Microsoft Data Science Bootcamp for Freshers

==========

Duration: 90 hours
Courseware: Unofficial PPT/PDF + AI-900 Official Courseware + AI-102 Official Courseware
Labs: Koenig DC + DP-100 LOD

==========

Module 01: Introduction to Python (18 hours)

- Introduction
- Data Types
- Variables
- Decision Control Files
- Operators
- List, Tuples, Sets, Dictionary
- Functions and Methods
- File Handling
- Module
- String
- Iterators and Generators
- Regular Expressions
- OO Programming Concepts
- NumPy
- Pandas

Module 02: Machine Learning Essentials – 40 hours

2.1 Introduction to Data Science & Machine Learning

- Need for Data Science and Machine Learning
- Types of Analytics
- Lifecycle of a Data Science project
- Skills for a Data Scientist role
- Types of Machine Learning

2.2 Python for Data Analysis & Pre-processing

- Python Libraries – NumPy, Pandas, matplotlib, Seaborn, scikit-learn, TensorFlow, Keras, PyTorch
- Exploratory Data Analysis (EDA)
- Data Cleaning Techniques, Handling Missing Data, Handling Categorical Data
- Introduction to EDA, 2D Scatter-plot, 3D Scatter-plot, Pair plots
- Univariate, Bivariate, and Multivariate Analysis, Box-plot
- Need for Data Pre-Processing
- Handling Missing Values
- Label-Encoding for Categorical Data
- One-Hot Encoding for Categorical Data Explained
- Need for Data Transformation
- Concept of Data Normalization
- Data Normalization Techniques - Standard Scalar & Minmax
- Train, Test & Validation of Data

2.3 Supervised Machine Learning – Regression

- Simple Linear Regression
- Concept of Linear Regression
- Ordinary Least Squares and Regression Errors
- Data Processing & Train and Test of Model
- Model Evaluation Parameters like R-squared, Score, RMSE, and their Interpretations
- Prediction Plot & its Interpretation
- Hands-on Problem
- Concept of Multiple Linear Regression
- Degrees of Freedom
- Adjusted R-Squared
- Assumptions of Multiple Linear Regression - Linearity, Multicollinearity, Autocorrelation, Endogeneity, Normality of Residuals, Homoscedasticity, etc.
- Concept of time-lag data in Autocorrelation
- Concept of Dummy Variable Trap
- Hands-on Problem

2.4 Supervised Machine Learning – Classification

- Concept of Logistic Regression
- Concept of Stratification
- Concept of Confusion Matrix
- Hands-on Problem
- Common Sensical Intuition of SVM
- Mathematical Intuition of SVM
- Different types of SVM Kernel Functions
- Hands-on Problem (Preferred: IRIS Classification Problem)
- Intuition Behind Decision Trees
- Optimal Model Selection Criterion in Decision Tree
- Hands-on Problem
- Ensemble Learning and Random Forests
- Bagging and Boosting

- Hands-on Problem
- Need for Evaluation and Accuracy Paradox
- Different Measures for Classification Models - Accuracy, Precision, Recall, F1 Score, etc.
- Threshold and Adjusting Thresholds
- AUC ROC Curve
- Hands-on Problem

2.5 Feature Selection and Dimensionality Reduction
- Feature Selection Importance
- Concept of Univariate Feature Selection
- F-Test for Regression and Classification
- Hands-on F-test (p-value analysis)
- Chi-Squared for Classification
- Feature Selection Techniques - SelectKBest, SelectPercentile & Generic Univariate Select
- Hands-on Chi-Squared (p-value analysis)
- Concept of Recursive Feature Elimination (RFE)
- Feature Importance Score/Feature Ranking
- Hands-on RFE
- Need to Reduce Dimensions and Importance of PCA
- Mathematical Intuition of PCA & Steps to Calculate PCA
- Hands-on PCA (Model Comparisons with PCA & without PCA recommended)

2.6 Cross Validation & Hyperparameter Tuning
- Importance of Cross Validation
- Parameter & Implementation of Cross Validation
- Hands-on Problem (Drawing inference from results)
- Concept of Hyperparameter Tuning
- Grid Search & Randomized Search
- Hands-on GridSearchCV (Analyze results)

2.7 Supervised Machine Learning – Natural Language Processing (NLP)
- Introduction to NLP
- Basic Concepts of NLP: Tokenization, Stop Words, Stemming, Lemmatization, etc.
- TF-IDF Vector and its Mathematical Intuition
- Recommendation System Example

2.8 Unsupervised Machine Learning – Clustering
- Introduction to Clustering
- Mathematical Intuition Behind Cluster Formation
- Elbow Method & Its Mathematical Intuition
- K-Means Clustering Implementation (Numerical)
- K-Means Clustering Implementation (Natural Language Processing)
- Introduction to Clustering

2.9 Introduction to Deep Learning

- Need & Applications of Deep Learning
- Working of Artificial Neural Network
- Backend (TensorFlow) & Frontend (Keras)
- Concept of Tensor
- Keras Model Building Overview - Construct, Compile & Evaluate
- Activation Function
- Loss Functions
- Optimization Techniques
- Evaluation Metrics for Deep Learning

Module 3: Designing and Implementing a Data Science Solution on Azure (DP-100) – 32 hours

3.1 Design a machine learning solution

- Design a data ingestion solution for machine learning projects
- Design a machine learning model training solution
- Design a model deployment solution
- Design a machine learning operations (MLOps) solution

3.2 Explore and configure the Azure Machine Learning workspace

- Explore the Azure Machine Learning workspace resources and assets
- Explore developer tools for workspace interaction
- Make data available in Azure Machine Learning
- Work with compute resources in Azure Machine learning
- Work with environments in Azure Machine Learning

3.3 Experiment with Azure Machine Learning

- Explore Automated Machine Learning
- Find the best classification model with Automated Machine Learning
- Track model training in notebooks with MLflow

3.4 Optimize model training with Azure Machine Learning

- Run a training script as a command job in Azure Machine Learning
- Track model training with MLflow in jobs
- Perform hyperparameter tuning with Azure Machine Learning
- Run pipelines in Azure Machine Learning

3.5 Manage and evaluate models in Azure Machine Learning

- Register an MLflow model in Azure Machine Learning
- Create and explore the Responsible AI dashboard

3.6 Deploy and consume models with Azure Machine Learning

- Deploy a model to a managed online endpoint
- Deploy a model to a batch endpoint