CSE 318 Assignment-04

Decision Tree

Dibya Jyoti Sarkar

2105084

1 Introduction

Decision Tree is a tree structure for conditional step by step decision making about a particular data from given attributes. It is used in binary or multi-classification problem. In this assignment, we used decision tree for classifying output label for two datasets: Iris.csv and Adult.data. Three criteria had been used for splitting attributes: Information Gain(IG), Information Gain Ratio(IGR), Normalized Weighted Information Gain (NWIG).

2 Output Label

- Adult.data: $\leq$50K, $< 50K$ were labels for Adult.data
- Iris.csv: Iris-setosa,Iris-virginica,Iris-Versicolor were labels for Iris.csv

3 Data Preprocessing

Decision Tree can work better in when data is preprocessed.

- **Adult.data:** In adult.data, there were "?" values in rows which could lead decision tree making wrong decisions. Again, there were some columns which were redundant to output label. These would lead to more complex calculations for decision tree as well as decrease the accuracy. For this, the redundant columns and irrelevant signs were removed.
- **Iris.csv:** In iris.csv, there were numerical values. Decision Tree works better when dealing with categorical values. So, the numerical values were discretized that increased the accuracy of decision tree significantly.

4 Graph Plots

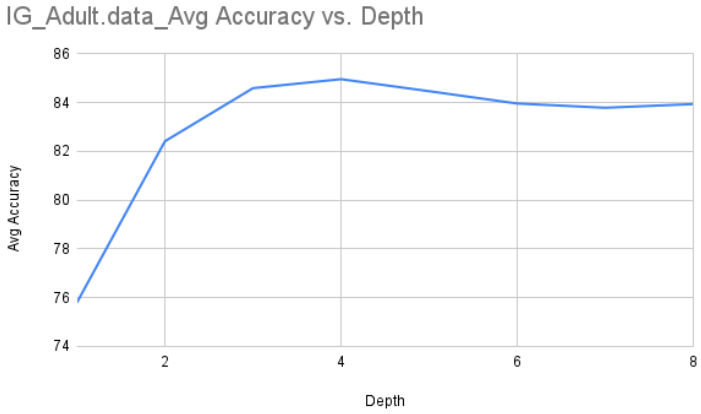
Three criteria had been chosen for splitting attributes for decision tree:

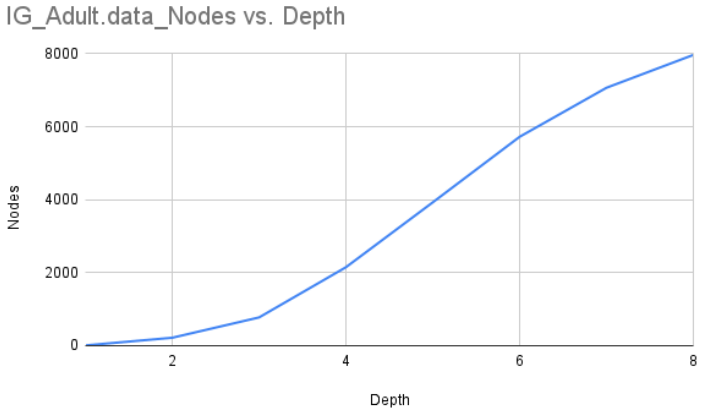
- Information Gain(IG)
- Information Gain Ratio(IGR)
- Normalized Weighted Information Gain (NWIG)

For each of these criteria, Average Accuracy vs Depth, Nodes vs Depth has been plotted. Lastly, plots of number of nodes, average accuracy, depths has been given for each criteria without using depth based pruning. The graph plots are given below:

