



Árboles de Decisión



Agenda

1. Motivación
2. Idea de un Árbol de Decisión
3. Aprendizaje de un Árbol de Decisión
 1. Criterio de Pureza
 2. Ganancia de Información
4. Uniendo las Piezas
5. Detalles Adicionales

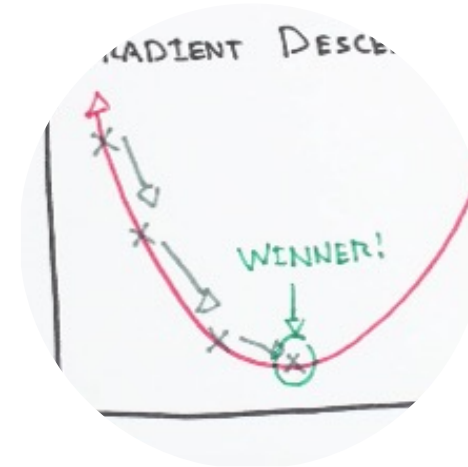


Motivación

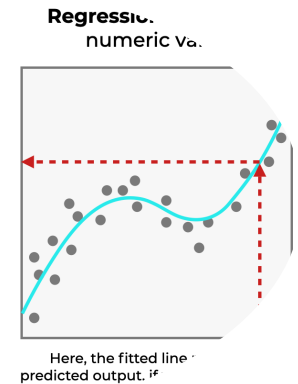
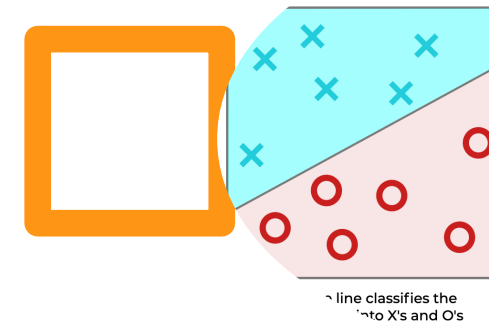
Motivación

Hasta el momento hemos estudiado modelos de clasificación y regresión. En particular:

- Gradiente descendiente
- Funciones de error o pérdida
- Problema de optimización



[Esta foto](#) de Autor desconocido está bajo licencia [CC BY-NC](#)



[Esta foto](#) de Autor desconocido está bajo licencia [CC BY-NC](#)

Motivación

Modelo de clasificación de gatos basado en algunas reglas.



Forma de la Oreja (x_1)	Forma de la Cara (x_2)	Bigotes (x_3)	¿Gato?
Punta	Redonda	Sí	1
Caídas	No redondas	Sí	1
Caídas	Redonda	No	0
Punta	No redonda	Sí	0
Punta	Redonda	Sí	1
Punta	Redonda	No	1
Caídas	No redonda	No	0
Punta	Redonda	No	1

X

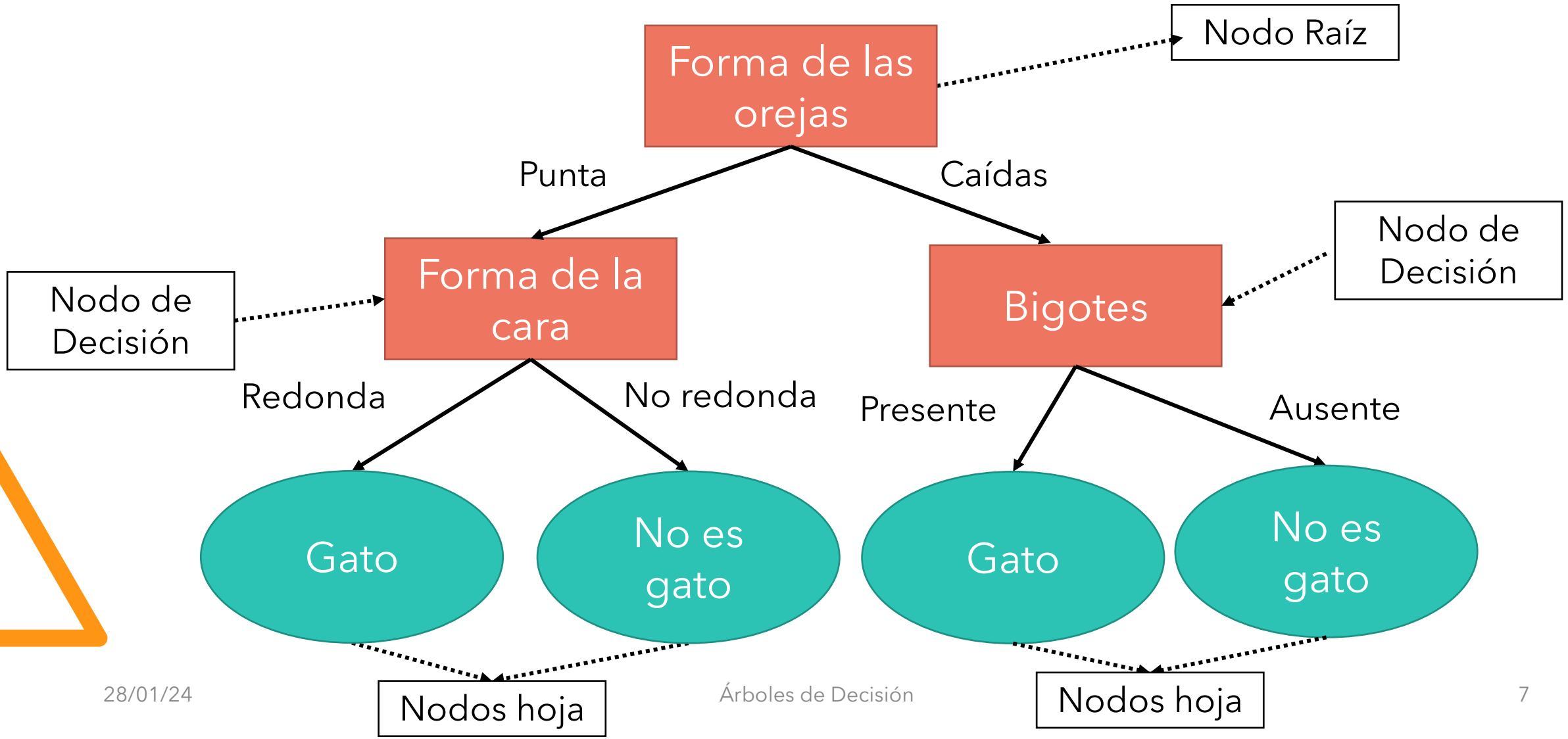
Árboles de Decisión

y



Idea de un Árbol de Decisión

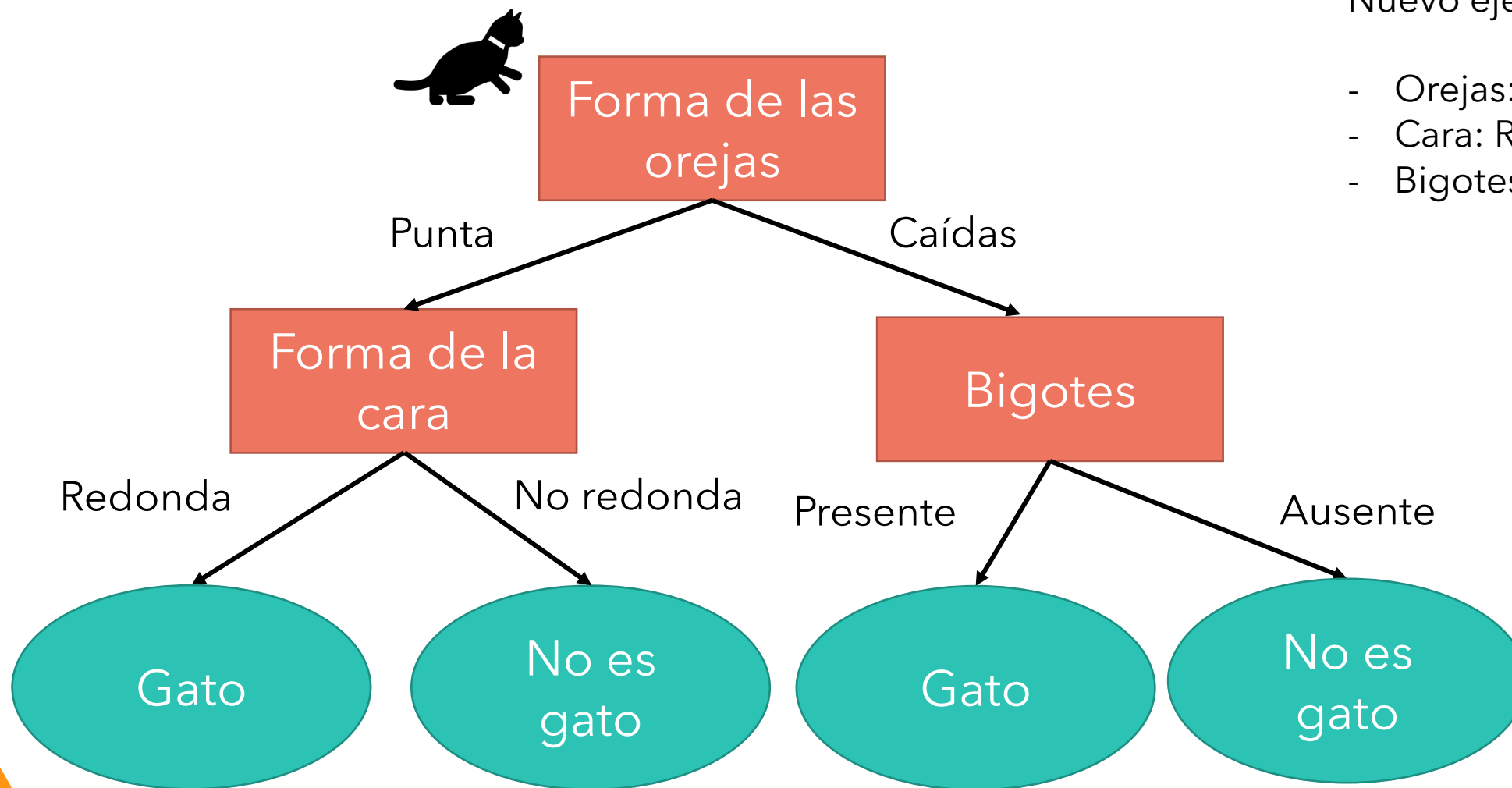
Idea de un Árbol de Decisión



Idea de un Árbol de Decisión

Nuevo ejemplo:

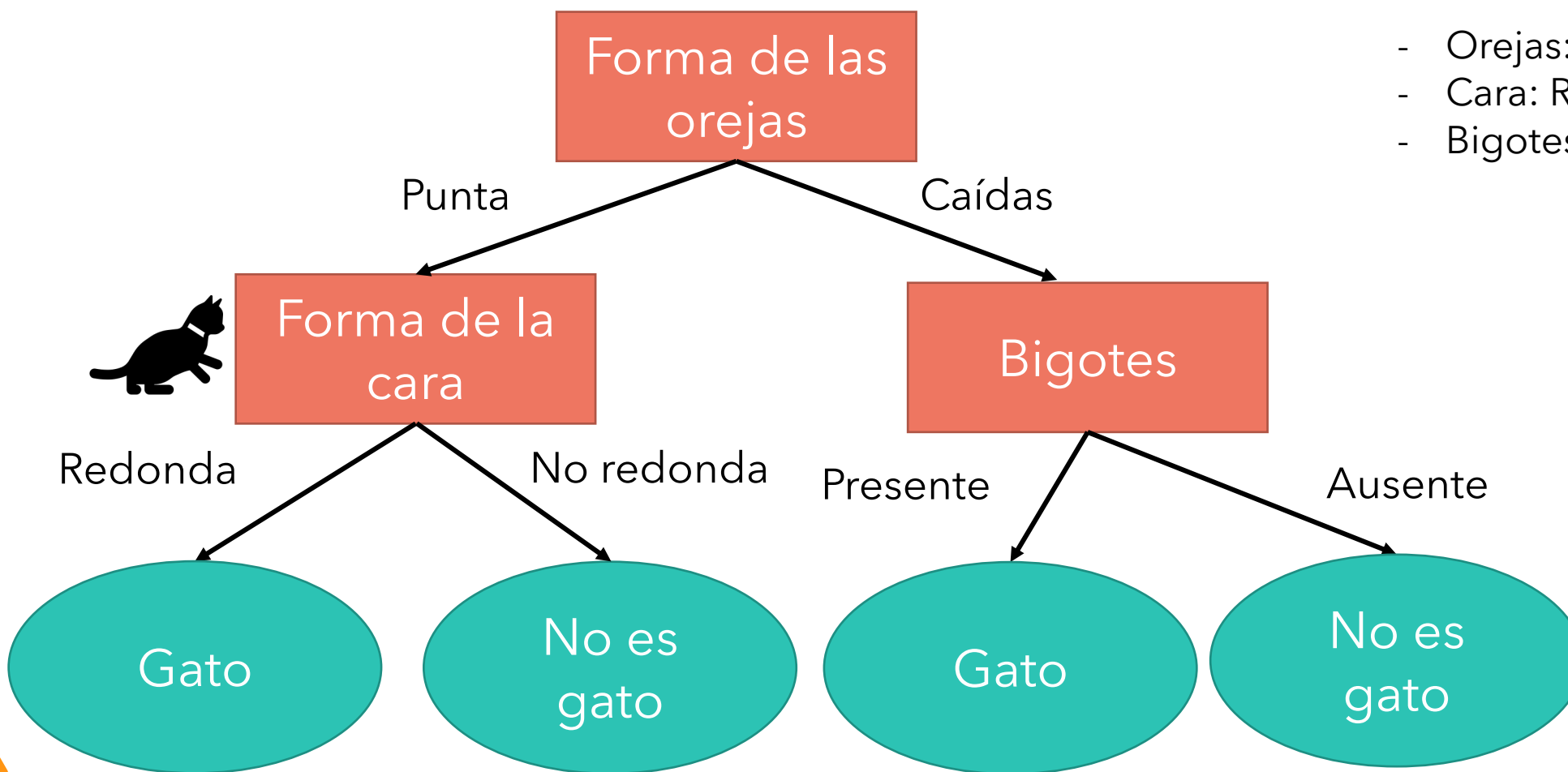
- Orejas: Punta
- Cara: Redonda
- Bigotes: Sí



Idea de un Árbol de Decisión

Nuevo ejemplo:

