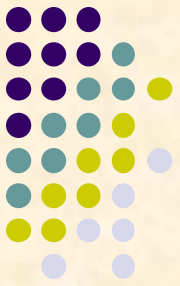


Data Mining con Árboles de Decisión

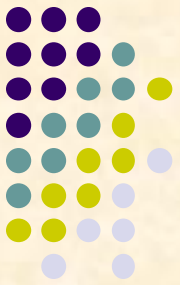
Jorge Martín Arevalillo
Profesor contratado doctor U.N.E.D
Consultor Estadístico / Data Mining





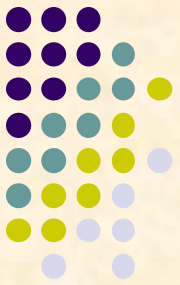
- ¿Qué es un árbol de decisión?
- Segmentación con modelos de árbol
- El algoritmo CHAID
- El algoritmo CART
- Árboles vs otras tecnologías
- Casos prácticos

Problema de credit scoring

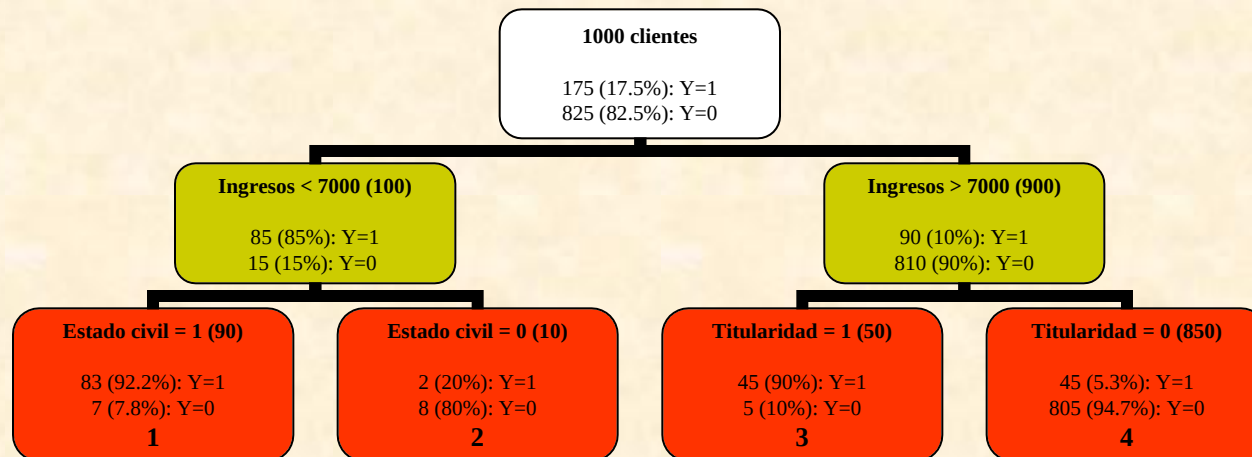


- **Definición del problema:** Búsqueda de perfiles de clientes con alto riesgo de morosidad.
- Formulación matemática. La variable criterio Y presenta dos estados: morosidad ($Y=1$) y no morosidad ($Y=0$)
- Situación inicial: si ingresos < 7000 € / año clasificar como moroso; en otro caso, clasificar como no moroso y conceder el crédito
- **Objetivo:** Mejorar la clasificación inicial y reducir pérdidas.

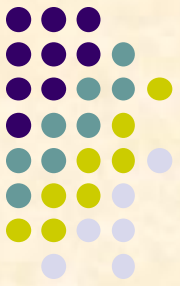
¿Qué es un modelo de árbol?



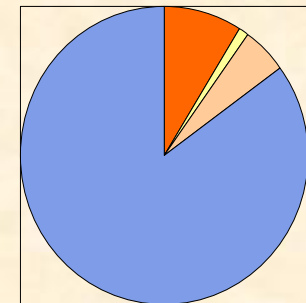
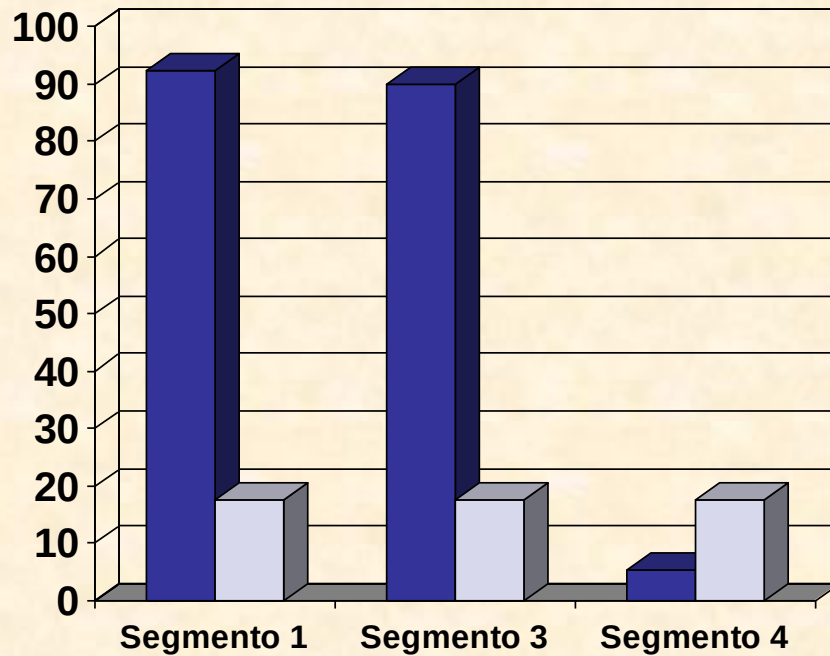
- El modelo de árbol se utiliza para clasificar cada solicitud de crédito en uno de los dos estados de Y
- El ajuste del modelo ha dado lugar a una nueva segmentación. Se obtuvieron los siguientes perfiles



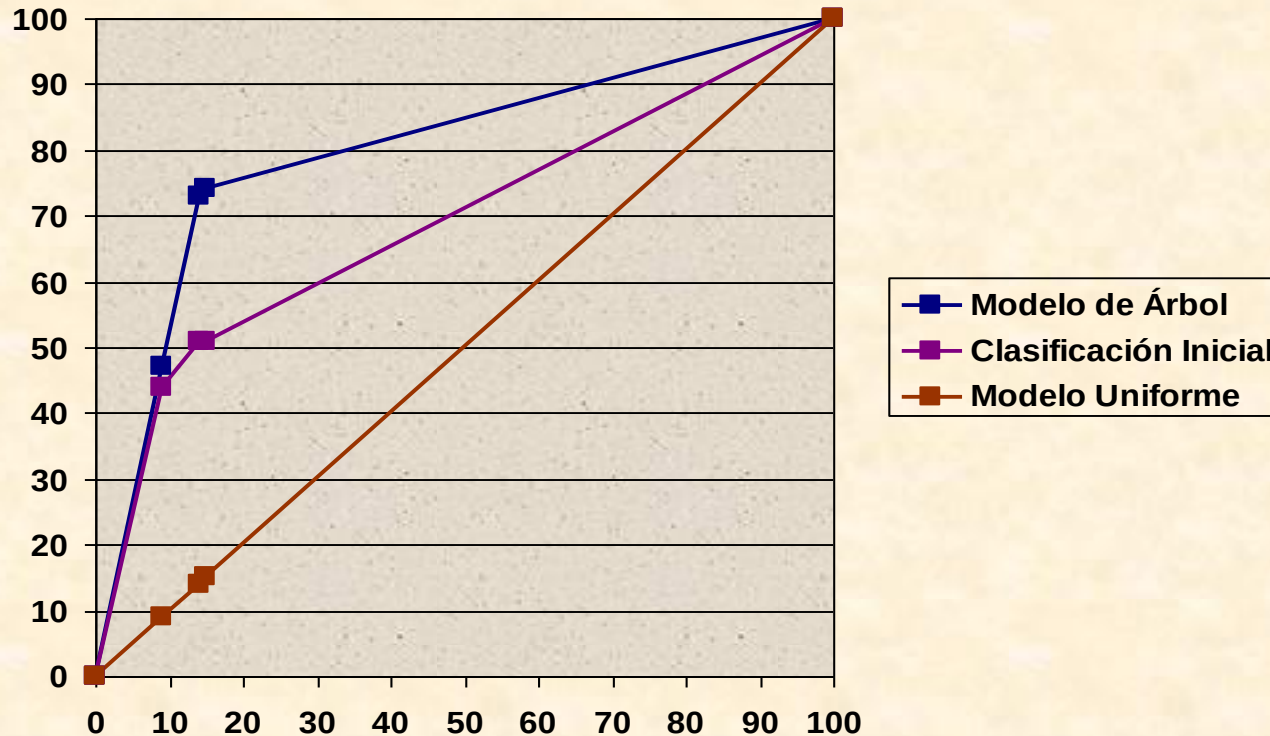
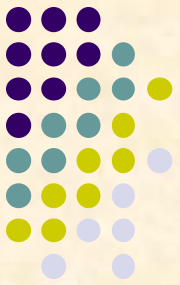
Resultados del modelo



- Los nodos terminales representan cuatro segmentos de la base de datos

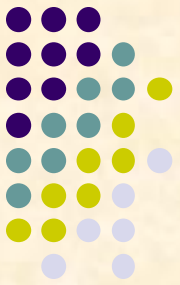


Ganancias



- Los nodos terminales 1 y 3 representan el 14% de los clientes. En ellos se localiza el 73% de la morosidad

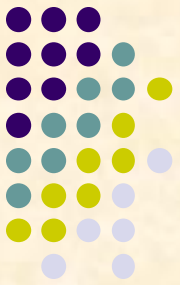
Sectores de aplicación



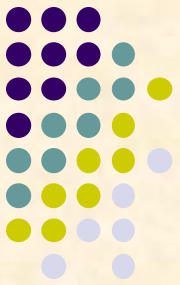
- Industria del seguro
- Credit scoring
- Detección de fraude
- Sector retail
- Optimización de campañas (telecos, bancos, gran consumo....)



