Q1. In bagging, what happens to the weights of the correctly classified sample?

A. It increases

B. It remains the same

C. It decreases

D. It doubles

✅ **Correct Answer: C. It decreases**

Q2. What is the result of combining weak classifiers in ADABOOST?

A. One weak classifier

B. One strong classifier

C. Many weak classifiers

D. Random output

✅ **Correct Answer: B. One strong classifier**

Q3. Which of the following ensemble methods reacts more to noisy data?

A. Bagging

B. Boosting

C. Random Forest

D. Stacking

✅ **Correct Answer: B. Boosting**

Q4. What does the final model in stacking learn?

A. Raw data features

B. Noise patterns

C. Predictions of base models

D. Random weights

✅ **Correct Answer: C. Predictions of base models**

Q5. Which of the following is true regarding ensemble methods?

A. They always perform worse than single models

B. They often outperform individual models

C. They require no training data

D. They cannot combine classifiers

✅ **Correct Answer: B. They often outperform individual models**

Q6. Choose the ensemble method that helps to reduce model variance.

A. Boosting

B. Bagging

C. Stacking

D. ECOC

✅ **Correct Answer: B. Bagging**

Q7. Which of the following is a potential disadvantage of stacking?

A. It trains too fast

B. It cannot be interpreted easily

C. It reduces accuracy
D. It uses less data
✅ Correct Answer: B. It cannot be interpreted easily

Q8. How does ECOC handle multi-class binary classification?
A. By merging classes together
B. By splitting into multiple binary models
C. By ignoring binary classes
D. By using one classifier for all classes
✅ Correct Answer: B. By splitting into multiple binary models

Q9. Select the correct reason that explains why overfitting is less in bagging.
A. It ignores errors
B. It averages the result of each model
C. It increases model complexity
D. It uses less data for training
✅ Correct Answer: B. It averages the result of each model

Q10. Identify the ensemble method that uses bootstrapped datasets.
A. AdaBoost
B. Gradient Boosting
C. Random Forest
D. Stacking
✅ Correct Answer: C. Random Forest

Q1. Briefly explain the trade-off between bias and variance.
Ans:

1. Bias: It is the error caused when a model is too simple and fails to capture the real pattern of data (underfitting).
2. Variance: It is the error caused when a model is too complex and fits even the noise in data (overfitting).
3. Trade-off: Increasing model complexity reduces bias but increases variance, and simplifying the model reduces variance but increases bias.
4. Goal: Achieve a balanced point where both bias and variance are low to get better accuracy on new data.
5. Example: Linear regression has high bias, low variance; decision tree has low bias, high variance — a mix is ideal.

Q2. How does hyperparameter tuning affect the performance of a Machine Learning model.
Ans:

1. Definition: Hyperparameters are settings that control how an ML model learns, like learning rate, tree depth, or number of clusters.
2. Better accuracy: Correct tuning helps the model learn patterns more effectively and improve accuracy.

3. **Avoid overfitting: Proper values prevent the model from becoming too complex.**
4. **Faster training: Good tuning reduces training time and improves stability.**
5. **Methods used: Grid Search, Random Search, and Bayesian Optimization are common tuning methods.**

Q3. Difference between K-Means algorithm and other clustering algorithms for unsupervised machine learning.

Ans:

Feature	K-Means	Other Clustering Algorithms (e.g., Hierarchical, DBSCAN)
Type	Partition-based	Hierarchical or density-based
Speed	Very fast and efficient	Slower for large datasets
Cluster Shape	Works best for circular/spherical clusters	Handles irregular cluster shapes
Number of Clusters	Must be given in advance (k value)	Can find number of clusters automatically
Noise Handling	Sensitive to outliers	DBSCAN and others handle noise better

Q4. Briefly explain how we can use Random Forest with AdaBoost in case of noisy datasets.

Ans:

1. **Random Forest: Reduces variance by averaging results from multiple decision trees using bagging.**
2. **AdaBoost: Combines weak learners and focuses on misclassified data to build a strong model.**
3. **Combination use: Random Forest can be used as the base learner inside AdaBoost to improve stability.**
4. **Handling noise: Random Forest reduces noise sensitivity, while AdaBoost improves learning accuracy.**
5. **Result: Together they produce a strong, balanced model that performs well even with noisy or imperfect data.**

Q1. Compare and evaluate Decision Tree and Random Forest algorithm for training purposes.