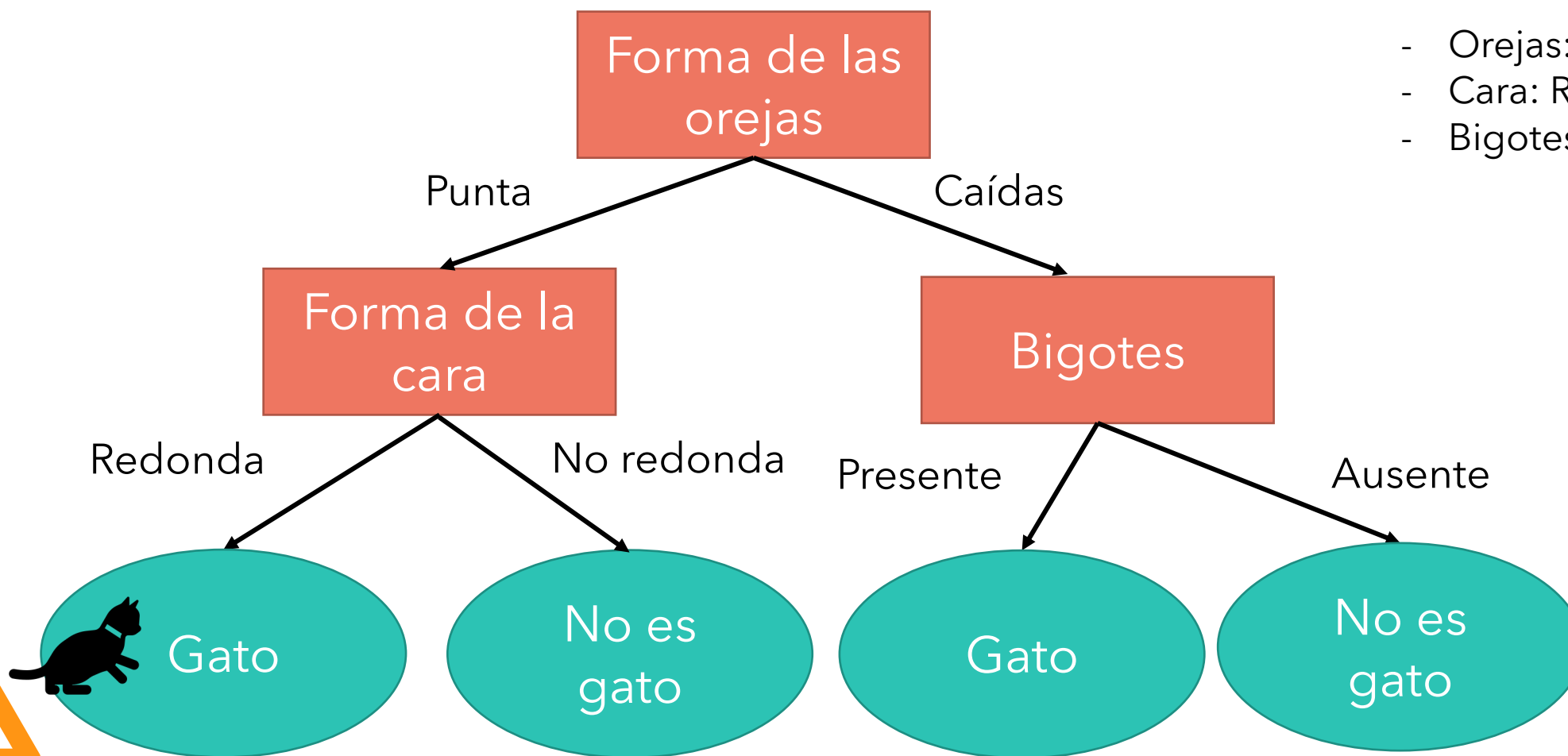
- Orejas: Punta
- Cara: Redonda
- Bigotes: Sí



Idea de un Árbol de Decisión

Nuevo ejemplo:

- Orejas: Punta
- Cara: Redonda
- Bigotes: Sí

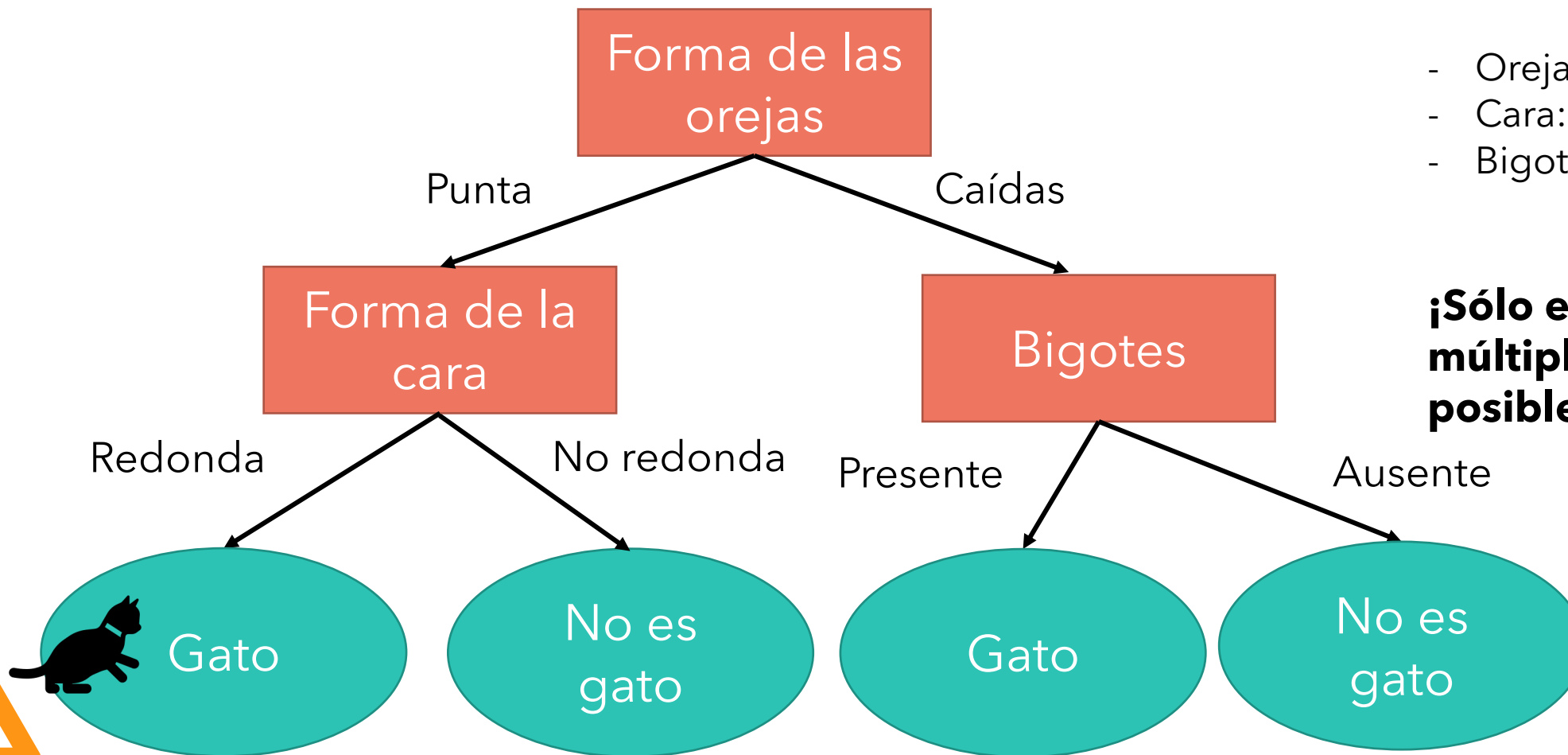


Idea de un Árbol de Decisión

Nuevo ejemplo:

- Orejas: Punta
- Cara: Redonda
- Bigotes: Sí

¡Sólo es uno de múltiples árboles posibles!



Idea de un Árbol de Decisión

Actividad: Propongan un árbol de decisión (uno por equipo) para resolver el problema de clasificación.

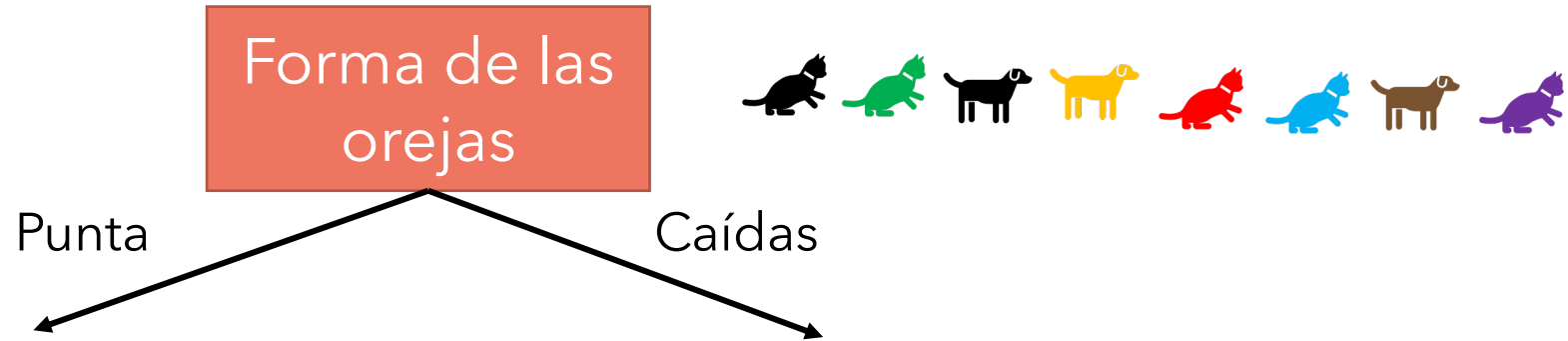


Forma de la Oreja (x_1)	Forma de la Cara (x_2)	Bigotes (x_3)	¿Gato?
Punta	Redonda	Sí	1
Caídas	No redondas	Sí	1
Caídas	Redonda	No	0
Punta	No redonda	Sí	0
Punta	Redonda	Sí	1
Punta	Redonda	No	1
Caídas	No redonda	No	0
Punta	Redonda	No	1

