Arboles de clasificación.

César Gómez

21 de octubre de 2020

Arboles de clasificación

- Un árbol de clasificación es similar a un árbol de regresión excepto que el árbol es utilizado para predecir una respuesta cualitativa (categórica) en vez de una respuesta cuantitativa.

Arboles de clasificación

- Un árbol de clasificación es similar a un árbol de regresión excepto que el árbol es utilizado para predecir una respuesta cualitativa (categórica) en vez de una respuesta cuantitativa.
- En un árbol de clasificación, la predicción para una observación particular que se encuentra en una determinada rama o nodo terminal del árbol corresponde a la clase que más ocurre dentro de dicha rama (*voto de mayoría*).

Arboles de clasificación

- Un árbol de clasificación es similar a un árbol de regresión excepto que el árbol es utilizado para predecir una respuesta cualitativa (categórica) en vez de una respuesta cuantitativa.
- En un árbol de clasificación, la predicción para una observación particular que se encuentra en una determinada rama o nodo terminal del árbol corresponde a la clase que más ocurre dentro de dicha rama (*voto de mayoría*).
- Al interpretar los resultados de un árbol de clasificación, no solo la clase predicha para un nodo o región terminal es de interés, si no también la proporción de cada clase dentro de dicho nodo o región.

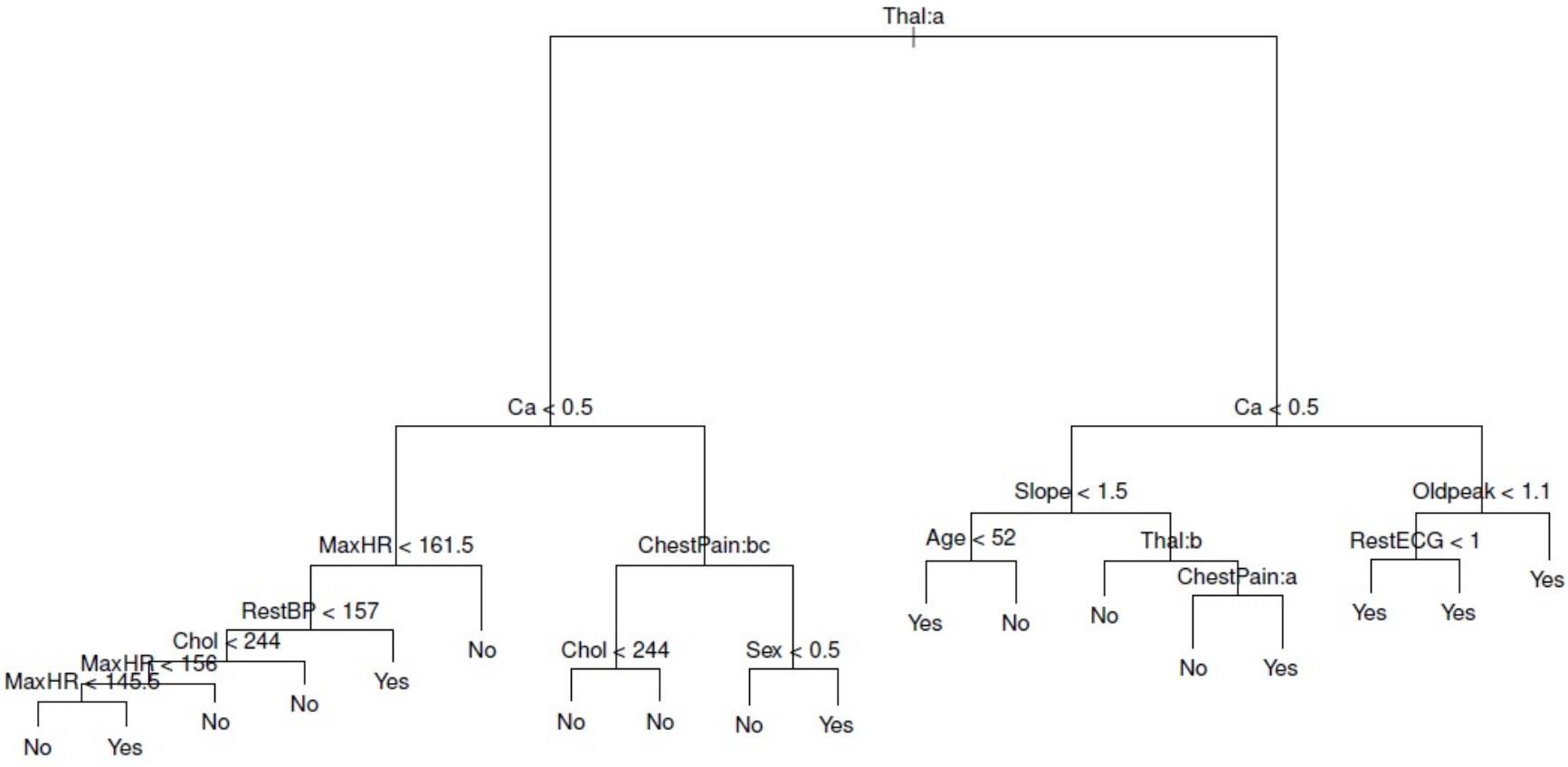

Figura 1: La base de datos "`Heart`" contiene una respuesta binaria `HD` para 303 pacientes que presentan dolor de pecho. "`yes`" indica la presencia de enfermedad coronaria. Hay 13 predictores: `Edad`, `Chol`(una medida de colesterol) y otras mediciones relacionadas con el funcionamiento del corazón y pulmones. La validación cruzada resulta en un árbol con 6 nodos terminales.

