INSAID

GLOBAL CERTIFICATE IN DATA SCIENCE

Online | 6 months | 6 Terms

PROGRAM SNAPSHOT

PRE-TERM PREP

- Python and Stats for Data Science

TERM 01

- Data Analysis basics with Python

TERM 02

- Data Visualization & EDA

TERM 03

- Machine Learning Foundation

PROJECT

- Capstone Project - I

TERM 04

- Machine Learning Intermediate

TERM 05 [ELECTIVE]

- Machine Learning Advanced [E-1]
- Data Analysis with R [E-2]
- Data visualization with Tableau [E-3]

TERM 06

- Capstone Project - II & Industry Immersion

Term 1:

DATA ANALYSIS BASICS WITH PYTHON

Module 1 : Data Science Fundamentals	• Thought Experiment: Data Science from a layman's perspective • Brief intro to Data Science • How companies use Data Science • Overview of Data Science project lifecycle • Walkthrough of data types and data challenges
Module 2 : Recap of Python for Data Science	• In class quiz for Python Basics • Common Python concepts and sample questions • Variable, Inbuilt datatyes, functions, modules and Packages • File operations and error handling
Module 3 : Recap of Statistics for Data Science	• In class quiz for Descriptive Statistics • Common charts used • In class quiz for Inferential Statistics • Probability, Central Limit theorem, Normal Distribution & Hypothesis testing
Module 4 : Mathematical operations using Numpy	• Introduction to Numpy Arrays • How to apply mathematical operations in Numpy • Array manipulation using Numpy • Broadcast values across Arrays using Numpy
Module 5 : Data manipulation with Pandas	• Types of Data Structures in Pandas • Clean data using Pandas • Manipulating data in Pandas • How to deal with missing values • Hands-on: Implement Numpy arrays and Pandas Dataframes

Mini Project :
Data manipulation
Projects

- Project Expectation setting
- Project Timelines
- Github for Project Submission
- Presentation guidelines

Term Projects

Indian Premier League 2008-2018

Analyse key success factors for top cricket team at IPL
The matches dataset contains 18 variables and 600+ observations . The deliveries dataset contains 21 variables and 160k+ observations of the IPL 2018 season.

Car Sales Advertisment

Analyse the car sale data in Ukraine.

The dataset contains 10 variables and 9k+ observations of the car sales data in Ukraine.

Olympic 1896- 2014

Analyse which country have won the most medals at Olympic games.

The dataset contains 9 variables and 31.2k observations of the summer olympic games (1896 - 2014)

1000 movies data

Analyse the IMDB 1000 most popular movies and come up with interesting insights.

The dataset contains 12 variables and 1000 observations of the top 1000 popular movies for past 10 years

Restaurant across America data

Analyse the availability and accessibility of food across America

The dataset contains 10 variables and 10k observation of the fast food restaurants in America.

Term 2:

DATA VISUALIZATION & EDA

Module 1 : Data Visualization in Python - 1

- Plotting basic statistical charts in Python
- Data visualization with Matplotlib
- Case study: Analysis of Wine dataset through visualizations

Module 2 : Data Visualization in Python - 2

- Statistical data visualization with Seaborn
- Interactive data visualization with Bokeh
- Case study: Analysis of Fifa data using Seaborn & Bokeh packages

Module 3 : Exploratory Data Analysis - 1

- Introduction to Exploratory Data Analysis (EDA) steps
- Plots to explore relationship between two variables
- Histograms, Box plots to explore a single variable
- Heat maps, Pair plots to explore correlations
- Case study: Perform EDA to explore survival using Titanic dataset

Module 4 : Industry immersion & project discussion

- Interaction with Industry experts
- QnA

Module 5 : Exploratory Data Analysis - 2

- Case study: Analyse mental health of IT professionals

Module 6:
Project presentation

- Top 3 teams present their projects details to the entire cohort

Term Projects

Indian Premier League 2008-2018

Analyse key success factors for top cricket team at IPL

The matches dataset contains 18 variables and 600+ observations . The deliveries dataset contains 21 variables and 160k+ observations of the IPL 2018 season.