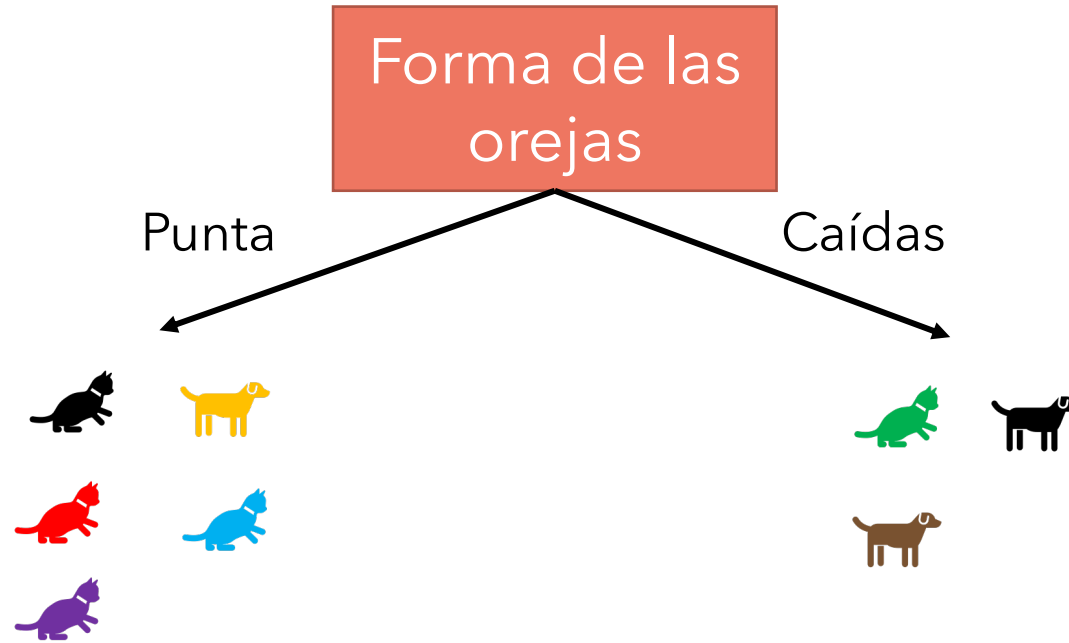
Construcción de un Árbol de Decisión



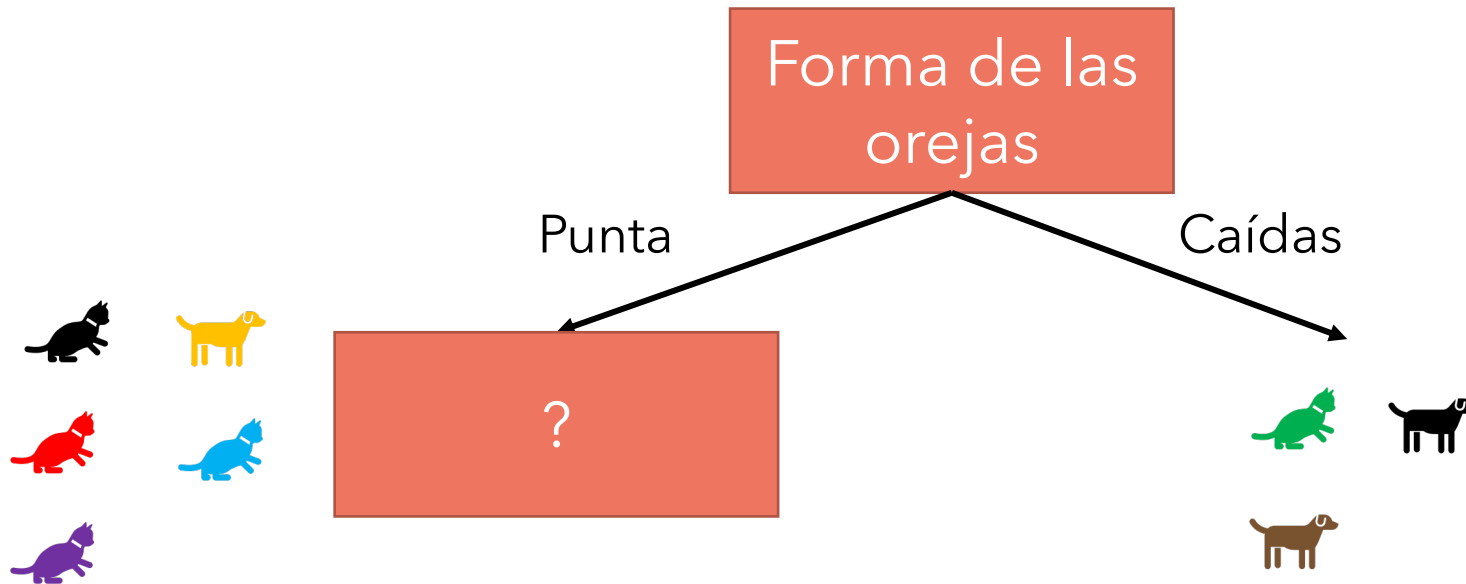
Construcción de un Árbol de Decisión



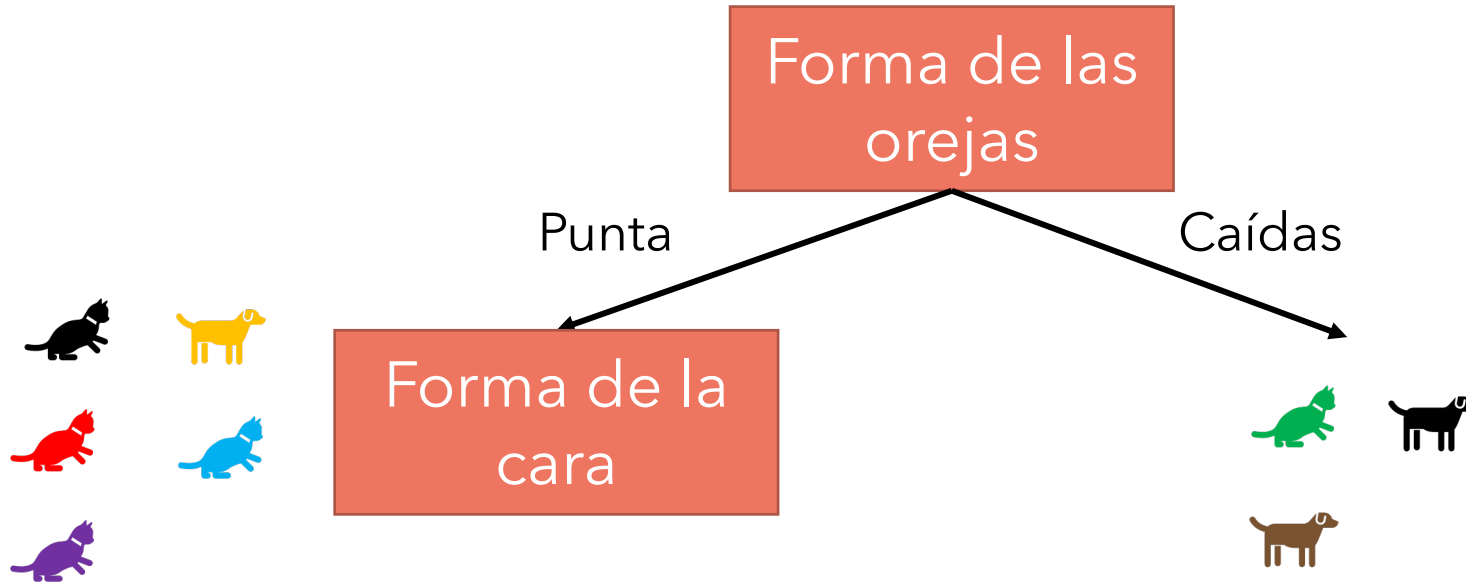
Construcción de un Árbol de Decisión



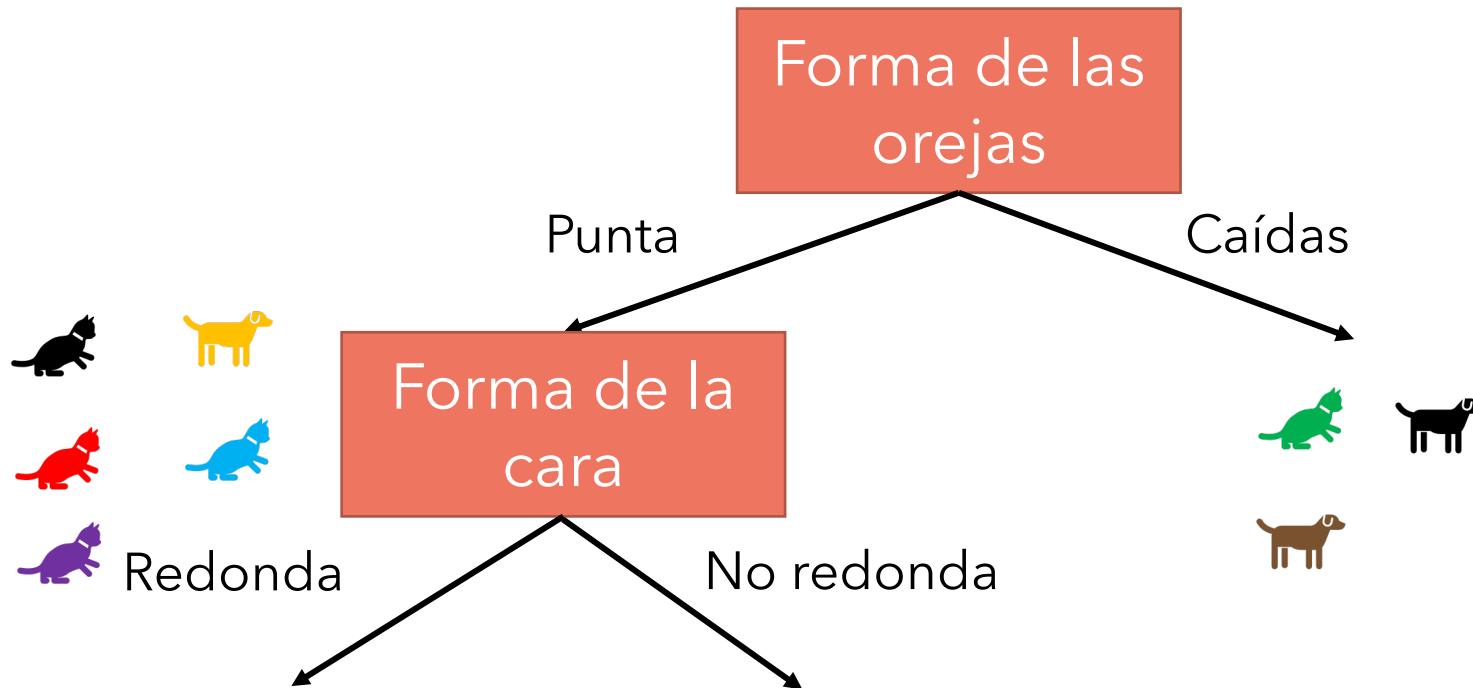
Construcción de un Árbol de Decisión



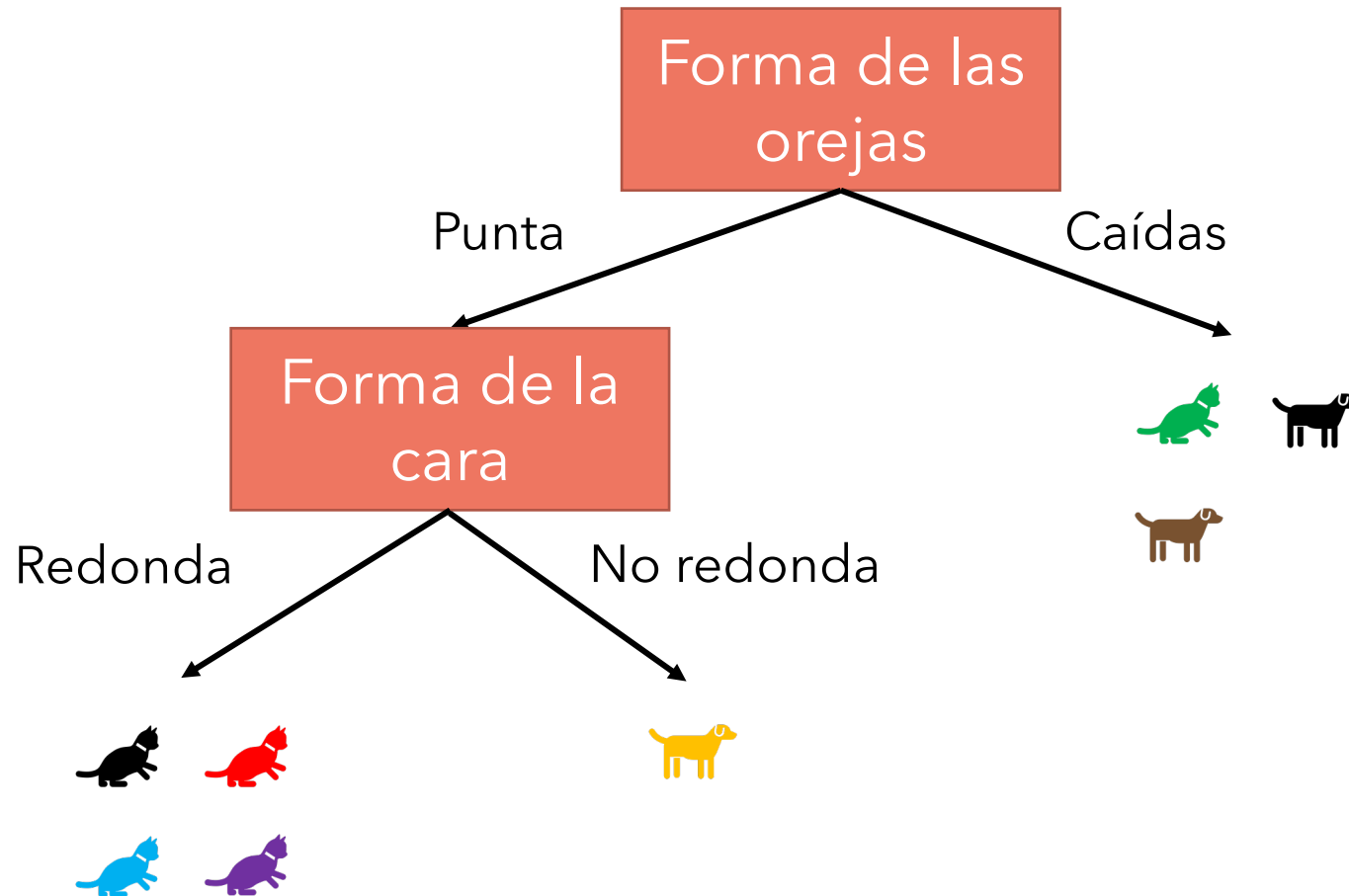
Construcción de un Árbol de Decisión



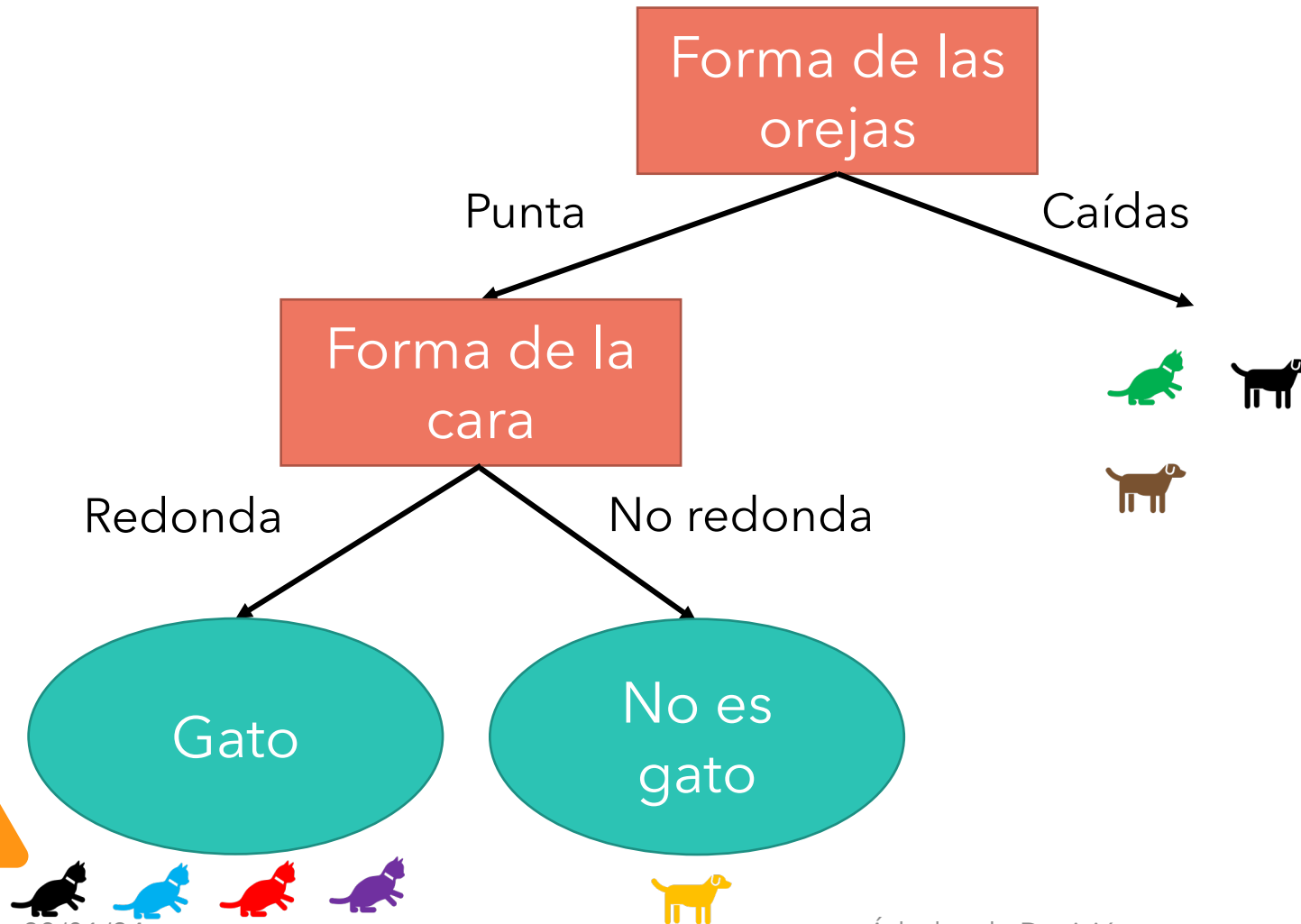
Construcción de un Árbol de Decisión



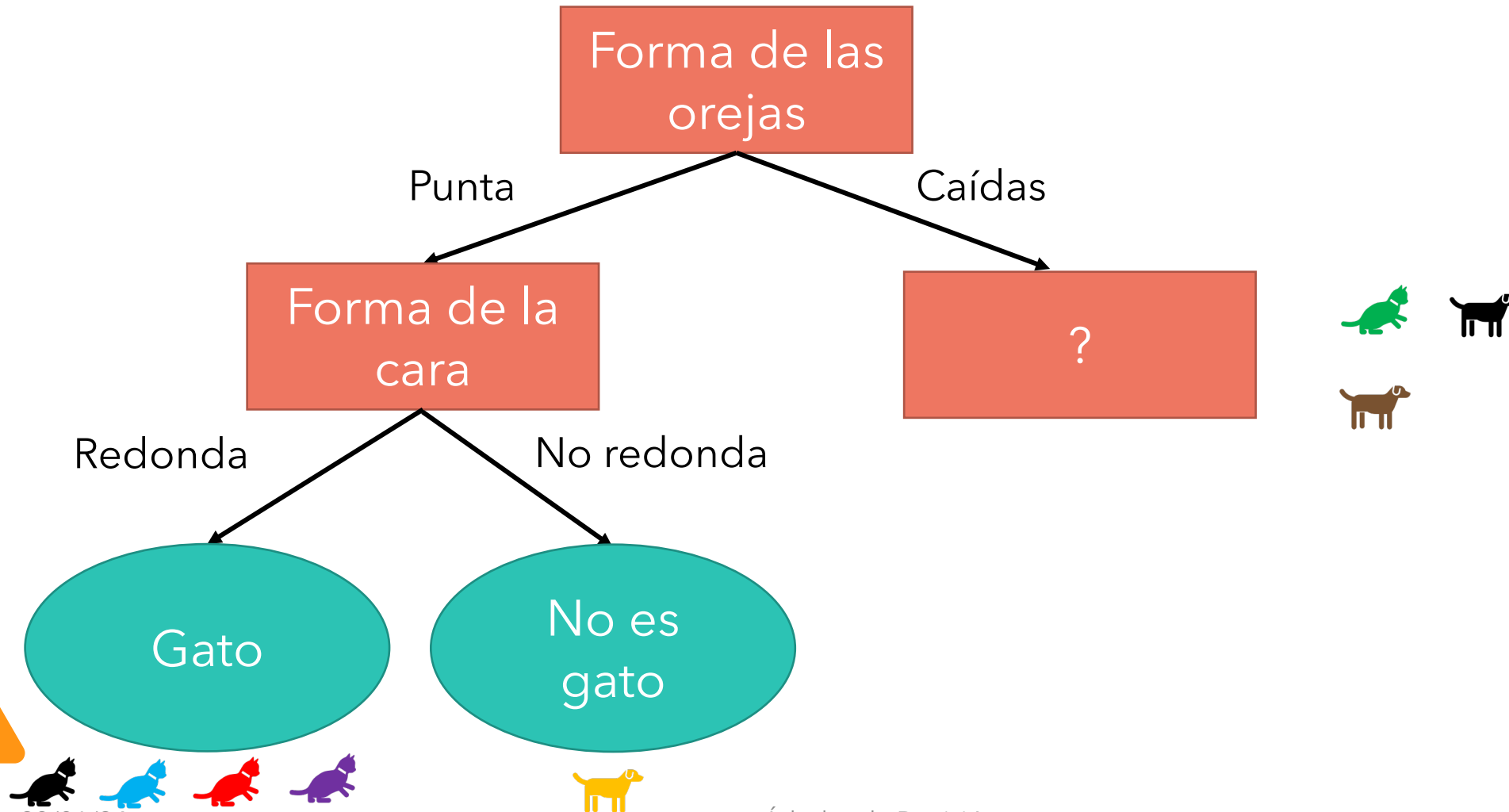
Construcción de un Árbol de Decisión



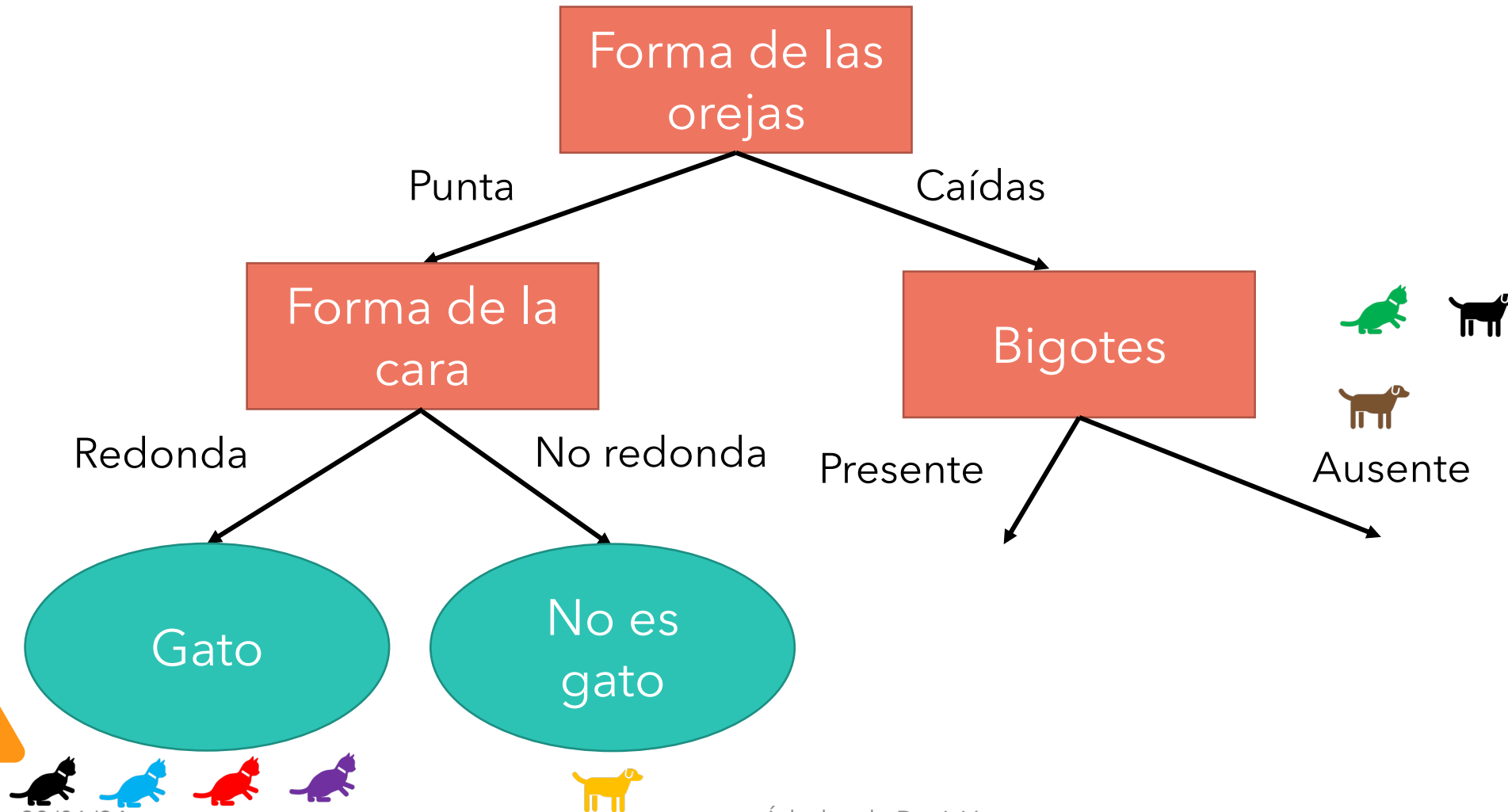
Construcción de un Árbol de Decisión



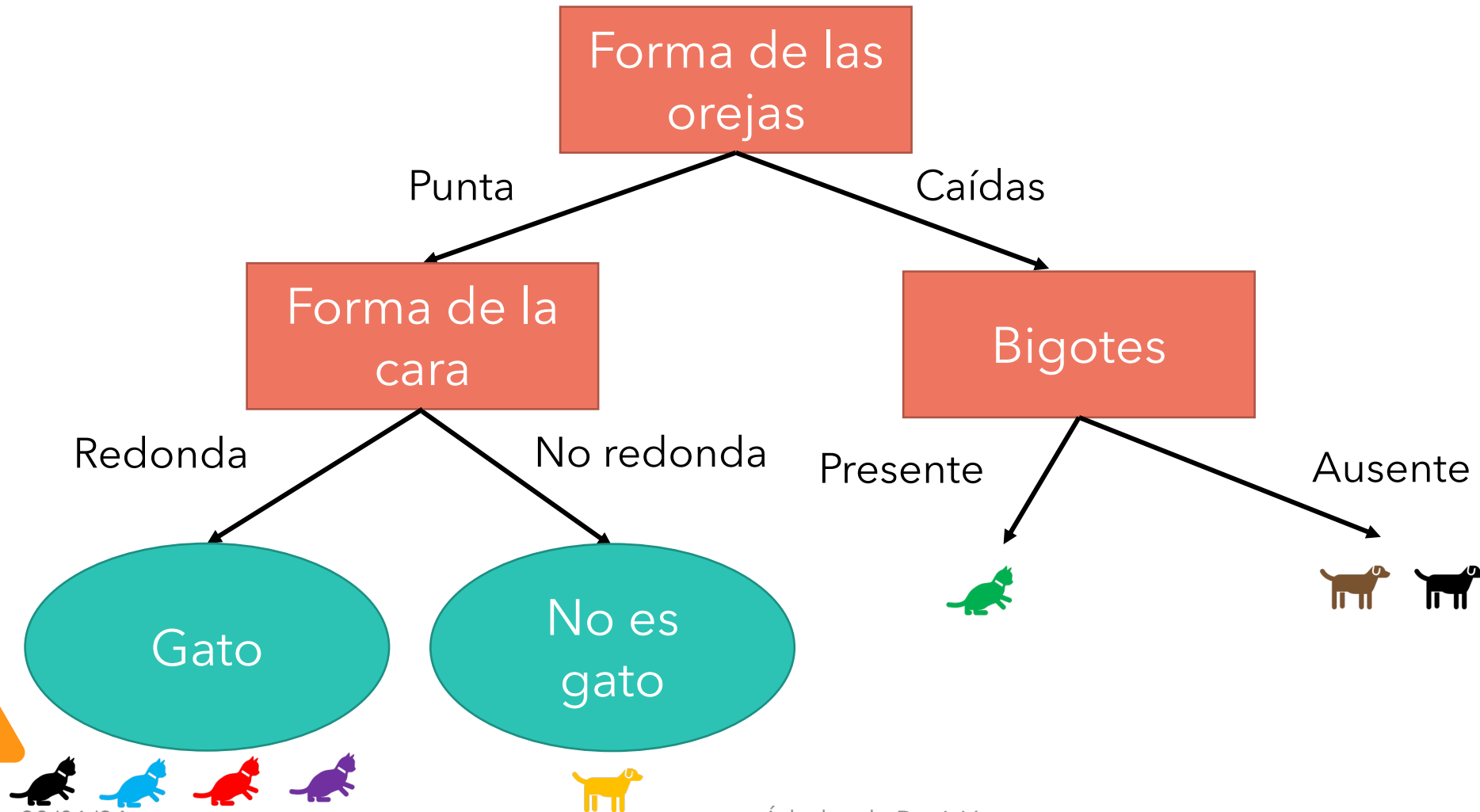
Construcción de un Árbol de Decisión



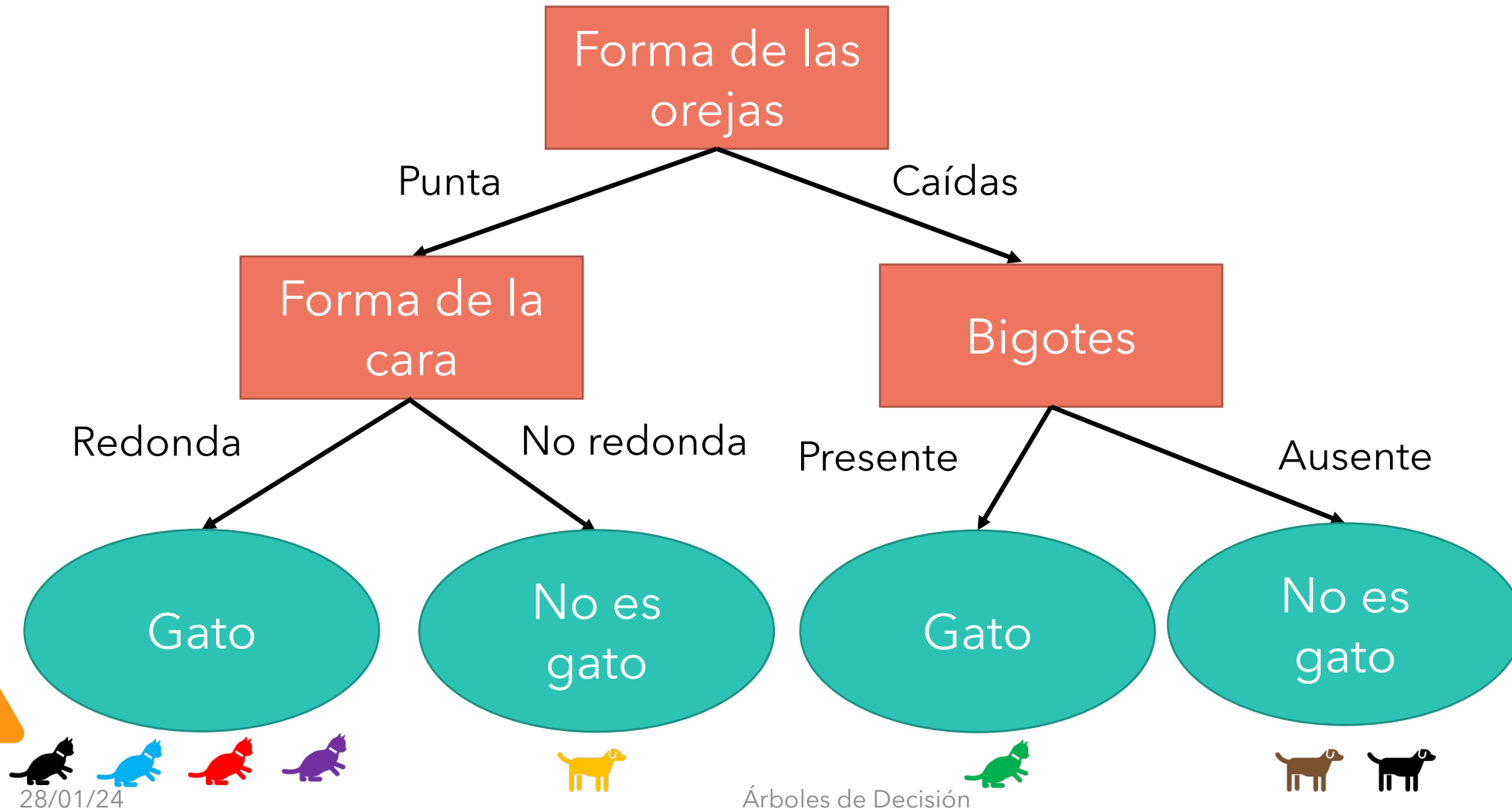
Construcción de un Árbol de Decisión



