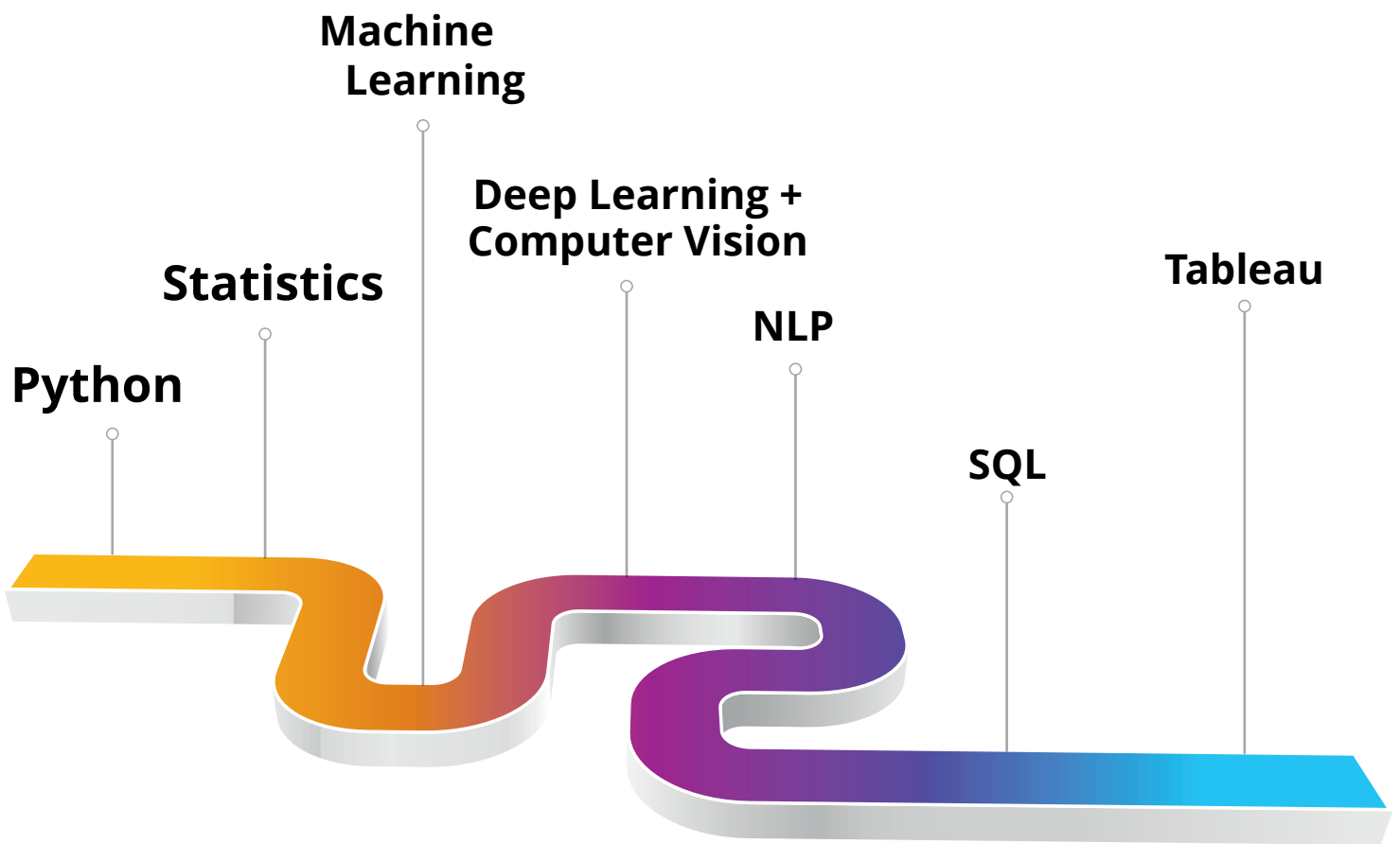


INNOMATICS[®]
RESEARCH LABS

DATA SCIENCE CURRICULUM



Course Objective

- To **understand the vital nature of data** for organizations.
- To learn the conceptual framework of machine learning.
- To **explore and analyze data** using supervised and unsupervised learning techniques.
- To **develop and deploy** knowledge learning models using Python.
- To Work on Unstructured Data Like Text processing them using **Nltk** and building Modules.
- Understanding **Neural Networks** and building deep networks using **Tensorflow and Keras** and working with **image processing** computer vision.

