Q1

1. **State Occam's razor principle.**
   Occam's razor states that among competing hypotheses, the one with the fewest assumptions should be selected.
2. **Define Data Analytics.**
   Data Analytics is the process of examining, transforming, and modeling data to extract insights, identify patterns, and support decision-making.
3. **What is supervised learning?**
   Supervised learning is a type of machine learning where a model is trained on labeled data, meaning each input has a corresponding correct output.
4. **What is TF-IDF?**
   TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used to evaluate the importance of a word in a document relative to a collection of documents.
5. **What is frequent itemset?**
   A frequent itemset is a set of items that appear together frequently in a dataset, commonly used in association rule mining.
6. **Define stemming.**
   Stemming is the process of reducing words to their root or base form by removing suffixes (e.g., "running" to "run").
7. **What is Link Prediction?**
   Link prediction is a technique used in network analysis to predict future or missing connections between entities in a graph.
8. **State applications of AI.**
   AI is used in healthcare (diagnostics), finance (fraud detection), customer service (chatbots), autonomous vehicles, and recommendation systems.
9. **State types of logistic regression.**
   The types of logistic regression are binary logistic regression, multinomial logistic regression, and ordinal logistic regression.
10. **Define precision.**
    Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, measuring a model's accuracy in identifying true positives.
11. **What is AVC & ROC curve?**
    ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates a classifier's performance. AUC (Area Under Curve) quantifies the overall ability of the model to distinguish between classes.
12. **Write any two applications of supervised machine learning.**
    Supervised machine learning is used in spam email detection and medical diagnosis prediction.
13. **Give the formula for support & confidence.**
    Support = (Transactions containing both X and Y) / (Total transactions)
    Confidence = (Transactions containing both X and Y) / (Transactions containing X)

14. **What is an outlier?**
    An outlier is a data point that significantly differs from the rest of the dataset, potentially indicating an

error or a rare event.

15. **State applications of NLP.**
    NLP is used in sentiment analysis, language translation, speech recognition, and chatbots.

16. **What is web scraping?**
    Web scraping is the automated process of extracting data from websites using scripts or tools.

17. **What is the purpose of n-gram?**
    N-gram helps in text analysis by capturing the sequence of 'n' words together to analyze patterns in language processing.

18. **Define classification.**
    Classification is a machine learning technique that categorizes data into predefined classes based on input features.

19. **Define recall.**
    Recall is the ratio of correctly predicted positive observations to all actual positives, measuring a model's ability to detect true positives.

20. **Define tokenization.**
    Tokenization is the process of breaking text into smaller units, such as words or sentences, for analysis.

21. **Define machine learning.**
    Machine learning is a branch of AI that enables computers to learn from data and improve performance without explicit programming.

22. **What is clustering?**
    Clustering is an unsupervised learning technique that groups similar data points together based on shared characteristics.

23. **What is data characterization?**
    Data characterization is the process of summarizing general features of data, often through statistical descriptions.

24. **What is Bag of Words?**
    Bag of Words (BoW) is a text representation method that converts documents into word frequency vectors without considering word order.

25. **What is text analytics?**
    Text analytics is the process of extracting meaningful insights from text data using techniques like sentiment analysis and keyword extraction.

26. **Define trend analytics.**
Trend analytics involves analyzing patterns over time to predict future movements and behaviors in data.

Q2

1. **State types of Machine Learning. Explain any one in detail.**
The three types of machine learning are supervised learning, unsupervised learning, and reinforcement learning.

    - **Supervised Learning:** In this approach, the model is trained using labeled data, meaning the input comes with the correct output. The algorithm learns to map inputs to desired outputs, making predictions on unseen data. Examples include spam detection and medical diagnosis prediction.

2. **How Receiver Operating Characteristic (ROC) curve is created?**
The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels. It helps evaluate a classifier's performance by showing the trade-off between sensitivity and specificity. The Area Under the Curve (AUC) quantifies how well the model distinguishes between positive and negative classes.

3. **What is association rule? Give one example.**
Association rule mining identifies relationships between items in large datasets. It is commonly used in market basket analysis to find items frequently bought together.
Example: If many customers who buy bread also buy butter, the rule {bread} → {butter} indicates an association between these items.

