Diti Technologies
Unlock Your Potential

Flat no: 103, Mahindra Residency, Beside Swathi Anukar Building, Ameerpet, Hyderabad-16

Data Science Course
Duration: 3 months 15 days

Mobile: 9100036789, 9100075868
Email: datascience@dititechnologies.com
Website: dititechnologies.com

Prerequisites

- Need basic Computer Skills
- Basic Mathematical Concepts

Course Objective

- Work with various data generation sources
- Analyze structured and unstructured data using different tools and techniques
- Develop an understand of Descriptive and Predictive Analytics
- Apply Data-driven, Machine Learning approaches for business decisions
- Build models for day-to-day applicability
- Perform Forecasting to take proactive business decisions
- Perform Text Mining to generate Sentiment Analysis
- Develop Use cases with Generative AI

Data Science - Preliminaries

- Getting Started with Data Science
- Differences and Interrelation of AI, ML, DL, Generative AI
- Data Science Skill Set
- End to End Data Science Project Life Cycle
- Different types of Data Science Tasks
- Introduction to Big Data Analytics and its uses
- Stages of Analytics - Descriptive, Predictive, Prescriptive, etc.
- Course outline, road map, and takeaways from the course
- Data Science Application Categories

Course Modules

Python

- Introduction to Python Programming
- Installation of Python
- Installation of Anaconda Distribution
- Setting Up Python Environment
- Python Editors & IDEs
- Getting Started with Jupyter notebook
- Concept of Packages/Libraries
- Installing & loading Packages
- Data Types
  - Integers
  - Float
  - String
  - Boolean Numbers
  - Complex Numbers
- Operators in Python
  - Arithmetic operators
  - Relational operators
  - Logical operators
  - Assignment operators
  - Bitwise operators
  - Membership operators
  - Identity operators
- Data structures
  - String Representation
  - Lists
  - Tuple
  - Sets
  - Dictionary
  - Matrix
  - Arrays
  - Series
  - Data Frames
- Date & Time Values
- Conditional Statements
  - if statement
  - if - else statement
  - if - elif statement
  - Nest if-else

Diti Technologies

- Multiple if
- Switch

- Loops
  - While loop
  - For loop
  - Range
  - Iterator and generator Introduction
  - For – else
  - Break
- Functions
  - Purpose of a function
  - Defining a function
  - Calling a function
  - Function parameter passing
    i. Formal arguments
    ii. Actual arguments
    iii. Positional arguments
    iv. Keyword arguments
    v. Variable arguments
    vi. Variable keyword arguments
    vii. *args, **kwargs
- Function call stack
  - Local
  - Global
- Modules
  - Python Code Files
  - Importing functions from another file
  - name: Preventing unwanted code execution
  - Folders Vs Packages
  - init.py
  - Namespace
  - Import *
- File Handling
- Exception Handling
- Oops concepts
- Classes and Objects
- Inheritance and Polymorphism
- Multi-Threading

Business Statistics

- Descriptive Statistics
  - Measures of Central Tendency
    Mean/Average, Median, Mode
  - Measures of Spread
    Variance, Standard Deviation, Range
- Inferential Statistics
  - Sampling
  - Need for Sampling?
  - Sampling Techniques
  - Probability & Probability Distribution
  - Continuous Probability Distribution / Probability Density Function
  - Discrete Probability Distribution / Probability Mass Function
  - Confidence interval
  - Normal Distribution and Characteristics of Normal Distribution
  - Standard Normal Distribution / Z distribution
  - Z scores and the Z table
  - Uniform Distribution
  - F-distribution
  - Binomial Distribution
  - Poisson Distribution
  - Bernoulli Distribution
  - Chi- Square Distribution
- Hypothesis Testing
  - Null and Alternative Hypothesis
  - Type I or Alpha Error and Type II or Beta Error
  - Reject or acceptance criterion
  - Confidence Level, Significance Level, Power of Test
- 1 Sample t-test, 2 Sample t-test and Paired t-test
- Z-test
- ANOVA
- Chi-Square test
- Correlation, Covariance, Associations, Odds Ratio, Relative Risk
- Spurious correlation
- Correlation vs. Causation
- Data Visualization using Python
  - Pie chart
  - Donut Chart
  - Histogram
  - Density Plot
  - Bar chart
  - Box plot
  - Scatter plot
  - Scatter plot matrix
  - Correlation Plot
  - Line Chart
  - Pairs Plot

