Data Science

Course Overview: This data science program follows the CRISP-DM Methodology. The premier modules are devoted to a foundational perspective of Statistics, Mathematics, Business Intelligence, and Exploratory Data Analysis. The successive modules deal with Probability Distribution, Hypothesis Testing, Data Mining Supervised, Predictive Modelling - Multiple Linear Regression, Lasso and Ridge Regression, Logistic Regression, Multinomial Regression, and Ordinal Regression. Later modules deal with Data Mining Unsupervised Learning, Recommendation Engines, Network Analytics, Machine Learning, Decision Tree and Random Forest, Text Mining, and Natural Language Processing. The final modules deal with Machine Learning - classifier techniques, Perceptron, Multilayer Perceptron, Neural Networks, Deep Learning Black-Box Techniques, SVM, Forecasting, and Time Series algorithms. This is the most enriching training program in terms of the array of topics covered.

Duration :16weeks

Module 1: Python Introduction
Module 2: SQL
Module 3: Data Science - Preliminaries
Module 4: Data Mining - Unsupervised Learning
Module 5: Data Mining - Supervised Learning
Module 6: Forecasting/Time Series
Module 7: Black Box method (ANN, CNN, RNN)
Module 8: Real-Time Data Science Projects
Module 9: Capstone Project

Module 1: Introduction to Data Science

Introduction to Python Programming
Installation of Python & Associated Packages
Graphical User Interface
Installation of Anaconda Python
Setting Up Python Environment
Data Types

Operators in Python:
Arithmetic operators, Relational operators, Logical operators, Assignment operators, Bitwise operators, Membership operators, Identity operators

Data structures:
Vectors, Matrix, Arrays, Lists, Tuple, Sets, String Representation, Arithmetic Operators, Boolean Values, Dictionary

Conditional Statements
if statement, if - else statement, if - elif statement, Nest if-else, 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: Formal arguments, Actual arguments, Positional arguments, Keyword arguments, Variable arguments, Variable keyword arguments, Use-Case *args, **kwargs

Function call stack
Locals (), Global ()

Stack frame
Modules
Python Code Files, importing functions from another file, __name__: Preventing unwanted code execution, Importing from a folder, Folders Vs Packages
__init__.py, Namespace, __all__, Import *, Recursive imports

File Handling
Exception Handling
Regular expressions
Oops concepts
Classes and Objects
Inheritance and Polymorphism
Multi-Threading

Module 2: SQL

What is a Database
Types of Databases
DBMS vs RDBMS
DBMS Architecture
Normalisation & Denormalization
Install PostgreSQL
Install MySQL
Data Models
DBMS Language
ACID Properties in DBMS
What is SQL
SQL Data Types
SQL commands
SQL Operators
SQL Keys
SQL Joins
GROUP BY, HAVING, ORDER BY
Subqueries with select, insert, update, delete statements
Views in SQL
SQL Set Operations and Types
SQL functions
SQL Triggers
Introduction to NoSQL Concepts
SQL vs NoSQL
Database connection SQL to Python

Module 3: Data Science Preliminaries

Data Visualization is a critical aspect of a data science course as it teaches students how to effectively communicate insights and patterns in data through visual representations.

CRISP-ML(Q) - Business & Data Understanding :

- Dos and Don'ts as a participant
- Introduction to Big Data Analytics
- Data and its uses – a case study (Grocery store)
- Interactive marketing using data & IoT – A case study
- Course outline, road map, and takeaways from the course
- Stages of Analytics - Descriptive, Predictive, Prescriptive, etc.
- Cross-Industry Standard Process for Data Mining

Data Preprocessing:

Typecasting, Handling Duplicates, Outlier Analysis/Treatment, Zero or Near Zero Variance Features, Missing Values, Discretization / Binning / Grouping, Encoding: Dummy Variable Creation, Transformation, Scaling: Standardization / Normalization

.

Exploratory Data Analytics(EDA):

The task is also called Descriptive Analytics or also known as exploratory data analysis. In this module, you also are introduced to statistical calculations which are used to derive information along with Visualizations to show the information in graphs/plots

Machine Learning project management methodology

- Data Collection - Surveys and Design of Experiments
- Data Types ,Further classification of data in terms of Nominal, Ordinal, Interval & Ratio types, Balanced vs Imbalanced datasets, Cross Sectional vs Time Series vs Panel / Longitudinal Data, Batch Processing vs Real Time Processing, Structured vs Unstructured vs Semi-Structured Data, Big vs Not-Big Data
- Data Cleaning / Preparation - Outlier Analysis, Missing Values Imputation Techniques, Transformations, Normalization / Standardization, Discretization
- Sampling techniques for handling Balanced vs. Imbalanced Datasets
- What is the Sampling Funnel and its application and its components?
  Population, Sampling frame, Simple random sampling, Sample
- Measures of Central Tendency & Dispersion:
  Population, Mean/Average, Median, Mode, Variance, Standard Deviation, Range

Feature Engineering:

The raw Data collected from different sources may have different formats, values, shapes, or characteristics. Cleansing, or Data Preparation, Data Munging, Data Wrapping, etc., are the next steps in the Data handling stage. The objective of this stage is to transform the Data into an easily consumable format for the next stages of development.