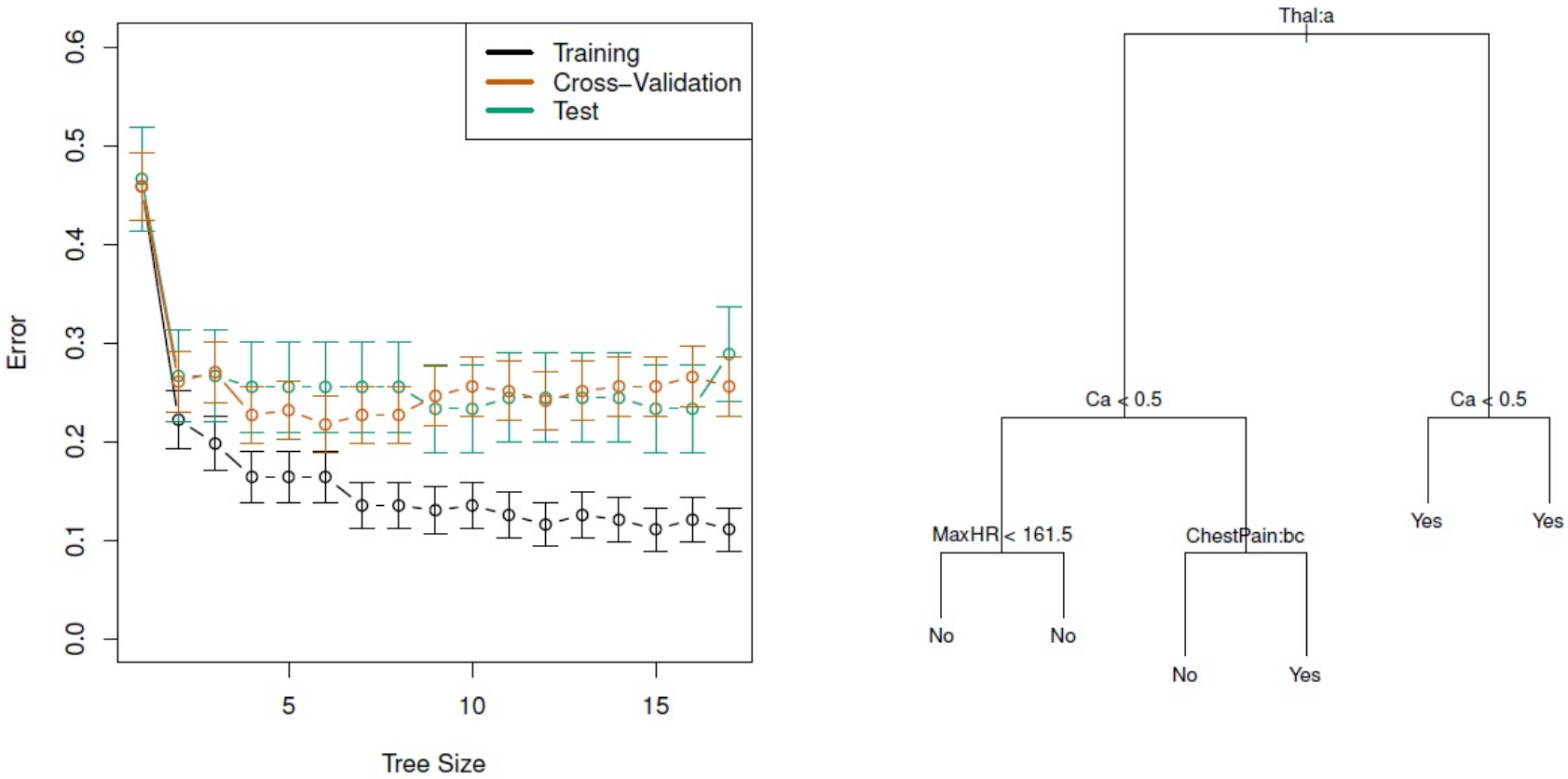

Figura 2: Validación cruzada de un árbol de clasificación sobre el conjunto de datos `Heart`. A la derecha el árbol podado correspondiente al menor valor del error de validación cruzada.

- Como en el caso de un árbol de regresión se utiliza *particionamiento binario recursivo* para generar un árbol de clasificación.

- Como en el caso de un árbol de regresión se utiliza *particionamiento binario recursivo* para generar un árbol de clasificación.
- Pero en este caso *RSS* no se puede utilizar como criterio para los particionamientos correspondientes a los nodos del árbol. En su lugar se debe utilizar la **tasa del error de clasificación.