Project Report

Decision Tree Classifier Implementation

1. Introduction

This project implements two decision tree classifiers, Quinlan's C4.5 and Breiman's CART. After implementation they will be experimented on the Credit Approval dataset from the University of California Irvine Machine Learning Repository.

2. Dataset

Dataset that is used is the UCI Credit Approval dataset (https://archive.ics.uci.edu/ml/datasets/Credit+Approval). It contains 690 examples over 15 attributes with each example labeled either positive("+") or negative("-"). Among the 15 attributes, 9 of them are categorical and 6 are continuous. The dataset is splited into training set of 80% (550 examples) and a test set of remaining 20% (140 examples). Both have similar distribution: 44% positive and 56% negative.

3. Implementation

Handling Missing Values

Continuous attributes- Missing values are replaced with the median.

Categorical attributes- Missing values are replaced with the alphabetically sorted median.

Decision Tree Classifiers

C4.5- Uses Gain Ratio to choose the best attribute.

CART- Uses Gini Index to determine splits. Model
Training and Evaluation

10-Fold Cross-validation:

The training set is divided into 10 sequential folds. Each fold is used for validation once, and the nice are used for training. The model with best F1-score is selected.

Final Model Evaluation

The best model is tested on the test set, after training it on the entire training set.

4. Results and Discussion

Cross-Validation Results (Training Set)

C4.5	0.8044
CART	0.8014

The model with higher F1-score, C4.5 is selected for final test.

Test Set Values

Accuracy	0.8429
F1-Score	0.8254

C4.5 performed better which tells that the Gain Ratio is more effective in handling categorical attributes.

Accuracy suggests that the model performs good to unseen data.

5. Conclusion

The best model we got is C4.5. The F1-score achieved on the test set is 0.8254.