Los árboles de clasificación y regresión constituyen una herramienta útil para la **predicción de variables cualitativas y cuantitativas**, respectivamente, de una manera **sencilla** y sin asunciones teóricas sobre los datos.

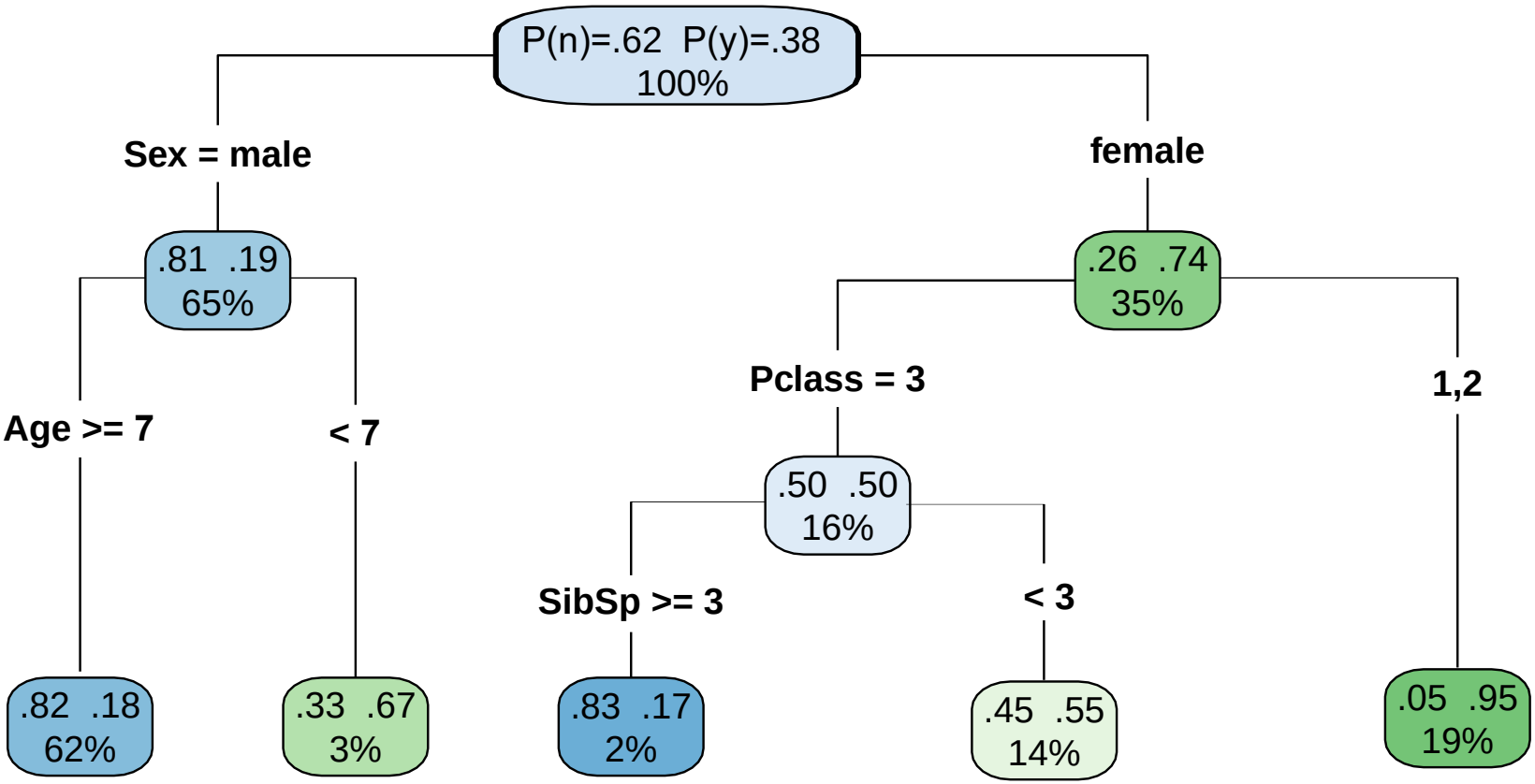
Los árboles representan una **segmentación de los datos** a partir de una serie de reglas simples, que se van aplicando de forma **jerárquica y secuencial**. De esta forma, se obtienen una serie de segmentos (llamados **nodos**) que contienen subconjuntos de la muestra. El **segmento original** contiene la totalidad de los datos y recibe el nombre de nodo **raíz**.