** Que corresponde a la fracción de clases que no corresponden a la clase de mayoría dominante en la correspondiente hoja.

$$E = 1 - \max_k(\hat{p}_{mk}), \tag{1}$$

acá $\hat{p}_{mk}$ corresponde a la proporción de observaciones de entrenamiento en la m-ésima región que corresponden a la k-ésima clase.

En el caso de árboles de clacificación no solo hay interés en la predicción de la clase en cada región, si no tambieén en las proporciones en cada clase

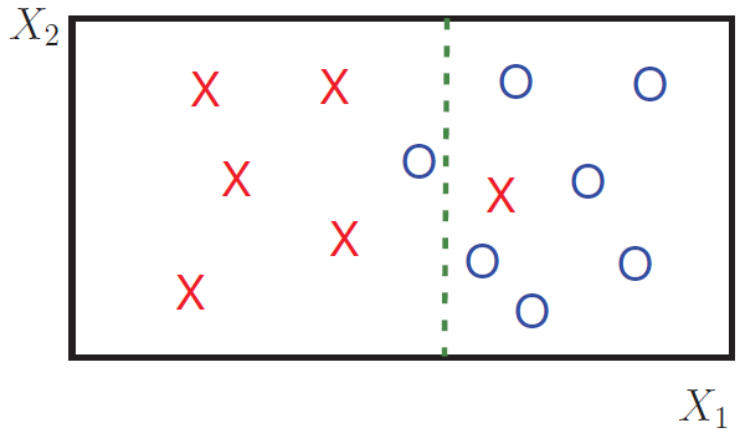

Figura 3:

Pero sucede que la tasa de error no es lo suficiente sensitiva al crecimiento de los árboles y en la práctica otras medidas son preferidas, en particular el *indice de Gini* y la *entropía*

