Theoretical Questions

1. **What is a Decision Tree, and how does it work?**
   A Decision Tree is a supervised learning algorithm used for classification and regression. It works by recursively splitting the dataset based on feature conditions to create branches, ultimately leading to decision nodes (classification outputs).
2. **What are impurity measures in Decision Trees?**
   Impurity measures determine how mixed the classes are in a node. The main impurity measures are:
   - **Gini Impurity** (used in CART)
   - **Entropy** (used in ID3 and C4.5)
3. **What is the mathematical formula for Gini Impurity?**

   $$Gini = 1 - \sum p_i^2$$

   where pi is the probability of class i in the node.
4. **What is the mathematical formula for Entropy?**

   $$Entropy = -\sum p_i \log_2(p_i)$$

   where pi is the probability of class i in the node.
5. **What is Information Gain, and how is it used in Decision Trees?**
   Information Gain measures the reduction in entropy after a split. It is calculated as:

   $$IG = Entropy(parent) - \sum \left(\frac{|child|}{|parent|} \times Entropy(child) \right)$$

   The feature with the highest Information Gain is chosen for splitting.
6. **What is the difference between Gini Impurity and Entropy?**
   - Gini is computationally faster than Entropy.
   - Entropy is based on logarithms, leading to a more nuanced measurement of impurity.
   - Both yield similar results in practice.
7. **What is the mathematical explanation behind Decision Trees?**
   Decision Trees use recursive binary splitting to minimize impurity (Gini/Entropy). The algorithm selects the best split by calculating the weighted sum of child impurities and continues until a stopping criterion is met.
8. **What is Pre-Pruning in Decision Trees?**
   Pre-pruning stops the tree from growing based on a predefined condition (e.g., maximum depth, minimum samples per leaf) to prevent overfitting.

9. **What is Post-Pruning in Decision Trees?**
   Post-pruning first allows the tree to grow fully and then prunes back unnecessary branches by evaluating performance on validation data.

10. **What is the difference between Pre-Pruning and Post-Pruning?**

    - Pre-Pruning stops tree growth early.
    - Post-Pruning removes parts of a fully grown tree to improve generalization.

11. **What is a Decision Tree Regressor?**
    A Decision Tree Regressor predicts continuous values instead of discrete classes by minimizing variance instead of impurity.

12. **What are the advantages and disadvantages of Decision Trees?**
    Advantages:

    - Simple and interpretable
    - Handles both numerical and categorical data
    - Requires minimal data preprocessing

    Disadvantages:

    - Prone to overfitting
    - Sensitive to noisy data
    - Greedy splitting may not find the best global solution

13. **How does a Decision Tree handle missing values?**

    - It can use surrogate splits to handle missing values.
    - Some implementations allow missing values by assigning the most probable class.

14. **How does a Decision Tree handle categorical features?**

    - Categorical features are typically encoded using one-hot encoding or label encoding before training the model.
    - Some algorithms (e.g., C4.5) handle categorical features natively.

15. **What are some real-world applications of Decision Trees?**

    1) **Fraud Detection**

       a) Used by banks and financial institutions to detect fraudulent transactions by analyzing spending patterns and unusual activities.

    2) **Medical Diagnosis**

       a) Helps doctors classify diseases based on symptoms, test results, and patient history.

       b) Example: Diagnosing diabetes, heart disease, or cancer risk.

3) **Credit Scoring and Loan Approval**

   a) Used by banks to determine whether a customer is eligible for a loan based on credit history, income, and other financial factors.

4) **Customer Segmentation**

   a) Businesses use Decision Trees to segment customers based on behavior, demographics, and purchase history to target marketing campaigns effectively.

5) **Churn Prediction**

   a) Telecom and subscription-based businesses use Decision Trees to identify customers likely to cancel their subscriptions and take preventive actions.

6) **Predictive Maintenance**

   a) Manufacturing companies use Decision Trees to predict when a machine or equipment is likely to fail and schedule maintenance accordingly.

7) **Spam Filtering**

   a) Email providers use Decision Trees to classify emails as spam or non-spam based on keywords, sender reputation, and message patterns.

8) **Product Recommendation Systems**

   a) E-commerce platforms use Decision Trees to recommend products based on user preferences, browsing history, and purchase behavior.

9) **Sentiment Analysis**
   - Used in social media monitoring and customer reviews to classify texts as positive, negative, or neutral.

10) **Stock Market Analysis**

   - Investors and financial analysts use Decision Trees to predict stock price movements based on historical data, company performance, and market trends.

11. **Energy Consumption Forecasting**

   - Used by energy companies to predict electricity demand based on weather conditions, historical usage, and time of day.

12. **Supply Chain Optimization**

   - Helps logistics companies determine the best route for deliveries based on traffic patterns, weather, and order priorities.

13. **Human Resource Analytics**

   - Used for employee performance evaluation, attrition prediction, and talent acquisition decisions.