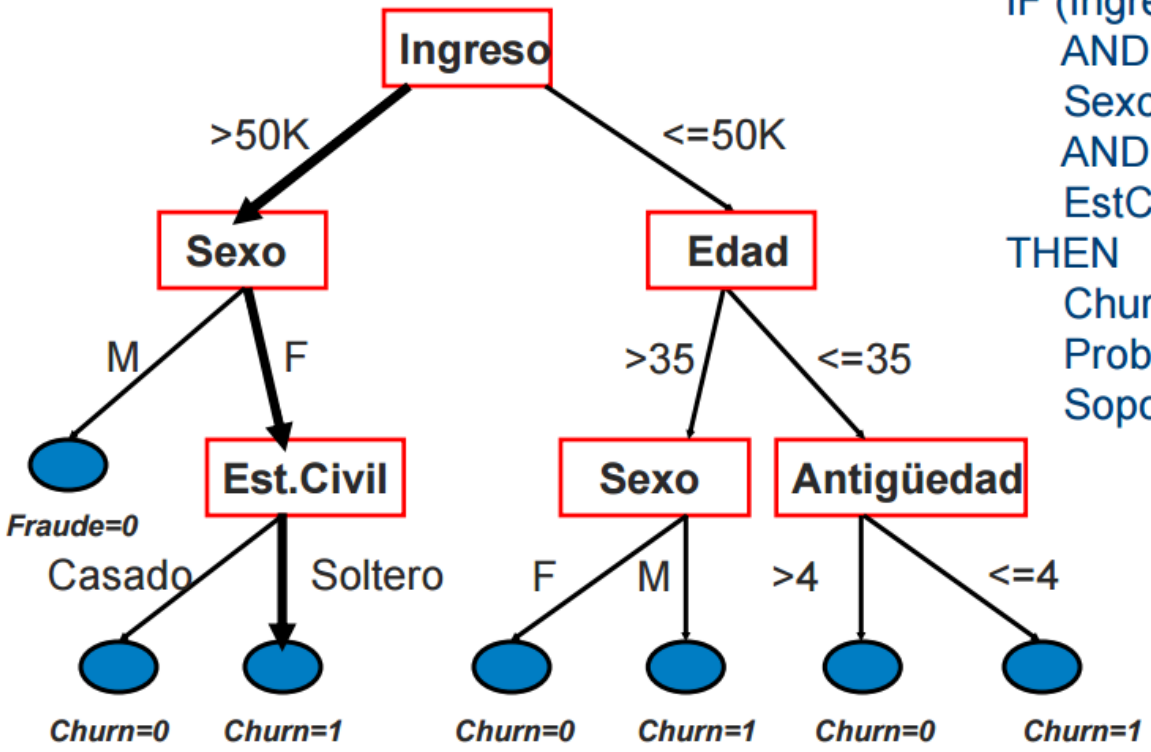
Árboles de decisión



RULE:
IF (Ingreso >50K
AND
Sexo = F
AND
EstCivil = Soltero)
THEN
Churn = 1
Prob = .77
Soporte = 250

Oscar Centeno Mora

Introducción

- Los algoritmos basados en árboles de decisión (clasificación o predicción) se consideran como adecuados, además de aptos para introducirse en el mundo de la minería de datos y el aprendizaje supervisado.

- Los métodos basados en árboles pueden ser utilizados desde los modelos predictivos como para la clasificación, brindando una aceptable, estabilidad y facilidad de interpretación.

- A diferencia de los modelos lineales, las relaciones no lineales se relacionan bastante bien. Son adaptables para resolver cualquier tipo de problema a mano, sin tener que pasar por la verificación de supuestos.

Introducción

- Los métodos como los árboles de decisión se considera se considera como la herramienta base de clasificación y predicción en los problemas de ciencia de datos. Por lo tanto, para cada analista, es importante aprender estos algoritmos y utilizarlos para el modelado.

- Aunque solo hablaremos de árboles dee decisión, se recomienda analizar conjuntamente la técnica de regresión (continua y dicotómica), de Random Forrest, etc, las cuales se complementan muy bien con la presente técnica.

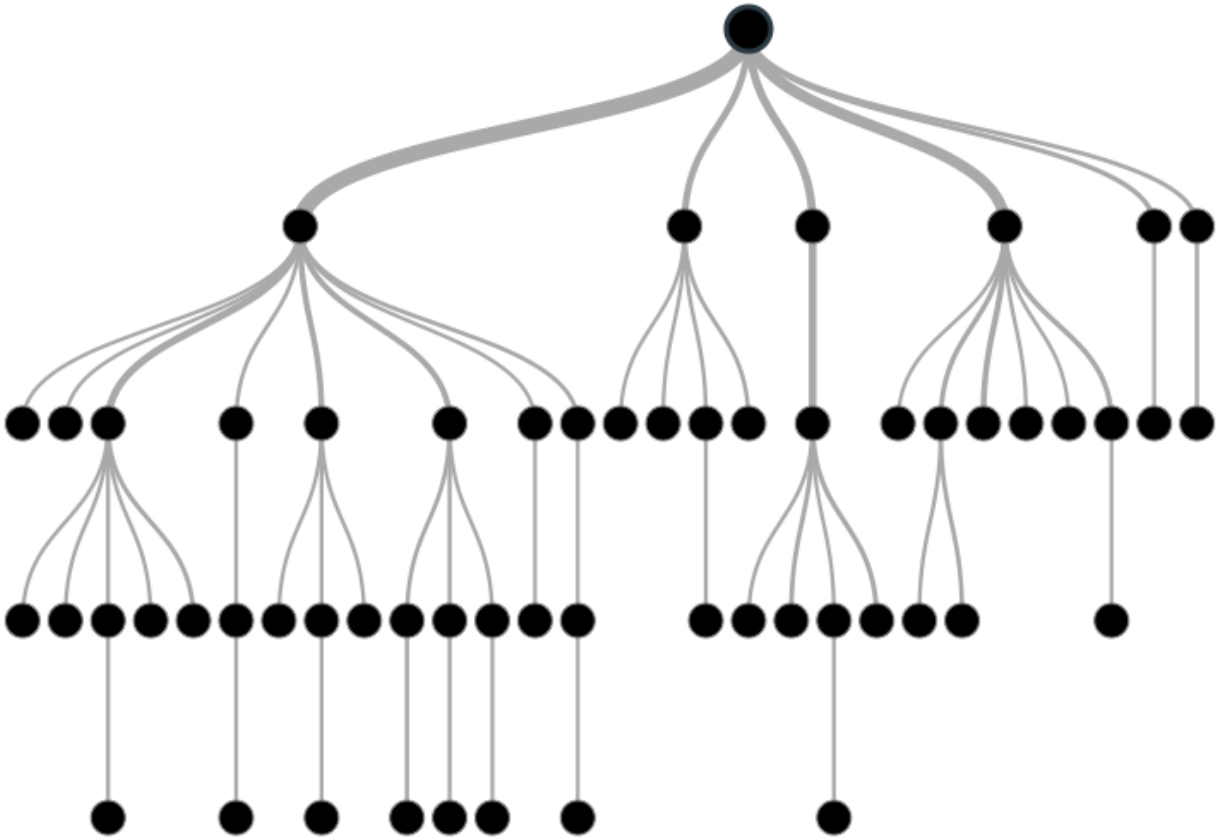
¿Qué es un árbol de decisión? ¿Cómo funciona ?

