Decision Tree
Assignment Questions

Theoretical Questions

1 What is a Decision Tree, and how does it work ?

Ans. A **Decision Tree** is a supervised machine learning algorithm used for classification and regression tasks. It mimics human decision-making by splitting data into branches based on feature conditions.

How it Works:

1. **Root Node**: The tree starts with a root node, representing the entire dataset.
2. **Splitting**: Data is split based on the best feature (using metrics like Gini impurity or entropy for classification, and variance reduction for regression).
3. **Decision Nodes**: Each split creates child nodes, further refining the data.
4. **Leaf Nodes**: The process continues until a stopping criterion is met (e.g., max depth, minimum samples per node).
5. **Prediction**: New data follows the tree's path to reach a leaf node, determining its final class or value.

Decision Trees are easy to interpret but can overfit, requiring pruning or ensemble methods like Random Forest to improve performance.

2 What are impurity measures in Decision Trees?

Ans Impurity measures in Decision Trees help determine the best feature to split the data by evaluating how mixed the classes are in a node.

Common Impurity Measures:

1. Gini Impurity: Measures the probability of incorrectly classifying a randomly chosen element. Lower Gini means purer nodes.
   Formula:

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

2. Entropy: Based on information gain, it quantifies uncertainty in a node. Lower entropy means purer splits. Formula:

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

3. Variance Reduction: Used in regression trees, it minimizes variance within splits for better predictions.
4. Mean Squared Error (MSE): Another regression metric, reducing MSE improves model accuracy.

Decision Trees use these measures to find the best splits, balancing accuracy and complexity.

3What is the mathematical formula for Gini Impurity

Ans The Gini Impurity measures how often a randomly chosen element would be incorrectly classified if randomly labeled according to class distribution. It is used in Decision Trees to determine the best feature for splitting.

Mathematical Formula:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

Where:

- ccc = Total number of classes
- pip_ipi = Proportion of samples belonging to class iii

Explanation:

- If all samples belong to one class (pi=1p_i = 1pi=1), Gini = 0 (pure node).
- If classes are evenly split, Gini is higher, indicating more impurity.
- Decision Trees aim to split at nodes that minimize Gini Impurity, leading to purer child nodes.

4 What is the mathematical formula for Entropy ?

Ans The **Entropy** in Decision Trees measures the impurity or randomness in a dataset. It helps determine the best feature for splitting by evaluating the disorder within a node.

Mathematical Formula:

$$Entropy = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Where:

- ccc = Total number of classes
- pip_ipi = Proportion of samples in class iii

Explanation:

- **Low Entropy (≈0)**: If all samples belong to one class, the node is pure.
- **High Entropy (≈1)**: If classes are evenly distributed, the node is highly impure.
- Decision Trees aim to split nodes to reduce entropy, leading to better classification.

5 What is Information Gain, and how is it used in Decision Trees

Ans **What is Information Gain?**

Information Gain (IG) measures the reduction in uncertainty (or impurity) after splitting a dataset based on a feature. It helps Decision Trees choose the best feature for splitting.

Mathematical Formula:

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

Where:

- **Entropy(parent)** = Entropy before splitting
- **Entropy(child)** = Weighted sum of entropy after splitting
- **N** = Number of samples

How is it Used in Decision Trees?

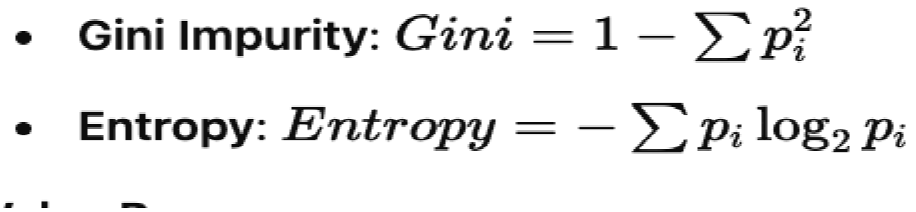
1. The tree calculates **Information Gain** for each feature.
2. The feature with the **highest IG** is chosen for splitting, reducing impurity.
3. The process repeats recursively until stopping criteria (e.g., max depth) are met.

A higher **Information Gain** leads to better decision-making, improving the tree's accuracy.