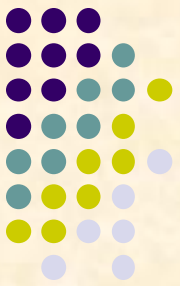
Aplicación de la tecnología



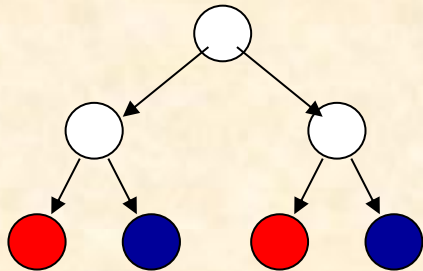
- Los Árboles de Decisión se pueden utilizar para modelizar problemas de
 - **Clasificación**
 - Binaria (fraude vs no fraude)
 - Multiclase (niveles de satisfacción: completamente, bastante, poco satisfecho, totalmente insatisfecho)
 - **Regresión**
 - Pagos que realiza una compañía de seguros
 - Gasto mensual de los clientes de una cadena de supermercados



- ¿Qué es un árbol de decisión?
- Segmentación con modelos de árbol
- El algoritmo CHAID
- El algoritmo CART
- Árboles vs otras tecnologías
- Casos prácticos



Segmentación con modelos de árbol



0	1
1	
	0

- La partición en cada nodo describe dos conjuntos disjuntos de la base de datos
- El corte viene dado por una o varias condiciones en una de las variables explicativas
- El particionamiento es recursivo. Se detiene en los nodos terminales
- A cada nodo terminal se le asigna uno de los estados de la variable criterio Y
- Para cada nueva observación, el estado de la variable respuesta se predice por el estado del nodo terminal al que dicha observación pertenece

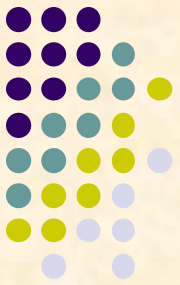
Algoritmos de segmentación



- El criterio que determina el corte en cada nodo, el mecanismo de segmentación y el criterio de parada (que permite decidir si un nodo es terminal o no) han dado lugar a distintos algoritmos de segmentación. Nos centramos en

- **CHAID.** Es el acrónimo de Chi-Squared Automatic Interaction Detector
- **CART.** Es el acrónimo de Classification and Regression Trees

Software



- CHAID lo comercializa IBM. Viene incorporado en la plataforma de Data Mining IBM SPSS Modeler

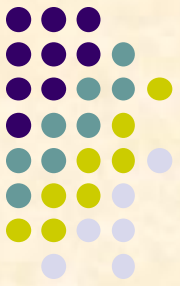


- La exclusividad de la comercialización del algoritmo CART la tiene Salford Systems. Lo distribuye en la herramienta CART[®]



- Andy Liaw y Matthew Wiener han implementado una versión libre de CART. Está en la librería **rpart** del entorno R de programación. La librería **party** implementa una variante de la tecnología





- ¿Qué es un árbol de decisión?
- Segmentación con modelos de árbol
- El algoritmo CHAID
- El algoritmo CART
- Árboles vs otras tecnologías
- Casos prácticos

El algoritmo CHAID



- Procede del ámbito de la Inteligencia artificial. Desarrollado por Kass a principios de los años 80
- Asume que las variables explicativas son categóricas u ordinales. Cuando no lo son, se discretizan
- Inicialmente se diseñó para el caso de variable respuesta Y categórica. Posteriormente se extendió a variables continuas
- Utiliza contrastes de la χ^2 de Pearson y la F de Snedecor
- El corte en cada nodo es multi-vía

Contraste de homogeneidad

χ^2 de Pearson



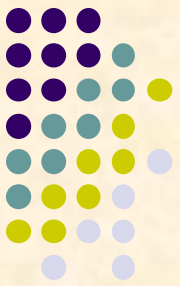
	Clases X				
Categorías Y	CX ₁	CX ₂	...	CX _r	Tot. filas
CY ₁	n_{11}	n_{12}	...	n_{1r}	m_1
CY ₂	n_{21}	n_{22}	...	n_{2r}	m_2
⋮	...	...	...	...	⋮
CY _s	n_{s1}	n_{s2}	...	n_{sr}	m_s
Tot. columnas	n_1	n_2	...	n_r	n

H_0 : Homogeneidad

H_1 : Heterogeneidad

Estadístico de contraste:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_i m_j / n)^2}{n_i m_j / n}$$



Contraste F de Snedecor

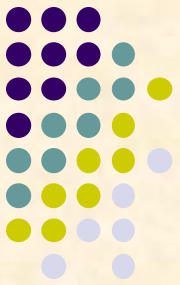
	Clases X			
	CX_1	CX_2	...	CX_r
Variable Criterio (Y)	y_{11}	y_{12}	...	y_{1r}
	y_{21}	y_{22}	...	y_{2r}
	...	...	...	...
	y_{n_11}	y_{n_22}	...	$y_{n_r r}$
Medias grupos	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_r$

H_0 : Igualdad de medias
 H_1 : Medias distintas

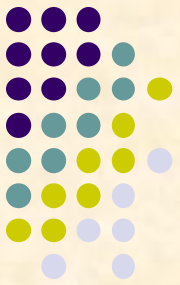
Estadístico de contraste:

$$F = \frac{\frac{\sum_{j=1}^r n_j (\bar{y}_j - \bar{y})^2}{r-1}}{\frac{\sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n-r}}$$

El criterio de corte CHAID



- CHAID considera todos los cortes posibles en todas las variables. Selecciona el corte que da el menor **p-valor** asociado a una medida de contraste estadístico
- Si la variable criterio es **categorica** la medida es la χ^2 de Pearson. Si es **continua** la medida es la del test de la F
- La búsqueda de la variable y el corte óptimo se lleva a cabo en dos fases: **merge** (fusión de categorías) y **split** (selección de la variable de corte)



Fase merge

- Agrupa estados o valores de las variables explicativas. Para cada variable, agrupará los estados de cuya unión se obtenga el de menor significación estadística del contraste; siempre que ésta supere un umbral α_{merge} , fijado de antemano

Fase merge

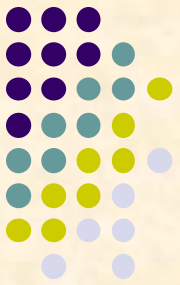
Paso 1. Encontrar el emparejamiento de categorías que conducen al mayor p-valor — p^* — para el test de la χ^2 o F

Paso 2. Comparar p^* con el umbral establecido α_{merge}

- Si $p^* > \alpha_{\text{merge}}$ agrupar las dos categorías en una sola. Volver a paso 1
- Si $p^* < \alpha_{\text{merge}}$ ir a paso 3

Paso 3. Ajustar el p-valor utilizando el multiplicador de Bonferroni:

$p_{\text{adj}} = p^* \cdot B$, siendo $B = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!}$, c el número original de categorías y r el número de categorías tras el agrupamiento



Fase split

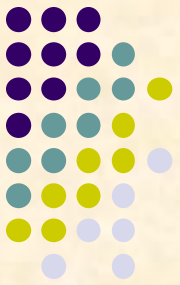
- De la fase merge se toma la agrupación en la variable con contraste más significativo (menor p-valor ajustado)
- Si la significación estadística es inferior a un mínimo α_{split} prefijado, se toma dicha agrupación como partición del nodo

Fase split

Paso 1. Encontrar la variable predictora con el menor p-valor ajustado $p^\dagger$

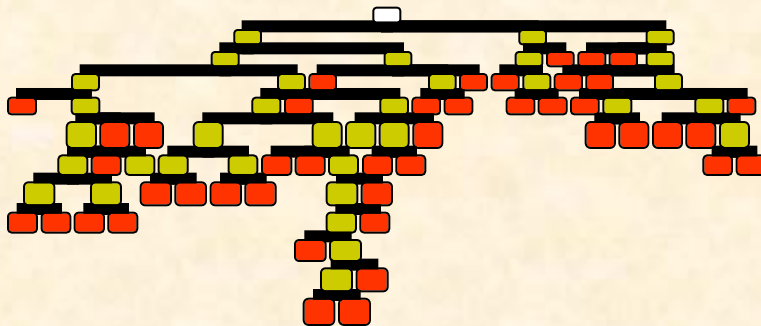
Paso 2. Comparar $p^\dagger$ con el umbral establecido α_{split}

- Si $p^\dagger < \alpha_{split}$ particionar el nodo utilizando el agrupamiento de categorías obtenido en la fase merge
- Si $p^\dagger > \alpha_{split}$ declarar el nodo terminal



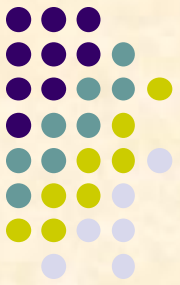
El criterio de parada CHAID

- Se fija de antemano por el experimentador. Depende de
 - El nivel α_{split}
 - Número de niveles de la estructura de árbol
 - Un umbral mínimo para el tamaño de los nodos descendientes



Complejidad

Caso práctico: consumo de vehículos

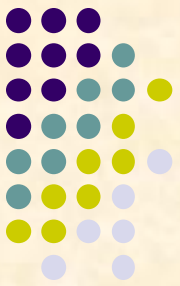


- Un estudio realizado en 1983 midió 8 variables sobre un total de 406 vehículos. Se eliminaron registros con NA y valor 3 o 5 en la variable cylinders. Quedaron 385 casos

VARIABLES	mpg	cylinders	displacement	horsepower	weight	acceleration	model.year	origin
TIPO	Numérica	Ordinal (4,6,8)	Numérica	Numérica	Numérica	Numérica	Numérica	Categórica

Fuente: <http://lib.stat.cmu.edu>

Consumo de vehículos (cont.)



