Total No. of Questions : 8]

SEAT No. :

PB4430

[Total No. of Pages : 3

[6262]-43

T.E. (Computer Engineering)

DATA SCIENCE AND BIG DATA ANALYTICS

(2019 Pattern) (Semester- II) (310251)

Time : 2½ Hours] *[Max. Marks : 70*

Instructions to the candidates:

1) *Answer Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.*
2) *Neat diagrams must be drawn wherever necessary.*
3) *Figures to the right side indicate full marks.*
4) *Assume suitable data if necessary.*
5) *Use of Scientific Calculator is permitted.*

Q1) a) What is the data Preparation phase in Data Analytics Lifecycle. What is the Analytics Sandbox and ETLT process in this phase? **[8]**

b) List out different stakeholders of an analytics project. What they usually expect at the conclusion (key outputs) of a project? **[8]**

OR

Q2) a) List out the activities to be carried out in model planning and model building phase. What are different tools used for these phases? **[8]**

b) What is linear regression, and what are its primary objectives? What is the difference between simple linear regression and multiple linear regression? How do you evaluate the performance of linear regression? **[8]**

Q3) a) What is logistic regression, and how does it differ from linear regression? What is the sigmoid function, and what role does it play in logistic regression? **[9]**

b) Suppose you are given a dataset containing information about whether emails are spam or not spam, along with two features: the presence of the word "offer" (1 for present, 0 for absent) and the presence of the word "free" (1 for present, 0 for absent). You are tasked with classifying a new email with the following feature values: "offer"=1 and "free"=1. **[9]**

P.T.O.

Given the training dataset:

Email	Offer	Free	Spam
1	1	0	No
2	0	1	Yes
3	1	1	Yes
4	0	1	No
5	1	1	Yes

Calculate the probability that the new email is spam using Naive Bayes.

OR

Q4) a) How does the Apriori algorithm discover frequent itemsets in a dataset? What is the role of support and confidence in the context of association rule mining using the Apriori algorithm? **[9]**

b) Explain the process of building a decision tree? What are the criteria used for splitting nodes in a decision tree? **[9]**

Q5) a) Suppose you have the following dataset containing the coordinates of points in a 2-dimensional space: **[9]**

Point	X Coordinate	Y Coordinate
A	2	3
B	4	7
C	3	5
D	6	9
E	8	6
F	7	8

Perform K-means clustering on this dataset with K = 2. Assume the initial centroids to be (2,3) and (8,6). Compute the new centroids after each iteration until convergence, and assign points to their nearest centroids.

b) How do you handle noise and irrelevant information in text data during preprocessing? Explain the terms bag of words and TF IDF in text analytics. **[9]**

OR

Q6) a) Explain how hierarchical clustering can be used for visualizing hierarchical relationships in data with suitable example? What are some real-world applications of hierarchical clustering? **[9]**

b) What is the holdout method, and how does it work? Explain the difference between training set, validation set, and test set in the holdout method. **[9]**

Q7) a) What is a histogram? How is it used to visualize the distribution of data? How is it different from a density plot? **[9]**

b) What is the Hadoop ecosystem, and what are its primary components? What is MapReduce, and how does it fit into the Hadoop ecosystem? **[9]**

OR

Q8) a) What is a box plot? Explain the different components of a box plot? How do you interpret the median, quartiles, and whiskers in a box plot? What does the interquartile range (IQR) represent in a box plot? **[9]**

b) Explain the role of Apache Pig in data processing workflows on Hadoop? What is Apache Spark, and how does it complement Hadoop for big data processing? **[9]**

Q1) a) What is the data Preparation phase in Data Analytics Lifecycle. What is the Analytics Sandbox and ETLT process in this phase?

