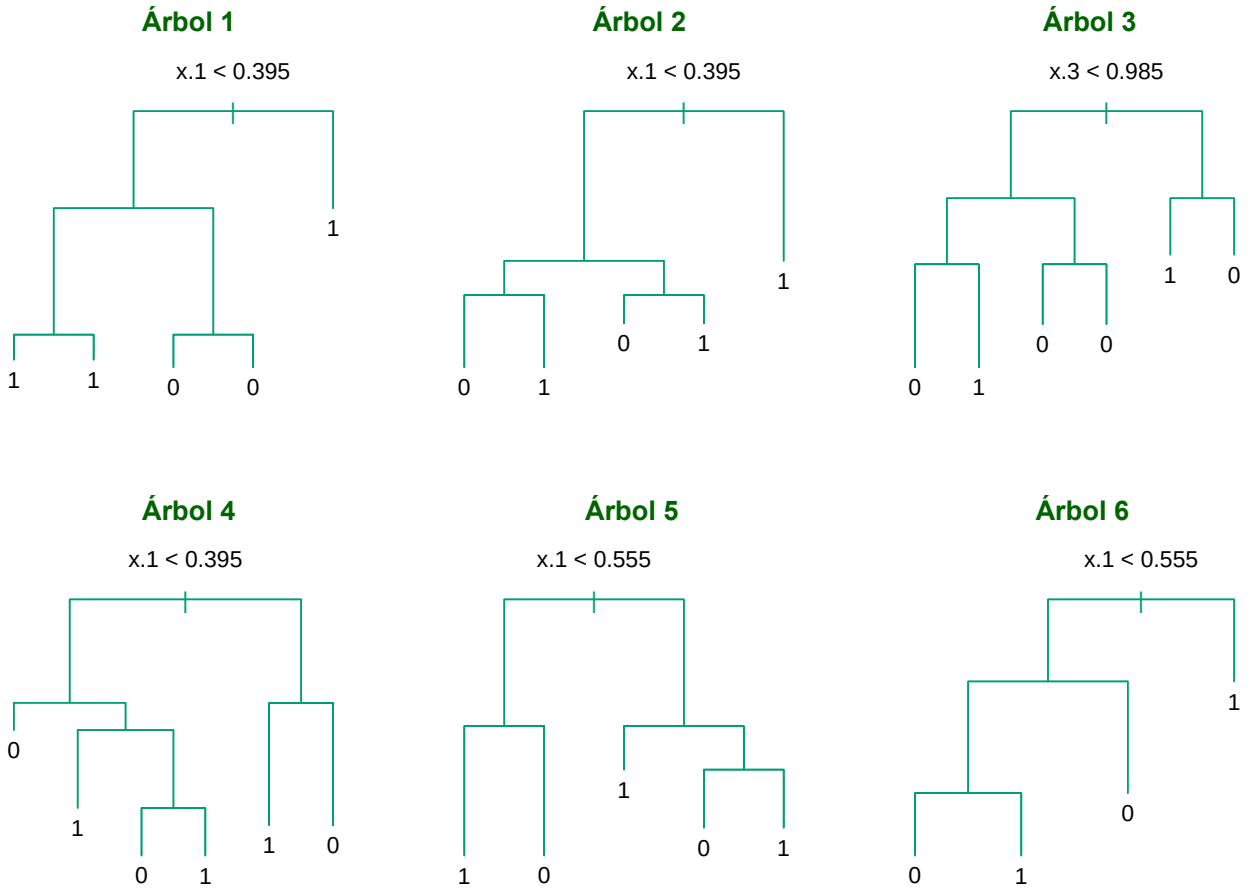
Una de las principales desventajas de los árboles de predicción es su **alta variabilidad**, entendida como el gran cambio que se produce en el mismo cuando se añaden o eliminan variables y/o observaciones (lo que se traduce en una **reducción de la credibilidad** de las predicciones obtenidas a través de los mismos).