Exploratory Data Analytics (EDA)

- Data Collection
- Data Types namely Continuous, Discrete, Categorical, Qualitative, Quantitative
- Classification of data in terms of Nominal, Ordinal, Interval & Ratio types
- Batch Processing vs Real Time Processing
- Structured versus Unstructured vs Semi-

Diti Technologies

Structured Data

- Balanced versus Imbalanced datasets
- Big Data vs Non-Big Data

Data Preprocessing

- Data Cleaning / Preparation - Outlier Analysis, Missing Values Imputation
- Data Manipulation - Sorting, Filtering, Duplicates, Merging, Appending, Sub setting, Derived variables, Typecasting, Renaming, Formatting etc.
- Uni variate, Bi variate, and Multivariate Analysis
- Encoding: Dummy Variable Creation and Label Encoding
- Scaling Techniques - Transformations, Normalization / Standardization
- Sampling techniques for handling Balanced vs. Imbalanced Datasets

Feature Engineering

- Feature Engineering on Numeric / Non-numeric Data
- Feature Extraction
- Feature Selection

Machine Learning

Supervised Learning

- Steps in Supervised Learning
- Difference between Regression and Classification
- Training, Validation and Testing data
- Evaluation Strategies
  - R-square, Adjusted R-square, MSE, RMSE, MAE
  - Confusion Matrix
  - F-1 Score, Accuracy, Precision and Recall
  - Sensitivity and Specificity
  - ROC and AUC
  - Hyper Parameters
  - Underfit and Overfit
  - Cross Validation

Linear Regression

- Principles of Linear regression
- Assumption & Steps in Linear regression
- Simple Regression and Multiple Linear Regression
- Variable Selection
- Gradient Descent Approach
- Ordinary least squares
- Cost Functions
- Model Development and interpretation
- Model Validation and Diagnostics
- Analysis of Regression results
- R-square, Adjusted R-square, MSE, RMSE, MAE
- Multicollinearity (Variance Inflation Factor)
- Homoscedasticity (Equal Variance) / Heteroscedasticity
- Advantages and Disadvantages

Logistic Regression

- Need for Logistic Regression
- Principles of Logistic Regression
- Assumption & Steps in Logistic Regression
- LOGIT link function
- Analysis of Logistic Regression results
  - Confusion matrix
  - False Positive, False Negative
  - True Positive, True Negative
  - Precision, Recall, Sensitivity, Specificity, F1 - Score
- Receiver operating characteristics curve (ROC)
- AUC
- Advantages and Disadvantages

Regularization Techniques

- Lasso Regression (L1 Regularization)
- Ridge Regression (L2 Regularization)
- Dropout (Used in Neural Networks)

Decision Trees

- Classification and Regression Trees
- Process of Tree building
- Measures of Impurity
- Entropy, Information Gain and GINI Index
- Choosing variables for Decision nodes
- Over fitting underfitting
- Pruning – Pre and Post Prune techniques
- Generalization and Regulation Techniques to avoid overfitting in Decision Tree
- Advantages and Disadvantages

Diti Technologies

Ensemble Techniques - Random Forest and Boosting

- Bagging, Boosting, Voting, Stacking

Random Forest

- Random Forest and understanding various arguments
- Checking for Underfitting and Overfitting in Random Forest
- Generalization and Regulation Techniques to avoid overfitting in Random Forest

Boosting

- Gradient Boosting Algorithm
- Extreme Gradient Boosting (XGB) Algorithm
- Checking for Underfitting and Overfitting
- Generalization and Regulation Techniques to avoid overfitting

KNN Classifier

- Deciding the K value
- Thumb rule in choosing the K value.
- Normalization of variables
- Building a KNN model by splitting the data
- Checking for Underfitting and Overfitting
- Generalization and Regulation Techniques to avoid overfitting

SVM – (Kernel Method)

- Hyperplanes
- Maximum Margin Line
- Cost Parameters
- SVM for Noisy Data
- Non- Linear Space Classification
- Non-Linear Kernel Tricks
- Linear Kernel
- Polynomial
- Sigmoid
- Gaussian RBF
- SVM for Multi-Class Classification