Una vez obtenida la segmentación "óptima", entendiendo así aquella que da lugar a nodos con comportamiento homogéneo respecto a la variable dependiente y heterogéneos entre si, se asigna un **valor de predicción** a aquellos nodos que no tienen sucesores (y que reciben el nombre de **hojas**) de forma que **todas las observaciones** pertenecientes a dicha hoja serán predichas a partir de dicho valor.

Así pues, podemos concluir que los árboles se comportan como un **análisis cluster** (por la creación de grupos homogéneos y diferentes entre sí) y a la vez como un **método predictivo**.

Aunque existen muchos tipos diferentes de árboles (como los árboles condicionales o los *CHAID*), en esta asignatura nos vamos a centrar en los denominados ***CART (Classification And Regression Trees)***, propuestos en Breiman et al. (1984).

Datos Titanic



FASES DE LA CONTRUCCIÓN DE UN ÁRBOL

1. Búsqueda de las **reglas** que definen los nodos:
   - I. A la hora de dividir un nodo es necesario decidir: a) **qué variable** utilizar y, b) **cómo realizar la división** de dicha variable. Este proceso se realiza seleccionando en primer lugar el **mejor punto de corte** de cada variable (sin perdida de generalidad, haremos divisiones binarias) para, a continuación, seleccionar la **mejor variable**. Nótese que el concepto mejor se traduce en que las observaciones de cada división sean homogéneas entre sí.
   - II. El proceso anterior se **repite** para cada nodo susceptible de ser dividido hasta que se cumpla una de las siguientes condiciones:
     - Ninguna de las variables disponibles produzca una **mejora**.
     - Todas las posibles divisiones den lugar a **hijos con pocas observaciones** (*minbucket*).
     - El número de **observaciones en el nodo** a dividir sea demasiado pequeño (*minsplit*). Se suele asumir como el triple de la cantidad anterior.
     - La **profundidad del árbol** (distancia, en número de nodos, entre el nodo raíz y la hoja más alejada) supere el umbral predefinido (*maxdepth*).

FASES DE LA CONSTRUCCIÓN DE UN ÁRBOL

2. Cálculo del **valor de predicción**: para variables dependientes cualitativas, son las **proporciones observadas** de cada una de sus categorías en el nodo (llamadas como en *NB* **probabilidades a posteriori**), y para las variables cuantitativas, las **medias observadas**.
3. Para evitar el posible sobreajuste, a partir de **validación cruzada**, se puede podar el árbol, es decir, buscar un **árbol más pequeño** que el anterior cuyo poder de predicción sea **similar** (o incluso mejor) que el árbol completo.

- El valor que toma el tamaño mínimo de hoja tiene un **gran impacto** en la capacidad predictiva del árbol resultante, pues determina los puntos de corte a evaluar, así como el número de hojas final.
- Este algoritmo **no asegura** obtener el mejor árbol posible, puesto que no se evalúan todas las **combinaciones factibles**, sino que se busca la mejor variable y el mejor punto de corte **en cada paso** (es un algoritmo *greedy*). Este tipo de proceso recibe el nombre de heurística, pues da lugar a **soluciones suficientemente buenas**, aunque no sean las óptimas.

DATOS FALTANTES

Destacan dos aspectos principalmente de la gestión de los datos faltantes en los árboles:

- Durante la **construcción del árbol**, sólo se ignoran aquellas observaciones que tienen dato faltante en la variable dependiente o en todas las predictoras.