Construcción de un Árbol de Decisión

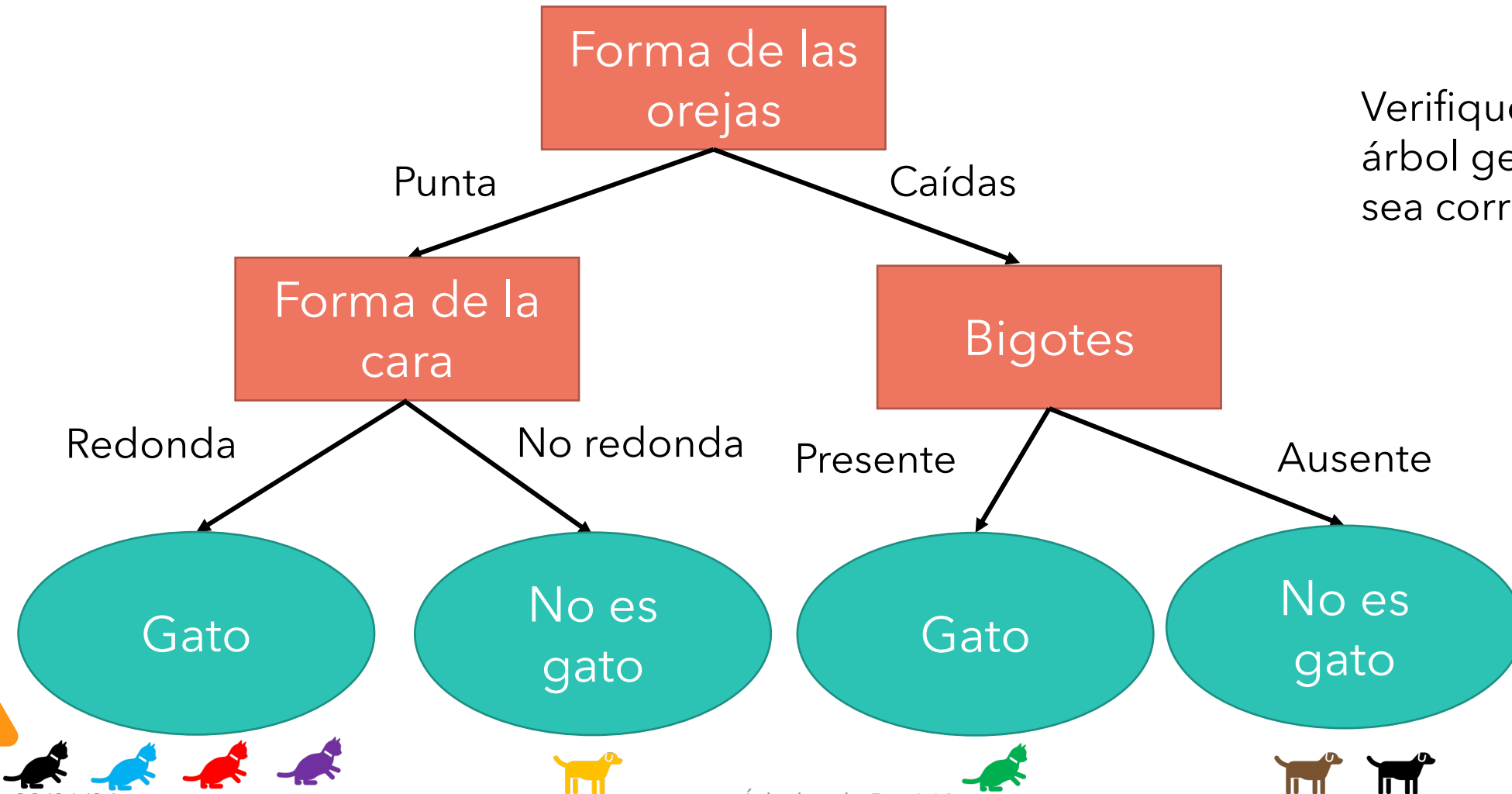


Construcción de un Árbol de Decisión



Construcción de un Árbol de Decisión

Verifiquen que el árbol generado sea correcto.



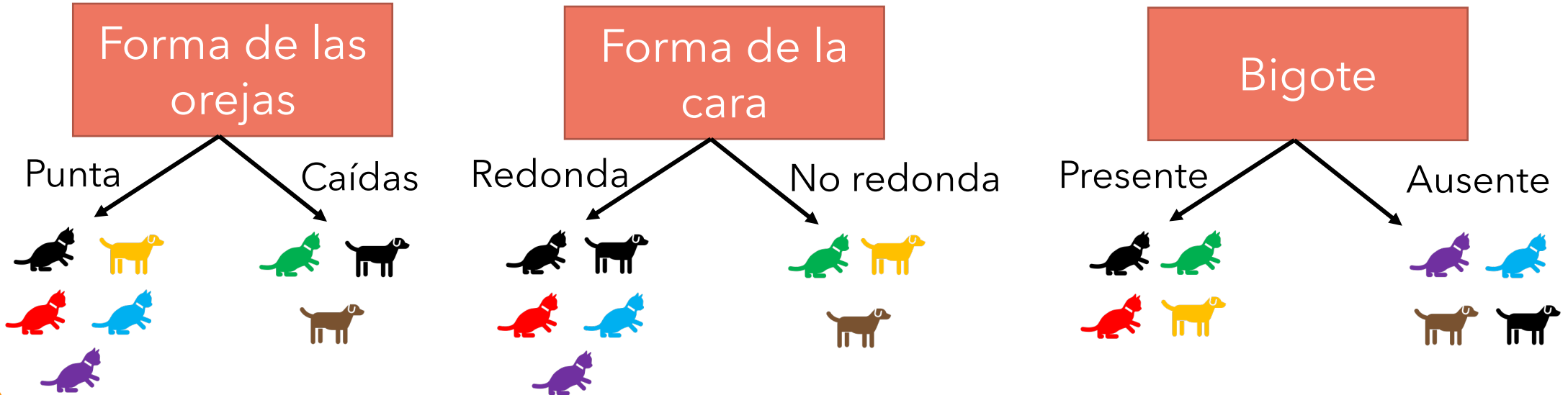


Aprendizaje de un Árbol de Decisión

Aprendizaje de un Árbol de Decisión

Detalle #1: ¿Cómo se debe elegir qué característica de los datos usar para hacer la partición en cada nodo?

Pureza de la hoja



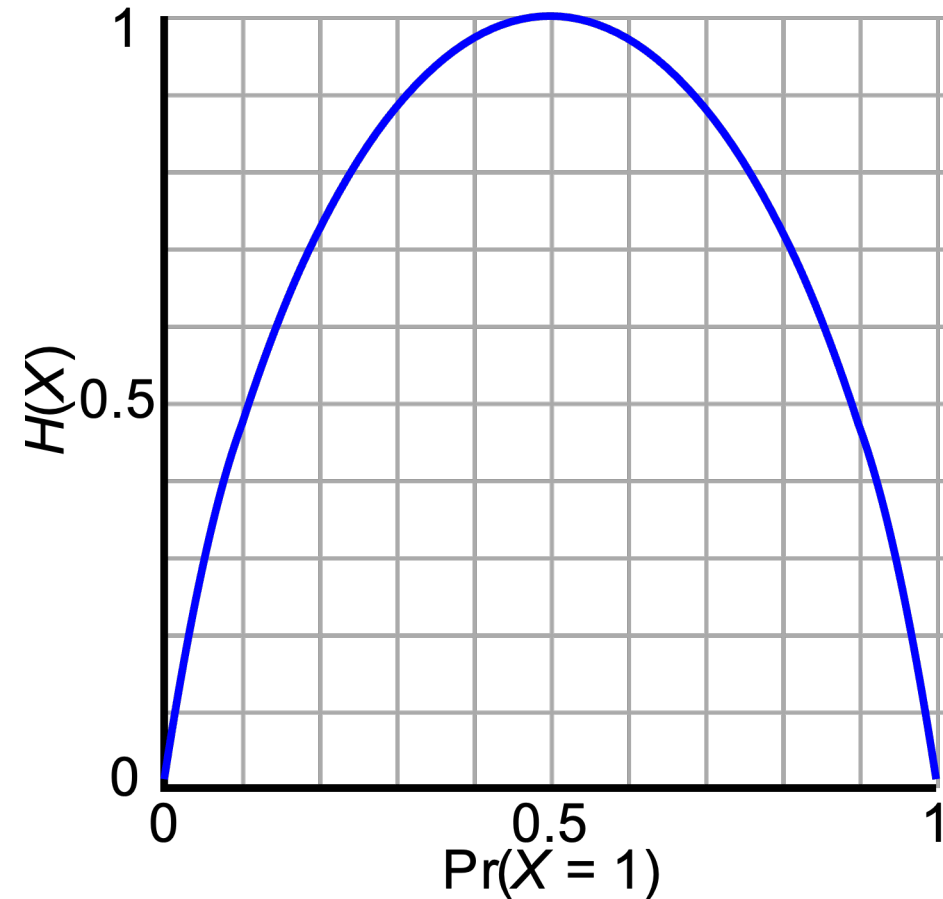
Aprendizaje de un Árbol de Decisión

Detalle #2: ¿Cómo **detener las particiones** en cada nodo?

- Si se llega a un 100% para cada clase.
- Si se llega a una profundidad máxima del árbol.
- Si al seguir expandiendo el árbol no se mejora el valor de pureza.
