REPORT

DECISION TREE

BY:

CHANDNI SHANKAR

CXS152830

Table of Contents

1. **Implementation Summary 3**
2. **Output 3**

2.1 Dataset 13

2.2 Dataset 24

1. **Conclusion5**
2. **IMPLEMENTATION SUMMARY**

Implemented the decision tree algorithm. For the given two datasets, built the decision tree using the training set and pruned the tree using the validation set to get a better performing and smaller tree if there exists one. The pruning was implemented using the post-pruning algorithm with given values of l and k. The accuracy of the decision tree was tested using the test data before pruning and after pruning the tree.

Input: Training set, Validation set, Test set and values of L and K for pruning the data.

Output: Accuracy of the built decision tree for the test set before pruning and after pruning

1. **OUTPUT**
   1. DATASET 1:

Before pruning:

Accuracy using Information gain heuristic: 75.7621189405

Accuracy using Variance Impurity heuristic: 59.1704147926

* + 1. L = 100 K = 20

Accuracy using Information gain heuristic: 77.8610694653

Accuracy using Variance Impurity heuristic: 65.0174912544

* + 1. L = 20 K = 10

Accuracy using Information gain heuristic: 77.7611194403

Accuracy using Variance Impurity heuristic: 63.0184907546

* + 1. L = 50 K = 50

Accuracy using Information gain heuristic: 77.5112443778

Accuracy using Variance Impurity heuristic: 64.5677161419

* + 1. L = 1000 K = 100

Accuracy using Information gain heuristic: 76.2618690655

Accuracy using Variance Impurity heuristic: 64.167916042

* + 1. L = 33 K = 100

Accuracy using Information gain heuristic: 76.311844078

Accuracy using Variance Impurity heuristic: 64.5677161419

* + 1. L = 56 K = 289

Accuracy using Information gain heuristic: 76.4117941029

Accuracy using Variance Impurity heuristic: 49.9750124938

* + 1. L = 300 K = 1000

Accuracy using Information gain heuristic: 76.211894053

Accuracy using Variance Impurity heuristic: 64.4177911044

* + 1. L = 170 K = 10

Accuracy using Information gain heuristic: 76.0619690155

Accuracy using Variance Impurity heuristic: 67.1664167916

* + 1. L = 5 K = 19

Accuracy using Information gain heuristic: 75.7621189405

Accuracy using Variance Impurity heuristic: 54.6726636682

* + 1. L = 999 K = 999

Accuracy using Information gain heuristic: 77.1614192904

Accuracy using Variance Impurity heuristic: 63.4182908546

* 1. DATASET 2:

Before pruning:

Accuracy using Information gain heuristic: 72.2129783694

Accuracy using Variance Impurity heuristic: 58.5690515807

* + 1. L = 100 K = 20

Accuracy using Information gain heuristic: 71.7138103161

Accuracy using Variance Impurity heuristic: 60.8985024958

* + 1. L = 20 K = 10

Accuracy using Information gain heuristic: 76.2063227953

Accuracy using Variance Impurity heuristic: 52.0798668885

* + 1. L = 50 K = 50

Accuracy using Information gain heuristic: 75.3743760399

Accuracy using Variance Impurity heuristic: 60.2329450915

* + 1. L = 1000 K = 100

Accuracy using Information gain heuristic: 73.7104825291

Accuracy using Variance Impurity heuristic: 57.7371048253

* + 1. L = 33 K = 100

Accuracy using Information gain heuristic: 75.3743760399

Accuracy using Variance Impurity heuristic: 56.9051580699

* + 1. L = 56 K = 289

Accuracy using Information gain heuristic: 72.0465890183

Accuracy using Variance Impurity heuristic: 61.5640599002

* + 1. L = 300 K = 1000

Accuracy using Information gain heuristic: 74.8752079867

Accuracy using Variance Impurity heuristic: 49.9168053245

* + 1. L = 170 K = 10

Accuracy using Information gain heuristic: 76.0399334443

Accuracy using Variance Impurity heuristic: 62.0632279534

* + 1. L = 5 K = 19

Accuracy using Information gain heuristic: 75.2079866889

Accuracy using Variance Impurity heuristic: 45.4242928453

* + 1. L = 999 K = 999

Accuracy using Information gain heuristic: 72.8785357737

Accuracy using Variance Impurity heuristic: 56.0732113145

1. **CONCLUSION**

The accuracy of the decision tree to predict the correct class for the data was calculated using the ID3 algorithm and was found to predict accurately for approximately 75% of the data using the Information Gain heuristic and approximately 50% using the Variance Impurity gain.