FUNDAMENTOS DEL *Bagging*

La técnica de *Bagging* tiene como objetivo **utilizar dicha desventaja** de los árboles para construir una herramienta de predicción más potente.

Para ello, basándose en el hecho de que **todos los árboles** que se pueden generar tienen cierto poder predictivo, se persigue aprovechar todos en lugar de centrarse en uno solo, para lo cual se **agrega la información de los mismos** en un único valor de predicción mediante la media aritmética.

De esta forma, se puede incluir en el proceso predictivo más información y **reducir la variabilidad**, pues recordemos que la varianza de la media de una m.a.s. es inferior a la varianza de cada uno de sus elementos:

$$Var(\bar{\hat{y}}) = Var\left(\frac{1}{B}\sum_{i=1}^{B}\hat{y}_i\right) = \frac{1}{B^2}\sum_{i=1}^{B}Var(\hat{y}_i) = \frac{Var(\hat{y})}{B}$$

FUNDAMENTOS DEL *Bagging*

Para **generar los distintos árboles** que formarán parte del ensamblado sin tener que obtener nuevas muestras (lo cual generalmente resulta infactible), se recurre a las denominadas **muestras *Bootstrap***, de las cuales procede el nombre de esta técnica: *Bootstrap* + *AGGregation*.

Una muestra *Bootstrap* no es más que una **submuestra con reemplazamiento** de los datos.

Esta técnica de remuestreo tan sencilla permite **simular el proceso de muestreo de la población** sin la necesidad de recurrir a nuevas muestras reales y obtener así mejores modelos sin aumentar el coste.

Aunque la idea de *Bagging* se puede aplicar teóricamente a **cualquier modelo de predicción**, es especialmente útil en árboles debido a su gran variabilidad, pues este proceso **permite reducir la variabilidad**, pero no siempre implica una mejora significativa del sesgo.

Por ello, lo recomendable es **construir árboles "grandes"** (gran profundidad y/o pequeño tamaño de hoja), pues son estos los que sufren de gran variabilidad y menor sesgo.

De hecho, si se construyen árboles "pequeños" es probable que **muchos de ellos sean iguales** (pues en las primeras etapas de construcción del árbol suelen aparecer pocas diferencias) y no se produzcan grandes mejoras con respecto a usar un único árbol.

Cuando se trabaja con **variables dependientes cualitativas**, la agregación de las predicciones de los árboles consiste en obtener la **proporción de árboles** que han predicho cada uno de los niveles de la variable dependiente (utilizando el criterio de la probabilidad máxima) y usar esas cantidades como las probabilidades a posteriori.

Importancia de variables

Uno de los mayores inconvenientes del *Bagging* es que **el modelo resultante no es interpretable** (pues resulta imposible visualizar un número tan elevado de árboles).

A pesar de lo anterior, lo que sí se puede generar es un *ranking* de las variables explicativas que permita hacerse una idea de los **factores qué más influyen** a la hora de predecir (aunque no se puede saber el "sentido" de dicha influencia).

Para ello, se obtiene la **importancia media de las variables a partir de los valores generados por los árboles** (que recordemos se obtienen como la mejora debida a cada variable en la SSE o el índice de Gini según trabajemos con una variable cuantitativa o cualitativa, respectivamente).

Parámetros a determinar

- **Tamaño de los árboles**: Como ya se ha comentado, para obtener mejores resultados, se deben construir árboles "grandes", por lo que **no se podarán** los mismos.

  No obstante, sí que habrá que fijar algún parámetro que **determine su tamaño**.

  Es habitual recurrir al **tamaño de hoja** (denominado previamente *minbucket*), es decir, al número mínimo de observaciones que debe tener una hoja, pues da lugar a árboles "grandes" cuando se fija en un valor pequeño.

  A pesar de lo anterior, el concepto "pequeño" varía según los datos que se vayan a modelizar, por lo que habrá que **buscar el valor óptimo** en cada caso.

- **Número de árboles**: El otro parámetro a fijar es el número de árboles a generar (usualmente denotado como B, pues coincide con el número de muestras *bootstrap*).

