Minería de Datos
IIC2433

Árboles de decisión

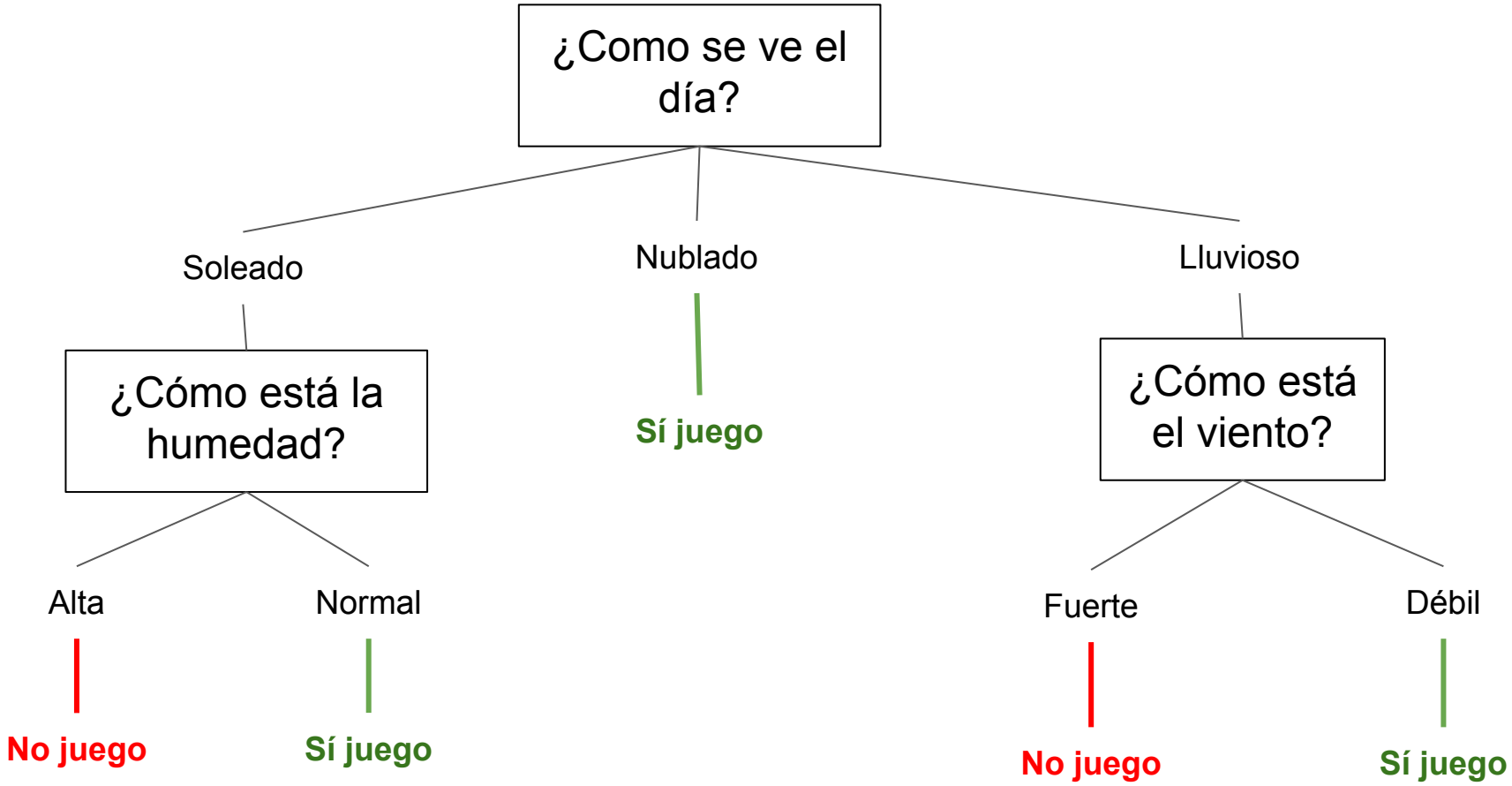
Vicente Domínguez

¿Qué veremos esta clase?