Key features in the Training

- **Duration: 5-6 Months**
- **Class Duration: 2 - Hrs (Monday to Friday)**
- **Online help on Doubt Clearance**, Monitoring Session, Career Guidance, Interview preparation & Mock interviews
- **Use Cases Covered: Python and Statistics : 4 , Machine Learning - 10, NLP - 2 , DL – 3.**
- One **Big Hackathon Challenge** on Machine Learning
- **Projects:** Python: Data Analysis Project, Machine Learning: Regression, Classification
NLP: Sentiment Analysis / Chatbot, DeepLearning: Face Recognition.
- Addition: **Assignments, Quizzes** for each Module From Python, Statistics, Machine Learning, NLP and Deep Learning + Computer vision topic wise assignments and quiz.
- **IBM Credentials** and Certification after completion of the Course.
- **Guaranteed In-house Internship**
- Nearly working on **20 use cases** during your course.
- 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.
- **Job Assistance** after completion of the course.

INTRODUCTION & WALK THROUGH THE COURSE

INTRODUCTION

- Introduction to Data Science
- Life cycle of data science
- Skills required for data science
- Applications of data science in different industries

MODULE 1: PYTHON CORE & ADVANCED

INTRODUCTION

- What is Python?
- Why does Data Science require Python?
- Installation of Anaconda
- Understanding Jupyter Notebook
- Basic commands in Jupyter Notebook
- Understanding Python Syntax

Data Types and Data Structures

- Variables and Strings
- Lists, Sets, Tuples and Dictionaries

Control Flow and Conditional Statements

- Conditional Operators, Arithmetic Operators and Logical Operators
- If, Elif and Else Statements
- While Loops
- For Loops
- Nested Loops and List and Dictionary Comprehensions

Functions

- What is function and types of functions
- code optimization and argument functions
- Scope
- Lambda Functions
- Map, Filter and Reduce

File Handling

- Create, Read, Write files and Operations in File Handling
- Errors and Exception Handling

Class and Objects

- Create a class
- Create an object
- The `__init__()`
- Modifying Objects
- Object Methods
- Self
- Modify the Object Properties
- Delete Object
- Pass Statements

MODULE 2: DATA ANALYSIS IN PYTHON

Numpy - NUMERICAL PYTHON

- Introduction to Array
- Creation and Printing of ndarray
- Basic Operations in Numpy
- Indexing
- Mathematical Functions of Numpy

Data Manipulation with Pandas

- Series and DataFrames
- Data Importing and Exporting through Excel, CSV Files
- Data Understanding Operations
- Indexing and slicing and More filtering with Conditional Slicing
- Groupby, Pivot table and Cross Tab
- Concatenating and Merging Joining
- Descriptive Statistics
- Removing Duplicates
- String Manipulation
- Missing Data Handling

DATA VISUALIZATION

Data Visualization using Matplotlib and Pandas

- Introduction to Matplotlib
- Basic Plotting
- Properties of plotting
- About Subplots
- Line plots
- pie chart and Bar Graph
- Histograms
- Box and Violin Plots
- Scatterplot

Case Study on Exploratory Data Analysis (EDA) and Visualizations

- What is EDA?
- Uni - Variate Analysis
- Bi - Variate Analysis
- More on Seaborn Based Plotting Including Pair Plots, Catplot, Heat Maps, Count plot along with matplotlib plots.

