Big Data Analytics

(Creating Best of Minds)

Professional Certificate in Big Data Analytics

Course Code: IT_CT_I_0007/25

Duration: 40 Hours

Delivery Format: Hybrid

Target Audience:

- Who is new to Data Science and want to build foundational skills.
- Have basic programming knowledge and want to apply it to data tasks.
- Seek practical experience with Python and machine learning.
- Aspire to be Data Scientists, Data Analysts, or transition into data science from other tech fields.
- No prior data science experience is required, but basic programming knowledge is helpful.

Program Objectives:

- Understand the fundamentals of Data Science and its real-world applications.
- Gain proficiency in Python and libraries like NumPy and pandas for data analysis.
- Master data visualization techniques using Matplotlib and Seaborn.
- Build and evaluate machine learning models.
- Learn advanced machine learning techniques and model deployment basics.
- Apply skills in a capstone project.

Detailed Syllabus

Module 1: Introduction to Data Science & Python for Data Science

Objective: To introduce the core concepts of Data Science and provide a foundation in Python programming for data analysis.

Topics:

Introduction to Data Science:

- What is Data Science?
- Applications of Data Science in real-world scenarios
- Career pathways in Data Science
- Key roles in Data Science: Data Analyst, Data Engineer, Data Scientist

Python Basics for Data Science:

- Introduction to Python and Jupyter Notebooks
- Data types, variables, loops, conditionals
- Functions and error handling
- Essential Python libraries: NumPy, pandas

Data Structures & Data Wrangling:

- Working with lists, tuples, dictionaries
- Introduction to pandas for data manipulation
- Loading, cleaning, and transforming data
- Handling missing data, duplicates, and outliers
- Data transformation techniques: merging, joining, reshaping

Activities:

- Discussions on the applications of data science in various industries.
- Hands-on exercises in Jupyter Notebooks covering Python basics, data structures, and pandas operations for data wrangling.
- Practical sessions on cleaning and transforming sample datasets.

Assessments:

- Quiz on basic data science concepts and Python fundamentals.
- Coding assignment using Python and pandas to perform data wrangling tasks on a given dataset.

Module 2: Data Visualization and Exploratory Data Analysis

Objective: To master data visualization techniques for exploratory data analysis and gain insights from data.

Topics:

Data Visualization:

- Importance of data visualization in Data Science
- Matplotlib and Seaborn basics
- Creating line plots, histograms, bar charts, scatter plots
- Customizing visualizations for better understanding

Exploratory Data Analysis (EDA):

- Understanding the distribution of data
- Descriptive statistics: mean, median, mode, standard deviation
- Identifying patterns and relationships in data
- Visualizing correlations and distributions

Hands-on Project: EDA and Visualization:

- Practical session to analyze a sample dataset and generate insights using EDA techniques

Activities:

- Creating various types of plots and charts using Matplotlib and Seaborn to visualize different aspects of data.
- Conducting exploratory data analysis on sample datasets to identify key features, patterns, and relationships.
- A practical project where participants apply EDA and visualization techniques to a dataset and present their findings.

Assessments:

- Assignment on creating different types of visualizations to represent data effectively.
- Project report on the EDA and visualization of a given dataset, including insights and interpretations.

Module 3: Machine Learning Fundamentals

Objective: To build and evaluate machine learning models, including linear regression and classification algorithms.

Topics:

Introduction to Machine Learning:

- What is Machine Learning? Overview of ML types: Supervised vs. Unsupervised
- Supervised learning algorithms: Regression and Classification
- Unsupervised learning: Clustering

Model Building: Linear Regression:

- Introduction to linear regression
- Building a linear regression model using Scikit-learn
- Evaluating model performance (MSE, RMSE)
- Feature selection and multicollinearity

Classification Algorithms: Logistic Regression:

- Understanding Logistic Regression for binary classification
- Model evaluation: Accuracy, Precision, Recall, F1-score
- Hands-on implementation using Scikit-learn

Activities:

- Building linear regression and logistic regression models using Scikit-learn.
- Evaluating model performance using appropriate metrics.
- Hands-on implementation of logistic regression for binary classification problems.

Assessments:

- Coding assignment to build and evaluate a linear regression model.
- Assignment on building and evaluating a logistic regression model for a classification task.

Module 4: Advanced Machine Learning & Model Deployment

Objective: To gain insights into advanced machine learning techniques and learn the basics of model deployment.

Topics:

Advanced Machine Learning Algorithms:

- Decision Trees and Random Forest
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- Model evaluation techniques: Cross-validation

Unsupervised Learning: Clustering Algorithms:

- K-Means clustering
- Hierarchical clustering
- Dimensionality reduction: PCA (Principal Component Analysis)

Model Deployment Basics:

- Introduction to model deployment
- Saving models with pickle
- Deploying models using Flask and Docker (Basic concepts)
- Overview of cloud-based deployment (e.g., AWS, Azure)

Capstone Project: End-to-End Data Science Project:

- Participants will work on a real-world problem, applying concepts learned to clean data, build a model, and visualize results.

Activities:

- Implementing advanced machine learning algorithms (Decision Trees, Random Forest, SVM, KNN) using Scikit-learn.
- Applying clustering algorithms to unsupervised learning problems.
- Hands-on introduction to saving models using pickle and basic deployment concepts with Flask and Docker.

Assessments:

- Assignment on applying advanced machine learning algorithms to a given dataset.
- Project proposal outlining the steps for deploying a machine learning model.

Passing Criteria: Minimum 50% overall to pass the program.

Assignments: One assignment per module, designed to reinforce key concepts and practical skills.

Project: One **Capstone Project** to demonstrate applied knowledge and integrate learning from all modules.