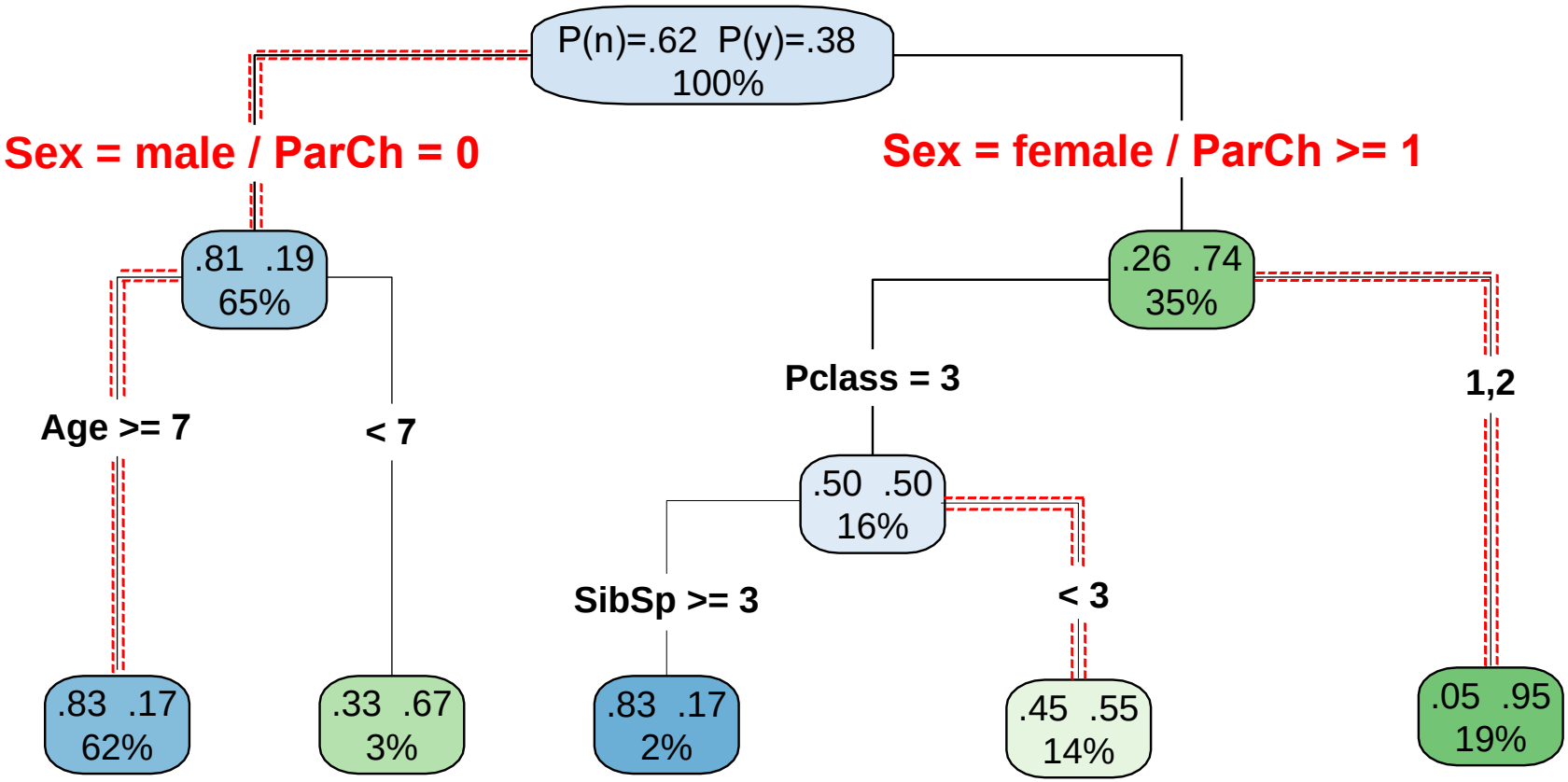
  Esto permite utilizar la información de **todas las observaciones con valor en cada variable predictora** a la hora de fijar el mejor punto de corte y su poder predictivo.

- Desde el punto de vista de la **predicción** de los árboles, se utilizan *reglas de sustitución* (*surrogate splits*) en todos los nodos padre que puedan utilizarse como **sustitutas de las originales** en el caso de que la observación tenga un missing y, por tanto, **no se pueda utilizar** la variable predictora que efectúa la división.

  Así, durante la construcción del árbol se pueden almacenar, para cada división, reglas sustitutas, basadas en otras variables predictoras, que den lugar a divisiones **lo más parecidas posible a las originales**. El **número máximo de estas reglas** se debe fijar de antemano.

  Alternativamente, si ninguna otra variable se parece lo suficiente, se puede aplicar la denominada **regla sustituta *básica***, consistente en ubicar la observación en el hijo **mayoritario** (aquel con una frecuencia mayor de observaciones) si no se dispone del valor de la variable que genera la regla principal.