Ans:

Feature	Decision Tree	Random Forest
Structure	Single tree-based model	Collection of multiple decision trees
Training Speed	Faster to train	Slower due to multiple trees
Overfitting	More prone to overfitting	Reduces overfitting by averaging results
Accuracy	Moderate	Higher and more stable accuracy
Interpretability	Easy to understand and visualize	Harder to interpret due to multiple trees

Q2. What steps would you take to train a Decision Tree model in Python?

Ans:

1. **Import libraries: Use sklearn.tree for model creation.**
2. **Load data: Import dataset using pandas or sklearn datasets.**
3. **Split data: Divide data into training and testing sets using train_test_split.**
4. **Train model: Use DecisionTreeClassifier().fit(X_train, y_train).**
5. **Evaluate model: Predict using .predict(X_test) and calculate accuracy or confusion matrix.**

Q3. What type of data structure is often used to implement Random Forest?

Ans:

- **Answer: Decision Tree**
- **Random Forest uses multiple Decision Trees as its building blocks, combining their results to make final predictions.**

Q4. What are the advantages of Decision Tree?

Ans:

1. **Easy to understand: Simple and visual representation of data.**
2. **Handles both data types: Works with numerical and categorical data.**

3. **No scaling needed: Does not require data normalization or standardization.**
4. **Feature importance: Helps identify which features affect predictions most.**
5. **Quick decision-making: Fast to train and interpret results.**

Q5. List some applications of Decision Tree.

Ans:

1. **Medical diagnosis: Predicts diseases based on patient symptoms.**
2. **Finance: Helps in loan approval and credit risk analysis.**
3. **Marketing: Used for customer segmentation and purchase prediction.**
4. **Education: Predicts student performance and dropout rates.**
5. **E-commerce: Suggests products based on buying patterns.**

Q6. How is the final prediction made in Random Forest for classification?

Ans:

1. **Each decision tree gives its own prediction.**
2. **All predictions are collected together.**
3. **The majority vote among all trees decides the final output.**
4. **This voting system ensures more reliable results.**
5. **It helps in improving accuracy and reducing model error.**

Q7. What aspect of a Random Forest Model reduces overfitting?

Ans:

- **Answer: Averaging across multiple trees**
- **Random Forest reduces overfitting by combining many decision trees and taking the average of their results, which balances variance and increases generalization.**

Q8. List some benefits of using Neural Networks.

Ans:

1. **High accuracy: Learns complex patterns and relationships in data.**
2. **Automation: Extracts features automatically without manual effort.**
3. **Adaptability: Works well with large and diverse datasets.**
4. **Non-linear modeling: Handles data that isn't linearly related.**
5. **Scalability: Can be expanded with more layers for better performance.**

Q9. What does a Neural Network mimic?

Ans:

1. **Mimics the working of the human brain.**
2. **Uses artificial neurons that pass signals through connections.**
3. **Each neuron processes information and passes it to the next layer.**
4. **It learns from data by adjusting connection weights.**
5. **Used for tasks like image recognition, voice processing, and predictions.**

Q10. Explain the purpose of an activation function in a Neural Network.
Ans:

1. **Non-linearity: Adds non-linear behavior to the model.**
2. **Decision-making: Helps neurons decide whether to activate or not.**
3. **Complex learning: Allows the network to learn complicated relationships.**
4. **Types: Common ones are Sigmoid, ReLU, and Tanh.**
5. **Result: Without activation, the network behaves like a simple linear model.**

Q11. What are the different layers of a Neural Network? Which layer adjusts the weights during training?
Ans:

Layer	Function
Input Layer	Takes data features into the network.
Hidden Layers	Process data and find relationships.
Output Layer	Gives final prediction result.
Adjusting Weights	Hidden layers adjust weights during backpropagation to reduce error.