UNSTRUCTURED DATA PROCESSING

Regular Expressions

- Structured Data and Unstructured Data
- Literals and Meta Characters
- How to Regular Expressions using Pandas?
- Inbuilt Methods
- Pattern Matching

PROJECT ON WEB SCRAPING : DATA MINING and EXPLORATORY DATA ANALYSIS

- Data Mining (WEB - SCRAPING)
This project starts completely from scratch which involves collection of Raw Data from different sources and converting the unstructured data to a structured format to apply Machine Learning and NLP models.
This project covers the main four steps of Data Science Life Cycle which involves
 - Data Collection
 - Data Mining
 - Data Preprocessing
 - Data VisualizationEx: Text, CSV, TSV, Excel Files, Matrices, Images

MODULE 3: ADVANCED STATISTICS

Data Types and Data Structures

- Statistics in Data science:
- What is Statistics?
- How is Statistics used in Data Science?
- Population and Sample
- Parameter and Statistic
- Variable and its types

Data Gathering Techniques

- Data types
- Data Collection Techniques
- Sampling Techniques:
 - Convenience Sampling, Simple Random Sampling
 - Stratified Sampling, Systematic Sampling and Cluster Sampling

Descriptive Statistics

- What is Univariate and Bi Variate Analysis?
- Measures of Central Tendencies
- Measures of Dispersion
- Skewness and Kurtosis
- Box Plots and Outliers detection
- Covariance and Correlation

Probability Distribution

- Probability and Limitations
- Discrete Probability Distributions
 - Bernoulli, Binomial Distribution, Poisson Distribution
- Continuous Probability Distributions
 - Normal Distribution, Standard Normal Distribution

Inferential Statistics

- Sampling variability and Central Limit Theorem
- Confidence Intervals
- Hypothesis Testing
 - Z-test, t-test
 - Chi – Square Test
 - F -Test and ANOVA

MODULE 4: SQL

SQL for Data Science

- **Introduction to Databases**
- Basics of SQL
 - DML, DDL, DCL and Data Types
 - Common SQL commands using SELECT, FROM and WHERE
 - Logical Operators in SQL
- SQL Joins
 - INNER and OUTER joins to combine data from multiple tables
 - RIGHT, LEFT joins to combine data from multiple tables
- Filtering and Sorting
 - Advanced filtering using IN, OR and NOT
 - Sorting with GROUPBY and ORDER BY
- SQL Aggregations
 - Common Aggregations including COUNT, SUM, MIN and MAX
 - CASE and DATE functions as well as work with NULL values
- Subqueries and Temp Tables
 - Subqueries to run multiple queries together
 - Temp tables to access a table with more than one query
- SQL Data Cleaning
 - Perform Data Cleaning using SQL

MODULE 5: MACHINE LEARNING SUPERVISED & UNSUPERVISED LEARNING

INTRODUCTION

- What Is Machine Learning?
- Supervised Versus Unsupervised Learning
- Regression Versus Classification Problems Assessing Model Accuracy

REGRESSION TECHNIQUES

Linear Regression

- Simple Linear Regression:
 - Estimating the Coefficients
 - Assessing the Coefficient Estimates
 - R Squared and Adjusted R Squared
 - MSE and RMSE

Multiple Linear Regression

- Estimating the Regression Coefficients
 - OLS Assumptions
 - Multicollinearity
 - Feature Selection
 - Gradient Descent

Evaluating the Metrics of Regression Techniques

- Homoscedasticity and Heteroscedasticity of error terms
 - Residual Analysis
 - Q-Q Plot
 - Cook's distance and Shapiro-Wilk Test
 - Identifying the line of best fit
 - Other Considerations in the Regression Model
 - Qualitative Predictors
 - Interaction Terms
 - Non-linear Transformations of the Predictors

Polynomial Regression

- Why Polynomial Regression
- Creating polynomial linear regression
- evaluating the metrics

Regularization Techniques

- Lasso Regularization
- Ridge Regularization
- ElasticNet Regularization

Case Study on Linear, Multiple Linear Regression, Polynomial, Regression using Python.

CAPSTONE PROJECT: A project on a use case will challenge the Data Understanding, EDA, Data Processing and above Regression Techniques.

