Universidad de los Andes | Educación Continua
Vicerrectoría Académica

Árboles de decisión y Bosques aleatorios

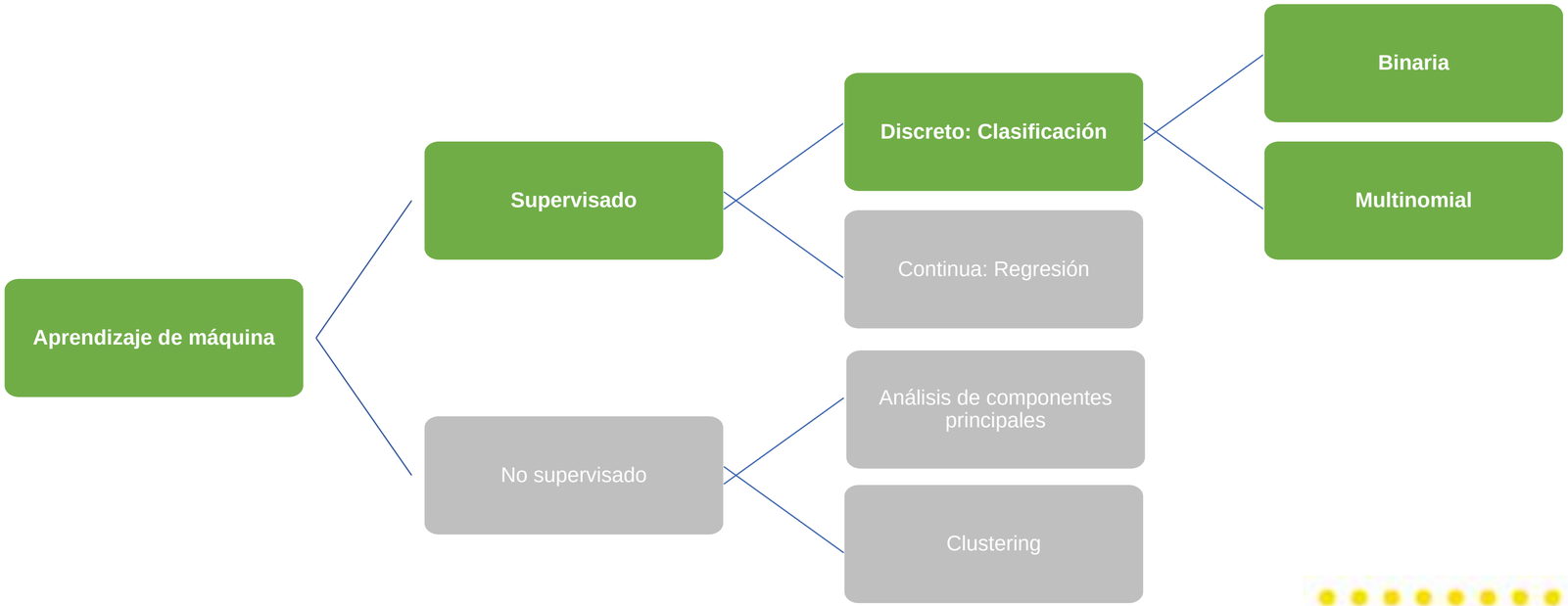
Al final de la clase de hoy

1. Identificar situaciones en las que es razonable probar árboles de decisión

1. Identificar la diferencia entre un Árbol de decisión y Bosques aleatorios.

2. Comprender los principales componentes del algoritmo, sus alcances y limitaciones.

3. Generar un modelo de Bosques Aleatorios sencillo en Python.



Un árbol de decisión es un modelo supervisado de ML.

Árboles de decisión

Los datos
- Clasificación binaria
- Visualización de los datos de hoy

Cómo clasifica un árbol
- Árbol de decisión

Cómo entrenamos el árbol
- La intuición
- Entropía
- Ganancia de información

Ejercicio

Casado	Edad	Ingresos
	24	3.5
	18	0.7
	26	4
	19	2
	36	5.6
	29	7.1
	42	6.2
	32	4.1
	33	6.2

Recordemos: si estamos en un problema de clasificación binaria, ¿qué valores debería tener nuestra columna a predecir?

