# GAZİ UNIVERSITY
# FACULTY OF ENGINEERING
# DEPARTMENT OF COMPUTER ENGINEERING

## CENG476 INTRODUCTION TO MACHINE LEARNING

Assignment II Part I Report

Midterm

171180758

Candan Baykan

May 2021

1. Please write about how you implement the decision tree and include a visual representation of the decision tree in the format provided in (A)

**Pseudocode:**

```
1. Create a node to add to the tree.
2. If all rows have same class, return the node with that class.
3. If there is no feature left, then return the node with most common class.
4. Else
    a. Choose the best feature to split tree using information gain or gini index.
    b. Remove the chosen feature from list.
    c. Set label to the best feature's name.
    d. For each unique value in best feature:
        i. Filter the rows that meet the condition (best_feature == value).
        ii. If there are no rows that meet the condition, create a subnode with the most common value in unfiltered data.
        iii. Else, create a new tree with filtered data under the new node.
5. Return the node.
```

**Visualization for Information Gain:**

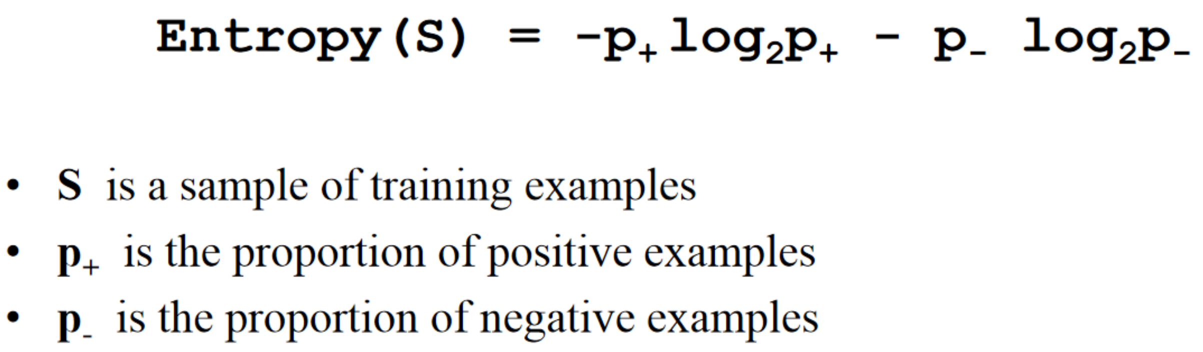
```
safety = med
| persons = 2 : no
| persons = 4
        | buying = high
                | maint = vhigh : no
                | maint = med : no
                | maint = low : yes
                | maint = high : yes
        | buying = vhigh
                | maint = high : no
                | maint = vhigh : no
                | maint = med : yes
                | maint = low : no
        | buying = med
```

```
                | maint = med : yes
                | maint = high : yes
                | maint = vhigh : yes
                | maint = low : yes
        | buying = low
                | maint = high : yes
                | maint = med : yes
                | maint = low : yes
                | maint = vhigh : no
| persons = more
        | buying = low
                | maint = low : yes
                | maint = vhigh : yes
                | maint = med : yes
                | maint = high : yes
        | buying = high
                | maint = high : yes
                | maint = low : yes
                | maint = med : no
                | maint = vhigh : no
        | buying = med
                | maint = high : yes
                | maint = vhigh : yes
                | maint = low : yes
                | maint = med : yes
        | buying = vhigh
                | maint = high : no
                | maint = med : yes
                | maint = low : yes
                | maint = vhigh : no
safety = high
| persons = 4
        | maint = high
                | buying = high : yes
                | buying = low : yes
                | buying = vhigh : no
                | buying = med : yes
        | maint = vhigh
                | buying = vhigh : no
                | buying = high : no
                | buying = med : yes
                | buying = low : yes
        | maint = low : yes
        | maint = med : yes
| persons = 2 : no
| persons = more
        | maint = high
                | buying = low : yes
                | buying = med : yes
                | buying = high : yes
                | buying = vhigh : no
        | maint = med
                | buying = low : yes
                | buying = vhigh : yes
                | buying = med : yes
                | buying = high : yes
        | maint = low
                | buying = vhigh : yes
                | buying = high : yes
                | buying = low : yes
                | buying = med : yes
        | maint = vhigh
                | buying = med : yes
                | buying = low : yes
                | buying = vhigh : no
                | buying = high : no
safety = low : no
```

