USN

Internal Assessment Test 2 – July 2024

Sub:	**Data Science and Visualization (Professional Elective)**			Sub Code:	21CS644	Branch:	CSE
Date:	10.7.2024	90 mins	Max Marks:	50	Sem/Sec:	VI/ A, B, C	OBE

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1	a. Explain the difference between filter methods and wrapper methods for feature selection?	5	CO3	L2

Answer: (5 marks)

In feature selection, **filter methods** and **wrapper methods** are two common techniques used to select the most relevant features in a dataset, but they differ in approach and evaluation criteria.

1. Filter Methods

- **Independent of any machine learning model**: Filter methods evaluate the relevance of features based on their statistical properties and do not involve any specific model.
- **Evaluation based on statistical scores**: Common techniques include correlation coefficients, chi-square tests, mutual information, and ANOVA.
- **Less computationally intensive**: These methods are generally faster and more efficient because they do not involve training a model.
- **Suitable for high-dimensional data**: Often used in scenarios with a large number of features.

Example: Using correlation to remove highly correlated features that might provide redundant information.

2. Wrapper Methods

- **Depend on a specific machine learning model**: Wrapper methods evaluate features based on model performance.
- **Iterative approach**: They involve training the model multiple times with different subsets of features and selecting the subset that yields the best performance.
- **Computationally intensive**: Since they involve multiple rounds of model training, they are more resource-intensive and slower.

- **Potentially more accurate**: By aligning feature selection with the specific model’s performance, they can lead to a feature set that enhances the model’s accuracy.

Example: Recursive Feature Elimination (RFE), which trains the model, eliminates the least important feature, and repeats until the optimal set is reached.

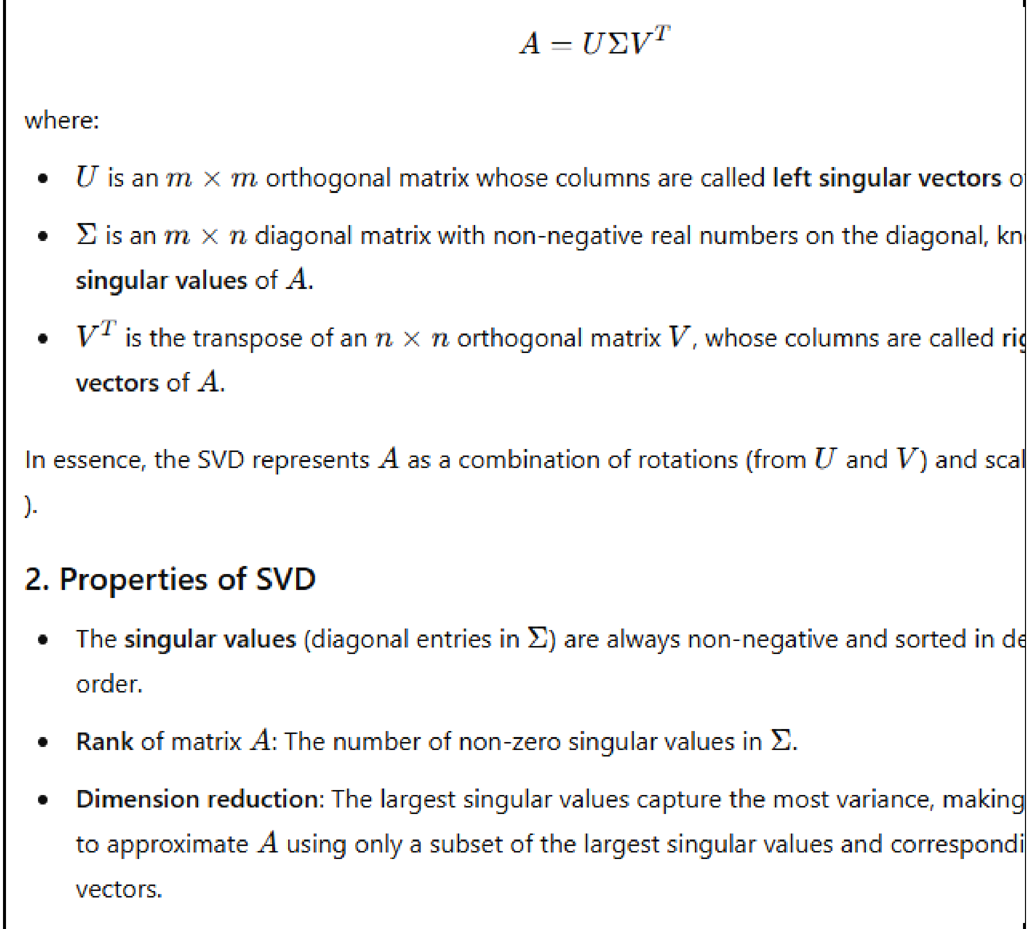
	b. Describe the Singular Value Decomposition (SVD) technique with relevant example and mathematical expressions	5	CO3	L 2

