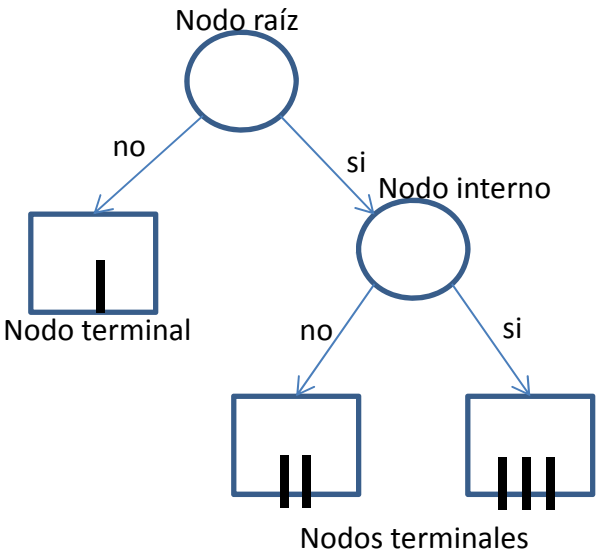
# ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN

Los árboles de clasificación y regresión (CART=Classification and Regression Trees) son una alternativa al análisis tradicional de clasificación/discriminación o a la predicción tradicional (regresión). Entre las ventajas de estos árboles CART podemos destacar su robustez a outliers, la invarianza en la estructura de sus árboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad.

Son árboles de regresión cuando la variable dependiente es continua y árboles de clasificación cuando la variable dependiente es de tipo cualitativo. En esencia, se trata de dar con un esquema de múltiples dicotomías o bifurcaciones, anidadas en forma de árbol , de manera que siguiendo cada una de las ramas del árbol obtengamos, al final , una predicción para la clase de pertenencia (clasificación) o para el valor que toman (regresión) los individuos que cumplen con las propiedades que se han ido exigiendo en las distintas bifurcaciones.

Los árboles de decisión se contruyen mediante un algoritmo conocido como segmentación recursiva, que es el proceso paso a paso para dicha construcción. Existen principalmente tres procedimientos: CHAID (Chi-Square Automatic Interaction Detector) ,QUEST (Quick Unbiased Efficient Statistical Tree ) y CART , en el que nos centraremos por ser relativamente sencillo y no fácil de implementar en R.

Si $Y$ es una variable respuesta y las $p$ variables predictoras son $x_1,x_2,\ldots x_p$, donde las $x$ son tomadas fijas y $Y$ es una variable aleatoria, el problema estadístico es establecer una relación entre $Y$ y las $x$ de tal forma que sea posible predecir $Y$ basado en los valores de $x$. Matemáticamente, se quiere estudiar la probabilidad condicional de la variable aleatoria $Y$. $P[Y=y \mid x_1,x_2,\ldots x_p]$ o una función de su probabilidad tal como la esperanza condicional. $E[\, Y \mid x_1,x_2,\ldots x_p]$, según se trate de un árbol de clasificación o de regresión, respectivamente. El árbol acaba teniendo la siguiente forma abreviada:




## Elementos del árbol

El árbol de la ilustración tiene tres niveles de nodos, el primer nivel tiene un único nodo en la cima llamado **nodo raíz**. Un **nodo interno** en el segundo nivel, y tres **nodos terminales** que están respectivamente en el segundo y tercer nivel. El nodo raíz y el nodo interno son particionados cada uno en dos nodos en el siguiente nivel los cuales son llamados **nodos hijos** (o ramas) izquierdo y derecho.

La completa **homogeneidad** de los nodos terminales es un ideal raramente alcanzado en el análisis de datos real. De esta manera, el objetivo del algoritmo de segmentación recursiva es hacer las variables resultantes en los nodos terminales tan homogéneas como sea posible.
Una medida cuantitativa de la homogeneidad es la noción de impureza.