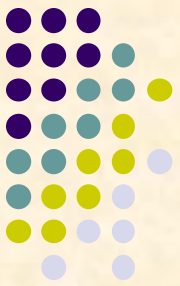
Objetivo: Explicar el nivel de consumo de los vehículos en función de sus características

ID	mpg	Consumo	cylinders	displacement	horsepower	Weight	acceleration	model.year	origin
hi 1200d	9,0	2	8	304,0	193,0	4732	18,5	70	1
ford f250	10,0	2	8	360,0	215,0	4615	14,0	70	1
chevy c20	10,0	2	8	307,0	200,0	4376	15,0	70	1
dodge d200	11,0	2	8	318,0	210,0	4382	13,5	70	1
mercury marquis	11,0	2	8	429,0	208,0	4633	11,0	72	1
chevrolet impala	11,0	2	8	400,0	150,0	4997	14,0	73	1
oldsmobile omega	11,0	2	8	350,0	180,0	3664	11,0	73	1
dodge monaco (sw)	12,0	2	8	383,0	180,0	4955	11,5	71	1
oldsmobile delta 88 royale	12,0	2	8	350,0	160,0	4456	13,5	72	1
mercury marquis brougham	12,0	2	8	429,0	198,0	4952	11,5	73	1

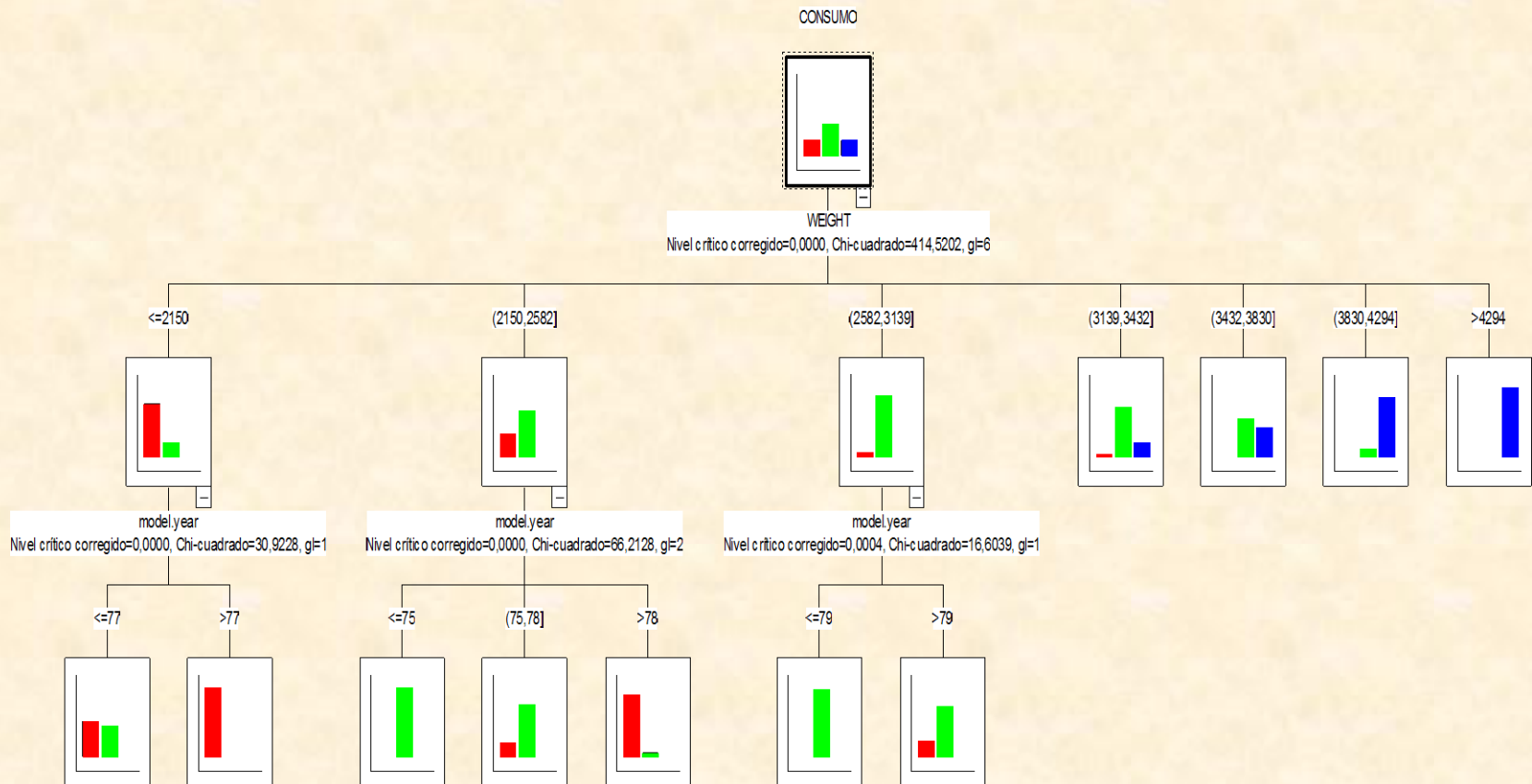
- Consumo

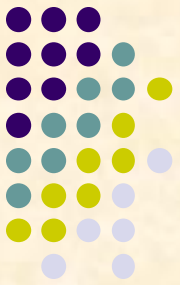
Bajo (**0**): $\text{mpg} > 29$. Medio (**1**): $17 < \text{mpg} \leq 29$. Alto (**2**): $\text{mpg} \leq 17$

Segmentación con CHAID



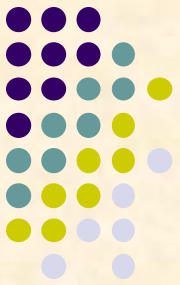
- Árbol exploratorio: $\alpha_{\text{merge}}=0.05$. $\alpha_{\text{split}}=0.01$. Parental=30. Filial=15





- ¿Qué es un árbol de decisión?
- Segmentación con modelos de árbol
- El algoritmo CHAID
- El algoritmo CART
- Árboles vs otras tecnologías
- Casos prácticos

El algoritmo CART

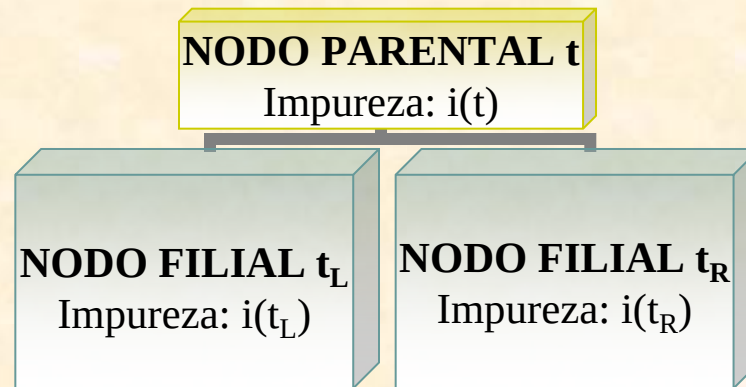


- Procede del ámbito de la Estadística. Desarrollado por matemáticos de la universidad de Berkeley y Stanford (Breiman, Friedman, Olshen y Stone) a mediados de los 80
- Trabaja con variables de todo tipo. No necesita discretizar las variables explicativas continuas
- El corte en cada nodo viene dado por reglas de tipo binario. Se pueden formular como preguntas: ¿Es $X_k < a$? ¿Pertenece X_k a un subconjunto E de estados?
- Da lugar a estructuras de árbol de mayor profundidad



El criterio de corte CART

- Se basa en la idea de impureza. CART selecciona el corte que conduce al mayor decrecimiento de la impureza. Así se consiguen descendientes homogéneos en la variable respuesta Y



$$\Delta i(t, t_L, t_R) = i(t) - p_L \cdot i(t_L) - p_R \cdot i(t_R).$$



Medidas de impureza

- Para el problema de clasificación:

- La entropía $i(t) = -p(t) \log p(t) - (1 - p(t)) \log(1 - p(t))$

- El índice de Gini $i(t) = p(t)(1 - p(t))$

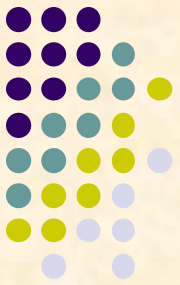
- El criterio del Twoing: Para un problema con J clases, busca maximizar la siguiente medida

$$\frac{PLPR}{4} \left(\sum_{j=0}^{J-1} |p(j|t_L) - p(j|t_R)| \right)^2$$

- Para el problema de regresión:

- La agregación de las varianzas de todos los nodos terminales

El criterio de parada

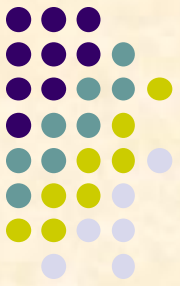


La alternativa de CART es no parar



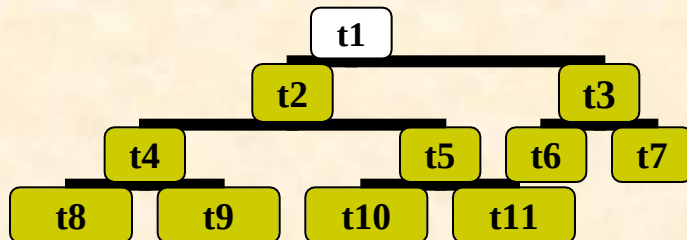
- CART propone segmentar la base de datos hasta obtener una estructura de árbol lo más compleja posible
- Un nodo se declara como terminal sólo si su tamaño es inferior a un umbral preestablecido (normalmente muy pequeño)
- La complejidad de un árbol se mide por el número de nodos terminales
- A continuación, se **poda** la estructura de árbol maximal que se ha obtenido