Ans:

Data Preparation Phase: -

- This stage involves collecting, processing and cleaning data. Here the focus shifts from business requirements to data requirements. In this early phase, data is collected but not analysed.
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often.

a) Preparing the analytic sandbox:

- Create the analytic sandbox. It also called a workspace. It allows the team to explore data without interfering with live production data.
- Sandbox collects all kinds of data. The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics.

b) Performing ETLT (Extract, Transform, Load, Transform):

- The team needs to execute Extract, Load and Transform (ELT) to get data into the sandbox.
- Extract, Transform, Load (ETL): It transforms the data based on a set of business rules before loading it into the sandbox.
- Extract, Load, Transform (ELT): It loads the data into the sandbox and then transforms it based on a set of business rules.
- Extract, Transform, Load, Transform (ETLT): It's the combination of ETL and ELT and has two transformation levels.

c) Learning about the data:

a. Data is captured through three main ways:
   i) Data acquisition: Obtaining existing data from outside sources.
   ii) Data entry: Creating new data values from data inputted within the organization.
   iii) iii. Signal reception: Capturing data created by devices.

d) Data Conditioning:

- Data conditioning includes cleaning data, normalizing datasets and performing transformations. It is often viewed as a preprocessing step prior to data analysis; it might be performed by data owner, IT department, DBA, etc.
- Best to have data scientists involved and data science teams prefer more data than too little.

e) Common tools for data preparation:

- Hadoop can perform parallel ingest and analysis.
- Alpine miner provides a graphical user interface for creating analytic workflows.
- Open Refine is a free, open-source tool for working with messy data.
- Similar to Open Refine, data wrangler is an interactive tool for data cleansing and transformation.

Q1) b) List out different stakeholders of an analytics project. What they usually expect at the conclusion (key outputs) of a project?

Ans:

Stakeholders in an Analytics Project:

- Analytics projects involve a variety of stakeholders, each playing a specific role and having unique expectations.
- Understanding the diverse expectations of stakeholders ensures that analytics projects deliver valuable, actionable, and usable outcomes. Each stakeholder measures success differently, so aligning goals early and reviewing key outputs regularly is critical.

1. Business Executives / Decision Makers:
   - Role: Provide business goals, budget, and strategic direction.
   - Expectation: Actionable insights, ROI, performance reports.

2. Project Managers:
   - Role: Oversee project timeline, resources, and team coordination.
   - Expectation: On-time delivery, milestone tracking, risk mitigation.

3. Data Scientists / Analysts:
   - Role: Perform data analysis, build models, and interpret data.
   - Expectation: Clean and well-prepared data, model performance metrics.

4. IT/Data Engineers:
   - Role: Handle data collection, storage, processing, and infrastructure.
   - Expectation: Efficient data pipelines, system scalability, data security.

5. Customers (Internal or External):
   - Role: Use the end product or benefit from insights.
   - Expectation: Improved experience, personalized services, better decision support.

6. Domain Experts:
   - Role: Provide domain-specific knowledge and validation.
   - Expectation: Accurate and relevant results, aligned with business context.

Key Outputs Expected at the Conclusion of a Project:

1. Business Insights:
   - Clear findings that support business decisions.
   - Example: Customer churn reasons, market trends.

2. Predictive/Prescriptive Models:
   - Working ML models for forecasting or recommendations.
   - Example: Sales prediction, fraud detection model.

3. Visual Dashboards and Reports:
   - User-friendly summaries with KPIs, charts, and graphs.

4. Technical Documentation:
   - Data sources, model assumptions, algorithm explanations.

5. Deployment Plan:
   - Strategy for integrating the solution into production systems.

6. Return on Investment (ROI) Analysis:
   - Evaluation of the project's impact on business performance.

Q2) a) List out the activities to be carried out in model planning and model building phase. What are different tools used for these phases?

Ans:

1. Model Planning Phase:

- This phase involves selecting the appropriate analytical techniques and designing a plan for how the data will be used to build the model.

Key Activities:

1. Identify modeling techniques:
    - Choose suitable methods (e.g., regression, classification, clustering).
2. Data exploration and visualization:
    - Understand data patterns, relationships, and outliers.
3. Feature selection and engineering:
    - Identify relevant variables and create new features.
4. Define data partitioning strategy:
    - Split data into training, testing, and validation sets.
5. Develop a modeling strategy:
    - Decide model evaluation metrics (accuracy, RMSE, etc.)

Tools Used:

- R and Python (for statistical planning)
- RapidMiner (visual workflow design)
- KNIME
- SAS Enterprise Miner
- IBM SPSS Modeler

Model Building Phase:

This phase focuses on the actual creation of the analytical models using selected algorithms and techniques.