Q12. Which component of Neural Network is used to adjust the weights?
Ans:

1. **Backpropagation Algorithm is used to update and adjust weights.**
2. **It calculates the error between actual and predicted output.**
3. **Then adjusts the weights to minimize the error using gradient descent.**
4. **This process repeats until the model achieves low loss.**
5. **It ensures the network learns effectively from data.**

Q13. Briefly explain the technique that helps to avoid overfitting in Neural Network.
Ans:

1. **Dropout: Randomly turns off some neurons during training.**
2. **Regularization: Adds penalty terms to reduce large weight values.**
3. **Early stopping: Stops training when validation loss starts rising.**
4. **Data augmentation: Increases data variety by adding more examples.**
5. **Cross-validation: Tests model performance on multiple data parts.**

Q14. Explain the concept of Neural Network in Machine Learning with an example.
Ans:

1. **Concept: A neural network is a system of connected layers (neurons) that process data and learn from examples.**
2. **Structure: Has input, hidden, and output layers that pass information forward.**
3. **Learning: It adjusts weights to reduce prediction errors using backpropagation.**
4. **Example: In handwriting recognition, it learns to identify letters by studying thousands of image samples.**
5. **Result: The model can accurately predict unseen handwritten characters.**

Q15. Compare and contrast the effectiveness of Neural Networks with traditional Machine Learning models.
Ans:

Feature	Neural Networks	Traditional ML Models
Data Handling	Works better with large datasets	Performs well on small datasets
Feature Extraction	Automatic feature learning	Needs manual feature selection
Performance	Very high accuracy with enough data	Moderate accuracy
Complexity	More complex and slower to train	Easier and faster
Interpretability	Harder to interpret	Easier to explain results

Q16. Evaluate the advantages and disadvantages of using Neural Networks for training large datasets.
Ans:

1. **Advantages:**
   - **Learns deep patterns and complex relations in large data.**
   - **Improves accuracy with more training data.**
   - **Automatically detects key features.**
2. **Disadvantages:**
   - **Needs large computational power and memory.**
   - **Training time is longer.**

- **Difficult to interpret or explain decisions.**

3. **Result: Very effective for tasks like image, text, and speech recognition when resources are available.**

Simple Linear Regression

y = mx + b

Where,

1. y is the predicted value(dependent variable)

2. x is the input(independent variable)

3. m is the slope of the (how much y changes when x changes)

4. b is the intercept(the value of y when x = 0)

M 3

Q1. Explain how to Preprocess a dataset and train it using at least two algorithm.
Ans:

1. **Data Cleaning:**
   - **Remove missing values, duplicates, and incorrect data entries.**
   - **Handle null values by replacing them with the mean, median, or mode.**
2. **Data Transformation:**
   - **Normalize or standardize data so all features are in a similar range.**
   - **Convert categorical data into numeric form using label encoding or one-hot encoding.**
3. **Data Splitting:**
   - **Divide the dataset into training (80%) and testing (20%) sets to evaluate performance.**
4. **Model Training (Algorithm 1 – Linear Regression):**
   - **Train the model to find the best-fitting line between dependent and independent variables.**
   - **Used for predicting continuous values like price or demand.**
5. **Model Training (Algorithm 2 – Decision Tree):**
   - **Builds a tree-like structure where each node splits data based on conditions.**
   - **Useful for classification tasks like “Yes” or “No” predictions.**

Q2. Critically evaluate the trade off between Variance and Bias in model Training.
Ans:

Term	Meaning	Effect if High	Desired Condition
Bias	Error due to simplifying the model too much	Model misses important patterns → underfitting	Low
Variance	Error due to model being too complex	Model learns noise → overfitting	Low
Trade-off	Balance between bias and variance	High bias → poor accuracy; High variance → poor generalization	Balanced
Goal	Minimize total error	Find a model that fits well but not too tightly	Moderate bias, low variance
Example	Linear regression (high bias) vs Decision tree (high variance)	Shows importance of balance	—

Q3. What is model over fitting and underfitting
Ans:

1. **Overfitting:**
   - **The model learns both useful patterns and noise from training data.**
   - **Works well on training data but fails on new unseen data.**
2. **Underfitting:**
   - **The model is too simple and fails to capture important patterns.**
   - **Performs poorly on both training and testing data.**
3. **Cause of Overfitting:**
   - **Too many features, complex model, or small dataset.**
4. **Cause of Underfitting:**
   - **Very simple algorithm, not enough training, or missing data features.**
5. **Goal:**
   - **Build a model that fits training data just enough to generalize well to new data.**

Q4. How do you identify over fitting in a trained model
Ans:

1. **Performance gap:**
   - **Training accuracy is very high, but test accuracy is much lower.**
2. **Validation score drop:**
   - **Model performs worse during cross-validation than in training.**
3. **Learning curve:**
   - **Training loss keeps decreasing, but validation loss starts increasing.**
4. **Unstable predictions:**
   - **Small change in input data leads to big changes in output.**
5. **Too complex model:**
   - **Many layers or parameters compared to dataset size cause overfitting.**

