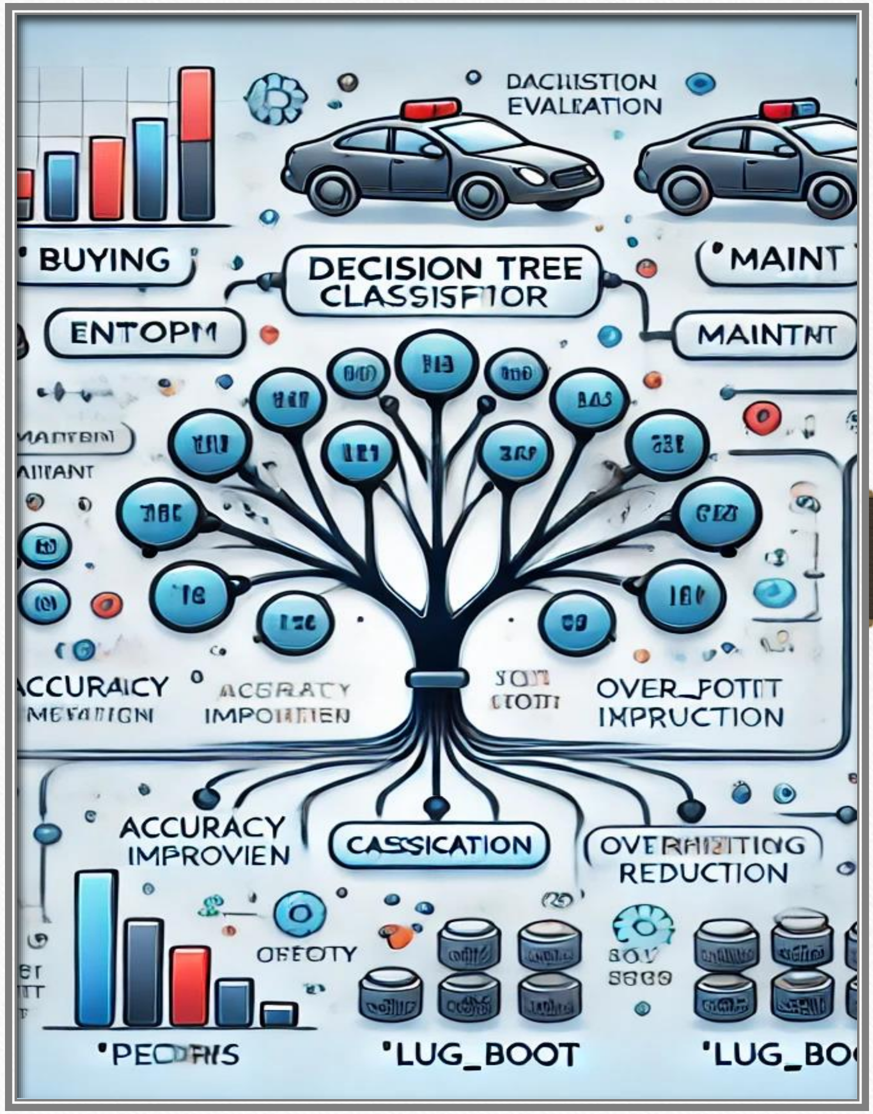
Decision Tree Classifier

Implementing a Decision Tree Classifier with Entropy-based Node Splitting and Pruning on the Kaggle Car Evaluation Dataset Using Tree Data Structures in C++

Group : 3

Mentor :- Dr. Dip Sankar Banerjee

Group Member Name :-
Shiv Jee Yadav [B23EE1095]
Sidharth Kumar Bhagat [B23EE1097]
Anshuman Parida [B23ES1008]
Sai Jagdeesh[B23EE1104]



What is a decision tree?

- A **decision tree** is a machine learning model used for classification and regression tasks. It splits data into subsets based on feature values, creating a tree-like structure of decisions. Each internal node represents a feature, and each branch represents a decision rule. The leaves of the tree show the final output (e.g., class label or predicted value).
- Why Decision trees?

  ->Easy to understand and interpret.
  ->Non-linear relationships.
  ->Handles both continuous and categorical data.

Function Used in Project

1. **Build Tree**
- **Purpose**: Recursively builds a decision tree based on information gain.

2. P**rint Tree**
- **Purpose**: Prints the decision tree in a structured format.

3. V**erify Prediction**
- **Purpose**: Verifies if the predictions match the actual values in the test data.

4. **Count Unique Attributes**
- **Purpose**: Counts unique values in each attribute of the dataset.

5. **Calculate Information_Gain**
- **Purpose**: Calculates the information gain for each attribute.

6. **ID3 And C-45 Algorithm**
- **Purpose**: Two Different Algorithm to increase Test Accuracy

ID3 Implementation-

Steps:

1.Calculate information gain for each attribute.

2.Calculate the entropy of the dataset.

3.Select the Recursive Call: Build child nodes using the same process.

4.Stopping Criteria: If all data points belong to the same class or no attributes remain.

C4.5 Implementation-

Steps:

1.Calculate the entropy for the dataset.

2.Calculate gain ratio instead of pure information gain.

3.Handle continuous attributes by finding an optimal threshold for splitting.

4.Prune branches to reduce overfitting.

Theoretical background

Entropy:

- Entropy is a measure of uncertainty or disorder in a set of data.
- Formula:

$$H(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Where pi is the probability of class i.

Information Gain:

measures the effectiveness of an attribute in classifying data.

Formula :

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \cdot \text{Entropy}(S_v)$$

Gini Index:

measure of impurity or purity used for splitting .

Formula:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{m} p_i^2$$

Contribution

All team members contributed equally and efficiently, working collaboratively to ensure the project's success.

1.Shiv Jee Yadav :**Build Tree , Print Tree , Count Unique Attributes , Information_Gain**

2.Sidharth Kumar Bhagat : **Verify Prediction , predict, input_dtc, statistical_error, Print Tree**

3.Anshuman Parida : **C-45 , Build Tree .**

4.Sai Jagdeesh : **Count Unique Attributes , Entropy .**