Key Activities:

1. Data preprocessing:
    - Handle missing values, normalization, and encoding.
2. Model training:
    - Apply algorithms on the training dataset.
3. Model testing and evaluation:
    - Use testing data to evaluate performance.
4. Hyperparameter tuning:
    - Optimize model settings for best results.
5. Model validation:
    - Ensure the model performs well on unseen data.

Tools Used:

- Python (Scikit-learn, TensorFlow, Keras)
- R (Caret, RandomForest, xgboost)
- Jupyter Notebooks
- Google Colab
- Weka
- H2O.ai

Q2) b) What is linear regression, and what are its primary objectives? What is the difference between simple linear regression and multiple linear regression? How do you evaluate the performance of linear regression?

Ans:

Linear Regression: -

- Linear Regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a straight line. It is one of the simplest forms of regression used in predictive modeling.
- The simplest form of regression to visualize is linear regression with a single predictor. A linear regression technique can be used if the relationship between X and Y can be approximated with a straight line.
- Linear regression with a single predictor can be expressed with the equation:
  $y = \theta_2 x + \theta_1 + e$
- The regression parameters in simple linear regression are the slope of the line (θ_2), the angle between a data point and the regression line and the y intercept (θ_1) the point where x crosses the y axis (X = 0).
- Model 'Y', is a linear function of 'X'. The value of 'Y' increases or decreases in linear manner according to which the value of 'X' also changes.

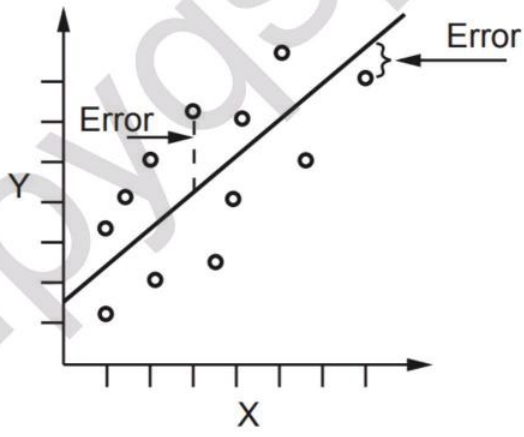

Fig. Linear Regression

Difference Between Simple and Multiple Linear Regression: -

Feature	Simple Linear Regression	Multiple Linear Regression
No. of Independent Variables	Only one	Two or more
Equation	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$
Visualization	Can be plotted as a straight line	Harder to visualize (multi-dimensional)
Use Case Example	Predicting height from age	Predicting house price from size, location, etc.

Performance Evaluation of Linear Regression:

1. R-squared (R^2):
    - Measures proportion of variance in the dependent variable explained by the model.
    - Ranges from 0 to 1 (higher is better).
2. Mean Squared Error (MSE):
    - Average of the squared differences between actual and predicted values.
    - Lower MSE indicates better fit.
3. Root Mean Squared Error (RMSE):
    - Square root of MSE; easier to interpret as it's in the same unit as output.
4. Mean Absolute Error (MAE):
    - Average of absolute differences between actual and predicted values.
5. Residual Analysis:
    - Checking residual plots for randomness to validate model assumptions.

Q3) a) What is logistic regression, and how does it differ from linear regression? What is the sigmoid function, and what role does it play in logistic regression?

Ans:

Logistic Regression: -

- Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous. A statistical method used to model dichotomous or binary outcomes using predictor variables.
- Logistic component: Instead of modeling the outcome, Y, directly, the method models the log odds (Y) using the logistic function.
- Regression component: Methods used to quantify association between an outcome and predictor variables. It could be used to build predictive models as a function of predictors.
- In simple logistic regression, logistic regression with 1 predictor variable.

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Sr.no	Linear regression	Logistic Regression
1.	Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
2.	Target is an interval variable.	Target is discrete variable.
3.	Solve regression problem	Solve classification problem
4.	Example: Relationship between number of hours worked with your salary	Example: whether they are male or female
5.	Example: What is temperature	Example: Will it rain or not?

Sigmoid Function

- The sigmoid function is a mathematical function that maps any real-valued number into the range (0, 1).

  Formula –

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

  Where $z = w^T x + b$

Role of Sigmoid in Logistic Regression: -

- Converts the linear output z into a probability.
- Helps in classification:
  If $\sigma(z) \geq 0.5$, classify as 1
  If $\sigma(z) < 0.5$, classify as 0
- Ensures the output is suitable for binary classification.