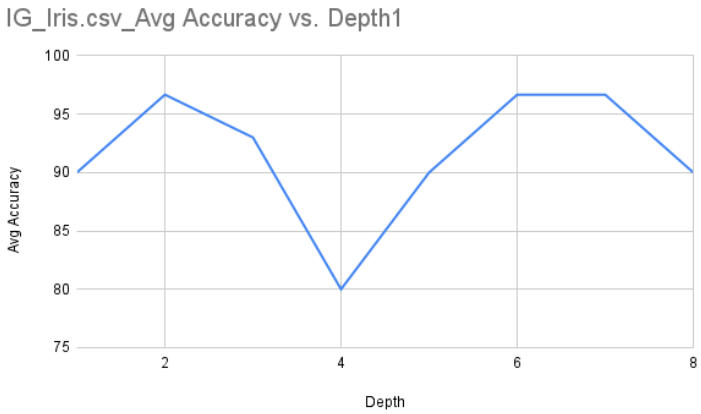
4.1 IG

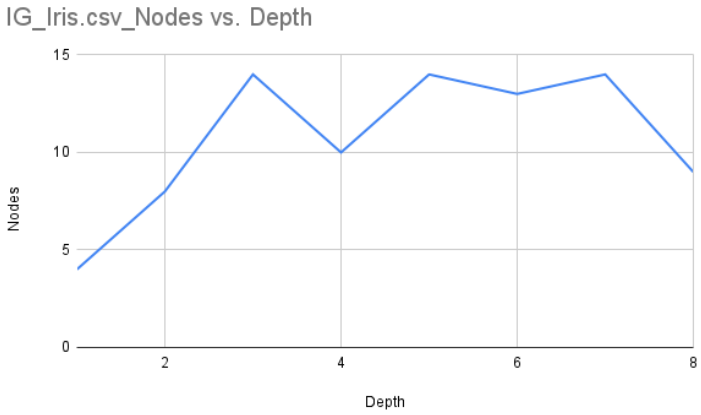
4.1.1 Adult.data dataset





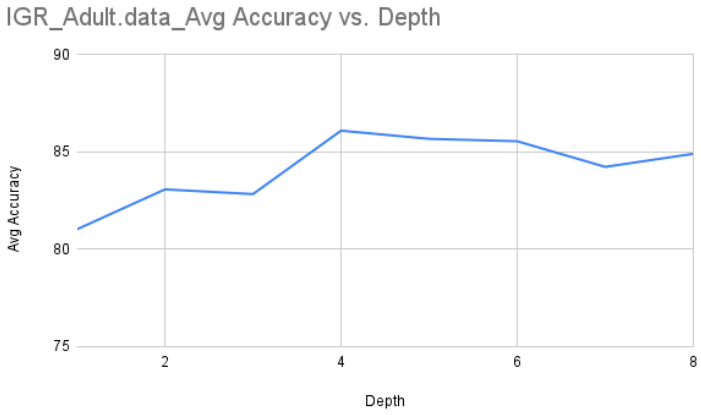
4.1.2 Iris.csv dataset

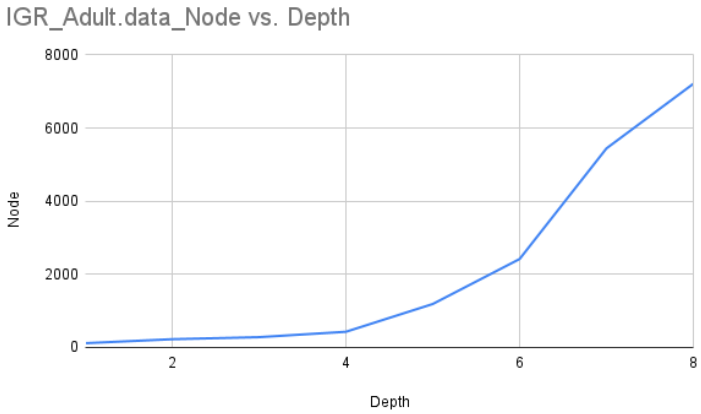




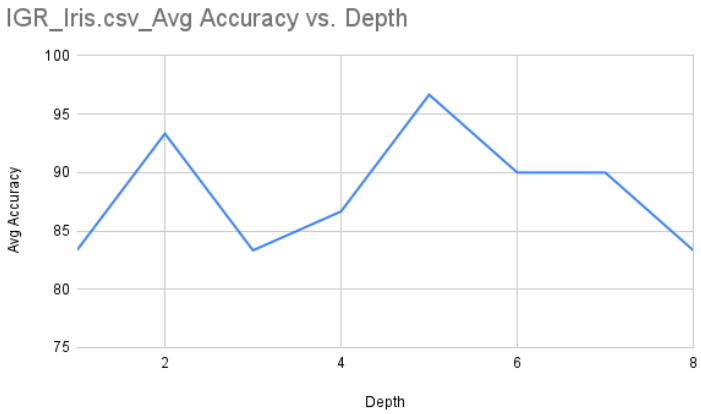
4.2 IGR

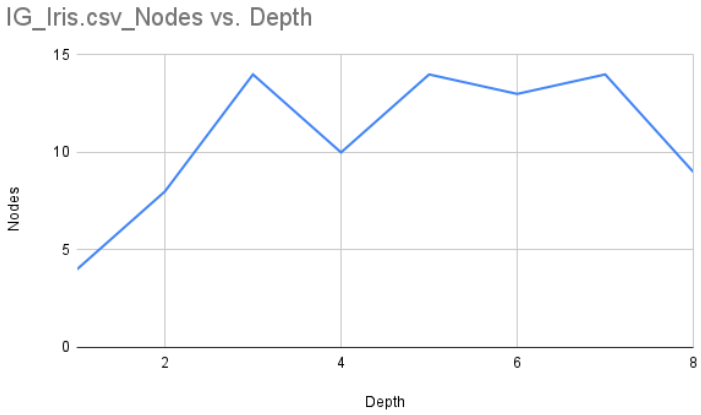
4.2.1 Adult.data dataset





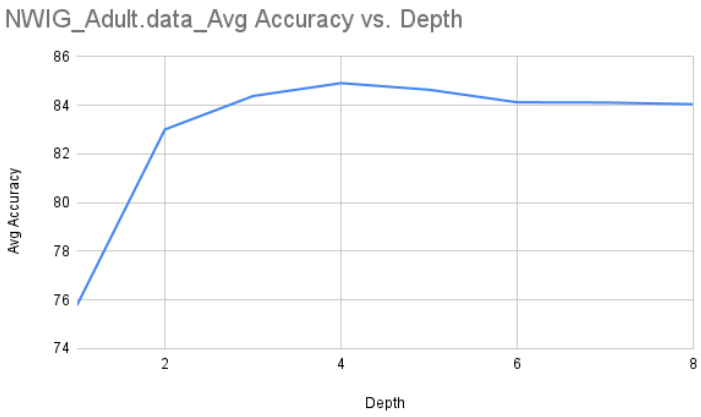
4.2.2 Iris.csv dataset

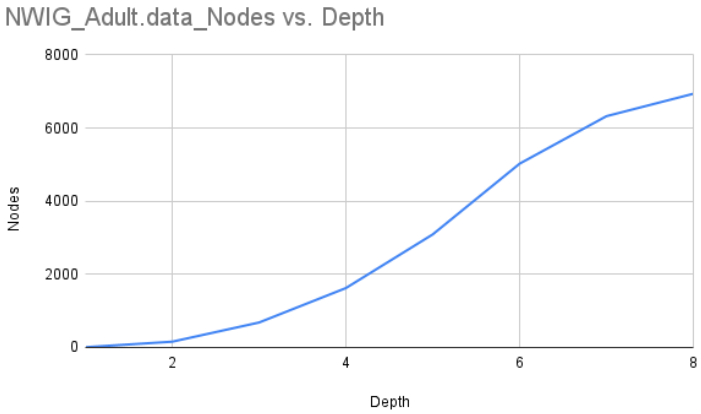




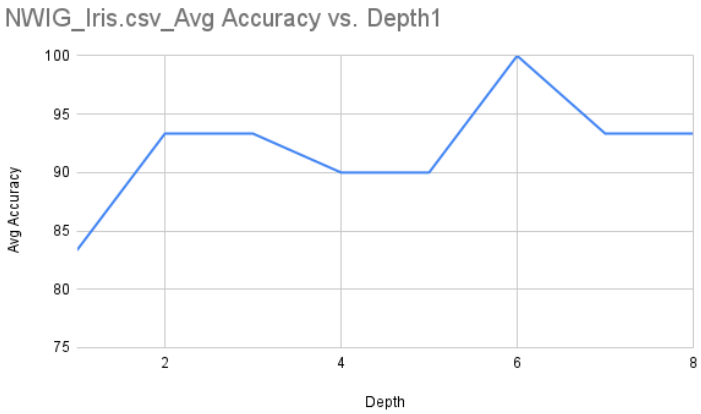
4.3 NWIG

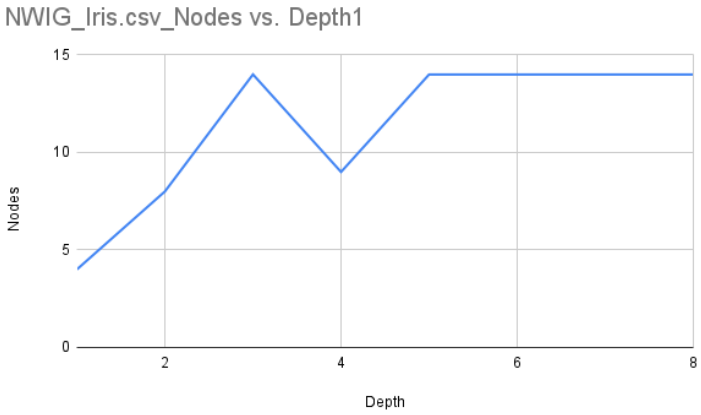
4.3.1 Adult.data dataset





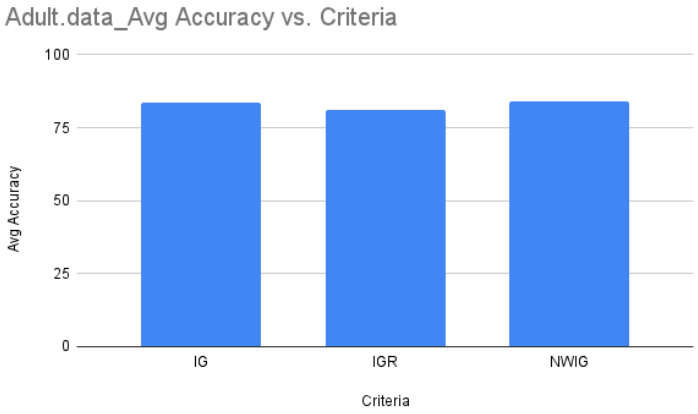
4.3.2 Iris.csv dataset

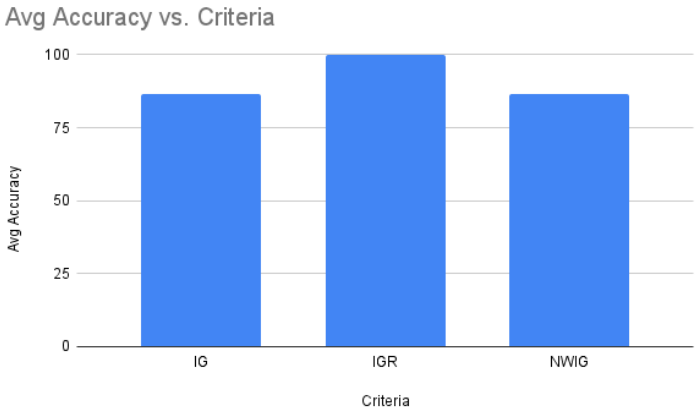




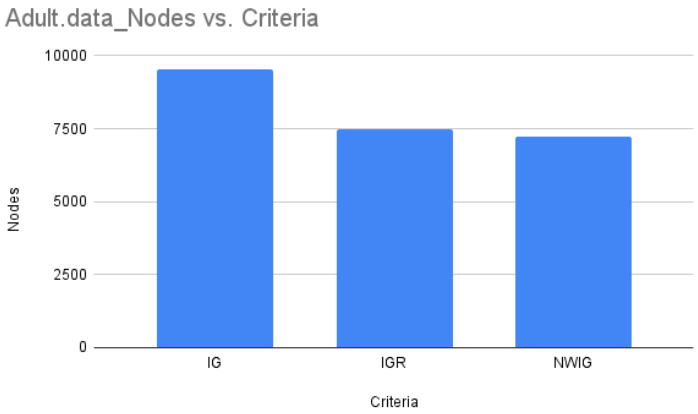
4.4 Without Pruning

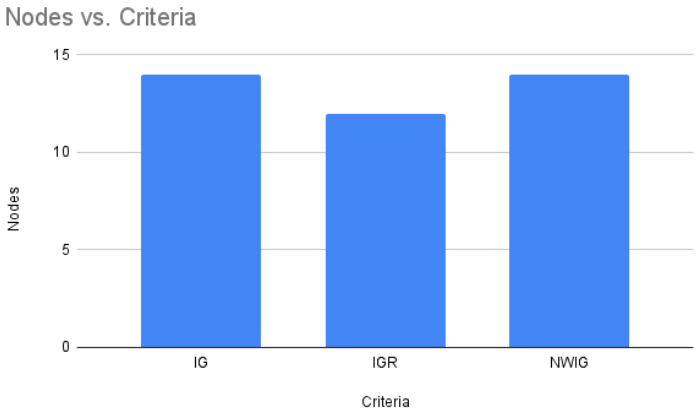
4.4.1 Average Accuracy vs Criteria



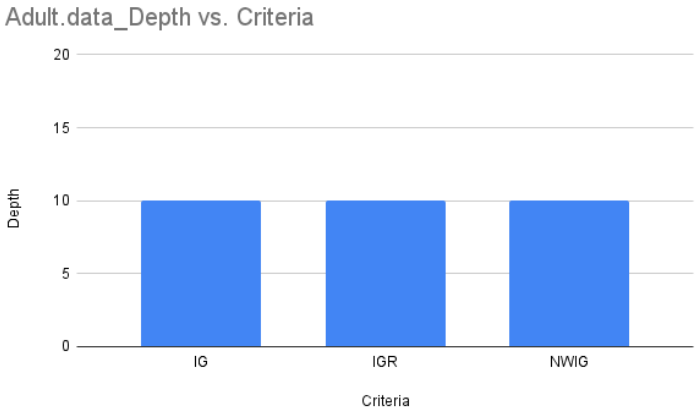


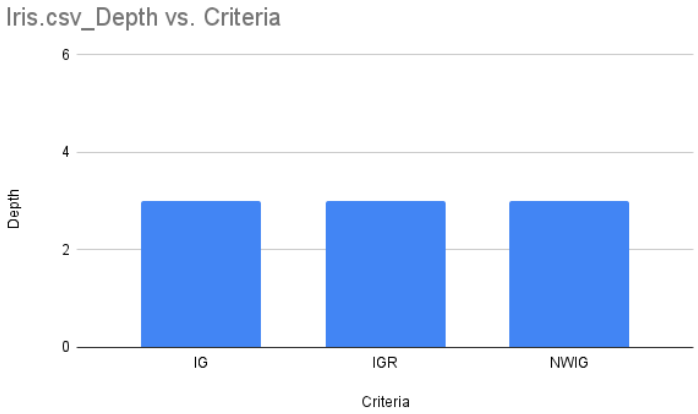
4.4.2 Nodes vs Criteria





4.4.3 Depths vs Criteria





5 Observation and Analysis

5.1 Performance in both datasets

5.1.1 Adult.data

