Data Science and Big Data Curriculum

Introduction (1.5hrs):

- What is Data Science?
- What is Big Data?
- What is Machine Learning?
- What is Analytics?
- What is Data analysis and Data Mining?
- Analytics project life cycle
- Real life applications, projects and career paths of Data Science and Big Data

- **Statistics:**
  - ➢ Definition and computation of probability
  - ➢ Measurement of central tendencies and it's applications
  - ➢ Spreads, Distributions(Normal, Z-distribution, Binomial, Poisson) and various types of probability distributions(Continuous and discrete)
  - ➢ Sampling and Sampling distributions
  - ➢ Measures of shape(Skewness and Kurtosis)
  - ➢ Measures of relationship between variables(Correlation, causation)
  - ➢ Hypothesis Testing(t-test, Chi-square, Anova)
  - ➢ Measures of Dispersion(Variance, Std. deviation, Range)
  - ➢ Prediction and Confidence interval-Computation and Analysis
  - ➢ Missing Value theorem

- **Exploratory Data Analysis(EDA) and Data Visualization**
  - ➢ What is EDA and why is it required?
  - ➢ Outlier treatment
  - ➢ Data distributions and transformations
  - ➢ Graphs
  - ➢ Bar charts
  - ➢ Histogram
  - ➢ Box-Whisker plot
  - ➢ Scatter plot
  - ➢ Variable selection
  - ➢ Bubble charts
  - ➢ Exam

- **Introduction to R :**
  - ➢ Why R and importance of R in Analytics?
  - ➢ Installation of R and R-studio
  - ➢ Data types
  - ➢ Variables
  - ➢ Operators
  - ➢ Decision making
  - ➢ Loops
  - ➢ Lists
  - ➢ Vectors
  - ➢ Strings
  - ➢ Matrices
  - ➢ Arrays
  - ➢ Factors
  - ➢ Functions (Built in and User defined functions)(aggregate,subset,merge,lapply, sapply, as.xxxx , which,sort,order- mandatory)
  - ➢ Importing Data from texts , spreadsheets and webdata
  - ➢ Extracting Tweets from Twitter using API
  - ➢ Data frames
  - ➢ Packages , libraries and their installation
  - ➢ Data manipulation and re-shaping
  - ➢ Data Visualization using R
  - ➢ Exam

- Introduction to Python programming:
  - ➢ What is Python?
  - ➢ History
  - ➢ Why is Python preferred for Data Science?
  - ➢ Installation of I python/Jupyter Notebook/ SPYDER

- Basics of Python:
  - ➢ Keywords
  - ➢ Built-in functions
  - ➢ String Formatting
  - ➢ Lists
  - ➢ Loops
  - ➢ Tuples
  - ➢ Indexing
  - ➢ Slicing
  - ➢ Sequences
  - ➢ Dictionaries
  - ➢ Sets
  - ➢ Importing and exporting data from python into various formats
- Functions
  - ➢ User defined functions
  - ➢ Parameters
  - ➢ Nested functions
  - ➢ Local and Global variables
  - ➢ Alternate Keys
  - ➢ Lambda functions
  - ➢ Sorting Lists and Dictionaries
  - ➢ Sorting Collections

- Error and Exception handling
  - ➢ Errors in Python
  - ➢ Abnormal termination
  - ➢ Exception handling methods
  - ➢ Ignoring Errors
  - ➢ Assertions and effective usage of assertions

- OOPS, Packages and Libraries in Python

- ➢ Methods and Inheritance
- ➢ Abstraction and Encapsulation
- ➢ Classes
- ➢ Walking Directory Trees
- ➢ Initializes
- ➢ Instance methods
- ➢ Class methods
- ➢ Data Static Methods
- ➢ Expressions
- ➢ Module Aliases
- ➢ Math functions
- ➢ Random Numbers
- ➢ Package Installation Methods
- ➢ Introduction to Numpy, Pandas and other libraries
- ➢ Plotting in Python
- ➢ Creating Data Frames
- ➢ Data Manipulation
- ➢ Slicing and Dicing

Machine Learning

- **Supervised Learning:**
  - ➢ What is supervised learning
  - ➢ Algorithms in Supervised learning
  - ➢ Steps in Supervised learning

✓ **Regression & Classification :**

- ➢ Regression vs classification
- ➢ Computation of co-relation coefficient and Analysis
- ➢ Performance and accuracy measurement of a Model
- ➢ Naive Baye's classifier
- ➢ Model Training, Validation and Testing
- ➢ Ordinary Least squares