- El árbol de decisión es un tipo de algoritmo de aprendizaje supervisado (que tiene una variable objetivo predefinida) que se utiliza principalmente en los problemas de clasificación (variable categorico) o predicción (variable continua).

- Funciona tanto para las variables categóricas como para las variables continuas de entrada y salida (Y dependiente y Xi independientes).

- En esta técnica, dividimos la población o la muestra en dos o más conjuntos homogéneos (o subpoblaciones) basados en el más significativo divisor / diferenciador en las variables de entrada.

¿Qué es un árbol de decisión? ¿Cómo funciona ?



¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Digamos que tenemos una muestra de 30 estudiantes con tres variables Género (Niño / Niña), Clase (IX / X) y Altura (5 a 6 pies). 15 de estos 30 jugar cricket en tiempo libre.

- Ahora, quiero crear un modelo para predecir quién jugará cricket durante el período de ocio.

- En este problema, necesitamos segregar a los estudiantes que juegan cricket en su tiempo de ocio basado en la variable de entrada altamente significativa entre los tres.

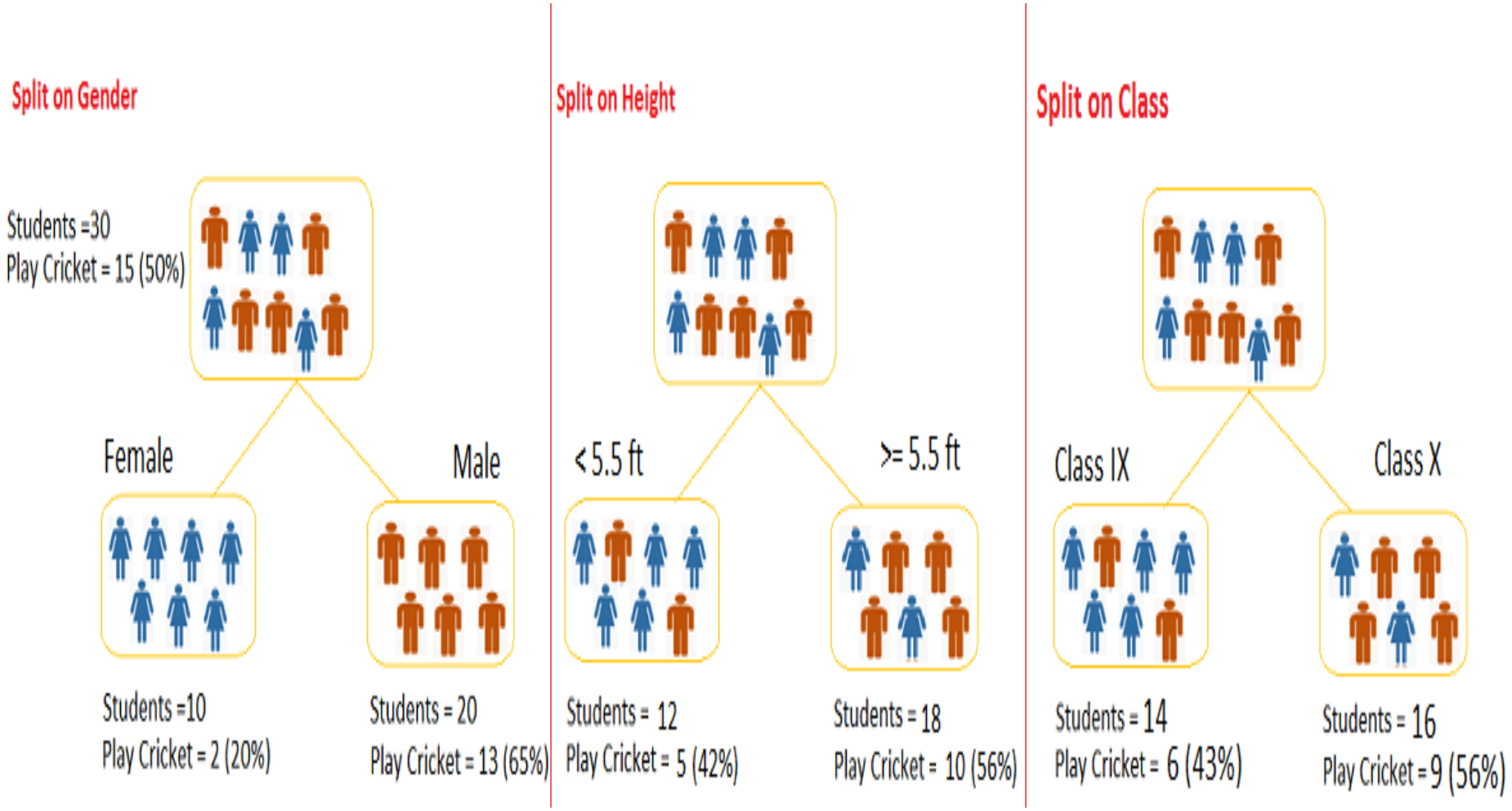
¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Aquí es donde el árbol de decisiones ayuda.

- Segregará a los estudiantes en base a todos los valores de tres variables e identificará la variable, lo que crea los mejores conjuntos homogéneos de estudiantes (que son heterogéneos entre sí).

- En la instantánea a continuación, puede ver que la variable Género es capaz de identificar los mejores conjuntos homogéneos en comparación con las otras dos variables.

