CS365 Lab C: Decision Tree Results

1. Introduction

This report presents the decision trees generated by our Python implementation of an ID3-like algorithm for Lab C. Three datasets were used to build and evaluate the models: tennis.txt, pets.txt, and titanic2.txt. For each dataset, the output includes:

- The decision tree structure
- The total number of nodes in the tree
- The training set accuracy
- The leave-one-out (LOO) cross validation accuracy

2. Results

2.1 Tennis Dataset

Decision Tree:

```
outlook?
  [outlook = sunny]
    humidity?
      [humidity = high]
        -> no
      [humidity = normal]
        -> yes
  [outlook = rain]
    wind?
      [wind = strong]
        -> no
      [wind = weak]
        -> yes
  [outlook = overcast]
    -> yes
```

Number of nodes in the tree: 8
Training set accuracy: 100.00%
Leave-One-Out Cross Validation Accuracy: 78.57%

2.2 Pets Dataset

Decision Tree:

```
size?
  [size = medium]
    color?
      [color = brown]
        -> yes
      [color = gray]
        -> no
  [size = large]
    -> no
  [size = enormous]
    -> no
  [size = tiny]
    color?
      [color = brown]
        -> no
      [color = white]
        -> no
  [size = small]
    color?
      [color = brown]
        -> no
      [color = orange]
        -> yes
      [color = gray]
        earshape?
          [earshape = folded]
            -> yes
          [earshape = pointed]
            tail?
              [tail = no]
                -> yes
              [tail = yes]
                -> no
```

Number of nodes in the tree: 17
Training set accuracy: 86.67%
Leave-One-Out Cross Validation Accuracy: 46.67%

2.3 Titanic2 Dataset

Decision Tree:

```
sex?
  [sex = male]
    pclass?
      [pclass = crew]
        -> no
      [pclass = 2nd]
        age?
          [age = adult]
            -> no
          [age = child]
            -> yes
      [pclass = 1st]
        age?
          [age = adult]
            -> no
          [age = child]
            -> yes
      [pclass = 3rd]
        age?
          [age = adult]
            -> no
          [age = child]
            -> no
  [sex = female]
    pclass?
      [pclass = crew]
        -> yes
      [pclass = 2nd]
        age?
          [age = adult]
            -> yes
          [age = child]
            -> yes
      [pclass = 1st]
        age?
          [age = adult]
            -> yes
          [age = child]
            -> yes
      [pclass = 3rd]
        age?
          [age = adult]
            -> no
          [age = child]
            -> no
```

Number of nodes in the tree: 23

Training set accuracy: 79.05%

Leave-One-Out Cross Validation Accuracy**:** 79.05%

3. Summary and Observations

• Tennis Dataset:

The tree is relatively simple with 8 nodes and achieves perfect training accuracy. However, the LOO cross validation accuracy of 78.57% suggests that the model may be somewhat overfitting to the small dataset.

• Pets Dataset:

A more complex tree with 17 nodes was generated. Training accuracy is moderate (86.67%), while the LOO cross validation accuracy drops to 46.67%, which could indicate class imbalance or overlapping attribute values affecting generalization.

• Titanic2 Dataset:

This dataset produced a decision tree with 23 nodes and a training accuracy of 79.05%. Additional evaluation metrics (such as LOO cross validation accuracy) would provide further insights into the generalization performance on this more complex, real-world dataset.

4. Conclusion

This report demonstrates the application of a decision tree algorithm across diverse datasets. The variations in tree structure and performance metrics highlight how dataset characteristics, such as size, feature overlap, and class balance, impact model accuracy and generalization. Future work may include implementing pruning techniques or exploring more advanced algorithms (e.g., C4.5) to improve performance and reduce overfitting.