- Si al expandir un nodo, el número de ejemplos se encuentra debajo de un límite establecido.

¿Cómo medir la pureza?

- Para medir la pureza vamos a utilizar el concepto de **entropía**.
- Específicamente, **entropía binaria** para este caso de *clasificación binaria*.
- La entropía mide el nivel de incertidumbre en un mensaje.

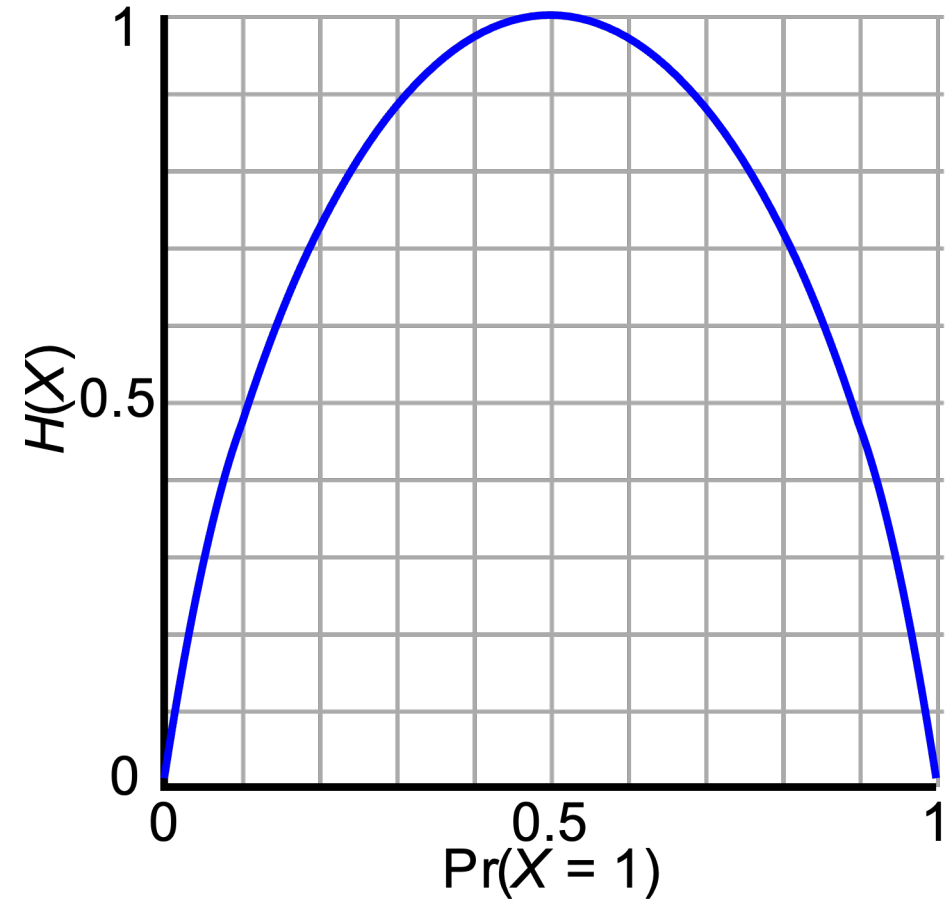


¿Cómo medir la pureza?

 $p_1 = 4/5$

 $p_1 = 3/5$

 $p_1 = 0/5$



¿Cómo medir la pureza?

Nota: definimos $0 \log(0) = 0$

p_1 = la fracción de ejemplos
que son gatos.

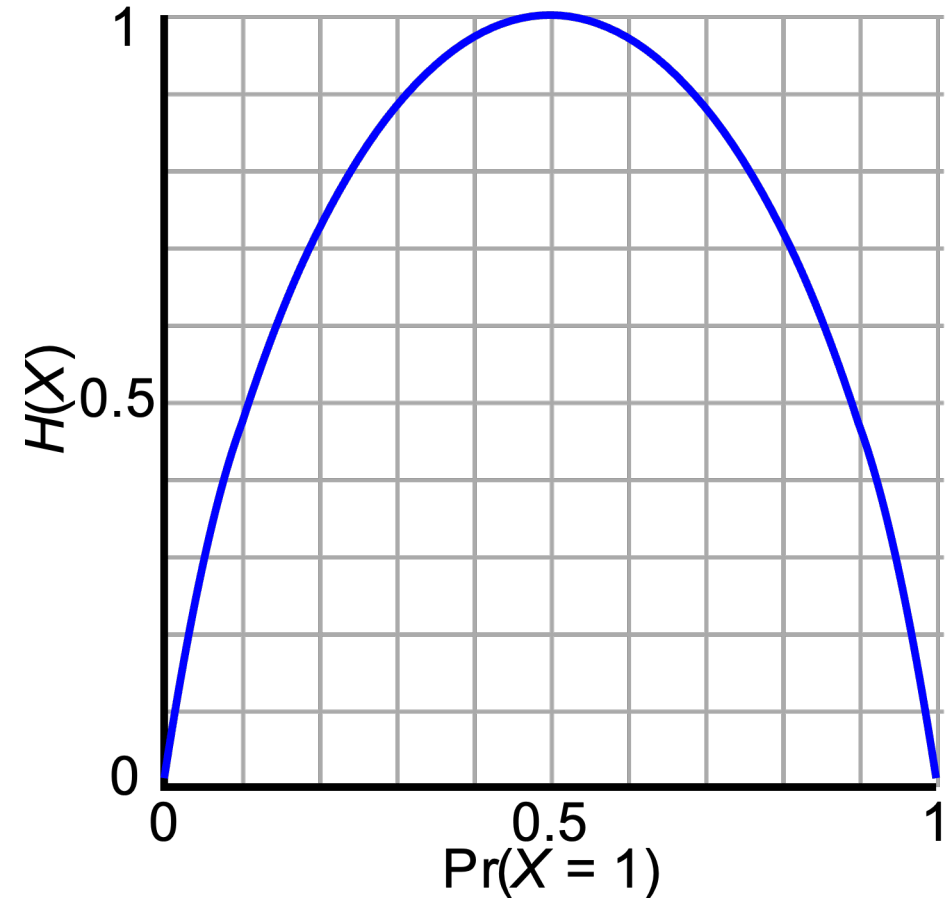
$$p_0 = 1 - p_1$$

La entropía binaria se define como

$$H(p_1) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

$$= -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$

¿Por qué logaritmos? La escala es *legible* y el máximo se encuentra en 1.