6 What is the difference between Gini Impurity and Entropy

Ans Difference Between Gini Impurity and Entropy

1. Definition:
    - ○ Gini Impurity measures the probability of misclassifying a randomly chosen sample.
    - ○ Entropy measures the level of disorder or impurity in a dataset based on information theory.
2. Formula:



- **Gini Impurity**: $Gini = 1 - \sum p_i^2$
- **Entropy**: $Entropy = - \sum p_i \log_2 p_i$

3. Value Range:
    - ○ Gini ranges from 0 (pure node) to a maximum of 0.5 (for two classes with equal probability).
    - ○ Entropy ranges from 0 (pure node) to 1 (for two classes with equal probability).
4. Computation Speed:
    - ○ Gini Impurity is faster as it does not involve logarithmic calculations.
    - ○ Entropy is slower because it involves logarithms.
5. Splitting Behavior:
    - ○ Gini tends to create larger class splits.

   - Entropy often results in more balanced splits.
6. Usage in Algorithms:

   - Gini Impurity is used in the CART (Classification and Regression Trees) algorithm.
   - Entropy is used in ID3, C4.5, and C5.0 algorithms.

8 What is Pre-Pruning in Decision Trees ?

Ans Mathematical Explanation of Decision Trees

Decision Trees split data using mathematical principles to minimize impurity and maximize information gain. The key concepts involved are impurity measures, splitting criteria, and recursive partitioning.

1. Impurity Measures

Decision Trees use impurity measures to determine how pure or impure a node is. The two most common measures are:

Gini Impurity

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

It measures the probability of misclassifying a randomly selected instance. Lower Gini values indicate purer nodes.

- Entropy:

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

- It measures the amount of uncertainty in a node. Lower entropy means a more homogeneous node.

2. Information Gain (IG)

To decide the best split, Decision Trees compute Information Gain, which measures the reduction in impurity after splitting:

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

The feature with the highest Information Gain is selected for splitting.

3. Splitting Criterion

At each step, the Decision Tree selects a feature X and a threshold t that best divides the dataset D into two subs

$$D_{left} = \{x \in D \mid X < t\}, \quad D_{right} = \{x \in D \mid X \geq t\}$$

\

The goal is to maximize Information Gain or minimize impurity.

4. Recursive Partitioning

- The tree starts at the root node (full dataset).
- It splits the data based on the best feature and threshold.
- This process repeats recursively until a stopping condition is met (e.g., max depth, minimum samples per node).

5. Stopping Criteria

To prevent overfitting, Decision Trees use:

- Max Depth: Stops splitting after a certain number of levels.
- Min Samples per Leaf: Ensures each leaf node has a minimum number of samples.
- Pruning: Removes unnecessary branches to simplify the tree.

Final Prediction

For classification, a new instance follows the tree's path until it reaches a leaf node, which assigns the most common class.
For regression, the prediction is the mean of target values in the leaf node.

Decision Trees are simple yet powerful models that use mathematical principles to make decisions efficiently.

9 What is Post-Pruning in Decision Trees ?

Ans Post-Pruning in Decision Trees

Post-Pruning is a technique used to reduce overfitting by trimming unnecessary branches after the tree is fully grown. It improves generalization by simplifying the model.

How It Works:

1. The Decision Tree is first grown completely without restrictions.
2. The algorithm evaluates nodes using a validation set or statistical measures (like cost-complexity pruning).
3. Nodes that do not improve accuracy significantly are removed (converted into leaf nodes).
4. The pruned tree is tested again to ensure better performance on unseen data.

Post-pruning makes the tree less complex, improves accuracy on new data, and reduces overfitting while maintaining interpretability.

10 What is the difference between Pre-Pruning and Post-Pruning

Difference Between Pre-Pruning and Post-Pruning

1. Definition:
    - Pre-Pruning stops tree growth early based on a predefined condition.
    - Post-Pruning first grows the tree fully and then removes unnecessary branches.
2. Execution Time:
    - Pre-Pruning occurs during tree construction.
    - Post-Pruning occurs after the tree is built.
3. Stopping Criteria:
    - Pre-Pruning stops splitting when conditions like max depth, minimum samples per node, or impurity threshold are met.

    - Post-Pruning removes branches based on validation set performance or statistical tests.

4. Risk of Underfitting/Overfitting:

    - Pre-Pruning may cause underfitting by stopping growth too early.
    - Post-Pruning reduces overfitting by simplifying a fully grown tree.

5. Performance:

    - Pre-Pruning is faster but may not always find the best tree.
    - Post-Pruning is slower but often results in a more optimal tree.

11 What is a Decision Tree Regressor ?

Ans A Decision Tree Regressor is a machine learning model used for regression tasks. Unlike classification trees, which predict discrete labels, a Decision Tree Regressor predicts continuous numerical values.