Q3) b) Suppose you are given a dataset containing information about whether emails are spam or not spam, along with two features: the presence of the word "offer" (1 for present, 0 for absent) and the presence of the word "free" (1 for present, 0 for absent). You are tasked with classifying a new email with the following feature values: "offer"=1 and "free"=1.

Given the training dataset:

Email	Offer	Free	Spam
1	1	0	No
2	0	1	Yes
3	1	1	Yes
4	0	1	No
5	1	1	Yes

Calculate the probability that the new email is spam using Naive Bayes.

Ans:

To calculate the probability that a new email is spam using the Naive Bayes classifier, given the features:

- Offer = 1
- Free = 1

Email	Offer	Free	Spam
1	1	0	No
2	0	1	Yes
3	1	1	Yes
4	0	1	No
5	1	1	Yes

Calculate Prior Probabilities

$$P(\text{Spam=Yes}) = \frac{3}{5}, \quad P(\text{Spam=No}) = \frac{2}{5}$$

Calculate Likelihoods (Using Naive Bayes Assumption): -

For Spam = Yes (3 examples):

- $P(\text{Offer}=1 \mid \text{Spam}=\text{Yes}) = \frac{2}{3}$
- $P(\text{Free}=1 \mid \text{Spam}=\text{Yes}) = \frac{3}{3} = 1$

For Spam = No (2 examples):

- $P(\text{Offer}=1 \mid \text{Spam}=\text{No}) = \frac{1}{2}$
- $P(\text{Free}=1 \mid \text{Spam}=\text{No}) = \frac{1}{2}$

Naive Bayes Formula: -

For Spam = Yes:

$$P(\text{Yes}) \cdot P(\text{Offer}=1 \mid \text{Yes}) \cdot P(\text{Free}=1 \mid \text{Yes}) = \frac{3}{5} \cdot \frac{2}{3} \cdot 1 = \frac{2}{5}$$

For Spam = No:

$$P(\text{No}) \cdot P(\text{Offer}=1 \mid \text{No}) \cdot P(\text{Free}=1 \mid \text{No}) = \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{10}$$

Normalize to Get Final Probabilities

- $P_1 = \frac{2}{5}$ (Spam = Yes)
- $P_2 = \frac{1}{10}$ (Spam = No)

$$P(\text{Spam}=\text{Yes} \mid \text{Offer}=1, \text{Free}=1) = \frac{P_1}{P_1 + P_2} = \frac{\frac{2}{5}}{\frac{2}{5} + \frac{1}{10}} = \frac{0.4}{0.4 + 0.1} = \frac{0.4}{0.5} = 0.8$$

Final Answer:

$$\boxed{P(\text{Spam} \mid \text{Offer} = 1, \text{Free} = 1) = 0.8}$$

Q4) a) How does the Apriori algorithm discover frequent itemsets in a dataset? What is the role of support and confidence in the context of association rule mining using the Apriori algorithm?

Ans:

Apriori Algorithm:

- The Apriori algorithm is used in association rule mining to discover frequent itemsets in a transactional database and to derive association rules.

Working Steps:

1. Generate Candidate Itemsets:
    - Start with individual items (1-itemsets).
    - Generate larger itemsets (k-itemsets) from (k–1)-itemsets using the Apriori property.
2. Prune Infrequent Itemsets:
    - Use the Apriori property: *If an itemset is frequent, then all its subsets must also be frequent.*
    - Eliminate itemsets that contain infrequent subsets.
3. Count Support:
    - Scan the dataset to count the support (frequency) of each candidate itemset.
4. Repeat:
    - Continue generating and pruning until no more frequent itemsets can be found.

1. Support:

- **Definition:** Proportion of transactions in which an itemset appears.
- **Formula:**

$$\text{Support}(A) = \frac{\text{Number of transactions containing } A}{\text{Total number of transactions}}$$

- **Use:** Filters out itemsets that occur infrequently.

2. Confidence:

- **Definition:** Measure of how often items in **B** appear in transactions that contain **A**.
- **Formula:**

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

- **Use:** Evaluates the strength of an association rule.

Q4) b) Explain the process of building a decision tree? What are the criteria used for splitting nodes in a decision tree?