Tomen 2 minutos para escribr en el chat y hablamos

Vamos a recuperar sus ideas y volvemos al final sobre esta pregunta

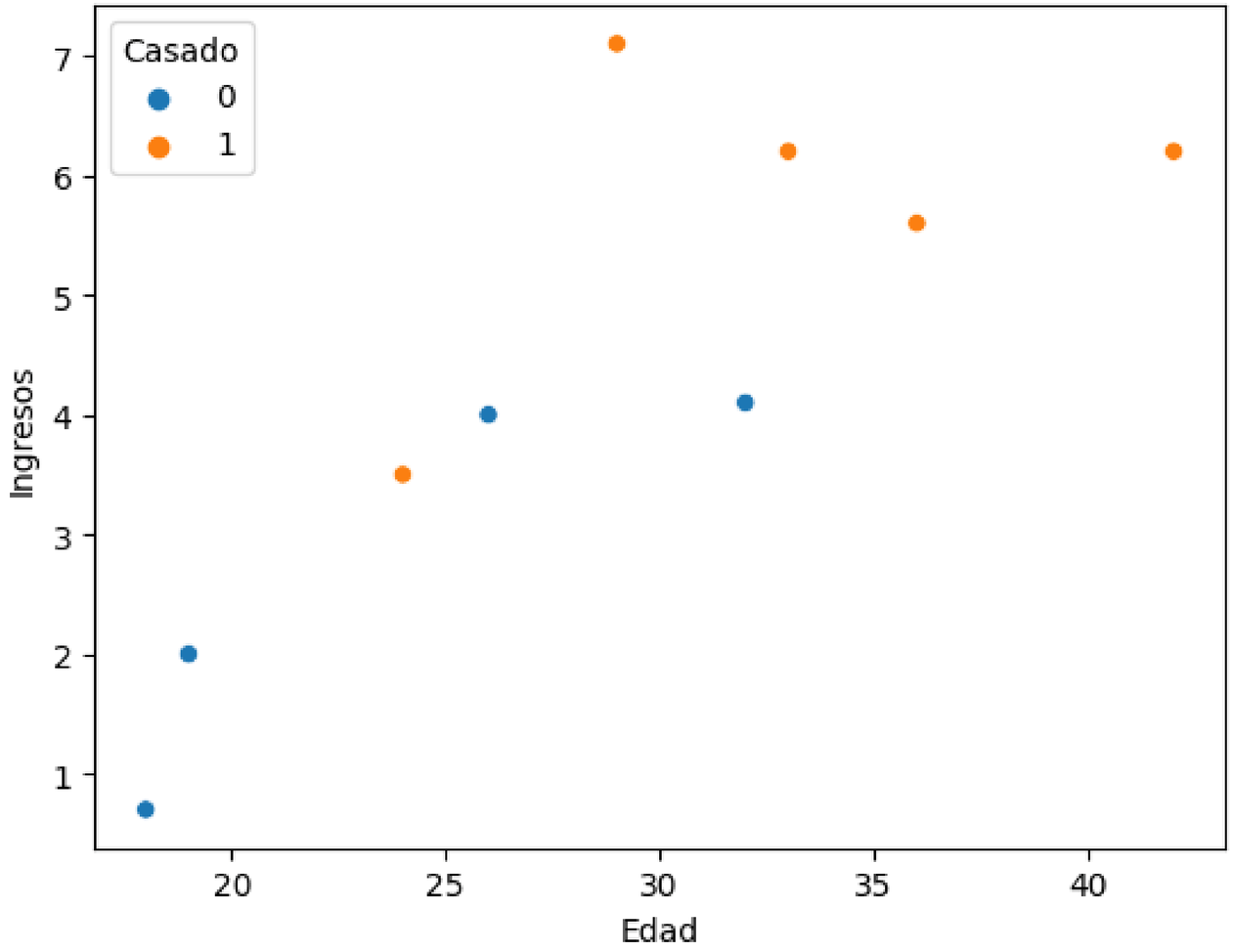
Volvemos

Ceros y unos. Variable *dummy* que indica la categoría a la que pertenece la observación.

Vamos a tratar de clasificar una persona entre **casada** y no casada a partir de la **edad** y los **ingresos**.

Casado	Edad	Ingresos
1	24	3.5
0	18	0.7
0	26	4
0	19	2
1	36	5.6
1	29	7.1
1	42	6.2
0	32	4.1
1	33	6.2