CLASSIFICATION TECHNIQUES

Logistic regression

- An Overview of Classification
- Difference Between Regression and classification Models.
- Why Not Linear Regression?
- Logistic Regression:
- The Logistic Model
- Estimating the Regression Coefficients and Making Predictions
- Logit and Sigmoid functions
- Setting the threshold and understanding decision boundary
- Logistic Regression for >2 Response Classes
- Evaluation Metrics for Classification Models:
- Confusion Matrix
- Accuracy and Error rate
- TPR and FPR
- Precision and Recall, F1 Score
- AUC – ROC
- Kappa Score

Naive Bayes

- Principle of Naive Bayes Classifier
- Bayes Theorem
- Terminology in Naive Bayes
- Posterior probability
- Prior probability of class
- Likelihood
- Types of Naive Bayes Classifier
- Multinomial Naive Bayes
- Bernoulli Naive Bayes and Gaussian Naive Bayes

TREE BASED MODULES

Decision Trees

- Decision Trees (Rule Based Learning):
- Basic Terminology in Decision Tree
- Root Node and Terminal Node
- Regression Trees and Classification Trees
- Trees Versus Linear Models
- Advantages and Disadvantages of Trees
- Gini Index
- Overfitting and Pruning
- Stopping Criteria
- Accuracy Estimation using Decision Trees

Case Study: A Case Study on Decision Tree using Python

- Resampling Methods:
- Cross-Validation
- The Validation Set Approach Leave-One-Out Cross-Validation
- k-Fold Cross-Validation
- Bias-Variance Trade-Off for k-Fold Cross-Validation

Ensemble Methods in Tree Based Models

- What is Ensemble Learning?
- What is Bootstrap Aggregation Classifiers and how does it work?

Random Forest

- What is it and how does it work?
- Variable selection using Random Forest

Boosting: AdaBoost, Gradient Boosting

- What is it and how does it work?
- Hyper parameter and Pro's and Con's

Case Study: Ensemble Methods - Random Forest Techniques using Python

DISTANCE BASED MODULES

K Nearest Neighbors

- K-Nearest Neighbor Algorithm
- Eager Vs Lazy learners
- How does the KNN algorithm work?
- How do you decide the number of neighbors in KNN?
- Curse of Dimensionality
- Pros and Cons of KNN
- How to improve KNN performance

Case Study: A Case Study on k-NN using Python

Support Vector Machines

- The Maximal Margin Classifier
- HyperPlane
- Support Vector Classifiers and Support Vector Machines
- Hard and Soft Margin Classification
- Classification with Non-linear Decision Boundaries
- Kernel Trick
- Polynomial and Radial
- Tuning Hyper parameters for SVM
- Gamma, Cost and Epsilon
- SVMs with More than Two Classes

Case Study: A Case Study on SVM using Python

CAPSTONE PROJECT: A project on a use case will challenge the Data Understanding, EDA, Data Processing and above Classification Techniques.

MODULE 6 : UNSUPERVISED LEARNING

- Why Unsupervised Learning
- How it Different from Supervised Learning
- The Challenges of Unsupervised Learning

Principal Components Analysis

- Introduction to Dimensionality Reduction and it's necessity
- What Are Principal Components?
- Demonstration of 2D PCA and 3D PCA
- EigenValues, EigenVectors and Orthogonality
- Transforming Eigen values into a new data set
- Proportion of variance explained in PCA

Case Study: A Case Study on PCA using Python

K-Means Clustering

- Centroids and Medoids
- Deciding optimal value of 'k' using Elbow Method
- Linkage Methods

Hierarchical Clustering

- Divisive and Agglomerative Clustering
- Dendrograms and their interpretation
- Applications of Clustering
- Practical Issues in Clustering

Case Study: A Case Study on clusterings using Python

Association Rules

- Market Basket Analysis

Apriori

- Metric Support/Confidence/Lift
- Improving Supervised Learning algorithms with clustering

Case Study: A Case Study on association rules using Python

CAPSTONE PROJECT : A project on a use case will challenge the Data Understanding, EDA, Data Processing and Unsupervised algorithms.