Car Sales Advertisment

Analyse the car sale data in Ukraine.

The dataset contains 10 variables and 9k+ observations of the car sales data in Ukraine.

Olympic 1896- 2014

Analyse which country have won the most medals at Olympic games.

The dataset contains 9 variables and 31.2k observations of the summer olympic games (1896 - 2014)

1000 movies data

Analyse the IMDB 1000 most popular movies and come up with interesting insights.

The dataset contains 12 variables and 1000 observations of the top 1000 popular movies for past 10 years

Restaurant across America data

Analyse the availability and accessibility of food across America

The dataset contains 10 variables and 10k observation of the fast food restaurants in America.

Term 3:

MACHINE LEARNING FOUNDATION

Module 1 : Introduction to Machine Learning (ML)

- What is Machine Learning ?
- Use Cases of Machine Learning
- Types of Machine Learning - Supervised to Unsupervised methods
- Machine Learning workflow

Module 2 : Linear Regression

- Introduction to Linear Regression
- Use cases of Linear Regression
- How to fit a Linear Regression model?
- Evaluating and interpreting results from Linear Regression models
- Case study: How linear regression helps determine demand

Module 3 : Logistic Regression

- Introduction to Logistic Regression
- Logistic Regression use cases
- Understand use of odds & Logit function to perform logistic regression
- Case study:Predicting default cases in the Banking Industry

Module 4 : Decision trees & Random Forests

- Introduction to Decision Trees & Random Forest
- Understanding criterion(Entropy & Information Gain) used in Decision Trees
- Using Ensemble methods in Decision Trees
- Applications of Random Forest
- Case study:Predict passengers survival in a Ship mishap

Module 5 : Model evaluation techniques

- Introduction to evaluation metrics and model selection in Machine Learning
- Importance of Confusion matrix for predictions
- Measures of model evaluation - Sensitivity, specificity, precision, recall & f-score
- Use AUC-ROC curve to decide best model
- Case study:Applying model evaluation techniques to prior case study

Term Projects

Predicting temperatures in World War II

Predict the maximum temperature when minimum temperatue is known.

The dataset contains 31 varibales and 100k+ observations of the weather conditions during World War II.

Candy Classification

Classify if a candy is a chocolate or not based on its features

The dataset contains 13 variables and 86 observations to predict the most or least popular Halloween candy.

Predicting Housing prices

Predict final sales price of each house of residential homes in Iowa.

The train dataset contains 81 variables and 1461 observations. The test dataset contains 80 variables and 1460 observations of the house. prices at any anonymous location.

Predicting risk in Life insurance

Develop a predictive model that accurately classifies risk, imapcting public perception of the industry.

The train dataset contains 128 variables and 59.4k observations. The test dataset contains 127 variables and 19.8k observations of the insurance applicants.

Credit Card Fraud Detection

Identify the fraudulent credit card transactions.

The dataset contains 31 variables and 285k observations of transactions made by credit cards in September 2013 by european cardholders.

Term 4:

MACHINE LEARNING INTERMEDIATE

Module 1 : Dimensionality Reduction using PCA

- Unsupervised Learning: Introduction to Curse of Dimensionality
- What is dimensionality reduction?
- Technique used in PCA to reduce dimensions
- Applications of Principle component Analysis (PCA)
- Case study: Optimize model performance using PCA on high dimension dataset

Module 2 : KNN (K- Nearest neighbours)

- Introduction to KNN
- Calculate neighbours using distance measures
- Find optimal value of K in KNN method
- Advantage & disadvantages of KNN
- Case Study:Classify malicious websites using close neighbour technique

Module 3 : Naïve Bayes classifier

- Introduction to Naïve Bayes classification
- Refresher on Probability theory
- Applications of Naive Bayes Algorithm in Machine Learning
- Case study : Classify Junk emails based on probability