Cómo se ven los *datos*

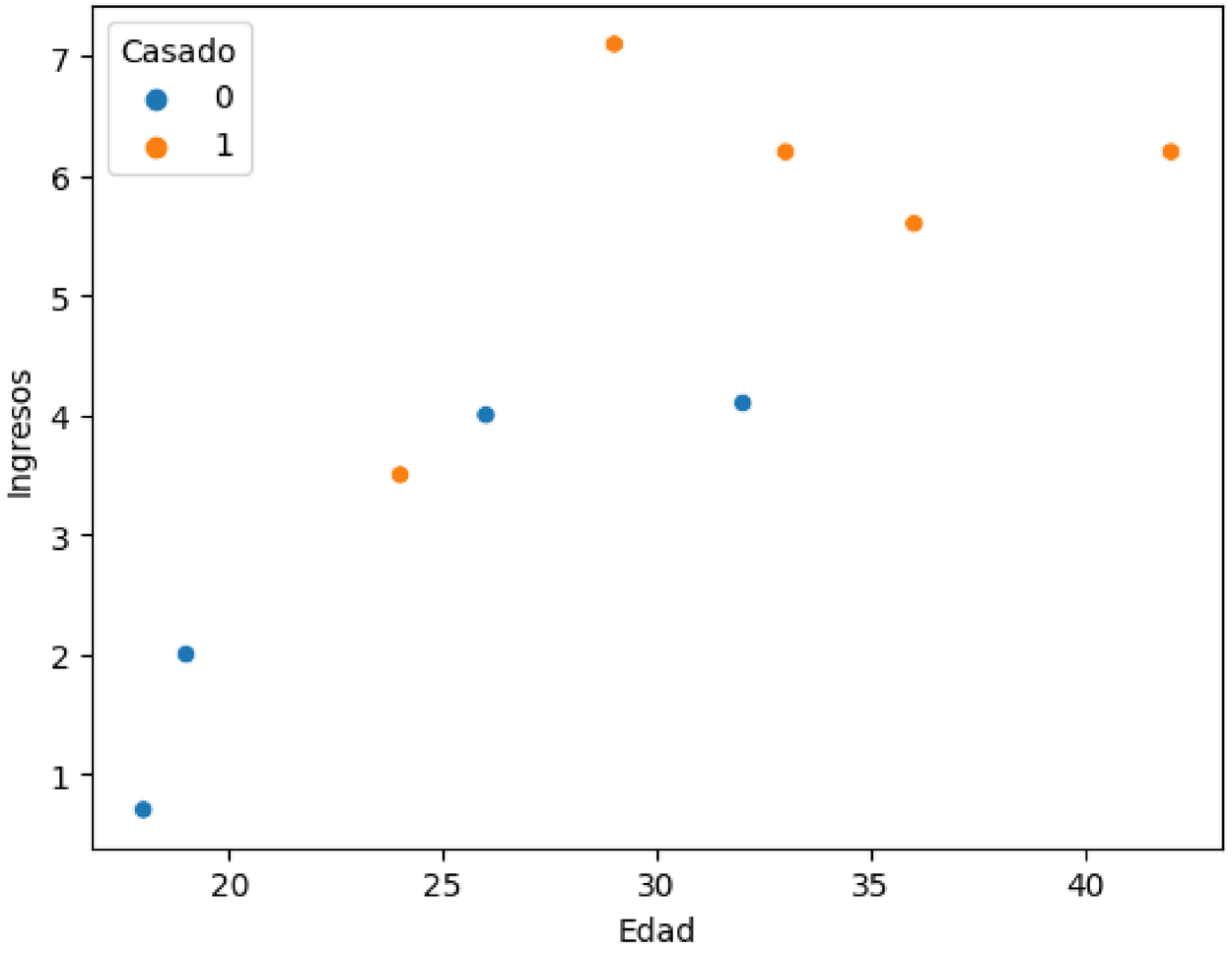


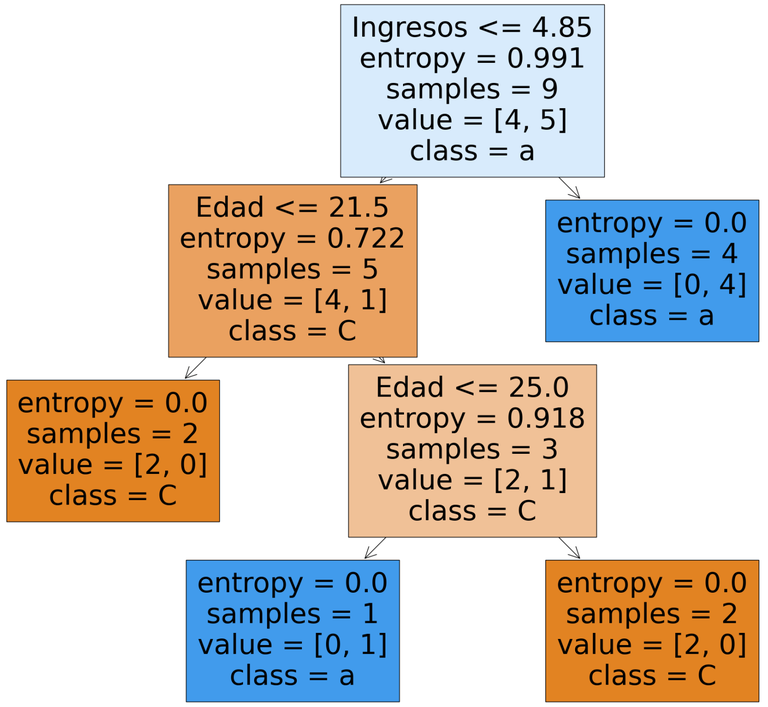
Esta visualización de los datos nos va a ser útil para entender cómo funcionan los árboles de decisión.

En los ejes tenemos dos variables predictoras (equis).