- Variable selection
- R-Square coefficient and RMSE as a strength of model
- Prediction and confidence interval determination and application
- Proviso of Regression
- Dummy variables
- Types of Regression: Linear and Logistic(Simple and multiple)
- Sum of least squares
- ROC and AUC curves
- Homoscedasticity and Heteroscedasticity
- Multicollinearity and vif
- Confusion matrix
- Techniques to improve accuracy and performance of regression models
- Assignment

Decision Trees and Random Forest Test

- Introduction to Decision tree Algorithms and it's applications
- Classification and regression trees-CART models,ID3,C4.5
- CHAID analysis
- Building Decision Trees using R
- Decision nodes and leaf nodes
- Variable Selection, Parent and child nodes branching
- Stopping Criterion
- Tree pruning
- Depth of a tree
- Overfitting
- Metrics for decision trees-Gini impurity, Information Gain, Variance Reduction
- Regression using decision tree
- Interpretation of a decision tree using If-else
- Pros and cons of a decision tree
- Introduction to Random forest test and it's applications

- ➢ Why Random forest test?
- ➢ Tree bagging
- ➢ Models and algorithms in Random Forest test
- ➢ Training Data set, Tree grouping and decision making on majority voting
- ➢ Boosting algorithms-Gradient Boosting, Adaptive boosting-Adaboost , Xgboost (Advanced)
- ➢ Accuracy estimation using cross validation

➢ **KNN-algorithm:**

- ➢ What is KNN and why do we use it?
- ➢ KNN-algorithm and regression
- ➢ Curse of dimensionality and brief introduction to dimension reduction
- ➢ KNN-outlier treatment and anomaly detection
- ➢ Cross Validation
- ➢ Pros and cons of KNN

➢ **Support Vector Machines**

- ➢ Linear and Non-Linear SVM's
- ➢ SVM regression
- ➢ Train time and Run time complexities
- ➢ Kernel Methods

- **Unsupervised Learning:**
  - ➢ What is unsupervised learning?
  - ➢ Algorithms in unsupervised learning
  - ➢ Steps in unsupervised learning

- **Dimensionality Reduction:**
  - ➢ Introduction to dimensionality reduction and it's necessity
  - ➢ Principal Component Analysis(PCA)
  - ➢ Singular Value Decomposition(SVD)
  - ➢ Kernel-PCA
  - ➢ Linear Discriminant Analysis
  - ➢ Feature extraction
  - ➢ Advantages and applications of Dimensionality reduction

- **Clustering**
  - ➢ Introduction to clustering
  - ➢ Real-life applications of clustering
  - ➢ Distance measurement methods
  - ➢ Hierarchical clustering
  - ➢ K-Means clustering and skew plot
  - ➢ Assignment

- **Text Mining**
  - ➢ Introduction to Text Mining
  - ➢ Applications
  - ➢ Structured and unstructured data
  - ➢ Extracting unstructured text from files and websites
  - ➢ Data cleaning and reshaping
  - ➢ Terminologies in Text Mining
  - ➢ Text clustering and categorization
  - ➢ Word cloud
  - ➢ N-gram charts
  - ➢ Sentiment Analysis
  - ➢ Twitter Analytics
  - ➢ Natural Language processing
  - ➢ Assignment

- Forecasting
  - ➢ Introduction to forecasting
  - ➢ Applications
  - ➢ Data Manipulation and Cleaning
  - ➢ Time Series
  - ➢ Time Series forecasting
  - ➢ Components of Time Series-Trend, Seasonality, Randomness
  - ➢ Trend Analysis
  - ➢ Forecasting methods
  - ➢ Smoothing Methods
  - ➢ Modeling Random Components
  - ➢ Modeling for stationary time series
  - ➢ ETS Model
  - ➢ Auto regressive Model
  - ➢ Moving Average Model
  - ➢ ARIMA Model
  - ➢ ETS Model
  - ➢ Anomaly Detection
  - ➢ Transformations
  - ➢ Growth curve
  - ➢ ARCH & GARCH Models

- **Association rules:**
  - ➢ Introduction
  - ➢ Importance of Association rules
  - ➢ Metrics of rules-Lift, Support, Confidence, Conviction
  - ➢ Apriority Model
  - ➢ Market Basket Analysis
  - ➢ Algorithm implementation and tuning
  - ➢ Applications
  - ➢ Assignment