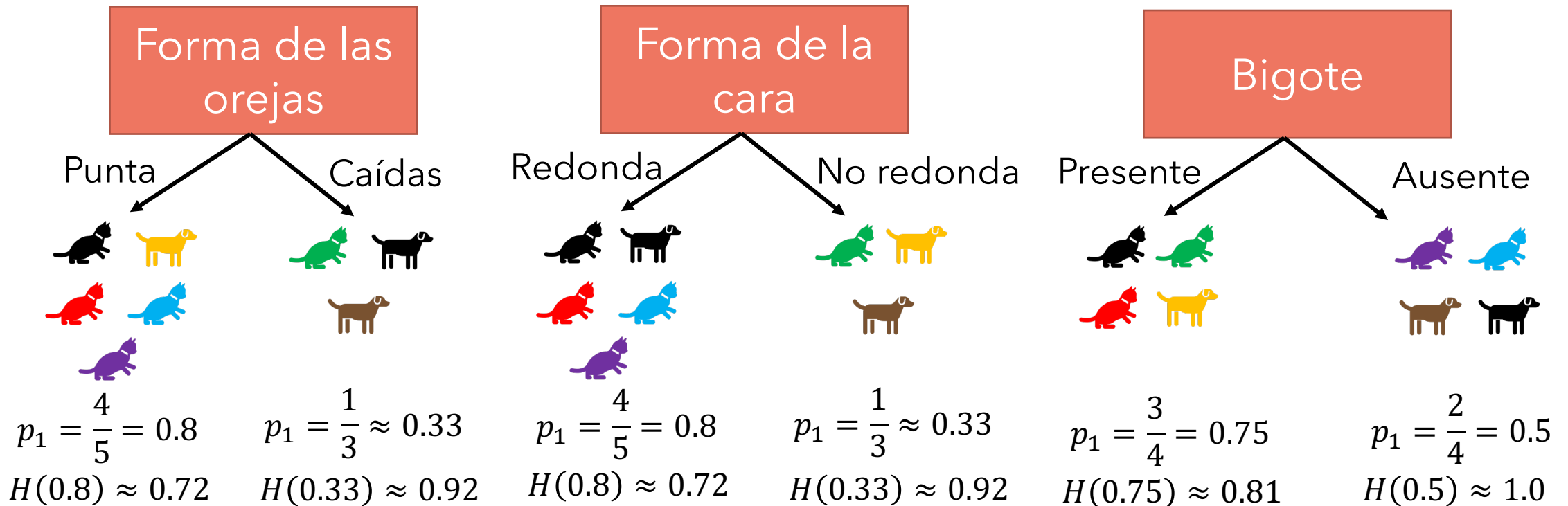


¿Cómo elegir la mejor partición del nodo?

- La idea principal es elegir la característica que maximice la pureza o reduce el valor de la entropía.
- La reducción de la entropía también se llama ganancia de información (en los Árboles de Decisión).

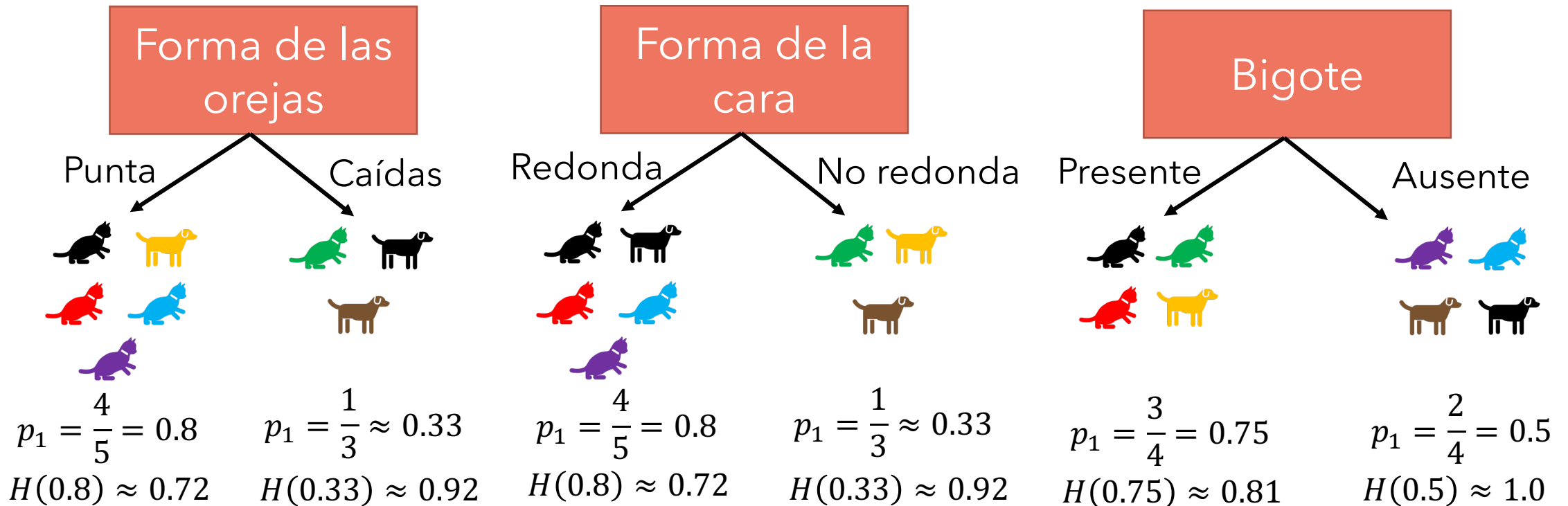
¿Cómo elegir la mejor partición del nodo?

$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$



¿Cómo elegir la mejor partición del nodo?

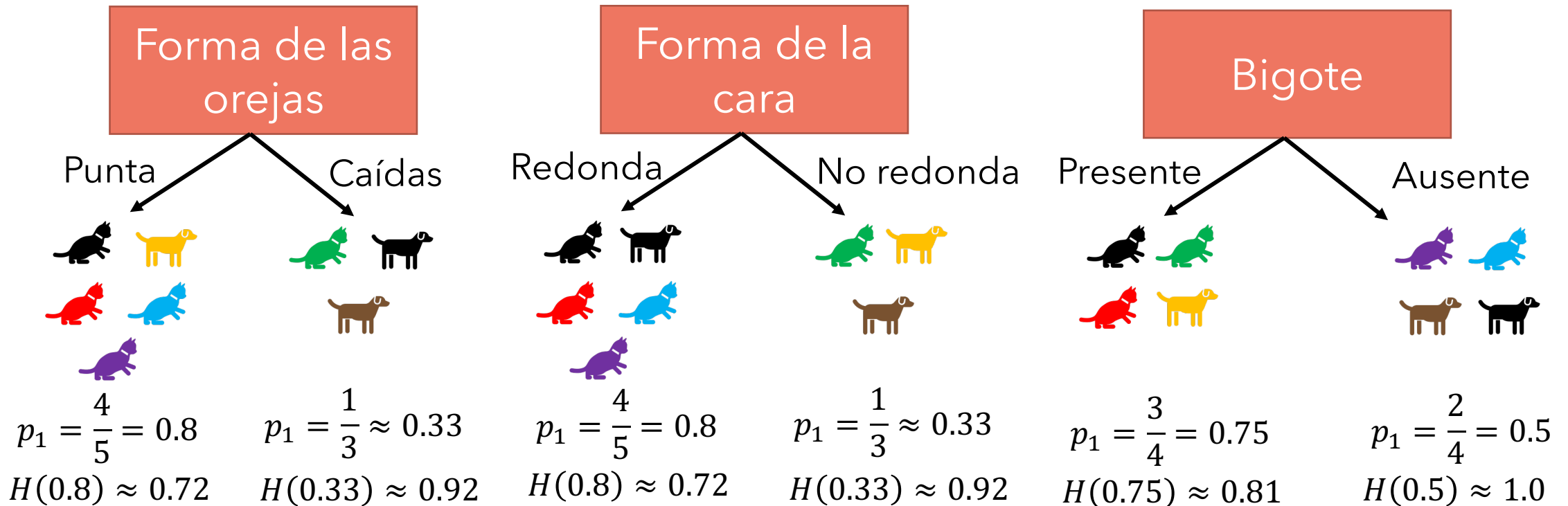
$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$



Verifiquen los resultados

¿Cómo elegir la mejor partición del nodo?

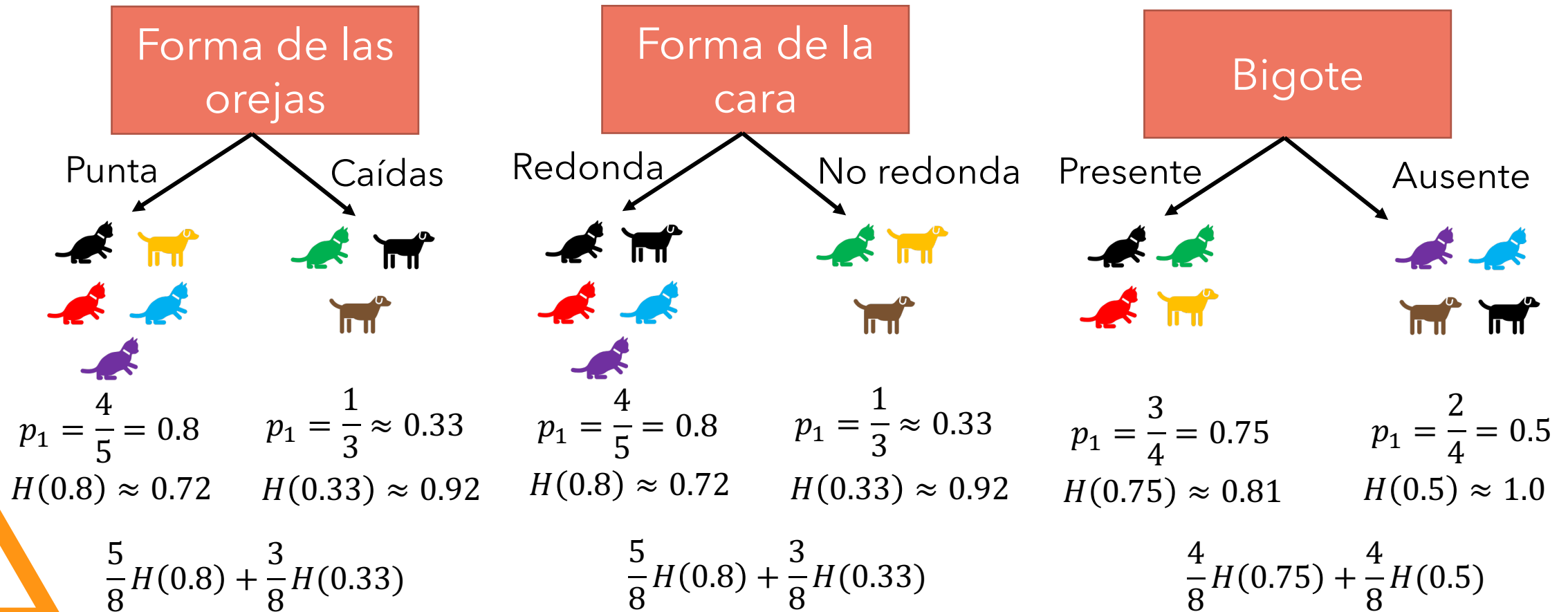
$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$



También debemos considerar el número de ejemplos que entran en cada nodo, por lo que se considera una ponderación. Esto ayuda a considerar un único valor en la decisión.

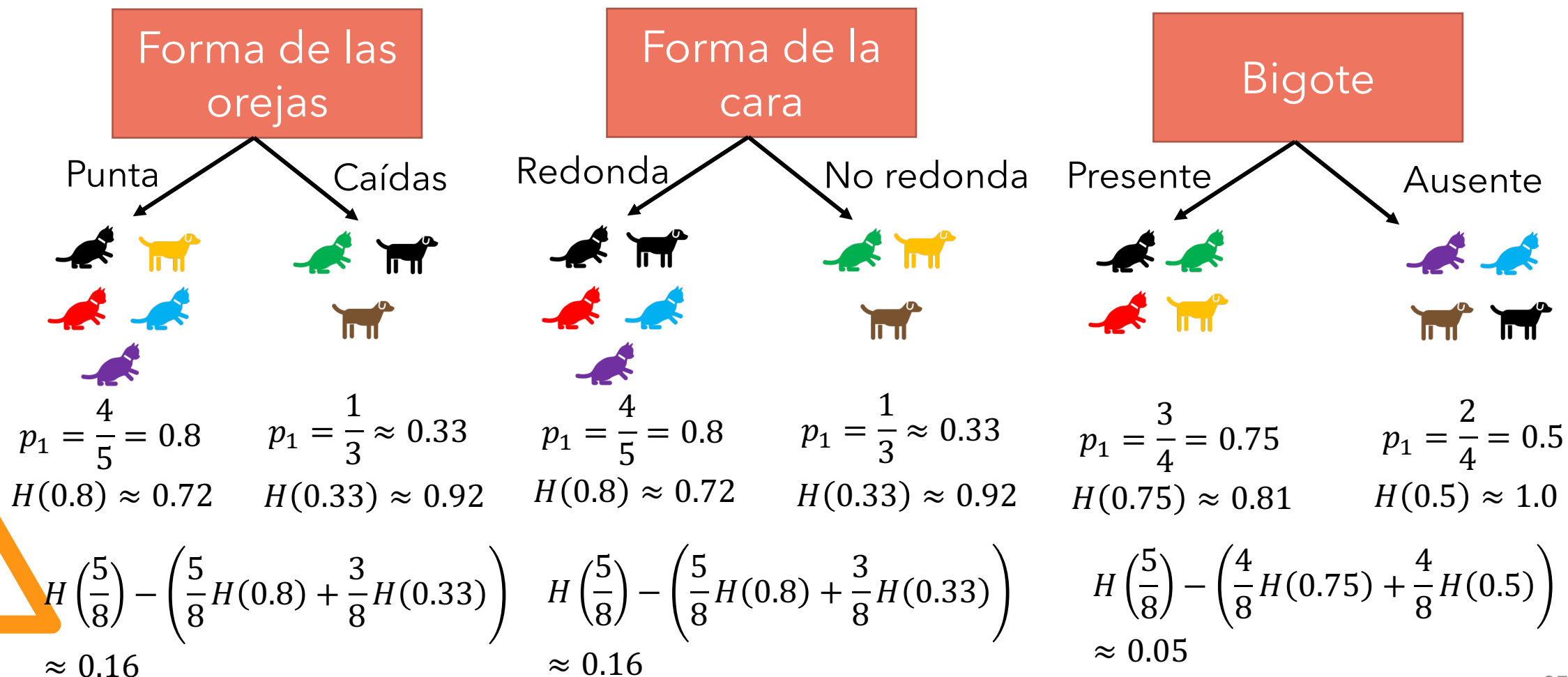
¿Cómo elegir la mejor partición del nodo?

Aquí ya se puede decidir: menor entropía.