¿Qué es un árbol de decisión? ¿Cómo funciona ?



¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Como se mencionó anteriormente, el árbol de decisiones identifica la variable más significativa y su valor que da los mejores conjuntos homogéneos de población.

- Ahora la pregunta que surge es, ¿cómo identifica la variable y la división? Para ello, el árbol de decisiones utiliza varios algoritmos, que discutiremos en la siguiente sección.

¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Los tipos de árbol de decisión se basan en el tipo de variable objetivo que tenemos. Puede ser de dos tipos:

1. Árbol de decisión de variables categóricas: Árbol de decisiones que tiene una variable de destino categórica y luego se llama como árbol de decisión de variable categórica (árbol de clasificación) . Ejemplo: En el escenario anterior del problema del estudiante, donde la variable objetivo era "El estudiante jugará al grillo o no", es decir, SI o NO.

2. Árbol de decisión variable continuo: El árbol de decisión tiene una variable de destino continua y luego se llama Árbol de decisión variable continuo (árbol de regresión). Ejemplo: estimación del salario.

¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Ejemplo:

- - Digamos que tenemos un problema para predecir si un cliente pagará su prima de renovación con una compañía de seguros (sí / no). Aquí sabemos que los ingresos del cliente es una variable significativa, pero la compañía de seguros no tiene detalles de ingresos para todos los clientes. Ahora, como sabemos que esto es una variable importante, entonces podemos construir un árbol de decisiones para predecir el ingreso del cliente basado en la ocupación, el producto y varias otras variables. En este caso, estamos prediciendo valores para la variable continua

¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Veamos la terminología básica utilizada con los árboles de decisión:

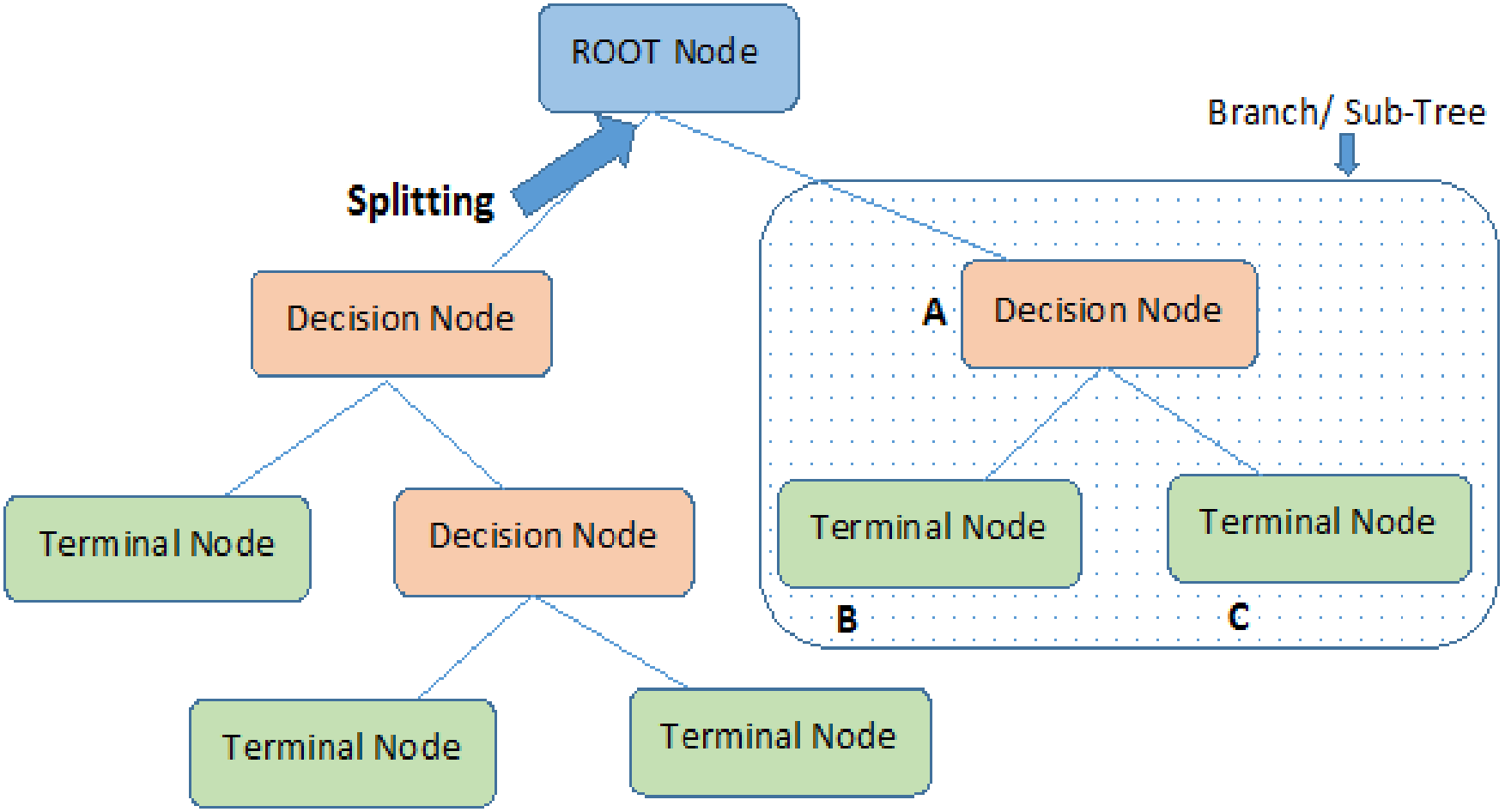
1. Nodo raíz: representa toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.

2. División: es un proceso de dividir un nodo en dos o más subnodos.

3. Nodo de decisión: cuando un sub-nodo se divide en subnodos adicionales, se llama nodo de decisión.

4. Nodo Leaf / Terminal: los nodos que no se dividen se llaman Leaf o nodo termial.

¿Qué es un árbol de decisión? ¿Cómo funciona ?



Note:- A is parent node of B and C.

¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Poda: Cuando eliminamos sub-nodos de un nodo de decisión, este proceso se llama poda. Se puede decir un proceso opuesto de división.
- Subdivisión / Sub-Árbol: Una sub sección de todo el árbol se llama rama o sub-árbol.
- Nodo padre e hijo: Un nodo, que se divide en sub-nodos, se llama nodo padre de sub-nodos donde como sub-nodos son el nodo secundario del nodo padre.