La poda de una rama

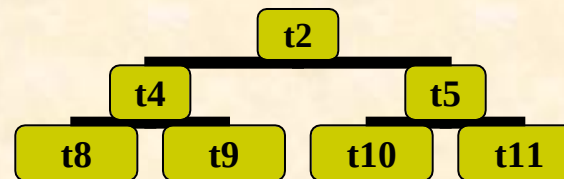


- Una rama del nodo t de un árbol T está formada por él y todos sus descendientes
- Podar la rama en t consiste en eliminar todos los descendientes del nodo t

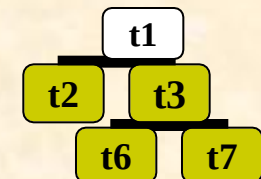
Árbol T

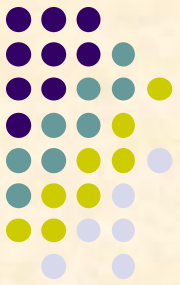


Rama de t_2



Árbol podado





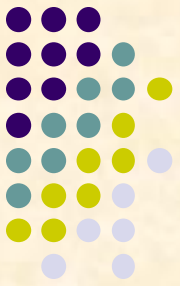
La poda del árbol maximal

- El proceso de poda se apoya en la siguiente medida:

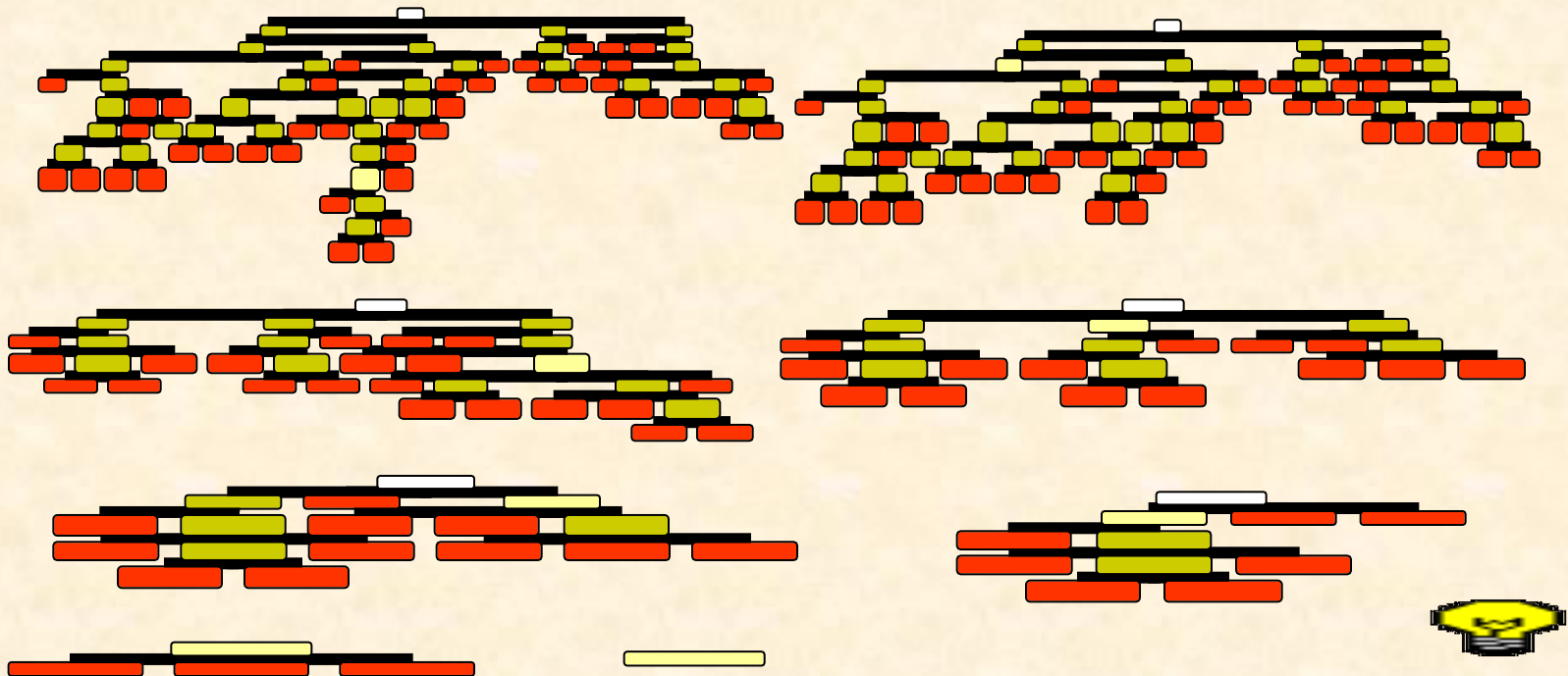
$$R_{\alpha}(T) = R(T) + \alpha|\tilde{T}|$$

- Combina el riesgo o coste de predicción y la complejidad.
- El primer sumando mide el riesgo de T (tasa de error si el problema es de clasificación o la suma de las varianzas residuales si es de regresión)
- El segundo sumando penaliza las estructuras de árbol complejas. El parámetro $\alpha \geq 0$ se denomina parámetro de complejidad

El proceso de poda

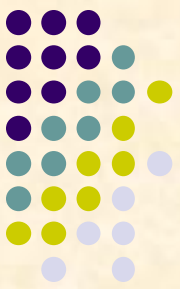


- Se realiza de una manera inteligente, eliminando las ramas más débiles. La idea es encontrar subárboles que minimicen $R_\alpha(T)$



Problema de seleccion

La selección del subárbol óptimo



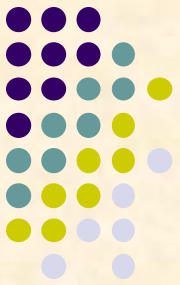
- La selección del subárbol óptimo es un problema de estimación del riesgo. Está relacionado con el problema del sobreajuste



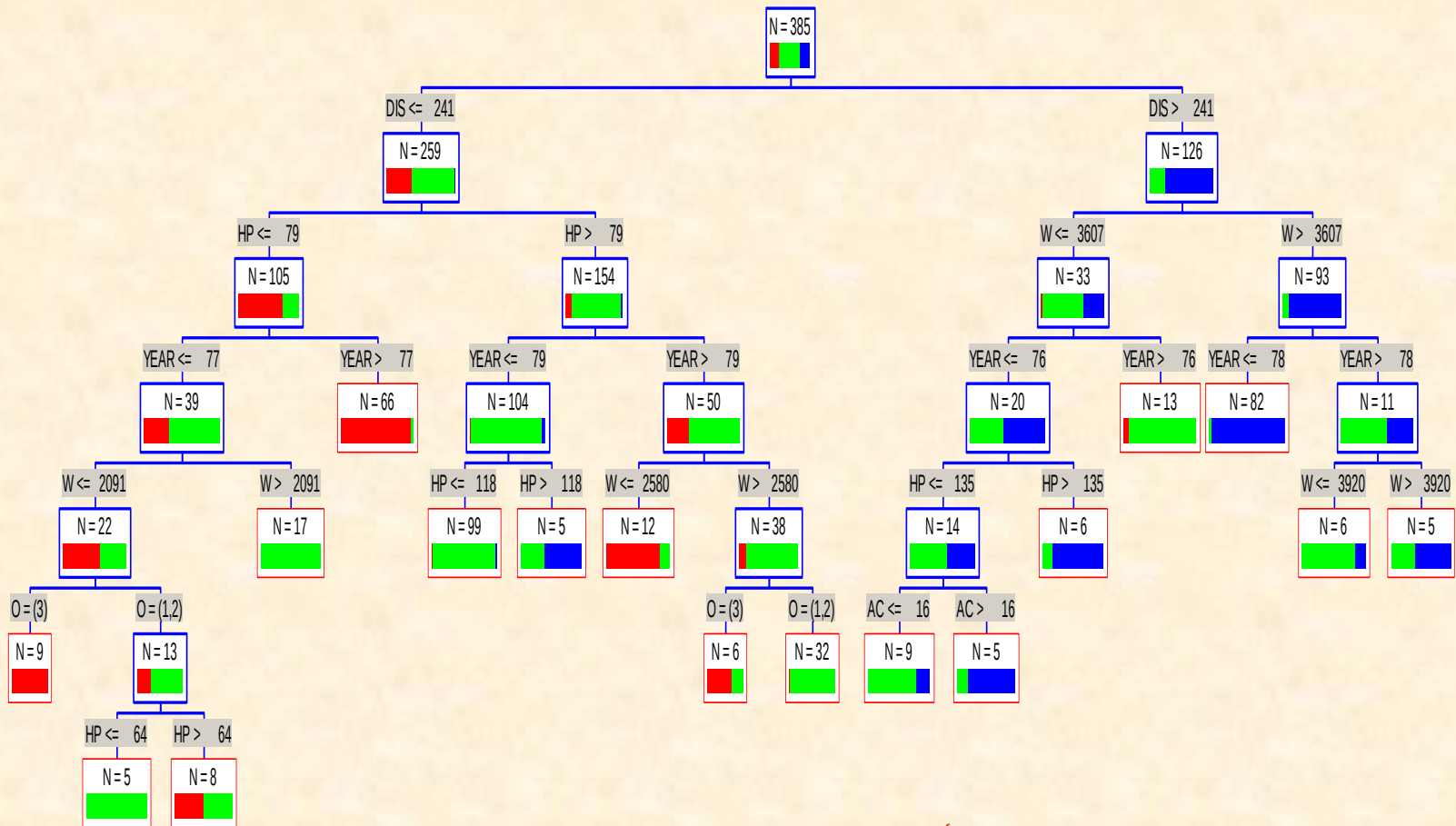
- Muestra de aprendizaje (learning o training)
- Muestra de validación (test)

Consumo vehículos.