Module 4 : K-means clustering technique

- Introduction to K-means clustering
- Decide clusters by adjusting centroids
- Find optimal 'k value' in kmeans
- Understand applications of clustering in Machine Learning
- Case study : Segment flower species in Iris flower data

Module 5 : Support vector machines (SVM)

- Introduction to SVM
- Figure decision boundaries using support vectors
- Identify hyperplane in SVM
- Applications of SVM in Machine Learning
- Case Study : Predicting wine quality without tasting the wine

Module 6 : Time series forecasting

- Introduction to Time Series analysis
- Stationary vs non stationary data
- Components of time series data
- Interpreting autocorrelation & partial autocorrelation functions
- Stationarize data and implement ARIMA model
- Case Study: Forecast demand for Air travel

Term Projects

SMS Spam collection data

Classify collection of spam messages tagged as spam or legitimate.

The dataset contains 5 variables and 5572 observations collected for SMS spam research.

Simplified Human Activity

Analyse the recordings of subjects performing activities while carrying inertial sensors.

The test dataset contains 562 variables and 1542 observations. the train dataset contains 563 variables and 3609 observations of subjects performing activities while carrying inertial sensors.

Gender recognition by voice

Identify a voice as male or female(SVM)

The dataset contains 21 variables and 3k+ observations to identify a voice as male or female using acoustic properties of voice and speech.

Store Item Demand Forecasting

Predict 3 months of item sales at different store.

The test dataset contains 4 variables and 45000 observations. the train dataset contains 4 variables and 90k observations of a 5 year of store-item-sales data.

Safe driver prediction

Predict the probability that an auto insurance policy holder files a claim.

The test dataset contains 58 variables and 80k+ observations. the train dataset contains 59 variables and 60k+ observations of driving data.

Term 5 [Elective-1]:

MACHINE LEARNING ADVANCED

Module 1 : Introduction to Apriori Algorithm

- Applications of Apriori algorithm
- Understand Association rule
- Developing product recommendations using association rules
- Case study : Analyse online data using association rules

Module 2 : Recommender Systems

- Introduction to Recommender systems
- Types of Recommender systems - collaborative, content based & Hybrid
- Types of similarity matrix (cosine , Jaccard, Pearson correlation)
- Case Study:Build Recommender systems on Movie data

Module 3 : Linear Discriminant Analysis (LDA)

- Recap of dimensionality reduction concepts
- Types of dimensionality reduction
- Dimensionality reduction using LDA
- Case Study : Apply LDA to determine Liquor Quality

Module 4 : Anomaly Detection

- Introduction to Anomaly detection
- How Anomaly detection works?
- Types of Anomaly detection: Density based, Clustering etc.
- Case Study:Detect anomalies on health data

Module 5 : Ensemble learning

- Introduction to Ensemble Learning
- What are Bagging and Boosting techniques?
- What is Bias variance trade off?
- Case study : Predict wage (annual income) classes from adult census data

Module 6 : Stacking

- Introduction to stacking
- Use Cases of stacking
- How stacking improves machine learning models?
- Case Study:Predict survivors in Titanic case

Module 7 : Optimization

- Introduction to optimization in ML
- Applications of optimization methods
- Optimization techniques: Linear Programming using Excel solver
- How Stochastic Gradient Descent(SGD) Works?
- Case study: Apply SGD on Regression data (sklearn dataset)

Module 8 : Neural Networks

- Introduction to Neural networks
- What are Perceptrons & Types of Perceptrons?
- Workflow of a Neural network & analogy with biological neurons
- Case Study : Apply computer vision for digit recognition on MNIST data

Term Projects

Apriori Algorithm(Market Basket Analysis)

Market Basket Analysis of e-commerce data of transaction of 2010 and 2011

The dataset contains 8 variables and 542k observations of all the transaction of 2011 and 2011 for a UK based and registered non - store online retail.

MovieLens Dataset

Predict the name of movies and based upon the reviews of the other critics having similar taste.