Answer: (5 marks)

Singular Value Decomposition (SVD) is a powerful matrix factorization technique commonly used in linear algebra, data compression, noise reduction, and dimensionality reduction. SVD decomposes a matrix into three simpler matrices, revealing important structure within the data.

1. SVD Definition and Formula

For a given matrix A of size m×n, SVD is defined as:



$$A = U\Sigma V^T$$

where:

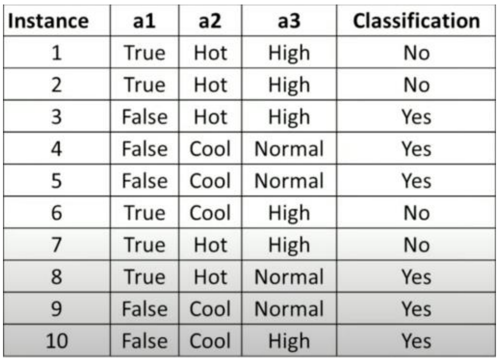
- U is an $m \times m$ orthogonal matrix whose columns are called **left singular vectors** o
- Σ is an $m \times n$ diagonal matrix with non-negative real numbers on the diagonal, kn **singular values** of A.
- V^T is the transpose of an $n \times n$ orthogonal matrix V, whose columns are called **ri vectors** of A.

In essence, the SVD represents A as a combination of rotations (from U and V) and scal).

2. Properties of SVD

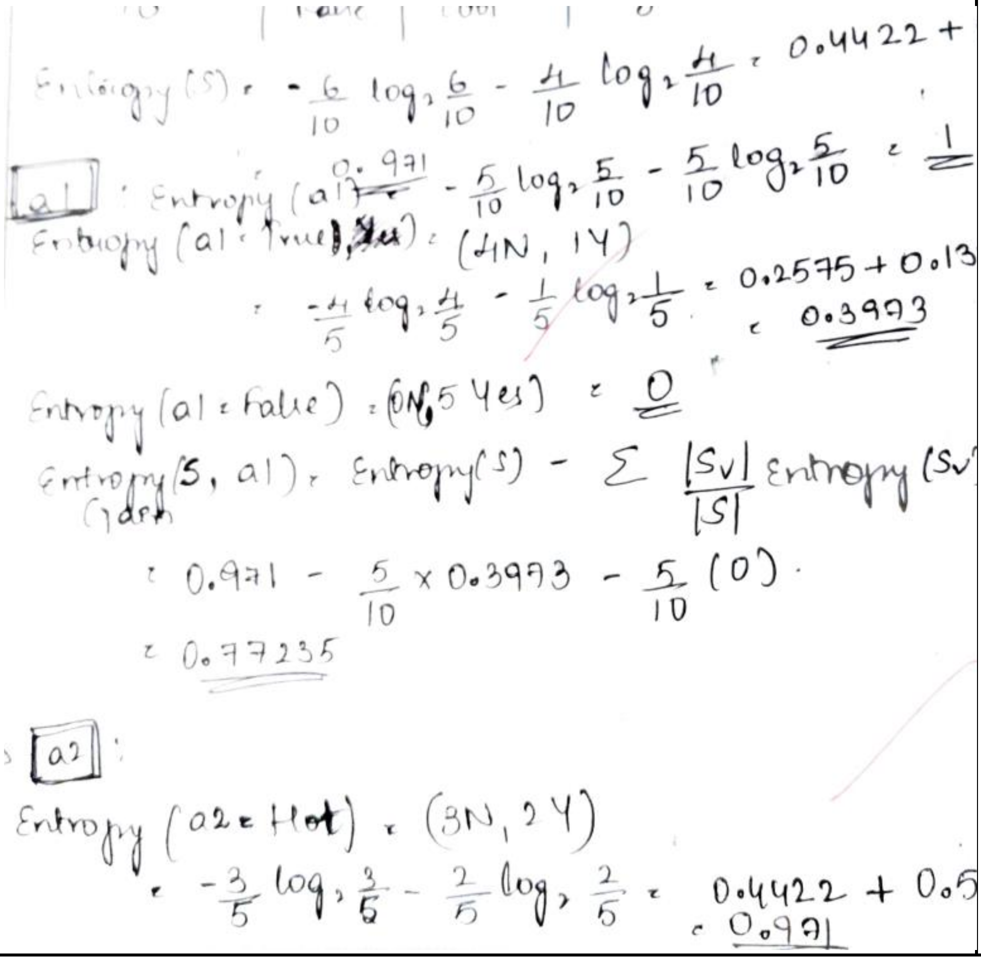
- The **singular values** (diagonal entries in Σ) are always non-negative and sorted in de order.
- **Rank** of matrix A: The number of non-zero singular values in Σ.
- **Dimension reduction**: The largest singular values capture the most variance, making to approximate A using only a subset of the largest singular values and correspondi vectors.