- Estos son los términos comúnmente utilizados para los árboles de decisión. Como sabemos que cada algoritmo tiene ventajas y desventajas, a continuación son los factores importantes que uno debe saber.

¿Qué es un árbol de decisión? ¿Cómo funciona ?

- Las ventajas son:

1. Fácil de entender: La salida del árbol de decisión es muy fácil de entender incluso para las personas de antecedentes no analíticos. No requiere ningún conocimiento estadístico para leerlos e interpretarlos. Su representación gráfica es muy intuitiva y los usuarios pueden relacionar fácilmente su hipótesis.

2. Útil en la exploración de datos: Árbol de decisión es una de las formas más rápidas de identificar las variables más significativas y la relación entre dos o más variables. Con la ayuda de árboles de decisión, podemos crear nuevas variables / características que tengan mejor poder para predecir la variable objetivo.

¿Qué es un árbol de decisión? ¿Cómo funciona ?

3. Menor limpieza de datos: requiere menos limpieza de datos en comparación con otras técnicas de modelado. No está influenciado por los valores atípicos y los valores perdidos en un grado razonable.

4. El tipo de datos no es una restricción: puede manejar variables numéricas y categóricas.

5. Método no paramétrico: Se considera que el árbol de decisión es un método no paramétrico. Esto significa que los árboles de decisión no tienen suposiciones acerca de la distribución del espacio y la estructura del clasificador.

¿Qué es un árbol de decisión? ¿Cómo funciona ?

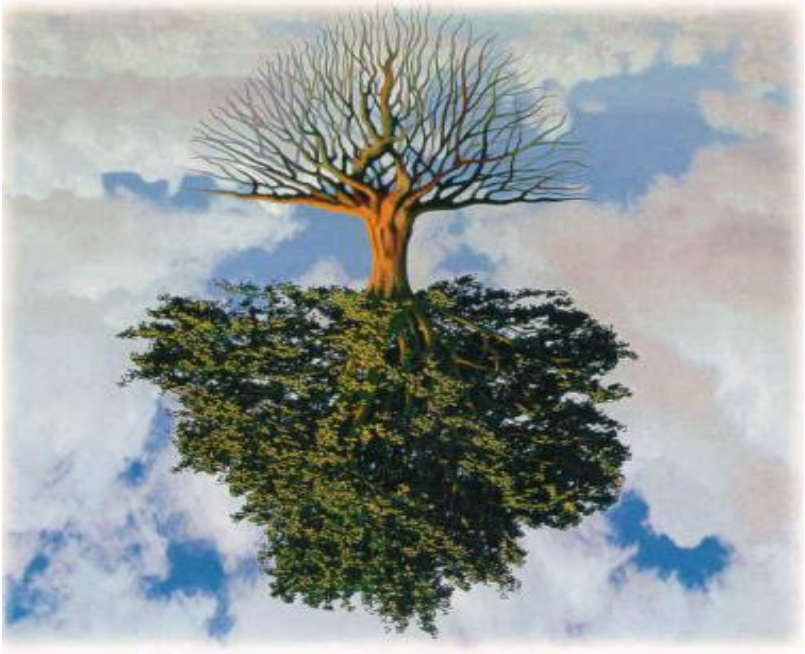
- Las desventajas:

1. Ajuste excesivo: La adaptación excesiva es una de las dificultades más prácticas para los modelos de árboles de decisión. Este problema se resuelve estableciendo restricciones sobre los parámetros del modelo y la poda (discutido en detalle a continuación).

2. No apto para variables continuas: Al trabajar con variables numéricas continuas, el árbol de decisión pierde información cuando categoriza variables en diferentes categorías.

Árboles de regresión vs árboles de clasificación

- Todos sabemos que los nodos terminales (o hojas) se encuentra en la parte inferior del árbol de decisión. Esto significa que los árboles de decisión suelen ser dibujados al revés de manera que las hojas son el fondo y las raíces son las cimas (se muestra a continuación).



Árboles de regresión vs árboles de clasificación

- Ambos árboles trabajan casi similares entre sí, veamos las principales diferencias y similitud entre la clasificación y los árboles de regresión:

- 1. Los árboles de regresión se usan cuando la variable dependiente es continua. Los árboles de clasificación se utilizan cuando la variable dependiente es categórica.

- 2. En el caso del árbol de regresión, el valor obtenido por los nodos terminales en los datos de entrenamiento es la respuesta media de la observación que cae en esa región. Por lo tanto, si una observación no observada de datos cae en esa región, haremos su predicción con valor medio.

Árboles de regresión vs árboles de clasificación

- 3. En el caso del árbol de clasificación, el valor (clase) obtenido por el nodo terminal en los datos de entrenamiento es el modo de las observaciones que caen en esa región. Por lo tanto, si una observación de datos no observados cae en esa región, haremos su predicción con el valor de modo.

- 4. Ambos árboles dividen el espacio predictor (variables independientes) en regiones distintas y no superpuestas. En aras de la simplicidad, se puede pensar en estas regiones como cajas o cajas de alta dimensión.

- 5. Ambos árboles siguen un enfoque codicioso de arriba hacia abajo conocido como división binaria recursiva. Lo llamamos "top-down" porque comienza desde la parte superior del árbol cuando todas las observaciones están disponibles en una sola región y divide sucesivamente el espacio predictor en dos nuevas ramas en el árbol. Se conoce como "codicioso" porque, el algoritmo cuida (busca la mejor variable disponible) sobre sólo la división actual, y no sobre las divisiones futuras que conducirán a un árbol mejor.

Árboles de regresión vs árboles de clasificación

- 6. Este proceso de división se continúa hasta que se alcanza un criterio de parada definido por el usuario. Por ejemplo: podemos decirle al algoritmo que se detenga una vez que el número de observaciones por nodo sea inferior a 50.

- 7. En ambos casos, el proceso de división da lugar a árboles completamente crecidos hasta que se alcanza el criterio de detención. Sin embargo, es probable que el árbol completamente desarrollado sobrepase los datos, lo que conduce a una escasa precisión en datos no vistos. Esto trae "poda". La poda es una de las técnicas que se utilizan para hacer sobrepeso. Aprenderemos más al respecto en la siguiente sección.

¿Cómo un árbol decide dónde dividir?

- La decisión de hacer divisiones estratégicas afecta en gran medida la precisión de un árbol. Los criterios de decisión son diferentes para árboles de clasificación y regresión.