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

Module 4: Data Mining – Unsupervised learning

Mathematical Foundations:

Learn the preliminaries of the Mathematical / Statistical concepts which are the foundation of techniques used for churning the Data. You will revise the primary academic concepts of foundational mathematics and Linear Algebra basics.

Topics: Data Optimization, Derivatives, Linear Algebra, Matrix Operations

Clustering/ Segmentation:

Data mining unsupervised techniques are used as EDA techniques to derive insights from the business data. In this first module of unsupervised learning, get introduced to clustering algorithms. Learn about different approaches for data segregation to create homogeneous groups of data

Topics: Clustering 101, Distance Metrics, Hierarchical Clustering, Non-Hierarchical Clustering, DBSCAN, Clustering Evaluation metrics

Dimension Reduction:

Dimension Reduction (PCA and SVD) / Factor Analysis Description: Learn to handle high dimensional data.

Topics: Principal Component Analysis (PCA), Singular Value Decomposition (SVD)

Association Rules:

Learn to measure the relationship between entities. Bundle offers are defined based on this measure of dependency between products.

Topics: Association rules mining 101,Measurement Metrics, Support, Confidence, Lift

Recommender Systems:

Topics: User Based Collaborative Filtering, Similarity Metrics, Item Based Collaborative Filtering, Search Based Methods, SVD Method.

Network Analytics:

The study of a network with quantifiable values is known as network analytics. The vertex and edge are the nodes and connection of a network, learn about the statistics used to calculate the value of each node in the network.

Topics: Entities of a Network, Properties of the Components of a Network, Measure the value of a Network, Community Detection Algorithms

Text Mining and Natural Language Processing(NLP):

Learn to analyse unstructured textual data to derive meaningful insights. Understand the language quirks to perform data cleansing, extract features using a bag of words and construct the key-value pair matrix called DTM. Learn to understand the sentiment of customers from their feedback to take appropriate actions.

Topics : Sources of data, Bag of words, Pre-processing, corpus Document Term Matrix (DTM) & TDM, Word Clouds, Corpus-level word clouds, Sentiment Analysis, Positive Word clouds, Negative word clouds, Unigram, Bigram, Trigram, Semantic network, Extract, user reviews of the product/services from Amazon and tweets from Twitter, Install Libraries from Shell, Extraction and text analytics in Python, LDA / Latent Dirichlet Allocation, Topic Modelling, Sentiment Extraction, Lexicons & Emotion Mining

Module 5: Data Mining – Supervised learning

Machine Learning:

Topics: Machine Learning primer**,** Difference between Regression and Classification**,** Evaluation Strategies**,** Hyper Parameters, Metrics**,** Overfitting and Underfitting

Naive bayes:

Revise Bayes theorem to develop a classification technique for Machine learning. In this tutorial, you will learn about joint probability and its applications.

Topics: Probability – Recap, 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.

Machine Learning - KNN Classifier:

k Nearest Neighbour algorithm is a distance-based machine learning algorithm. The KNN Classifier also known as a lazy learner is a very popular algorithm and one of the easiest for application.

Topics: Deciding the K value, Thumb rule in choosing the K value, Building a KNN model by splitting the data, Checking for Underfitting and Overfitting in KNN, Generalization and Regulation Techniques to avoid overfitting in KNN.

Confidence Interval

To identify the properties of a continuous random variable, statisticians have defined a variable as a standard, learning the properties of the standard variable and its distribution.

Topics: Probability & Probability Distribution, Continuous Probability Distribution / Probability Density Function, Discrete Probability Distribution / Probability Mass Function, Normal Distribution, Standard Normal Distribution / Z distribution, Z scores and the Z table, QQ Plot / Quantile - Quantile plot,Sampling Variation, Central Limit Theorem, Sample size calculator, Confidence interval - concept, Confidence interval with sigma, T-distribution Table / Student's-t distribution / T table, Confidence interval, Population parameter with Standard deviation known, Population parameter with Standard deviation not known.

