Escuela Politécnica Superior (EPS)

UNIVERSIDAD D CÓRDOBA
Departamento de Informática y Análisis Numérico

Práctica 4.
Ejemplo de creación de un árbol de decisión

Juan Carlos Fernández Caballero (jfcaballero@uco.es)

Introducción al Aprendizaje Automático (IAA)
3º de Grado de Ingeniería Informática
Especialidad en Computación
mayo de 2020

GRUPO DE INVESTIGACIÓN AYRNA
APRENDIZAJE **Y** **R**EDES **N**EURONALES **A**RTIFICIALES
uco.es/ayrna

Entrenamiento de los árboles de decisión

Para crear y entrenar un árbol se usa el conjunto de entrenamiento o ***training***, y se realizan **procesos iterativos** (***Greedy* o Voraz**, **Divide y Vencerás**, etc) sobre los datos, que van obteniendo ramas y subconjuntos de datos homogéneos respecto a la clase. Estos procesos iterativos se basan en la **ganancia de información**, que depende de la **Entropía**.

- **Entropía:** Medida de incertidumbre, relacionada con:
  - **1) Pureza:** Cómo de cerca está un conjunto de pertenecer a una misma clase.
  - **2) Impureza (desorden):** Cómo de cerca está un conjunto de la incertidumbre total.
- La Entropía es una medida:
  - **1) Directamente proporcional** a la impureza, incertidumbre, irregularidad.
  - **2) Inversamente proporcional** a la pureza, certidumbre, redundancia.
- **Objetivo:** Minimizar la Entropía (desorden o error)!

Entrenamiento de los árboles de decisión

A partir del conjunto de ***entrenamiento*** y de forma general, los pasos son los siguientes:

1. Elección del **nodo raíz**: Se escogerá aquél que provoque una mejor separación de las clases.
   Se hace uso de la **Ganancia de información**, que a su vez usa la **Entropía de la clase** y la **Entropía de la clase condicionada a un valor**.
   - Cantidad de información mutua o ganancia de información:

$$I(C, X_i) = H(C) - H(C|X_i)$$

   - Entropía de una variable:

$$H(C) = -\sum_{c=1}^{n} p(c) \log_2 p(c)$$

Entrenamiento de los árboles de decisión

- Entropía de una variable condicionada a un valor:

$$H(C|X) = -\sum_{i=1}^{n} p(x|c)\log_2 p(c|x) = -\sum_{c}\sum_{x} p(x|c)\log_2 p(c|x)$$

- Se elige como primer nodo aquél que maximiza la expresión: $I(C, X_i)$, es decir, **minimiza la Entropía (error o desorden).**

2. Cuando se dividen las ramas, se crean ***nuevos subconjuntos de training*** para cada una de ellas.
3. Se repite el proceso (puntos 1 y 2) para cada rama generada.

Ejemplo de entrenamiento (construcción) de un árbol de decisión

A partir de la siguiente tabla con datos de ***training***, usando las expresiones de **Entropía** y **Cantidad de información**, cree un árbol de decisión.

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

Cálculo de la entropía de la variable *decision (clase)*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

1. Elección del **nodo raíz**.
   - Paso 1: Calcular la Entropía de la clase:

$$H(C) = -\sum_{c=1}^{n} p(c) \log_2 p(c)$$

$$H(Decision) = -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$$

$$H(Decision) = -(\frac{6}{10} \log_2 \frac{6}{10}) - (\frac{4}{10} \log_2 \frac{4}{10}) = 0{,}971$$

Cálculo de la entropía condicionada para la *variable Outlook*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

- Paso 2 : Calcular la Entropía de clase condicionada a un valor:

$$H(C|X) = -\sum_{i=1}^{n} p(x|c)\log_2 p(c|x) =$$

$$-\sum_{c}\sum_{x} p(x,c)\log_2 p(c|x)$$

$$\begin{aligned} H(Decision|Outlook) = &- p(Rain, Yes)\log_2 p(Yes|Rain) - p(Rain, No)\log_2 p(No|Rain) \\ &- p(Sun, Yes)\log_2 p(Yes|Sun) - p(Sun, No)\log_2 p(No|Sun) \\ &- p(Overcast, Yes)\log_2 p(Yes|Overcast) - p(Overcast, No)\log_2 p(No|Overcast) \end{aligned}$$

$$\begin{aligned} H(Decision|Outlook) = &- (\frac{3}{10}\log_2\frac{3}{4}) - (\frac{1}{10}\log_2\frac{1}{4}) \\ &- (\frac{1}{10}\log_2\frac{1}{4}) - (\frac{3}{10}\log_2\frac{3}{4}) \\ &- (\frac{2}{10}\log_2\frac{2}{2}) - (\frac{0}{10}\log_2\frac{0}{2}) = 0{,}325 + 0{,}325 + 0 = 0{,}65 \end{aligned}$$

