

DATA SCIENCE

**with DATA ANALYTICS, MACHINE LEARNING,
DEEP LEARNING & ARTIFICIAL INTELLIGENCE
using PYTHON, R & Weka**

INTRODUCTION TO DATA SCIENCE:

- What is Data Science?
- Who is Data Scientist and who can become a Data Scientist?
- Real time process of Data Science
- Data Science Applications
- Technologies used in Data Science
- Prerequisites knowledge to learn Data Science

INTRODUCTION TO MACHINE LEARNING:

- What is Machine Learning?
- How Machine will learn like Human Learning?
- Traditional Programming vs. machine learning
- Machine Learning engineer responsibilities
- Types of learning
 - Supervised learning
 - Un-supervised learning
- Machine learning algorithms: KNN, Naïve-bayes, Decision trees, Classification rules, Regression (Linear Regression, Logistic Regression), K-means clustering, Association rules, Support Vector Machine, Random Forest.

PYTHON PROGRAMMING:

- What is Python? History of Python
- Python Features, Applications of Python
- Downloading and Installing Python
- Python IDE: Jupyter Notebook & Spyder
- What is Anaconda Navigator?
- Downloading and Installing Anaconda, Jupyter Notebook & Spyder
- Python Programming vs. Existing Programming
- Interactive Mode Programming & Script Mode Programming
- Python Identifiers, Reserved Words
- Lines and Indentations, Quotations, Comments
- Assigning values to variables

- Operators - Arithmetic Operators, Comparison (Relational) Operators, Assignment Operators, Logical Operators, Bitwise Operators, Membership Operators, Identity Operators
- Decision Making and Loops
- Flavors in Python, Python Versions
- Data Types: int, float, complex, bool, str
- List, Tuple, Range, Bytes & Bytearray
- Set, Frozenset, Dict, None
- Inbuilt Functions in Python, Slice operator - Indexing
- Mutable vs. Immutable, Modules and Packages
- Database Connection - PyMySQL, Defining & Manipulating

NumPy with Python:

- NumPy Environment setup in Python, Features of NumPy
- Array Creation, Indexing & Slicing, Array Manipulation
- Mathematical Functions, Statistical Functions

Pandas with Python:

- Pandas Environment setup in Python
- Features of Pandas, Data Structures
- Series - Create Series, Accessing Data from Series with Position
- DataFrame - Features of DataFrame, Create DataFrame, DataFrame from List, Dict, Row & Column Selecting, Adding & Deleting
- Panel - Create and select data from Panel
- Indexing & Selecting Data, Statistical Functions
- Merging / Joining, Categorical Data

R PROGRAMMING:

- R Programming Introduction
- R Programming vs. Existing Programming
- Downloading and Installing R, What is CRAN?
- R Programming IDE: RStudio, Downloading and Installing RStudio
- Variable Assignment - Displaying & Deleting Variables
- Comments – Single Line and Multi Line Comments
- Data Types – Logical, Integer, Double, Complex, Character
- Operators - Arithmetic Operators, Relational Operators, Logical Operators, Assignment Operators, R as Calculator, Performing different Calculations
- Functions – Inbuilt Functions and User Defined Functions
- STRUCTURES – Vector, List, Matrix, Data frame, Array, Factors
- Inbuilt Constants & Functions

Setting Environment:

- Search Packages in R Environment
- Search Packages in Machine with inbuilt function and manual searching
- Attach Packages to R Environment

- Install Add-on Packages from CRAN
- Detach Packages from R Environment
- Functions and Packages Help

Vectors:

- Vector Creation, Single Element Vector, Multiple Element Vector
- Vector Manipulation, Sub setting & Accessing the Data in Vectors

Lists:

- Creating a List, Naming List Elements, Accessing List Elements
- Manipulating List Elements, Merging Lists, Converting List to Vector

Matrix:

- Creating a Matrix, Accessing Elements of a Matrix
- Matrix Manipulations, Dimensions of Matrix, Transpose of Matrix

Data Frames:

- Create Data Frame, Vector to Data Frame
- Why Characters are Converting into Factors? – stringsAsFactors
- Convert the columns of a data frame to characters
- Extract Data from Data Frame
- Expand Data Frame, Column Bind and Row Bind
- Merging / Joining Data Frames – Inner Join, Outer Join & Cross Join

Arrays:

- Create Array with Multiple Dimensions, Naming Columns and Rows
- Accessing Array Elements, Manipulating Array Elements
- Calculations across Array Elements

Factors:

- Factors in Data Frame, Changing the Order of Levels
- Generating Factor Levels, Deleting Factor Levels

Loading and Reading Data:

- **DATA EXTRACTION FROM CSV**
 - Getting and Setting the Working Directory
 - Input as CSV File, Reading a CSV File
 - Analyzing the CSV File, Writing into a CSV File
- **DATA EXTRACTION FROM URL**
- **DATA EXTRACTION FROM CLIPBOARD**
- **DATA EXTRACTION FROM EXCEL**
 - Install “xlsx” Package
 - Verify and Load the “xlsx” Package, Input as “xlsx” File
 - Reading the Excel File, Writing the Excel File
- **DATA EXTRACTION FROM DATABASES**
 - RMySQL Package, Connecting to MySQL
 - Querying the Tables, Query with Filter Clause
 - Updating Rows in the Tables, Inserting Data into the Tables
 - Creating Tables in MySQL, Dropping Tables in MySQL
 - Using dplyr and tidyr package

STATISTICS:

- Mean, Median and Mode
- Data Variability: Range, Quartiles, IQR, Calculating Percentiles
- Variance, Standard Deviation, Statistical Summaries
- Types of Distributions – Normal, Binomial, Poisson
- Probability Distributions, Skewness, Outliers
- Data Distribution, 68–95–99.7 rule (Empirical rule)
- Descriptive Statistics and Inferential Statistics
- Statistics Terms and Definitions, Types of Data
- Data Measurement Scales, Normalization
- Measure of Distance, Euclidean Distance
- Probability Calculation – Independent & Dependent
- Hypothesis Testing, Analysis of Variance

DATA VISUALIZATION:

- Data Visualization with Matplotlib and Seaborn
- Data Visualization with Graphics and GrDevices
- High Level Plotting and Low Level Plotting
- Pie Charts - Title, Colors, Slice Percentages, Chart Legend
- 3D Pie Charts
- Box Plots - Outliers, Ranges, IQR, Quartiles, Median, Data Distribution Analysis, 68–95–99.7 rule (Empirical rule)
- Bar Charts - Label, Title, Colors, Group Bar, Stacked Bar Charts
- Histograms - Range of X and Y Values
- Line Graphs - Types: Points, Lines, Both, Overplotted, Steps
- Scatterplots
- Combining Plots - Par and Layout

LAZY LEARNING – CLASSIFICATION USING NEAREST NEIGHBORS:

- **Understanding Classification Using Nearest Neighbors**
 - The KNN algorithm
 - Calculating distance
 - Choosing an appropriate k
 - Preparing data for use with KNN
 - Why is the KNN algorithm lazy?
- **Diagnosing breast cancer with the KNN algorithm**
 - Collecting data
 - Exploring and preparing the data
 - Transformation-normalizing numeric the data
 - Data preparing –creating training and test datasets
 - Training a model on the data
 - Evaluating model performance
 - Improving model performance
 - Transformation –z-score standardization
 - Testing alternative values of k

PROBABILISTIC LEARNING – CLASSIFICATION USING NAÏVE BAYES:

- **Understanding Naïve-Bayes**
 - Basic concepts of Bayesian methods
 - Probability
 - Joint probability
 - Conditional probability with Bayes' theorem
- **The Naïve Bayes Algorithm**
 - The Naïve Bayes classification
 - The Laplace estimator
 - Using numeric features with Naïve Bayes
- **Filtering Mobile Phone Spam with the Naïve-Bayes Algorithm**
 - Collecting data
 - Exploring and preparing the data
 - Data preparation –processing text data for analysis
 - Data preparation –creating training and test datasets
 - Visualizing text data-word clouds
 - Data preparation-creating indicator features for frequent words
 - Training a model on the data
 - Evaluating model performance
 - Improving model performance

