Course Content

**Unit 1 Linux     -    5 hours**

Introduction to Linux  – Establishing the fundamental knowledge of how linux works and how you can begin with Linux OS. Linux Basics – File Handling, data extraction, etc.

**Unit 2 Version Control  - 4 hours**

What is version control, types, SVN. GIT, Git Lifecycle, Common Git commands, Working with branches in Git. Github collaboration (pull request), Github Authentication (ssh and Http), Merging branches, Resolving merge conflicts, Git workflow

**Unit 3 Data Analysis With MS-Excel  -  8 hours**

Reading the Data, Referencing in formulas , Name Range, Logical Functions, Conditional Formatting, Advanced Validation, Dynamic Tables in Excel, Sorting and Filtering. Working with Charts in Excel, Pivot Table, Dashboards, Data And File Security. VBA Macros, Ranges and Worksheet in VBA. IF conditions, loops, Debugging, etc. Handling Text Data, Splitting, combining, data imputation on text data, Working with Dates in Excel, Data Conversion, Handling Missing Values, Data Cleaning, Working with Tables in Excel, etc. Charts, Pie charts, Scatter and bubble charts. Bar charts, Column charts, Line charts, Maps. Multiples: A set of charts with the same axes, Matrices, Cards, Tiles. Power Pivot, Power Query and Power View.

**Unit 4 Data Analysis with SQL – 8 hours**

Fundamentals of Structured Query Language, SQL Tables, Joins, Variables, Advanced SQL – SQL Functions, Subqueries, Rules, Views, Nested Queries, string functions, pattern matching, Mathematical functions, Date-time functions, etc. , User Defined Functions, Types of UDFs, Inline table value, multi-statement table. Stored procedures, rank function, SQL ROLLUP, etc. SQL Optimization and Performance -  Record grouping, searching, sorting, etc., Clustered indexes, common table expressions.

**Unit 5 Python for Data Science – 15 hours**

Introduction to Python and IDEs – The basics of the python programming language, how you can use various IDEs for python development like Jupyter, Pycharm, etc.

Python Basics – Variables, Data Types, Loops, Conditional Statements, functions, decorators, lambda functions, file handling, exception handling ,etc.

Object Oriented Programming – Introduction to OOPs concepts like classes, objects, inheritance, abstraction, polymorphism, encapsulation, etc.

 Extract Transform Load - Web Scraping, Interacting with APIs

Data Handling with NumPy- NumPy Arrays, CRUD Operations, etc. Linear Algebra – Matrix multiplication, CRUD operations, Inverse, Transpose, Rank, Determinant of a matrix, Scalars, Vectors, Matrices.

Data Manipulation Using Pandas- Loading the data, data frames, series, CRUD operations, splitting the data, etc.

Data Preprocessing - Exploratory Data Analysis, Feature engineering, Feature scaling, Normalization, standardization, etc. Null Value Imputations, Outliers Analysis and Handling, VIF, Bias-variance trade-off, cross validation techniques, train-test split, etc.

Data Visualization - Bar charts, scatter plots, count plots, line plots, pie charts, donut charts, etc. with Python matplotlib. Regression plots, categorical plots, area plots, etc, with Python seaborn.

**Unit 6 Linear Algebra and Statistics for Data Science**

Descriptive Statistics –           Measure of central tendency, a measure of spread, five points summary, etc.

Probability - Probability Distributions, bayes theorem, central limit theorem.

Inferential Statistics –  Correlation, covariance, confidence intervals, hypothesis testing, F-test, Z-test, t-test, ANOVA, chi-square test, etc.

