Week 8

I used online calculators to calculate entropy, entropy average, weighted entropy average, Information Gain, Gini Impurity, Gini Impurity average, and weighted Gini Impurity average.

1. Calculate Entropy and Information Gain:

Calculate Entropy for Entire Dataset:

E(D) = 1

Calculate Entropy for Each Attribute:

Weather Condition

E(Rain) = 0.92

E(Snow) = 0.92

E(Clear) = 1

Road Condition

E(Bad) = 0.81

E(Average) = 1

E(Good) = 0.81

Traffic Condition

E(High) = 0.82

E(Normal) = 0.92

E(Light) = 0.92

Engine Problem

E(Yes) = 0.81

E(No) = 0.92

Calculate the Weighted Entropy Average for Each Attribute:

Weather Condition = 0.95

Road Condition = 0.85

Traffic Condition = 0.88

Engine Problem Condition = 0.88

Calculate the Information Gain for Each Attribute:

Weather Condition = 0.05

Road Condition = 0.15

Traffic Condition = 0.12

Engine Problem = 0.12

Choose Best Attribute as Root Node:

Road Condition has highest information gain value.

Road Condition is root node.

2. Calculate Gini Impurity:

Calculate Gini Impurity for Entire Dataset:

Gini(D) = 0.5

Calculate Gini Impurity for Each Attribute:

Weather Condition

Gini(Rain) = 0.44

Gini(Snow) = 0.44

Gini(Clear) = 5

Road Condition

Gini(Bad) = 0.38

Gini(Average) = 0.5

Gini(Good) = 0.38

Traffic Condition

Gini(High) = 0.38

Gini(Normal) = 0.44

Gini(Light) = 0.44

Engine Problem

Gini(Yes) = 0.5

Gini(No) = 0.5

Calculate the Weighted Gini Impurity Average for Each Attribute:

Weather Condition = 0.47

Road Condition = 0.4

Traffic Condition = 0.42

Engine Problem Condition = 0.5

Choose Best Attribute as Root Node:

Road Condition has lowest Gini impurity value.

Road Condition is root node.

3. Construct the Decision Tree:

Determine first layer:

Root node = Road Condition

Branch 1 = Bad

Branch 2 = Average

Branch 3 = Good

Determine Branch 1 with Information Gain:

Weather Condition = 0.31

Traffic Condition = 0

Engine Problem = 0.31

Use Weather or Engine

Intermediate node = Weather Condition

Branch 1.1 = Rain

Branch 1.2 = Snow

Branch 1.3 = Clear

Determine Branch 2 with Information Gain:

Weather Condition = 0

Traffic Condition = 0

Engine Problem = 1

Intermediate node = Engine Problem

Branch 2.1 = Yes

Branch 2.2 = No

Determine Branch 3 with Information Gain:

Weather Condition = 0.31

Traffic Condition = 0

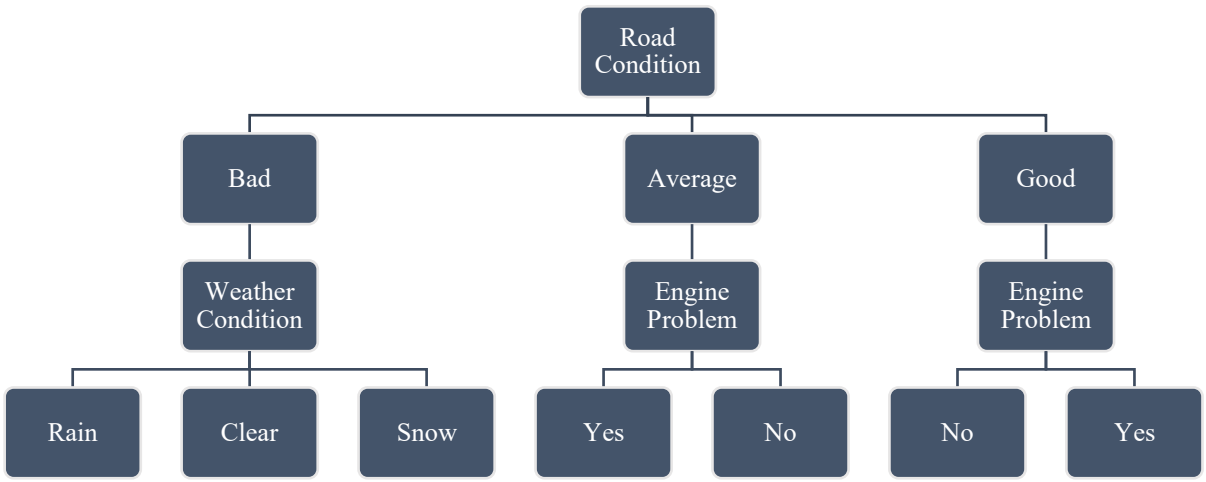
Engine Problem = 0.81

Intermediate node = Engine Problem

Branch 3.1 = No

Branch 3.2 = Yes

Final Decision Tree:



Reflection

For this week, I learned two different methods for manually building a decision tree. The first method involved calculating the entropy of all instances '*D*', followed by calculating the entropy of each subset for each attribute, '*D sub'*. Next the weighted average entropy for each attribute is computed, '*D average'*. Finally, the information gain is calculated for each attribute as the difference between '*D*' and '*D average'*. Road condition had the highest information gain, making it the root node.

The other method of Gini Impurity was used to display an alternative way to construct a decision tree. First the impurity of the entire dataset is calculated. This is followed by the impurity of each attribute's subsets. Lastly, the average weighted impurity of each attribute is calculated. Unlike information gain, a lower value using the Gini Impurity method indicates the root node. The lowest impurity was road condition, meaning both methods chose the same attribute as the root node.

I chose the information gain method to build my decision tree. Road condition was chosen as the root node because it had the highest information gain. The branches were thus its values of bad, average, and good. The information gain process was then repeated with the remaining attributes. Traffic condition at each branch provided no information gain as was thus discarded. The end result is shown in the graphic above.