- Los árboles de decisión utilizan múltiples algoritmos para decidir dividir un nodo en dos o más subnodos. La creación de subnodos aumenta la homogeneidad de los subnodos resultantes. En otras palabras, podemos decir que la pureza del nodo aumenta con respecto a la variable objetivo. El árbol de decisiones divide los nodos en todas las variables disponibles y, a continuación, selecciona la división que da lugar a los subnodos más homogéneos.

- La selección del algoritmo también se basa en el tipo de variables objetivo. Veamos los cuatro algoritmos más utilizados en el árbol de decisiones: Gini, Chi-Cuadrado, Ganacia de Información, Reducción de variancia.

¿Cómo un árbol decide dónde dividir? Indice de Gini

- El índice de Gini dice, si seleccionamos dos artículos de una población al azar entonces deben ser de la misma clase y la probabilidad para esto es 1 si la población es pura.

1.Funciona con la variable objetivo categórica "Éxito" o "Fracaso".

2. Realiza solamente divisiones binarias

3. Mayor el valor de Gini mayor la homogeneidad.

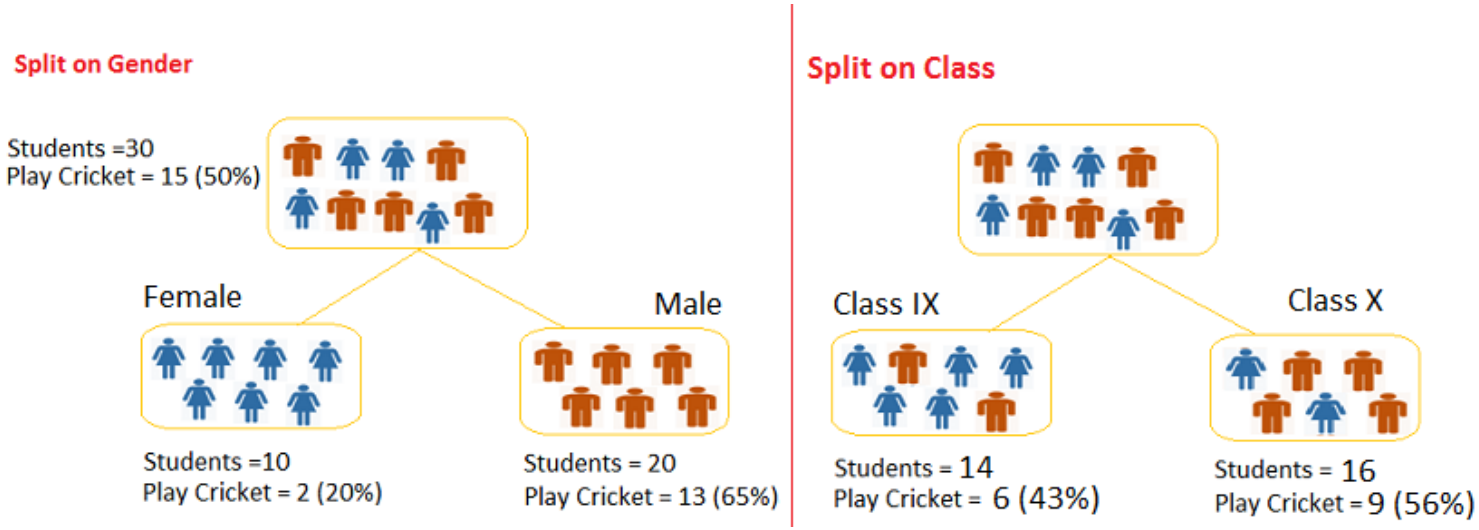
4. CART (Árbol de clasificación y de regresión) utiliza el método Gini para crear divisiones binarias.

¿Cómo un árbol decide dónde dividir? Indice de Gini

- Pasos para calcular Gini para una división

1. Calcular Gini para subnodos, usando la fórmula suma de cuadrado de probabilidad de éxito y fracaso (p ^ 2 + q ^ 2).

2. Calcular el Gini para la división mediante el puntaje de Gini ponderado de cada nodo de esa división



¿Cómo un árbol decide dónde dividir? Chi-cuadrado

- Es un algoritmo para descubrir la significación estadística entre las diferencias entre los subnodos y el nodo padre. La medimos por suma de cuadrados de diferencias estandarizadas entre las frecuencias observadas y esperadas de la variable objetivo.

- 1. Funciona con la variable objetivo categórica "Éxito" o "Fracaso".
- 2. Puede realizar dos o más divisiones.
- 3. Mayor el valor de Chi-Cuadrado mayor la significación estadística de las diferencias entre el sub-nodo y el nodo padre.
- 4. Chi-cuadrado de cada nodo se calcula mediante la fórmula,
- 5. Chi-cuadrado = ((Actual - Esperado) ^ 2 / Esperado) ^ 1/2
- 6. Genera un árbol llamado CHAID (Chi-square Automatic Interaction Detector)

¿Cómo un árbol decide dónde dividir? Chi-cuadrado

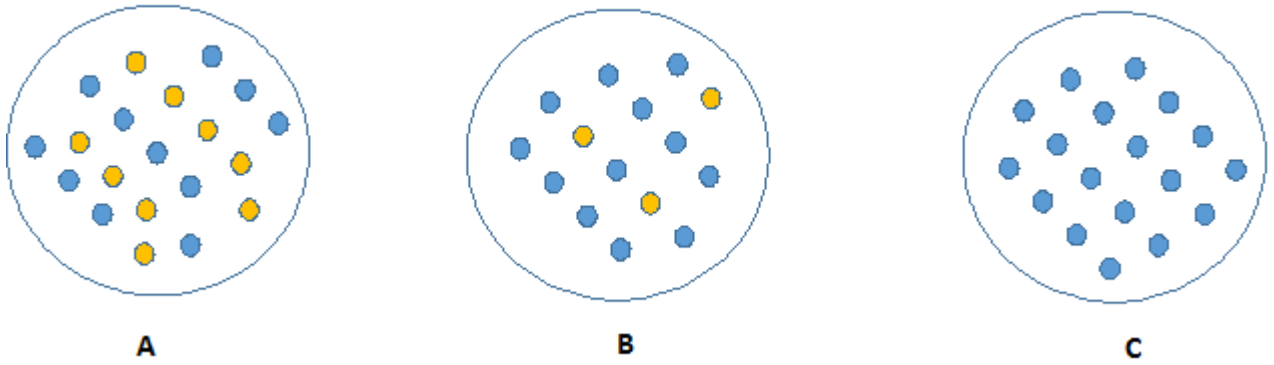
- Pasos para calcular el Chi cuadrado para una división:
- 1. Calcular Chi-cuadrado para el nodo individual calculando la desviación para el éxito y el fracaso ambos
- 2. Calcular el Chi-cuadrado de Split usando la Suma de todo Chi-cuadrado de éxito y Fracaso de cada nodo de la división