Datos Titanic



IMPORTANCIA DE VARIABLES

Cuando se construyen modelos predictivos suele resultar de interés **obtener un *ranking* de la importancia de variables** que nos permita concluir cuáles resultan de mayor utilidad y de cuáles se podría prescindir (como con el análisis de tipo II en los modelos paramétricos).

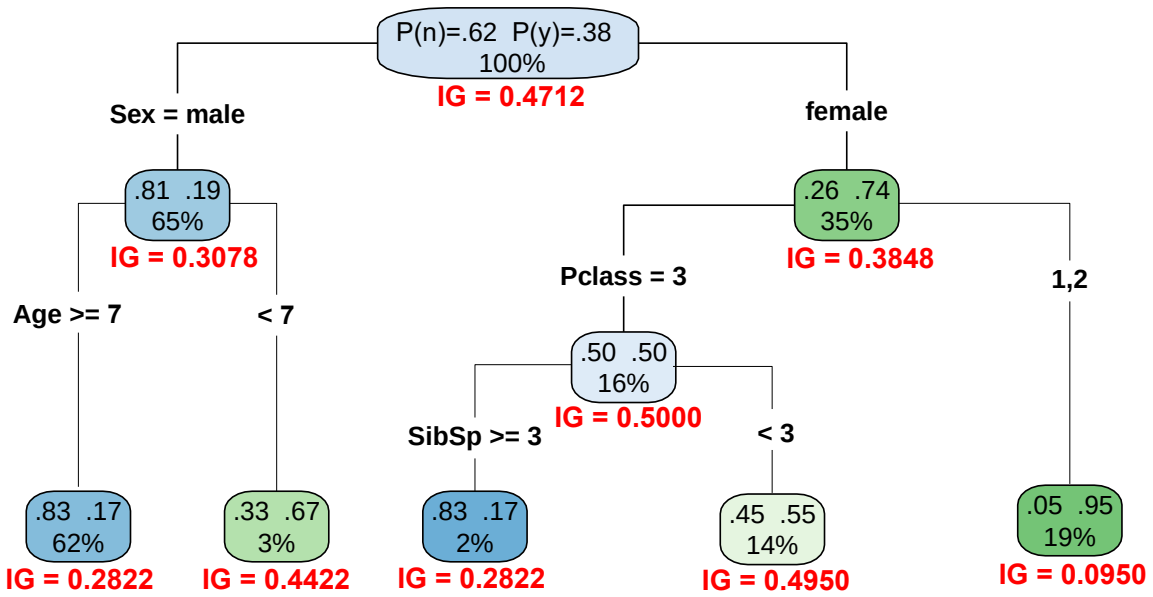
Dado que **cada vez que se divide un nodo padre se calcula la mejora que esto implica** (en términos de la métrica de homogeneidad que corresponda), esa cantidad se puede utilizar como un indicador de la importancia de la variable implicada (pues grandes mejoras implicarán que la variable explicativa utilizada es importante para la predicción).

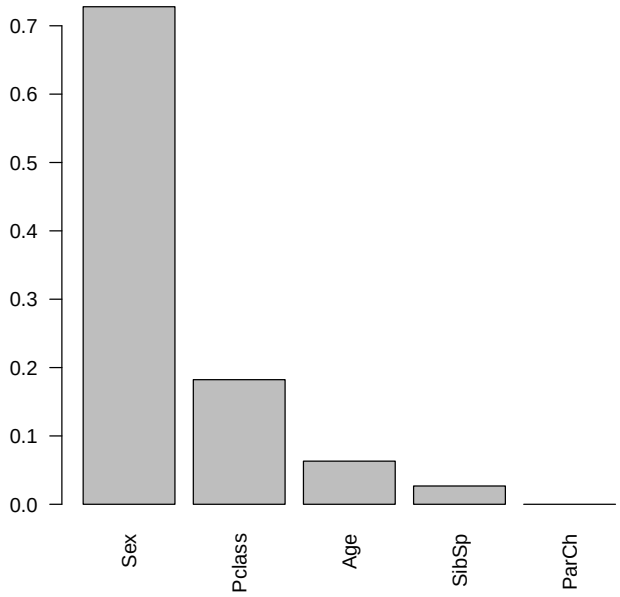
Por ello, los árboles almacenan durante su proceso de creación la **mejora generada en cada división por la variable correspondiente**, multiplican dicha cantidad por el número de observaciones en el nodo que se está dividiendo, y agregan dichas cantidades si las variables se han utilizado en más de una división.

