

WEEK 8 LAB

Dataset Overview

SNo.	Weather Condition	Road Condition	Traffic Condition	Engine Problem	Accident
1	Rain	Bad	High	No	Yes
2	Snow	Average	Normal	Yes	Yes
3	Clear	Bad	Light	No	No
4	Clear	Good	Light	Yes	Yes
5	Snow	Good	Normal	No	No
6	Rain	Average	Light	No	No
7	Rain	Good	Normal	No	No
8	Snow	Bad	High	No	Yes
9	Clear	Good	High	Yes	No
10	Clear	Bad	High	Yes	Yes

Student Exercises

1. Calculate Entropy and Information Gain:
 - Compute the Information Gain for each attribute in the dataset.

Entropy of Dataset D (Training set of instances with class labeled)

Formula to calculate entropy, $Ent(D) = - \sum (p_i * \log_2(p_i))$

Where p_i is the probability of each class.

Total Instances = 10

Accident = Yes → 5

Accident = No → 5

Therefore probabilities accident equals to yes and no ;

$p(\text{Yes}) = 5/10 = 0.5$

$p(\text{No}) = 5/10 = 0.5$

Using formula to calculate entropy;

$$\begin{aligned}
 \text{Entropy}(D) &= -(0.5 * \log_2(0.5) + 0.5 * \log_2(0.5)) \\
 &= -(0.5 * (-1) + 0.5 * (-1)) \\
 &= -(-0.5 - 0.5) \\
 &= 1
 \end{aligned}$$

$$\text{Entropy}(D) = 1.000 \text{ bit}$$

1.2. Calculating entropy for each attributes in a dataset:

----- **Weather Condition**-----

For Rain: instance=3, Yes=1, No =2,

$$\begin{aligned}
 \text{Using formula, Ent(rain)} &= -(1/3 * \log_2(1/3) + 2/3 * \log_2(2/3)) \\
 &= 0.9183
 \end{aligned}$$

For Snow : instance=3, Yes=2, No=1;

$$\begin{aligned}
 \text{Using formula, Ent(snow)} &= -(1/3 * \log_2(1/3) + 2/3 * \log_2(2/3)) \\
 &= 0.9183
 \end{aligned}$$

For Clear : instance=4, Yes=2, No=2;

$$\begin{aligned}
 \text{Using formula, Ent(clear)} &= -(2/4 * \log_2(2/4) + 2/4 * \log_2(2/4)) \\
 &= 1
 \end{aligned}$$

Calculating weight average entropy of weather condition:

$$\begin{aligned}
 \text{Using formula, Ent(weather)} &= 3/10 * 0.9183 + 3/10 * 0.9183 + 4/10 * 1 \\
 &= 0.951
 \end{aligned}$$

Calculating information gain of weather condition;

$$\text{Using formula, Gain(weather)} = 1 - 0.951 = 0.049$$

----- **Road Condition**-----

For Good; instance=4, Yes=1, No=3;

$$\text{Using formula, Ent(good)} = -(1/4 * \log_2(1/4) + 3/4 * \log_2(3/4))$$

$$=0.811$$

For Bad; instance=4, Yes=3, No=1;

$$\begin{aligned}\text{Using formula, Ent(bad)} &= -(3/4 \cdot \log_2(3/4) + 1/4 \cdot \log_2(1/4)) \\ &= 0.811\end{aligned}$$

For Average: instance=2, Yes=1, No=1;

$$\begin{aligned}\text{Using formula, Ent(average)} &= -(1/2 \cdot \log_2(1/2) + 1/2 \cdot \log_2(1/2)) \\ &= 1.0000\end{aligned}$$

Calculating weight average entropy of Road condition:

$$\begin{aligned}\text{Using formula, Ent(road)} &= 4/10 \cdot 0.811 + 4/10 \cdot 0.811 + 2/10 \cdot 1 \\ &= 0.849\end{aligned}$$

Calculating information gain of Road condition;

$$\text{Using formula, Gain(road)} = 1 - 0.849 = 0.151$$

-----Traffic Condition-----

For High; instance=4, Yes=3, No=1;

$$\begin{aligned}\text{Using formula, Ent(high)} &= -(3/4 \cdot \log_2(3/4) + 1/4 \cdot \log_2(1/4)) \\ &= 0.811\end{aligned}$$

For Normal; instance=3, Yes=1, No=2;

$$\begin{aligned}\text{Using formula, Ent(normal)} &= -(1/3 \cdot \log_2(1/3) + 2/3 \cdot \log_2(2/3)) \\ &= 0.918\end{aligned}$$

For Light; instance=3, Yes=1, No=2;

$$\text{Using formula, Ent(light)} = -(1/3 \cdot \log_2(1/3) + 2/3 \cdot \log_2(2/3))$$

$$=0.918$$

Calculating weight average entropy of Traffic condition:

$$\begin{aligned}\text{Using formula, Ent(traffic)} &= 4/10 \cdot 0.811 + 3/10 \cdot 0.918 + 3/10 \cdot 0.918 \\ &= 0.875\end{aligned}$$