Naive Bayes

- Conditional Probability
- Bayes Rule
- Naïve Bayes Classifier
- Text Classification using Naive Bayes
- Checking for Underfitting and Overfitting in Naive Bayes
- Generalization and Regulation Techniques to avoid overfitting in Naive Bayes

Unsupervised Learning

Clustering / Segmentation

- Distance Metrics
- K means Clustering
- Hierarchical Clustering
- DBSCAN
- Clustering Evaluation metrics
- Elbow Curve / Scree Plot

Dimensionality Reduction Techniques

- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)

Association Rules

- Market Basket Analysis
- APRIORI Algorithm
- Association rules mining
- Measurement Metrics
  - Support
  - Confidence
  - Lift

Recommender Systems

- User Based Collaborative Filtering
- Item Based Collaborative Filtering
- Similarity Metrics
- Search Based Methods

Text Mining and Natural Language Processing (NLP)

- Sources of data
- Pre-processing, corpus, DocumentTerm Matrix (DTM) & TDM
- Tokenization, Stemming, Lemmatization, Chunking, Lexicons, Polarity, Subjectivity
- Stop words
- Regular Expressions
- Bag of words
- Word Clouds
- Unigram, Bigram, Trigram
- Text Classification and Sentiment Analysis
- Topic Modelling

DiT Technologies

Time Series / Forecasting (Model driven and data driven methods)

- Introduction to time series data
- Steps to forecasting
- Components to time series data
- Lag Plot
- ACF - Auto-Correlation Function / Correlogram
- Errors in the forecast and it metrics - ME, MAD, MSE, RMSE, MPE, MAPE
- Stationary Time Series
- Trend, Seasonality, Randomness
- Moving Averages
- Exponential Smoothing
- AR (Auto-Regressive) model for errors
- \Moving Average
- Exponential Smoothing
- Holt's / Double Exponential Smoothing
- Winters / Holt-Winters
- De-seasoning and de-trending
- Seasonal Indexes
- ARMA (Auto-Regressive Moving Average), Order p and q
- ARIMA (Auto-Regressive Integrated Moving Average), Order p, d, and q
- Multivariate Time Series Analysis (VAR - Vector Autoregression)

Deep Learning (ANN, CNN, RNN)

- Artificial Neural Networks
- Introduction to Perceptron and Multilayer Perceptron
- Neurons of a Biological Brain
- Artificial Neuron
- Perceptron
- Perceptron Algorithm
- Artificial Neural Networks (ANN)
    - Integration functions
    - Activation functions (Sigmoid, Tanh, Relu etc.)
    - Weights
    - Bias
    - Learning Rate - Shrinking Learning Rate, Decay Parameters
    - Error functions - Entropy, Binary Cross Entropy, Categorical Cross Entropy, KL Divergence, etc.
    - Gradient Descent Algorithm
    - Backward Propagation
    - Network Topology
    - Principles of Gradient Descent (Manual Calculation)
    - Learning Rate (eta)
    - Batch Gradient Descent
    - Stochastic Gradient Descent
    - Minibatch Stochastic Gradient Descent
    - Optimization Methods: Adagrad, Adadelta, Adam
- Convolution Neural Network (CNN)
    - Image Processing
- Recurrent Neural Network
    - Text Analytics and Sentiment Analysis
    - Long Short-Term Memory (LSTM)
    - Gated Recurrent Network (GRU)

PySpark

- Introduction
- Spark Framework
- RDD
- Pyspark
- Models

R Programming

- Importance of R
- R and R-studio installation
- Getting started with R

Azure, AWS, GCP

Databricks

Snowflake

Flask and Django

Generative AI Use cases (ChatGPT Models)

SQL

- What is a Database
- Types of Databases
- DBMS vs RDBMS
- DBMS Architecture
- Normalization
- Install PostgreSQL
- Install MySQL
- Data Models
- DBMS Language

Dit Technologies

- ACID Properties in DBMS
- What is SQL
- SQL Data Types, commands, Operators, Keys, Joins
- Subqueries with select, insert, update, delete statements
- GROUP BY, HAVING, ORDER BY
- Views in SQL
- Set Operations and Types
- functions
- Triggers
- Introduction to NoSQL Concepts
- SQL vs NoSQL
- Database connection SQL to Python