Ese resultado se puede representar en un **gráfico de barras** de manera que aquellas variables que han conseguido una mejora mayor serán las más importantes. Así mismo, aquellas variables que no hayan sido utilizadas, obtendrán una **importancia de cero**.

En *R*, el aporte de las variables utilizadas para la creación de las reglas sustitutas se tiene en cuenta para el cálculo de la importancia de variables por lo que hay que tener cuidado con ello a la hora de interpretar los resultados.

Datos Titanic





Los árboles de regresión son aquellos cuya variable dependiente es **cuantitativa**.

CRITERIO PARA SELECCIÓN DE PUNTO DE CORTE Y VARIABLE

Como se ha comentado, se utilizará la **media** de la variable dependiente de las observaciones de las hojas como **valor de predicción**. Por ello, cada vez que se realiza una predicción se incurre en un **error** que equivale a la diferencia entre el valor real y la predicción: $y_i - \hat{y}_i$.

Para resumir los errores cometidos en cada nodo, utilizamos la expresión:

$$MSE(nodo_j) = \frac{1}{nobs\{nodo_j\}} \sum_{i \in nodo_j} (y_i - \hat{y}_i)^2 = \frac{1}{nobs\{nodo_j\}} \sum_{i \in nodo_j} (y_i - \bar{y}_j)^2$$

Se persigue, por tanto, una partición de los datos que **minimice la varianza** pues, en ese caso, la diferencia (al cuadrado) entre el valor real y el predicho también se minimiza.

Nótese que al realizar una partición se generan tantos $MSE(nodo_j)$ como nodos hijos (habitualmente $\ell = 2$), por lo que será necesario **agregarlos**. La cantidad a minimizar será:

$$MSE = \sum_{j=1}^{\ell} \frac{nobs\{nodo_j\}}{n} MSE(nodo_j) = \frac{1}{n} \sum_{j=1}^{\ell} \sum_{i \in nodo_j} (y_i - \bar{y}_j)^2$$

Para evaluar la bondad del árbol, se obtiene el estadístico R^2 de bondad del modelo como: $1 - MSE_{arbol}/MSE_{raiz}$, que tomará valores cercanos a 1 cuanto más preciso sea.

Los árboles de clasificación son aquellos cuya variable dependiente es **cualitativa**.

CRITERIO PARA SELECCIÓN DE PUNTO DE CORTE Y VARIABLE

Aunque no es la única opción, el criterio más habitual es el **Indice de Gini**, el cual permite establecer cómo de **homogéneas** son las observaciones de cada nodo con respecto a una variable categórica. El Índice de Gini del nodo viene dado por:

$$IG(nodo_j) = 1 - \sum_{k=1}^{K} \left(p_k^j\right)^2,$$

donde p_k^j representa la proporción de observaciones de la categoría k en el nodo j.

Un nodo "puro" tendrá, por tanto, un índice de **Gini igual a 0** y un nodo con igual frecuencia de las K categorías, tendrá un índice de $1 - \frac{1}{K}$.

Al igual que ocurre con el criterio MSE en árboles de regresión, es necesario **agregar los índices** de Gini que componen la partición. Para ello, se pondera el IG de cada nodo por la proporción de observaciones del nodo hijo j sobre el total del padre, obteniendo un IG total a **minimizar**.