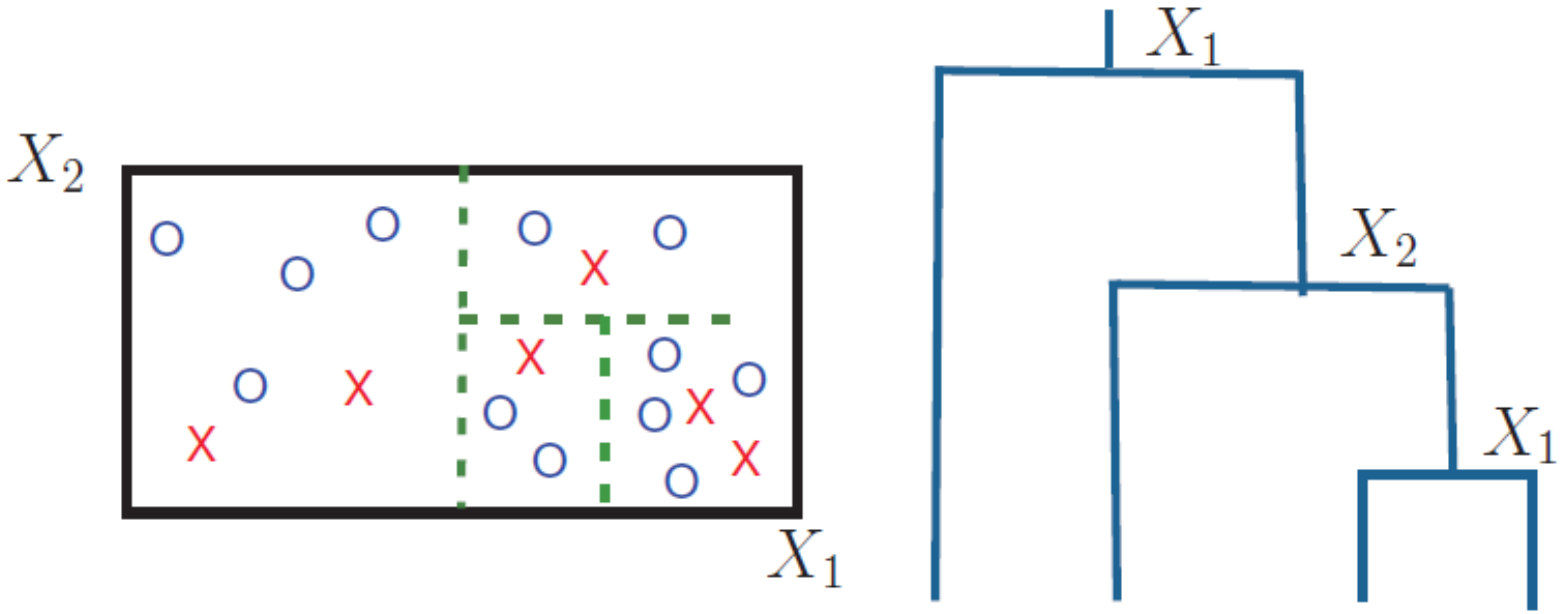

Figura 4:

Indice de Gini, entropía

1. El **índice de Gini** es definido por

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2}$$

Es una medida de la varianza total de todas las clases. El índice de Gini toma valores pequeños si todas las proporciones $\hat{p}_{mk}$ toman valores cercanos a cero 0 ó uno 1. Por esta razón el índice de Gini es preferido como medida de la pureza de un nodo. Un valor pequeño indica que se tiene un nodo terminal en que predominan observaciones de una misma clase.

2. Una alternativa al **índice de Gini** la constituye la **entropía** que viene dada por

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{3}$$

Como $0 \leq \hat{p}_{mk} \leq 1$, se tiene que

$$0 \leq -\hat{p}_{mk} \log \hat{p}_{mk},$$

la *entropía* toma un valor pequeño cada vez que las proporciones $\hat{p}_{mk}$ toman valores cercanos a cero 0 o uno 1.

El índice de Gini y la entropía son numéricamente similares.

Ejemplo. El conjunto de datos `Heart`

Heart

Filter

	X	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
1	1	63	1	typical	145	233	1	2	150	0	2.3	3	0	fixed	No
2	2	67	1	asymptomatic	160	286	0	2	108	1	1.5	2	3	normal	Yes
3	3	67	1	asymptomatic	120	229	0	2	129	1	2.6	2	2	reversable	Yes
4	4	37	1	nonanginal	130	250	0	0	187	0	3.5	3	0	normal	No
5	5	41	0	nontypical	130	204	0	2	172	0	1.4	1	0	normal	No
6	6	56	1	nontypical	120	236	0	0	178	0	0.8	1	0	normal	No
7	7	62	0	asymptomatic	140	268	0	2	160	0	3.6	3	2	normal	Yes
8	8	57	0	asymptomatic	120	354	0	0	163	1	0.6	1	0	normal	No
9	9	63	1	asymptomatic	130	254	0	2	147	0	1.4	2	1	reversable	Yes
10	10	53	1	asymptomatic	140	203	1	2	155	1	3.1	3	0	reversable	Yes

El conjunto de datos `Heart`

- Estos datos contienen un resultado binario de `HD` para 303 pacientes que presentaron dolor de pecho.
- Un valor de `HD` de `Yes` indica la presencia de enfermedad cardíaca según una prueba angiográfica, mientras que `No` significa que no hay enfermedad.
- Hay 13 predictores incluyendo `Age`,`Sex`,`Chol`(una medida de colesterol) y otros factores cardíacos y mediciones de la función pulmonar.
- La validación cruzada resulta en un árbol con seis nodos terminales.

