1. **Data Science**: Data science is a field that involves extracting insights and knowledge from large sets of structured and unstructured data. It combines elements of statistics, computer science, and domain knowledge to analyze data and make informed decisions.

2. **Confusion Matrix:** A confusion matrix is a table that is used to evaluate the performance of a classification model. It compares the actual values of a dataset to the predicted values generated by the model, showing the number of true positives, true negatives, false positives, and false negatives.

3. **Linear & Logistic Regression:** Linear regression is a statistical method used to analyze the relationship between two continuous variables. It aims to find the best-fitting straight line that predicts the value of one variable based on the value of another. Logistic regression, on the other hand, is used for binary classification problems, where the outcome variable is categorical.

4. **Data Analytics Life Cycle:** The data analytics lifecycle consists of several stages, including data collection, data preparation, data analysis, interpretation of results, and decision-making based on those results. It's a systematic approach to extracting insights from data to solve business problems or make informed decisions.

5. **Data Wrangling:** Data wrangling, also known as data munging, is the process of cleaning, structuring, and enriching raw data into a format suitable for analysis. It involves tasks like handling missing values, removing duplicates, transforming data types, and combining data from different sources.

6. **Techniques for Handling Missing Values:** Techniques for handling missing values include imputation, where missing values are replaced with estimated values based on other observations, and deletion, where observations with missing values are removed from the dataset.

7. **Outliers:** Outliers are data points that significantly differ from other observations in a dataset. They can skew statistical analyses and lead to inaccurate results if not properly addressed.

8. **Techniques to detect outliers:** Techniques to detect outliers include visual methods such as scatter plots and box plots, statistical methods such as z-score and interquartile range (IQR), and machine learning algorithms designed to identify anomalies in data.

9. **Hypothesis Testing**: Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves formulating a null hypothesis, which states that there is no significant difference or relationship, and an alternative hypothesis, which states the opposite.

10. **Null Hypothesis, Alternative Hypothesis:** The null hypothesis (H0) is a statement that there is no significant difference or relationship between variables, while the alternative hypothesis (H1) is a statement that there is a significant difference or relationship.

11. **Data Visualization:** Data visualization is the graphical representation of data to visually communicate insights and patterns. It includes charts, graphs, maps, and other visualizations that help stakeholders understand complex data sets more easily.

12. **Tools & Techniques for Data Visualization**: Tools for data visualization include software like Tableau, Power BI, and Python libraries such as Matplotlib and Seaborn. Techniques include selecting appropriate chart types, choosing color schemes, and designing interactive dashboards.

13. **Decision Tree, KNN:** Decision tree is a supervised learning algorithm used for classification and regression tasks. It recursively splits the dataset into subsets based on the most significant attribute, creating a tree-like structure to make predictions. K-Nearest Neighbors (KNN) is a classification algorithm that predicts the class of a data point by majority voting among its k nearest neighbors in the feature space.

14. **Naive Bayes:** Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that the features in a dataset are independent of each other, hence the term "naive." Despite this simplifying assumption, Naive Bayes can be surprisingly effective in many real-world classification tasks, particularly in text classification and spam filtering

15. **Bayes Theorem:** Bayes' theorem is a fundamental concept in probability theory that describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is expressed mathematically as P(A|B) = [P(B|A) * P(A)] / P(B), where P(A|B) is the probability of event A given that event B has occurred, P(B|A) is the probability of event B given that event A has occurred, P(A) is the prior probability of event A, and P(B) is the prior probability of event B.

16. Predictive, Prescriptive, Descriptive:

- Descriptive Analytics: Descriptive analytics focuses on describing past events and understanding historical data. It helps to summarize and visualize what has happened in the past, providing insights into trends, patterns, and anomalies. Examples include summary statistics, dashboards, and data visualization techniques.
- Predictive Analytics: Predictive analytics aims to forecast future outcomes or trends based on historical data and statistical algorithms. It involves using machine learning and statistical modeling techniques to identify patterns and make predictions about future events. Predictive analytics can be used for various purposes, such as forecasting sales, predicting customer behavior, and identifying potential risks.
- Prescriptive Analytics: Prescriptive analytics goes beyond predicting future outcomes by providing recommendations on actions to take to achieve desired outcomes. It combines

data analysis, optimization, and simulation techniques to generate actionable insights and suggest the best course of action. Prescriptive analytics can help organizations make informed decisions and optimize business processes in real-time.

-

17. Hadoop, MapReduce:

- Hadoop: Hadoop is an open-source framework designed for distributed storage and processing of large datasets across clusters of commodity hardware. It consists of two main components: the Hadoop Distributed File System (HDFS) for storing data across multiple nodes, and the MapReduce programming model for processing and analyzing data in parallel.
- MapReduce: MapReduce is a programming model and processing engine used to process and analyze large datasets in parallel across a distributed cluster. It divides a task into two main phases: the Map phase, where data is processed in parallel across multiple nodes to generate intermediate key-value pairs, and the Reduce phase, where the intermediate results are aggregated and combined to produce the final output. MapReduce is the core processing framework in Hadoop and is widely used for batch processing tasks like data cleaning, transformation, and analysis.You

18) Apriori

Apriori is a popular algorithm used in data mining and association rule learning. It is primarily used for discovering frequent item sets and generating association rules from transactional data. The algorithm works based on the concept of support and confidence.

- Support: Support measures the frequency of occurrence of an item set in a dataset. It indicates how often an item set appears in the transactions.
- Confidence: Confidence measures the reliability of the association rule. It indicates the likelihood that an item B is purchased when item A is purchased.

The Apriori algorithm works in iterations to discover frequent item sets. It starts by finding all frequent individual items (singletons) in the dataset and then iteratively generates larger item sets by combining frequent item sets found in the previous iteration. During each iteration, candidate item sets are pruned if they do not meet a minimum support threshold.

Apriori algorithm is widely used in market basket analysis, where it helps identify relationships between items frequently purchased together. It has applications in recommendation systems, cross-selling strategies, and retail analytics.