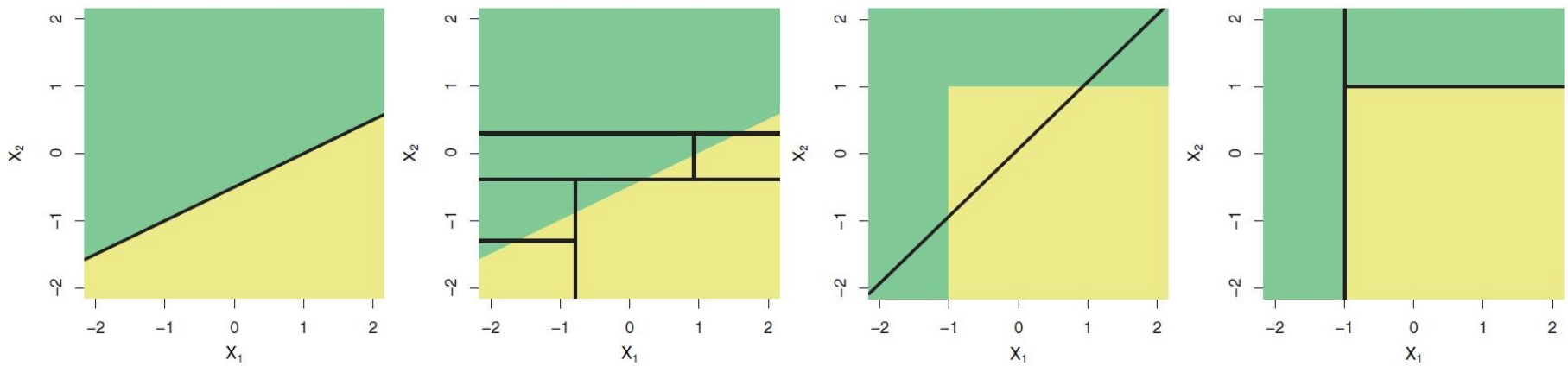
Nótese que los árboles de clasificación generan las probabilidades a posteriori, a partir de las cuales debe generarse la clasificación. La estrategia más habitual es asignar a la categoría mayoritaria, pero en el caso binario, se podría modificar el "punto de corte".

VENTAJAS

- Los resultados son simples y se comprenden fácilmente.
- Permiten encontrar interacciones y reglas difíciles de detectar con otros métodos.
- No hay asunciones teóricas, ni sobre la distribución de las variables ni sobre el tipo de relación entre ellas.
- Permiten tratar los datos ausentes de una manera eficiente.
- Son bastante robustos frente a valores atípicos.
- No mejoran al aplicar transformaciones monótonas a las variables predictoras.
- Por todo lo anterior, requieren poca preparación de los datos (a diferencia de lo que ocurre con otras técnicas que a menudo requieren la normalización de datos, utilización de variables ficticias o imputación de valores ausentes).
- Son capaces de manejar tanto datos numéricos como categóricos.
- Aportan medidas de importancia de las variables.
- La selección de variables forma parte del proceso de creación.

DESVENTAJAS

- Las predicciones son “toscas” (mismo valor para todas las observaciones del nodo), lo que puede implicar una reducción en la capacidad predictiva de los mismos.
- Los modelos resultantes tienen gran variabilidad: añadir una variable nueva o un nuevo conjunto de observaciones puede alterar mucho el árbol resultante.
- No logran capturar correctamente relaciones claramente lineales entre la variable dependiente y una predictora.



PODA DE ÁRBOLES

Los árboles tienen tendencia al **sobreajuste**, es decir, tienden a crear más hojas de las necesarias (sobre todo si el tamaño mínimo de hoja es pequeño), por lo que se obtienen modelos que no son capaces de generalizar correctamente. Para solucionar este problema se debe **podar el árbol** obtenido (denominado maximal) y seleccionar alguno de sus subárboles.

En línea con la selección de variables en modelos de regresión basado en los estadísticos AIC/BIC, para podar los árboles se define el **riesgo de los subárboles** $R_\alpha(T)$ y se busca aquel que minimice esta cantidad:

$$R_\alpha(T) = error(T) + |T|\ \alpha\ error(T_1),$$

donde T se refiere al subárbol en evaluación, T_1 es la raíz del árbol, α es el parámetro de penalización, $|T|$ es el número de hojas de T y $error(T)$ se define como la *SSE* de T si este es de regresión o el número de observaciones mal clasificadas si es de clasificación.

Nótese que el riesgo consta de dos partes: la primera evalúa el **error** cometido por el subárbol y el segundo es una **penalización** debido al tamaño del mismo. De esta forma, sólo compensará seleccionar un árbol más grande si la **reducción** del error es superior al aumento de la penalización (que es proporcional al $100 \cdot \alpha\ \%$ del error inicial).

