Decision Tree

Ejercicio 1: Una persona registra los 13 espectáculos de comedia de la ciudad en un día con las características del comediante en cada espectáculo (años, experiencia, ranking y nacionalidad) y también si asistió o no al evento.

Age	Experience	Rank	Nationality	Nationality	Attended	Attended
36	10	9	UK	1	NO	0
42	12	4	USA	2	NO	0
23	4	6	OTHER	0	NO	0
52	4	4	USA	2	NO	0
43	21	8	USA	2	YES	1
44	14	5	UK	1	NO	0
66	3	7	OTHER	0	YES	1
35	14	9	UK	1	YES	1
52	13	7	OTHER	0	YES	1
35	5	9	OTHER	0	YES	1
24	3	5	USA	2	NO	0
18	3	7	UK	1	YES	1
45	9	9	UK	1	YES	1

Step 0: Proceso de **Codificacion Ordinal** (***Ordinal encoding***) para Nationality y **Codificación Binaria** (***Binary enconding***) para Attended.

Step 1: Calcular **Índice de Gini** (***Gini Index***) y **Entropía de Shannon** (***Shannon Entropy***) del nodo raíz [*todo el dataset*]

$$\begin{cases} No\ attended = 6 \\ Attended = 7 \end{cases} \rightarrow N = 13$$

$$\cdot\ Gini\ Index = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = \frac{84}{169} \approx \mathbf{0.497} \quad \cdot\ Shannon\ Entropy = -\frac{7}{13} log_2\left(\frac{7}{13}\right) - \frac{6}{13}\log_2\left(\frac{6}{13}\right) \approx \mathbf{0.99}$$

Step 2: Calcular los puntos de corte de cada variable y cuál es el punto de corte qué tiene mayor eficiencia (*menor Gini*)

Age

Dataset ordenado

Age_n	*Attended*
18	1
23	0
24	0
35	1
35	1
36	0
42	0
43	1
44	0
45	1
52	1
52	0
66	1

Hay que buscar el punto de corte con menor promedio de Gini
Hay que ir probando y observar

Punto de corte 35.5:

$Gini_{Left} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$

$Gini_{Right} = 1 - \left(\frac{4}{8}\right)^2 - \left(\frac{4}{8}\right)^2 = 0.5$

$Gini_a = \frac{5}{13} \cdot 0.48 + \frac{8}{13} \cdot 0.5 \approx 0.4923$

Punto de corte 39.5:

$Gini_{Left} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$

$Gini_{Right} = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$

$Gini_a = \frac{6}{13} \cdot 0.5 + \frac{7}{13} \cdot 0.489 \approx 0.4945$

Experience

Dataset ordenado

Experience_n	*Attended*
3	0
3	1
3	1
4	0
4	0
5	1
9	1
10	0
12	0
13	1
14	0
14	1
21	1

Hay que buscar el punto de corte con menor promedio de Gini
Hay que ir probando y observar

Punto de corte 9.5:

$Gini_{Left} = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$

$Gini_{Right} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$

$Gini_a = \frac{7}{13} \cdot 0.489 + \frac{6}{13} \cdot 0.5 \approx 0.4945$

Rank

Dataset ordenado

Rank_n	*Attended*
4	0
4	0
5	0
5	0
6	0
7	1
7	1
7	1
8	1
9	0
9	1
9	1
9	1

Hay que buscar el punto de corte con menor promedio de Gini
Hay que ir probando y observar

Punto de corte 4.5:

$Gini_{Left} = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{0}\right)^2 = 0$

$Gini_{Right} = 1 - \left(\frac{4}{11}\right)^2 - \left(\frac{7}{11}\right)^2 = 0.4628$

$Gini_a = \frac{2}{13} \cdot 0 + \frac{11}{13} \cdot 0.4628 \approx 0.39$

Punto de corte 6.5:

$Gini_{Left} = 1 - \left(\frac{5}{5}\right)^2 - \left(\frac{0}{5}\right)^2 = 0$

$Gini_{Right} = 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = 0.21875$

$Gini_a = \frac{5}{13} \cdot 0 + \frac{8}{13} \cdot 0.21875 \approx 0.13$

Nationality

Dataset ordenado

Nationality_n	*Attended*
0	0
0	1
0	1
0	1
1	0
1	0
1	1
1	1
1	1
2	0
2	0
2	0
2	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 2 puntos (<=0.5 o <=1.5)

Punto de corte 0.5:

$Gini_{Left} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$

$Gini_{Right} = 1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2 = 0.49$

$Gini_a = \frac{4}{13} \cdot 0.375 + \frac{9}{13} \cdot 0.49 \approx 0.4572$

Punto de corte 1.5:

$Gini_{Left} = 1 - \left(\frac{3}{9}\right)^2 - \left(\frac{6}{9}\right)^2 = 0.\hat{4}$

$Gini_{Right} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$

$Gini_{Av} = \frac{9}{13} \cdot 0.\hat{4} + \frac{4}{13} \cdot 0.5 \approx 0.4615$

*El árbol eligió la **columna rank con un punto de corte 6.5** porque es donde se alcanza el menor Índice de Gini, logrando la mejor separación de clases. En el lado izquierdo (<= 6.5), el Gini es 0, lo que indica que todas las muestras pertenecen a una sola clase, por lo que este nodo se cierra. En el lado derecho (> 6.5), el Gini es 0.21875, lo que refleja una mezcla parcial de clases, permitiendo seguir dividiendo este nodo para mejorar la separación.*

Step 3: Recortar el dataset (*valores descartados*)

Age	Experience	Rank	Nationality	Nationality	Attended	Attended
36	10	9	UK	1	NO	0
42	12	4	USA	2	NO	0
23	4	6	OTHER	0	NO	0
52	4	4	USA	2	NO	0
43	21	8	USA	2	YES	1
44	14	5	UK	1	NO	0
66	3	7	OTHER	0	YES	1
35	14	9	UK	1	YES	1
52	13	7	OTHER	0	YES	1
35	5	9	OTHER	0	YES	1
24	3	5	USA	2	NO	0
18	3	7	UK	1	YES	1
45	9	9	UK	1	YES	1

Step 4: Calcular los puntos de corte de cada variable y cuál es el punto de corte qué tiene mayor eficiencia (*menor Gini*)

Age

Dataset ordenado

Age_n	Attended
18	1
35	1
35	1
36	0
43	1
45	1
52	1
66	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 2 puntos

Punto de corte 35.5:

$$Gini_{Left} = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gini_{Right} = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$$

$$Gini_a = \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0.32 = 0.2$$

Punto de corte 39.5:

$$Gini_{Left} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini_{Right} = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$Gini_a = \frac{4}{8} \cdot 0.375 + \frac{4}{8} \cdot 0 = 0.1875$$

Experience

Dataset ordenado

Experience_n	Attended
3	1
3	1
5	1
9	1
10	0
13	1
14	1
21	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 2 puntos

Punto de corte 9.5:

$$Gini_{Left} = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$Gini_{Right} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini_a = \frac{4}{8} \cdot 0 + \frac{4}{8} \cdot 0.375 = 0.1875$$

Punto de corte 11.5:

$$Gini_{Left} = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$$

$$Gini_{Right} = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gini_a = \frac{5}{8} \cdot 0.32 + \frac{3}{8} \cdot 0 = 0.2$$

Rank

Dataset ordenado

Rank_n	Attended
7	1
7	1
7	1
8	1
9	0
9	1
9	1
9	1

Hay que buscar el punto de corte con menor promedio de Gini
Solo puede haber 1 punto

Punto de corte 8.5:

$$Gini_{Left} = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$Gini_{Right} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini_a = \frac{4}{8} \cdot 0 + \frac{4}{8} \cdot 0.375 = 0.1875$$