Q5. Apply cross validation to validate model performance.
Ans:

1. **Definition:**
   - **Cross-validation divides data into multiple parts (folds) to test model performance fairly.**
2. **Process:**
   - **Split data into *k* equal folds (e.g., k=5).**
   - **Train on (k-1) folds and test on the remaining one.**
   - **Repeat this process *k* times.**
3. **Average accuracy:**
   - **Calculate the mean accuracy of all folds for a balanced performance measure.**
4. **Reduces bias:**
   - **Ensures model is tested on all data parts and prevents overfitting.**
5. **Common type:**
   - **K-Fold Cross Validation is the most widely used method for model validation.**

Q6. What are the methods to reduce over fitting.
Ans:

1. **Cross Validation:**
   - **Use k-fold cross-validation to get a balanced accuracy score.**
2. **Regularization:**
   - **Add penalty terms (L1, L2) to reduce the influence of large weights.**
3. **Feature Selection:**
   - **Keep only important features and remove unnecessary ones.**
4. **Early Stopping:**
   - **Stop training when validation loss starts increasing.**
5. **Data Augmentation:**
   - **Increase data variety by adding more samples or modifying existing ones.**

Q7. Evaluate the performance of a trained classification model using various matrix like accuracy, precision, recall, etc.
Ans:

Metric	Formula	Meaning	High Value Indicates
Accuracy	**(TP + TN) / (TP + TN + FP + FN)**	**How many predictions were correct**	**Overall good performance**
Precision	**TP / (TP + FP)**	**Correct positive predictions among total predicted positives**	**Low false positives**
Recall (Sensitivity)	**TP / (TP + FN)**	**Correctly predicted positives among all actual positives**	**Low false negatives**
F1 Score	**2 × (Precision × Recall) / (Precision + Recall)**	**Balance between precision and recall**	**Balanced performance**
Confusion Matrix	**Table showing TP, FP, TN, FN**	**Visual summary of prediction results**	**Easy error identification**

TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative

M 4

Q1. Describe how Machine Learning supports business decision making.
Ans:

1. **Data-driven insights**: ML analyzes large amounts of data to find useful patterns that help managers make smart decisions.
2. **Prediction ability**: It helps forecast sales, demand, and market trends to plan future strategies.
3. **Automation of analysis**: ML automates report generation, saving time and reducing human errors.
4. **Customer understanding**: It studies customer behavior to improve marketing and product recommendations.

5. **Faster decision-making**: Real-time data processing allows quick and accurate decisions in dynamic business environments.

Q2. Why is it important to focus on a business problem before selecting any machine learning model.

Ans:

1. **Clear goal**: Understanding the business problem ensures the model is built with a clear purpose.
2. **Model selection**: Different problems need different models; knowing the problem helps choose the right one.
3. **Resource saving**: Avoids wasting time and money on building models that don't solve real issues.
4. **Better results**: When focused on a real problem, ML solutions give more meaningful outcomes.
5. **Alignment with business needs**: Ensures the model supports company goals instead of just showing technical performance.

Q3. How can machine learning help in improving patient health outcomes.

Ans:

1. **Early disease detection**: ML can analyze medical data to identify health problems before they become serious.
2. **Personalized treatment**: Helps doctors create treatment plans based on each patient's unique data.
3. **Predicting risks**: Predicts chances of disease reoccurrence or side effects in advance.
4. **Medical image analysis**: Detects patterns in X-rays, MRIs, and CT scans faster and more accurately than humans.
5. **Remote monitoring**: ML in smart devices tracks patient health in real-time for better continuous care.

Q4. Explain the role of Machine Learning in detecting fraud in banking system.

Ans:

1. **Pattern recognition**: ML studies past transaction data to identify unusual behavior.
2. **Real-time monitoring**: Detects and blocks suspicious transactions as they happen.
3. **Risk scoring**: Assigns a risk score to each transaction to measure fraud probability.
4. **Adaptive learning**: The model keeps learning from new fraud techniques and improves itself.
5. **Customer safety**: Reduces financial loss and builds trust between banks and customers.

Q5. What is the importance of choosing the best fit learning model for a business case.

Ans:

1. **Accurate results**: A suitable model increases accuracy and reliability of predictions.
2. **Efficiency**: The right model uses less data and computing power, saving resources.
3. **Better decision-making**: Gives precise insights that help in strategic business moves.