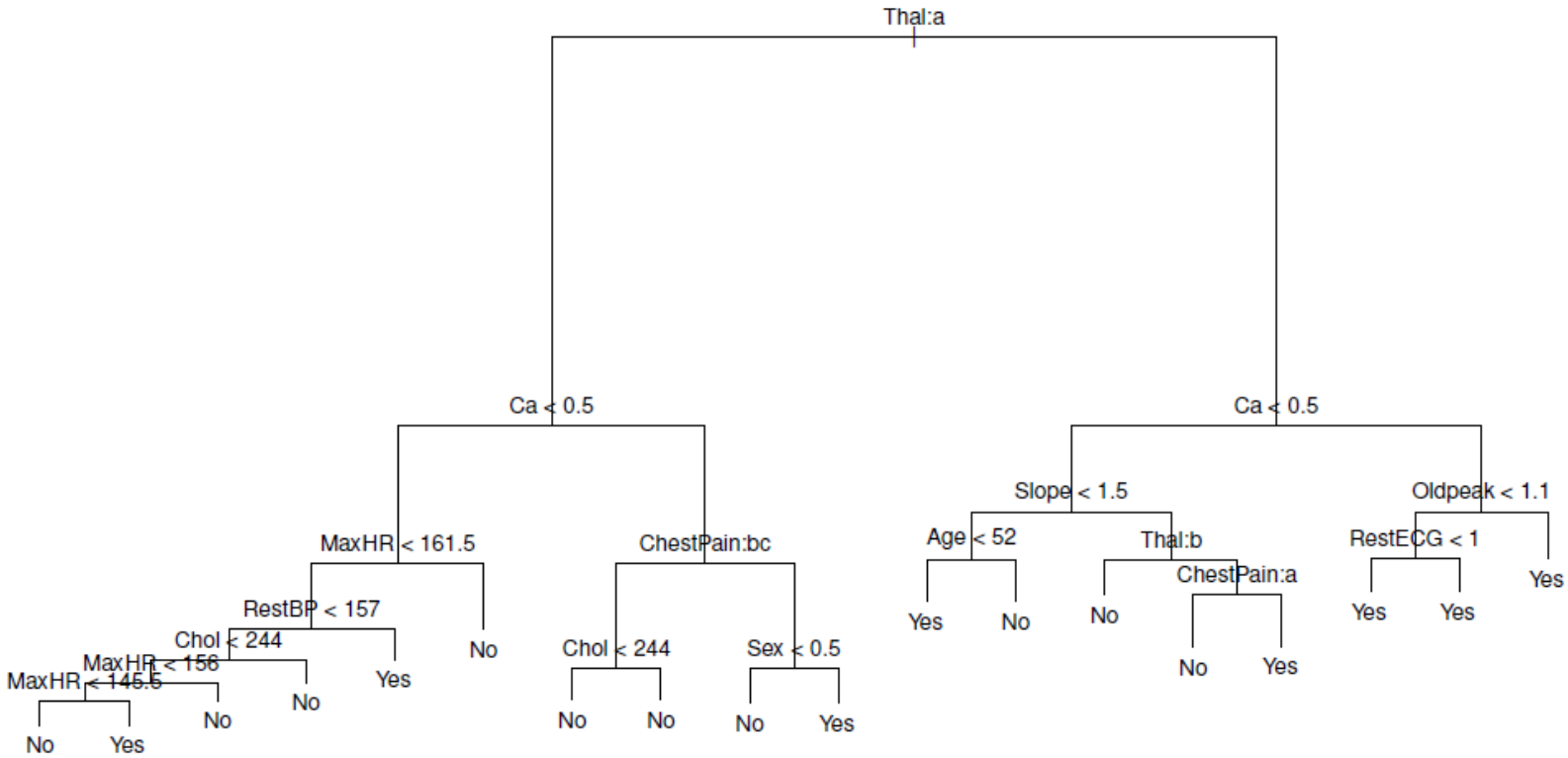

Figura 5: El árbol no podado ajustado al conjunto de datos `Heart`.