The combined dataset consists of 4 different dataset. The links dataset have 3 variables and 9k+ observation, the movies dataset have 3 variables and 9k observation. The ratings dataset have 4 variables and 100k observation. The tags dataset have 4 variables and 1297 observations.

Pokemon Dataset(LDA)

Build a pokemon dream team of 6 pokemon that inflicts the most damage while remaining impervious to any other team of 6 pokemon.

The dataset contains 41 variables and 801 observation of data on pokemon from all 7 generations.

Property Inspection prediction

Predict a transformed count of hazards or pre-existing damages using dataset of property information.

Predict a transformed count of hazards or pre-existing damages using dataset of property information.

Letter recognition

Predict the letter category based on its attributes.

The test dataset contains 17 variables and 4000 observations. the train dataset contains 18 variables and 16k observation of 26 capital letters of english alphabet based on their different attributes.

Term 5 [Elective-2]:
Data Analysis with R

Module	Topics
Module 1 : Data Science Fundamentals	Thought Experiment: Data science @ Google Introduction to Data Science Real world use-cases of Data Science Walkthrough of data types Data Science project lifecycle
Module 2 : Introduction to programming in R	Installing R and R Studio Basic Commands in R Installing packages Setting working directory Exercises: Basic exercises in R Programming
Module 3 : Playing around with Data objects in R	Data structures Basic Data management Loops and Functions Saving output Exercises: Loops and functions in R
Module 4 : Descriptive statistics - 1	Introduction to Statistics Descriptive Statistics Measures of central tendency Measures of Dispersion and shape Case Study: Investigation of Crime statistics in Beaufort
Module 5 : Descriptive statistics - 2	Introduction to Probability Probability Distributions used in Data Science Quantiles, percentiles, and standard score Case Study : Analyse student's performance at school

Module 6 : Inferential Statistics - 1

- Introduction to Inferential Statistics
- Population and Samples
- Central Limit Theorem
- Case Study: Sampling data for Business analysis

Module 7 : Inferential Statistics - 2

- Introduction to Hypothesis Testing
- Confidence Intervals
- Tests of significance: p-value
- Case Study: Apply Inferential statistics & Central limit theorem using Python

Module 8 : Intermediate R: Importing data

- Loading data from R libraries
- Importing data from Excel and CSV files
- Connecting SQL databases
- Webscraping using R
- Case study: Webscraping websites using scrapy package

Module 9 : Intermediate R : Data Manipulation using Tidyverse

- Identifying NULL values in datasets
- Introduction to data imputation methods
- Creating new variables and recoding variables
- Type conversions
- Case Study: Using Tidyverse in Data Manipulation

Module 10 : Intermediate R: Restructuring Data

- Managing Date values
- Numerical and Character functions
- Aggregating & Restructuring data
- Sorting, merging datasets
- Exercises : Subsetting datasets for use in Predictive analytics

Module 11 : Intermediate R : Exploratory data analytics using ggplot2

- Introduction to basic graphs: Barplots, Scatterplots & line graphs
- Using Boxplots in univariate analysis
- Applications of Histograms
- Using ggplot2 for advanced visualizations

Term Projects

Kickstarter funding patterns

Derive insights on successful and failed projects on Kickstarter platform

The dataset contains 15 variables and around 400,000 observations

How bad is the Air Quality in metropolitans?

Analyse worsening airquality in metropolitan cities

The dataset contains 6 variables and over 100 observations

Marketing strategies in Retail banking

Derive insights on how sucessful are the direct marketing campaigns of a Portuguese Bank

The dataset contains 17 variables and over 45000 observations

What are the Characters in superhero comics

Identify the good, bad and the ugly nature of characters in Marvel comics

The dataset contains 11 variables and over 20000 observations

What caused International crisis

Identify all factors which caused major international crisis events in the last 100 years

The dataset contains 96 Variables and over 1000 observations

Term 5 [Elective-3]:

DATA VISUALIZATION WITH TABLEAU

Module 1 : Introduction to Visual Analytics

- Introduction to data visualization
- Understanding Tableau ecosystem in industry
- Loading data files in Tableau
- Creating first visualizations
- Case Study: Sales performance Analysis

Module 2 : Data Visualization using Tableau

- Introduction to graphs - bar graph and line graph
- Working with continuous measures & discrete variables
- Heat maps and Geographical data visualizations
- Creating map Views
- Case Study: Analyse Natural calamity trend and effect

Module 3 : Data joining & blending in Tableau

- Introduction to SQL joins
- Performing data blending in Tableau
- Creating dual axis charts in Tableau
- Introduction to descriptive statistics and Visual analytics
- Case Study : Analyse trends in Retail businesses

Module 4 : Predicitve Analytics using Tableau and R

- Introduction to R programming tool & R studio
- Installing R and R studio
- Applications of linear regression in prediction
- Data crunching: Creating groups, sets & parameters
- Case Study: Forecast revenues

Module 5 : Interactive Dashboard Design

- Introduction to principles of dashboard design
- Custom geocoding in Tableau
- Developing dashboard products using Tableau
- Introduction to writing storyline in Tableau
- Case Study: Customer segmentation dashboard

Module 6 : Advanced Calculations using Tableau

- Introduction to calculations: Date calculations
- Using LOD calculations: INCLUDE, EXCLUDE & FIXED functions
- Working with Table calculations
- Exporting data from Tableau
- Case Study: Analyse sales across geographies, products & customers

Module 7 : Applications of advanced Calculations using Tableau

- Introduction to customer churn analysis
- Estimating customer life time value
- Applications of context filtering
- Applications of logical functions in Tableau
- Case Study: Analyse retail sales data to predict customer behaviour

Module 8 : Revision of concepts and Project discussion

- Revision of key concepts: data blending, writing calculations, LOD calcuations etc.
- Review of Tableau project portfolio
- Communicating data insights using reporting tools
- Tableau Interview prep
- Discussing EDA objectives of final project

Term Projects

Hubway data visualization challenge

Produce visualizations that reveal interesting user patterns about how people in Boston gets around on Hubway

The dataset contains 1 million observations on bike usage by residents of Boston

Term 6:

CAPSTONE PROJECT - I

James Telco Bond

In this capstone project, students will be provided with data collected by a major Telecom operator on the demographic behaviour of users using different handsets.

Students are required to do the initial bit of data cleansing, pre-processing and then upload this data to SQL server via a web hosting platform that will be provided to them.

This data from SQL server will be used to create a dashboard for the company using D3.js scripts. D3,js scripts will be provided to students upfront. These dashboards are reflective of how interactive visualizations can help companies make strategies such as what demographies to cater to, how men and women customers behave differently, which geographies are popular and ones that need more investment from the company in terms of finance and marketing?

Term 6:

CAPSTONE PROJECT - II

Demand Planners

This capstone project will focus more on applying machine learning concepts rather than data gathering and storing aspects. Students will be provided with data collected by a major Taxi Aggregator of taxi bookings done in a leading city. As budding data science consultants, students are required to do exploratory data analysis & present an initial report.

After that the students are required to create an UI that displays the observations regarding taxi usage across the city from the analysis and the website should also have a provision for the company to forecast demand for taxis at a specific time in the day.

The taxi bookings data provided will be in csv format and dashboards for the company need to be created using D3.js scripts. The D3.js scripts will be provided to the students beforehand.

Need to know

Program Start

First week of every month

Duration

- 06 months (Incl. Capstone Projects)

Prerequisite

- Background in Programming (Not Mandatory)
- Laptop with 4 GB RAM

Program Fee

INR 2 Lakh + GST

Scholarships

50 scholarships (each with 70% tuition waiver) for professionals passionate about making a career in Data Science & furthering INSAID's mission of putting India on the global AI map.

Talk to our Admissions team today or **attend the next Data Science MasterClass** to know more

For Further details, write to us info@insaid.co