2	Describe the decision tree technique and find the root node of the following below using ID 3.	10	CO3	L 3



Instance	a1	a2	a3	Classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	Yes
4	False	Cool	Normal	Yes
5	False	Cool	Normal	Yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Normal	Yes
9	False	Cool	Normal	Yes
10	False	Cool	High	Yes

Solution :



Entropy(S) = $-\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$ = 0.4422 +

$\boxed{a1}$: Entropy (a1) $\overset{0.971}{=}$ $-\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$ = $\underline{\underline{1}}$

Entropy (a1 = True), ~~Yes~~) = (4N, 1Y)

= $-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$ = 0.2575 + 0.13

= $\underline{\underline{0.3973}}$

Entropy (a1 = False) = (0N, 5 Yes) = $\underline{\underline{0}}$

Entropy (S, a1) = Entropy(S) − $\sum \frac{|S_v|}{|S|}$ Entropy (S_v

(gain)

= 0.971 − $\frac{5}{10}$ × 0.3973 − $\frac{5}{10}$ (0).

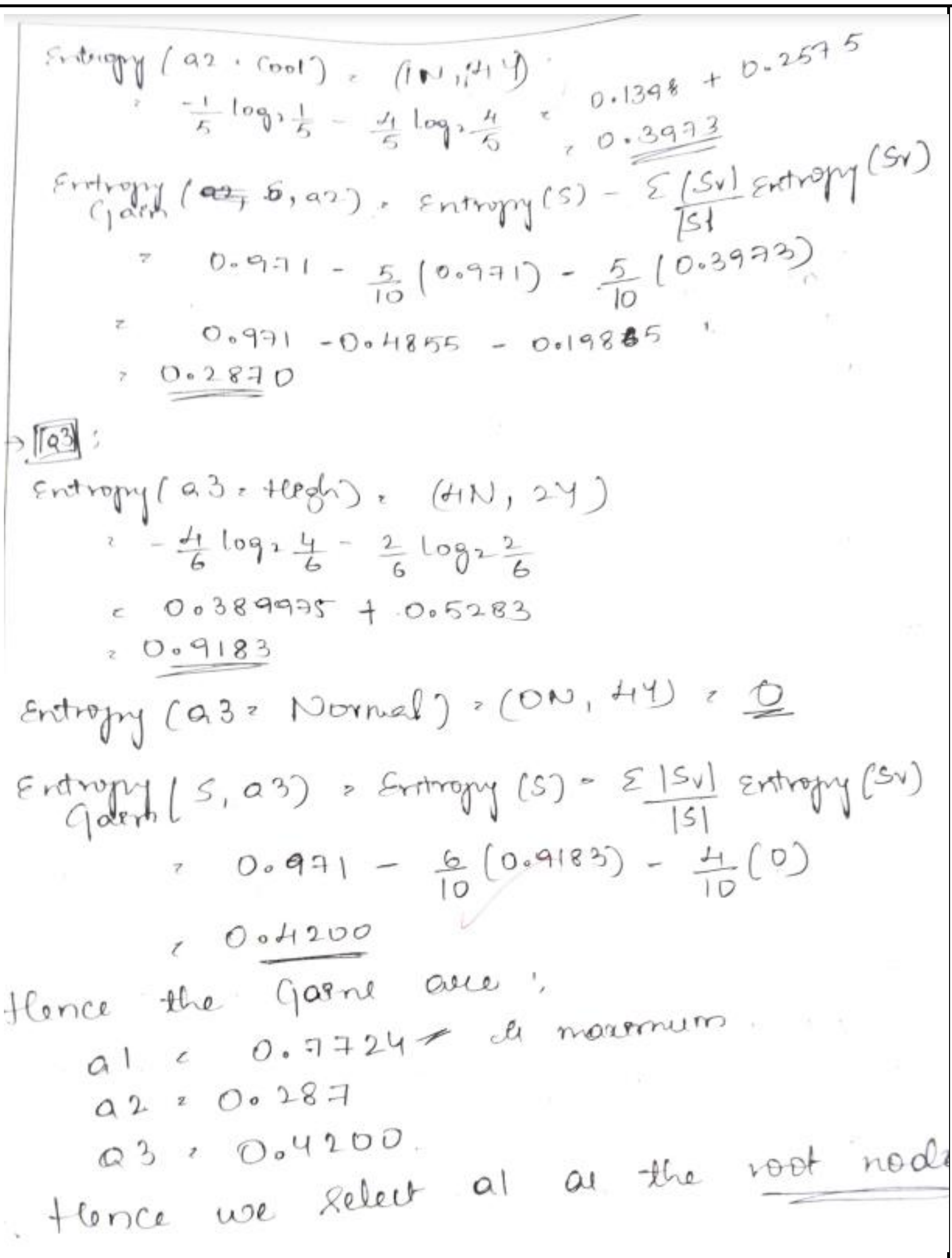
= $\underline{\underline{0.77235}}$

$\boxed{a2}$:

Entropy (a2 = Hot) = (3N, 2Y)

= $-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$ = 0.4422 + 0.5

= $\underline{0.971}$

	 Entropy (a2 = Cool) = (1N, 4Y) = 0.1398 + 0.2575 $= -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$ = 0.3973 Entropy Gain (~~a2~~ S, a2) = Entropy(S) − $\sum \frac{(S_v)}{\lvert S\rvert}$ Entropy (S_v) $= 0.971 - \frac{5}{10}(0.971) - \frac{5}{10}(0.3973)$ = 0.971 − 0.4855 − 0.19865 = 0.2870 → $\boxed{a3}$: Entropy (a3 = High) = (4N, 2Y) $= -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$ = 0.389975 + 0.5283 = 0.9183 Entropy (a3 = Normal) = (0N, 4Y) = 