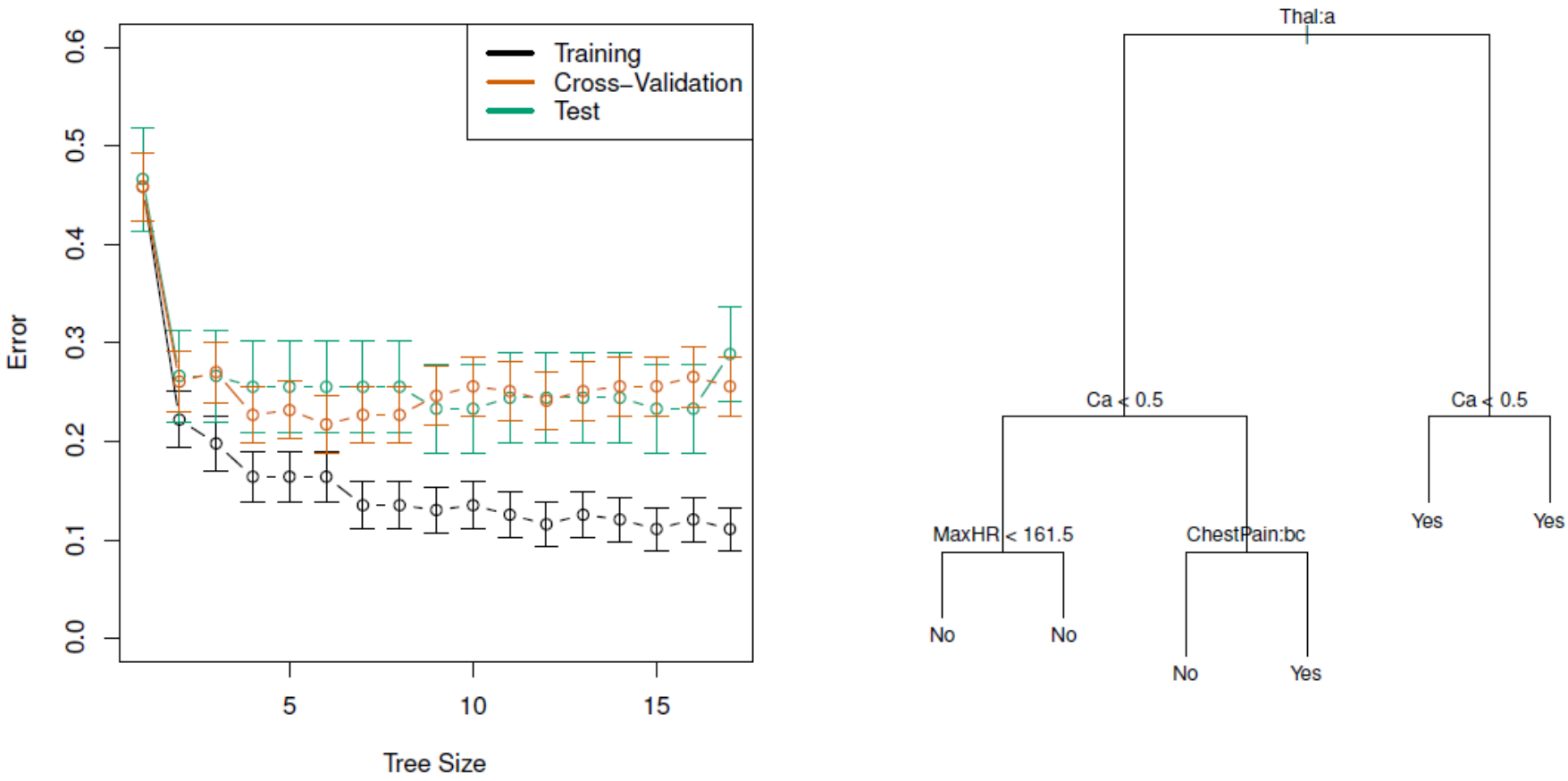

Figura 6: **Izquierda** Errores de validación cruzada, de entrenamiento y de prueba para diferentes tamaños del árbol podado. **Derecha** El árbol podado correspondiente al menor error de validación cruzada.

- Al revisar La figura (5) sorprende que algunas de las divisiones produscan dos nodos terminales que con el mismo valor de predicción.
- Por ejemplo, considere la división $\texttt{RestECG} < 1$ cerca de la parte inferior derecha del árbol sin podar. Independientemente del valor de `RestECG`, se predice un valor de respuesta de Sí para esas observaciones.
- ¿Por qué, entonces, se realiza la división? La división se realiza porque conduce a una mayor pureza del nodo.

