Kailash Bhanushali

Programming Assignment 2 1001398090

## Decision Tree

Decision trees is used to classify the data. Classification of the data is done to accommodate items of a similar types of data. The aim for classification technique is to classify and predict items in a dataset accurately.

Decision trees are simple, flexible and immune to noise and ambiguity. It can classify the data doesn't matter whether it is discrete or continuous data. A decision tree has a root node i.e. top node, leaf nodes and branches. The sub nodes are the conditions on the attributes, branches are the results and leaf are class labels.

There are many different algorithms used to implement decision tree such as ID3, CART, C4.5 etc. Below are the methods which can be used to reduce the impurity out of the dataset. They are as follows:

1) Entropy

$$\text{Entropy} = -\sum_{i} P_i \log_2 P_i$$

where Pi stands for prediction

2) Gini

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

3) Misclassification Error

$$Error(t) = 1 - \max_i P(i \mid t)$$

Kailash Bhanushali

Programming Assignment 2 1001398090

## Dependencies

## Data Description

```
from sklearn.datasets import load_wine
```

```
data = load_wine()
df=pd.DataFrame(data=np.c_[data['data'],data['target']],columns=data['feature_names']+['target'])
```

We use the following command to get the description of the dataset:-

```
print(raw_data['DESCR'])
```

OUTPUT:-

```
Data Set Characteristics:

                :- class:
                 - class_0
                 - class_1
                 - class_2
```

Out of the many things we observe that there are three classes (creatively named `'class_0'`, `'class_1'`, and `'class_2'`).

### How Do Decision Trees Work?

Below are the steps involved in the building of a decision tree.

1. **Splitting:** The process of segregating the data set into subsets. For each split, two determinations are made: the predictor variable used for the split, i.e. the splitting variable, and Splitting at the point between left and right is called as split point. Above methods are the techniques which help us to decide the node to split into left and right node. The leaf node, also called a terminal node.

2. **Pruning:** Pruning is the process of minimizes the size of the tree by turning some branch nodes into leaf nodes, and deleting the leaf nodes under the original branch. It is useful as because classification may act bad when new data is processed to it. Pruning facilitates you to find the next biggest tree and minimize the problem. A simpler tree often avoids over-fitting.

3. **Tree Selection:** The process of finding the smallest tree that fits the data. Usually this is the tree that yields the lowest cross-validated error.

### Splitting of data:

In order to efficiently train as well as test the model, we will need to split the data into a training set which we will feed to our model along the the training labels. Then after training the model, we will test it on the 'test' data, so that we can simulate the real-world applicability of the model.

Kailash Bhanushali

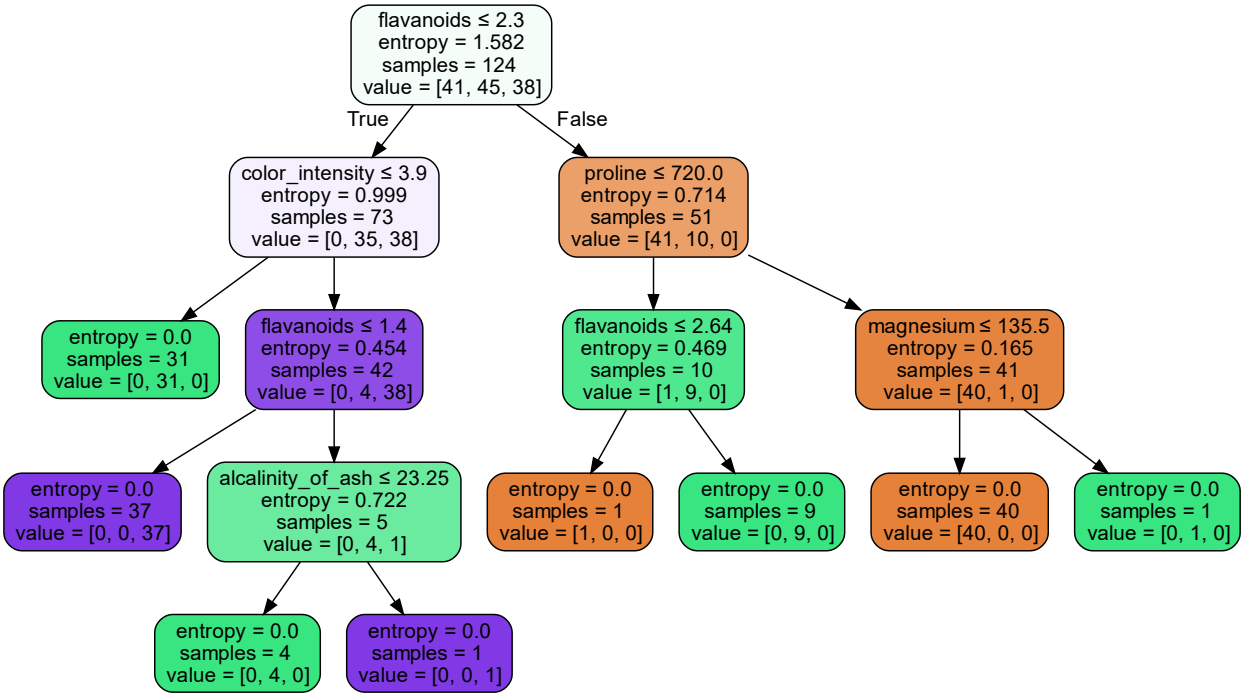
Programming Assignment 2 1001398090

We will be using train_test_split() and test_Size to test the dataAs given we will be using 30% to test the data.

```
data_train, data_test, label_train, label_test = \
train_test_split(data['data'],data['target'],test_size=0.3)
warnings.filterwarnings("ignore")
print(len(data_train),' samples in training data\n',len(data_test),' samples in test data\n')
```