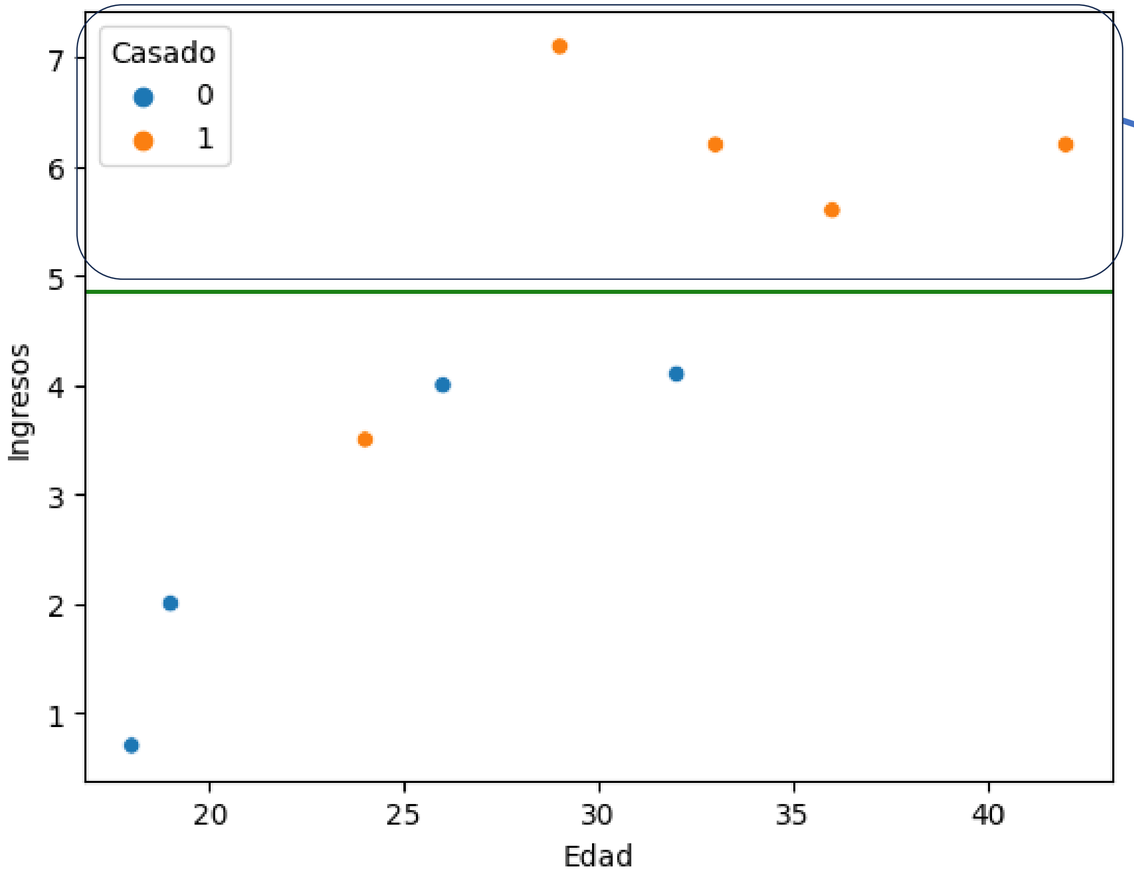
Indicado por el color, tenemos la categoría que queremos predecir.