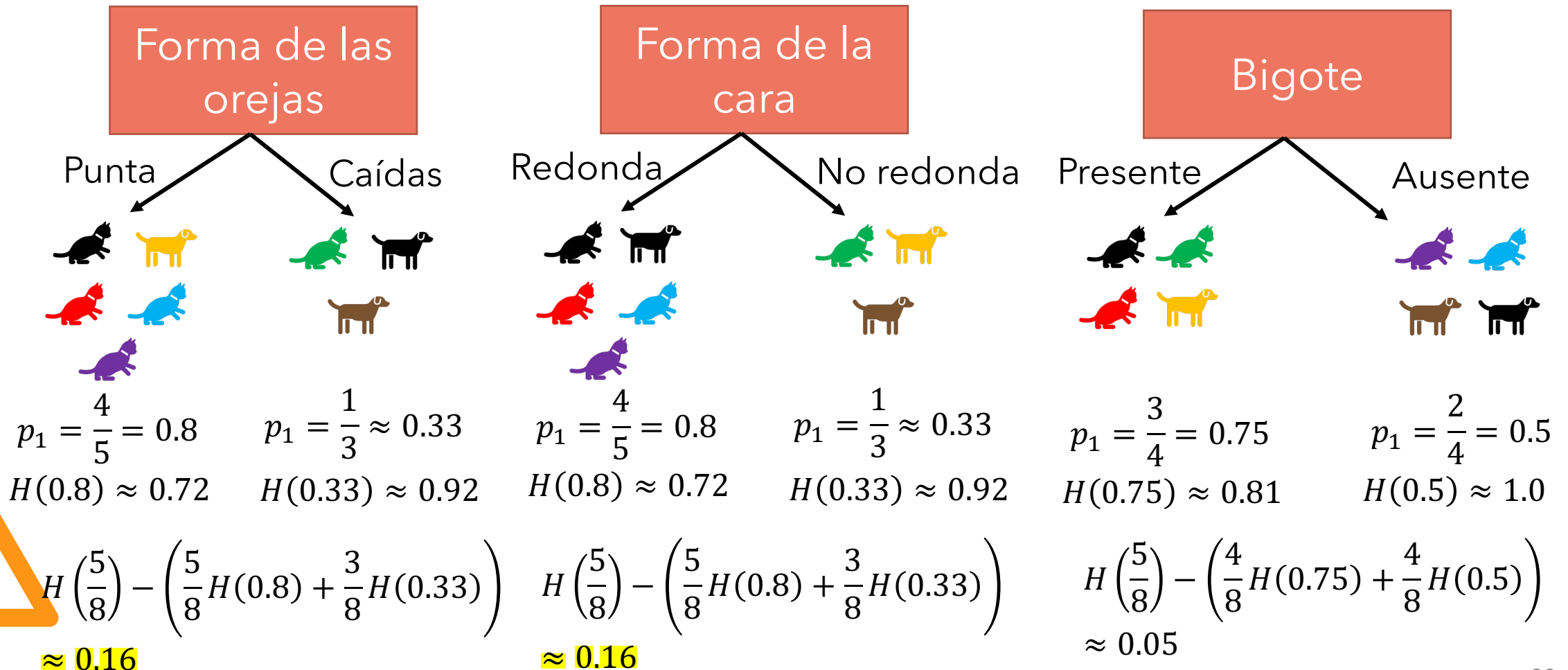
¿Cómo elegir la mejor partición del nodo?

Esto es **ganancia de información**.



¿Cómo elegir la mejor partición del nodo?

Esto es **ganancia de información**.

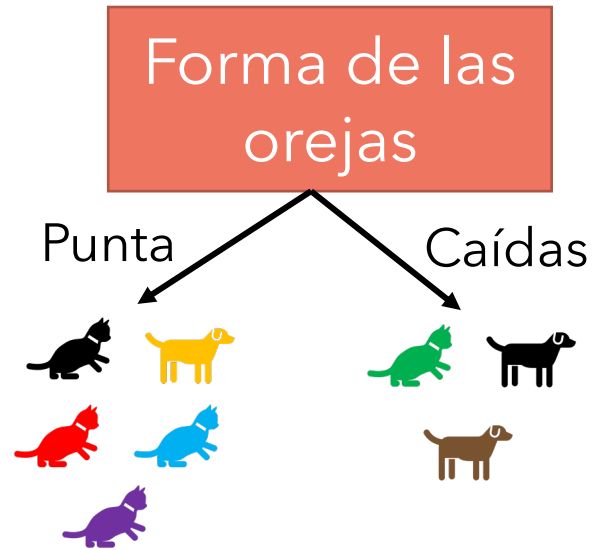


¿Cómo elegir la mejor partición del nodo?



$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$

$$p_1^{raiz} = \frac{5}{8} = 0.625$$



Ganancia de información

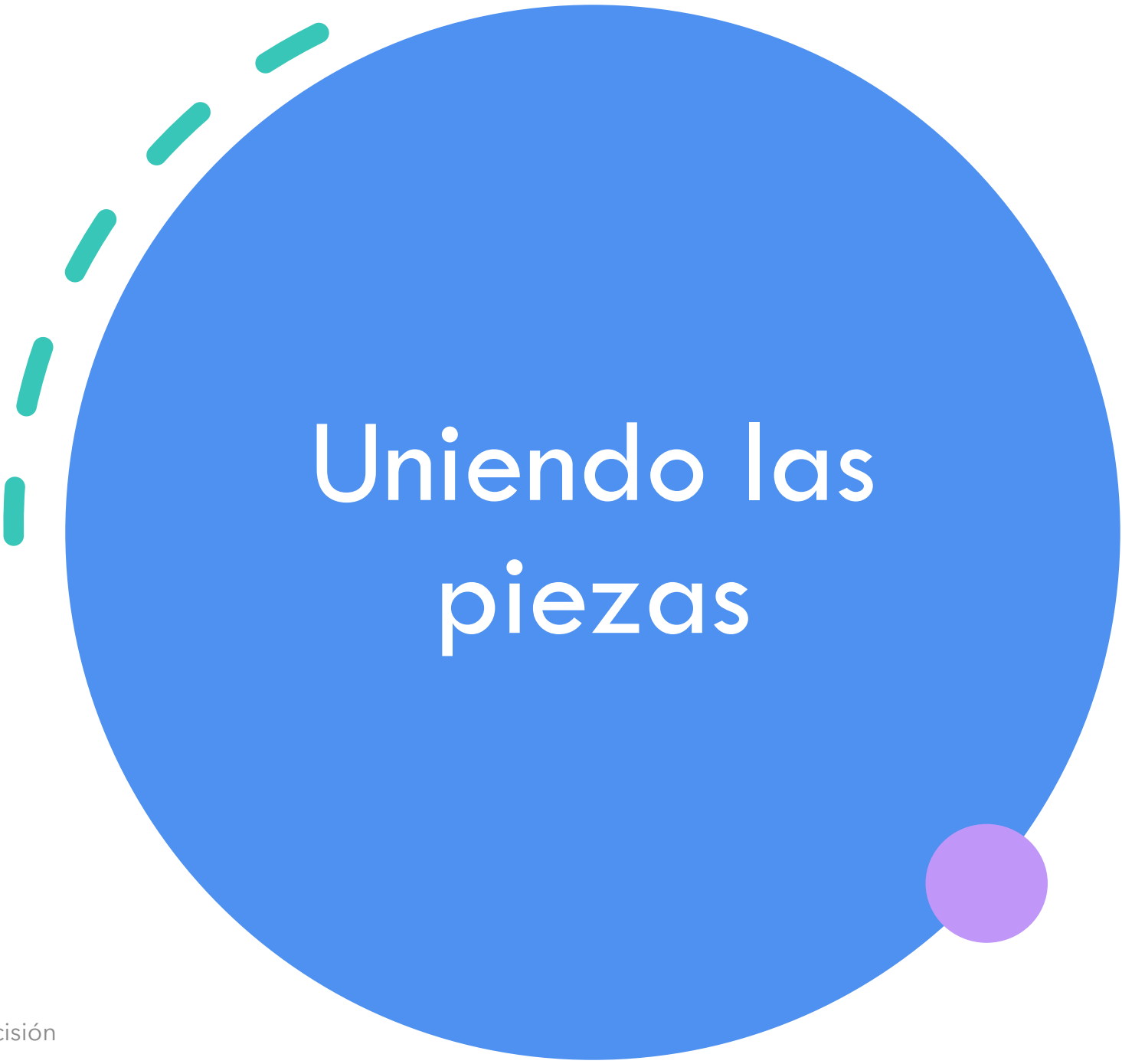
$$H(p_1^{raiz}) - (w^{izq} H(p_1^{izq}) + w^{der} H(p_1^{der}))$$

$$p_1^{izq} = \frac{4}{5} = 0.8$$

$$p_1^{der} = \frac{1}{3} \approx 0.33$$

$$w^{izq} = \frac{5}{8} = 0.625$$

$$w^{der} = \frac{3}{8} = 0.375$$



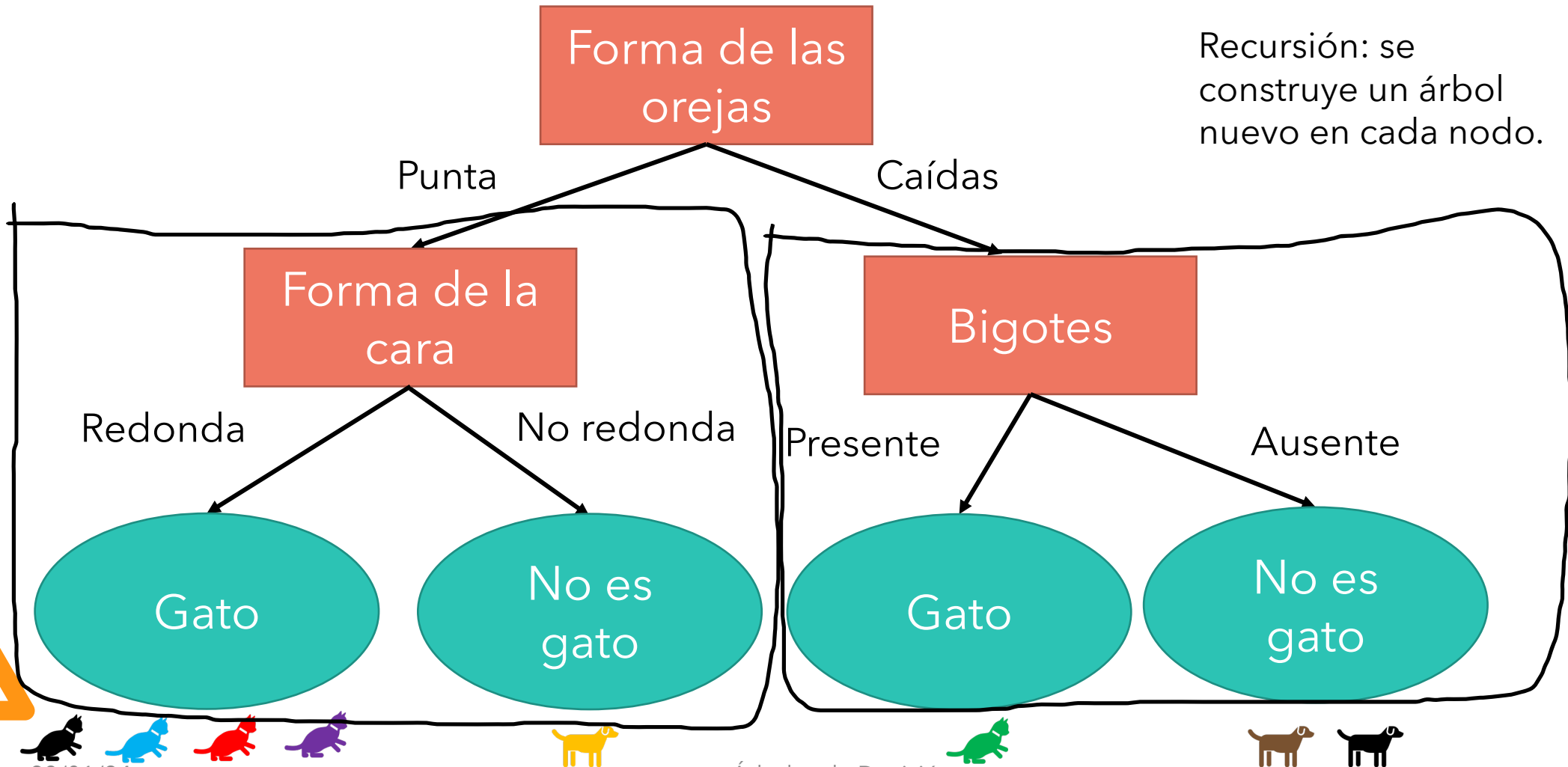
Uniendo las piezas

Árboles de Decisión

1. Se comienza con todos los datos en el nodo raíz.
2. Se calcula la ganancia de información para todas las posibles características. Se elige la que tenga el mayor valor.
3. Partir el conjunto de datos según la característica elegida y crear las ramas izquierda y derecha.
4. Recursión: se inicia 1 con los datos del nodo anterior, y se repite 2 y 3 hasta que:
 - Un nodo sea 100% de una clase.
 - Si al partir un nodo se excede la profundidad máxima establecida.
 - La ganancia de información de particiones subsecuentes es menor que un límite establecido.
 - Si el número de ejemplos en un nodo es menor que un límite establecido.

Construcción de un Árbol de Decisión

Recursión: se construye un árbol nuevo en cada nodo.







Detalles Adicionales

Características no binarias



Forma de la Oreja (x_1)	Forma de la Cara (x_2)	Bigotes (x_3)	¿Gato?
Punta	Redonda	Sí	1
Caídas	No redondas	Sí	1
Ovalada	Redonda	No	0
Punta	No redonda	Sí	0
Ovalada	Redonda	Sí	1
Punta	Redonda	No	1
Caídas	No redonda	No	0
Ovalada	Redonda	No	1

One hot encoding




	Forma de la Oreja	Orejas Punta (x_1)	Orejas Caídas (x_2)	Orejas Ovaladas (x_3)	Forma de la Cara (x_4)	Bigotes (x_5)	¿Gato?
	Punta	1	0	0	Redonda	Sí	1
	Caídas	0	1	0	No redondas	Sí	1
	Ovalada	0	0	1	Redonda	No	0
	Punta	1	0	0	No redonda	Sí	0
	Ovalada	0	0	1	Redonda	Sí	1
	Punta	1	0	0	Redonda	No	1
	Caídas	0	1	0	No redonda	No	0
	Ovalada	0	0	1	Redonda	No	1

One Hot Encoding

La idea del One Hot Encoding es que, si una variable categórica puede tomar k valores, **se pueden crear k características binarias** (0 y 1).