DIVIDE AND CONQUER – CLASSIFICATION USING DECISION TREES AND RULES:

- **Understanding decision trees**
 - Divide conquer
 - The C5.0 decision tree algorithm
 - Choosing the best split
 - Pruning the decision tree
- **Identifying risky bank loans using C5.0 decision trees**
 - Collect data
 - Exploring and preparing the data
 - Data preparation-creating random training and test datasets
 - Training a model on the data
 - Evaluating model performance
 - Improving model performance
 - Boosting the accuracy of decision trees
 - Making some mistakes more costly than others
- **Understanding classification rules**
 - Separate and conquer
 - The one rule algorithm
 - The RIPPER algorithm
 - Rules from decision trees

➤ **Identifying poisonous mushrooms with rule learners**

- Collecting data
- Exploring and preparing data
- Training a model on the data
- Evaluating model performance
- Improving model performance

FORECASTING NUMERIC DATA – REGRESSION METHODS:

➤ **Understanding regression**

- Simple linear regression
- Ordinary least squares estimation
- Correlations
- Multiple linear regressions

➤ **Predicting medical expenses using linear regression**

- Collecting data
- Exploring and preparing data
 - Exploring relationships among features- the correlation matrix
 - Visualizing relationships among features –the scatter plot matrix
- Training a model on the data
- Evaluating model performance
- Improving model performance
 - Model specification –adding non-linear relationships
 - Transformation –converting a numeric variable to a binary indicator
 - Model specification –adding interaction effects
 - Putting it all together-an improved regression model

➤ **Understanding regression trees and model trees**

- Adding regression to trees

➤ **Estimating the quality of wines with regression trees and model trees**

- Collecting data
- Exploring and preparing the data
- Training a model on the data
 - Visualizing decision trees
- Evaluating model performance
 - Measuring performance with mean absolute error
- Improving model performance

FINDING PATTERNS - MARKET BASKET ANALYSIS USING ASSOCIATION RULES:

➤ **Understanding Association Rules**

- The Apriori algorithm for association rule learning
 - Measuring rule interest –support and confidence

- Building a set of rules with the Apriori
- **Identifying frequently purchased groceries with association rules**
 - Collecting data
 - Exploring and preparing the data
 - Data preparation – creating a sparse matrix for transaction data
 - Visualizing item support –item frequency plots
 - Visualizing transaction data-plotting the sparse matrix
 - Training a model on the data
 - Evaluating model performance
 - Improving model performance
 - Sorting the set of association rules
 - Taking subsets of association rules
 - Saving association rules to a file or data frame

FINDING GROUPS OF DATA - CLUSTERING WITH K-MEANS:

- **Understanding Clustering**
 - Clustering as a machine learning task
 - The K-means algorithm for clustering
 - Using distance to assign and update cluster
 - Choosing the appropriate number of cluster
- **Finding teen market segments using K-means clustering**
 - Collecting data
 - Exploring and preparing the data
 - Data preparation –dummy coding missing values
 - Data preparing –imputing missing values
 - Training a model on the data
 - Evaluating model performance
 - Improving model performance

EVALUATING MODEL PERFORMANCE:

- **Measuring Performance for Classification**
 - Working with classification prediction data in R
 - A closer look at confusion matrices
 - Using confusion matrices to measure performance
 - Beyond accuracy – other measure of performance
 - The kappa statistic
 - Sensitivity and specificity
 - Precision and recall
 - The F- measure
 - Visualizing performance TRADEOFFS
 - ROC curves
- **Estimating future performance**
 - The holdout method
 - Cross-validation

- Bootstrap sampling

IMPROVING MODEL PERFORMANCE:

- **Tuning Stock Models for Better Performance**
 - Using caret for automated parameter tuning
 - Creating a simple tuned model
 - Customizing the tuning process
- **Improving Model Performance with Meta – Learning**
 - Understanding ensembles
 - Bagging
 - Boosting
 - Random forests
 - Training random forests
 - Evaluating random forest performance

DEEP LEARNING:

- Installation of Theano, TensorFlow, Keras, OpenCV
- Relating Deep Learning and Traditional Machine Learning
- Basics of Neural Networks
- Artificial Neural Networks
- Deep Neural Networks
- Convolutional Neural Networks
- Recurrent Neural Networks
- Deep learning with Theano
- Deep Learning with TensorFlow
- Deep Learning with Keras
- Deep Learning with OpenCV
- Implementation of Deep learning

ARTIFICIAL INTELLIGENCE:

- AI Introduction
- AI Intelligent Systems
- AI Popular Search Algorithms
- AI Fuzzy Logic Systems
- AI Natural Language Processing
- AI Robotics
- AI Neural Networks

INTRODUCTION TO WEKA

- **EXPLORE WEKA MACHINE LEARNING TOOLKIT**
 - Installation of WEKA
 - Features of WEKA Toolkit
 - Explore & Load data sets in Weka
- **PERFORM DATA PREPROCESSING TASKS**
 - Apply Filters on data sets

- **PERFORMING CLASSIFICATION ON DATA SETS**
 - J48 Classification Algorithm
 - Decision Trees Algorithm
 - K-NN Classification Algorithm
 - Naive-bayes Classification Algorithm
 - Comparing Classification Results
- **PERFORMING REGRESSION ON DATA SETS**
 - Simple Linear Regression Model, Multi Linear Regression Model
 - Logistic Regression Model, Cross-Validation and Percentage Split
- **PERFORMING CLUSTERING ON DATA SETS**
 - Clustering Techniques in Weka
 - Simple K-means Clustering Algorithm
 - Association Rule Mining on Data Sets
 - Apriori Association Rule Algorithm
 - Discretization in the Rule Generation Process
- **GRAPHICAL VISUALIZATION IN WEKA**
 - Visualization Features in Weka
 - Visualize the data in various dimensions
 - Plot Histogram, Derive Interesting Insights



DATAhill

Trainer: Mr. Srinivas Reddy

- Trainer received Masters of Technology in Computer Science & Engineering from JNTU, MICROSOFT Certified Professional, Certified from IIT Kanpur & IIT Ropar.
- Having 10+ Years of Experience in Software & Training.
- His experience Includes Managing, Data Processing, Data Cleaning, Predicting and Analyzing of Large volume of Business Data.
- Expertise in Data Science, Data Analytics, Machine Learning, Deep Learning, Artificial Intelligence, Python, R, Weka, Data Management & BI Technologies.
- Having publications and patents in various fields such as machine learning, data security, and data science technologies.
- Professionally, he is Data Science management consultant with over 7+ years of experience in finance, retail, transport and other industries.

KEY FEATURES IN THIS TRAINING

- Best training materials are provided with Lab Exercises, Data sets, Codes, Quizzes, Case studies on real data.
- For every online session Recorded video & live running notes will provide.
- Real time Training with live Scenarios and Applications.
- Support in Resume preparation and Interview preparation.
- Conduct Mock interviews through Skype and Telephonic after course completion.
- You can shift the batch to weekday batches (morning or evening) and weekend batches.
- Any number of batches can be attend in a year without any extra fees
- Job support for 1 month after successfully placing the candidates.
- Online help on Doubt Clearance, Career Guidance, Resume Preparation and Interview Preparation.



DATA SCIENCE



**with Data Analytics, Machine Learning,
Deep Learning & Artificial Intelligence
using Python, R & Weka**

Classroom & Online Training



Good value for money – Charged less than any other training institutions

Contact: 9292005440

Mail: datahills7@gmail.com

DATAhill Solutions, Near Malabar Gold, KPHB, Hyderabad.

Ph: +91 9292005440, +91 7780163743, info@datahill.in, www.datahill.in