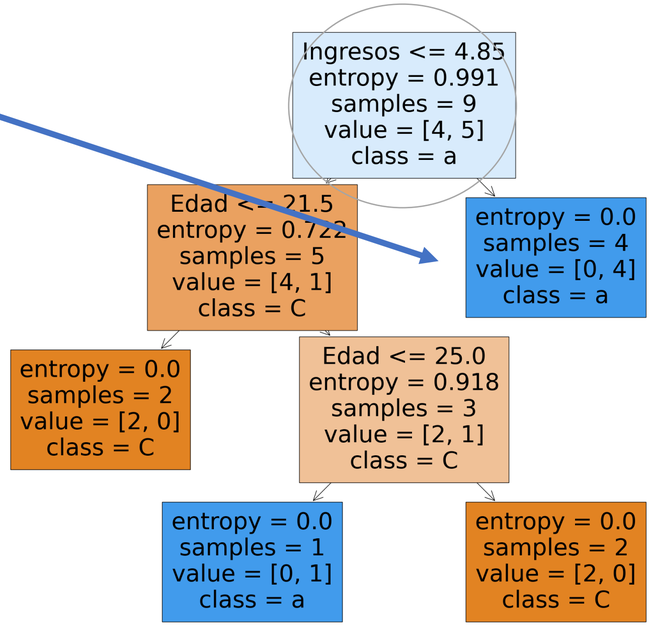
Hay nodos hojas y nodos de decisión



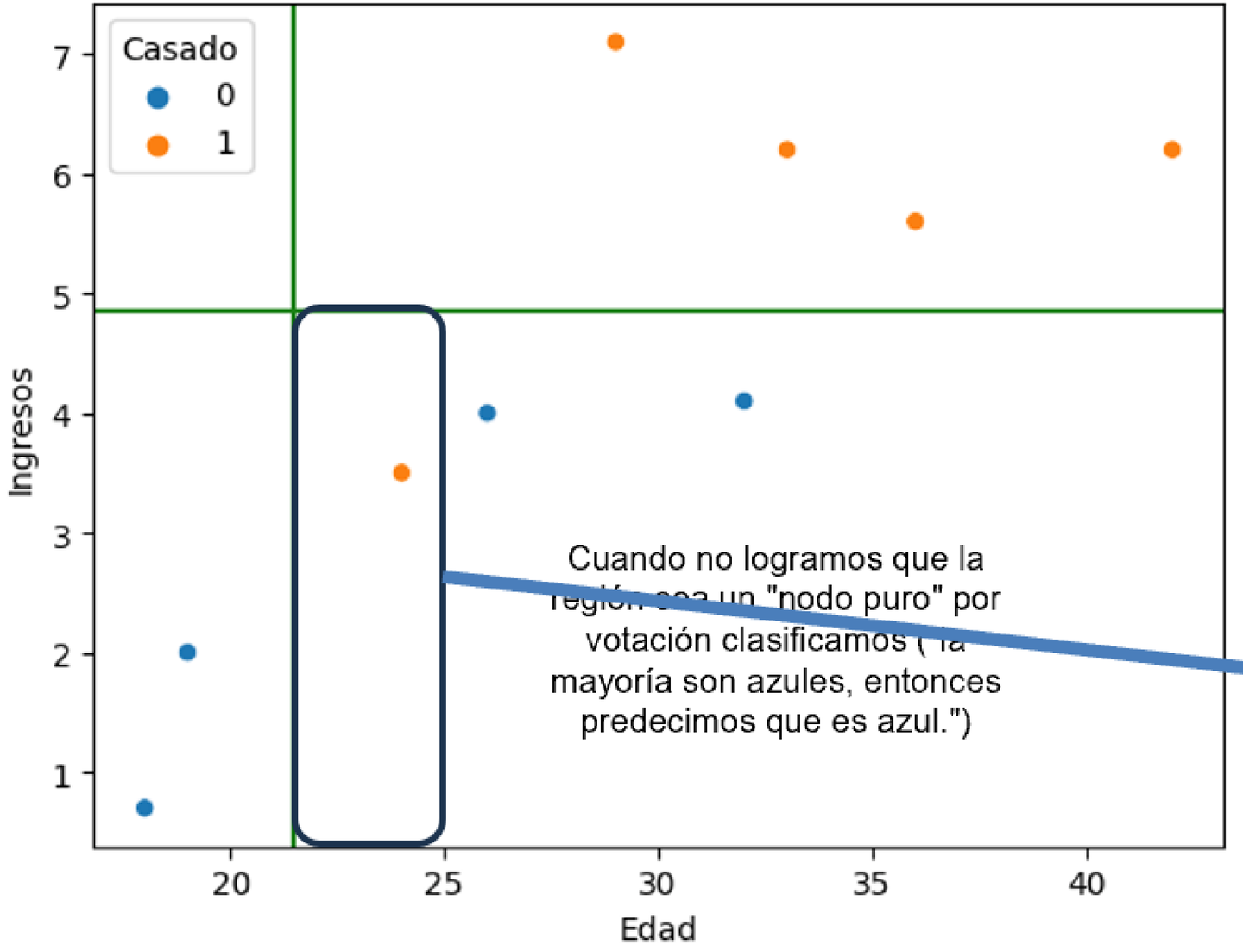


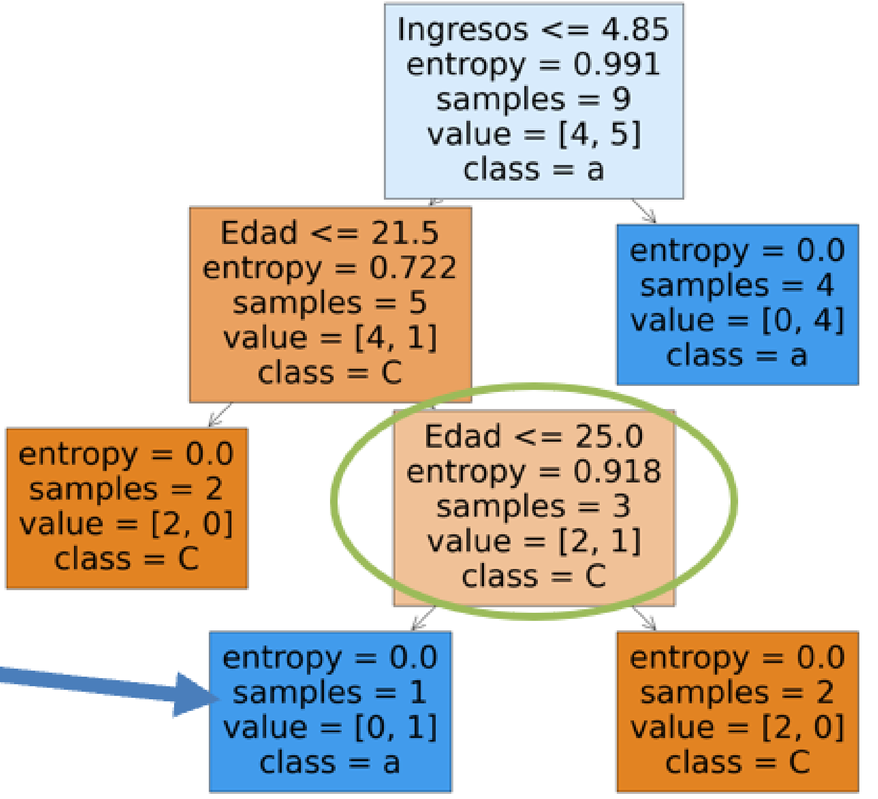
El primer nodo de decisión





El tercer nodo de decisión



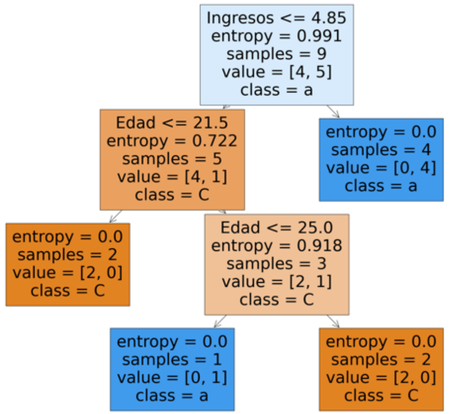


Noten que el árbol de decisión separa en regiones el espacio de clasificación.