Node	Play Cricket	Not Play Cricket	Total	Expected Play Cricket	Expected Not Play Cricket	Deviation Play Cricket	Deviation Not Play Cricket	Chi-Square	
								Play Cricket	Not Play Cricket
Female	2	8	10	5	5	-3	3	1.34	1.34
Male	13	7	20	10	10	3	-3	0.95	0.95
							Total Chi-Square	4.58	

Node	Play Cricket	Not Play Cricket	Total	Expected Play Cricket	Expected Not Play Cricket	Deviation Play Cricket	Deviation Not Play Cricket	Chi-Square	
								Play Cricket	Not Play Cricket
IX	6	8	14	7	7	-1	1	0.38	0.38
X	9	7	16	8	8	1	-1	0.35	0.35
							Total Chi-Square	1.46	

¿Cómo un árbol decide dónde dividir? Ganancia de información

- Mira la imagen de abajo y piensa qué nodo se puede describir fácilmente. Estoy seguro, su respuesta es C porque requiere menos información ya que todos los valores son similares. Por otro lado, B requiere más información para describirlo y A requiere la máxima información. En otras palabras, podemos decir que C es un nodo Puro, B es menos impuro y A es más impuro.



¿Cómo un árbol decide dónde dividir? Ganancia de información

- Ahora, podemos construir una conclusión de que menos nodo impuro requiere menos información para describirlo. Y, nodo más impuro requiere más información. La teoría de la información es una medida para definir este grado de desorganización en un sistema conocido como Entropía. Si la muestra es completamente homogénea, entonces la entropía es cero y si la muestra es igualmente dividida (50% - 50%), tiene entropía de uno.

- La entropía se puede calcular usando la fórmula:

$$\text{Entropy} = -p \log_2 p - q \log_2 q$$

¿Cómo un árbol decide dónde dividir? Ganancia de información

- Aquí p y q es la probabilidad de éxito y fracaso, respectivamente, en ese nodo. La entropía también se utiliza con la variable objetivo categórica. Elige la división que tiene la entropía más baja en comparación con el nodo padre y otras divisiones. Cuanto menor sea la entropía, mejor será.

- Pasos para calcular la entropía para una división:

- 1.Cálculo de la entropía del nodo padre
- 2. Calcule la entropía de cada nodo individual de división y calcule el promedio ponderado de todos los subnodos disponibles en la división.

¿Cómo un árbol decide dónde dividir? Reducción en variancia

- Hasta ahora, hemos discutido los algoritmos para la variable objetivo categórica. La reducción de la varianza es un algoritmo utilizado para las variables objetivo continuas (problemas de regresión). Este algoritmo utiliza la fórmula estándar de varianza para elegir la mejor división. La división con menor varianza se selecciona como criterio para dividir la población:

$$\text{Variance} = \frac{\Sigma(X - \overline{X})^2}{n}$$

- Por encima de X-bar es la media de los valores, X es real y **n** es el número de valores.

¿Cómo un árbol decide dónde dividir? Reducción en variancia

- Pasos para calcular la varianza:

1. Calcule la varianza para cada nodo.

2. Calcule la varianza para cada división como promedio ponderado de cada varianza del nodo.

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- La sobre o sub estimación (Overfitting) es uno de los desafíos claves enfrentados mientras aplicamos los árboles. Si no hay un conjunto de límites de un árbol de decisión, le dará 100% de precisión en el entrenamiento puesto que en el peor de los casos terminará haciendo 1 hoja por cada observación. Por lo tanto, la prevención de **overfitting** es fundamental mientras se modela un árbol de decisión y se puede hacer de dos maneras:

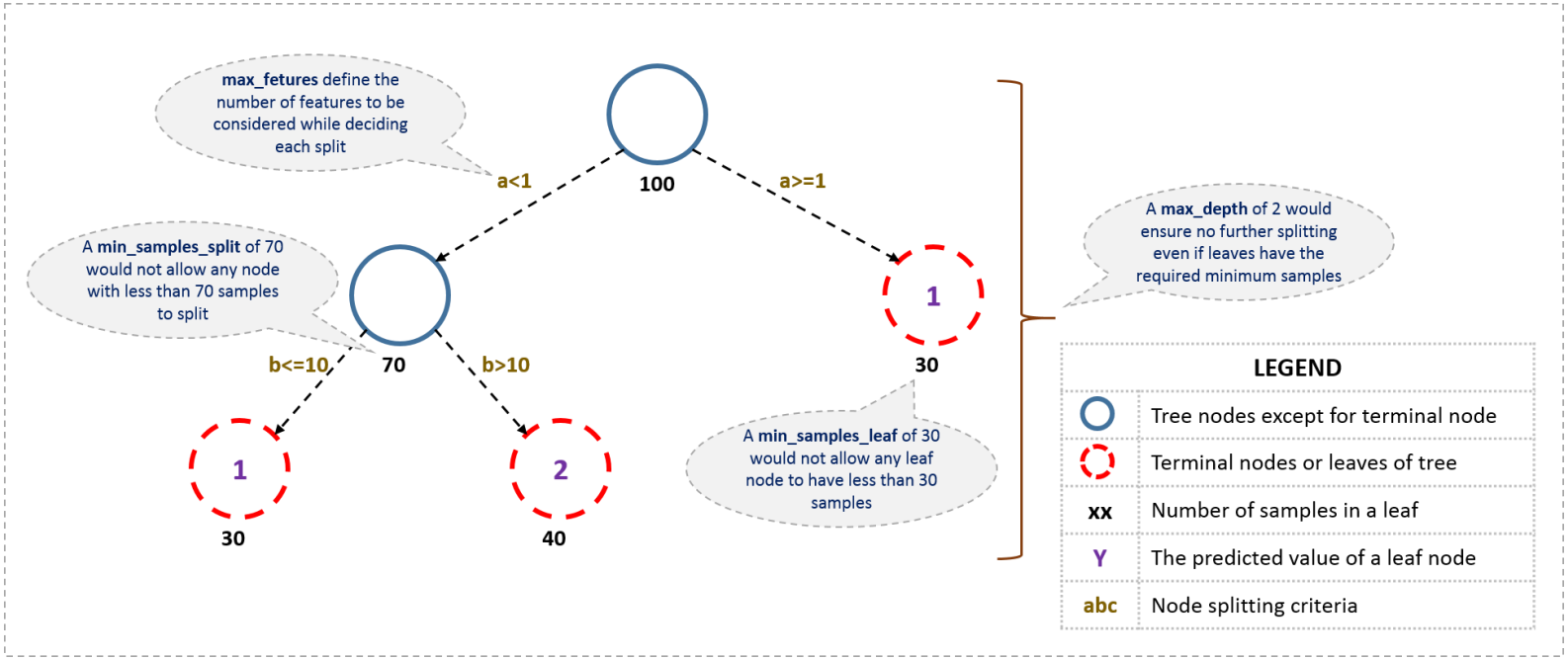
- 1.Configuración de restricciones en el tamaño del árbol
- 2.La Poda