Nationality

Dataset ordenado

Nationality_n	Attended
0	1
0	1
0	1
1	0
1	1
1	1
1	1
2	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 2 puntos

Punto de corte 0.5:

$$Gini_{Left} = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gini_{Right} = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$$

$$Gini_a = \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0.32 = 0.2$$

Punto de corte 1.5:

$$Gini_{Left} = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{6}{7}\right)^2 = 0.2449$$

$$Gini_{Right} = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$$

$$Gini_a = \frac{7}{8} \cdot 0.2449 + \frac{1}{8} \cdot 0 \approx 0.2142$$

Hubo un triple emplate entre las columnas age, experience y rank, todas con un Índice de Gini de 0.1875 *en su punto de corte más óptimo.*

Se podría seguir desarrollando el árbol con cualquiera de las 3 opciones, pero según las reglas internas de Sklearn, ***se seleccionó la columna age para la división*** *(igual que en el archivo de Python), donde en el lado izquierdo (<= 39.5), el Índice de Gini resultante fue 0.375, indicando una mezcla parcial de clases, por lo que este nodo seguirá desarrollándose. Y en el lado derecho (> 39.5), el Gini fue 0, lo que significa que todas las muestras pertenecen a una sola clase, cerrando ese nodo.*

Step 5: Recortar el dataset (*valores descartados*)

Age	Experience	Rank	Nationality	Nationality	Attended	Attended
36	10	9	UK	1	NO	0
42	12	4	USA	2	NO	0
23	4	6	OTHER	0	NO	0
52	4	4	USA	2	NO	0
43	21	8	USA	2	YES	1
44	14	5	UK	1	NO	0
66	3	7	OTHER	0	YES	1
35	14	9	UK	1	YES	1
52	13	7	OTHER	0	YES	1
35	5	9	OTHER	0	YES	1
24	3	5	USA	2	NO	0
18	3	7	UK	1	YES	1
45	9	9	UK	1	YES	1

Step 6: Calcular los puntos de corte de cada variable y cuál es el punto de corte qué tiene mayor eficiencia (*menor Gini*)

Age

Dataset ordenado

Age_n	*Attended*
18	1
35	1
35	1
36	0

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 1 punto

Punto de corte 35.5:

$Gini_{Left} = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$

$Gini_{Right} = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$

$Gini_a = \frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0$

Experience

Dataset ordenado

Experience_n	*Attended*
3	1
5	1
10	0
14	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 2 puntos

Punto de corte 7.5:

$Gini_{Left} = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$

$Gini_{Right} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$

$Gini_a = \frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 0.5 = 0.25$

Punto de corte 12:

$Gini_{Left} = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.\hat{4}$

$Gini_{Right} = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$

$Gini_a = \frac{1}{4} \cdot 0.\hat{4} + \frac{3}{4} \cdot 0 = 0.\hat{3}$

Rank

Dataset ordenado

Rank_n	*Attended*
7	1
9	0
9	1
9	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 1 punto

Punto de corte 8:

$Gini_{Left} = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$

$Gini_{Right} = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.\hat{4}$

$Gini_a = \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0.\hat{4} = 0.\hat{3}$

Nationality

Dataset ordenado

Nationality_n	*Attended*
0	0
0	1
0	1
2	1

Hay que buscar el punto de corte con menor promedio de Gini
Sólo puede haber 1 punto

Punto de corte 1.0:

$Gini_{Left} = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.\hat{4}$

$Gini_{Right} = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$

$Gini_a = \frac{3}{4} \cdot 0.\hat{4} + \frac{1}{4} \cdot 0 = 0.\hat{3}$

En la columna age se encontró un punto de corte en 35.5, donde el Índice de Gini es 0 tanto en el lado izquierdo (<=35.5) como en el lado derecho (> 35.5). Esto indica que ambos grupos son completamente puros, es decir, todas las muestras de cada grupo pertenecen a una sola clase. Por lo tanto, no es necesario seguir dividiendo, y el árbol de decisión se da por finalizado.