Definido así, se puede ver que los **valores extremos de** α son 0 (lo que implica seleccionar el árbol maximal) y 1 (lo que implica seleccionar la raíz).

PODA DE ÁRBOLES

Aunque α puede tomar cualquier valor entre 0 y 1, el **número de posibles subárboles es finito** por lo que, en la práctica, basta con considerar aquellos valores "críticos" que implican un aumento/disminución de tamaño del subárbol.

Por ese motivo, la forma habitual de proceder consiste en, durante el proceso de creación del árbol maximal, **registrar los valores α decrecientes que dan lugar a un crecimiento del árbol**. En particular, cada vez que se divide un nodo (lo que implica pasar de un árbol con ℓ hojas, que denotamos como T_ℓ, a uno con $\ell + 1$) se registra el valor:

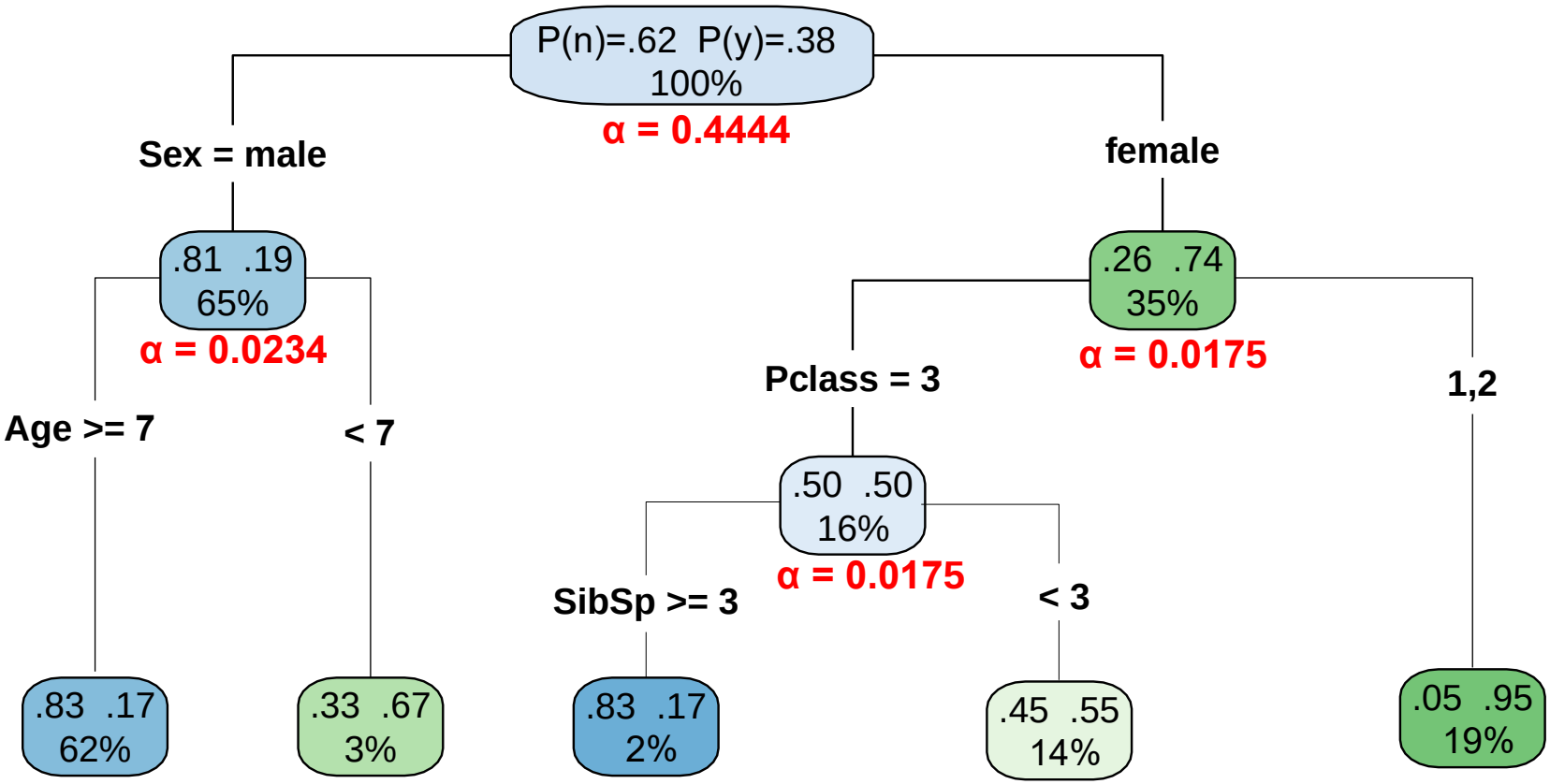
$$\alpha = \frac{error(T_\ell) - error(T_{\ell+1})}{error(T_1)}$$

