Decision Tree – Entropy/Information Gain

Necessary Formulas:

1. Entropy, $E = -\sum p_i \log_2 p_i$; $i = 1$ to k, where k = number of classes.
2. Average Entropy, $E_{New} = (\sum - V_{ij} \log_2 V_{ij} + \sum S_j \log_2 S_j) / N$;
   $i = 1$ to k, where k = number of classes and
   $j = 1$ to n, where n = number of unique values for an attribute and
3. Information Gain, $I_g = E_{Start} - E_{New}$

Iteration 1 (For Selecting the Root Node)

We have 3 classes. So, The Value of Initial Entropy, E_{Start} will be:

$E_{Start} = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - p_3 \log_2 p_3$

There are 4 instances with classification 1, 5 instances with classification 2 and 15 instances with classification 3. So, $p_1 = (4/24)$, $p_2 = (5/24)$ and $p_3 = (15/24)$.

$$\begin{aligned} E_{Start} &= -(4/24) \log_2 (4/24) - (5/24) \log_2 (5/24) - (15/24) \log_2 (15/24) \\ &= 0.4308 + 0.4715 + 0.4238 \\ &= 1.3261 \text{ bits} \end{aligned}$$

Now, we need to calculate E_{New} for each of the attributes.

Frequency Table for Age

	Age = 1	Age = 2	Age = 3
Class 1	2	1	1
Class 2	2	2	1
Class 3	4	5	6
Sum	8	8	8

$$\begin{aligned} E_{New}\text{(Age)} = &(-2 \log_2 2 - 1 \log_2 1 - 1 \log_2 1 \\ &- 2 \log_2 2 - 2 \log_2 2 - 1 \log_2 1 \\ &- 4 \log_2 4 - 5 \log_2 5 - 6 \log_2 6 \\ &+ 8 \log_2 8 + 8 \log_2 8 + 8 \log_2 8) / 24 \\ = &1.2867 \end{aligned}$$

Frequency Table for SpecRx

	SpecRx = 1	SpecRx = 2
Class 1	3	1
Class 2	2	3
Class 3	7	8
Sum	12	12

$$\begin{aligned} E_{New}\text{(SpecRx)} = &(-3 \log_2 3 - 1 \log_2 1 - 2 \log_2 2 \\ &- 3 \log_2 3 - 7 \log_2 7 - 8 \log_2 8 \\ &+ 12 \log_2 12 + 12 \log_2 12)/24 \\ = &1.2866 \end{aligned}$$

Frequency Table for Astig

	Astig = 1	Astig = 2
Class 1	0	4
Class 2	5	0
Class 3	7	8
Sum	12	12

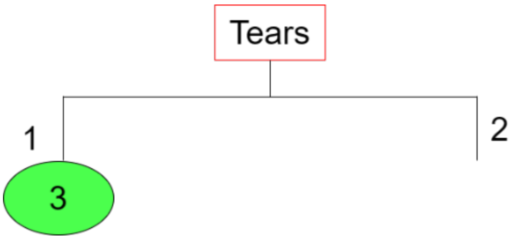
$$\begin{aligned} E_{New}\text{(Astig)} = &(-0 - 4 \log_2 4 - 5 \log_2 5 - 0 \\ &- 7 \log_2 7 - 8 \log_2 8 + 12 \log_2 12 \\ &+ 12 \log_2 12)/24 \\ = &0.9491 \end{aligned}$$

Frequency Table for Tears

	Tears = 1	Tears = 2
Class 1	0	4
Class 2	0	5
Class 3	12	3
Sum	12	12

$$\begin{aligned} E_{New}\text{(Tears)} = &(-0 - 4 \log_2 4 - 0 - 5 \log_2 5 \\ &- 12 \log_2 12 - 3 \log_2 3 + 12 \log_2 12 \\ &+ 12 \log_2 12)/24 \\ = &0.7773 \end{aligned}$$