- Árboles de decisión

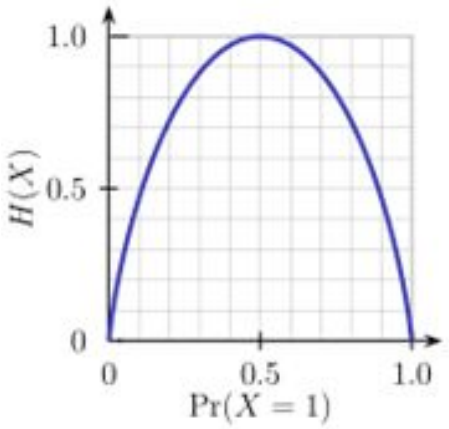
Árbol de decisión

Árbol para decidir si jugar tenis o no



Árbol de decisión

- Los nodos del árbol representan variables, las ramas representan valores de las variables que permiten clasificar
- Las hojas del árbol corresponden a la clasificación
- En la construcción del árbol, se testea una variable a la vez, usando el concepto de Entropía (número de bits necesarios para transmitir un mensaje - o - nivel de incerteza respecto a un evento)



$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

¿Cuál de los siguientes escenarios es más incierto?

- Predicción del tiempo
  - Llueve: 50%, Sol: 50%
  - Llueve: 100%, Sol: 0%
- Calcular la entropía de cada uno:

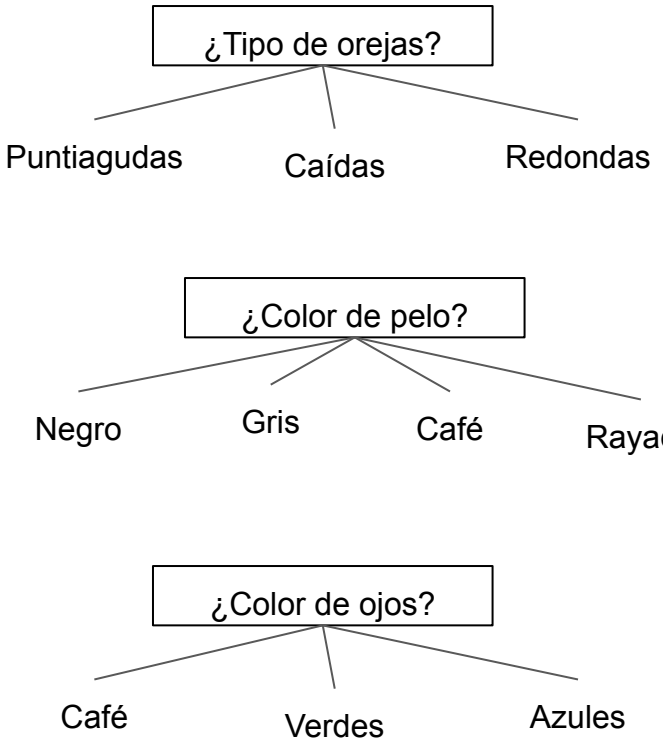
$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Árbol de decisión

Construcción a partir de un dataset de entrenamiento

ID	Tipo orejas	Color pelo	Color ojos	**Clase**
1	Puntiagudas	Negro	Café	**Gato**
2	Caídas	Gris	Verdes	**Perro**
3	Redondas	Café	Azules	**Perro**
4	Puntiagudas	Rayado	Verdes	**Gato**
5	Caídas	Negro	Café	**Perro**

¿Qué atributo elegimos para comenzar?

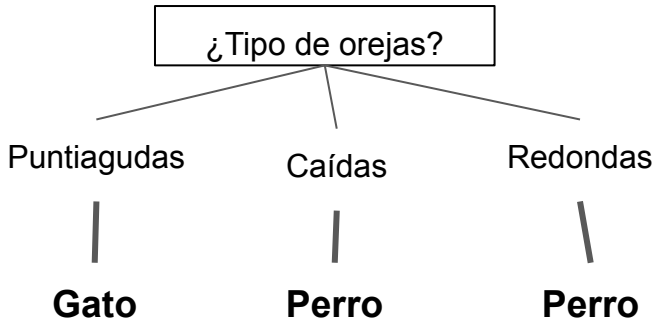


Árbol de decisión

Construcción a partir de un dataset de entrenamiento

ID	Tipo orejas	Color pelo	Color ojos	**Clase**
1	Puntiagudas	Negro	Café	**Gato**
2	Caídas	Gris	Verdes	**Perro**
3	Redondas	Café	Azules	**Perro**
4	Puntiagudas	Rayado	Verdes	**Gato**
5	Caídas	Negro	Café	**Perro**

¿Qué atributo elegimos?



Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	Tipo orejas	Color pelo	Color ojos	**Clase**
1	Puntiagudas	Negro	Café	**Gato**
2	Caídas	Gris	Verdes	**Perro**
3	Redondas	Café	Azules	**Perro**
4	Puntiagudas	Rayado	Verdes	**Gato**
5	Caídas	Negro	Café	**Perro**

P(gato) = ⅖
P(perro) = ⅗

Entropía

H(S) = -
(P(gato)*$\log_2$(P(gato)) +
P(perro)*$\log_2$((P(perro))

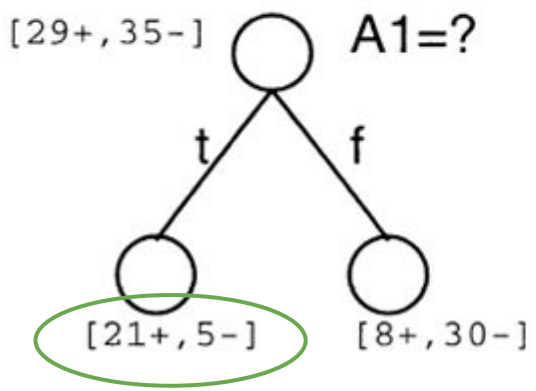
0.97

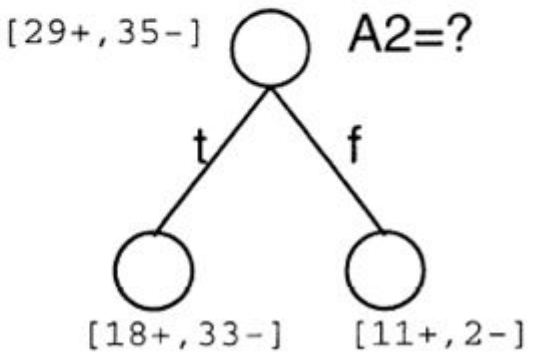
Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

- El que reduce en mayor grado la entropía inicial

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$





Proporciones en clase objetivo

Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	**Tipo orejas**	Color pelo	Color ojos	**Clase**
1	**Puntiagudas**	Negro	Café	**Gato**
2	**Caídas**	Gris	Verdes	**Perro**
3	**Redondas**	Café	Azules	**Perro**
4	**Puntiagudas**	Rayado	Verdes	**Gato**
5	**Caídas**	Negro	Café	**Perro**

Ganancia de información por **tipo de orejas**

0.97

Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	**Tipo orejas**	Color pelo	Color ojos	**Clase**
1	**Puntiagudas**	Negro	Café	**Gato**
2	**Caídas**	Gris	Verdes	**Perro**
3	**Redondas**	Café	Azules	**Perro**
4	**Puntiagudas**	Rayado	Verdes	**Gato**
5	**Caídas**	Negro	Café	**Perro**

Ganancia de información por **tipo de orejas**

0.97 - 0

Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	**Tipo orejas**	Color pelo	Color ojos	**Clase**
1	**Puntiagudas**	Negro	Café	**Gato**
2	**Caídas**	Gris	Verdes	**Perro**
3	**Redondas**	Café	Azules	**Perro**
4	**Puntiagudas**	Rayado	Verdes	**Gato**
5	**Caídas**	Negro	Café	**Perro**

Ganancia de información por **tipo de orejas**

0.97 - 0

Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	**Tipo orejas**	Color pelo	Color ojos	**Clase**
1	**Puntiagudas**	Negro	Café	**Gato**
2	**Caídas**	Gris	Verdes	**Perro**
3	**Redondas**	Café	Azules	**Perro**
4	**Puntiagudas**	Rayado	Verdes	**Gato**
5	**Caídas**	Negro	Café	**Perro**

Ganancia de información por **tipo de orejas**

0.97 - 0 - 0

Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	**Tipo orejas**	Color pelo	Color ojos	**Clase**
1	**Puntiagudas**	Negro	Café	**Gato**
2	**Caídas**	Gris	Verdes	**Perro**
3	**Redondas**	Café	Azules	**Perro**
4	**Puntiagudas**	Rayado	Verdes	**Gato**
5	**Caídas**	Negro	Café	**Perro**

Ganancia de información por **tipo de orejas**

0.97 - 0 - 0

Árbol de decisión

*¿Cómo **calculamos** cuál es el mejor atributo?*

ID	**Tipo orejas**	Color pelo	Color ojos	**Clase**
1	**Puntiagudas**	Negro	Café	**Gato**
2	**Caídas**	Gris	Verdes	**Perro**
3	**Redondas**	Café	Azules	**Perro**
4	**Puntiagudas**	Rayado	Verdes	**Gato**
5	**Caídas**	Negro	Café	**Perro**