From the graph plots, it can be observed that in terms of average accuracy vs depth, IG,NWIG provided nearly the same results, where as IGR performed consistently better in adult.data. IGR's performance was better because it penalizes attributes that split data in too many subsets. So, it creates more balanced trees compared to IG. On the other hand, NWIG sometimes overpenalizes high cardinality attributes.

However, in terms of no depth based pruning, IGR performs slightly less than IG or NWIG. This indicates that eventually it can fall into overfitting issues.

5.1.2 Iris.csv

From the graph plots, it can be observed that in terms of average accuracy vs depth, NWIG performed consistently better than IG, IGR. IG put up a good performance till depth 4, it went down from 96 percent to 80 percent, indicating that it might have chosen wrong attribute to split, leading to collapse although it recovered in next depths. IGR performed moderately in this dataset although in increased depths, it's accuracy fell down due to overfitting issue.

5.2 Overfitting issue

In all three criteria, we can see that after a certain depth, decision tree tends to give less accurate results. This happens due to overfitting. It is beacuse decision trees starts to memorizing instead of generalizing. That's why it can be seen from graphs that depth based pruning can increase the average accuracy.

In adult.data, all three criteria tried to contain overfitting and we can see that in graphs too. However, in iris.csv, NWIG contained overfitting significantly compared to IGR and IG.We knwo that in higher depths, sample size becomes small and cardinality of distinct values increases. For this, NWIG penalizes small sample size $|S|$ as well as attributes with many distinct values thus containing the overfitting.

5.3 Unexpected Patterns

There is an unexpected pattern where IG criteria has been used for iris.csv. Here, we can see that the accuracy suddenly dropped to 80 percent from 96 percent. This happened beacuse IG might have chosen wrong attribute to split on. Same thing can be noticed in IGR too. However, they both recovered from that in next depths.

5.4 Preprocessing Helps

Preprocessing the data before feeding it to decision tree improve tree's decision making accuracy significantly.

In adult.data, there were redundant information that could lead to complex trees. This complexity could lead to decrease in tree's decision making accuracy.

In iris.csv, the data were numericals. Decision Tree works better with categorical values. So, discretizing the numerical values helped increasing accuracy a lot.

6 Discussion

We can make several key observations regarding the performance of decision tree based on three criteria: IG,IGR,NWIG. For example, IG sometimes perform less than expected due to it's information bias. IGR performs consistently and moderately due to penalizing attributes that lead to too many subsets. NWIG performs nearly same as IG for large sample size but performs significantly better in smaller datasets.

Additionally, it can be observed that dataset size and preprocessing can help improving tree's decision making accuracy. Redundant information as well as missing data can make sample size large,complex and lead to making decision tree error prone. Discretizing numerical values to categorical values can also help in improving tree's accuracy.

In short, it can be said that choosing criteria based on dataset size, choosing correct depth for pruning and preprocessing dataset can help decision tree improve their decision making accuracy.