Ans:

Decision Tree: -

- A Decision Tree is a supervised learning algorithm used for classification and regression tasks.
- It splits the dataset into smaller subsets while forming a tree structure where each node represents a decision based on a feature.

Process of Building a Decision Tree:

1. **Start at the Root Node:**
    - Begin with the entire dataset.
2. **Choose the Best Feature to Split:**
    - Use a **splitting criterion** (e.g., Information Gain, Gini Index) to select the best attribute.
3. **Split the Dataset:**
    - Divide the dataset into subsets based on the selected feature.
4. **Repeat Recursively:**
    - For each subset (child node), repeat the process using only the data in that subset.
5. **Stopping Conditions:**
    - All data in a node belongs to the same class.
    - No remaining features to split.
    - Tree reaches a specified maximum depth or minimum samples per node.
6. **Assign Class Labels:**
    - Leaf nodes are labelled with the most common class in that subset.

Criteria for Splitting Nodes in a Decision Tree:

The goal is to maximize purity (i.e., make child nodes as homogeneous as possible). Common criteria include:

a) Information Gain (used in ID3):

- Measures the reduction in entropy after the split.
- **Formula:**

$$\text{Information Gain} = \text{Entropy(parent)} - \sum \frac{n_i}{n} \cdot \text{Entropy(child}_i\text{)}$$

b) Gini Index (used in CART):

- Measures impurity of a node.
- Lower Gini = more pure.
- **Formula:**

$$\text{Gini}(D) = 1 - \sum_{i=1}^{C} p_i^2$$

c) Gain Ratio (used in C4.5):

- Adjusts information gain by considering the intrinsic information of a split.
- Used to penalize splits with many branches.

Q5) a) Suppose you have the following dataset containing the coordinates of points in a 2-dimensional space:

Point	X Coordinate	Y Coordinate
A	2	3
B	4	7
C	3	5
D	6	9
E	8	6
F	7	8

Perform K-means clustering on this dataset with K = 2. Assume the initial centroids to be (2,3) and (8,6). Compute the new centroids after each iteration until convergence, and assign points to their nearest centroids.

Ans:

Initial Centroids:

- Centroid 1 (C1): (2, 3)
- Centroid 2 (C2): (8, 6)

Iteration 1: Assign points to nearest centroid

Use Euclidean distance formula:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Point	Distance to C1 (2,3)	Distance to C2 (8,6)	Assigned Cluster
A (2,3)	0.00	6.71	C1
B (4,7)	4.47	4.12	C2
C (3,5)	2.24	5.10	C1
D (6,9)	7.21	3.16	C2
E (8,6)	6.71	0.00	C2
F (7,8)	6.40	2.24	C2

Clusters after Iteration 1:

- **C1:** A, C
- **C2:** B, D, E, F

New Centroids (Iteration 2):

Cluster C1: Points A (2,3), C (3,5)

$$\text{New C1} = \left(\frac{2+3}{2}, \frac{3+5}{2}\right) = (2.5, 4)$$

Cluster C2: Points B (4,7), D (6,9), E (8,6), F (7,8)

$$\text{New C2} = \left(\frac{4+6+8+7}{4}, \frac{7+9+6+8}{4}\right) = (6.25, 7.5)$$

Point	Distance to New C1 (2.5, 4)	Distance to New C2 (6.25, 7.5)	Assigned Cluster
A	1.12	5.80	C1
B	3.20	2.29	C2
C	1.12	4.30	C1
D	6.10	2.06	C2
E	5.70	2.06	C2
F	5.15	1.25	C2

Clusters remain the same:

- **C1:** A, C
- **C2:** B, D, E, F

Since the clusters did not change, the algorithm has converged.

Final Clusters:

- **Cluster 1:** A (2,3), C (3,5) — **Centroid = (2.5, 4)**
- **Cluster 2:** B (4,7), D (6,9), E (8,6), F (7,8) — **Centroid = (6.25, 7.5)**

Q5) b) How do you handle noise and irrelevant information in text data during preprocessing? Explain the terms bag of words and TF IDF in text analytics.

