Name: Sumit Avinash Solanke  
Roll No: TYITB121  
Assignment No.: 4  
AIM: Assignment on Decision Tree on cancer dataset  
PREREQUISITE: Python programming

THEORY:

The Decision Tree classification technique is a supervised learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure, making it easy to understand and interpret. It works by splitting the dataset into subsets based on the value of input features, resulting in a tree of decision nodes and leaf nodes.

Decision Trees are widely used in various fields due to their simplicity and ability to handle both numerical and categorical data. They are useful for extracting meaningful patterns from data and are often used in decision analysis, customer segmentation, and predictive modeling.

Concept of Decision Tree Classification

To understand Decision Tree classification, imagine a process of making decisions based on a sequence of questions. Each internal node in the tree represents a test on an attribute (feature), each branch represents the outcome of that test, and each leaf node represents a class label (decision).

For example, in a weather dataset, a decision tree might first split the data based on whether it is raining or not. Then, it may split further based on temperature or humidity to finally decide whether a person will play tennis or not. The tree continues to split the data until it reaches a decision or cannot be split further.

The tree grows by selecting features that best split the data using measures like Gini Index, Entropy, or Information Gain.

Working Mechanism of Decision Tree

The Decision Tree works by recursively partitioning the data into smaller subsets:

1. Select Best Feature to Split
   * Choose the feature that results in the best data split using metrics like Information Gain or Gini Index.
2. Split the Dataset
   * Based on the selected feature, divide the data into subsets.
3. Repeat Recursively
   * Repeat the process for each subset until a stopping condition is met (e.g., all records belong to one class or no more features are left).
4. Build Final Tree
   * The result is a tree with decision nodes and leaves representing the classification outcome.

Applications of Decision Tree

Decision Trees are used in many real-world applications, including:

* Customer Segmentation: Identify customer groups for targeted marketing.
* Credit Risk Assessment: Evaluate loan applicants based on financial attributes.
* Medical Diagnosis: Determine the presence of a disease based on symptoms.
* Business Decision Making: Support strategic decisions using historical data.

Advantages of Decision Tree

* Easy to Understand: Tree structures are simple and intuitive to interpret.
* Handles Both Types of Data: Can work with both categorical and numerical data.
* No Need for Data Normalization: Does not require feature scaling.
* Feature Importance: Helps identify the most significant features for prediction.

Disadvantages of Decision Tree

* Overfitting: Trees can become too complex and fit noise in the data.
* Unstable: Small changes in data can result in a completely different tree.
* Biased Towards Dominant Classes: May favor features with more levels.
* Less Accurate Alone: Often outperformed by ensemble methods like Random Forest.

**Data Description:**

**Overview:**

This dataset contains features computed from digitized images of breast mass fine needle aspirates (FNA). The features describe characteristics of the cell nuclei present in the image.

**Target Variable:**

| **Value** | **Meaning** |
| --- | --- |
| 0 | Malignant |
| 1 | Benign |

**Dataset Characteristics:**

| **Attribute** | **Value** |
| --- | --- |
| **Samples** | 569 |
| **Features** | 30 numeric features |
| **Classes** | 2 (Malignant, Benign) |
| **Missing Values** | None |
| **Data Type** | Numerical |
| **Balanced Classes** | Slightly Imbalanced (212 Malignant, 357 Benign) |

**Feature Categories:**

Each sample has **30 numeric features**, grouped into three types for each of 10 measurements:

1. **Mean** (average of values)
2. **Standard error** (SE)
3. **"Worst"** or largest (mean of the three largest values)

**Measurements:**

| **Measurement** | **Description** |
| --- | --- |
| radius | Distance from center to points on the perimeter |
| texture | Standard deviation of gray-scale values |
| perimeter | Perimeter of the mass |
| area | Area of the mass |
| smoothness | Local variation in radius lengths |
| compactness | Perimeter² / Area - 1.0 |
| concavity | Severity of concave portions of the contour |
| concave points | Number of concave portions of the contour |
| symmetry | Symmetry of the mass |
| fractal dimension | "Roughness" of the contour |

Each of these 10 measurements has:

* mean (e.g., radius\_mean)
* standard error (e.g., radius\_se)
* worst (e.g., radius\_worst)

So in total: 10 × 3 = **30 features**

**Sample Features (first 5 columns):**

| **radius\_mean** | **texture\_mean** | **perimeter\_mean** | **area\_mean** | **smoothness\_mean** |  |
| --- | --- | --- | --- | --- | --- |
| 17.99 | 10.38 | 122.8 | 1001.0 | 0.1184 |  |

**Usage:**

This dataset is ideal for:

* Binary classification tasks
* Model testing (e.g., Logistic Regression, SVM, Random Forest, etc.)
* Feature importance analysis
* ROC curve and AUC evaluation

CONCLUSION:

Decision Tree is a versatile and powerful algorithm that mimics human decision-making. Its clear structure makes it a great tool for understanding and visualizing the decision process. However, it may overfit and be sensitive to small data changes. Despite its limitations, it remains one of the most widely used algorithms in machine learning for classification and regression tasks.