Ig (Age) = $E_{Start} – E_{New}$ (Age) = 1.3261 – 1.2867 = 0.0394
Ig (SpecRx) = $E_{Start} – E_{New}$ (SpecRx) = 1.3261 – 1.2866 = 0.0395
Ig (Astig) = $E_{Start} – E_{New}$ (Astig) = 1.3261 – 0.9491 = 0.377
Ig (Tears) = $E_{Start} – E_{New}$ (Tears) = 1.3261 – 0.7773 = 0.5488



Iteration 2 (For Branch Tears = 2)

There are 4 instances with classification 1, 5 instances with classification 2 and 3 instances with classification 3. So, $p_1 = (4/12)$, $p_2 = (5/12)$ and $p_3 = (3/12)$.

$$\begin{aligned} E_{Start} &= – (4/12) \log_2 (4/12) – (5/12) \log_2 (5/12) – (3/12) \log_2 (3/12) \\ &= 0.5283 + 0.5263 + 0.5 \\ &= 1.5546 \text{ bits} \end{aligned}$$

Now, we need to calculate E_{New} for each of the attributes.

Frequency Table for Age

	Age = 1	Age = 2	Age = 3
Class 1	2	1	1
Class 2	2	2	1
Class 3	0	1	2
Sum	4	4	4

$$\begin{aligned} E_{New} \text{ (Age)} &= (– 2 \log_2 2 – 1 \log_2 1 – 1 \log_2 1 – 2 \log_2 2 – 2 \log_2 2 – 1 \log_2 1 – 0 – 1 \log_2 1 – 2 \log_2 2 + 4 \log_2 4 + 4 \log_2 4 + 4 \log_2 4)/12 \\ &= 1.3333 \end{aligned}$$

Frequency Table for SpecRx

	SpecRx = 1	SpecRx = 2
Class 1	3	1
Class 2	2	3
Class 3	1	2
Sum	6	6

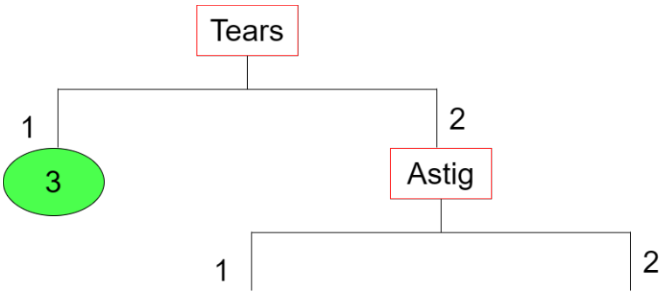
$$\begin{aligned} E_{New} \text{ (SpecRx)} &= (– 3 \log_2 3 – 1 \log_2 1 – 2 \log_2 2 – 3 \log_2 3 – 1 \log_2 1 – 2 \log_2 2 + 6 \log_2 6 + 6 \log_2 6) / 12 \\ &= 1.4592 \end{aligned}$$

Frequency Table for Astig

	Astig = 1	Astig = 2
Class 1	0	4
Class 2	5	0
Class 3	1	2
Sum	6	6

$$\begin{aligned} E_{New} \text{ (Astig)} &= (0 – 4 \log_2 4 – 5 \log_2 5 – 0 – 1 \log_2 1 – 2 \log_2 2 + 6 \log_2 6 + 6 \log_2 6)/12 \\ &= 0.7842 \end{aligned}$$

Ig (Age) = $E_{Start} – E_{New}$ (Age) = 1.5546 – 1.3333 = 0.2213
Ig (SpecRx) = $E_{Start} – E_{New}$ (SpecRx) = 1.5546 – 1.4592 = 0.0954
Ig (Astig) = $E_{Start} – E_{New}$ (Astig) = 1.5546 – 0.7842 = 0.7704



Iteration 3 (For Branch Astig = 1)

There are 5 instances with classification 2 and 1 instance with classification 3. So, $p_1 = (5/6)$ and $p_2 = (1/6)$.

$$
\begin{aligned}
E_{Start} &= -(5/6)\log_2(5/6) - (1/6)\log_2(1/6) \\
&= 0.2192 + 0.4308 \\
&= 0.65 \text{ bits}
\end{aligned}
$$

Now, we need to calculate E_{New} for each of the attributes.

