CMR
INSTITUTE OF
TECHNOLOGY

USN □□□□□□□□□□

## Internal Assessment Test - II

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Sub: | **Exploratory Data Analysis for Business** | | | | | | | Code: | **22MBABA304** |
| Date: | **17-04-2025** | Duration: | **90 mins** | Max Marks: | **50** | Sem: | **III** | Branch: | **MBA** |
| **SET- I** | | | | | | | | | |

**1.a. List 3 key differences between Ridge regression & LASSO regression.**

| Criteria | Ridge Regression | LASSO Regression |
|---|---|---|
| Penalty term | Ridge Regression uses L2 regularization(Squared penalty term), which shrinks coefficients but does not force them to be exactly zero. | LASSO (Least Absolute Shrinkage and Selection Operator) uses L1 regularization(absolute penalty term), which can shrink some coefficients to exactly zero, effectively performing feature selection. |
| Feature selection | Ridge Regression does not perform feature selection; it only reduces coefficient magnitudes. | LASSO can eliminate less important features by setting their coefficients to zero |
| Multicollinearity (high correlation among feature variables) | Ridge Regression is more effective in handling multicollinearity (highly correlated predictors) by distributing coefficient values among correlated variables. | LASSO may arbitrarily select one feature from a group of highly correlated variables and shrink the rest to zero, which might lead to loss of important information. |

**1.b. Explain the process and steps in constructing a decision tree.**

Steps to Construct a Decision Tree:

1. Selecting the Best Feature (Splitting Criterion)

- Choose a feature to split the dataset based on a measure of impurity:
  - For Classification: Use Gini Impurity or Entropy (Information Gain).
  - For Regression: Use Mean Squared Error (MSE) or Variance Reduction.
- The feature that provides the most significant reduction in impurity is selected for the split.

2. Splitting the Dataset

- The selected feature is used to split the dataset into subsets.
- The process repeats recursively for each subset.

3. Stopping Criteria (Tree Growth Limitation)

A tree keeps splitting until it meets one of the stopping conditions:

- Maximum Depth Reached: Limits tree growth to prevent overfitting.
- Minimum Samples per Leaf: Ensures a node has a minimum number of samples before splitting.
- Impurity Threshold: Stops if the impurity is below a predefined threshold.

4. Assigning Class Labels (For Classification) or Values (For Regression)

- For Classification: Each leaf node is assigned the majority class from its subset.
- For Regression: Each leaf node is assigned the mean/median of the target values in that subset.

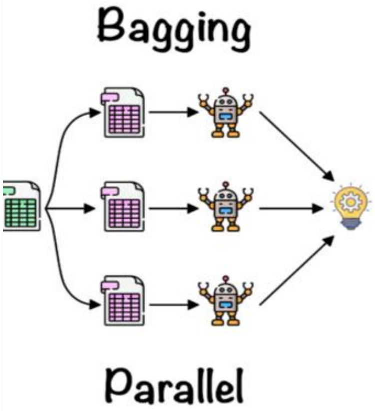
5. Pruning the Tree (Optimizing Performance)

Pruning reduces overfitting by removing unnecessary branches:

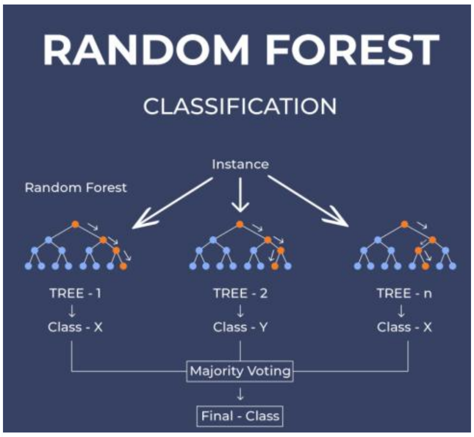
- Pre-Pruning (Early Stopping): Stops tree growth early based on predefined constraints.
- Post-Pruning (Prune After Full Growth): Removes branches that do not improve model performance (done using cross-validation).