0 Entropy Gain (S, a3) = Entropy (S) − $\sum \frac{\lvert S_v\rvert}{\lvert S\rvert}$ Entropy (S_v) $= 0.971 - \frac{6}{10}(0.9183) - \frac{4}{10}(0)$ = 0.4200 Hence the Gains are : a1 = 0.7724 → is maximum a2 = 0.287 a3 = 0.4200. Hence we select a1 as the root nod			
3	a. Discuss the importance of data visualization in the context of data exploration and decision-making. b. Compare and contrast at least three different types of comparison plots, providing examples of when each might be most effectively used. Solution: a) Data visualization is crucial in both data exploration and decision-making, as it allows individuals to understand complex data quickly, identify patterns, spot anomalies, and ultimately make informed decisions. Here are key reasons why data visualization is vital in these contexts:	5+5	CO4	L 2

1. Enhances Data Exploration

- **Pattern Recognition**: Visualizations like scatter plots, heatmaps, and histograms allow data analysts and scientists to spot patterns, relationships, and trends that might not be obvious in raw data.
- **Outlier Detection**: Visualizations can reveal outliers or unusual data points that may indicate errors, data entry issues, or novel insights.
- **Efficient Summarization**: Large volumes of data can be summarized in an understandable format, helping analysts grasp the data's overall structure and main characteristics without diving into individual values.
- **Correlation Identification**: For example, in multi-variable data, visual tools like correlation matrices or pair plots make it easier to see how variables relate to each other, guiding further analysis.