Ans:

Handling Noise and Irrelevant Information in Text Data:

- Preprocessing is a critical step in Natural Language Processing (NLP) to clean and prepare text data. Common techniques include:

a) Lowercasing:
Converts all text to lowercase to maintain uniformity.
Example: "The", "the", and "THE" become "the".

b) Removing Punctuation and Special Characters:
Eliminates symbols like!, ?, @, etc., which usually carry no meaning in many applications.

c) Removing Stop Words:
Stop words (e.g., "and", "the", "is") are common words that do not add much meaning and can be removed.

d) Tokenization:
Splits text into individual words or tokens.
Example: "I love NLP" → ["I", "love", "NLP"]

e) Stemming/Lemmatization:
Reduces words to their base or root form.
Example: "running" → "run" (stemming)
"better" → "good" (lemmatization)

f) Removing Numbers and Rare Words:
Numbers and infrequent words are often removed to reduce dimensionality and noise**.**

Bag of Words (BoW):

Definition:

- BoW is a simple representation where each document is converted into a vector based on word frequency, ignoring grammar and word order.

Example:

Document	Text
D1	"I love NLP"
D2	"I love AI"

Vocabulary: ["I", "love", "NLP", "AI"]

	I	love	NLP	AI
D1	1	1	1	0
D2	1	1	0	1

TF-IDF (Term Frequency – Inverse Document Frequency):

Definition:

- TF-IDF assigns a weight to each word based on how frequently it appears in a document (TF) and how rare it is across all documents (IDF).

- TF (Term Frequency):

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in } d}{\text{Total terms in } d}$$

- IDF (Inverse Document Frequency):

$$IDF(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$

- TF-IDF Score:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

Q6) a) Explain how hierarchical clustering can be used for visualizing hierarchical relationships in data with suitable example? What are some real-world applications of hierarchical clustering?

Ans:

Hierarchical Clustering:

- Hierarchical clustering is an unsupervised learning technique used to build a hierarchy of clusters. It does not require the number of clusters to be specified in advance.

There are two types:

- Agglomerative (bottom-up): Each data point starts as its own cluster and merges with others.
- Divisive (top-down): All points start in one cluster and are split recursively.

Visualization Using Dendrogram:

- A **dendrogram** is a tree-like diagram that visualizes the hierarchical relationship between clusters.

Example:

Suppose we have 4 animals with features (weight, height):

- Cat: (3 kg, 25 cm)
- Dog: (10 kg, 50 cm)
- Elephant: (5000 kg, 300 cm)
- Tiger: (220 kg, 90 cm)

The dendrogram will first group similar animals (e.g., cat and dog), then merge those groups with others based on similarity.

The vertical axis shows the **distance (dissimilarity)** at which merges happen.

Advantages:

- Easy to visualize using dendrograms.
- No need to pre-specify number of clusters.
- Works well for **hierarchical or nested structures** in data.

Real-World Applications of Hierarchical Clustering:

- Document Clustering — Grouping news articles or research papers by topic.
- Retail/Marketing — Customer segmentation based on purchase behaviour.
- Genomics/Biology — Creating phylogenetic trees to show species evolution.
- Business Intelligence — Organizing product categories based on similarity.
- Web Usage Mining — Grouping users based on browsing patterns.

Q6) b) What is the holdout method, and how does it work? Explain the difference between training set, validation set, and test set in the holdout method.

Ans:

Holdout Method: -

- The holdout method is a simple technique used to evaluate the performance of machine learning models by splitting the dataset into separate parts:
  - Training Set
  - Validation Set (optional)
  - Test Set
- This method helps ensure that the model generalizes well to unseen data.

Working of Holdout Method: -

- Split the dataset (usually 60–80% for training, 20–40% for testing/validation).
- Train the model on the training set.
- Tune hyperparameters (if needed) using the validation set.
- Evaluate final performance on the test set.

Differences Between Training, Validation, and Test Sets:

Set Type	Purpose	Used During
Training Set	Used to fit the model (learn patterns).	Model training
Validation Set	Used to tune parameters and avoid overfitting.	Model tuning (optional)
Test Set	Used to evaluate final model performance.	Final evaluation

Advantages of Holdout Method:

- Simple and fast.
- Suitable for large datasets.

Disadvantages:

- Performance may vary depending on the split.
- Not ideal for small datasets (cross-validation is better).

Q7) a) What is a histogram? How is it used to visualize the distribution of data? How is it different from a density plot?