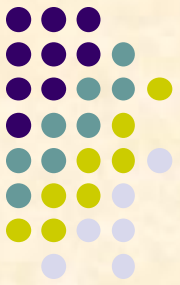
Segmentación con CART



- Árbol exploratorio: Gini. Nodo parental=10. Nodo filial=5

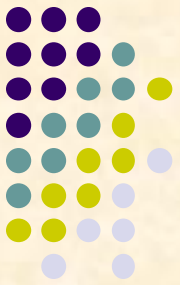


CHAID vs CART

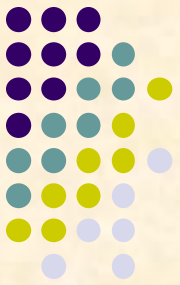


- El particionamiento CHAID es multivía; el de CART es binario
- Por tanto, las estructuras de árbol CHAID suelen ser más simples que las dadas por CART
- CART no requiere discretizar variables. CHAID sí
- Análisis experimentales han demostrado que CHAID es más vulnerable a generar falsos positivos
- CART suele tener mayor capacidad predictiva
- **Recomendación:** prueba varios ajustes CHAID y CART y selecciona el de mayor utilidad de negocio

Guión

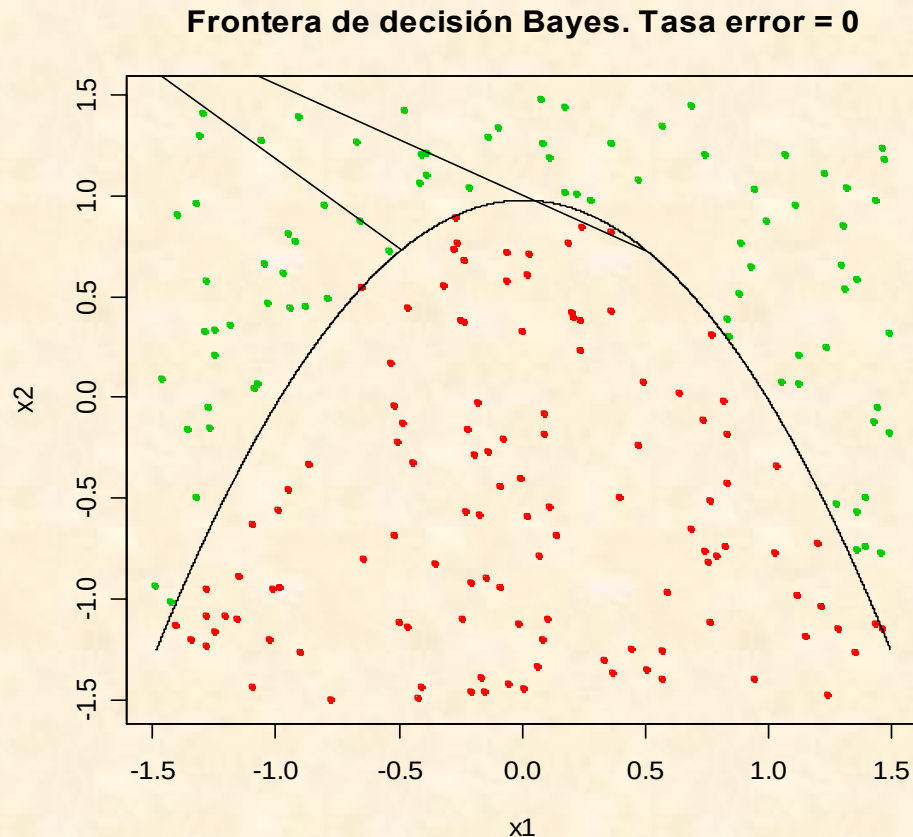


- ¿Qué es un árbol de decisión?
- Segmentación con modelos de árbol
- El algoritmo CHAID
- El algoritmo CART
- Árboles vs otras tecnologías
- Casos prácticos



Ejemplo 1

- Problema de clasificación binaria. Frontera de decisión cuadrática



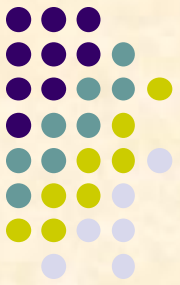
- Se simularon 200 observaciones del par (x_1, x_2) ; ambas componentes son independientes con distribución uniforme en $(-1.5, 1.5)$

- La clasificación viene dada por:

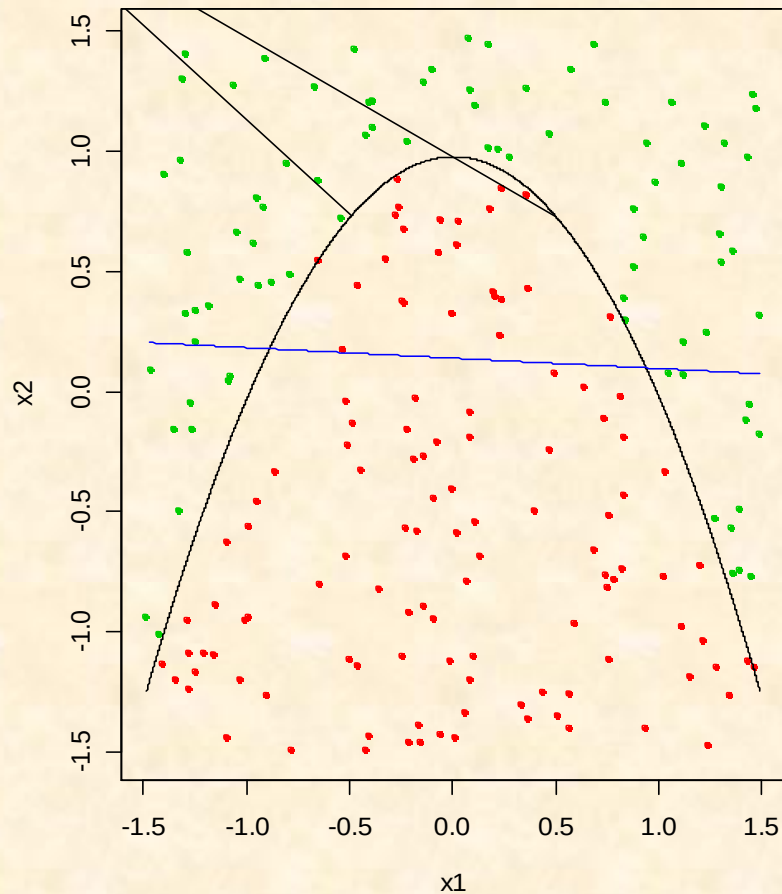
Clase **2**: $x_2 > 1 - x_1^2$

Clase **1**: $x_2 < 1 - x_1^2$

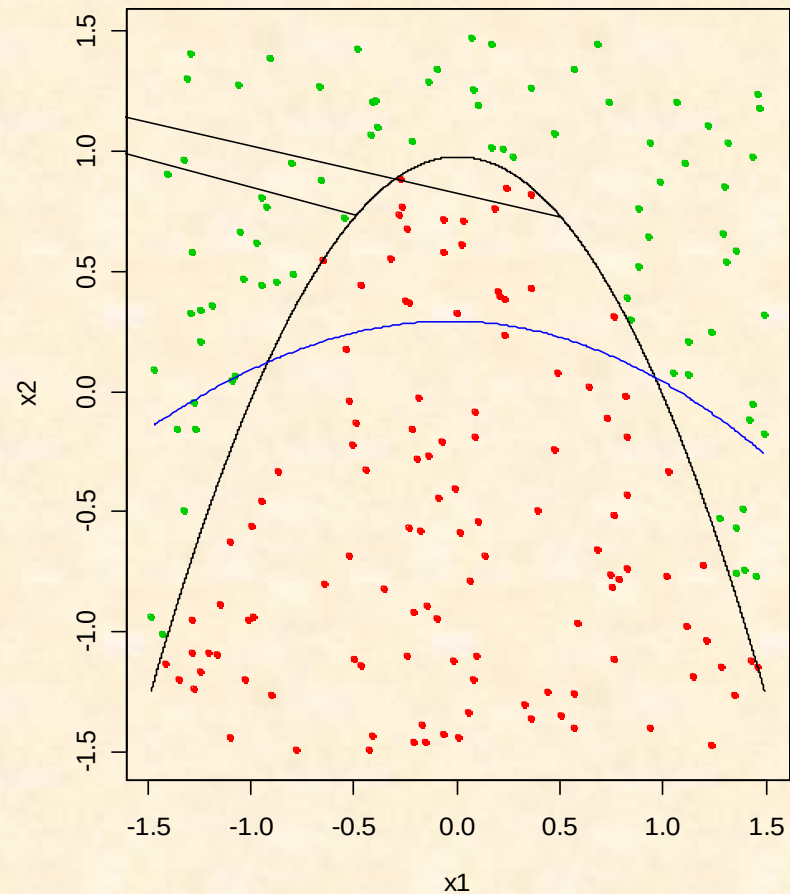
Soluciones lda y qda



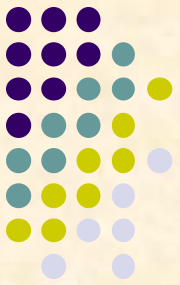
Frontera decisión LDA. Tasa error = 19.91