**1.c. What is Bagging? Justify Random Forest as a way for prediction accuracy**.

Bagging (Bootstrap Aggregating) is an ensemble learning technique designed to improve the stability and accuracy of machine learning algorithms, especially decision trees. It works by combining the predictions of multiple models trained on different subsets of the training data.



Random Forest is a specific implementation of bagging applied to decision trees with an extra twist: when splitting a node during tree construction, only a random subset of features is considered rather than all features.
This increases diversity among the trees and leads to better performance.



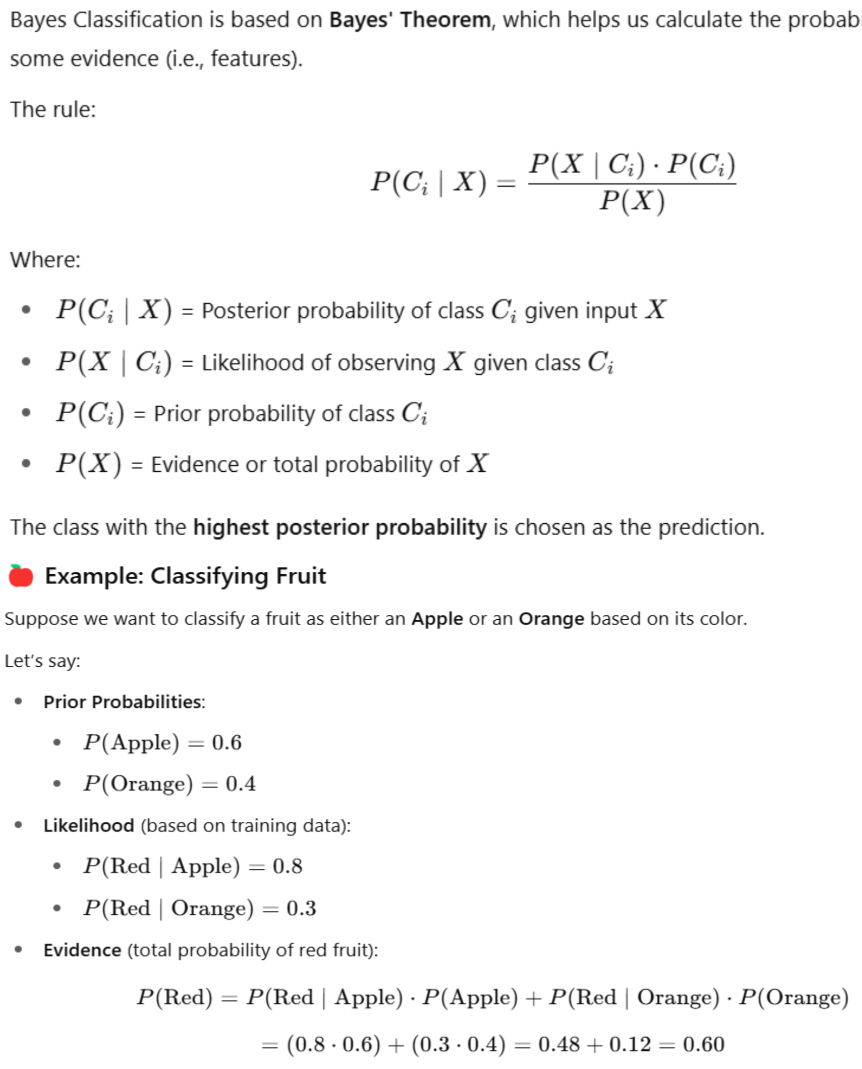
Why Random Forest Improves Prediction Accuracy

1. Reduces Overfitting: By combining many decision trees, random errors and overfitting of individual trees are averaged out.
2. Handles High Dimensionality: The random selection of features at each split helps deal with irrelevant features and makes it more robust.
3. Robust to Noise: Since it averages many trees, it's less sensitive to outliers or noise in the training data.
4. Works Well Out-of-the-Box: Minimal parameter tuning is needed to get good results.
5. Parallelizable: Each tree is built independently, making training fast with the right infrastructure.