  Algunos autores sugieren fijar este parámetro en un valor bajo (50), mientras que otros argumentan que es recomendable usar un número elevado (500 ó 1000) pues una de las ventajas del *Bagging* es que **no sufre de excesivo sobreajuste cuando este parámetro es grande** (lo que ocurre es que el modelo no mejora).

  Debido al coste computacional que implica almacenar más árboles de los necesarios, se recomienda recurrir a un valor de B grande y **estudiar posteriormente si se puede reducir** sin que esto suponga una gran pérdida de poder predictivo.

Observaciones *OOB*

A la hora de construir modelos *Bagging* se deben **fijar los valores óptimos** de los parámetros.

Como ya se ha comentado previamente, es recomendable recurrir a alguna **técnica de remuestreo** (como la validación cruzada) para llevar a cabo este proceso, pero ésta implicaría un gran coste computacional.

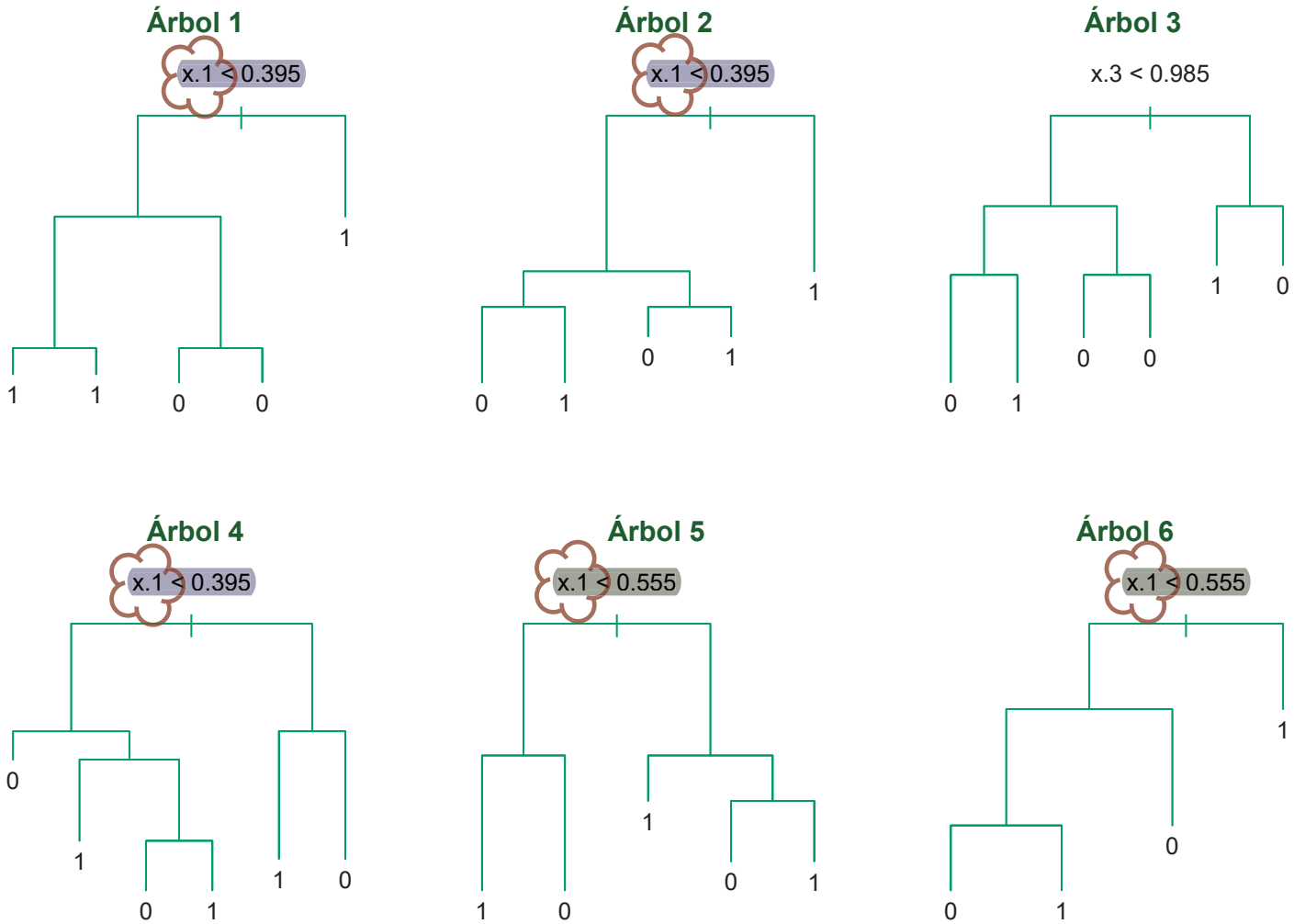
En su lugar, se puede recurrir a las denominadas observaciones ***OOB* (*Out Of Bag*)**, que son aquellas que no se utilizan en cada árbol, ya que no forman parte de la muestra *bootstrap*.

