WEEK 8

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)

Fromula to calculate entropy, $Ent(D) = - \Sigma (pi * \log2(pi))$

Where pi is the probability of each class.

Total Instances = 10
Accident = Yes → 5
Accident = No → 5

Therefore probabilities accident equals to yes and no ;

p(Yes) = 5/10 = 0.5
p(No) = 5/10 = 0.5

Using formula to calcualte entropy;

Entropy(D) = -(0.5 * log2(0.5) + 0.5 * log2(0.5))
= -(0.5* (-1)+0.5(-1))
= -(-0.5-0.5)
= 1
Entropy(D) = 1.000 bit

1.2. Calculating entropy for each attributes in a dataset:

- **Weather Condition**

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

Using formula, Ent(rain) = -(1/3*log2(1/3)+2/3*log2(2/3))
=0.9183

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

Using formula, Ent(snow) = -(1/3*log2(1/3)+2/3*log2(2/3))
= 0.9183

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

Using formula, Ent(clear) = -(2/4*log2(2/4)+2/4*log2(2/4))
= 1

Calculating weight average entropy of weather condition:

Using formula, Ent(weather)= 3/10*0.9183+3/10*0.9183+4/10*1
= 0.951

Calculating information gain of weather condition;

Using formula, Gain(weather)= 1-0.951 = 0.049

- **Road Condition**

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

Using formula, Ent(good) = -(1/4*log2(1/4)+3/4*log2(3/4))
=0.811

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

Using formula, Ent(bad) = -(3/4*log2(3/4)+1/4*log2(1/4))

= 0.811

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

Using formula, Ent(average) = -(1/2*log2(1/2)+1/2*log2(1/2))

= 1.0000

Calculating weight average entropy of Road condition:

Using formula, Ent(road)= 4/10*0.811+4/10*0.811+2/10*1

= 0.849

Calculating information gain of Road condition;

Using formula, Gain(road)= 1-0.849 = 0.151

- **Traffic Condition**

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

Using formula, Ent(high) = -(3/4*log2(3/4)+1/4*log2(1/4))

=0.811

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

Using formula, Ent(normal) = -(1/3*log2(1/3)+2/3*log2(2/3))

=0.918

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

Using formula, Ent(light) = -(1/3*log2(1/3)+2/3*log2(2/3))
=0.918

Calculating weight average entropy of Traffic condition:

Using formula, Ent(traffic)= 4/10*0.811+3/10*0.918+3/10*0.918

= 0.875

Calculating information gain of weather condition;

Using formula, Gain(traffic)= 1-0.875 = 0.125

- **Engine Problem**

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

Using formula, Ent(yes) = -(3/4*log2(3/4)+1/4*log2(1/4))

=0.811

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

Using formula, Ent(no) = -(2/6*log2(2/6)+4/6*log2(4/6))

=0.918

Calculating weight average entropy of Traffic condition:

Using formula, Ent(engine)= 4/10*0.811+6/10*0.918

= 0.875

Calculating information gain of weather condition;

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:
Gini(D) = 1 - Σ(pi^2)

For Full Dataset :
p(Yes)=0.5, p(No)=0.5,
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

→ Gini(rain) = 1-[(1/3*1/3)+(2/3*2/3)]

= 0.4444

· Snow: Yes=2 No=1

→ Gini(snow) = 1-[(2/3*2/3)+(1/3*1/3)]

= 0.4444

· Clear: Yes=2 No=2

→ Gini (clear)= 1-[(2/4*2/4)+(2/4*2/4)]

= 0.5

Weight gini(weather)= 3/10*(0.444)+3/10*(0.444)+4/10*(0.5) =0.4667

Gain (weather) = gini(D)-weight gini(weather) =0.0333

- Road Condition

· Bad: Yes=3, No=1 → Gini(bad) = 1-[(3/4*3/4)+(1/4*1/4)] = 0.375

· Average: Yes=1, No=1 → Gini(average) = 1-[(1/2*1/2)+(1/2*1/2)] = 0.5

· Good: Yes=1, No=3 → Gini (good)= 1-[(1/4*1/4)+(3/4*3/4)] = 0.375

Weight gini(road)= 0.4(0.375)+0.2(0.5)+0.4(0.375) =0.40

Gain (road) = gini(D)-weight gini(road) = 0.5-0.40 = 0.10

- Traffic Condition

· High: Yes=2, No=2 → Gini(high) = 1-[(2/4*2/4)+(2/4*2/4)] = 0.5

· Normal: Yes=1, No=2 → Gini(normal) = 1-[(1/3*1/3)+(2/3*2/3)] = 0.4444

· Light: Yes=1, No=2 → Gini (light)= 1-[(1/3*1/3)+(2/3*2/3)] = 0.444

Weight gini(traffic)= 0.4(0.5)+0.3(0.4444)+0.3(0.4444)

= 0.4667

Gain (traffic) = gini(D)-weight gini(traffic) =0.0333

- **Engine Problem**

YES: Yes=3, No=1 → Gini(high) = 1-[(3/4*3/4)+(1/4*1/4)] = 0.375

NO: Yes=2, No=4 → Gini(normal) = 1-[(2/6*2/6)+(4/6*4/6)] = 0.4444

Weight gini(engine)= 0.4(0.375)+0.3(0.4444)

=0.417

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