Entrenamiento de un árbol

CoronaNet

¿Qué magia tiene?



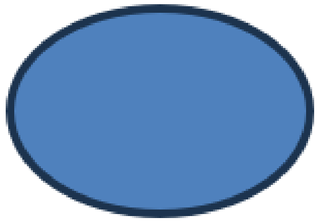
- ¿No es acaso un árbol de decisión un montón de expresiones de tipo if... else...?
- Sí. Es exactamente eso.
- **Lo difícil es escoger las preguntas correctas**.

Ingresos <= 4.85

¿Por qué no 3, 6, 5...? Para esto usamos los datos. Esto significa **entrenar**.

Criterios

- Hay varios criterios que usamos para seleccionar un árbol de decisión. Un criterio es **un valor (una cantidad) que queremos maximizar**.



Nivel de ____: 0.6

- Sea cual sea el criterio, la intuición es la siguiente:
  - Medimos el nivel en el que está el criterio inicialmente.
  - Probamos con **todas las posibles combinaciones de preguntas** (hacemos preguntas sobre todas las X, y con todos los umbrales).
  - Escogemos la pregunta que maximiza **la ganancia** en nuestro criterio: **el criterio evaluado en el nodo padre, menos todos los valores ponderados en los nodos hijos.**

Si preguntamos $X_1 < 5$ El nivel de ____ queda en **0.73**

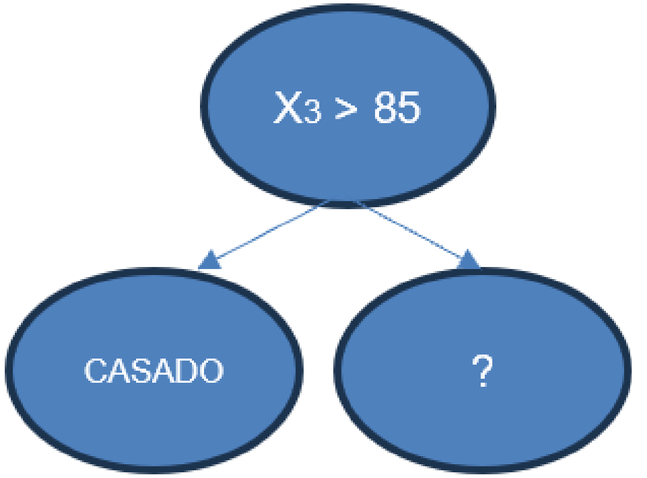
Mejor hagamos esta pregunta

Si preguntamos $X_3 > 85$ El nivel de ____ queda en **0.82**

Criterios

- Hay varios criterios que usamos para seleccionar un árbol de decisión. Un criterio es **un valor (una cantidad) que queremos maximizar**.

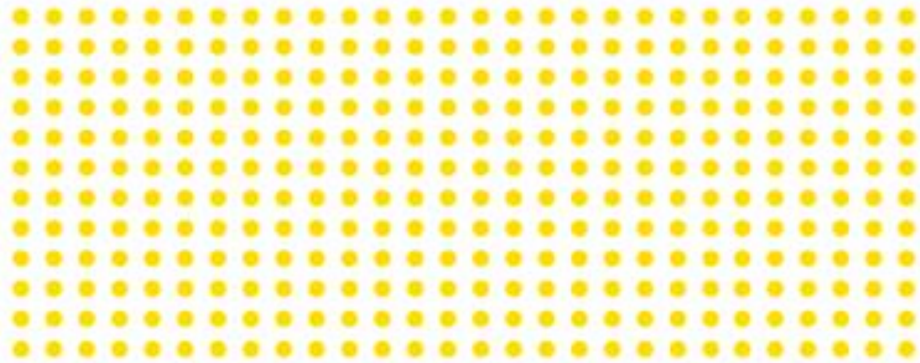
- Sea cual sea el criterio, la intuición es la siguiente:
  - Medimos el nivel en el que está el criterio inicialmente.
  - Probamos con **todas las posibles combinaciones de preguntas** (hacemos preguntas sobre todas las X, y con todos los umbrales).
  - Escogemos la pregunta que maximiza **la ganancia** en nuestro criterio: **el criterio evaluado en el nodo padre, menos todos los valores ponderados en los nodos hijos.**



Y repetimos el proceso hasta que nuestras hojas sean puras, o alcancemos la cantidad máxima de niveles indicados.

¿Por qué Bosques?, ¿Por qué Aleatorios? y ¿Por qué me habría de importar?

