**Name: Sumit Avinash Solanke**

**Roll No: TYITB121**

**Assignment No. 3**

**AIM:** Assignment on Decision Trees

**PREREQUISITE:** Python programming

**THEORY:**

Decision trees are a fundamental tool in machine learning, widely used for classification and regression tasks. They enable decision-making by structuring data into a tree-like model, where each internal node represents a decision based on an attribute, each branch represents an outcome of that decision, and each leaf node signifies a final classification or prediction.

**Machine Learning: Supervised Learning**

Supervised learning involves training a model on labeled data, meaning that the input-output pairs are known in advance. This allows the model to learn from the data and make predictions about new, unseen inputs. Within supervised learning, there are two key tasks:

* **Regression:** Used when the output is a continuous variable.
* **Classification:** Used when the output belongs to a discrete category.

Decision trees are particularly useful for classification problems. By analyzing different attributes in the dataset, they determine the best splits to categorize data into distinct classes.

**How Decision Trees Work**

A decision tree operates by:

1. **Identifying Attributes:** Evaluating different features in the dataset.
2. **Splitting Data:** Finding the best attributes to divide the data.
3. **Creating Decision Nodes:** Determining logical steps in the decision process.
4. **Classifying Data:** Assigning final outcomes at the leaf nodes.

A common example involves classifying fruits based on features like size and texture. Given enough labeled data, the decision tree can determine whether an input belongs to a certain category, such as an apple or an orange.

**Types of Decision Tree Algorithms**

Several algorithms exist to construct decision trees, including:

* **ID3 (Iterative Dichotomiser 3):** Uses entropy and information gain to select the best attribute for splitting.
* **C4.5:** An extension of ID3, it introduces gain ratios and handles continuous data better.
* **CART (Classification and Regression Trees):** Utilizes the Gini impurity criterion for splitting nodes.

**Choosing the Best Attribute for Splitting**

To construct an optimal decision tree, it is essential to determine which attribute provides the most informative split. Two popular methods used are:

* **Information Gain:** Measures how well a particular attribute separates the data into different categories.
* **Gini Impurity:** Evaluates the likelihood of misclassification if an attribute is used for splitting.

The attribute that maximizes information gain or minimizes Gini impurity is selected at each decision node.

**Advantages and Disadvantages of Decision Trees**

**Advantages:**

* **Easy to Understand:** The tree-like structure makes it intuitive and interpretable.
* **Handles Missing Data:** Decision trees can handle datasets with missing values.
* **Works with Various Data Types:** Can process numerical and categorical data.
* **Versatile:** Used for both classification and regression problems.

**Disadvantages:**

* **Prone to Overfitting:** Complex trees may fit the training data too well, reducing generalization to new data.
* **Sensitive to Small Changes:** Slight variations in data can lead to drastically different tree structures.
* **Computational Cost:** Finding the best attribute splits requires significant computational resources.

**Dataset Description: Iris Dataset**

**Dataset Name:** *Iris Dataset*  
**Source:** Introduced by Sir Ronald A. Fisher in 1936, this dataset is widely used for classification and clustering tasks in machine learning.

**Overview:**

The Iris dataset contains **150 records** of iris flowers, with **four numeric features** and **one categorical target class**. These records are equally divided among **three species** of iris flowers:

* *Iris setosa*
* *Iris versicolor*
* *Iris virginica*

Each row in the dataset represents the measurements of one flower.

**Features and Descriptions:**

| **Feature Name** | **Data Type** | **Unit** | **Description** |
| --- | --- | --- | --- |
| sepal\_length | Float | Centimeters (cm) | Length of the sepal |
| sepal\_width | Float | Centimeters (cm) | Width of the sepal |
| petal\_length | Float | Centimeters (cm) | Length of the petal |
| petal\_width | Float | Centimeters (cm) | Width of the petal |
| species (target) | Categorical | N/A | Species of the iris flower (setosa, versicolor, or virginica) |

**Sample Records:**

| **sepal\_length** | **sepal\_width** | **petal\_length** | **petal\_width** | **species** |
| --- | --- | --- | --- | --- |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 7.0 | 3.2 | 4.7 | 1.4 | versicolor |
| 6.3 | 3.3 | 6.0 | 2.5 | virginica |

**Dataset Characteristics:**

* **Total Instances:** 150
* **Classes:** 3 (50 samples per class)
* **Balanced:** Yes
* **Missing Values:** None
* **Type:** Multivariate, numerical features with a categorical target
* **Use Cases:** Classification, clustering, dimensionality reduction (e.g., PCA)

**Conclusion:**

Decision trees are a powerful tool in machine learning, offering an intuitive way to model data for classification and regression tasks. While they have advantages in interpretability and flexibility, they also present challenges like overfitting and sensitivity to data changes. By understanding and optimizing their construction using algorithms such as ID3, C4.5, and CART, decision trees can provide valuable insights for decision-making in various applications.