2. Supports Decision-Making

- **Clarity in Communication**: Data visualization helps in communicating insights to non-technical stakeholders by translating complex analyses into clear, digestible visuals. Decision-makers can thus interpret insights without needing technical knowledge.
- **Evidence-Based Decisions**: Visualizations provide concrete, visually compelling evidence that can support or refute hypotheses, helping decision-makers make choices based on data rather than intuition.
- **Risk and Opportunity Identification**: By showing data in real-time or as trends over time, visualizations can highlight areas of potential risk or opportunity, enabling proactive decision-making.
- **Comparative Analysis**: Visualization allows for the easy comparison of different scenarios, trends, or key performance indicators (KPIs). Decision-makers can use tools like bar charts, line charts, and dashboards to evaluate different options at a glance.

3. Facilitates Interactive Analysis

- **Dynamic Exploration**: Interactive data visualizations (like dashboards) allow users to filter, drill down, or zoom into specific data subsets, empowering them to explore data from multiple perspectives.
- **Real-Time Insights**: With real-time data visualizations, decision-makers can monitor ongoing processes, adjust strategies promptly, and optimize outcomes as data flows in.
- **Increased Engagement**: Interactive tools engage users in the exploration process, fostering better understanding and retention of information, which leads to more confident decision-making.

4. Encourages Storytelling

- **Narrative Creation**: Data visualization helps to transform raw data into a compelling story, guiding stakeholders through an exploration of findings and potential implications.

- **Engagement and Persuasion**: When data is visually presented as part of a cohesive narrative, it is often more persuasive and memorable, helping leaders rally support for decisions.

Example: Business Decision-Making

In a business setting, a sales team might use data visualizations like line graphs to monitor monthly sales trends or geographical heatmaps to understand sales distribution. These insights can help them identify peak periods, focus on high-performing regions, and allocate resources effectively.

b) Comparison plots are essential in data visualization, allowing users to understand differences, similarities, and trends across datasets or categories. Here are three widely used comparison plots: **bar charts**, **line charts**, and **box plots**, each serving different purposes.

