**Name: Abhishek Rajaram Walkoli**

**PRN: 123B2F155**

**Assignment No 4**

**AIM:** Assignment on Decision Tree.

**PREREQUISITE:**

1.Python environment (Jupyter Notebook/IDE).

2.Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn.

3.Fundamental knowledge of machine learning and classification.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**THEORY:**

The role of an information system is to sift useful information out of raw data. Data science is the inter-disciplinary fusion of machine learning, big data, and statistics to examine and interpret data. Machine learning provides decision-making with little or no human intervention through learning from past data.

**Machine Learning: Supervised Learning**

Supervised learning is a machine learning algorithm in which an algorithm learns a function (f) to map input (X) to output (y) by leveraging labeled datasets. It involves two principal tasks:

● **Regression:** Regresses continuous numerical values.

● **Classification:** Predicts categorical values (e.g., spam email vs. non-spam email).

**Regression vs. Classification**

**● Regression:** Output is continuous (e.g., forecasting house prices).

● **Classification:** Output is categorical (e.g., classifying fruit as an apple or an orange).

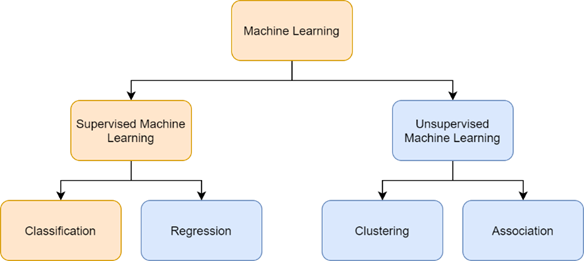
**Decision Trees**

A Decision Tree is a supervised learning model employed for classification as well as regression. It is a tree-like structure in which:

●Internal nodes are tests on attributes.

●Branches are results of the tests.

●Leaf nodes are class labels.



**How It Works:**

1. 1.The algorithm chooses the best attribute to split the dataset.
2. 2.It continues dividing the dataset into subsets using recursive partitioning.
3. 3.The algorithm continues until all the samples in one subset are of the same class.
4. 4.The final tree structure is utilized for classification or prediction.

**Types of Decision Trees**

**● ID3 (Iterative Dichotomiser 3):** Employ entropy and information gain for splits.

**● C4.5:** ID3 improvement that takes into account gain ratios.

**● CART (Classification and Regression Trees):** Employ Gini impurity for splits selection.

**Selecting the Best Attribute to Split**

Decision trees employ various criteria for the best attribute to split:

**1. Entropy & Information Gain:**

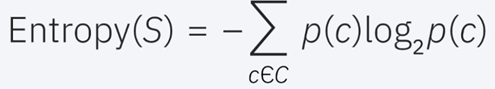
○ Entropy measures the dataset's impurity.

○Information gain calculates the reduction in entropy after a split.

○The attribute with the highest information gain is chosen for splitting.

**2.Formula for Entropy:**

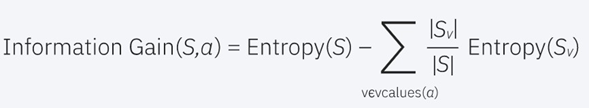
Entropy(S)=−∑p(c)log⁡2p(c)Entropy(S) = -\sum p(c) \log\_2 p(c)



**Formula for Information Gain:**

IG(S,a)=Entropy(S)−∑∣Sv∣∣S∣Entropy(Sv)IG(S, a) = Entropy(S) - \sum \frac{|S\_v|}{|S|} Entropy(S\_v)

**3.Gini Impurity:**

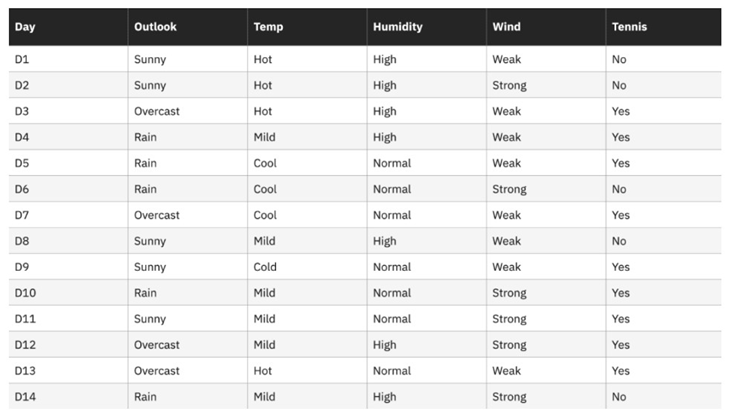
****

○Measures the probability of misclassification.

○Lower Gini impurity indicates a better split.

**4.Formula for Gini Impurity:**

Gini(S)=1−∑p(c)2Gini(S) = 1 - \sum p(c)^2



**Advantages and Disadvantages of Decision Trees**

**Advantages:**

● Easy to understand and visualize.

● Needs minimal data preprocessing.

● Supports both numerical and categorical data.

**Disadvantages:**

● Susceptible to overfitting, particularly with deep trees.

● Sensitive to minor changes in data.

● Training can be computationally costly.

CONCLUSION:



The Decision Tree classifier effectively proves its capability to:

● Learn from data by recursively splitting features based on Gini Impurity and Information Gain.

● Visualize the decision-making process through an interpretable tree structure.

● Generate new outcomes with reasonable accuracy, validated using performance metrics such as precision, recall, and confusion matrix.