Cálculo de la entropía condicionada para la *variable Temperature*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

$$\begin{aligned} H(Decision|\,Temperature) = &- p(Mild, Yes)\log_2 p(Yes|Mild) - p(Mild, No)\log_2 p(No|Mild) \\ &- p(Cool, Yes)\log_2 p(Yes|Cool) - p(Cool, No)\log_2 p(No|Cool) \\ &- p(Hot, Yes)\log_2 p(Yes|Hot) - p(Hot, No)\log_2 p(No|Hot) \end{aligned}$$

$$\begin{aligned} H(Decision|\,Temperature) = &- (\frac{2}{10}\log_2\frac{2}{3}) - (\frac{1}{10}\log_2\frac{1}{3}) \\ &- (\frac{3}{10}\log_2\frac{3}{4}) - (\frac{1}{10}\log_2\frac{1}{4}) \\ &- (\frac{1}{10}\log_2\frac{1}{3}) - (\frac{2}{10}\log_2\frac{2}{3}) = 0{,}275 + 0{,}325 + 0{,}275 = 0{,}875 \end{aligned}$$

Cálculo de la entropía condicionada para las *variables Humidity y Wind*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

$$H(Decision|Humidity) = -\ p(Normal, Yes)\log_2 p(Yes|Normal) - p(Normal, No)\log_2 p(No|Normal)$$
$$- p(High, Yes)\log_2 p(Yes|High) - p(High, No)\log_2 p(No|High)$$

$$H(Decision|Humidity) = -\ (\frac{4}{10}\log_2\frac{4}{5}) - (\frac{1}{10}\log_2\frac{1}{5})$$
$$-\ (\frac{2}{10}\log_2\frac{2}{5}) - (\frac{3}{10}\log_2\frac{3}{5}) = 0{,}361 + 0{,}485 = 0{,}846$$

$$H(Decision|Wind) = -\ p(Weak, Yes)\log_2 p(Yes|Weak) - p(Weak, No)\log_2 p(No|Weak)$$
$$- p(Strong, Yes)\log_2 p(Yes|Strong) - p(Strong, No)\log_2 p(No|Strong)$$

$$H(Decision|Wind) = -\ (\frac{5}{10}\log_2\frac{5}{7}) - (\frac{2}{10}\log_2\frac{2}{7})$$
$$-\ (\frac{1}{10}\log_2\frac{1}{3}) - (\frac{2}{10}\log_2\frac{2}{3}) = 0{,}604 + 0{,}275 = 0{,}879$$

Cálculo de la ganancia de información

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

- Cantidad de información mutua o ganancia de información (a maximizar):

$$I(C, X_i) = H(C) - H(C|X_i)$$

$$\mathbf{I(Decision,Outlook) = H(Decision) - H(Decision|Outlook) = 0{,}971 - 0{,}650 = 0{,}321}$$

$$I(Decision, Temperature) = H(Decision) - H(Decision|Temperature) = 0{,}971 - 0{,}875 = 0{,}096$$