**Visualization for Gini Index:**

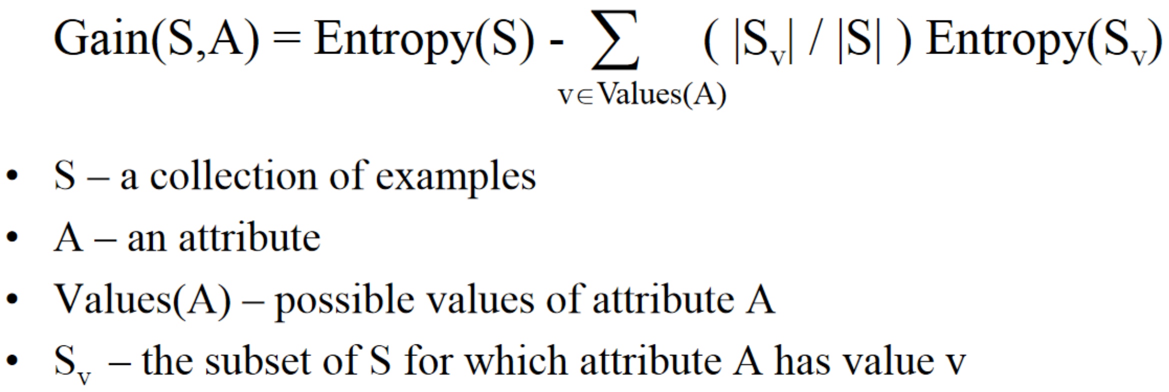
```
safety = med
| persons = 2 : no
| persons = 4
        | buying = high
                | maint = vhigh : no
                | maint = med : no
                | maint = low : yes
                | maint = high : yes
        | buying = vhigh
                | maint = high : no
                | maint = vhigh : no
                | maint = med : yes
                | maint = low : no
        | buying = med
                | maint = med : yes
                | maint = high : yes
                | maint = vhigh : yes
                | maint = low : yes
        | buying = low
                | maint = high : yes
                | maint = med : yes
                | maint = low : yes
                | maint = vhigh : no
| persons = more
        | buying = low
                | maint = low : yes
                | maint = vhigh : yes
                | maint = med : yes
                | maint = high : yes
        | buying = high
                | maint = high : yes
                | maint = low : yes
                | maint = med : no
                | maint = vhigh : no
        | buying = med
                | maint = high : yes
                | maint = vhigh : yes
                | maint = low : yes
                | maint = med : yes
        | buying = vhigh
                | maint = high : no
                | maint = med : yes
                | maint = low : yes
                | maint = vhigh : no
safety = high
| persons = 4
        | maint = high
                | buying = high : yes
                | buying = low : yes
                | buying = vhigh : no
                | buying = med : yes
        | maint = vhigh
                | buying = vhigh : no
                | buying = high : no
                | buying = med : yes
                | buying = low : yes
        | maint = low : yes
        | maint = med : yes
| persons = 2 : no
| persons = more
        | maint = high
                | buying = low : yes
                | buying = med : yes
                | buying = high : yes
                | buying = vhigh : no
        | maint = med
                | buying = low : yes
                | buying = vhigh : yes
                | buying = med : yes
                | buying = high : yes
        | maint = low
                | buying = vhigh : yes
                | buying = high : yes
                | buying = low : yes
                | buying = med : yes
        | maint = vhigh
                | buying = med : yes
                | buying = low : yes
                | buying = vhigh : no
                | buying = high : no
safety = low : no
```

2. Please write about what you have done at this stage, how do you use the information gain and the Gini index to decide how to split your tree? Also include information on the value of Information Gain and Gini Index of the root node using your model and scikit-learn.