## 2.a. How PCA is different from discriminant analysis.

- PCA (Principal Component Analysis) is an unsupervised technique that reduces dimensionality by maximizing variance without considering class labels.
- Discriminant Analysis (like LDA) is a supervised technique that maximizes class separability using label information.
- PCA focuses on feature variance, while Discriminant Analysis focuses on maximizing the between-class variance and minimizing within-class variance.

## 2.b. Illustrate Bayes Classification rule with an example.



Bayes Classification is based on **Bayes' Theorem**, which helps us calculate the probability of a class given some evidence (i.e., features).

The rule:

$$P(C_i \mid X) = \frac{P(X \mid C_i) \cdot P(C_i)}{P(X)}$$

Where:

- $P(C_i \mid X)$ = Posterior probability of class $C_i$ given input $X$
- $P(X \mid C_i)$ = Likelihood of observing $X$ given class $C_i$
- $P(C_i)$ = Prior probability of class $C_i$
- $P(X)$ = Evidence or total probability of $X$

The class with the **highest posterior probability** is chosen as the prediction.

### 🍎 Example: Classifying Fruit

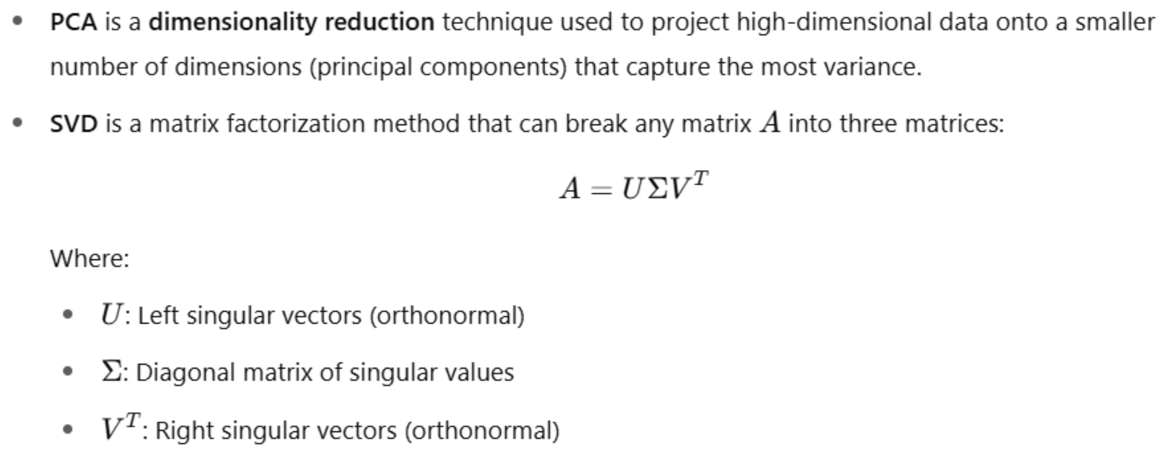
Suppose we want to classify a fruit as either an **Apple** or an **Orange** based on its color.

Let’s say:

- **Prior Probabilities**:
  - $P(\text{Apple}) = 0.6$
  - $P(\text{Orange}) = 0.4$
- **Likelihood** (based on training data):
  - $P(\text{Red} \mid \text{Apple}) = 0.8$
  - $P(\text{Red} \mid \text{Orange}) = 0.3$
- **Evidence** (total probability of red fruit):

$$P(\text{Red}) = P(\text{Red} \mid \text{Apple}) \cdot P(\text{Apple}) + P(\text{Red} \mid \text{Orange}) \cdot P(\text{Orange})$$
$$= (0.8 \cdot 0.6) + (0.3 \cdot 0.4) = 0.48 + 0.12 = 0.60$$