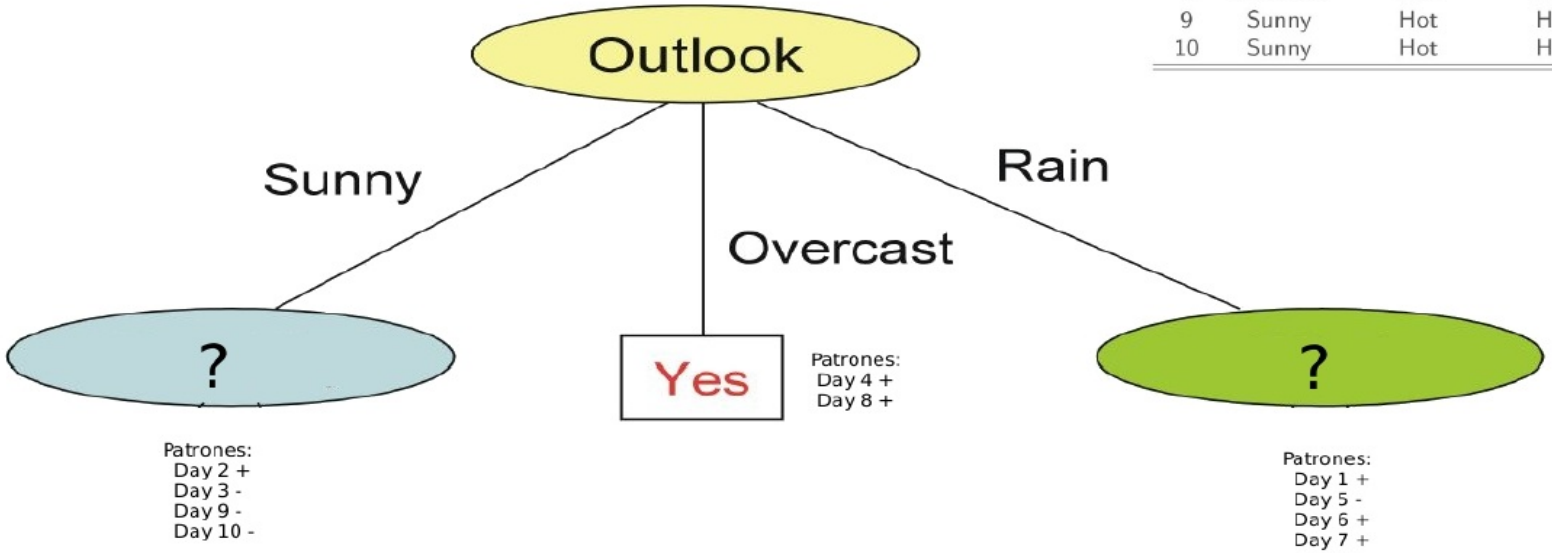
$$I(Decision, Humidity) = H(Decision) - H(Decision|Humidity) = 0{,}971 - 0{,}846 = 0{,}125$$

$$I(Decision, Wind) = H(Decision) - H(Decision|Wind) = 0{,}971 - 0{,}879 = 0{,}092$$

- Por tanto, el atributo que maximiza la ganancia de información es ***Outlook***

Árbol parcial obtenido

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
4	Overcast	Cool	Normal	Strong	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes
8	Overcast	Hot	High	Weak	Yes
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No



Nuevos conjuntos de datos

2. Se crean ***nuevos subconjuntos de training*** para cada rama.

Conjunto de datos para la rama ***Sunny***

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

Conjunto de datos para la rama ***Rain***

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes

Cálculo de la entropía de la *variable decision (clase), rama Sunny*

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

1. Elección del **nodo raíz rama Sunny**.
   - Entropía de la clase:

$$H(C) = -\sum_{c=1}^{n} p(c) \log_2 p(c)$$

$$H(Decision) = -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$$

$$H(Decision) = -(\frac{1}{4} \log_2 \frac{1}{4}) - (\frac{3}{4} \log_2 \frac{3}{4}) = 0{,}811$$

Cálculo de la entropía condicionada para la *variable Temperature*

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

- Entropía de la clase condicionada a un valor:

$$H(C|X) = -\sum_{i=1}^{n} p(x|c)\log_2 p(c|x) =$$

$$-\sum_{c}\sum_{x} p(x,c)\log_2 p(c|x)$$

$$\begin{aligned} H(Decision|\,Temperature) = &- p(Mild, Yes)\log_2 p(Yes|Mild) - p(Mild, No)\log_2 p(No|Mild) \\ &- p(Cool, Yes)\log_2 p(Yes|Cool) - p(Cool, No)\log_2 p(No|Cool) \\ &- p(Hot, Yes)\log_2 p(Yes|Hot) - p(Hot, No)\log_2 p(No|Hot) \end{aligned}$$

$$\begin{aligned} H(Decision|\,Temperature) = &- (\frac{0}{4}\log_2\frac{0}{1}) - (\frac{1}{4}\log_2\frac{1}{1}) \\ &- (\frac{1}{4}\log_2\frac{1}{1}) - (\frac{0}{4}\log_2\frac{0}{1}) \\ &- (\frac{0}{4}\log_2\frac{0}{2}) - (\frac{2}{4}\log_2\frac{2}{2}) = 0+0+0=0 \end{aligned}$$

