Data Science Syllabus

DATA SCIENCE with Python

Module 1: Introduction to Data Science

- **What is Data Science?**
  - Overview of data science, data analysis, and machine learning.
  - Applications of data science across industries.
- **Python Basics for Data Science**
  - Introduction to Python programming (variables, data types, functions).
  - Setting up the environment: Jupyter notebooks, Python IDEs.
  - Libraries for data science: **NumPy**, **Pandas**, **Matplotlib**, **Seaborn**, **SciPy**.

Module 2: Python Fundamentals for Data Science

- **NumPy for Numerical Computation**
  - Arrays, matrices, and multi-dimensional arrays.
  - Indexing and slicing.
  - Vectorized operations and broadcasting.
- **Pandas for Data Manipulation**
  - DataFrames and Series.
  - Importing/exporting data (CSV, Excel, SQL).
  - Handling missing values and duplicates.
  - Data aggregation, groupby operations, and merging datasets.

Module 3: Data Exploration and Visualization

- **Data Cleaning and Preprocessing**
  - Handling missing data: imputation vs. removal.
  - Data transformation: normalization and standardization.
  - Encoding categorical variables: one-hot encoding, label encoding.
- **Data Visualization**
  - **Matplotlib** and **Seaborn** for static plots.
  - Types of plots: bar, line, histogram, scatter, boxplot, etc.
  - Plotting time series, distributions, and relationships.
  - Customizing plots (labels, legends, themes).

Module 4: Statistics for Data Science

- **Descriptive Statistics**
  - Mean, median, mode, variance, standard deviation.
  - Data distribution and percentiles.
- **Inferential Statistics**
  - Probability theory basics.
  - Hypothesis testing (t-tests, chi-square tests, p-values).
  - Confidence intervals and significance levels.
- **Correlation and Covariance**
  - Understanding correlation vs. causation.
  - Pearson's correlation coefficient.

Module 5: Data Modeling and Machine Learning

- **Introduction to Machine Learning**
  - Supervised vs unsupervised learning.
  - Key machine learning algorithms: regression, classification, clustering.
- **Supervised Learning**

- **Linear Regression**: Simple and multiple linear regression.
- **Logistic Regression**: For binary classification.
- **K-Nearest Neighbors (KNN)**: Classification and regression.
- **Decision Trees** and **Random Forest**: Building and evaluating models.
- Model evaluation: confusion matrix, accuracy, precision, recall, F1-score.

- **Unsupervised Learning**
  - **K-Means Clustering**: Cluster analysis.
  - **Hierarchical Clustering**.
  - **Principal Component Analysis (PCA)**: Dimensionality reduction.

Module 6: Model Evaluation and Selection

- **Cross-Validation**
  - K-fold cross-validation and train-test split.
  - Bias-variance trade-off.
- **Hyperparameter Tuning**
  - Grid Search and Random Search.
  - Model selection and fine-tuning.
- **Overfitting and Underfitting**
  - Identifying and mitigating overfitting.
  - Regularization: L1, L2 regularization (Ridge, Lasso).

Module 7: Advanced Topics in Data Science

- **Deep Learning with Python**
  - Introduction to Neural Networks.
  - Using **TensorFlow** or **Keras** for basic neural networks.
  - Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).
- **Natural Language Processing (NLP)**
  - Text processing: tokenization, stemming, lemmatization.
  - **Bag of Words** and **TF-IDF** techniques.
  - Text classification with machine learning.
  - Sentiment analysis and topic modeling.

Module 8: Big Data and Advanced Tools

- **Big Data Tools**
  - Introduction to Hadoop and Spark for large-scale data processing.
  - Using **PySpark** with Python for distributed computing.
- **Database Management**
  - SQL Basics and querying databases with Python.
  - Working with databases using **SQLAlchemy** and **Pandas**.

Module 9: Projects and Case Studies

- **End-to-End Data Science Project**
  - From data collection to cleaning, analysis, and visualization.
  - Applying machine learning models to real-world datasets.
- **Capstone Project**
  - Working on a real-world dataset (e.g., Kaggle datasets).
  - Presenting results and insights through visualizations and a written report.

Tools and Libraries Covered in the Course:

- **Pandas**: Data manipulation and analysis.
- **NumPy**: Numerical computing.

- **Matplotlib** and **Seaborn**: Data visualization.
- **Scikit-learn**: Machine learning library.
- **TensorFlow/ Keras**: Deep learning.
- **SQLAlchemy**: Database connection and manipulation.
- **PySpark**: Big data processing.

Assessment and Learning Activities:

- Quizzes and assignments after each module.
- Hands-on projects using real-world datasets.
- Weekly practice problems and coding challenges.
- Collaborative group projects or presentations.