**Practical- Linear algebra with Python and Excel**

**Unit 7 Machine learning  - 25 hours**

**Supervised Learning**- Linear Regression – Creating linear regression models for linear data using statistical tests, data preprocessing, standardization, normalization, etc.            Logistic Regression – Creating logistic regression models for classification problems – such as if a person is diabetic or not, if there will be rain or not, etc.            Decision Tree – Creating decision tree models on classification problems in a tree-like format with optimal solutions. Random Forest – Creating random forest models for classification problems in a supervised learning approach. Support Vector Machine – SVM or support vector machines for regression and classification problems. Gradient Descent – The gradient descent algorithm is an iterative optimization approach to finding local minimum and maximum of a given function. K-Nearest Neighbours – A simple algorithm that can be used for classification problems. Time Series Forecasting – Making use of time series data, gathering insights and useful forecasting solutions using time series forecasting.

**Unsupervised Learning** - K-means – The k-means algorithm that can be used for clustering problems in an unsupervised learning approach.        Dimensionality reduction – Handling multi-dimensional data and standardizing the features for easier computation.  Linear Discriminant Analysis –  LDA or linear discriminant analysis to reduce or optimize the dimensions in the multidimensional data. Principal Component Analysis – PCA follows the same approach in handling multidimensional data.

**Regression**- Introduction classification problems, Identification of a regression problem, dependent and independent variables.   Training, Evaluation and optimization of regression Model.

**Classification**- Introduction to classification problems, Identification of a classification problem, and dependent and independent variables. Train, evaluate, and optimize classification problems.

**Clustering**- Introduction to clustering problems, Identification of a clustering problem, and dependent and independent variables. Train, evaluate, and optimize clustering problems.

**Performance Metrics -** Classification reports – To evaluate the model on various metrics like recall, precision, f-support, etc. Confusion matrix – To evaluate the true positive/negative, and false positive/negative outcomes in the model. r2, adjusted r2, mean squared error, etc.

**Practical: Machine Learning with Python, Excel, scikit-learn, Keras**

**Unit 8  Deep Learning -  10 hours**

Artificial Intelligence Basics, Introduction to keras API and tensorflow

Neural Networks - Neural networks, Multi-layered Neural Networks, Artificial Neural Networks

Deep neural networks-          Convolutional Neural Networks, Recurrent Neural Networks, GPU in deep learning, Autoencoders, restricted boltzmann machine

**Practical :  Deep Learning using Keras and tensorflow**

**Unit 9 Natural Language Processing – 15 hours**

 Text Mining, Cleaning, and Pre-processing**-**Various Tokenizers, Tokenization, Frequency Distribution, Stemming, POS Tagging, Lemmatization, Bigrams, Trigrams & Ngrams, Lemmatization, Entity Recognition.

**Text classification, NLTK, sentiment analysis, etc**- Overview of Machine Learning, Words, Term Frequency, Countvectorizer, Inverse Document Frequency, Text conversion, Confusion Matrix, Naive Bayes Classifier.

**Sentence Structure, Sequence Tagging, Sequence Tasks, and Language Modeling** - Language Modeling, Sequence Tagging, Sequence Tasks, Predicting Sequence of Tags, Syntax Trees, Context-Free Grammars, Chunking, Automatic Paraphrasing of Texts, Chinking.

**AI Chatbots and Recommendations Engine** - Using the NLP concepts, build a recommendation engine and an AI chatbot assistant using AI.

**Practical: NLP using R/Python**

**Unit 10 Data Science At Scale with Pyspark – 10 hours**

Introduction to Big Data And Spark- Apache spark framework, RDDs, Stopgaps in existing computing methodologies.

RDDs- RDD persistence, caching, General operations: Transformation, Actions, and Functions. Concept of Key-Value pair in RDDs, Other pair, two pair RDDs. RDD Lineage, RDD Persistence, WordCount Program Using RDD Concepts. RDD Partitioning & How it Helps Achieve Parallelization

Advanced Concepts & Spark-Hive- Passing Functions to Spark, Spark SQL Architecture, SQLContext in Spark SQL. User-Defined Functions, Data Frames, Interoperating with RDDs. Loading Data through Different Sources, Performance Tuning.   Spark-Hive Integration