## 2.c. Analyze how Singular Value Decomposition (SVD) is used in Principal Component Analysis (PCA).

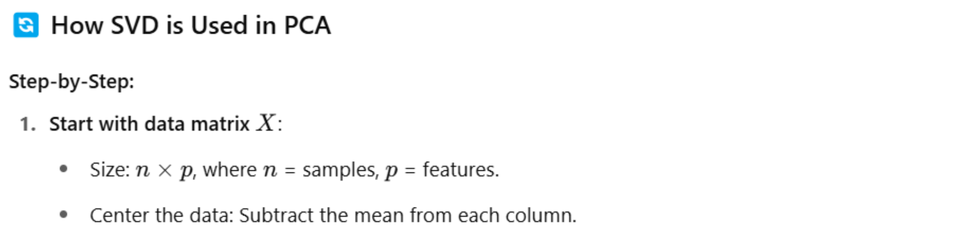


- **PCA** is a **dimensionality reduction** technique used to project high-dimensional data onto a smaller number of dimensions (principal components) that capture the most variance.
- **SVD** is a matrix factorization method that can break any matrix $A$ into three matrices:

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

Where:

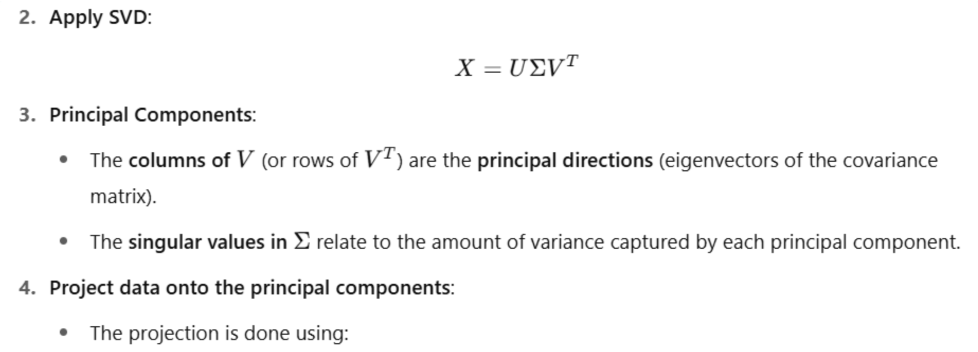
- $U$: Left singular vectors (orthonormal)
- $\Sigma$: Diagonal matrix of singular values
- $V^T$: Right singular vectors (orthonormal)



### 🔄 How SVD is Used in PCA

**Step-by-Step:**

1. **Start with data matrix $X$**:
   - Size: $n \times p$, where $n$ = samples, $p$ = features.
   - Center the data: Subtract the mean from each column.

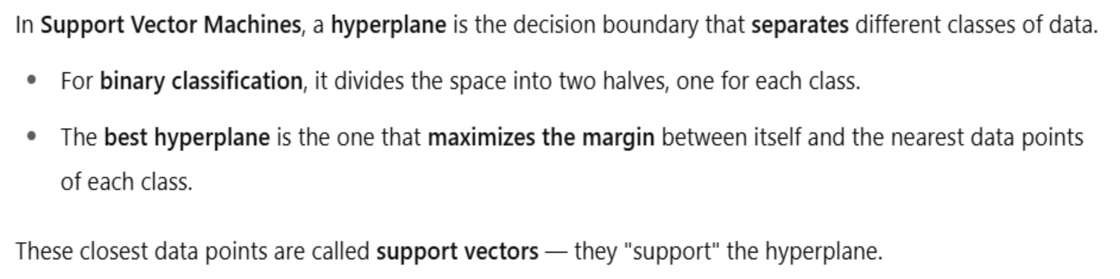


2. **Apply SVD**:

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

3. **Principal Components**:
   - The **columns of $V$** (or rows of $V^T$) are the **principal directions** (eigenvectors of the covariance matrix).
   - The **singular values in $\Sigma$** relate to the amount of variance captured by each principal component.
4. **Project data onto the principal components**:
   - The projection is done using:

## 3.a. What do you mean by hyperplane?



In **Support Vector Machines**, a **hyperplane** is the decision boundary that **separates** different classes of data.