Hypothesis Testing:

Learn to frame business statements by making assumptions. Understand how to perform testing of these assumptions to make decisions for business problems. Formulating a Hypothesis

Topics: Choosing Null and Alternative Hypotheses, Type I or Alpha Error and Type II or Beta Error, Confidence Level, Significance Level, Power of Test, Comparative study of sample proportions using Hypothesis testing,2 Sample t-test,ANOVA,2 Proportion test, Chi-Square test.

Supervised Learning- Regression techniques:

Data Mining supervised learning is all about making predictions for an unknown dependent variable using mathematical equations explaining the relationship with independent variables

Topics: Scatter diagram, Correlation analysis, Correlation coefficient, Ordinary least squares, Principles of regression, Simple Linear Regression, Exponential Regression, Logarithmic Regression, Quadratic or Polynomial Regression, Confidence Interval versus Prediction Interval, Heteroscedasticity / Equal Variance.

Multiple Linear Regression - Predictive Modelling

Learn about the conditions and assumptions to perform linear regression analysis and the workarounds used to follow the conditions. Understand the steps required to perform the evaluation of the model and to improvise the prediction accuracies

Topics: LINE assumption, Linearity, Independence, Normality,Equal Variance / Homoscedasticity, Collinearity (Variance Inflation Factor),Multiple Linear Regression, Model Quality metrics, Deletion Diagnostics.

Logistic Regression Binary Value Prediction, MLE

Learn about the principles of the logistic regression model, understand the sigmoid curve, and the usage of cut-off value to interpret the probable outcome of the logistic regression model. Learn about the confusion matrix and its parameters to evaluate the outcome of the prediction model.

Topics: Principles of Logistic regression, Types of Logistic regression, Assumption & Steps in Logistic regression, Analysis of Simple logistic regression results, Multiple Logistic regression, Confusion matrix, False Positive, False Negative, True Positive, True Negative, Sensitivity, Recall, Receiver operating characteristics curve (ROC curve), Precision Recall (P-R) curve, Lift charts and Gain charts.

Lasso and Ridge Regressions:

We need to strike the right balance between overfitting and underfitting, learn about regularization techniques L1 norm and L2 norm used to reduce these abnormal conditions. Understanding Overfitting (Variance) vs. Underfitting (Bias)

Topics: Generalization error and Regularization techniques, Different Error functions, Loss functions, or Cost functions, Lasso Regression, Ridge Regression

Multinomial and Ordinal Logistic Regression

Extension to logistic regression We have multinomial and Ordinal Logistic regression techniques used to predict multiple categorical outcomes. Understand the concept of multi-logit equations, baseline, and making classifications using probability outcomes. Learn about handling multiple categories in output variables including nominal as well as ordinal data.

Topics: Logit and Log-Likelihood, Category Baselining, Modeling Nominal categorical data, Handling Ordinal Categorical Data, Interpreting the results of coefficient values

Advanced Regression for Count Data:

These regression techniques are used to analyze the numeric data known as count data. Based on the discrete probability distributions namely Poisson, negative binomial distribution the regression models try to fit the data to these distributions.

Topics: Poisson Regression, Poisson Regression with Offset, Negative Binomial Regression, Treatment of data with Excessive Zeros, Zero-inflated Poisson, Zero-inflated Negative Binomial, Hurdle Model

Kernel Method - SVM

Support Vector Machines / Large-Margin / Max-Margin Classifier

Topics: Hyperplanes, Best Fit "boundary", Linear Support Vector Machine using Maximum Margin, SVM for Noisy Data, Non- Linear Space Classification, Non-Linear Kernel Tricks, Linear Kernel, Polynomial, Sigmoid, Gaussian RBF, SVM for Multi-Class Classification, One vs. All, One vs. One, Directed Acyclic Graph (DAG) SVM.

Survival Analytics

Survival analysis is about analyzing the duration of time before the event. Learn how survival analysis techniques can be used to understand the effect of the features on the event using the Kaplan-Meier survival plot.

Topics: Examples of Survival Analysis, Time to event, Censoring, Survival, Hazard, and Cumulative Hazard Functions, Introduction to Parametric and non-parametric functions

Decision Tree

Decision Tree models are some of the most powerful classifier algorithms based on classification rules. Elements of classification tree - Root node, Child Node, Leaf Node, etc.

Topics: Greedy algorithm, Measure of Entropy, Attribute selection using Information gain, Decision Tree C5.0 and understanding various arguments, Checking for Underfitting and Overfitting in Decision Tree, Pruning – Pre and Post Prune techniques, Generalization and Regulation Techniques to avoid overfitting in Decision Tree, Random Forest and understanding various arguments, Checking for Underfitting and Overfitting in Random Forest, Generalization and Regulation Techniques to avoid overfitting in Random Forest

Ensemble Techniques:

The parallel and sequential approaches taken in Bagging and Boosting methods are discussed in this module. Random forest is yet another ensemble technique constructed using multiple Decision trees and the outcome is drawn from the aggregating the results obtained from these combinations of trees. You will also learn about stacking methods.