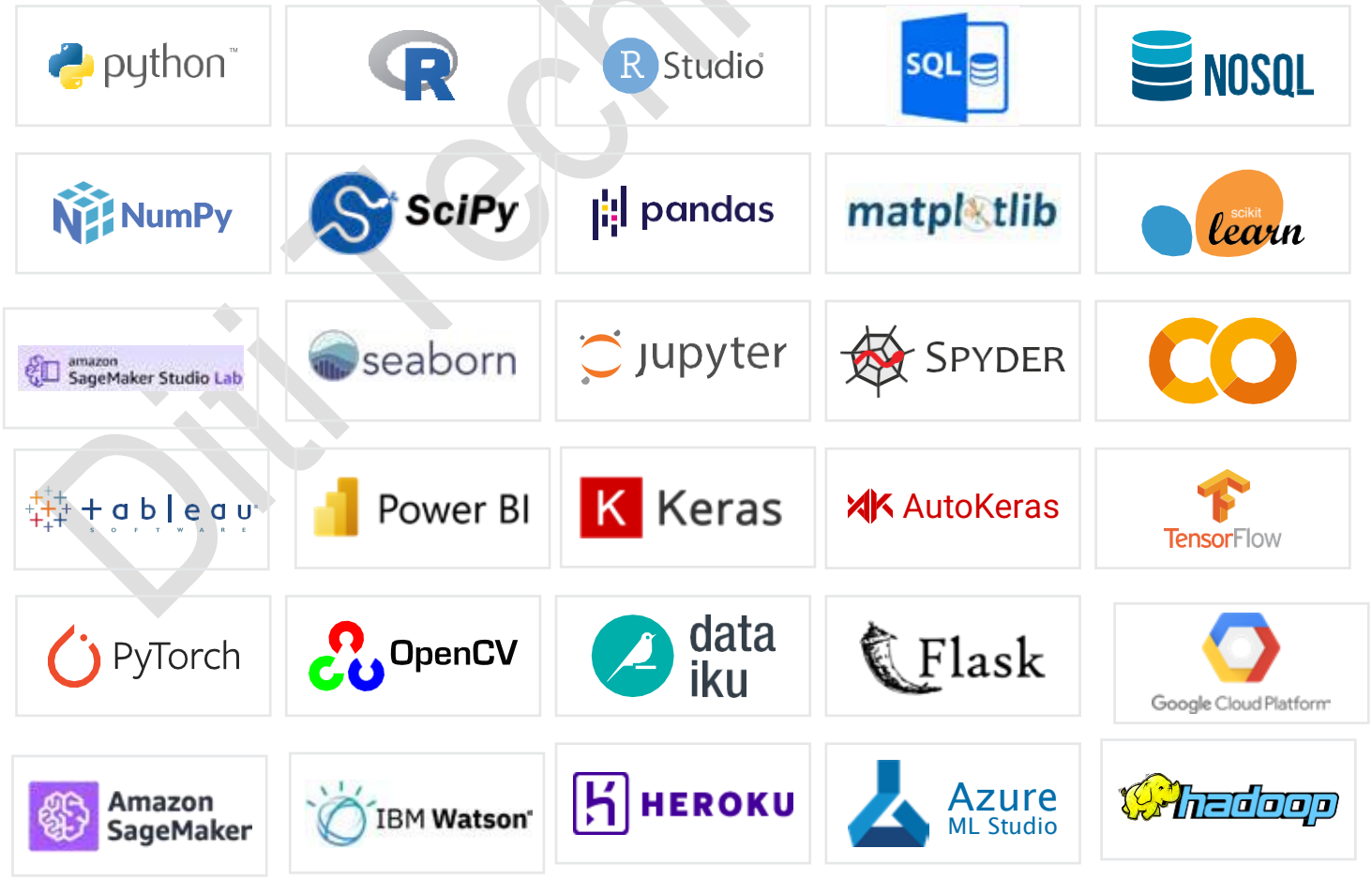
PowerBI

- Installation and Introduction to PowerBI
- Transforming Data using Power BI Desktop
- Importing data
- Changing Database
- Data Types in PowerBI
- Basic Transformations
- Managing Query Groups
- Splitting Columns
- Changing Data Types
- Working with Dates
- Removing and Reordering Columns
- Conditional Columns
- Custom columns
- Connecting to Files in a Folder
- Merge Queries
- Transforming Less Structured Data
- Column profiling
- Query Performance Analytics

Tableau

- Installation and Introduction to PowerBI
- Workbooks
- Dashboards

Tools Covered



Dit Technologies