Unit 1: Basics and need of Data Science and Big Data

- Data Science: Interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data.
- Big Data: Refers to datasets that are too large or complex for traditional data processing applications to handle.
- Applications of Data Science: Span various industries including healthcare, finance, marketing, and more, to make data-driven decisions.
- Data explosion: Rapid increase in the volume, velocity, variety, veracity, and value of data.
- 5 V's of Big Data: Volume, Velocity, Variety, Veracity, and Value.
- Relationship between Data Science and Information Science: Data Science focuses on extracting insights from data, while Information Science deals with the organization and retrieval of information.
- Business intelligence versus Data Science: Business Intelligence primarily deals with historical data analysis for decision-making, while Data Science involves predictive and prescriptive analytics.
- Data Science Life Cycle: Process comprising data collection, data wrangling, data exploration, model building, interpretation, and deployment.
- Data Types: Categorical, numerical, ordinal, and nominal.
- Data Collection: Gathering data from various sources using methods like surveys, sensors, and web scraping.
- Need of Data Wrangling: Process of cleaning, transforming, and enriching raw data for better analysis.
- Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.

Unit 2: Need of statistics in Data Science and Big Data Analytics

- Statistics in Data Science: Provides tools and techniques for analyzing and interpreting data.
- Measures of Central Tendency: Mean, Median, Mode, Mid-range.
- Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation.
- Bayes Theorem: Probability theorem used to update the probability of a hypothesis based on new evidence.
- Hypothesis Testing: Process of making inferences about a population parameter based on sample data.
- Pearson Correlation: Measure of the linear relationship between two variables.

- Sample Hypothesis Testing: Testing hypotheses about population parameters using sample data.
- Chi-Square Tests: Statistical tests used to determine if there is a significant association between categorical variables.
- t-test: Statistical test used to determine if there is a significant difference between the means of two groups.

Unit 3: Introduction to Big Data

- Sources of Big Data: Data originates from various sources such as social media, sensors, and IoT devices, leading to large and diverse datasets.
- Data Analytic Lifecycle:
- Discovery: Identifying data sources and understanding their characteristics.
- Data Preparation: Cleaning, integrating, and transforming raw data for analysis.
- Model Planning: Defining objectives, selecting appropriate models, and designing analysis strategies.
- Model Building: Constructing and training analytical models using algorithms and techniques.
- Communication of Results: Presenting findings and insights derived from data analysis to stakeholders.
- Operationalization: Integrating analytical solutions into operational systems for real-world applications.

Unit 4: Data Preprocessing and Analytics

- Essential Python Libraries: Pandas, NumPy, Matplotlib are essential libraries for data manipulation, numerical computing, and visualization in Python.
- Data Preprocessing:
- Removing Duplicates: Identifying and eliminating duplicate entries to maintain data integrity.
- Transformation of Data: Converting data into a suitable format for analysis using functions or mapping.
- Handling Missing Data: Strategies for dealing with missing values, such as imputation or deletion.
- Analytics Types:
- Predictive Analytics: Forecasting future trends and outcomes based on historical data and statistical techniques.
- Descriptive Analytics: Summarizing and visualizing past data to gain insights into trends and patterns.

- Prescriptive Analytics: Recommending actions or decisions based on analytical insights to optimize outcomes.
- Association Rules: Algorithms like Apriori and FP-growth for discovering relationships and patterns in large datasets.
- Regression: Techniques such as Linear Regression and Logistic Regression for modeling the relationship between variables and making predictions.
- Classification: Algorithms like Naïve Bayes and Decision Trees for categorizing data into predefined classes or categories.

Unit 5: Clustering, Text Analysis, and Model Evaluation

- Clustering Algorithms:
- K-Means: Partitioning data into clusters based on similarity, aiming to minimize intra-cluster variance.
- Hierarchical Clustering: Creating a hierarchy of clusters by recursively merging or splitting data points.
- Time-series Analysis: Analyzing temporal data to identify patterns, trends, and anomalies over time.
- Introduction to Text Analysis:
- Text Preprocessing: Cleaning and tokenizing text data, removing stopwords, and stemming or lemmatizing words.
- Bag of Words: Representing text data as a matrix of word frequencies for analysis.
- TF-IDF: Term Frequency-Inverse Document Frequency weighting to measure the importance of terms in a document corpus.
- Introduction to Social Network Analysis: Analyzing the structure and dynamics of social networks to understand relationships and behaviors.
- Introduction to Business Analysis: Applying analytical techniques to solve business problems and optimize processes.
- Model Evaluation and Selection:
- Metrics for Evaluating Classifier Performance: Measures like accuracy, precision, recall, and F1-score to assess model performance.
- Holdout Method and Random Subsampling: Techniques for splitting data into training and testing sets to evaluate model performance.
- Parameter Tuning and Optimization: Adjusting model parameters to improve performance using techniques like grid search or cross-validation.
- Result Interpretation: Interpreting model outputs and insights to make informed decisions or recommendations.

Unit 6: Data Visualization and Tools

- Introduction to Data Visualization: Data visualization is the graphical representation of data to facilitate understanding, analysis, and communication of insights. It includes various techniques and methods to present data effectively.
- Challenges to Big Data Visualization: Big data visualization faces challenges such as scalability, complexity, and heterogeneity due to the large volume, variety, and velocity of data. Overcoming these challenges requires specialized tools and techniques.
- Types of Data Visualization: Data can be visualized in various forms, including charts, graphs, maps, and dashboards, each suitable for different types of data and analysis purposes.
- Data Visualization Techniques: Techniques such as bar charts, line plots, scatter plots, histograms, heatmaps, and box plots are commonly used to visualize different aspects of data, such as distributions, trends, correlations, and relationships.
- Visualizing Big Data: Visualizing large and complex datasets requires specialized tools and techniques that can handle the scale and complexity of big data, such as parallel processing, interactive visualization, and summarization techniques.
- Tools used in Data Visualization: There are various tools available for data visualization, ranging from general-purpose tools like Matplotlib, Seaborn, and Plotly in Python to specialized tools like Tableau, Power BI, and D3.js. These tools offer different features and capabilities for creating interactive and visually appealing visualizations.
- Hadoop Ecosystem: The Hadoop ecosystem is a collection of open-source software tools and frameworks for distributed storage and processing of big data. It includes components like HDFS, MapReduce, Pig, Hive, HBase, Spark, and more, each serving different purposes in the big data processing pipeline.
- MapReduce: MapReduce is a programming model and framework for processing and analyzing large datasets in parallel across distributed clusters of computers. It divides the processing tasks into two phases: map and reduce, and is commonly used in Hadoop-based systems.
- Pig: Apache Pig is a high-level platform for creating and executing data flow programs for processing and analyzing large datasets. It provides a simple and expressive scripting language called Pig Latin, which abstracts away the complexities of MapReduce programming.
- Hive: Apache Hive is a data warehouse infrastructure built on top of Hadoop for querying and analyzing large datasets stored in Hadoop HDFS. It provides a SQL-like query language called HiveQL, which allows users to perform ad-hoc queries and analysis on structured data stored in Hadoop.

- Analytical Techniques used in Big Data Visualization: Various analytical techniques, such as clustering, classification, regression, and anomaly detection, can be combined with visualization to gain deeper insights and uncover hidden patterns in big data. These techniques enable data-driven decision-making and business intelligence.