Frequency Table for Age

	Age = 1	Age = 2	Age = 3
Class 1	0	0	0
Class 2	2	2	1
Class 3	0	0	1
Sum	2	2	2

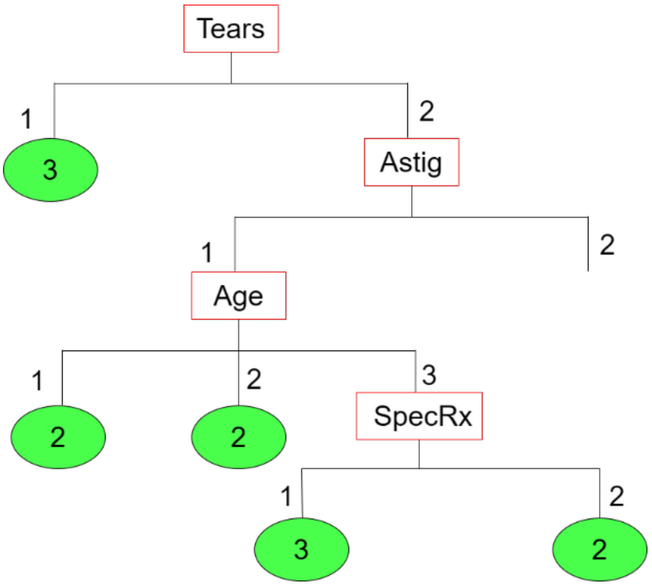
$$
\begin{aligned}
E_{New}(\text{Age}) &= (0 - 0 - 0 - 2\log_2 2 - 2\log_2 2 - 1\log_2 1 - 0 - 0 - 1\log_2 1 + 2\log_2 2 + 2\log_2 2 + 2\log_2 2)/6 \\
&= 0.3333
\end{aligned}
$$

Frequency Table for SpecRx

	SpecRx = 1	SpecRx = 2
Class 1	0	0
Class 2	2	3
Class 3	1	0
Sum	3	3

$$
\begin{aligned}
E_{New}(\text{SpecRx}) &= (0 - 0 - 2\log_2 2 - 3\log_2 3 - 1\log_2 1 - 0 + 3\log_2 3 + 3\log_2 3)/6 \\
&= 0.4592
\end{aligned}
$$

$Ig(\text{Age}) = E_{Start} - E_{New}(\text{Age}) = 0.6500 - 0.3333 = 0.3167$
$Ig(\text{SpecRx}) = E_{Start} - E_{New}(\text{SpecRx}) = 0.6500 - 0.4592 = 0.1908$



Iteration 4 (For Branch Astig = 2)

There are 4 instances with classification 1 and 2 instance with classification 3. So, $p_1 = (4/6)$ and $p_2 = (2/6)$.

$$
\begin{aligned}
E_{Start} &= -(4/6)\log_2(4/6) - (2/6)\log_2(2/6) \\
&= 0.3900 + 0.5283 \\
&= 0.9183 \text{ bits}
\end{aligned}
$$

Now, we need to calculate E_{New} for each of the attributes.