4. **What is Influence Maximization?**
Influence Maximization is the process of identifying key individuals in a social network who can maximize the spread of information or influence. It is widely used in viral marketing, political campaigns, and social awareness programs to optimize outreach strategies.

5. **Explain Knowledge Discovery in Database (KDD) process.**
The KDD process involves extracting useful knowledge from large datasets. It consists of several steps: data selection, preprocessing (cleaning and transformation), data mining (pattern discovery), and interpretation/evaluation. This process is used in business intelligence, fraud detection, and healthcare analytics.

6. **Explain the concept of underfitting & overfitting.**
Underfitting occurs when a model is too simple to capture patterns in the data, leading to poor performance on both training and test sets. Overfitting happens when a model learns noise and specific details from the training data, resulting in high accuracy on training data but poor generalization to new data. A balanced model should avoid both issues.

7. **What is linear regression? What type of Machine Learning applications can be solved with linear regression?**
   Linear regression is a supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables using a linear equation. It is used in applications like predicting house prices, sales forecasting, and analyzing the impact of marketing campaigns.

8. **What is Social Media Analytics?**
   Social Media Analytics involves collecting, analyzing, and interpreting data from social media platforms to gain insights into user behavior, trends, and public opinion. It is used for brand monitoring, sentiment analysis, and understanding customer preferences.

9. **What are the advantages of FP-growth Algorithm?**
   The FP-growth algorithm is an efficient method for frequent itemset mining. It eliminates the need for candidate generation, making it faster than the Apriori algorithm. It also compresses the dataset using a compact structure called the FP-tree, reducing memory usage and improving processing speed.

10. **What are dependent & independent variables?**
    In a mathematical model, an independent variable is the input or predictor that influences the outcome, while a dependent variable is the outcome that depends on the independent variable. For example, in predicting house prices, features like square footage and location are independent variables, while the house price is the dependent variable.

11. **What is confusion matrix?**
    A confusion matrix is a table used to evaluate the performance of a classification model. It includes True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), helping to calculate metrics like precision, recall, and accuracy.

12. **Define support and confidence in association rule mining.**

- **Support** measures how frequently an itemset appears in a dataset. Formula:
  Support = (Transactions containing both X and Y) / (Total transactions)
- **Confidence** measures the likelihood that item Y is purchased given that item X is purchased. Formula:
  Confidence = (Transactions containing both X and Y) / (Transactions containing X)

13. **Explain any two Machine Learning (ML) Applications.**

- **Fraud Detection:** ML models analyze transaction patterns to detect anomalies and fraudulent activities in banking and finance.
- **Recommendation Systems:** Platforms like Netflix and Amazon use ML to recommend movies and products based on user preferences and browsing history.

14. **Write a short note on stop words.**
    Stop words are common words such as "is," "the," "and," and "in" that are often removed from text during preprocessing in Natural Language Processing (NLP). These words do not carry significant meaning and are ignored to improve efficiency in text analysis and search engines.

15. **Define Supervised Learning and Unsupervised Learning.**

- **Supervised Learning:** The model is trained using labeled data, where inputs are paired with corresponding outputs. It is used in classification and regression problems.
- **Unsupervised Learning:** The model finds patterns and structures in unlabeled data without predefined outputs. It is used in clustering and anomaly detection.

Q3,4

1. What is prediction? Explain any one regression model in detail.

Prediction refers to the process of using historical data and statistical techniques to forecast future outcomes. It is widely used in data science, finance, healthcare, and various other domains.

Linear Regression Model:
Linear regression is a supervised learning algorithm that models the relationship between a dependent variable (target) and one or more independent variables (predictors) using a linear equation. The equation is given as:
$Y=mX+CY = mX + C$
where **Y** is the dependent variable, **X** is the independent variable, **m** is the slope, and **C** is the intercept.

Example: In real estate, linear regression can predict house prices based on factors like area size, number of rooms, and location. The model finds the best-fit line that minimizes the error between actual and predicted values.