Bosques Aleatorios

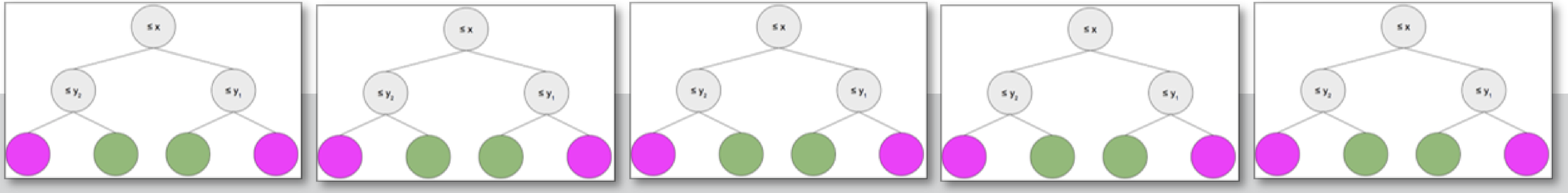


¿Por qué bosques?

Los árboles de decisión son altamente susceptibles a la información del entrenamiento, lo que los hace tener una variación muy alta y afecta la posibilidad de generalizar.

Entonces vamos a crear un bosque, en otras palabras… muchos árboles de decisión.

De esta forma no es tan sensible al entrenamiento.



¿Por qué aleatorios?

Al crear un bosque aleatorio vamos a dividir nuestro conjunto de datos (DataSet) en pequeñas partes.

Muestreo con reemplazo (Bootstrapping)

Consiste en un método de remuestreo, donde se asume que la población se puede representar por un nuevo muestreo de los datos.

Las nuevas muestras van a tener la misma cantidad de registros que el original.

Se genera un muestreo con reemplazo, es decir que podemos duplicar ciertos datos.

¿Por qué aleatorios?

Al crear un bosque aleatorio vamos a dividir nuestro conjunto de datos (DataSet) en pequeñas partes.

Muestreo de características

Se selecciona un subconjunto aleatorio de características del conjunto.

Esto ayuda a que un conjunto de características dominantes no influya demasiado en un árbol.

Esto va a ayudar a la capacidad de generalizar.

Bootstrapping

ID	X1	X2	X3	X4	X5	Y
1	74	81	1	37	97	31
2	11	6	54	95	6	75
3	90	4	81	30	98	95
4	92	68	43	10	86	78
5	74	10	43	91	85	93
6	31	71	22	77	18	76
7	43	35	93	73	54	48
8	55	64	42	60	8	80
9	97	1	19	64	69	65
10	26	5	61	48	39	88

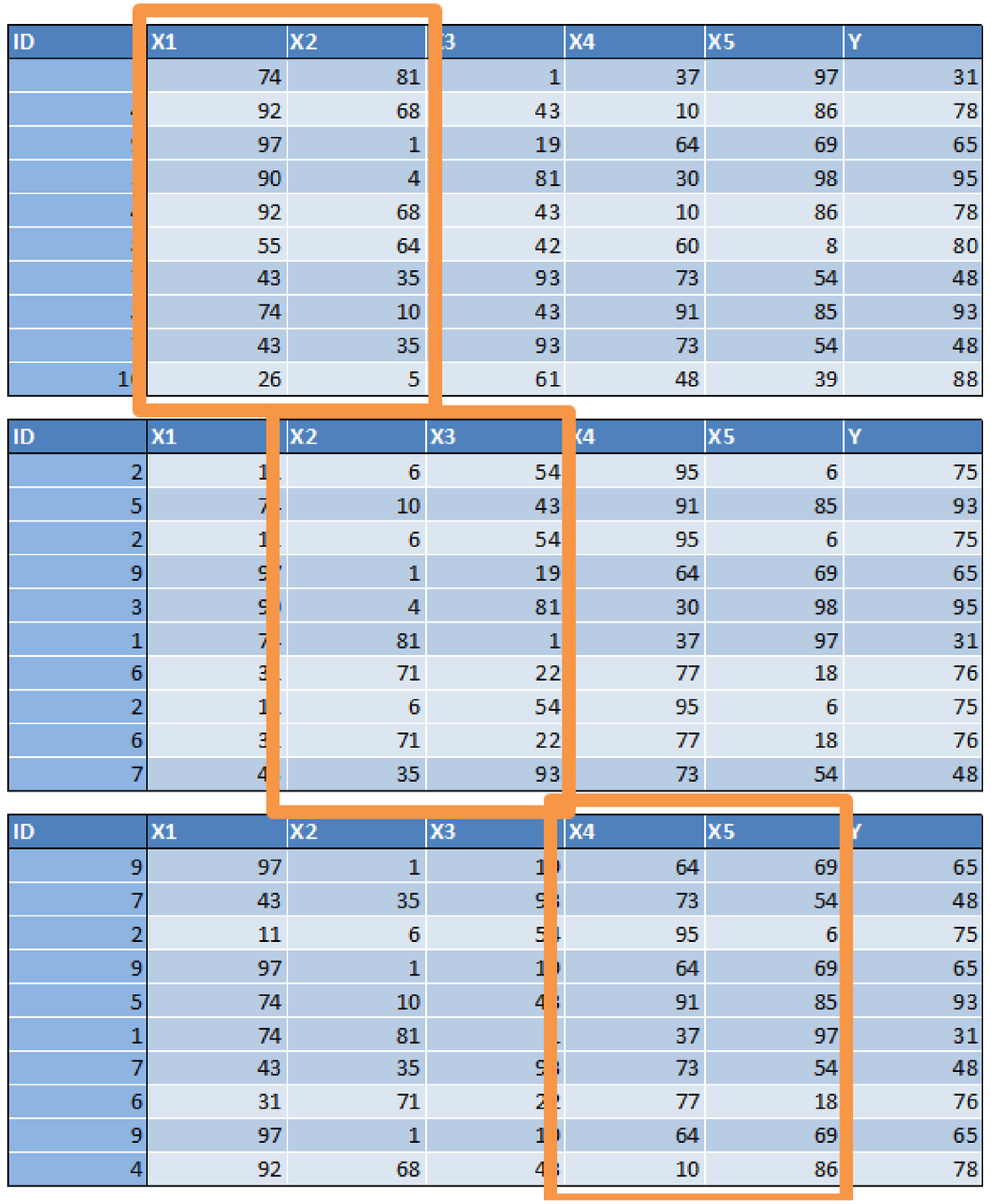
ID	X1	X2	X3	X4	X5	Y
1	74	81	1	37	97	31
4	92	68	43	10	86	78
9	97	1	19	64	69	65
3	90	4	81	30	98	95
4	92	68	43	10	86	78
8	55	64	42	60	8	80
7	43	35	93	73	54	48
5	74	10	43	91	85	93
7	43	35	93	73	54	48
10	26	5	61	48	39	88