Cálculo de la entropía condicionada para las *variabless Humidity y Wind*

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

$$H(Decision|Humidity) = -\ p(Normal, Yes)\log_2 p(Yes|Normal) - p(Normal, No)\log_2 p(No|Normal)$$
$$- p(High, Yes)\log_2 p(Yes|High) - p(High, No)\log_2 p(No|High)$$

$$H(Decision|Humidity) = -\ (\frac{1}{4}\log_2\frac{1}{1}) - (\frac{0}{4}\log_2\frac{0}{1})$$
$$- (\frac{0}{4}\log_2\frac{0}{3}) - (\frac{3}{4}\log_2\frac{3}{3}) = 0 + 0 = 0$$

$$H(Decision|Wind) = -\ p(Weak, Yes)\log_2 p(Yes|Weak) - p(Weak, No)\log_2 p(No|Weak)$$
$$- p(Strong, Yes)\log_2 p(Yes|Strong) - p(Strong, No)\log_2 p(No|Strong)$$

$$H(Decision|Wind) = -\ (\frac{1}{4}\log_2\frac{1}{3}) - (\frac{2}{4}\log_2\frac{2}{3})$$
$$- (\frac{0}{4}\log_2\frac{0}{1}) - (\frac{1}{4}\log_2\frac{1}{1}) = 0{,}689 + 0 = 0{,}689$$

Cálculo de la ganancia de información

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No

- Cantidad de información mutua o ganancia de información:

$$I(C, X_i) = H(C) - H(C|X_i)$$

$$I(Decision, Temperature) = H(Decision) - H(Decision|Temperature) = 0{,}811 - 0 = 0{,}811$$

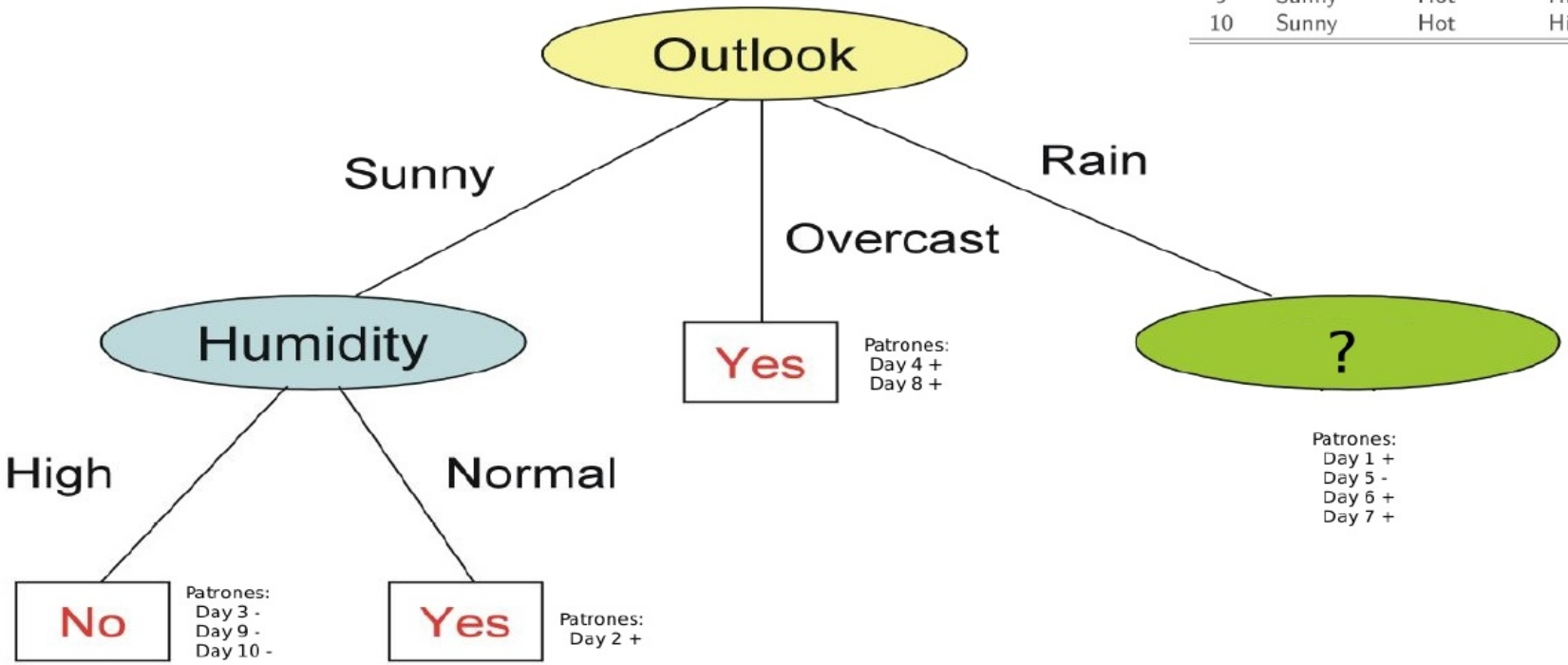
$$\mathbf{I(Decision,Humidity) = H(Decision) - H(Decision|Humidity) = 0.811 - 0 = 0.811}$$

$$I(Decision, Wind) = H(Decision) - H(Decision|Wind) = 0{,}811 - 0{,}689 = 0{,}122$$

- En este caso los atributos que maximizan la ganancia de información son ***Temperature*** y ***Humidity***.
- Se selecciona ***Humidity*** porque con sólo dos ramas clasifica a todos los patrones.