2. Differentiate between Stemming and Lemmatization.

Feature	Stemming	Lemmatization
Definition	Reduces a word to its base/root form by removing suffixes.	Converts a word to its dictionary form (lemma) by considering meaning.
Accuracy	Less accurate as it uses simple rules.	More accurate as it considers word context.
Example	"Running" → "Run", "Happily" → "Happi"	"Running" → "Run", "Better" → "Good"
Approach	Uses heuristic rules.	Uses linguistic dictionaries.
Speed	Faster since it applies basic rules.	Slower as it involves complex processing.

3. Describe types of Data Analytics.

Data analytics is categorized into four types:

- **Descriptive Analytics:** Summarizes past data to understand trends and patterns. Example: Sales reports showing monthly revenue.
- **Diagnostic Analytics:** Identifies reasons behind past outcomes. Example: Analyzing why customer churn increased.
- **Predictive Analytics:** Uses statistical models and machine learning to forecast future events. Example: Predicting stock prices.
- **Prescriptive Analytics:** Suggests optimal actions based on predictive insights. Example: Recommending personalized product offers.

4. Which are the challenges in social media analytics?

Social media analytics faces several challenges:

- **Data Volume & Velocity:** Enormous data generated every second requires efficient storage and processing.
- **Data Quality:** Noise, fake profiles, and bots make it difficult to extract reliable insights.
- **Sentiment Analysis Complexity:** Sarcasm, slang, and multiple languages make it hard to analyze emotions.
- **Privacy Concerns:** Analyzing user data raises ethical and legal issues.
- **Real-time Analysis:** Detecting trends instantly is challenging due to continuous content updates.

5. Explain Reinforcement Learning.

Reinforcement learning (RL) is a type of machine learning where an agent learns by interacting with an environment and receiving rewards or penalties. It follows a trial-and-error approach to maximize cumulative rewards.

Key Components:

- **Agent:** The entity making decisions.
- **Environment:** The system in which the agent operates.
- **Actions:** Choices made by the agent.
- **Rewards:** Feedback received for actions taken.

Example: In gaming AI, RL is used to train bots to play chess by learning optimal moves through self-play and rewards.

6. What are frequent itemsets & association rules? Describe with example.

- **Frequent Itemsets:** These are sets of items that appear together frequently in a dataset. They are identified using techniques like the Apriori algorithm.
- **Association Rules:** These rules describe relationships between items in a dataset and are defined using support and confidence metrics.

Example: In a supermarket, if many customers buy bread and butter together, the association rule could be:

- {Bread} → {Butter} (if you buy bread, you are likely to buy butter).
  This helps businesses with product recommendations and store arrangements.

7. What is Bag of Words & POS tagging in NLP?

- **Bag of Words (BoW):** A text representation model that converts documents into word frequency vectors without considering grammar or word order. Example: "I love data science" and "Science loves data" will have the same representation.
- **POS Tagging:** Part-of-Speech (POS) tagging assigns grammatical categories to words in a sentence, such as nouns, verbs, and adjectives. Example: "The cat is sleeping" → "The (Determiner), cat (Noun), is (Verb), sleeping (Verb)".

8. What is Logistic Regression? Explain it with example.

Logistic regression is a classification algorithm used to predict categorical outcomes (e.g., yes/no, true/false). It estimates probabilities using the logistic (sigmoid) function:
P(Y=1)=11+e−(b0+b1X)P(Y=1) = \frac{1}{1 + e^{-(b0 + b1X)}}

Example: In medical diagnosis, logistic regression can predict whether a patient has diabetes (1) or not (0) based on factors like age, BMI, and glucose levels. The output probability determines the classification.

9. Write a short note on community detection.

Community detection is the process of identifying groups of interconnected nodes in a network, such as social media users or biological systems. It helps in understanding relationships and segmenting data.

Example: In social networks, it identifies groups of people with similar interests, such as tech enthusiasts or fitness communities. Algorithms like modularity optimization and hierarchical clustering are used for community detection.

10. Explain Apriori algorithm.

The Apriori algorithm is a frequent pattern mining algorithm used in association rule learning. It identifies frequent itemsets in transactional databases using the **Apriori Property**, which states that a subset of a frequent itemset must also be frequent.

Steps:

1. Identify frequent individual items using a minimum support threshold.
2. Generate candidate itemsets by combining frequent items.
3. Filter out infrequent itemsets and generate association rules.