4. **Avoiding overfitting**: Prevents the model from performing well only on training data but poorly on new data.
5. **Improved performance**: The correct model ensures faster processing and better overall business performance.

Q6. What are the key phases of applying Machine Learning to solve any business problem.
Ans:

1. **Problem understanding**: Identify and define the business goal clearly.
2. **Data collection**: Gather clean and relevant data for the problem.
3. **Data preprocessing**: Remove errors, fill missing values, and prepare data for training.
4. **Model training and testing**: Choose the right algorithm, train it, and test accuracy.
5. **Deployment and monitoring**: Apply the model in the real system and keep updating it regularly.

Q7. How can using Machine Learning help in Managing IT issue.
Ans:

1. **Predictive maintenance**: ML predicts hardware or software failure before it happens.
2. **Automated troubleshooting**: Suggests or applies quick fixes for common IT problems.
3. **Log analysis**: Analyzes huge IT logs to find the root cause of system errors.
4. **Security alerts**: Detects unusual activities that could lead to cyber threats.
5. **Resource optimization**: Helps manage system resources efficiently and reduces downtime.

Q8. Why is domain knowledge important in applying Machine Learning to business problem.
Ans:

1. **Understanding data**: Helps interpret the meaning of data correctly.
2. **Model relevance**: Ensures ML models are built with features that matter to the business.
3. **Error detection**: Experts can identify mistakes in data or results easily.
4. **Effective communication**: Makes collaboration between technical and business teams smoother.
5. **Better decisions**: Leads to more practical and useful solutions for real-world issues.

Q9. Explain the contribution of Machine Learning to reduce risk in enterprises.
Ans:

1. **Fraud detection**: Identifies suspicious transactions to prevent financial loss.
2. **Predictive analysis**: Forecasts future risks like market crashes or supply issues.
3. **Operational safety**: Detects system failures before they occur.
4. **Customer analysis**: Finds risky customers or defaulters through behavior patterns.
5. **Data security**: Helps in threat detection and keeps sensitive data safe.

LAQ

Q1. Create a system that uses Machine Learning to recommend personalized treatment plans for patients.

Ans:

1. **Data Collection:**
   - **Collect patient data such as age, gender, medical history, lifestyle, test reports, and past treatments.**
   - **Use hospital databases and wearable health devices for real-time data gathering.**
2. **Data Preprocessing:**
   - **Clean and organize the data by removing missing or incorrect values.**
   - **Convert patient records into numerical form so that ML algorithms can process them.**
3. **Model Selection:**
   - **Use algorithms like Decision Tree, Random Forest, or Neural Networks to analyze patient patterns.**
   - **The model predicts which treatment plan works best for similar patients.**
4. **Prediction and Recommendation:**
   - **The system compares the patient's data with past cases and suggests the most effective treatment plan.**
   - **Recommendations can include medicines, diet, or therapy schedules.**
5. **Continuous Learning:**
   - **As more patient data is added, the system learns and improves accuracy.**
   - **Helps doctors save time and offer faster, data-backed treatment decisions.**

Q2. Design a Machine Learning approach for predict risk analysis in banking system with data driven inside.

Ans:

1. **Data Gathering:**
   - **Collect data such as transaction history, account balances, credit scores, and customer details.**
   - **Include both normal and fraudulent transaction records.**
2. **Feature Engineering:**
   - **Identify key factors like spending habits, loan repayment patterns, and sudden large withdrawals.**
   - **Prepare these factors as input features for the ML model.**
3. **Model Building:**
   - **Use classification algorithms like Logistic Regression or Random Forest to predict risk levels.**
   - **The model assigns a "risk score" to each customer or transaction.**
4. **Prediction & Monitoring:**
   - **Predicts high-risk accounts that may default or perform fraudulent actions.**
   - **Alerts the bank for quick action.**
5. **Decision Making:**
   - **Helps banks plan better loan approvals, fraud prevention, and customer monitoring strategies using data-driven insights.**

Q3. Propose a Framework where Machine Learning helps in managing server downtimes in cloud based IT system.