## División de un nodo

Para dividir el nodo raíz en dos nodos homogéneos, se debe seleccionar entre los rangos de todas las variables predictoras el valor de la división que más se acerque al límite pureza para cada nodo hijo. La idea es que si el nodo A se divide en $A_L$ ( rama izquierda) y $A_R$ (rama derecha), la pureza de los dos nodos hijos debe ser mayor que la del nodo A. O su impureza menor. Impureza que suele medirse por la mínima probabilidad, la entropía o el índice de Gini.

## Nodos terminales

El proceso de segmentación recursiva continúa hasta que el árbol sea saturado en el sentido de que los sujetos en los nodos descendientes no se pueden partir en una división adicional. El número total de divisiones permitidas para un nodo disminuye cuando aumentan los niveles del árbol. Cualquier nodo que no pueda o no sea dividido es un nodo terminal.

### Impureza del nodo

Sea *Y* una variable dicotómica con valores 0 y 1. El nodo τ es más impuro cuando su impureza es máxima con $P[Y = \text{correcto}] = \frac{1}{2}$. La función de impureza tiene una forma cóncava y de puede definir formalmente como: $i(\tau) = \phi(P\{Y = \text{correcto}\})$. Donde $\phi$ tiene las siguientes propiedades.

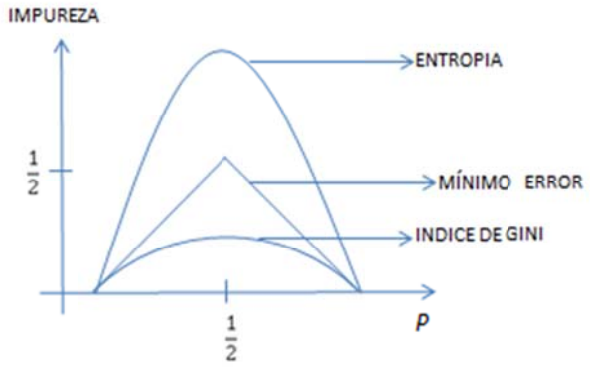
(i) $\phi \geq 0$ y (no negativa)

(ii) para cualquier $p \in (0,1)$, $\phi(p) = \phi(1-p)$ (simétrica) y $\phi(0) = \phi(1) < \phi(p)$ mínima para el éxito y el fracaso aboluto.

Las elecciones más comunes de funciones de impureza para la construcción de árboles de clasificación son:

- $\phi(p) = min(p,1-p)$, *(mínimo error o error de Bayes)*
- $\phi(p) = -p\ log(p) - (1-p)\ log(1-p)$, *(entropía)*
- $\phi(p) = p(1-p)$, *(índice de Gini)*

donde se define 0 log(0) = 0.




## MÉTODO CART

Esta metodología consiste de tres pasos:

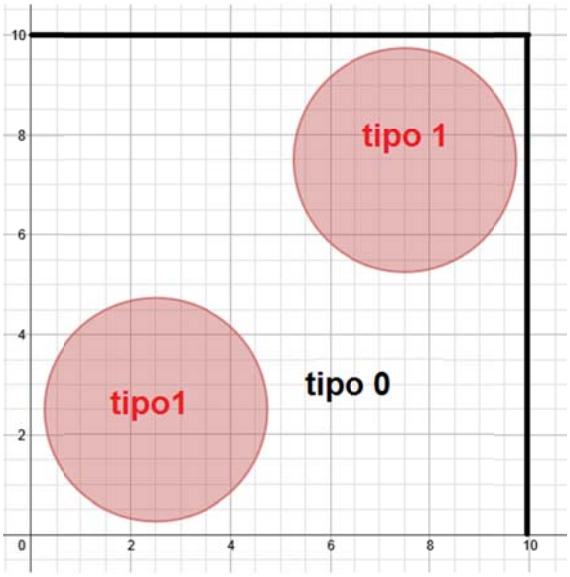
- Construcción del árbol saturado
- Elección del tamaño correcto
- Clasificación de nuevos datos a partir del árbol construido.

Una de las ventajas de los árboles de clasificación en relación con las técnicas tradicionales de análisis de datos multivariantes, especialmente con el Análisis Discriminante es su considerable mejor (más eficiente) comportamiento ante situaciones de estructura discriminante muy alejadas de la linealidad.