Estos valores de α generan una secuencia de subárboles "óptima" en el sentido de que cada valor crítico fija qué hojas deben formar parte del árbol (sólo aquellas que han sido obtenidas mediante una división con una reducción del error superior) y, por ende, su tamaño.

Dicha **secuencia óptima de subárboles**, se debe evaluar a partir de **validación cruzada** pues, de lo contrario, siempre se seleccionaría el árbol maximal como mejor.

Para ello, además, en el caso de los árboles de clasificación podemos **recurrir a otros estadísticos**, como el índice *kappa* o el *AUC*, pues son más sencillos de interpretar.

Datos Titanic



Estos árboles se construyen para variables dependientes de tipo **ordinal**.

Por supuesto, para esas variables se pueden construir árboles de clasificación, pero estos ignorarán el orden, por lo que **la calidad resultante suele ser peor**.

Los árboles ordinales comparten las características de los anteriores; solo debe utilizarse una **métrica de homogeneidad apropiada**.

CRITERIO PARA SELECCIÓN DE PUNTO DE CORTE Y VARIABLE

El criterio más habitual es el **Indice de Gini generalizado (IGg)**, el cual permite establecer cómo de **homogéneas** son las observaciones de cada nodo con respecto a una variable categórica, teniendo en cuenta que las **categorías adyacentes** se pueden considerar más homogéneas que aquellas más distantes.

Como su nombre indica, es una generalización del Índice de Gini, por lo que recordamos su fórmula y la reescribimos de una manera alternativa más apropiada a este caso:

$$IG(nodo_j) = 1 - \sum_{k=1}^{K} \left(p_k^j\right)^2 = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i \neq k}}^{K} p_k^j p_i^j$$

CRITERIO PARA SELECCIÓN DE PUNTO DE CORTE Y VARIABLE

El IGg parte del IG anterior e **incluye una función de coste** en los sumandos que penaliza las categorías más alejadas:

$$IGg(nodo_j) = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i \neq k}}^{K} C(i,k)\ p_k^j p_i^j,$$

donde $C(i,k) = |i-k|$ en el caso que nos ocupa, pero puede ser de otra forma en otros contextos. Así, un nodo "puro" tendrá un **IGg igual a 0**.

Nótese que para poder calcular el IGg, es necesario **codificar la variable dependiente ordinal** con números, que respeten el orden natural de la variable.

Al igual que ocurre con los criterios anteriores, es necesario **agregar los IGg** componen la partición. Para ello, se pondera el IGg de cada nodo por la proporción de observaciones del nodo hijo j sobre el total del padre, obteniendo un IGg total a **minimizar**.

A diferencia de lo que ocurre con los árboles de clasificación, los árboles ordinales **no proporcionan probabilidades** a posteriori, sino que facilitan la clasificación directamente.

Debido al carácter ordenado de la variable y al uso del valor absoluto como función de coste, el valor de predicción generado es la **mediana de la variable dependiente** de cada hoja.

- Los árboles ordinales proporcionan también una medida de **importancia de las variables**. Esta se basa en la mejora producida en el IGg en cada división, ponderando dicha cantidad por el número de observaciones ubicadas en el nodo a dividir.
- La **gestión de los ausentes** es exactamente la misma que en los árboles de regresión y clasificación.
- En cuanto a la poda, el error que se utiliza en el cálculo del riesgo de los subárboles R_α se basa en el **coste de clasificación errónea**, que no es más que la proporción de errores de clasificación cometidos, ponderados por la función de coste utilizada en el IGg.
- Una vez generada la secuencia de los valores de α, lo recomendable es evaluar los subárboles mediante el **índice kappa ponderado**.