Esta fracción de las observaciones, que suponen aproximadamente un tercio en cada repetición, puede ser utilizada para conocer de una manera realista la capacidad predictiva de cada árbol permitiendo **no tener que recurrir a validación cruzada** a la hora de parametrizar el modelo, lo que reduce significativamente el coste computacional.

Para ello, durante la construcción del modelo *Bagging* **se registra qué observaciones han sido utilizadas** en cada árbol y, al finalizar, se obtiene la predicción de todas las observaciones **agregando únicamente las predicciones obtenidas en árboles para cuya construcción no se utilizaron** (aproximadamente un tercio del total de árboles).

Si el número de árboles es suficientemente grande (lo que por definición siempre ocurre), se puede demostrar que la estimación del error obtenida mediante *OOB* es **equivalente a la validación cruzada** *Leave-One-Out*.

Aunque los modelos *bagging* **mejoran la capacidad predictiva de los árboles** al sacar provecho de la variabilidad de los mismos, la aleatoriedad introducida en el modelo mediante las muestras *bootstrap* no suele ser suficiente para **generar árboles muy diferentes**:



- La causa del comportamiento anterior es que generalmente existen unas **pocas variables con mayor capacidad predictiva** que forman parte siempre de las primeras fases de división del árbol, lo que implica que muchas de **las variables con capacidades predictivas "medias" queden eclipsadas** por las primeras y acaben siendo infrautilizadas (lo que implica una pérdida de información).
- La consecuencia es que las predicciones obtenidas mediante los distintos árboles no son lo suficientemente diferentes, lo que se traduce en que las predicciones están correladas y, por tanto, **la variabilidad del modelo resultante (su agregación) no se reduce tanto como se desearía** (el segundo término no desaparece):

$$Var(\bar{\hat{y}}) = Var\left(\frac{1}{B}\sum_{i=1}^{B}\hat{y}_i\right) = \frac{Var(\hat{y})}{B} + \frac{1}{B^2}\sum_{i\neq j} Cov(\hat{y}_i, \hat{y}_j)$$

Para solucionar este problema Breiman propone en el año 2001 los bosques aleatorios (más conocidos por su nombre en inglés - *Random Forest*) con el objetivo de "descorrelar" los árboles que forman parte de los modelos *bagging*.

FUNDAMENTOS DE LOS BOSQUES ALEATORIOS

Como acabamos de ver, el problema de los modelos *bagging* es que los árboles que componen el ensamble son **"demasiado similares"**.

Los modelos de bosque aleatorio (RF) surgen como una **generalización de los modelos *bagging*** a los cuales se les añade una **nueva fuente de aleatoriedad** de cara a lograr que los árboles sean más diferentes entre sí.

Dado que la causa del parecido es la **existencia de variables "dominantes"**, los modelos RF tienen como objetivo impedir que dichas variables formen parte de todos los árboles.