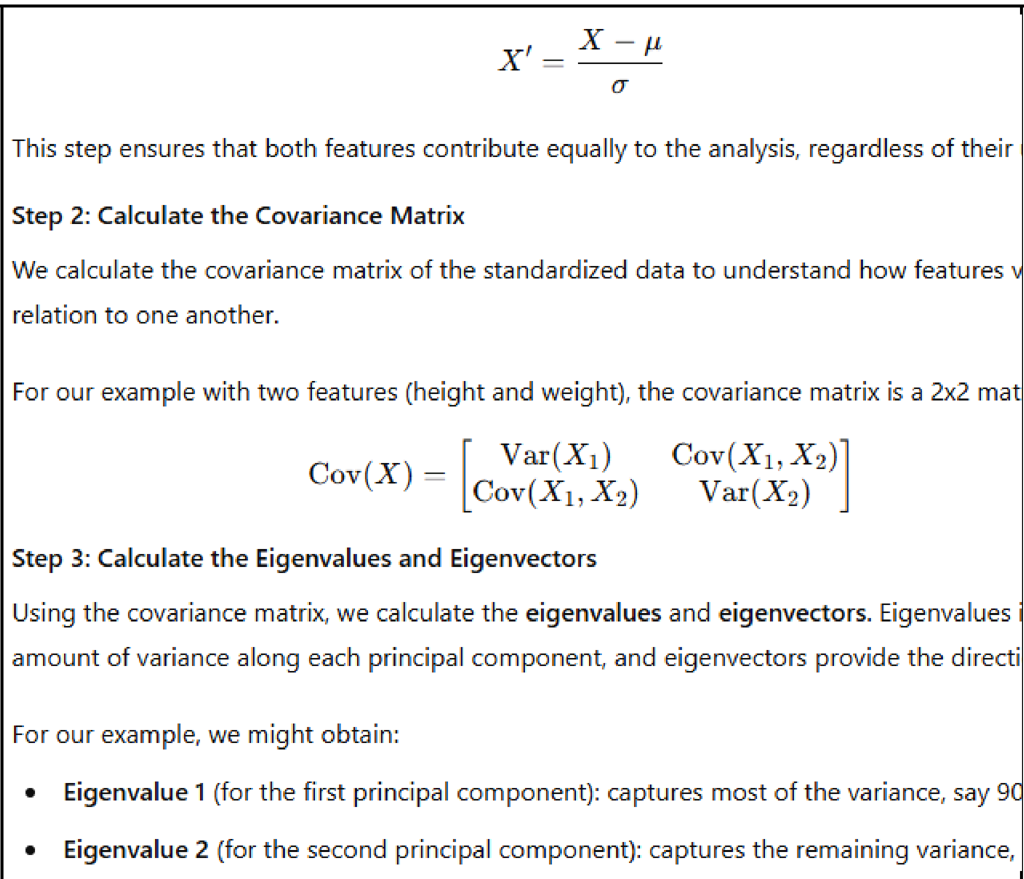
1. Bar Chart

- **Description**: A bar chart uses rectangular bars to show comparisons across discrete categories. The length of each bar is proportional to the value it represents.
- **Use Case**: Bar charts are ideal for comparing quantities across distinct categories, such as sales by product or counts by region.
- **Example**: Suppose a business wants to compare quarterly revenue for different product categories. A bar chart with each quarter as separate bars would provide a clear visual comparison across categories.
- **Strengths and Limitations**:
    - **Strengths**: Bar charts are easy to interpret and effective for comparing values across discrete categories.
    - **Limitations**: They can become cluttered with too many categories, making it difficult to interpret the differences.

2. Line Chart

- **Description**: A line chart displays data points connected by lines to show trends over continuous intervals, often time-based.
- **Use Case**: Line charts are effective for visualizing trends over time, such as tracking changes in stock prices, website traffic, or temperature fluctuations.
- **Example**: A line chart could be used to compare monthly sales revenue over the past two years, with each year as a separate line. This format clearly shows seasonality, upward or downward trends, and variations between years.
- **Strengths and Limitations**:
    - **Strengths**: Line charts are excellent for showing changes over time and are easy to interpret for time-series data.
    - **Limitations**: Not ideal for comparing categorical data without a natural ordering, and overlapping lines can reduce clarity when comparing multiple categories.

	3. Box Plot (Box-and-Whisker Plot) • **Description**: A box plot shows the distribution of data based on quartiles, displaying the median, lower and upper quartiles, and potential outliers. • **Use Case**: Box plots are useful for comparing the spread and distribution of a variable across categories, highlighting central tendency, spread, and any outliers. • **Example**: A box plot could be used to compare exam scores across different classes or departments, showing not only the median scores but also the range and variability within each class. • **Strengths and Limitations**: ○ **Strengths**: Box plots are very effective for highlighting distribution characteristics like spread, skewness, and outliers. ○ **Limitations**: They are not as intuitive for audiences unfamiliar with quartile-based representations and may not clearly show trends over time.			
4	Illustrate the PCA algorithm with a suitable example? Solution: (Explanation -7 marks , example-3 marks) Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify complex datasets while preserving as much variance as possible. PCA achieves this by transforming the original features into a new set of orthogonal features called **principal components**, which capture the directions of maximum variance in the data. **Steps in the PCA Algorithm** Here's how PCA works, along with an example to illustrate each step: *Example Scenario* Suppose we have a dataset with two features, **X1 (height)** and **X2 (weight)**, of a group of individuals. We want to reduce these two features to a single dimension while capturing as much variance as possible in the data. *Step 1: Standardize the Data* PCA is affected by scale, so we first standardize each feature to have a mean of 0 and a standard deviation of 1.	10	CO3	L 2



$$X' = \frac{X - \mu}{\sigma}$$

This step ensures that both features contribute equally to the analysis, regardless of their

Step 2: Calculate the Covariance Matrix

We calculate the covariance matrix of the standardized data to understand how features v relation to one another.