## Entropy

Entropy splits the data into left and right based on the sample training value generated.



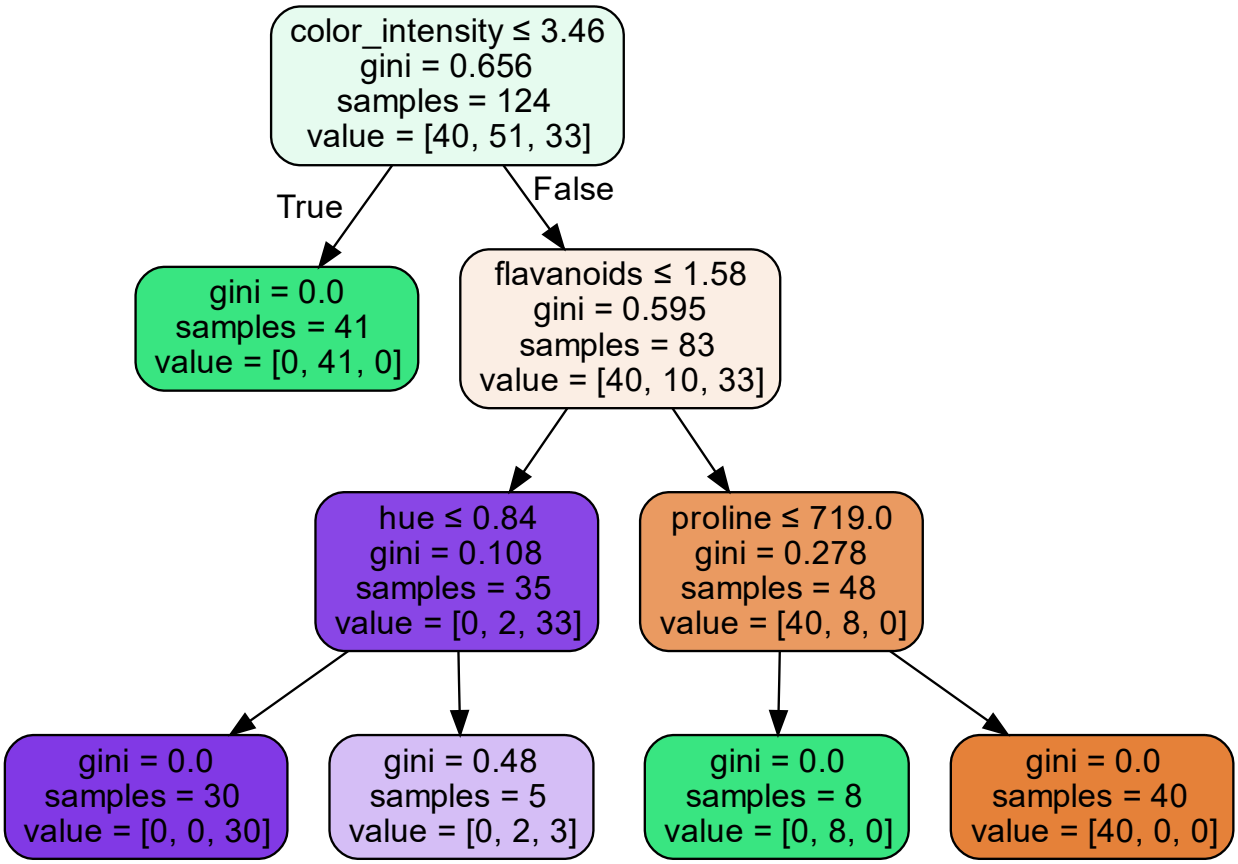
Kailash Bhanushali

Programming Assignment 2

1001398090

## Gini Impurity:

Another method to reduce and find the impurity in the given dataset.




## Comparison between Gini and Entropy:

The values for precision, Recall, Fscore are marginally same with entropy having an edge over Gini classification. So we can conclude that Gini impurity and Information Gain Entropy are pretty much the same. And people do use the values interchangeably.

Gini impurity doesn't require us to compute logarithmic functions, which are computationally intensive.

## References:

1) https://www.saedsayad.com/decision_tree.htm
2) https://gerardnico.com/data_mining/entropy

3) http://www.simafore.com/blog/bid/94454/A-simple-explanation-of-how-entropy-fuels-a-decision-tree-model