Vamos a discutir ambos brevemente.

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

Estableciendo limitaciones en el tamaño del árbol

Esto se puede hacer usando varios parámetros que se utilizan para definir un árbol. Primero, veamos la estructura general de un árbol de decisión:



¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- A continuación se explican los parámetros utilizados para definir un árbol. Los parámetros descritos a continuación son independientes de la herramienta.

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- 1. Muestras mínimas para una división de nodos

-Define el número mínimo de muestras (u observaciones) que se requieren en un nodo a considerar para la división.

-Usado para controlar el ajuste excesivo. Valores más altos impiden que un modelo de aprendizaje de las relaciones que pueden ser muy específicos para la muestra particular seleccionada para un árbol.

-Demasiado valores altos pueden conducir a bajo ajustada por lo tanto, debe ajustarse utilizando CV.

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- 2. Muestras mínimas para un nodo terminal (hoja)
- 3. Profundidad máxima del árbol (profundidad vertical)
- 4. Número máximo de nodos terminales
- 5. Características máximas a considerar para la división

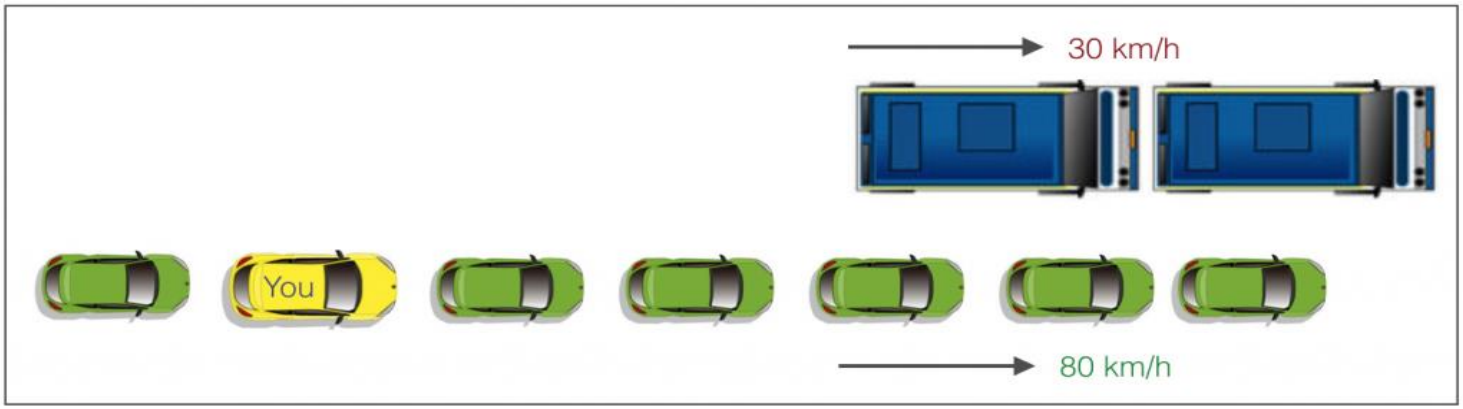
Refierase a la siguiente página:

https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- **La poda**

Como se discutió anteriormente, la técnica de establecer la restricción es un enfoque codicioso. En otras palabras, comprobará la mejor división instantáneamente y avanzará hasta que se alcance una de las condiciones de parada especificadas. Consideremos el siguiente caso cuando se conduce:



¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- Hay 2 carriles:

1. Un carril con coches que se mueven a 80km / h
2. Un carril con camiones que se mueven a 30km / h

En este instante, usted es el coche amarillo y tiene 2 opciones:

1. Tome a la izquierda y adelantar a los otros 2 coches rápidamente
2. Manténgase en movimiento en el carril actual

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- Vamos a analizar estas opciones. En la primera opción, adelantarás inmediatamente el coche y llegarás detrás del camión y empezarás a moverte a 30 km / h, buscando una oportunidad de retroceder. Todos los coches originalmente detrás de usted avanzan en el mientras tanto. Esta sería la opción óptima si su objetivo es maximizar la distancia recorrida en los siguientes 10 segundos. En la última opción, la venta a través de la misma velocidad, camiones cruzados y luego superar tal vez dependiendo de la situación por delante. Codicioso usted!
- Esta es exactamente la diferencia entre el árbol de decisión normal y poda. Un árbol de decisión con restricciones no verá el camión por delante y adoptará un enfoque codicioso tomando una izquierda. Por otro lado si usamos la poda, en efecto miramos algunos pasos adelante y tomamos una decisión.

¿Cuáles son los parámetros clave del modelado de árboles y cómo podemos evitar el exceso de ajuste en árboles de decisión?

- Así que sabemos que la poda es mejor. ¿Pero cómo implementarlo en el árbol de decisión? La idea es simple.

1. Primero hacemos el árbol de la decisión a una profundidad grande.

2. Luego empezamos en la parte inferior y empezamos a quitar las hojas que nos están dando resultados negativos cuando se comparan desde la parte superior.

3. Supongamos que una división nos está dando una ganancia de -10 (pérdida de 10) y luego la siguiente división en que nos da una ganancia de 20. Un árbol de decisión simple se detendrá en el paso 1, pero en la poda, vamos a ver que La ganancia total es de +10 y mantener ambas hojas.