Topics: Overfitting, Underfitting, Voting, Stacking, Bagging, Random Forest, Boosting, AdaBoost / Adaptive Boosting Algorithm, Checking for Underfitting and Overfitting in AdaBoost, Generalization and Regulation Techniques to avoid overfitting in AdaBoost, Gradient Boosting Algorithm, Checking for Underfitting and Overfitting in Gradient Boosting, Generalization and Regulation Techniques to avoid overfitting in Gradient Boosting, Extreme Gradient Boosting (XGB) Algorithm, Checking for Underfitting and Overfitting in XGB, Generalization and Regulation Techniques to avoid overfitting in XGB

Module 6: Forecasting/ Time series

Model-Driven Algorithms:

Time series analysis is performed on the data which is collected with respect to time. The response variable is affected by time

Topics: Introduction to time series data, Steps to forecasting, Components to time series DataScatter plot and Time Plot ,Lag Plot, ACF - Auto-Correlation Function / Correlogram, Visualization principles ,Naïve forecast methods, Errors in the forecast and it metrics - ME, MAD, MSE, RMSE, MPE, MAPE, Model-Based approaches, Linear Model, Exponential Model, Quadratic Model, Additive Seasonality, Multiplicative Seasonality, Model-Based approaches Continued, AR (Auto-Regressive) model for errors, Random walk

Data-Driven Algorithms:

In this continuation module of forecasting learn about data-driven forecasting techniques. Learn about ARMA and ARIMA models which combine model-based and data-driven techniques. Understand the smoothing techniques and variations of these techniques.

Topics: ARMA (Auto-Regressive Moving Average), Order p and q, ARIMA (Auto-Regressive Integrated Moving Average), Order p, d, and q, A data-driven approach to forecasting, Smoothing techniques, Moving Average, Exponential Smoothing, Holt's / Double Exponential Smoothing, Winters / Holt-Winters, De-seasoning and de-trending, Seasonal Indexes

Module 7: Black Box Method

Introduction to Perceptron and Multilayer Perceptron:

The Perceptron Algorithm is defined based on a biological brain model. You will talk about the parameters used in the perceptron algorithm which is the foundation of developing much complex neural network models for AI applications. Understand the application of perceptron algorithms to classify binary data in a linearly separable scenario.

Topics: Neurons of a Biological Brain ,Artificial Neuron, Perceptron, Perceptron Algorithm, Use case to classify a linearly separable data, Multilayer Perceptron to handle non-linear data

Building Blocks of Neural Network - ANN:

Neural Network is a black box technique used for deep learning models. Learn the logic of training and weights calculations using various parameters and their tuning. Understand the activation function and integration functions used in developing a Artificial Neural Network.

Topics: Integration functions, Activation functions, Weights, Bias, Learning Rate (eta) - Shrinking Learning Rate, Decay Parameters, Error functions - Entropy, Binary Cross Entropy, Categorical Cross Entropy, KL Divergence, etc.

Deep Learning Primer

Topics: Artificial Neural Networks,ANN Structure, Error Surface, 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, RMSprop, Adam. Convolution Neural Network (CNN),ImageNet Challenge – Winning Architectures, Parameter Explosion with MLPs, Convolution Networks, Recurrent Neural Network, Language Models, Traditional Language Model, Disadvantages of MLP, Back Propagation Through Time, Long Short-Term Memory (LSTM),Gated Recurrent Network (GRU)

Module 8: Real-Time Data Science Projects
Students will work on a series of real-time data science projects, progressively increasing in complexity.
Projects may include:

- Predictive Modeling for Customer Churn
- Image Classification with CNNs
- Sentiment Analysis on Social Media Data
- Time Series Forecasting for Financial Markets
- Recommender Systems

Module 9: Capstone Project Students will work on an extensive capstone project, bringing together all the concepts learned throughout the course to solve a challenging real-world problem.

Prerequisites:

Basic knowledge of programming (Python recommended).
Familiarity with mathematics concepts such as linear algebra, calculus, and statistics.
Resources:

Online tutorials, articles, and resources for self-study and reference.
Access to real-world datasets and Jupyter Notebooks for hands-on practice.
The duration of the course will depend on the number of hours of instruction per week and the availability of class time for projects and assignments. Additionally, it's essential to adapt the pace based on the students' progress and provide support to help them grasp the concepts effectively