- For **binary classification**, it divides the space into two halves, one for each class.
- The **best hyperplane** is the one that **maximizes the margin** between itself and the nearest data points of each class.

These closest data points are called **support vectors** — they "support" the hyperplane.

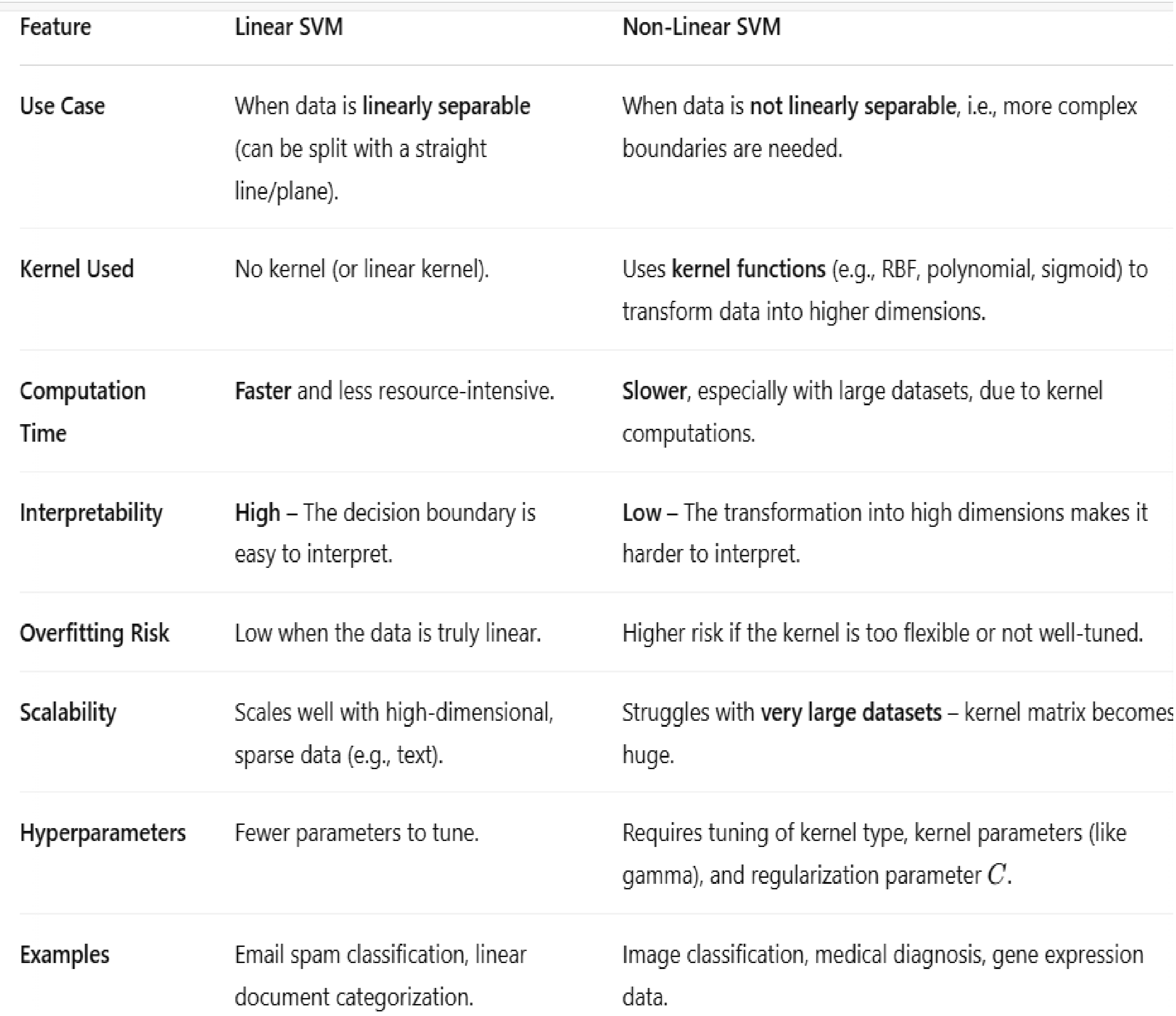
## 3.b. Give an overview of SVM and its application areas.