For our example with two features (height and weight), the covariance matrix is a 2x2 mat

$$\mathrm{Cov}(X) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{bmatrix}$$

Step 3: Calculate the Eigenvalues and Eigenvectors

Using the covariance matrix, we calculate the **eigenvalues** and **eigenvectors.** Eigenvalues i amount of variance along each principal component, and eigenvectors provide the directi

For our example, we might obtain:

- **Eigenvalue 1** (for the first principal component): captures most of the variance, say 90
- **Eigenvalue 2** (for the second principal component): captures the remaining variance,

The eigenvector corresponding to Eigenvalue 1 gives us the direction of maximum variance.

Step 4: Sort Eigenvalues and Select Principal Components

We sort the eigenvalues in descending order and select the top kkk eigenvectors as our **principal components**. Here, we want a 1-dimensional output, so we choose the eigenvector with the highest eigenvalue (representing 90% of the variance).

Step 5: Project the Data onto the Principal Components

Finally, we project the original data onto the selected principal component(s) to obtain the reduced-dimension dataset.

If our principal component is represented by the vector P1P_1P1, then the projected data ZZZ for each data point XXX is:

Z=X·P1Z = X \cdot P_1Z=X·P1

In our example, this new 1-dimensional representation captures the primary structure of the data (e.g., a combination of height and weight), summarizing the individuals along the main axis of variance.

	Visualizing the Result In a 2D plot, the points might form an elongated shape, showing a strong correlation between height and weight. The first principal component would be a line passing through this shape in the direction of maximum variance. The data points, when projected onto this line, provide a simplified 1-dimensional representation that retains most of the original variability.			
5	a. Describe the different types of geo plots and their respective use cases. b. Give an example of how a choropleth map can be used to visualize customer retention data across different regions. Solution: (% types-5 marks) Geo plots are specialized visualizations that display data on maps, providing spatial context to the data. They are particularly useful in fields like geography, urban planning, epidemiology, and logistics. Here are common types of geo plots and their respective use cases: **1. Choropleth Map** • **Description**: A choropleth map uses color gradients or shading to represent data values across geographic regions (such as countries, states, or counties). • **Use Case**: Ideal for displaying regional data, such as population density, average income, or unemployment rates. For example, a choropleth map can show COVID-19 case rates across different countries or states by using color intensity to indicate the level of cases. **2. Heat Map** • **Description**: A heat map displays data as color intensity over a map, with darker or lighter colors representing higher or lower concentrations of data points. Unlike choropleths, heat maps do not rely on distinct geographic boundaries. • **Use Case**: Useful for visualizing the density of events or occurrences in a specific area, such as accident hotspots, crime rates, or customer locations. For instance, a heat map could highlight where most traffic accidents occur in a city, with darker regions indicating higher frequency. **3. Scatter Plot Map (Point Map)** • **Description**: A scatter plot map uses points to represent specific locations, often with the option to encode additional data through point color, size, or shape. • **Use Case**: Suitable for displaying individual events or data points that have specific geographic coordinates, such as store locations, user check-ins, or landmarks. A retail company, for example, might use a scatter plot map to display all store locations, with point size representing each store’s revenue.	[5+5]	CO4	L2

4. Bubble Map

- **Description**: A bubble map is similar to a scatter plot map but uses bubbles or circles of varying sizes to represent quantitative data at specific locations.
- **Use Case**: Useful for comparing quantities at specific locations, such as population size, number of facilities, or economic metrics. For example, a bubble map could display cities worldwide with bubble size representing population, providing a quick visual comparison of city sizes.

5. Flow Map

- **Description**: A flow map shows movement between locations, with arrows or lines indicating the direction and sometimes the volume of flow.
- **Use Case**: Commonly used to illustrate migration patterns, trade routes, traffic flows, or supply chains. For instance, a flow map can show international trade by connecting countries with arrows indicating the direction of trade flow and line thickness representing trade volume.

b) A **choropleth map** can be a powerful tool for visualizing customer retention data across different regions. For instance, consider a retail company with stores or service coverage across multiple states or countries. The company wants to understand where customer retention rates are high or low to develop targeted strategies.