Veámoslo con un ejemplo extremo. Suponemos una situación en la que contamos con dos variables (discriminadoras) para intentar clasificar dos grupos de individuos. La razón para sólo considerar dos dimensiones es su fácil visualización.

Supongamos que las dos variables (X,Y) tienen una distribución uniforme en el rectángulo [0,10]×[0,10].

Supongamos que los individuos se agrupan en dos clases o pertenecen a dos grupos o tienen o no cierta propiedad. La variable dicotómica Clase, con campo de variación {0,1}, da cuenta de ello. Y que la situación es tal que si el individuo pertenece al círculo de radio $\sqrt{5}$ y centro el punto (7.75; 7.75) o bien si pertenece al círculo de radio $\sqrt{5}$ y centro el punto (2.25; 2.25) tendrá la propiedad considerada o pertenecerá al grupo 1, y si no está en el círculo no. Obviamente la estructura discriminante en el espacio X,Y es fuertemente no lineal tal como se muestra en el gráfico:




Una vez generados 1000 individuos con valores de *x* e *y* aleatorios según sendas distribuciones U(0,10) y clasificados como 1 o 0 según si su zona de pertenencia procedemos a discriminar los 1000 individuos según un análisis discriminante canónico y obtenemos :

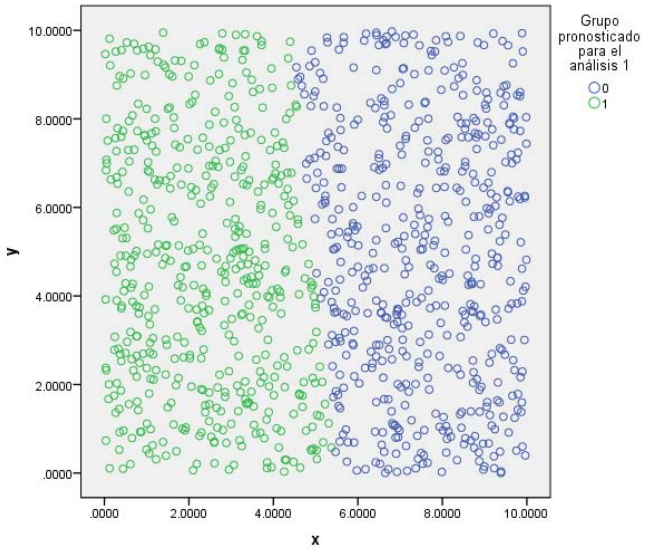
Una función canónica discriminante (no estandarizada) FCD= -1.881 + 0.338x + 0.038 y que conlleva a una clasificación ciertamente muy poco acurada con sólo un 53.8 % de individuos correctamente re-clasificados como muestra esta tabla:

**Resultados de la clasificación[a]**

| | | clas | Grupo de pertenencia pronosticado 0 | 1 | Total |
|---|---|---|---|---|---|
| Original | Recuento | 0 | 373 | 326 | 699 |
| | | 1 | 136 | 165 | 301 |
| | % | 0 | 53.4 | 46.6 | 100.0 |
| | | 1 | 45.2 | 54.8 | 100.0 |

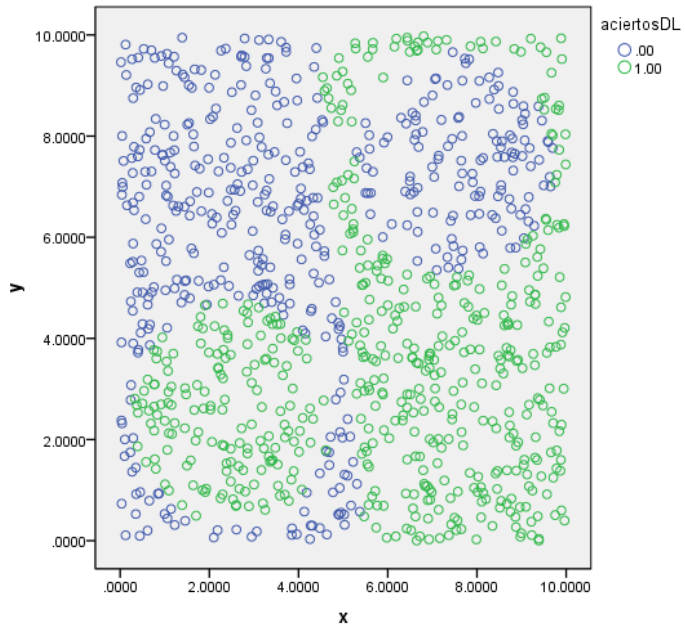
Clasificados correctamente el 53.8% de los casos agrupados originales