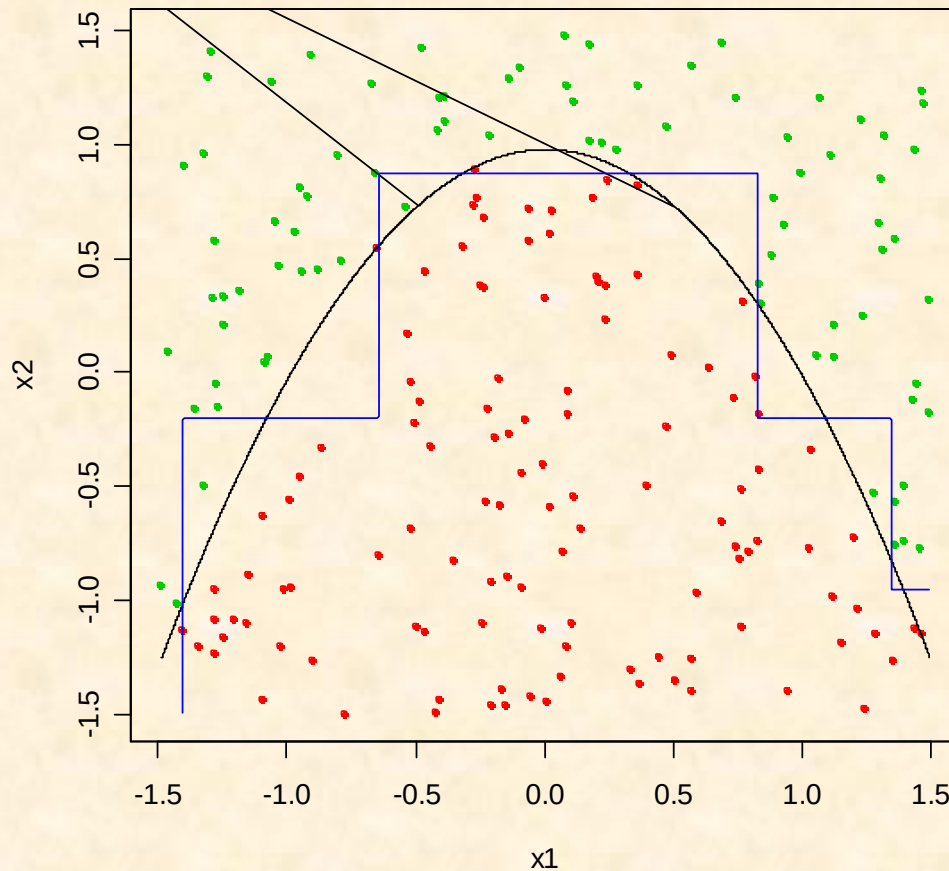
Frontera decisión QDA. Tasa error = 15.75



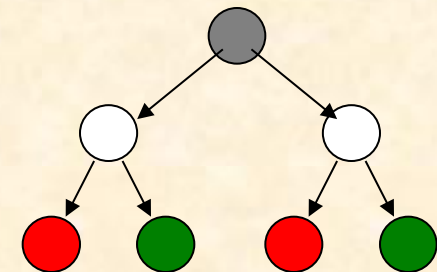
Solución árbol de decisión

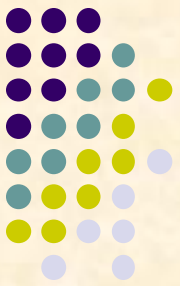


Frontera decisión CART. Tasa error = 8.1



Solución obtenida
mediante CART.
Árbol sin podar

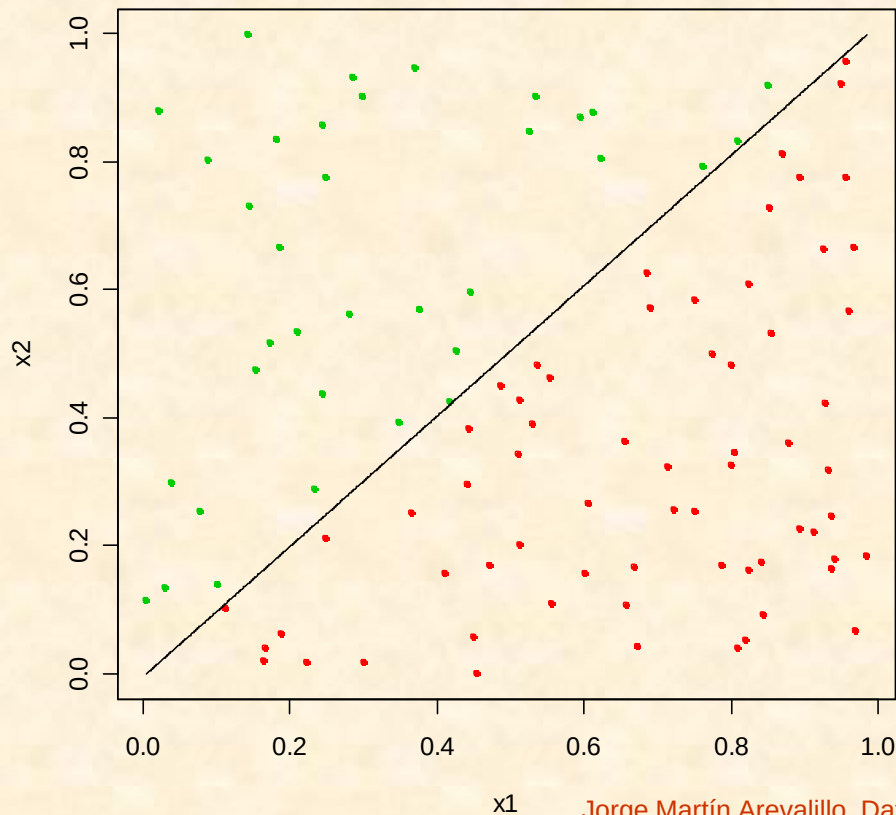




Ejemplo 2

- Problema de clasificación binaria. Frontera de decisión lineal

Frontera de decisión Bayes. Tasa error = 0



- Se simularon 100 observaciones del par (x_1, x_2) ; ambas componentes son independientes con distribución uniforme en $(0,1)$
- La clasificación viene dada por:

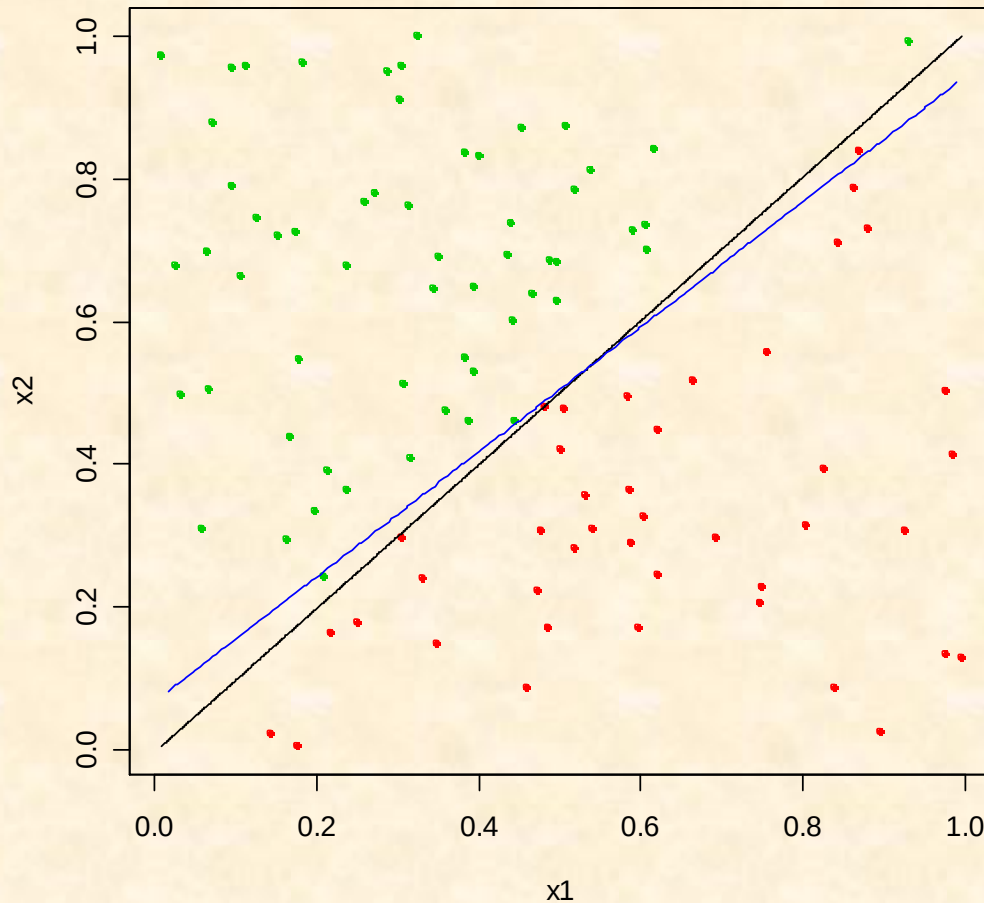
Clase **2**: $x_2 < x_1$

Clase **1**: $x_2 > x_1$

Solución Ida



Frontera decisión LDA. Tasa error = 3.06

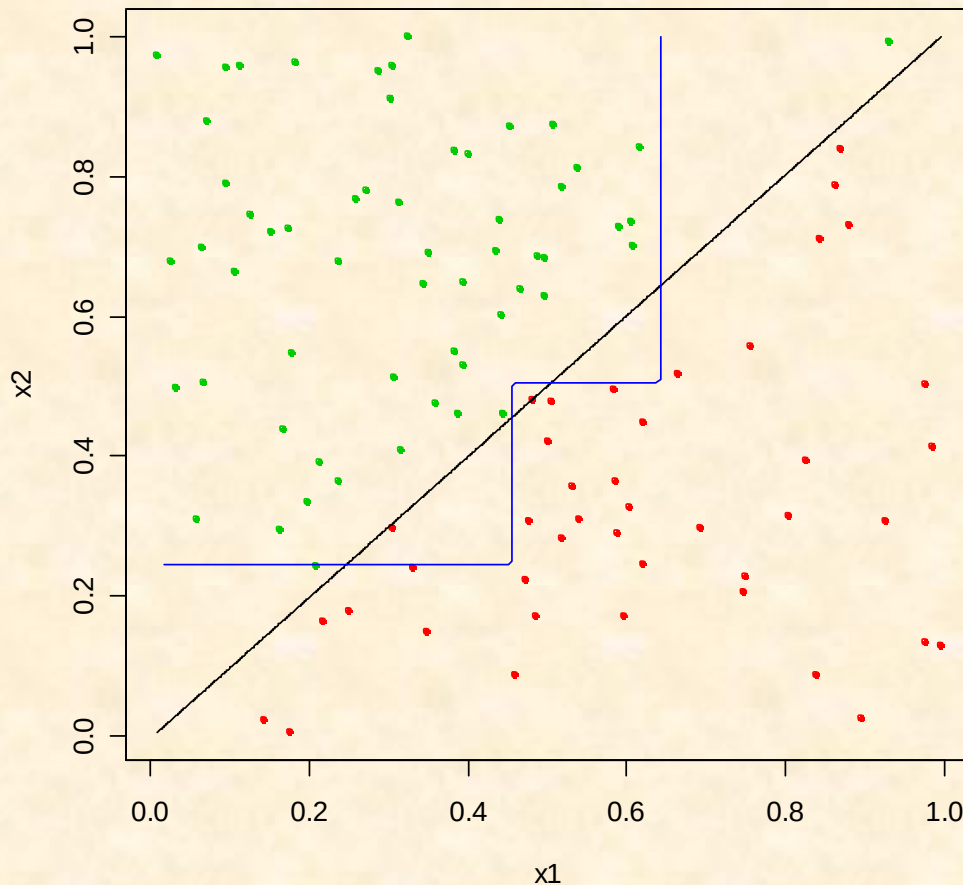


A pesar de que falla la hipótesis de normalidad, la solución lda es una buena aproximación

