Practical No.8

Title: Write a program to demonstrate the working of the decision tree based ID3 algorithm. Use an appropriate data set for building the decision tree and apply this knowledge to classify a new sample

Objective

The objective of this practical is to implement a **Decision Tree Classifier** using the **ID3 (Iterative Dichotomiser 3) algorithm**. The model will be trained on a dataset, and a new sample will be classified based on the constructed decision tree.

Introduction to Decision Trees and the ID3 Algorithm

A **Decision Tree** is a supervised learning algorithm used for classification and regression tasks. It is structured like a flowchart, where each internal node represents a **decision based on an attribute**, each branch represents an **outcome**, and each leaf node represents a **class label**.

ID3 Algorithm Overview

The **ID3 algorithm** builds a decision tree using a **top-down, recursive, and greedy approach** by selecting the best attribute at each node using **information gain**.

Key Concepts of ID3 Algorithm

1. **Entropy (HHH)**
    - Measures the impurity of a dataset.
    - Given by the formula:

H(S)=−∑pilog⁡2piH(S) = -\sum p_i \log_2 p_iH(S)=−∑pilog2pi

where pip_ipi is the proportion of class iii in the dataset.

    - If entropy = 0, all samples belong to the same class (pure dataset).
2. **Information Gain (IGIGIG)**
    - Measures the reduction in entropy after splitting the dataset on an attribute.
    - Given by:

IG(S,A)=H(S)−∑|Sv||S|H(Sv)IG(S, A) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)IG(S,A)=H(S)−∑|S||Sv|H(Sv)

where SvS_vSv represents subsets of SSS after splitting on attribute AAA.

- The attribute with the **highest information gain** is chosen as the **splitting criterion**.

Why Use ID3?

- ID3 is simple and efficient for small datasets.
- It **automatically selects the best attributes** for splitting.
- The resulting decision tree is **interpretable** and easy to visualize.

Dataset Description

An **appropriate dataset** is chosen for building the decision tree. The dataset consists of:

- **Multiple attributes (independent variables)** used for classification.
- **A categorical target variable (dependent variable)** representing different class labels.
- The dataset should contain **discrete categorical values**, as ID3 does not handle continuous values directly (preprocessing may be required).

Implementation Steps

Step 1: Load and Explore the Dataset

- Read the dataset from a **CSV file**.
- Display dataset information, including feature names and class distribution.
- Check for and handle missing values.

Step 2: Data Preprocessing

- Convert numerical values to categorical if needed.
- Encode categorical variables if required for processing.
- Split the dataset into **training (80%)** and **testing (20%)** sets.

Step 3: Build the Decision Tree Using ID3 Algorithm

- Compute **entropy** of the dataset.
- Calculate **information gain** for each attribute.
- Select the attribute with the **highest information gain** as the root node.
- Recursively repeat the process for each subset until a stopping condition is met:
    - If all instances belong to the same class, stop splitting.
    - If no attributes remain, assign the most common class label.

Step 4: Visualize the Decision Tree

- Display the generated decision tree structure.
- Use a graphical representation if possible.

Step 5: Classify a New Sample

- Provide a new sample with attribute values.
- Traverse the decision tree from root to leaf based on attribute conditions.
- Assign a class label to the new sample.

Step 6: Compute Model Accuracy

The accuracy of the model is computed by testing it on the test dataset:

Accuracy=Correct PredictionsTotal Predictions×100Accuracy = \frac{Correct\ Predictions}{Total\Predictions}\times 100Accuracy=Total PredictionsCorrect Predictions×100

Additionally, a **confusion matrix** is used to analyze correct and incorrect classifications.

Expected Output

- **Constructed Decision Tree** in textual or graphical format.
- **Predicted class labels** for test samples.
- **Classification result for a new sample** based on the decision tree.
- **Accuracy score and confusion matrix** for model evaluation.

Conclusion

This practical demonstrates the implementation of the **ID3 Decision Tree Algorithm** for classification. The model was trained, tested, and used to classify a new sample.