Ans:

Histogram: -

- A histogram is a graphical representation used to show the distribution of a dataset.
- It divides the data into intervals or bins, and displays the frequency (count) of data points in each bin using bars.
- X-axis: Represents the data intervals (bins).
- Y-axis: Represents the number of data points (frequency) in each bin.
- Histograms help identify:
  - Shape of the data (e.g., normal, skewed)
  - Central tendency (where most data lie)
  - Spread or variability
  - Outliers or gaps
- A histogram of exam scores can show whether most students scored between 60–80, and whether the distribution is symmetric or skewed.

Difference Between Histogram and Density Plot:

Feature	Histogram	Density Plot
Type	Bar chart	Smooth curve (line plot)
Data Grouping	Groups data into bins	Estimates distribution using kernel smoothing
Y-Axis	Shows frequency or count	Shows probability density (area under curve = 1)
Appearance	Discrete bars	Continuous, smooth line
Usefulness	Easy to interpret raw counts	Better for comparing multiple distributions

Q7) b) What is the Hadoop ecosystem, and what are its primary components? What is MapReduce, and how does it fit into the Hadoop ecosystem?

Ans:

Hadoop Ecosystem: -

- Hadoop ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems.
- The Hadoop ecosystem refers to the various components of the Apache Hadoop software library, as well as to the accessories and tools provided by the Apache software foundation for these types of software projects and to the ways that they work together.
- Hadoop is a Java-based framework that is extremely popular for handling and analysing large sets of data. The idea of a Hadoop ecosystem involves the use of different parts of the core Hadoop set such as MapReduce, a framework for handling vast amounts of data and the Hadoop Distributed File System (HDFS), a sophisticated file-handling system. There is also YARN, a Hadoop resource manager.
- In addition to these core elements of Hadoop, Apache has also delivered other kinds of accessories or complementary tools for developers.
- Some of the most well-known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.

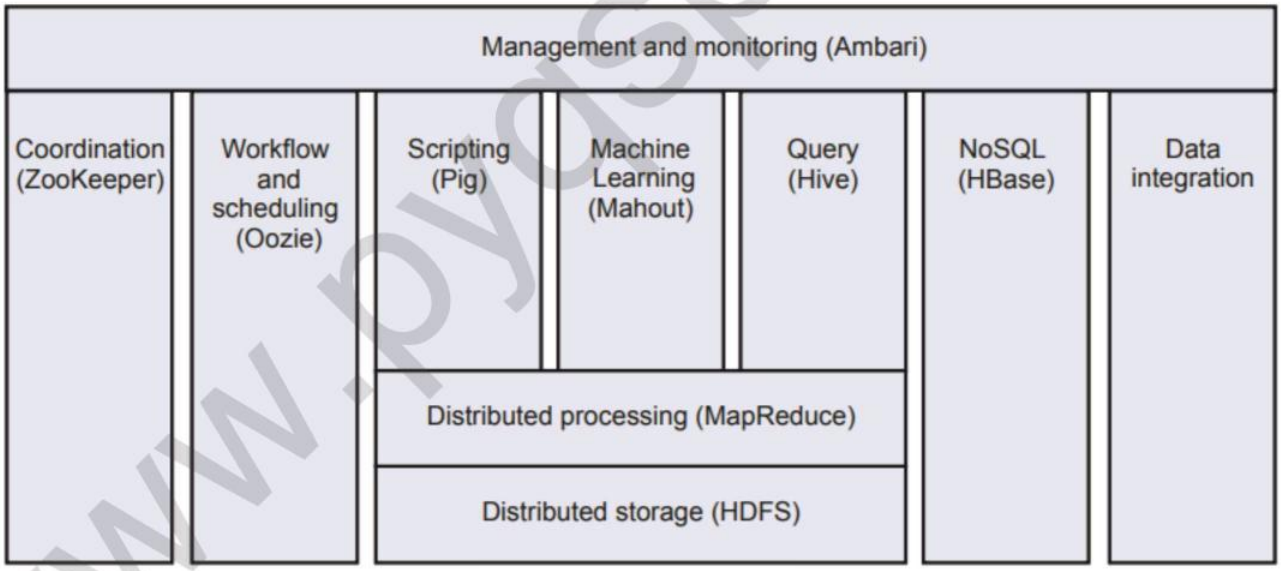

Fig. Hadoop Ecosystem

MapReduce: -

- MapReduce is a programming model used for processing and generating large datasets in a distributed computing environment.