Solución árbol de decisión

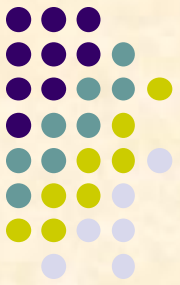


Frontera decisión CART. Tasa error = 12.79



CART (sin poda). Peor que el discriminante lineal

Árboles vs otras tecnologías



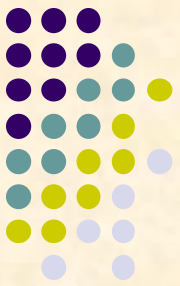
Some characteristics of different learning methods.

Key: ● = good, ● = fair, and ● = poor.

Characteristic	Neural Nets	SVM	CART	GAM	KNN, kernels	MART
Natural handling of data of "mixed" type	●	●	●	●	●	●
Handling of missing values	●	●	●	●	●	●
Robustness to outliers in input space	●	●	●	●	●	●
Insensitive to monotone transformations of inputs	●	●	●	●	●	●
Computational scalability (large N)	●	●	●	●	●	●
Ability to deal with irrelevant inputs	●	●	●	●	●	●
Ability to extract linear combinations of features	●	●	●	●	●	●
Interpretability	●	●	●	●	●	●
Predictive power	●	●	●	●	●	●

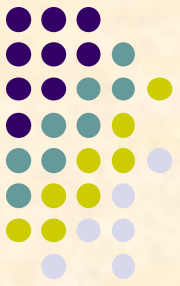
Fuente





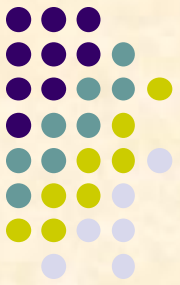
Resumen

- Los árboles de decisión son una herramienta no paramétrica de clasificación y regresión
- Proporcionan un conjunto de reglas de interpretación simple y que son fáciles de traducir en acciones de negocio
- Son eficaces para buscar perfiles o patrones en las bases de datos
- Sin embargo, son ineficaces para descubrir relaciones lineales en los datos
- Producen segmentaciones muy inestables
- Tienen poca capacidad predictiva en comparación con otras tecnologías como redes neuronales, GAM, RF, SVM, boosting....



- ¿Qué es un árbol de decisión?
- Segmentación con modelos de árbol
- El algoritmo CHAID
- El algoritmo CART
- Árboles vs otras tecnologías
- Casos prácticos

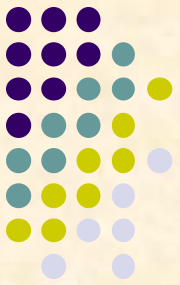
Credit Scoring



Y	Cuenta	Mes	Ppag	Uso	DM	Sexo	Estc
buen	no	18	pre buen pagador	privado	1049	mujer	vive solo
buen	no	9	pre buen pagador	profesional	2799	hombre	no vive solo
buen	bad running	12	pre buen pagador	profesional	841	mujer	vive solo
buen	no	12	pre buen pagador	profesional	2122	hombre	no vive solo
buen	no	12	pre buen pagador	profesional	2171	hombre	no vive solo
buen	no	10	pre buen pagador	profesional	2241	hombre	no vive solo
buen	no	8	pre buen pagador	profesional	3398	hombre	no vive solo
buen	no	6	pre buen pagador	profesional	1361	hombre	no vive solo
buen	good running	18	pre buen pagador	privado	1098	mujer	vive solo
buen	bad running	24	pre buen pagador	privado	3758	mujer	vive solo
buen	no	11	pre buen pagador	profesional	3905	hombre	no vive solo
buen	no	30	pre buen pagador	privado	6187	mujer	no vive solo
buen	no	6	pre buen pagador	privado	1957	mujer	vive solo
buen	bad running	48	pre buen pagador	profesional	7582	hombre	no vive solo
buen	no	18	pre buen pagador	privado	1936	mujer	no vive solo
buen	no	6	pre buen pagador	privado	2647	hombre	no vive solo
buen	no	11	pre buen pagador	profesional	3939	hombre	no vive solo
buen	bad running	18	pre buen pagador	privado	3213	mujer	no vive solo
buen	bad running	36	pre buen pagador	privado	2337	hombre	no vive solo

- Banco local del sur de Alemania
- **Objetivo:** Predecir la morosidad

Descripción de los datos



Datos: matriz con 1000 filas y 8 columnas

Filas: cada fila o registro corresponde a un cliente. Los datos contienen 700 no morosos y 300 morosos

Columnas: cada campo es una variable medida sobre todos los clientes

Partición: 2/3 (training) y 1/3 (test)

Y	Cuenta	Mes	Ppag	Uso	DM	Sexo	Estc
Variable criterio binaria (buen y mal)	Calidad de la cuenta (no, good running, bad running)	Duración en meses	Historial del cliente (pre buen pagador, pre mal pagador)	Uso que se hace del crédito (privado, profesional)	Cuantía del crédito	Sexo del cliente (hombre, mujer)	Estado civil (no vive solo, vive solo)

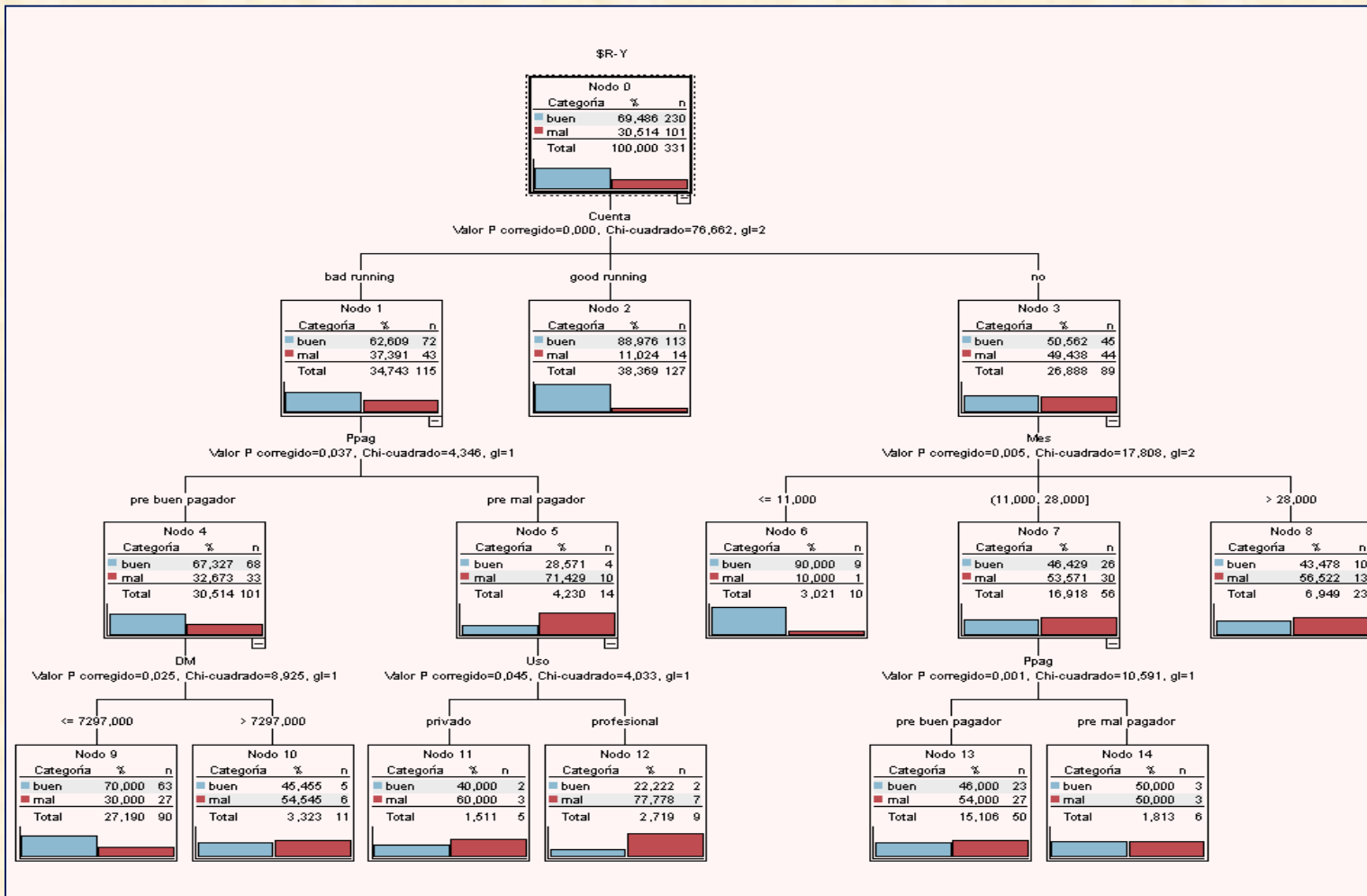
Fuente: <http://www.r-project.org> (librería Fahrmeir)