How It Works:

1. Splitting: The dataset is recursively split into subsets based on the feature that minimizes the variance in the target variable.
2. Impurity Measure: It uses Mean Squared Error (MSE) or Mean Absolute Error (MAE) to

    find the best split

$$MSE = \frac{1}{n}\sum(y_i - \bar{y})^2$$

3. :
4. Leaf Nodes: When further splitting does not significantly reduce variance, a leaf node is created. The average of target values in that node is used for prediction.
5. Prediction: For new data, it follows the tree's path and outputs the mean target value of the final leaf node.

Advantages:

- Handles non-linear relationships well.
- Interpretable and easy to visualize.

Limitations:

- Prone to overfitting, requiring pruning or ensemble methods (e.g., Random Forest).
- Sensitive to noisy data.

12 What are the advantages and disadvantages of Decision Trees

Ans Advantages and Disadvantages of Decision Trees

Advantages:

1. Easy to Understand & Interpret – Decision Trees are simple and intuitive, making them easy to visualize and explain.
2. Handles Both Numerical & Categorical Data – Can be used for classification and regression tasks.
3. Requires Little Data Preprocessing – No need for feature scaling or normalization.
4. Captures Non-Linear Relationships – Can model complex relationships without needing transformations.
5. Feature Selection Built-in – Automatically selects important features by splitting on the most relevant ones.
6. Works Well with Large Datasets – Efficient for handling large datasets with multiple features.

Disadvantages:

1. Prone to Overfitting – Can create very deep trees that fit training data too well, leading to poor generalization.
2. Sensitive to Noisy Data – Small changes in data can lead to different tree structures.
3. Greedy Algorithm – Chooses the best split at each step without considering the global optimal solution.
4. Biased with Imbalanced Data – Tends to favor dominant classes if not handled properly.
5. Limited Extrapolation in Regression – Decision Tree Regressors predict based on observed values, making them poor at extrapolating beyond training data.

To overcome these issues, pruning and ensemble methods (e.g., Random Forest, Gradient Boosting)

13 How does a Decision Tree handle missing values ?

Ans Decision Trees can handle missing values in several ways to maintain model accuracy and robustness.

1. Ignoring Missing Values During Splitting

- If a feature has missing values, the tree can only use available data for calculating impurity and determining the best split.

2. Assigning the Most Common Value (for Categorical Data)

- The missing value is replaced with the most frequent category in that feature.

3. Assigning the Mean/Median (for Numerical Data)

- The missing value is replaced with the mean or median of that feature.

4. Using Surrogate Splitting

- When a feature with missing values is chosen for a split, the tree finds an alternative (surrogate) feature highly correlated with the original feature to guide the split.

5. Using Probabilistic Assignment

- Missing values are assigned to branches based on probability, considering the distribution of available data.

14 How does a Decision Tree handle categorical features ?

Decision Trees can effectively handle categorical features using various techniques:

1. One-Hot Encoding (Binary Splitting)

- Each category is converted into a separate binary feature (0 or 1).
- Example: A "Color" feature with values {Red, Blue, Green} becomes three binary features: Color_Red, Color_Blue, Color_Green.

2. Label Encoding (Ordinal Representation)

- Each category is assigned a numerical value (e.g., Red = 1, Blue = 2, Green = 3).
- Works well for ordinal data but can introduce false ordinal relationships in nominal data.

3. Decision Tree’s Built-in Handling (For Some Algorithms)

- Some Decision Tree implementations (e.g., CART, C4.5) can directly handle categorical data by splitting based on category groups instead of numerical values.
- Example: If splitting on "Color," the tree may create branches like {Red, Blue} vs. {Green}.

4. Frequency or Target Encoding

- Replaces categories with their frequency in the dataset or the mean target value for regression tasks.

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

Ans Healthcare – Used for disease diagnosis, predicting patient risk levels, and recommending treatments based on symptoms and medical history.

1. Finance & Banking – Helps in credit scoring, fraud detection, loan approval, and assessing financial risks.

2. E-commerce & Retail – Used for customer segmentation, product recommendations, and predicting customer churn.

3. Manufacturing & Quality Control – Identifies defect patterns and optimizes production processes to minimize waste.

4. Human Resources – Helps in employee attrition prediction, recruitment decision-making, and performance evaluation.

5. Marketing & Advertising – Optimizes targeted advertising, customer response predictions, and campaign effectiveness analysis.

6. Education – Used for student performance prediction, dropout risk assessment, and personalized learning recommendations.

7. Telecommunications – Helps in network fault detection, customer churn prediction, and optimizing service plans.

8. Energy Sector – Predicts energy consumption trends, detects faults in power grids, and optimizes energy distribution.

9. Fraud Detection & Cybersecurity – Identifies suspicious transactions, malware detection, and network security monitoring.

Decision Trees are widely used due to their interpretability, efficiency, and ability to handle complex decision-making in various industries.