Frequency Table for Age

	Age = 1	Age = 2	Age = 3
Class 1	2	1	1
Class 2	0	0	0
Class 3	0	1	1
Sum	2	2	2

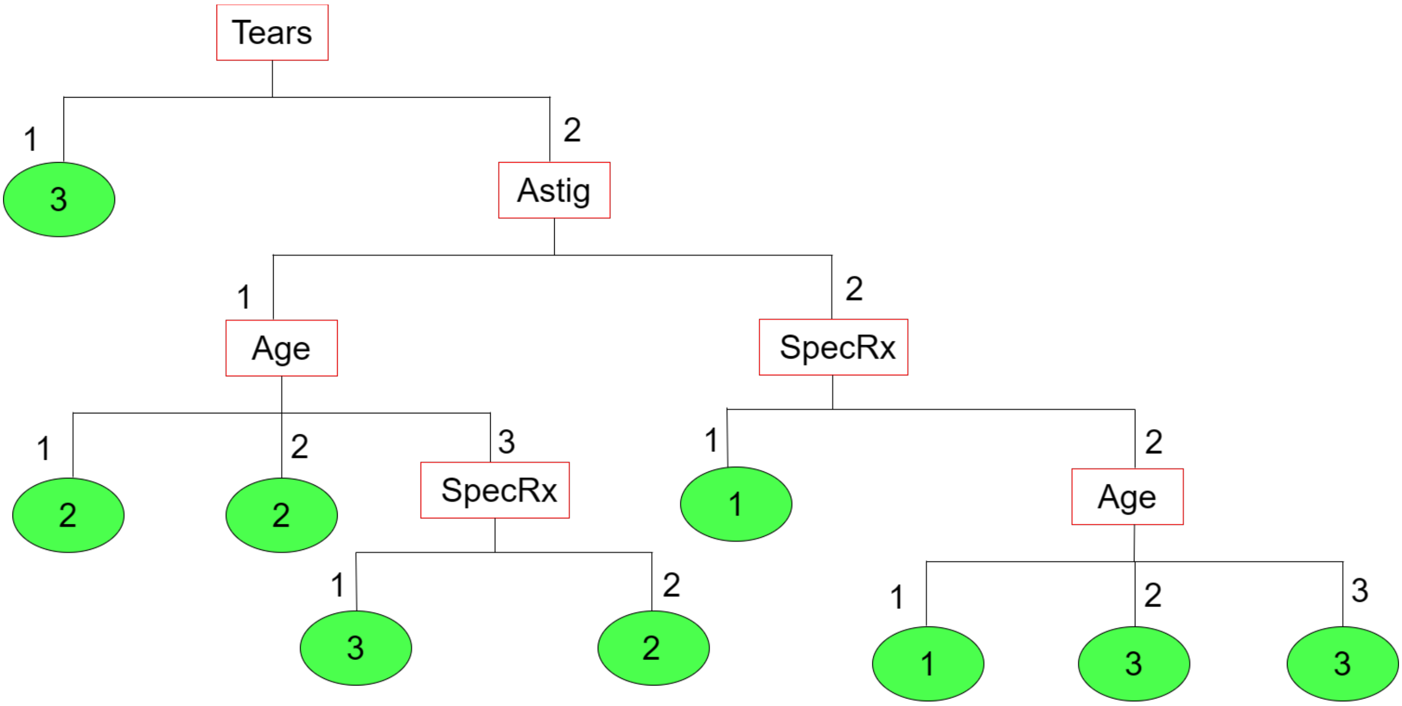
$$
\begin{aligned}
E_{New}(\text{Age}) &= (-2\log_2 2 - 1\log_2 1 - 1\log_2 1 - 0 - 0 - 0 - 0 - 1\log_2 1 - 1\log_2 1 + 2\log_2 2 + 2\log_2 2 + 2\log_2 2)/6 \\
&= 0.6667
\end{aligned}
$$

Frequency Table for SpecRx

	SpecRx = 1	SpecRx = 2
Class 1	3	1
Class 2	0	0
Class 3	0	2
Sum	3	3

$$
\begin{aligned}
E_{New}(\text{SpecRx}) &= (-3\log_2 3 - 1\log_2 1 - 0 - 0 - 0 - 2\log_2 2 + 3\log_2 3 + 3\log_2 3)/6 \\
&= 0.4592
\end{aligned}
$$

$Ig(\text{Age}) = E_{Start} - E_{New}(\text{Age}) = 0.9183 - 0.6667 = 0.2516$
$Ig(\text{SpecRx}) = E_{Start} - E_{New}(\text{SpecRx}) = 0.9183 - 0.4592 = 0.4591$


Tears
1
3
2
Astig
1
Age
2
SpecRx
1
2
2
2
3
SpecRx
1
1
2
Age
1
3
2
2
1
1
2
3
3
3