It consists of two main functions:

- **Map Function:**
  Takes input data and converts it into a set of intermediate key-value pairs.
- **Reduce Function:**
  Merges or aggregates intermediate values with the same key to produce the final output.

- **Input Splitting**: Large data is split into chunks.
- **Mapping**: Each chunk is processed in parallel to generate key-value pairs.
- **Shuffling**: Intermediate data is grouped by key and moved to reducers.
- **Reducing**: Aggregated values are processed to produce final results.
- **Output**: Final output is written to the file system (HDFS).

MapReduce in the Hadoop Ecosystem:

MapReduce is one of the core components of the Hadoop ecosystem, along with:

HDFS	Hadoop Distributed File System — stores large volumes of data.
MapReduce	Processes data stored in HDFS using the Map and Reduce paradigm.
YARN	Resource manager that schedules MapReduce jobs on the cluster.

Q8) a) What is a box plot? Explain the different components of a box plot? How do you interpret the median, quartiles, and whiskers in a box plot? What does the interquartile range (IQR) represent in a box plot?

Ans:

Box plot: -

- A Box Plot (also called Box-and-Whisker Plot) is a graphical representation used to display the distribution, spread, and skewness of a numerical dataset.
- It shows the five-number summary of the data:
  Minimum
  First Quartile (Q1)
  Median (Q2)
  Third Quartile (Q3)
  Maximum

Components of a Box Plot:

Component	Description
Box	Represents the interquartile range (Q1 to Q3).
Median Line	A line inside the box indicating the median (Q2).
Whiskers	Lines extending from the box to minimum and maximum (excluding outliers).
Outliers	Data points outside the whiskers, shown as dots or stars.

Interpretation:

Element	Meaning
Median (Q2)	Middle value; divides data into two equal halves.
Q1 (25%)	25% of data falls below this value.
Q3 (75%)	75% of data falls below this value.
Whiskers	Show the spread of the data (excluding outliers).
Outliers	Extreme values that lie beyond 1.5 × IQR from Q1 or Q3.

Uses of Box Plot:

1. Shows central tendency (via median).
2. Reveals spread and variability of data (via IQR).
3. Helps detect outliers.
4. Useful for comparing distributions across different groups.

Interquartile Range (IQR):

- IQR = Q3 − Q1
- Represents the middle 50% of the data.
- A measure of variability: the larger the IQR, the more spread out the central data.

Q8) b) Explain the role of Apache Pig in data processing workflows on Hadoop? What is Apache Spark, and how does it complement Hadoop for big data processing?

Ans:

Apache Pig in Hadoop

- Apache Pig is a high-level platform for processing large datasets in Hadoop.
- It uses a scripting language called Pig Latin.

Role in Data Processing Workflows:

- Simplifies the development of MapReduce programs.
- Translates Pig Latin scripts into MapReduce jobs that run on Hadoop (HDFS).
- Ideal for data transformation, ETL (Extract-Transform-Load) tasks, and analysis.

Features:

- Less complex than Java MapReduce.
- Supports joins, filters, groupings, etc.
- Suitable for semi-structured and structured data.

Example Use Case:

- Parsing web server logs, filtering specific entries, and aggregating results using Pig scripts.

Apache Spark and Its Role in Big Data

- Apache Spark is an open-source, in-memory distributed computing framework.
- Processes large-scale data much faster than traditional MapReduce.

Key Features:

- In-memory computation (faster than disk-based MapReduce).
- Supports multiple languages: Python, Scala, Java.
- Has built-in libraries for:
  - Spark SQL (structured data)
  - MLlib (machine learning)
  - GraphX (graph processing)
  - Spark Streaming (real-time data)

Spark Complements Hadoop:

Aspect	Apache Hadoop (MapReduce)	Apache Spark
Processing	Batch, disk-based	Batch + Real-time, in-memory
Speed	Slower due to disk I/O	Much faster due to memory usage
Ease of Use	More code (Java)	Less code (Python/Scala APIs)
Use With Hadoop	Spark can use HDFS for storage	Spark can run on top of Hadoop ecosystem