INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Department of Computer Science and Engineering

CS60050 – Machine Learning

Project - 1

Decision Tree Implementation

Submitted By: (GROUP – 15)

- **Kamal Kyal - 19CE10035**
- **Rishi Suman - 19EC39045**
- **Shailesh Chaudhary - 22CS60R37**

Introduction

- Decision Trees are powerful machine learning algorithms capable of performing regression and classification tasks. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

- Entropy measures the randomness or disorders in a system. In terms of data, we can define it as the randomness in the information we are processing. The higher the randomness, the higher the entropy. Hence, harder to conclude from that information.

- Information gain (IG) measures the amount of information provided by a given feature or attribute about a particular target class. While creating a decision tree, our goal is to find the attribute having the highest Information Gain, and conversely, the lowest entropy. Mathematically, it is calculated as the difference of the initial and final entropy.

Algorithm

- While training, our decision tree model evaluates all possible splits across all possible columns and picks the split with the highest information gain.

- With the first split, all the data according to a specific condition falls towards either the left or the right of the root node.

- Now, for each side of training data under the root node, all possible splits are calculated again and the split with the highest I.G is chosen. The process repeats for both left and right sides till we reach the terminating nodes representing a class in the target column.

- This way, our decision tree grows iteratively, layer by layer.

Results and Classification Reports

One of the performance assessment measures for a classification-based machine learning model is the Classification Report. The precision, recall, F1 score, and support of your model are shown. It gives us a clearer picture of our trained model's overall performance.

Metrics	Definition
Precision	Precision is defined as the ratio of true positives to the sum of true and false positives.
Recall	Recall is defined as the ratio of true positives to the sum of true positives and false negatives.
F1 Score	The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

Metrics	Definition
Support	Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models; it just diagnoses the performance evaluation process.

Classification Reports (Without Pruning)

1. First Fold Validation (Our Model):

Class	Precision	Recall	F1-Score	Support
US	0.87	0.87	0.87	31
Japan	0.58	0.58	0.58	12
Europe	0.60	0.60	0.60	10

Accuracy of the Model = 75%

First Fold Validation (Scikit-learn):

Class	Precision	Recall	F1-Score	Support
US	0.91	0.91	0.91	35
Japan	0.33	0.60	0.43	5
Europe	0.71	0.45	0.56	11

Accuracy of the Model = 78%

2. Second Fold Validation (Our Model):

Class	Precision	Recall	F1-Score	Support
US	0.87	0.93	0.90	28
Japan	0.64	0.64	0.64	11
Europe	0.91	0.77	0.83	13

Accuracy of the Model = 83%

Second Fold Validation (Scikit-learn):

Class	Precision	Recall	F1-Score	Support
US	0.93	0.93	0.93	30
Japan	0.57	0.31	0.40	13

Europe	0.43	0.75	0.55	8

Accuracy of the Model = 75%

3. Third Fold Validation (Our Model):

Class	Precision	Recall	F1-Score	Support
US	0.97	0.87	0.92	39
Japan	0.57	0.44	0.50	9
Europe	0.30	0.75	0.43	4

Accuracy of the Model = 79%

Third Fold Validation (Scikit-learn):

Class	Precision	Recall	F1-Score	Support
US	0.79	0.92	0.85	24
Japan	1.00	0.53	0.69	17
Europe	0.64	0.90	0.75	10

Accuracy of the Model = 78%

4. Fourth Fold Validation (Our Model):

Class	Precision	Recall	F1-Score	Support
US	0.87	0.93	0.90	29
Japan	0.62	0.56	0.59	9
Europe	0.85	0.79	0.81	14

Accuracy of the Model = 83%

Fourth Fold Validation (Scikit-learn):

Class	Precision	Recall	F1-Score	Support
US	0.94	0.94	0.94	32
Japan	0.71	0.50	0.59	10
Europe	0.50	0.67	0.57	9

Accuracy of the Model = 80%

5. Fifth Fold Validation (Our Model):

Class	Precision	Recall	F1-Score	Support
US	0.91	0.91	0.91	35
Japan	0.62	0.71	0.67	7
Europe	0.78	0.70	0.74	10

Accuracy of the Model = 85%

Fifth Fold Validation (Scikit-learn):

Class	Precision	Recall	F1-Score	Support
US	0.95	1.00	0.97	36
Japan	0.71	0.83	0.77	6
Europe	0.83	0.56	0.67	9

Accuracy of the Model = 90%

6. Average of 5-Folds Validation (Our Model):

Class	Precision	Recall	F1-Score	Support
US	0.898	0.902	0.90	32
Japan	0.606	0.586	0.596	10
Europe	0.688	0.722	0.682	10

Average Accuracy of the Model = 81%

Average of 5-Folds Validation (Scikit-learn):

Class	Precision	Recall	F1-Score	Support
US	0.904	0.94	0.92	31
Japan	0.664	0.554	0.576	10
Europe	0.60	0.666	0.62	9

Average Accuracy of the Model = 80.2%

Classification Reports (With Pruning)

1. Pruning at Depth = 5

Class	Precision	Recall	F1-Score	Support
US	0.93	0.87	0.90	31
Japan	0.58	0.58	0.58	12
Europe	0.42	0.50	0.45	10

Accuracy of the Model = 74%

2. Pruning at Depth = 10

Class	Precision	Recall	F1-Score	Support
US	0.87	0.87	0.87	31
Japan	0.58	0.58	0.58	12
Europe	0.60	0.60	0.60	10

Accuracy of the Model = 75%

3. Pruning at Depth = 15

Class	Precision	Recall	F1-Score	Support
US	0.87	0.87	0.87	31
Japan	0.58	0.58	0.58	12
Europe	0.60	0.60	0.60	10

Accuracy of the Model = 75%

4. Pruning at Depth = 20

Class	Precision	Recall	F1-Score	Support
US	0.87	0.87	0.87	31
Japan	0.58	0.58	0.58	12
Europe	0.60	0.60	0.60	10

Accuracy of the Model = 75%

5. Pruning at Depth = 25

Class	Precision	Recall	F1-Score	Support
US	0.87	0.87	0.87	31
Japan	0.58	0.58	0.58	12
Europe	0.60	0.60	0.60	10

Accuracy of the Model = 75%