Para ello, durante la fase de construcción de los árboles, en lugar de permitir que los árboles seleccionen la mejor variable y el mejor punto de corte de entre todas las variables predictoras disponibles, a la hora de dividir cada nodo **se limita dicha elección a un subconjunto aleatorio de dichas variables**, que varía en cada división (no en cada árbol).

De esta forma, las variables dominantes no siempre están disponibles a la hora de llevar a cabo las divisiones, lo que implica una **mayor variabilidad de árboles** (en el bosque) y, por ende, una menor correlación y una **mejora en la capacidad predictiva** del modelo final.

PARÁMETROS A DETERMINAR

Dado que los modelos RF son una generalización de los modelos *bagging*, se repiten los parámetros a determinar anteriores y se añade uno nuevo:

- **Tamaño de los árboles**: Como en el caso de *bagging*, el tamaño de los árboles que componen el bosque es determinante para la capacidad predictiva final.

  De nuevo, se recurre al **tamaño de hoja**, procurando establecer un valor suficientemente pequeño para aprovechar todo el potencial predictivo, pero no demasiado como para generar sobreajuste e implicar un tiempo de cálculo excesivo.

- **Número de árboles**: El segundo parámetro a fijar es el número de árboles a generar.

  Como ya se vió en *bagging*, este tipo de modelos **no suele sufrir sobreajuste cuando este parámetro es grande** por lo que se puede establecer un valor relativamente grande por defecto y estudiar si se puede disminuir posteriormente.

  Es importante destacar que los modelos RF **tienden a requerir más árboles que los modelos *bagging*** debido a las restricciones sobre el uso de las variables predictoras.

PARÁMETROS A DETERMINAR

- **Número de variables explicativas disponibles**: Este parámetro suele recibir el nombre de *mtry* y se corresponde con el tamaño de los subconjuntos aleatorios de variables explicativas entre las cuales deben elegir los árboles a la hora de dividir sus hojas.

  Dependiendo del número de variables explicativas existente, lo relacionadas que estén éstas entre sí y de la relación que exista entre éstas y la dependiente, **será preferible un valor de *mtry* mayor o menor**, por lo que se recomienda probar varios valores.

  En el artículo original, se sugiere utilizar **un tercio** de las variables disponibles para los RF de regresión y de la **raíz cuadrada**, para los RF de clasificación.

  Cabe destacar que el hecho de limitar el conjunto de elección de los árboles implica una **reducción importante en el tiempo computacional** (con respecto a los modelos *bagging*) requerido para su obtención debido a que no se deben realizar tantos cálculos cuando se tiene que dividir una hoja.

  Por último, resaltamos que en el **caso particular** de que el *mtry* corresponda con el número total de variables explicativas disponibles, el modelo resultante sería un *bagging* (pues no se estaría imponiendo ninguna restricción en cuanto a la elección de variables).

Dado que los modelos RF son una generalización de los *bagging*, comparten las siguientes características (ya vistas previamente pero que viene bien recordar):

- El uso de paralelización permite reducir el tiempo computacional requerido para su construcción, pues cada *core* puede generar un árbol independientemente del resto.
- El uso de múltiples árboles impide su representación gráfica y su representación, pero sigue siendo posible obtener medidas de la importancia de las variables (se hace utilizando el método explicado en la página 34).
- Las observaciones *OOB* (explicadas en la página 36) permiten determinar de una manera realista la calidad del modelo resultante, lo que implica no tener que recurrir a validación cruzada a la hora de parametrizar el modelo, lo cuál sería computacionalmente muy intenso.
- Las predicciones de variables cuantitativas se obtienen como la media de las predicciones individuales generadas por cada árbol.
- Las probabilidades predichas de las variables cualitativas objetivo se obtienen como la proporción de votos obtenidos para cada categoría por el conjunto de árboles.
- Además, tanto en los modelos RF como en los *bagging*, es posible recurrir a los *Partial Dependency Plots* (PDP) estudiamos para KNN.