La reclasificación obtenida se muestra en este gráfico:




Que conlleva una muy pobre aproximación a la situación real. Ello es lógico si pensamos que establecer un “frontera lineal” entre los grupos tipo 1 y tipo 0 es algo ni siquiera posible con un grado razonable de aproximación.

Este otro gráfico muestra  los aciertos y errores de clasificación:

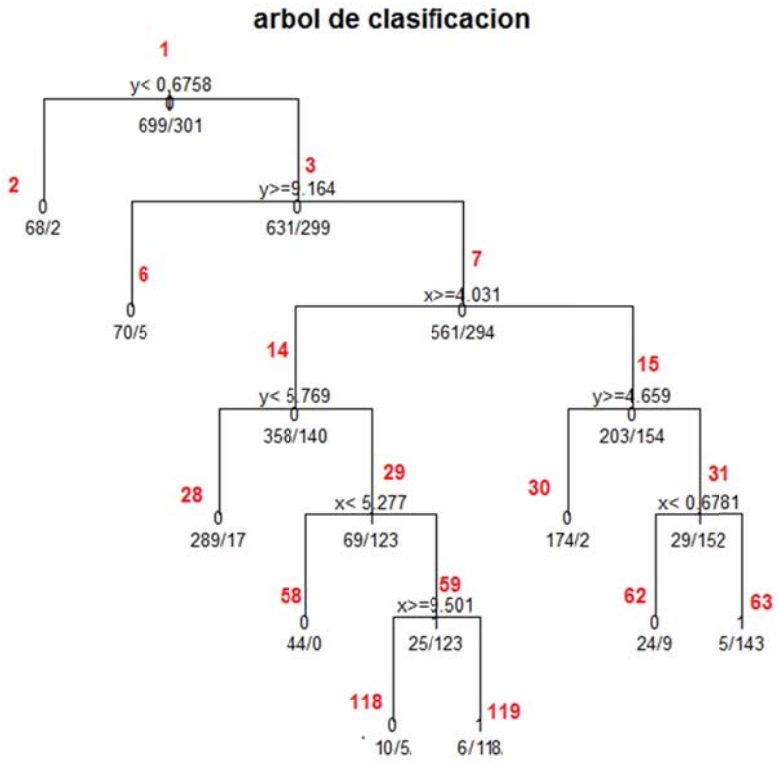



En cambio si realizamos una clasificación por arboles CART obtenemos el siguiente esquema discriminativo o discriminante:

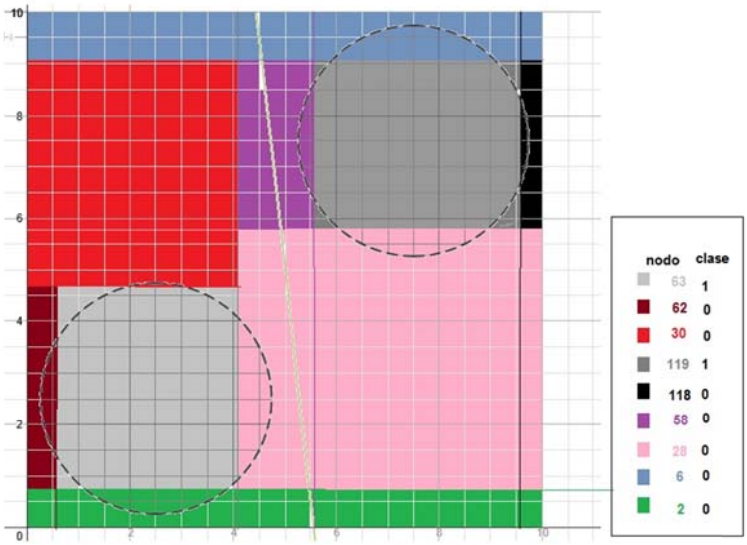
node), split, n, loss, yval, (yprob)  * denotes terminal node