- Es decir, las 9 observaciones correspondiente a la hoja de la derecha tienen un valor de respuesta de `Yes`, mientras que 7/11 de los correspondientes a la hoja de la izquierda tambień tienen un valor de respuesta `Yes`.

- Es decir, las 9 observaciones correspondiente a la hoja de la derecha tienen un valor de respuesta de `Yes`, mientras que 7/11 de los correspondientes a la hoja de la izquierda tambień tienen un valor de respuesta `Yes`.
- ¿Por qué es importante la pureza del nodo? Supongase que se tiene una observación de prueba que pertenece a la región dada por esa hoja de la derecha. Entonces uno puede estar bastante seguro de que su valor de respuesta es `Yes`.

- Es decir, las 9 observaciones correspondiente a la hoja de la derecha tienen un valor de respuesta de `Yes`, mientras que 7/11 de los correspondientes a la hoja de la izquierda tambień tienen un valor de respuesta `Yes`.
- ¿Por qué es importante la pureza del nodo? Supongase que se tiene una observación de prueba que pertenece a la región dada por esa hoja de la derecha. Entonces uno puede estar bastante seguro de que su valor de respuesta es `Yes`.
- Por el contrario, si una observación de prueba pertenece a la región dada por la hoja de la izquierda, entonces su respuesta probablemente sea Sí, pero estamos mucho menos seguros.

- Es decir, las 9 observaciones correspondiente a la hoja de la derecha tienen un valor de respuesta de `Yes`, mientras que 7/11 de los correspondientes a la hoja de la izquierda tambień tienen un valor de respuesta `Yes`.
- ¿Por qué es importante la pureza del nodo? Supongase que se tiene una observación de prueba que pertenece a la región dada por esa hoja de la derecha. Entonces uno puede estar bastante seguro de que su valor de respuesta es `Yes`.
- Por el contrario, si una observación de prueba pertenece a la región dada por la hoja de la izquierda, entonces su respuesta probablemente sea Sí, pero estamos mucho menos seguros.
- Aunque la división $\texttt{RestECG} < 1$ no reduce el error de clasificación, mejora el Índice de Gini y la entropía cruzada, que son más sensibles a la pureza de un nodo.

Árboles vs Modelos lineales

Un modelo de regresión lineal, posee la forma

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{4}$$

mientras que un árbol de regresión

$$f(X) = \sum_{m=1}^{M} c_m \cdot 1_{(X \in R_m)}, \tag{5}$$

¿Cual de los 2 modelos es mejor?
Rta. **Depende del problema que se tenga en manos.**

- Si la relación entre la respuesta y los predictores es bien aproximada por un modelo lineal, entonces el uso de un modelo lineal funcionará bien y tendrá un mejor desempeño que un método basado en árboles de decisión que no explotan esta estructura lineal de los datos.

- Si la relación entre la respuesta y los predictores es bien aproximada por un modelo lineal, entonces el uso de un modelo lineal funcionará bien y tendrá un mejor desempeño que un método basado en árboles de decisión que no explotan esta estructura lineal de los datos.
- Más si en cambio, la relación entre la variable respuesta y los predictores es marcadamente no-lineal y más compleja los árboles de decisión tendrán mejor desempeño que los métodos clásicos.

- Si la relación entre la respuesta y los predictores es bien aproximada por un modelo lineal, entonces el uso de un modelo lineal funcionará bien y tendrá un mejor desempeño que un método basado en árboles de decisión que no explotan esta estructura lineal de los datos.
- Más si en cambio, la relación entre la variable respuesta y los predictores es marcadamente no-lineal y más compleja los árboles de decisión tendrán mejor desempeño que los métodos clásicos.
- Por supuesto hay otras consideraciones más allá de optimizar un error de prueba a la hora de seleccionar un método de aprendizaje estadístico; por ejemplo, en ciertos contextos utilizar un árbol puede ser más ventajoso en relación a la interpretabilidad y la visualización.