$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

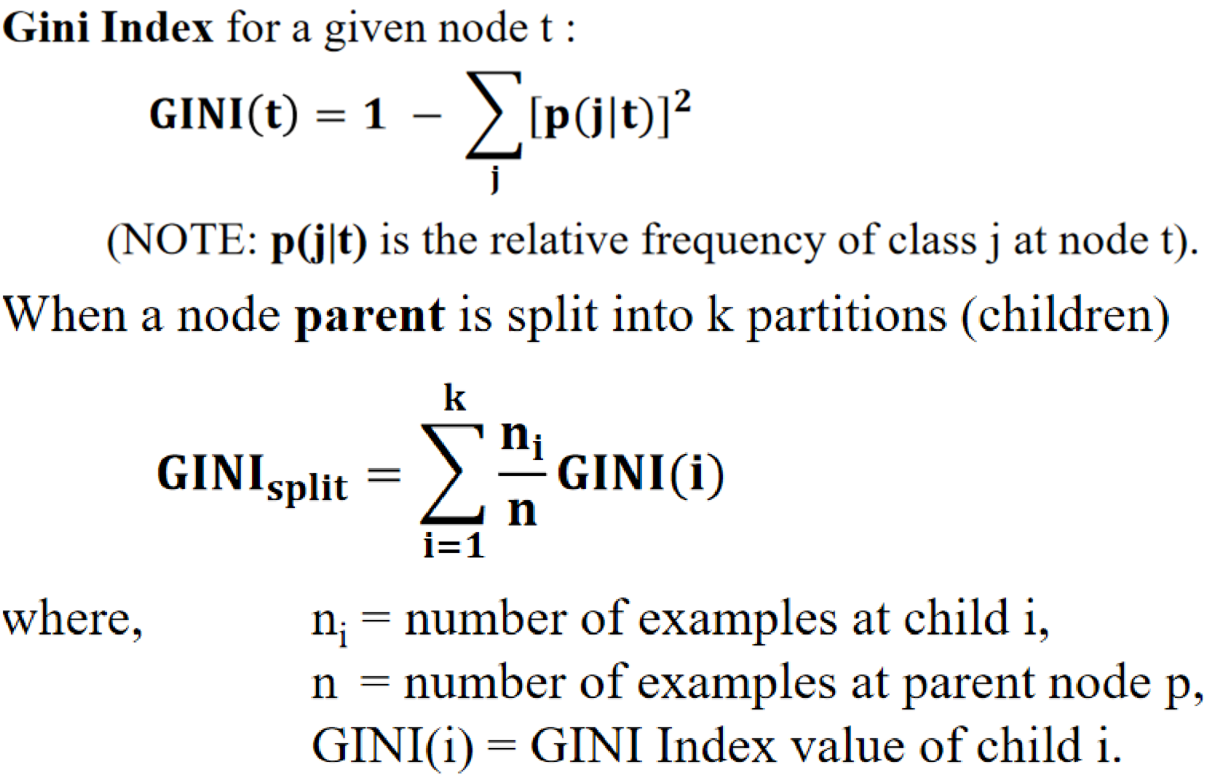
- **S** is a sample of training examples
- $\mathbf{p_+}$ is the proportion of positive examples
- $\mathbf{p_-}$ is the proportion of negative examples



$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} (|S_v| / |S|)\ \text{Entropy}(S_v)$$

- S – a collection of examples
- A – an attribute
- Values(A) – possible values of attribute A
- $S_v$ – the subset of S for which attribute A has value v

For each feature, I applied the formulas above. Then I chose the feature that has the greatest information gain. (Çiçekli, Decision Tree Learning, n.d.)



**Gini Index** for a given node t :

$$\mathbf{GINI(t) = 1 - \sum_{j} [p(j|t)]^2}$$

(NOTE: **p(j|t)** is the relative frequency of class j at node t).

When a node **parent** is split into k partitions (children)

$$\mathbf{GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)}$$

where, $n_i$ = number of examples at child i,
n = number of examples at parent node p,
GINI(i) = GINI Index value of child i.

For each feature, I applied the formulas above. Then I chose the feature that has the least gini index. (Çiçekli, Classification, n.d.)

**My model:**

Root Node's Information Gain: 0.23139045239171413

Root Node's Gini Index: 0.32264253932307974

**scikit-learn:**

Root Node's Information Gain: 0.882

Root Node's Gini Index: 0.42

My model’s and scikit-learn’s root node have different information gain and gini index because my tree has multiway splits while scikit-learn only has binary splits.

3. Please explain your experiments and findings. The labels generated on the test data and accuracy on the test data using your model and scikit-learn.

My findings are below:

- If the safety is low, car is not profitable.
- If the person capacity is 2, car is not profitable.
- Higher safety, lower buying price, lower maintenance cost and higher person capacity makes car profitable.

**My model:**

Information Gain: 0.91

Gini Index: 0.91

**scikit-learn:**

Information Gain: 0.91

Gini Index: 0.91

There were no differences between the predictions of all implementations.

# REFERENCES

Çiçekli, İ. (n.d.). *Classification*. Retrieved from Hacettepe University: https://web.cs.hacettepe.edu.tr/~ilyas/Courses/VBM684/lec05_Classification_DecisionTree.pdf

Çiçekli, İ. (n.d.). *Decision Tree Learning*. Retrieved from Hacettepe University: https://web.cs.hacettepe.edu.tr/~ilyas/Courses/BIL712/lec02-DecisionTree.pdf