1) root 1000 301 0 (0.69900000 0.30100000)
2) y< 0.6757896 70 2 0 (0.97142857 0.02857143) *
3) y>=0.6757896 930 299 0 (0.67849462 0.32150538)
6) y>=9.163761 75 5 0 (0.93333333 0.06666667) *
7) y< 9.163761 855 294 0 (0.65614035 0.34385965)
14) x>=4.031183 498 140 0 (0.71887550 0.28112450)
28) y< 5.769003 306 17 0 (0.94444444 0.05555556) *
29) y>=5.769003 192 69 1 (0.35937500 0.64062500)
58) x< 5.277384 44 0 0 (1.00000000 0.00000000) *
59) x>=5.277384 148 25 1 (0.16891892 0.83108108)
118) x>=9.501409 24 5 0 (0.79166667 0.20833333) *
119) x< 9.501409 124 6 1 (0.04838710 0.95161290) *
15) x< 4.031183 357 154 0 (0.56862745 0.43137255)
30) y>=4.658871 176 2 0 (0.98863636 0.01136364) *
31) y< 4.658871 181 29 1 (0.16022099 0.83977901)
62) x< 0.6780751 33 9 0 (0.72727273 0.27272727) *
63) x>=0.6780751 148 5 1 (0.03378378 0.96621622) *

En forma de árbol:

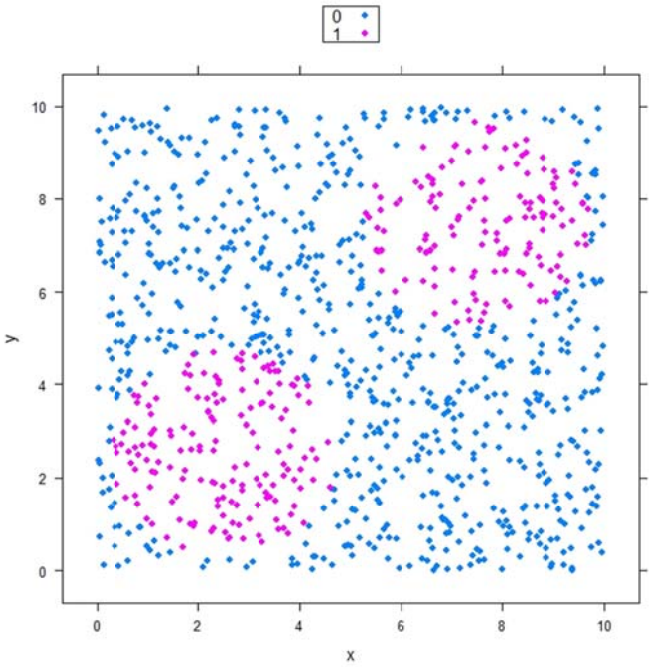



Si en el espacio de las variables *XY* desarrollamos la partición que genera este árbol de clasificación podemos observar cómo produce un resultado de re-clasificación notablemente más eficiente que la discriminación canónica:

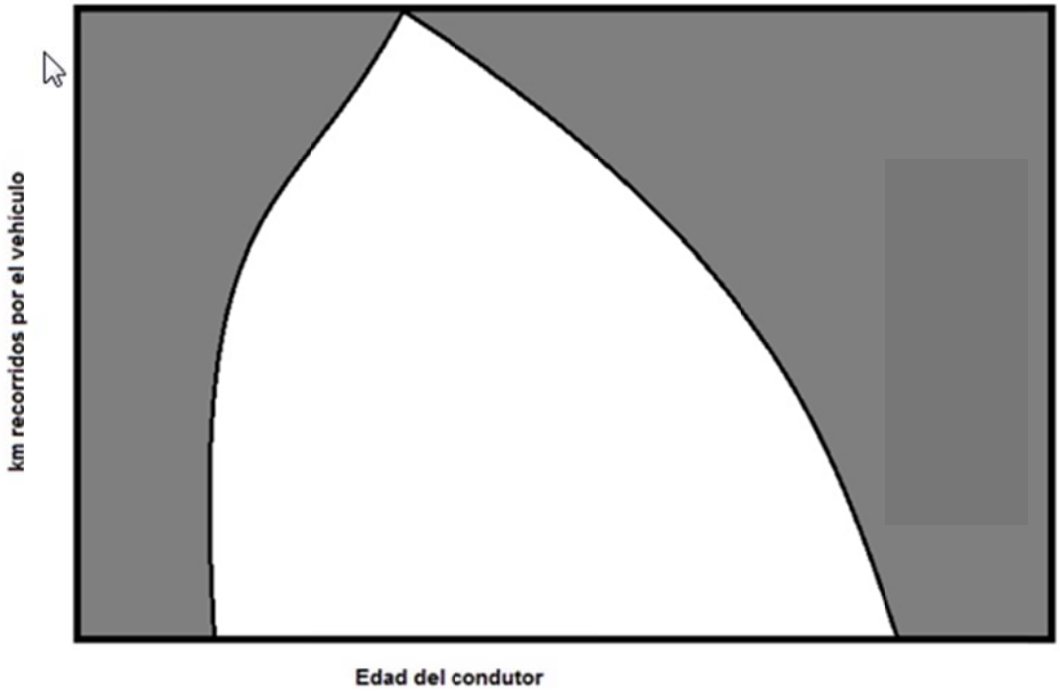



Con un mínimo porcentaje de errores en cuanto a su número (6.1 % en comparación con el 46.2 % de error de la discriminación canónica) y situándose estos en áreas muy concretas que

se corresponden con la aproximación rectangular de los círculos como se ve en este gráfico que representa las clases pronosticadas según el árbol

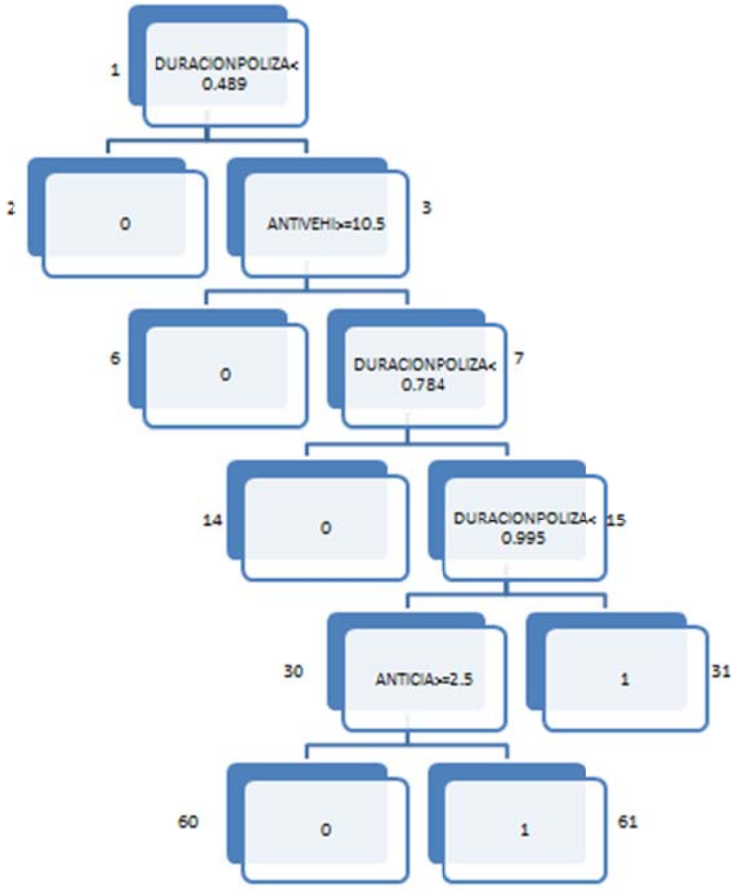



Se puede argumentar que este es un caso muy extremo, y es cierto, pero otras situaciones de alta "no-linealidad" en su estructura discriminante son perfectamente concebibles en un nuestro campo de trabajo. Así por ejemplo considerando la siniestralidad como categoría discriminable en función de los factores edad y km del vehículo podríamos encontrarnos con una situación similar a esta, en la que la zona sombreada sería la de riesgo de siniestro:

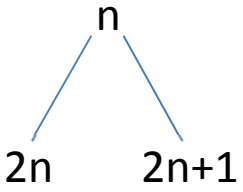


## UN EJEMPLO DE ÁRBOL DE CLASIFICACIÓN

Para ilustrar cómo se especifica y cómo funciona un CART proponemos a continuación un ejemplo de árbol de clasificación, en el que la variable dependiente "siniestro" es una variable dicotómica que toma valor 0 si no ha existido siniestro y 1 si ha existido algún siniestro:




La numeración de los nodos que nos proporciona el R es la siguiente:




Siendo los cálculos que proporciona R para cada uno de los nodos los presentados a continuación:

En cada línea se representan en orden:

El número del nodo), la condición, el número de (individuos) clientes del nodo (ni) , el número de (individuos) clientes del tipo erróneo, el tipo asignado, y (entre paréntesis): las probabilidades (condicionadas a caer en el nodo del tipo asignado y del tipo erróneo )

nº nodo) condición, nº indiv, nº indiv erróneos, el tipo asignado (prob condicionada caer en el tipo asignado, prob cond caer en el tipo erróneo)

1) root 92510 25191 0 (0.7276943 0.2723057)
2) duracionpoliza< 0.48984 44743 6503 0 (0.8546588 0.1453412)
3) duracionpoliza>=0.48984 47767 18688 0 (0.6087676 0.3912324)
6) antivehi>=10.5 15440 4486 0 (0.7094560 0.2905440)
7) antivehi< 10.5 32327 14202 0 (0.5606768 0.4393232)
14) duracionpoliza< 0.784817 18515 6900 0 (0.6273292 0.3726708)
15) duracionpoliza>=0.784817 13812 6510 1 (0.4713293 0.5286707)
30) duracionpoliza< 0.99532 11994 5898 1 (0.4917459 0.5082541)
60) anticia>=2.5 5472 2456 0 (0.5511696 0.4488304)
61) anticia< 2.5 6522 2882 1 (0.4418890 0.5581110)
31) duracionpoliza>=0.99532 1818 612 1 (0.3366337 0.6633663)

Con este tipo de árboles creamos una mejor segmentación, segmentación expresada como técnica multivariante, que debe encuadrarse entre los métodos de dependencia, ya que se establece una distinción entre variables cuyo comportamiento se desea explicar, o variables dependientes (como la cuantía del siniestro o el número de siniestros), y aquellas que se utilizan para explicar las anteriores, o variables independientes (factores de riesgo).

## ARBOL DE REGRESIÓN

Un modelo de árbol de regresión es una descripción condicional de *Y* dado *X*. Los dos componentes fundamentales del modelo son: un árbol binario $b$, nodos terminales y el vector de parámetros Θ = $(\theta_1, \theta_2, \dots, \theta_b)$, donde el parámetro $\theta_i$ está asociado al nodo terminal $N_i$. Un árbol de regresión crea un modelo explicativo y predictivo para una variable cuantitativa dependiente basada en variables explicativas cuantitativas y cualitativas.

### Profundidad de un nodo

La profundidad *d* de un nodo se define como el número de nodos de división que se encuentran por arriba de dicho nodo en el árbol.

### Regla de división

Para dividir un nodo de un árbol binario se consideran dos pasos:

1. Se selecciona una variable $X_i$
2. Si la variable seleccionada resulta cuantitativa, se selecciona aleatoriamente un valor *r* y se asignan al nodo hijo izquierdo las observaciones que cumplan la condición $X_i \leq r$, las restantes sin asignadas al nodo hijo derecho. Si la variable seleccionada resulta cualitativa, se selecciona un subconjunto *A* de las categorías de la variable $X_i$ y las observaciones con valores pertenecientes al conjunto *A* se asignan al nodo hijo izquierdo, las restantes al nodo hijo derecho