RECOMENDATION SYSTEMS

- What are recommendation engines?
- How does a recommendation engine work?
- Data collection
- Data storage
- Filtering the data
- Content based filtering
- Collaborative filtering
- Cold start problem
- Matrix factorization
- Building a recommendation engine using matrix factorization
- Case Study

MODULE 7: DEEP LEARNING

Introduction to Neural Networks

- Introduction to Perceptron & History of Neural networks
- Activation functions
 - a) Sigmoid b) Relu c) Softmax d) Leaky Relu e) Tanh
- Gradient Descent
- Learning Rate and tuning
- Optimization functions
- Introduction to Tensorflow
- Introduction to keras
- Back propagation and chain rule
- Fully connected layer
- Cross entropy
- Weight Initialization
- Regularization

TensorFlow 2.0

- Introducing Google Colab
- Tensorflow basic syntax
- Tensorflow Graphs
- Tensorboard

Artificial Neural Network with Tensorflow

- Neural Network for Regression
- Neural Network for Classification
- Evaluating the ANN
- Improving and tuning the ANN
- Saving and Restoring Graphs

MODULE 8: CNN & COMPUTER VISION

UNIT 1: Working with images & CNN Building Blocks

- Working with Images_Introduction
- Working with Images - Reshaping understanding, size of image understanding pixels Digitization, Sampling, and Quantization
- Working with images - Filtering
- Hands-on Python Demo: Working with images
- Introduction to Convolutions
- 2D convolutions for Images
- Convolution - Backward
- Transposed Convolution and Fully Connected Layer as a Convolution
- Pooling: Max Pooling and Other pooling options

UNIT 2: CNN Architectures and Transfer Learning

- CNN Architectures and LeNet Case Study
- Case Study: AlexNet
- Case Study: ZFNet and VGGNet
- Case Study: GoogleNet
- Case Study: ResNet
- GPU vs CPU
- Transfer Learning Principles and Practice
- Hands-on Keras Demo: SVHN Transfer learning from MNIST dataset
- Transfer learning Visualization (run package, occlusion experiment)
- Hands-on demo -T-SNE

UNIT 3: Object Detection

- CNN's at Work - Object Detection with region proposals
 - CNN's at Work - Object Detection with Yolo and SSD
 - Hands-on demo- Bounding box regressor
- #Need to do a semantic segmentation project

UNIT 4: CNN's at Work - Semantic Segmentation

- CNNs at Work - Semantic Segmentation
- Semantic Segmentation process
- U-Net Architecture for Semantic Segmentation
- Hands-on demo - Semantic Segmentation using U-Net
- Other variants of Convolutions
- Inception and MobileNet models

UNIT 5: CNN's at work- Siamese Network for Metric Learning

- Metric Learning
- Siamese Network as metric learning
- How to train a Neural Network in Siamese way
- Hands-on demo - Siamese Network

MODULE 9: NATURAL LANGUAGE PROCESSING

Unit 1: Introduction to Statistical NLP Techniques

- Introduction to NLP
- Preprocessing , NLP Tokenization, stop words, normalization, Stemming and lemmatization
- Preprocessing in NLP Bag of words,TF-IDF as features
- Language model probabilistic models, n-gram model and channel model
- Hands on NLTK

Unit 2 : Word embedding

- Word2vec
- Golve
- POS Tagger
- Named Entity Recognition(NER)
- POS with NLTK
- TF-IDF with NLTK

Unit 3: Sequential Models

- Introdcution to sequential models
- Introduction to RNN
- Intro to LSTM
- LSTM forward pass
- LSTM backprop through time
- Hands on keras LSTM

Unit 4 : Applications

- Sentiment Analysis
- Sentence generation
- Machine translation
- Advanced LSTM structures
- Keras- machine translation
- ChatBot

MODULE 10 : TABLEAU

Tableau for Data Science

- Install Tableau for Desktop 10
- Tableau to Analyze Data
- Connect Tableau to a variety of dataset
- Analyze, Blend, Join and Calculate Data
- Tableau to Visualize Data
- Visualize Data In the form of Various Charts, Plots and Maps
- Data Hierarchies
- Work with Data Blending in Tableau
- Work with Parameters
- Create Calculated Fields
- Adding Filters and Quick Filters
- Create Interactive Dashboards
- Adding Actions to Dashboards

Tools & Technologies



matplotlib



RegEx