ID	X1	X2	X3	X4	X5	Y
2	11	6	54	95	6	75
5	74	10	43	91	85	93
2	11	6	54	95	6	75
9	97	1	19	64	69	65
3	90	4	81	30	98	95
1	74	81	1	37	97	31
6	31	71	22	77	18	76
2	11	6	54	95	6	75
6	31	71	22	77	18	76
7	43	35	93	73	54	48

ID	X1	X2	X3	X4	X5	Y
9	97	1	19	64	69	65
7	43	35	93	73	54	48
2	11	6	54	95	6	75
9	97	1	19	64	69	65
5	74	10	43	91	85	93
1	74	81	1	37	97	31
7	43	35	93	73	54	48
6	31	71	22	77	18	76
9	97	1	19	64	69	65
4	92	68	43	10	86	78

Muestreo de características

ID	X1	X2	X3	X4	X5	Y
1	74	81	1	37	97	31
2	11	6	54	95	6	75
3	90	4	81	30	98	95
4	92	68	43	10	86	78
5	74	10	43	91	85	93
6	31	71	22	77	18	76
7	43	35	93	73	54	48
8	55	64	42	60	8	80
9	97	1	19	64	69	65
10	26	5	61	48	39	88



ID	X1	X2	X3	X4	X5	Y
	74	81	1	37	97	31
	92	68	43	10	86	78
	97	1	19	64	69	65
	90	4	81	30	98	95
	92	68	43	10	86	78
	55	64	42	60	8	80
	43	35	93	73	54	48
	74	10	43	91	85	93
	43	35	93	73	54	48
1	26	5	61	48	39	88

ID	X1	X2	X3	X4	X5	Y
2	1	6	54	95	6	75
5	7	10	43	91	85	93
2	1	6	54	95	6	75
9	9	1	19	64	69	65
3	9	4	81	30	98	95
1	7	81	1	37	97	31
6	3	71	22	77	18	76
2	1	6	54	95	6	75
6	3	71	22	77	18	76
7	4	35	93	73	54	48

ID	X1	X2	X3	X4	X5	Y
9	97	1	19	64	69	65
7	43	35	93	73	54	48
2	11	6	54	95	6	75
9	97	1	19	64	69	65
5	74	10	43	91	85	93
1	74	81	1	37	97	31
7	43	35	93	73	54	48
6	31	71	22	77	18	76
9	97	1	19	64	69	65
4	92	68	43	10	86	78

¿Por qué me habría de importar?

Ahora vamos a crear un árbol de cada una de estas muestras.

En este ejemplo tendríamos 3 árboles, pero comúnmente se puede hacer el modelo con 100 o más árboles.

Luego lo que faltaría es juntar los diferentes modelos para dar un veredicto al ingresar información nueva.

Combinación de modelos

Para sacar el resultado final de nuestro Bosque aleatorio debemos combinar nuestros modelos independientes.

La forma más simple es:

- En clasificación: Se obtiene la moda de los resultados
- En regresión: Se obtiene el promedio de los resultados

Existen otras formas más complejas, pero eso les queda de tarea.

Ventajas

- Reducción del sobreajuste
- Buena generalización
- Limita el efecto de características dominantes
- Limita el efecto de datos atípicos o de ruido
- Interpretables (bueno más o menos)
- Puede predecir variables categóricas y cuantitativas

Desventajas

- Mayor gasto computacional
- Menos interpretables que un árbol
- Mayor gasto en la memoria
- Para problemas donde los datos y la relación entre estos es compleja no es buena opción

Con todo esto…

1. Conocemos la diferencia entre árboles de decisión y bosques aleatorios.

2. Entendemos que se pueden combinar modelos para generar un nuevo modelo mejor.

3. Entendemos cómo trabajar con bosques aleatorios.

¡Gracias!

Aprendiendo juntos a lo largo de la vida

educacioncontinua.uniandes.edu.co

Síguenos: **EdcoUniandes**

Universidad de los Andes | Educación Continua
Vicerrectoría Académica