Example Scenario

Imagine a telecommunications company with customers across the United States. They have calculated the **customer retention rate** (percentage of customers who continue to use their services) for each state.

Steps to Create and Use the Choropleth Map

1. **Data Collection**: Collect the customer retention rates for each state. For example:
   - California: 85%
   - Texas: 75%
   - New York: 90%
   - Florida: 70%
   - Illinois: 65%
   - And so forth for each state.
2. **Map Visualization**: Create a choropleth map of the United States, where:
   - **Color Intensity**: Use a color gradient (e.g., shades of green) to represent retention rates. Darker shades of green might indicate higher retention rates, while lighter shades represent lower rates.
3. **Interpretation**:
   - Regions with **high retention rates** (e.g., New York and California) will appear in darker green, indicating a strong customer base.

	o Regions with **low retention rates** (e.g., Florida and Illinois) will appear in lighter green, signaling potential issues or areas for improvement. 4. **Actionable Insights**: o The company can investigate the factors contributing to lower retention in states like Florida and Illinois. They may then decide to run targeted customer loyalty programs, improve service offerings, or enhance customer support in these regions to boost retention. o Similarly, regions with high retention rates can serve as benchmarks to understand what practices might be replicated elsewhere.			
6	Explain the concept of a Random Forest and its features with relevant example Solution: (Concept, features and examples- 4+4+2 marks) **Random Forest** is an ensemble learning technique used for both classification and regression tasks. It operates by building multiple decision trees during training and aggregating their outputs to make predictions. The goal of Random Forest is to improve the predictive accuracy and control overfitting that may occur with individual decision trees. **Key Concepts and Features of Random Forest** 1. **Ensemble of Decision Trees**: Random Forest consists of many decision trees. Each tree is trained on a different subset of the data (using **bagging**), and their results are averaged (for regression) or voted on (for classification). 2. **Randomness in Tree Building**: o **Bootstrap Sampling**: Each tree is trained on a random sample of the training data, allowing trees to capture different aspects of the data. o **Feature Randomness**: At each split in a tree, only a random subset of the features is considered. This feature randomness reduces the likelihood of trees becoming too similar, improving generalization. 3. **Aggregating Predictions**: Once the forest is built, predictions are made by taking the **average of predictions** from each tree (for regression) or the **majority vote** (for classification). 4. **Reduced Overfitting**: By averaging many decision trees, Random Forest generally has lower variance than a single decision tree, making it more robust and less prone to overfitting. 5. **Feature Importance**: Random Forest provides a measure of feature importance by evaluating how much each feature contributes to reducing the overall error across all trees. **Example: Predicting Loan Approval** Suppose a bank wants to predict whether a loan applicant will default. They have a dataset with features like **income level, credit score, age, employment status, loan amount**, etc. 1. **Training Phase**:	10	CO3	L 2

  - A Random Forest is built by creating, say, 100 decision trees. Each tree is trained on a random subset of applicants and only considers a random subset of features at each split.
  - The randomness ensures that each tree learns a different aspect of the dataset, leading to diverse decision boundaries.

2. **Making Predictions**:
   - For a new applicant, each tree in the forest will output either "approve" or "reject" as its prediction.
   - The Random Forest takes the majority vote across all trees. For example, if 70 out of 100 trees say "approve" and 30 say "reject," the Random Forest will predict "approve."
3. **Evaluating Feature Importance**:
   - The bank can also interpret which features are most important in predicting loan defaults by examining the feature importance scores given by the Random Forest. For instance, if **credit score** and **income level** are the most important, the bank can prioritize these features when assessing risk.

Advantages of Random Forest

- **High Accuracy**: By aggregating multiple trees, Random Forest often achieves higher accuracy compared to single decision trees.
- **Robustness**: It is less sensitive to outliers and noisy data than single decision trees.
- **Feature Importance**: It provides insights into which features are most influential in making predictions, helpful for interpretability.

Disadvantages of Random Forest

- **Computationally Intensive**: Training many trees can be slow, especially for large datasets.
- **Less Interpretability**: Although it shows feature importance, understanding individual predictions is difficult compared to simpler models.