Ganancia de información por **tipo de orejas**

0.97 - 0 - 0 - 0 = **0.97**

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

0.97

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

0.97 - ⅖ * 1

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

$0.97 - \frac{2}{5} * 1$

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

$0.97 - \frac{2}{5} * 1 - \frac{1}{5} * 0$

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

$0.97 - \frac{2}{5} * 1 - \frac{1}{5} * 0$

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

$$0.97 - \tfrac{2}{5} * 1 - \tfrac{1}{5} * 0 - \tfrac{1}{5} * 0$$

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

$$0.97 - \tfrac{2}{5} * 1 - \tfrac{1}{5} * 0 - \tfrac{1}{5} * 0$$

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	**Color pelo**	Color ojos	**Clase**
1	Puntiagudas	**Negro**	Café	**Gato**
2	Caídas	**Gris**	Verdes	**Perro**
3	Redondas	**Café**	Azules	**Perro**
4	Puntiagudas	**Rayado**	Verdes	**Gato**
5	Caídas	**Negro**	Café	**Perro**

Ganancia de información por **color de pelo**

0.97 - ⅖ * 1 - ⅕ * 0 - ⅕ * 0 - ⅕ * 0 = **0.57**

Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	Color pelo	**Color ojos**	**Clase**
1	Puntiagudas	Negro	**Café**	**Gato**
2	Caídas	Gris	**Verdes**	**Perro**
3	Redondas	Café	**Azules**	**Perro**
4	Puntiagudas	Rayado	**Verdes**	**Gato**
5	Caídas	Negro	**Café**	**Perro**

Ganancia de información por **color de ojos**

0.97 - ⅖ * 1 - ⅖ * 1 - ⅕ * 0 =
0.16

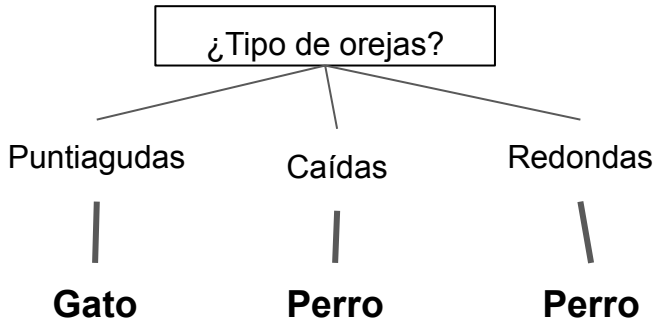
Árbol de decisión

¿Cómo calculamos el mejor atributo?

ID	Tipo orejas	Color pelo	Color ojos	**Clase**
1	Puntiagudas	Negro	Café	**Gato**
2	Caídas	Gris	Verdes	**Perro**
3	Redondas	Café	Azules	**Perro**
4	Puntiagudas	Rayado	Verdes	**Gato**
5	Caídas	Negro	Café	**Perro**

GI(tipo orejas) = 0.97
GI(color pelo) = 0.57
GI(color ojos) = 0.16

Elegir atributo con **mayor ganancia de información.**



Ejercicio

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Atributos con muchos valores

- Si un atributo tiene muchos valores, probablemente la métrica de ganancia de información lo seleccionará
- Una forma de solucionar el problema es usar la razón de ganancia (GainRatio)

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$

$$SplitInformation(S, A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Ejemplo

rid	age	income	student	credit_rating	Class: buys_computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

$$\text{SplitInfo}(S, Income) = -\frac{4}{14} log_2 \frac{4}{14} - \frac{6}{14} log_2 \frac{6}{14} - \frac{4}{14} log_2 \frac{4}{14} = 1.5567$$

Poda del Árbol

- Pre-podar: Este tipo de poda consiste en detener la expansión de un nodo en un momento dado de la construcción del árbol.
  - Una vez que se detiene la expansión se genera un nodo hoja con la clasificación más frecuente en el subconjunto de tuplas correspondiente.
- Post-Podar: Se genera el árbol completo y luego se buscan sub-ramas a podar, la poda se realiza de la misma forma que en el caso anterior

Poda del Árbol

- Al momento de analizar cada nodo N compara la complejidad del sub-árbol desde el nodo N y la complejidad si se reemplaza el sub-árbol por una hoja.
- Se deben definir criterios de poda relativos a complejidad del árbol y reducción del error del set de test.