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and sometimes regression.

SVMs are known for their high accuracy, especially in high-dimensional spaces and small to medium-sized datasets.

1. Text Classification / Spam Detection
    - Classifying emails as spam or not spam
    - Sentiment analysis (positive/negative reviews)
2. Image Classification
    - Facial recognition
    - Object detection (e.g., cat vs. dog classifier)
3. Bioinformatics
    - Gene expression classification
    - Disease diagnosis based on medical data
4. Finance
    - Credit scoring
    - Fraud detection

### 3.c. Critically analyze the implementation of linear & non-linear support vector machines.



| Feature | Linear SVM | Non-Linear SVM |
|---|---|---|
| **Use Case** | When data is **linearly separable** (can be split with a straight line/plane). | When data is **not linearly separable**, i.e., more complex boundaries are needed. |
| **Kernel Used** | No kernel (or linear kernel). | Uses **kernel functions** (e.g., RBF, polynomial, sigmoid) to transform data into higher dimensions. |
| **Computation Time** | **Faster** and less resource-intensive. | **Slower**, especially with large datasets, due to kernel computations. |
| **Interpretability** | **High** – The decision boundary is easy to interpret. | **Low** – The transformation into high dimensions makes it harder to interpret. |
| **Overfitting Risk** | Low when the data is truly linear. | Higher risk if the kernel is too flexible or not well-tuned. |
| **Scalability** | Scales well with high-dimensional, sparse data (e.g., text). | Struggles with **very large datasets** – kernel matrix becomes huge. |
| **Hyperparameters** | Fewer parameters to tune. | Requires tuning of kernel type, kernel parameters (like gamma), and regularization parameter $C$. |
| **Examples** | Email spam classification, linear document categorization. | Image classification, medical diagnosis, gene expression data. |

4. **Customer Churn Prediction:**
A telecom company wants to reduce customer churn. They have collected customer data, including demographics, service usage, and past complaints.

- What are the key features you would consider for building the model?
- Build a decision tree and interpret the model.

## Customer Churn Prediction- Decision Tree

## Features

### A. Demographics

- Age 18–25, 26–35, etc.
- Gender Male, Female, Other
- Income Level Low, Medium, High
- Region / Location Urban, Suburban, Rural
- Marital Status Single, Married, Divorced
- Occupation Student, Professional, Retired
- Location (Urban/Rural, region, etc.)

### B. Service Usage

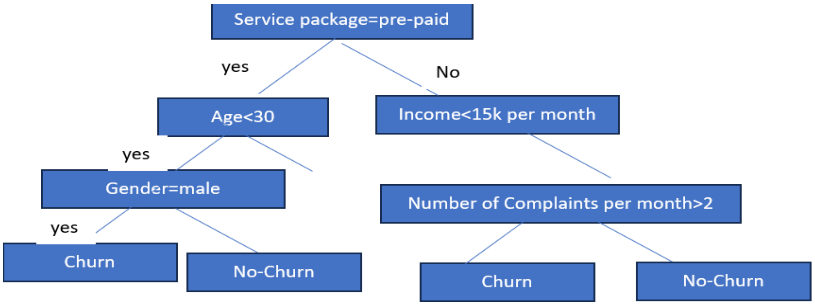
- Number of services subscribed
- Frequency of use
- Monthly spending
- Contract type (e.g., prepaid, postpaid)
- Tenure (how long they've been a customer)

### C. Past Complaints

- Number of complaints
- Complaint categories (e.g., billing, service downtime)
- Resolution time
- Whether complaints were resolved successfully

## Target Variable

- **Churn**: Binary variable → churned/ Not-churned



- Younger customers might have less brand loyalty or lower patience with service issues. → Consider loyalty programs or onboarding journeys targeted at younger users.
- Lower-income groups may churn due to pricing. → Introduce budget-friendly options or flexible plans.