Nuevo árbol parcial obtenido

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Cool	Normal	Weak	Yes
3	Sunny	Mild	High	Weak	No
9	Sunny	Hot	High	Strong	No
10	Sunny	Hot	High	Weak	No



Cálculo de la entropía de la *variable decision (clase), rama Rain*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes

1. Elección del **nodo raíz rama Rain**.

- Entropía de la clase:

$$H(C) = -\sum_{c=1}^{n} p(c) \log_2 p(c)$$

$$H(Decision) = -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$$

$$H(Decision) = -(\frac{3}{4} \log_2 \frac{3}{4}) - (\frac{1}{4} \log_2 \frac{1}{4}) = 0{,}811$$

Cálculo de la entropía condicionada para la *variable Temperature*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes

- Entropía de la clase condicionada a un valor:

$$H(C|X) = -\sum_{i=1}^{n} p(x|c)\log_2 p(c|x) =$$

$$-\sum_{c}\sum_{x} p(x,c)\log_2 p(c|x)$$

$$H(Decision|Temperature) = -\,p(Mild, Yes)\log_2 p(Yes|Mild) - p(Mild, No)\log_2 p(No|Mild)$$
$$-\,p(Cool, Yes)\log_2 p(Yes|Cool) - p(Cool, No)\log_2 p(No|Cool)$$

$$H(Decision|Temperature) = -\,(\frac{2}{4}\log_2\frac{2}{2}) - (\frac{0}{4}\log_2\frac{0}{2})$$
$$-\,(\frac{1}{4}\log_2\frac{1}{2}) - (\frac{1}{4}\log_2\frac{1}{2}) = 0 + 0{,}5 = 0{,}5$$

Cálculo de la entropía condicionada para la *variables Humidity y Wind*

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes

$$H(Decision|Humidity) = -\,p(Normal, Yes)\log_2 p(Yes|Normal) - p(Normal, No)\log_2 p(No|Normal) - p(High, Yes)\log_2 p(Yes|High) - p(High, No)\log_2 p(No|High)$$

$$H(Decision|Humidity) = -\,(\frac{2}{4}\log_2\frac{2}{3}) - (\frac{1}{4}\log_2\frac{1}{3}) - (\frac{1}{4}\log_2\frac{1}{1}) - (\frac{0}{4}\log_2\frac{0}{1}) = 0{,}689 + 0 = 0{,}689$$

$$H(Decision|Wind) = -\,p(Weak, Yes)\log_2 p(Yes|Weak) - p(Weak, No)\log_2 p(No|Weak) - p(Strong, Yes)\log_2 p(Yes|Strong) - p(Strong, No)\log_2 p(No|Strong)$$

$$H(Decision|Wind) = -\,(\frac{3}{4}\log_2\frac{3}{3}) - (\frac{0}{4}\log_2\frac{0}{3}) - (\frac{0}{4}\log_2\frac{0}{1}) - (\frac{1}{4}\log_2\frac{1}{1}) = 0 + 0 = 0$$

Cálculo de la ganancia de información

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rain	Mild	Normal	Weak	Yes
5	Rain	Cool	Normal	Strong	No
6	Rain	Cool	Normal	Weak	Yes
7	Rain	Mild	High	Weak	Yes

- Cantidad de información mutua o ganancia de información:

$$I(C, X_i) = H(C) - H(C|X_i)$$

$$I(Decision, Temperature) = H(Decision) - H(Decision|Temperature) = 0{,}811 - 0{,}5 = 0{,}311$$

$$I(Decision, Humidity) = H(Decision) - H(Decision|Humidity) = 0{,}811 - 0{,}689 = 0{,}122$$

$$\mathbf{I(Decision,Wind) = H(Decision) - H(Decision|Wind) = 0.811 - 0 = 0.811}$$

- Por tanto, el atributo que maximiza la ganancia de información es ***Wind***

Árbol definitivo obtenido

Árbol construido a partir del conjunto de entrenamiento o ***training***.