Ans:

1. **Data Collection:**
   - **Collect server logs, CPU usage, memory load, and network traffic data continuously.**
   - **Include past downtime and failure data.**
2. **Preprocessing:**
   - **Filter and clean log data to remove unnecessary information.**
   - **Convert time-series data into readable numeric patterns.**
3. **Model Development:**
   - **Use ML algorithms like LSTM or Support Vector Machines to predict server performance.**
   - **The model learns patterns that usually happen before a server crash.**
4. **Prediction and Alert System:**
   - **The system sends automatic alerts when it predicts high chances of downtime.**
   - **Admins can take preventive steps before actual failure happens.**
5. **Self-healing Framework:**
   - **Integrate automation tools to restart or reallocate servers automatically.**
   - **Reduces downtime, saves costs, and increases cloud system reliability.**

Q4. Propose a solution to enhance customer retention in a Telecom Business using Machine Learning.

Ans:

1. **Data Collection:**
   - **Collect customer call logs, recharge patterns, complaint records, and service usage data.**
   - **Include customer feedback and churn history.**
2. **Feature Analysis:**
   - **Identify important factors like call drop rate, recharge frequency, and customer satisfaction score.**
   - **Convert them into numerical features for ML processing.**
3. **Model Training:**
   - **Use models like Logistic Regression or Decision Trees to predict which customers are likely to leave (churn).**
   - **The model classifies customers as “high-risk” or “safe”.**
4. **Action Plan:**
   - **For high-risk customers, provide personalized offers, discounts, or improved service quality.**
   - **Use chatbots and recommendation systems to engage customers.**
5. **Continuous Feedback Loop:**
   - **Keep updating the model with new data to improve prediction accuracy.**
   - **Increases retention rate and reduces customer loss over time.**

Q5. Create a pilot Machine Learning Project to detect failure in IT Infrastructure before they occur.

Ans:

1. **Data Gathering:**
   - **Collect system logs, hardware health reports, CPU temperature, and memory usage data.**
   - **Include both normal and failure cases.**
2. **Data Preprocessing:**
   - **Clean and label data as "Normal" or "Failure" to train the ML model.**
   - **Use time-stamped data for sequential pattern detection.**
3. **Model Development:**
   - **Apply algorithms like Random Forest or Anomaly Detection models to learn failure patterns.**
   - **Train the model to detect small irregularities that signal upcoming issues.**
4. **Prediction and Alert Mechanism:**
   - **Predicts failures in advance and triggers automatic alerts to system administrators.**
   - **Prevents costly downtime or damage.**
5. **Pilot Deployment:**
   - **Test the model on a small part of the IT network.**
   - **Once successful, expand to full infrastructure for predictive maintenance and improved performance.**

Q1. Analyze the workflow of a typical Machine Learning process from data collection to model evaluation.

Ans:

1. **Data Collection:**
   - **The first step is gathering relevant data from various sources such as databases, sensors, files, or online repositories.**
   - **Data quality and quantity directly affect the model's performance.**
2. **Data Preprocessing:**
   - **Raw data often contains missing values, duplicates, and noise.**
   - **Steps include cleaning, normalization, encoding categorical data, and splitting into training and testing sets.**
3. **Model Selection and Training:**
   - **Choose an appropriate algorithm based on the type of problem (e.g., regression, classification, clustering).**
   - **Train the model on the training dataset so it learns the relationship between input and output variables.**
4. **Model Testing and Evaluation:**
   - **Use the test dataset to evaluate how well the model performs on unseen data.**
   - **Calculate metrics like accuracy, precision, recall, and F1-score to check model performance.**
5. **Model Deployment and Monitoring:**
   - **Once validated, the model is deployed in real-world applications.**
   - **Regular monitoring ensures that it continues to perform well as new data arrives.**

Q2. Demonstrate how to preprocess any data step.

Ans:

1. **Data Cleaning:**
   - **Handle missing values by replacing them with mean, median, or mode.**
   - **Example: Replace missing salary values with the average salary.**
2. **Data Transformation:**
   - **Convert categorical values into numerical using Label Encoding or One-Hot Encoding.**
   - **Example: Convert "Male/Female" to 1/0.**
3. **Feature Scaling:**
   - **Normalize or standardize data so all features are in the same range (important for models like KNN or SVM).**
   - **Example: Convert values to a 0–1 range using Min-Max scaling.**
4. **Feature Selection:**
   - **Remove irrelevant or redundant features to improve accuracy and reduce overfitting.**
   - **Example: Drop unnecessary columns like "Customer_ID".**
5. **Data Splitting:**
   - **Divide the dataset into training (80%) and testing (20%) sets using train_test_split() in Python.**
   - **This helps evaluate how well the model generalizes to unseen data.**