End Data Science Project

- Problem Definition and Dataset Selection
- Data Preprocessing and Feature Engineering

- Model Selection, Evaluation, and Optimization
- Model Deployment and Presentation

Additional Resources

Key Python Libraries

- Scikit-Learn, Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, Keras

Reading Materials

- Key Books, Articles, and Research Papers on Data Science and Machine Learning