Calculating information gain of weather condition;

$$\text{Using formula, Gain(traffic)} = 1 - 0.875 = 0.125$$

-----Engine Problem-----

For YES; instance=4, Yes=3, No=1;

$$\begin{aligned}\text{Using formula, Ent(yes)} &= -(3/4 \cdot \log_2(3/4) + 1/4 \cdot \log_2(1/4)) \\ &= 0.811\end{aligned}$$

For NO; instance=6, Yes=2, No=4;

$$\begin{aligned}\text{Using formula, Ent(no)} &= -(2/6 \cdot \log_2(2/6) + 4/6 \cdot \log_2(4/6)) \\ &= 0.918\end{aligned}$$

Calculating weight average entropy of Traffic condition:

$$\begin{aligned}\text{Using formula, Ent(engine)} &= 4/10 \cdot 0.811 + 6/10 \cdot 0.918 \\ &= 0.875\end{aligned}$$

Calculating information gain of weather condition;

$$\text{Using formula, Gain(engine)} = 1 - 0.875 = 0.125$$

Determine the root of the decision tree.

=> The root of the decision tree is Road condition with the highest information gain of 0.151

2. Calculate Gini Impurity:

- Perform the same split analysis using Gini Impurity.

Gini Impurity Formula:

$$\text{Gini}(D) = 1 - \sum (p_i^2)$$

For Full Dataset :

$$p(\text{Yes})=0.5,$$

$$p(\text{No})=0.5,$$

$$\text{Gini}(D) = 1 - (0.5*0.5 + 0.5*0.5)$$

$$= 1 - (0.25 + 0.25)$$

$$= 1 - 0.5 = 0.5$$

Gini Impurity of all subsets:

-----weather-----

- Rain: Yes=1 No=2

$$\rightarrow \text{Gini}(\text{rain}) = 1 - [(1/3*1/3) + (2/3*2/3)]$$

$$= 0.4444$$

- Snow: Yes=2 No=1

$$\rightarrow \text{Gini}(\text{snow}) = 1 - [(2/3*2/3) + (1/3*1/3)]$$

$$= 0.4444$$

- Clear: Yes=2 No=2

$$\rightarrow \text{Gini}(\text{clear}) = 1 - [(2/4*2/4) + (2/4*2/4)]$$

$$= 0.5$$

$$\text{Weight gini}(\text{weather}) = 3/10*(0.444) + 3/10*(0.444) + 4/10*(0.5)$$

$$= 0.4667$$

$$\text{Gain (weather)} = \text{gini(D)} - \text{weight gini(weather)} = 0.0333$$

----- Road Condition -----

- Bad: Yes=3, No=1 $\rightarrow \text{Gini(bad)} = 1 - [(3/4 * 3/4) + (1/4 * 1/4)] = 0.375$
- Average: Yes=1, No=1 $\rightarrow \text{Gini(average)} = 1 - [(1/2 * 1/2) + (1/2 * 1/2)] = 0.5$
- Good: Yes=1, No=3 $\rightarrow \text{Gini (good)} = 1 - [(1/4 * 1/4) + (3/4 * 3/4)] = 0.375$

$$\text{Weight gini(road)} = 0.4(0.375) + 0.2(0.5) + 0.4(0.375)$$

$$= 0.40$$

$$\text{Gain (road)} = \text{gini(D)} - \text{weight gini(road)} = 0.5 - 0.40 = 0.10$$

-----Traffic Condition -----

- High: Yes=2, No=2 $\rightarrow \text{Gini(high)} = 1 - [(2/4 * 2/4) + (2/4 * 2/4)] = 0.5$
- Normal: Yes=1, No=2 $\rightarrow \text{Gini(normal)} = 1 - [(1/3 * 1/3) + (2/3 * 2/3)] = 0.4444$
- Light: Yes=1, No=2 $\rightarrow \text{Gini (light)} = 1 - [(1/3 * 1/3) + (2/3 * 2/3)] = 0.444$

$$\text{Weight gini(traffic)} = 0.4(0.5) + 0.3(0.4444) + 0.3(0.4444)$$

$$= 0.4667$$

$$\text{Gain (traffic)} = \text{gini(D)} - \text{weight gini(traffic)} = 0.0333$$

-----Engine Problem-----

- YES: Yes=3, No=1 $\rightarrow \text{Gini(high)} = 1 - [(3/4 * 3/4) + (1/4 * 1/4)] = 0.375$
- NO: Yes=2, No=4 $\rightarrow \text{Gini(normal)} = 1 - [(2/6 * 2/6) + (4/6 * 4/6)] = 0.4444$

$$\text{Weight gini(engine)} = 0.4(0.375) + 0.3(0.4444)$$

=0.417

Gain (engine) = gini(D)-weight gini(engine) =0.083

-

3. Construct the Tree:

- Build the decision tree manually based on your calculations.

Road

GOOD

AVERAGE

BAD