Example: In market basket analysis, it finds that customers who buy "milk" and "bread" often buy "butter" too.

11. Explain phases in Natural Language Processing (NLP).

- **Lexical Analysis:** Tokenization and breaking text into words.
- **Syntactic Analysis:** Checking grammar and sentence structure.
- **Semantic Analysis:** Understanding meaning and relationships.
- **Pragmatic Analysis:** Interpreting text based on context.
- **Discourse Integration:** Linking sentences to form meaning.

Example: In chatbots, NLP phases help process user queries and generate responses.

12. Explain Exploratory Data Analysis (EDA).

Exploratory Data Analysis (EDA) is the process of analyzing data sets to summarize main characteristics using statistical techniques and visualization.

Steps in EDA:

- **Data Cleaning:** Handling missing values and outliers.
- **Summary Statistics:** Mean, median, variance, and correlation analysis.
- **Data Visualization:** Using histograms, scatter plots, and box plots to identify patterns.

Example: In sales data, EDA helps understand trends and customer behavior before applying predictive models.

13. Explain the life cycle of Social Media Analytics.

- **Data Collection:** Extracting data from platforms like Twitter, Facebook, or Instagram.
- **Data Processing:** Cleaning and structuring the data for analysis.
- **Analysis & Insights:** Using statistical techniques to extract patterns and trends.
- **Visualization & Reporting:** Creating dashboards and charts for decision-making.
- **Decision Making:** Applying insights for marketing, brand monitoring, or customer engagement.

Example: Businesses analyze customer feedback on social media to improve products and services.

Q5

1. Write a short note on Text Analytics.

Text Analytics is the process of converting unstructured textual data into meaningful insights using machine learning, natural language processing (NLP), and statistical techniques. It involves tasks like sentiment analysis, topic modeling, entity recognition, and keyword extraction.

Applications:

- Sentiment analysis in social media to determine customer opinions.
- Email filtering to classify spam and important messages.
- Healthcare industry for analyzing patient records and medical reports.

- Legal document analysis for extracting relevant case details.

Example: A company analyzing customer feedback reviews to understand market trends.

2. Define the terms:

- **Confusion Matrix:** A table used to evaluate the performance of a classification model. It consists of four elements—True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

- **Accuracy:** Measures how often the model correctly predicts outcomes. It is calculated as: $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$

- **Precision:** Measures how many of the predicted positive cases are actually positive. It is calculated as: $\text{Precision} = \frac{TP}{TP + FP}$

3. What is Machine Learning? Explain its types.

Machine Learning (ML) is a subset of artificial intelligence (AI) that enables computers to learn patterns from data and make predictions or decisions without being explicitly programmed.

Types of Machine Learning:

Type	Description	Example
Supervised Learning	Uses labeled data where the algorithm learns from input-output pairs.	Spam email classification
Unsupervised Learning	Works with unlabeled data to find patterns and relationships.	Customer segmentation
Reinforcement Learning	Uses reward-based learning where an agent interacts with an environment to maximize rewards.	Self-driving cars

4. Write a short note on Support Vector Machine (SVM).

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It works by finding an optimal hyperplane that separates data points into different categories with the maximum margin.

Key Features:

- Effective in high-dimensional spaces.
- Works well for binary and multi-class classification problems.
- Uses kernel tricks (linear, polynomial, RBF) to handle non-linear data.

Example: SVM can be used for handwriting recognition, where it classifies letters based on pixel features.

5. Explain the life cycle of Data Analytics.

The data analytics life cycle consists of six key stages:

- **Data Collection:** Gathering structured and unstructured data from various sources like databases, IoT devices, and social media.
- **Data Cleaning & Preprocessing:** Handling missing values, outliers, and noise to ensure data quality.
- **Exploratory Data Analysis (EDA):** Using visualization techniques and statistical methods to understand data patterns.
- **Model Building & Analysis:** Applying machine learning or statistical models to extract insights.
- **Interpretation & Visualization:** Presenting insights using dashboards, graphs, and reports.
- **Decision Making & Action:** Applying the results for business improvements, predictions, or automation.

Example: In e-commerce, the data analytics lifecycle is used to predict customer buying behavior and optimize marketing strategies.