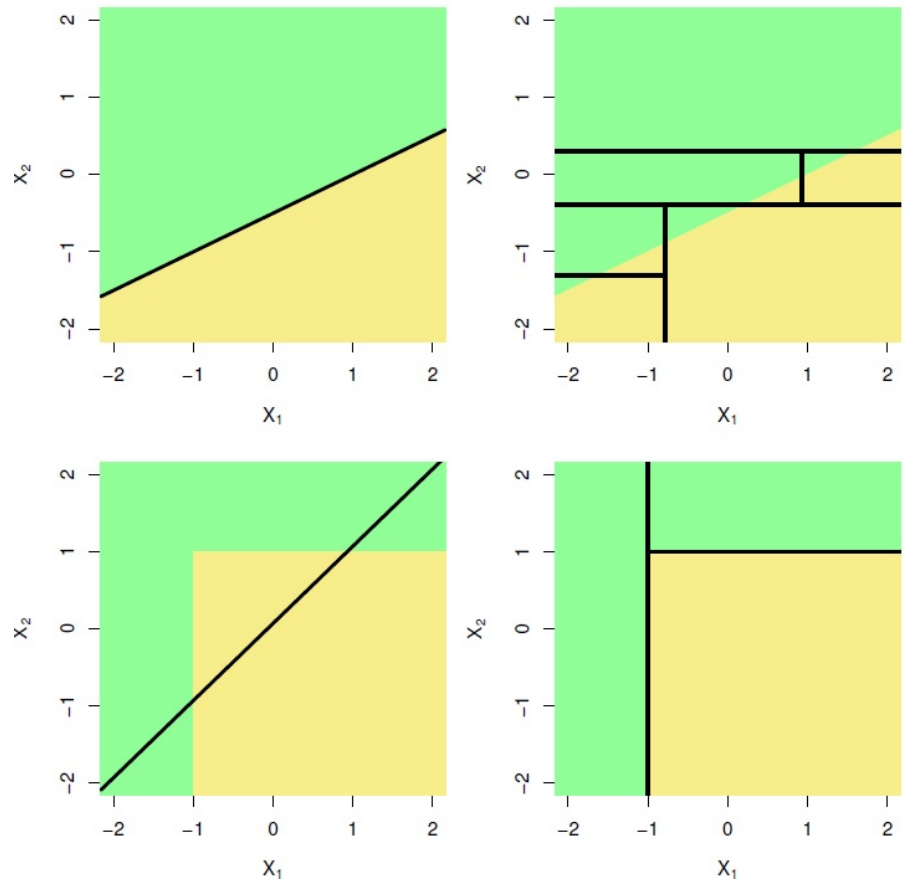

Figura 7: Un ejemplo de clacificación con 2 predictores, en unos casos un modelo lineal supera un árbol de clacificación y viceversa.

Ventajas y Desventajas de los árboles de decisión

1. Los árboles son fáciles de explicar. De hecho, son más fáciles de explicar que la regresión lineal.

Ventajas y Desventajas de los árboles de decisión

1. Los árboles son fáciles de explicar. De hecho, son más fáciles de explicar que la regresión lineal.
2. Algunas personas piensan que los árboles de decisión simulan más de cerca el proceso de toma de decisiones de los humanos que los abordajes de predicción y clasificación vistos con anterioridad en el curso.

Ventajas y Desventajas de los árboles de decisión

1. Los árboles son fáciles de explicar. De hecho, son más fáciles de explicar que la regresión lineal.
2. Algunas personas piensan que los árboles de decisión simulan más de cerca el proceso de toma de decisiones de los humanos que los abordajes de predicción y clasificación vistos con anterioridad en el curso.
3. Los árboles pueden ser representados gráficamente y son fáciles de interpretar por no expertos, particularmente si los árboles son pequeños.

Ventajas y Desventajas de los árboles de decisión

1. Los árboles son fáciles de explicar. De hecho, son más fáciles de explicar que la regresión lineal.
2. Algunas personas piensan que los árboles de decisión simulan más de cerca el proceso de toma de decisiones de los humanos que los abordajes de predicción y clasificación vistos con anterioridad en el curso.
3. Los árboles pueden ser representados gráficamente y son fáciles de interpretar por no expertos, particularmente si los árboles son pequeños.
4. Los árboles pueden manejar predictores cualitativos sin la necesidad de crear variables dummy o indicadoras.

5. Desafortunadamente, los árboles no poseen el mismo nivel de precisión a la hora de predecir que tienen algunos de los otros métodos de regresión y clasificación que se han visto.
6. Adicionalmente los árboles son poco robustos. En otras palabras, un cambio pequeño en los datos puede causar un cambio grande en el árbol estimado.