¿Son modelos basados en árboles mejores que modelos lineales?

- "Si puedo usar la regresión logística para problemas de clasificación y regresión lineal para problemas de regresión, ¿por qué hay necesidad de usar árboles"?

- Muchos de nosotros tenemos esta pregunta. Y, esto es válido también.

- En realidad, puede utilizar cualquier algoritmo. Depende del tipo de problema que esté resolviendo. Echemos un vistazo a algunos factores clave que le ayudarán a decidir qué algoritmo utilizar:

¿Son modelos basados en árboles mejores que modelos lineales?

1. Si la relación entre variable dependiente y independiente es bien aproximada por un modelo lineal, la regresión lineal superará al modelo basado en árboles.

2. Si existe una alta no linealidad y relación compleja entre variables dependientes e independientes, un modelo de árbol superará a un método de regresión clásico.

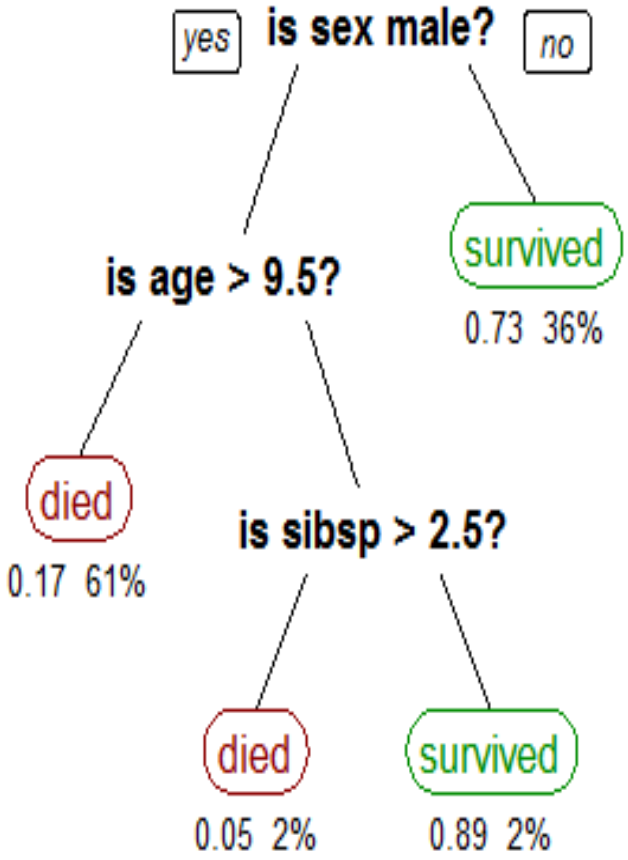
3. Si necesita construir un modelo que sea fácil de explicar a la gente, un modelo de árbol de decisión siempre será mejor que un modelo lineal. Los modelos de árboles de decisión son aún más fáciles de interpretar que la regresión lineal.

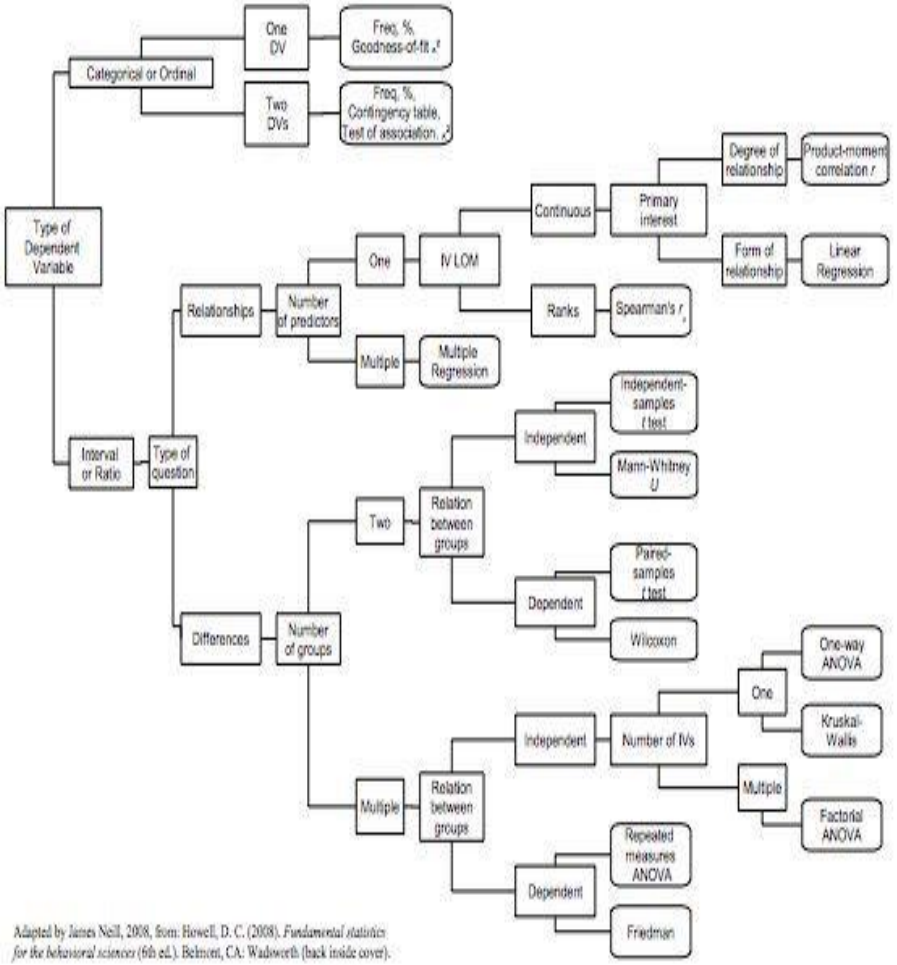
¿Qué obtenemos al final del análisis?

- Recapitulando, podemos obtener 3 cosas a partir de los árboles de decisión:

1. Predicciones o clasificaciones
2. Conocimiento de la estructura del árbol
3. Descripción de los nodos.
4. Toma de decisiones a partir de todo lo anterior.

¿Qué obtenemos al final del análisis?





Etapas de un análisis por árboles de decisión

- Siempre se recomienda:

1. Análisis descriptivo de las variables.
2. Estimación del modelo (predicción o clasificación).
3. La poda o las restricciones de los árboles.
4. La predicción o clasificación de las variables.
5. Interpretación de todo lo anterior.


The End