Decision Tree:

Assignment Questions

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

- A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks.
- It works by recursively partitioning the dataset into subsets based on feature values.
- The structure resembles an inverted tree, with:
    - **Nodes:** Representing tests on features.
    - **Branches:** Representing the outcome of the tests.
    - **Leaves:** Representing the predicted outcome (class or value).
- The algorithm selects the best feature to split on at each node, aiming to create subsets that are as "pure" as possible with respect to the target variable.

2. What are impurity measures in Decision Trees?

- Impurity measures quantify the degree of "mixedness" of classes within a subset of data.
- They help determine the best feature to split on by measuring how much the split reduces impurity.
- Common impurity measures include Gini impurity and Entropy.

3. What is the mathematical formula for Gini Impurity?

- For a dataset with *C* classes, the Gini impurity is calculated as:
    - `Gini = 1 - Σ (p_i)^2`
    - Where `p_i` is the proportion of samples belonging to class *i*.

4. What is the mathematical formula for Entropy?

- Entropy measures the disorder or uncertainty in a dataset.
- For a dataset with *C* classes, it's calculated as:
    - `Entropy = - Σ p_i * log2(p_i)`
    - Where `p_i` is the proportion of samples belonging to class *i*.

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

- Information Gain measures the reduction in entropy (or Gini impurity) achieved by splitting a dataset on a particular feature.
- It's calculated as:
    - `Information Gain = Entropy(parent) - Σ [ ( |child| / |parent| ) * Entropy(child) ]`
    - Where `|parent|` and `|child|` are the sizes of the parent and child nodes, respectively.
- Decision Trees use Information Gain to select the feature that maximizes the reduction in impurity, leading to the most informative splits.

6. What is the difference between Gini Impurity and Entropy?

- Both Gini impurity and Entropy measure impurity, but they differ in their mathematical formulation and sensitivity.
- Entropy uses logarithms, making it slightly more computationally expensive.
- Gini impurity is generally faster to compute and is often the default choice.
- In practice, the results produced by decision trees using either Gini impurity or Entropy are often very similar.

7. What is the mathematical explanation behind Decision Trees?

- At each node, the algorithm aims to find the feature and threshold that optimally split the data.
- This optimization is done by maximizing Information Gain or minimizing impurity.
- The process is recursive, meaning it's repeated for each subset until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf).
- The final result is a set of rules, that can be expressed in if then statements, that classify or predict the value of new data points.

8. What is Pre-Pruning in Decision Trees?

- Pre-pruning involves stopping the tree's growth early, before it fully fits the training data.
- This is done by setting constraints on the tree's parameters, such as:
    - Maximum depth.
    - Minimum samples per leaf.
    - Minimum samples per split.
- It aims to prevent overfitting by creating simpler trees.

9. What is Post-Pruning in Decision Trees?

- Post-pruning involves growing the tree fully and then removing branches that do not improve performance on a validation set.
- Techniques like cost-complexity pruning are used to identify and remove less informative branches.
- It aims to simplify the tree and improve its generalization ability.

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

- **Pre-pruning:** Stops the tree's growth early, based on predefined criteria.

- **Post-pruning:** Grows the tree fully and then removes branches based on performance on a validation set.
- Pre-pruning is faster, but may underfit. Post-pruning is more computationally expensive, but tends to produce more accurate trees.

11. What is a Decision Tree Regressor?

- A Decision Tree Regressor is a variant of the decision tree algorithm used for regression tasks (predicting continuous values).
- Instead of classifying data into categories, it predicts a numerical value at each leaf node.
- The prediction is typically the average or median of the target values in the leaf.
- The impurity measure used is usually Mean Squared Error (MSE) or Mean Absolute Error (MAE).

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

- **Advantages:**
    - Easy to understand and interpret.
    - Can handle both numerical and categorical data.
    - Requires minimal data preprocessing.
    - Can handle non-linear relationships.
    - Relatively fast to train and predict.
- **Disadvantages:**
    - Prone to overfitting.
    - Can be sensitive to small variations in the data.
    - Can create biased trees if some classes dominate.
    - Not always the most accurate algorithm.

13. How does a Decision Tree handle missing values?

- Decision trees can handle missing values in several ways:
    - **Surrogate splits:** Use other features to approximate the split when the primary feature is missing.
    - **Imputation:** Fill in missing values with estimated values (e.g., mean, median, mode).
    - Some implementations will send the missing value down all branches, and weight the final result based on the probability of each branch.
    - Some implementations will treat missing values as their own category.

14. How does a Decision Tree handle categorical features?

- Decision trees can handle categorical features directly.
- For binary categorical features, the split is straightforward (e.g., "yes" or "no").
- For multi-category features, the algorithm can create splits based on subsets of the categories.
- Some implementations will use one hot encoding, prior to creating the tree.

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

- **Medical diagnosis:** Predicting patient risk or diagnosing diseases.
- **Financial risk assessment:** Evaluating creditworthiness or detecting fraud.
- **Customer churn prediction:** Identifying customers likely to cancel subscriptions.
- **Image recognition:** Classifying objects in images.
- **Recommendation systems:** Suggesting products or content.
- **Manufacturing quality control:** Detecting defects in products.
- **Gameplay AI:** Creating decision making logic for game AI.