- 0 indica que no está presente esa característica
- 1 indica que sí está presente esa característica

One hot encoding

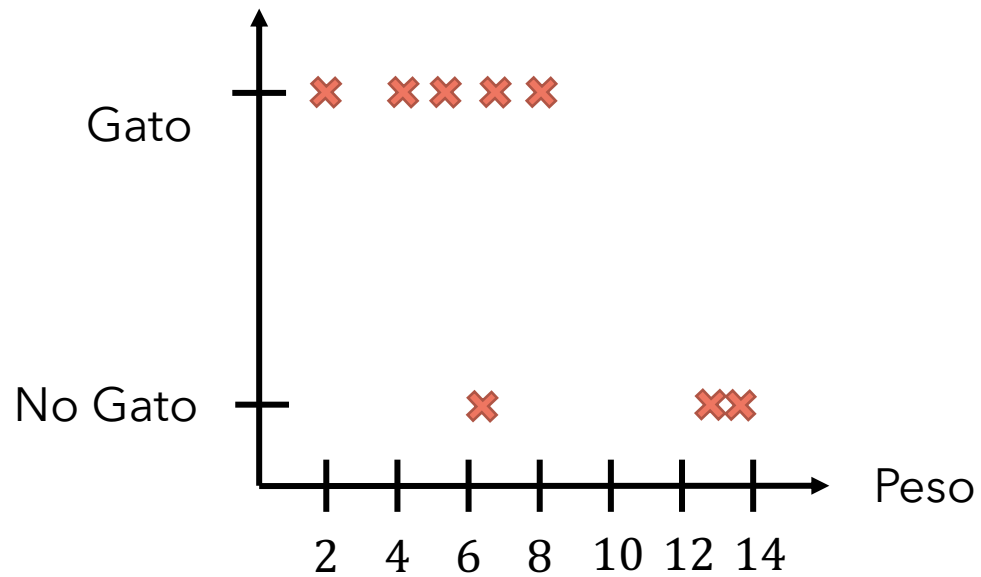
	Forma de la Oreja	Orejas Punta (x_1)	Orejas Caídas (x_2)	Orejas Ovaladas (x_3)	Forma de la Cara (x_4)	Bigotes (x_5)	¿Gato?
	Punta	1	0	0	Redonda	Sí	1
	Caídas	0	1	0	No redondas	Sí	1
	Ovalada	0	0	1	Redonda	No	0
	Punta	1	0	0	No redonda	Sí	0
	Ovalada	0	0	1	Redonda	Sí	1
	Punta	1	0	0	Redonda	No	1
	Caídas	0	1	0	No redonda	No	0
	Ovalada	0	0	1	Redonda	No	1

Árboles de Decisión y Valores Continuos



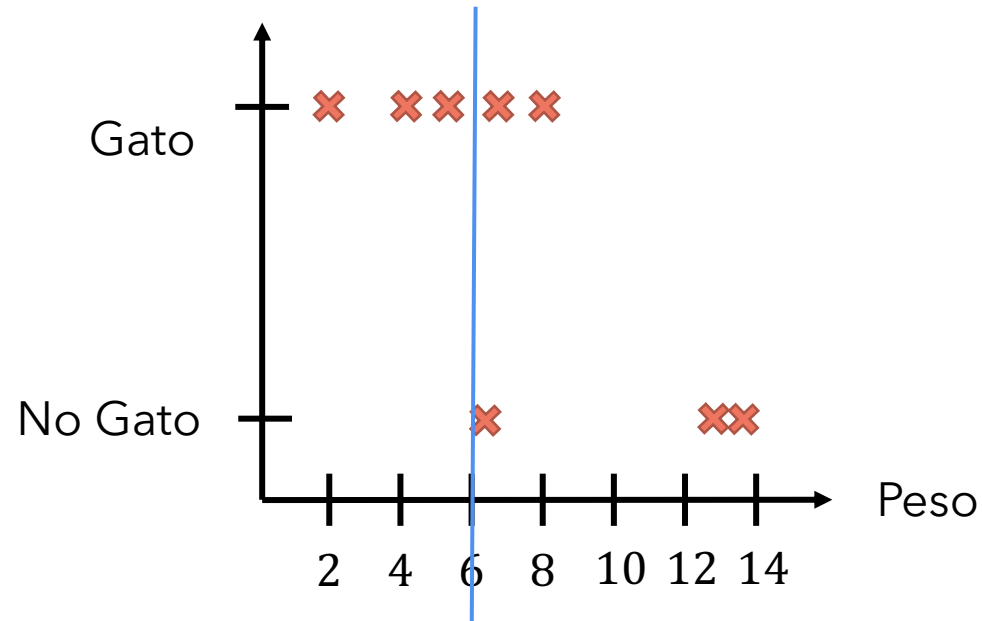
Forma de la Oreja (x_1)	Forma de la Cara (x_2)	Bigotes (x_3)	Peso kg (x_4)	¿Gato?
Punta	Redonda	Sí	2.3	1
Caídas	No redondas	Sí	4.5	1
Caídas	Redonda	No	6.4	0
Punta	No redonda	Sí	13.4	0
Punta	Redonda	Sí	5.6	1
Punta	Redonda	No	6.6	1
Caídas	No redonda	No	13.9	0
Punta	Redonda	No	8.5	1

Árboles de Decisión y Valores Continuos



Árboles de Decisión y Valores Continuos

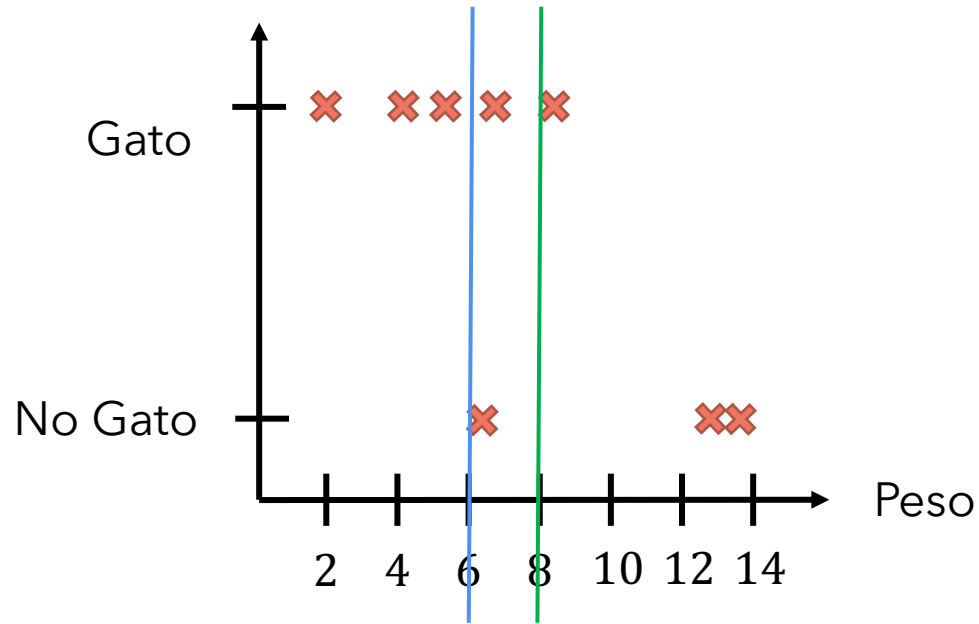
$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$



$$H\left(\frac{5}{8}\right) - \left(\left(\frac{3}{8}\right) H\left(\frac{3}{3}\right) + \left(\frac{5}{8}\right) H\left(\frac{2}{5}\right) \right) \approx 0.35$$

Árboles de Decisión y Valores Continuos

$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$

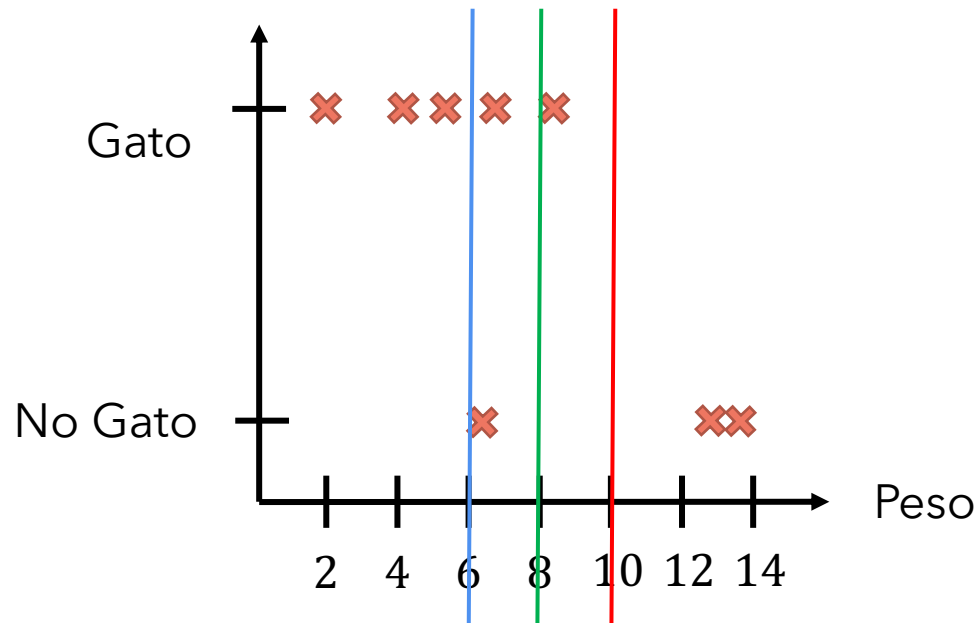


$$H\left(\frac{5}{8}\right) - \left(\left(\frac{3}{8}\right) H\left(\frac{3}{3}\right) + \left(\frac{5}{8}\right) H\left(\frac{2}{5}\right) \right) \approx 0.35$$

$$H\left(\frac{5}{8}\right) - \left(\left(\frac{5}{8}\right) H\left(\frac{4}{5}\right) + \left(\frac{3}{8}\right) H\left(\frac{1}{3}\right) \right) \approx 0.16$$

Árboles de Decisión y Valores Continuos

$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$



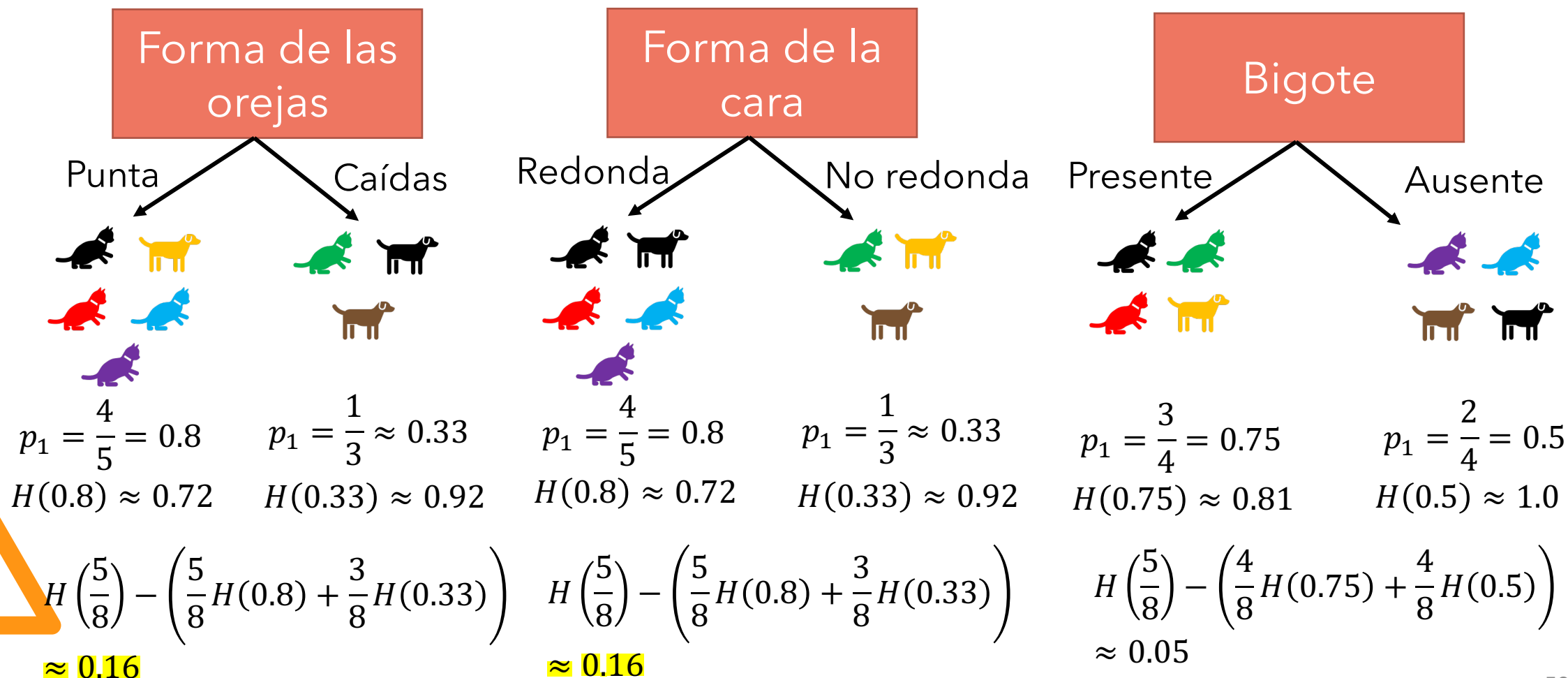
$$H\left(\frac{5}{8}\right) - \left(\left(\frac{3}{8}\right) H\left(\frac{3}{3}\right) + \left(\frac{5}{8}\right) H\left(\frac{2}{5}\right) \right) \approx 0.35$$

$$H\left(\frac{5}{8}\right) - \left(\left(\frac{5}{8}\right) H\left(\frac{4}{5}\right) + \left(\frac{3}{8}\right) H\left(\frac{1}{3}\right) \right) \approx 0.16$$

$$H\left(\frac{5}{8}\right) - \left(\left(\frac{6}{8}\right) H\left(\frac{5}{6}\right) + \left(\frac{2}{8}\right) H\left(\frac{0}{2}\right) \right) \approx 0.47$$

¿Cómo elegir la mejor partición del nodo?

Esto es **ganancia de información**.



Árboles de Decisión y Valores Continuos

¿Cómo se eligen esos límites para determinar la ganancia de información?

- Los valores se ordenan de menor a mayor.
- Los puntos medios entre los valores se eligen como límites para evaluar la ganancia de información.
- E.g., si los puntos son (20,29,40,50), los límites serían (24.5,34.5, 45).

Más detalles

- Es válido que en una partición de un nodo se elija la misma característica en ambas ramas.
- ¿Cuándo existe **alto sesgo**? Si la complejidad del modelo es baja, que se da cuando **los árboles son poco profundos**.
- ¿Cuándo existe **alta varianza**? Si la complejidad del modelo es alta, que se da cuando **los árboles son muy profundos**.
- Esta forma de crear Árboles de Decisión se conoce como **C4.5**.
- Los Árboles de Decisión pueden funcionar con poca información. Además, no requieren escalar las características o centrar los datos.

Tareas

1. ¿Cómo funcionan los Árboles de Decisión para el problema de regresión? Investigar su uso y particularidades.
 - E.g. No se usa entropía, más bien varianza como criterio de ganancia de información.
2. Investigar en qué consiste el criterio de pureza de Gini para los Árboles de Decisión.
3. Demostrar que para la función

$$H(p_1) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

su máximo se encuentra cuando $p_1 = p_0 = 1/2$.



Gracias

Luis Zúñiga

p40887@correo.uia.mx

<https://lzun.github.io>