Fahrmeir Ludwig, Tutz G. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer. 2000

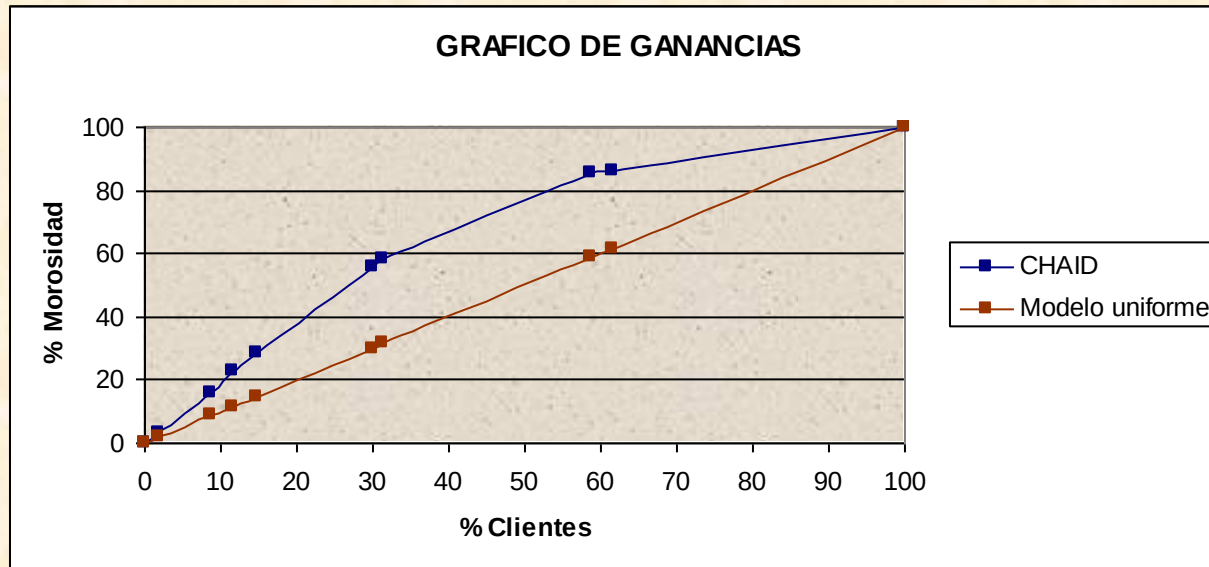
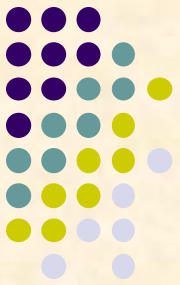
Credit scoring (CHAID)



$\alpha_{\text{merge}}=0.01$. $\alpha_{\text{split}}=0.05$. Parental = 20. Filial = 10. Profundidad = 3 niveles

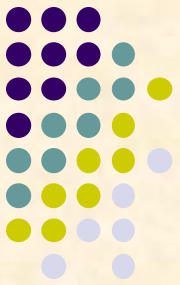


Ganancias CHAID

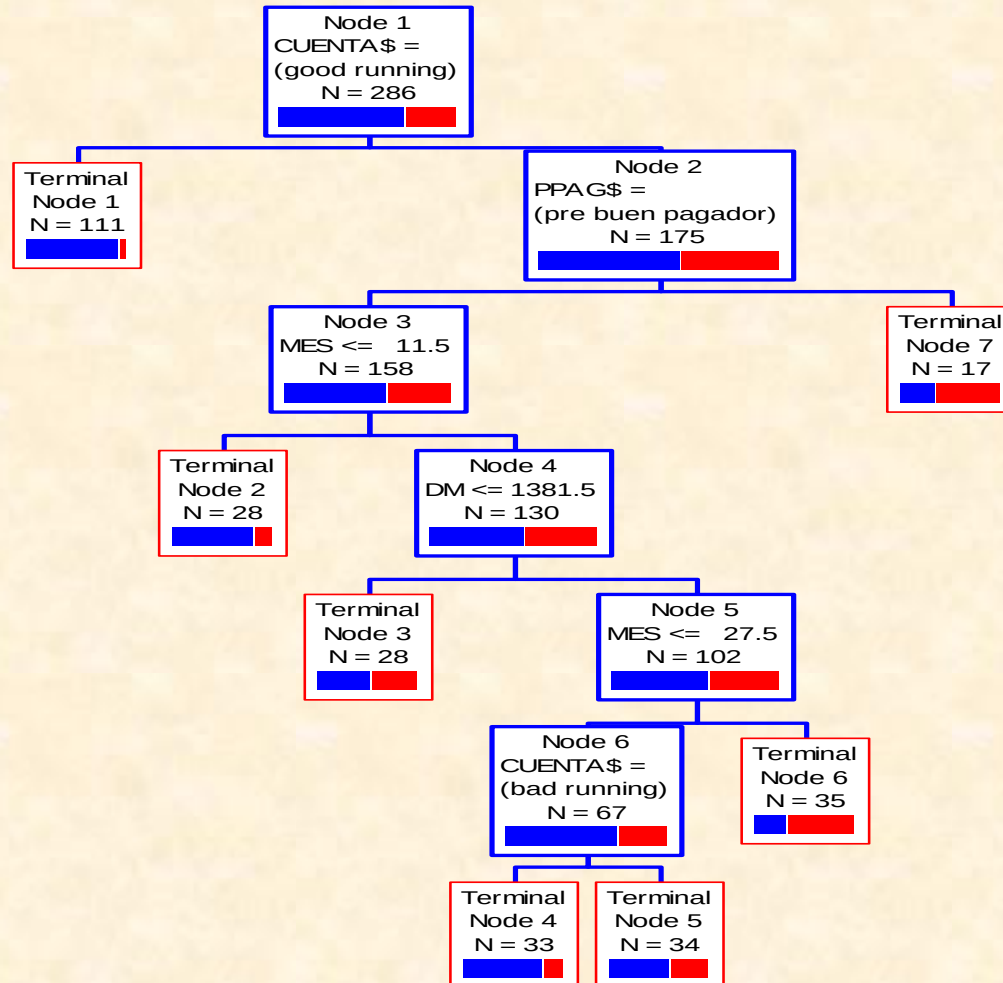


Nodos	Nodo: n	Nodo (%)	Ganancia: n	Ganancia (%)	Respuesta (%)	Indice (%)	Uniforme
0	0	0	0	0	0	0	0
14	6	1,81	3	2,97	50	163,86	1,81
8	29	8,76	16	15,84	55,17	180,81	8,76
12	38	11,48	23	22,77	60,53	198,36	11,48
10	49	14,8	29	28,71	59,18	193,96	14,8
13	99	29,91	56	55,45	56,57	185,38	29,91
11	104	31,42	59	58,42	56,73	185,92	31,42
9	194	58,61	86	85,15	44,33	145,28	58,61
6	204	61,63	87	86,14	42,65	139,76	61,63
2	331	100	101	100	30,51	100	100

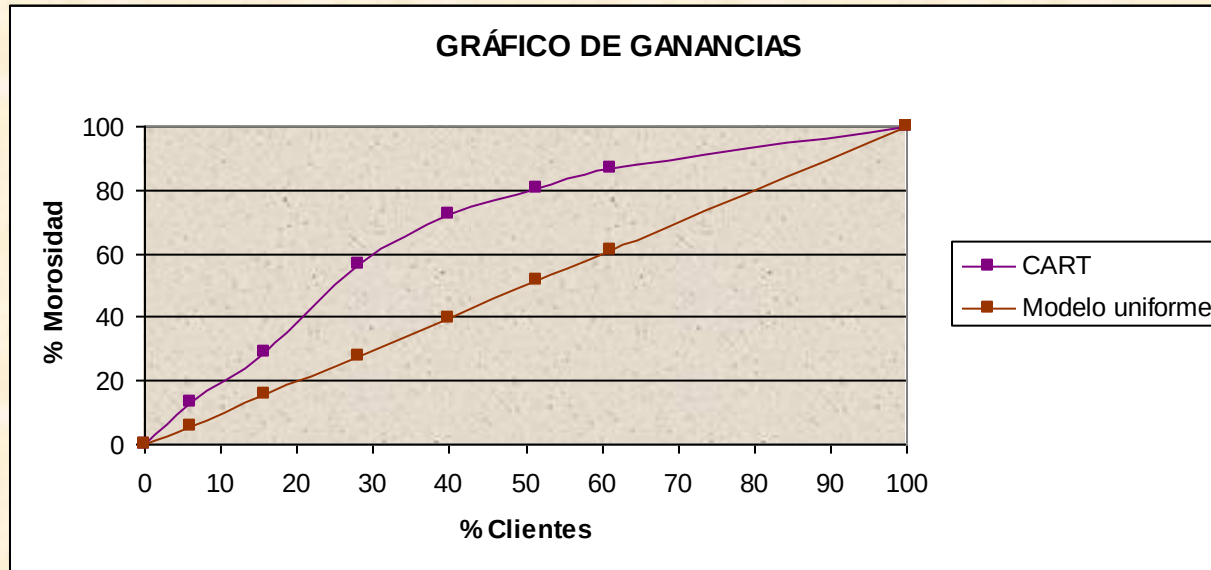
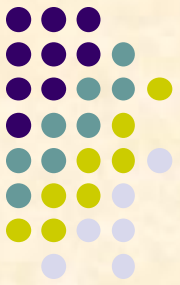
Credit scoring (CART)



Method = Entropy. Parental = 100. Filial = 50. Profundidad = AUTO. Priors = EQUAL

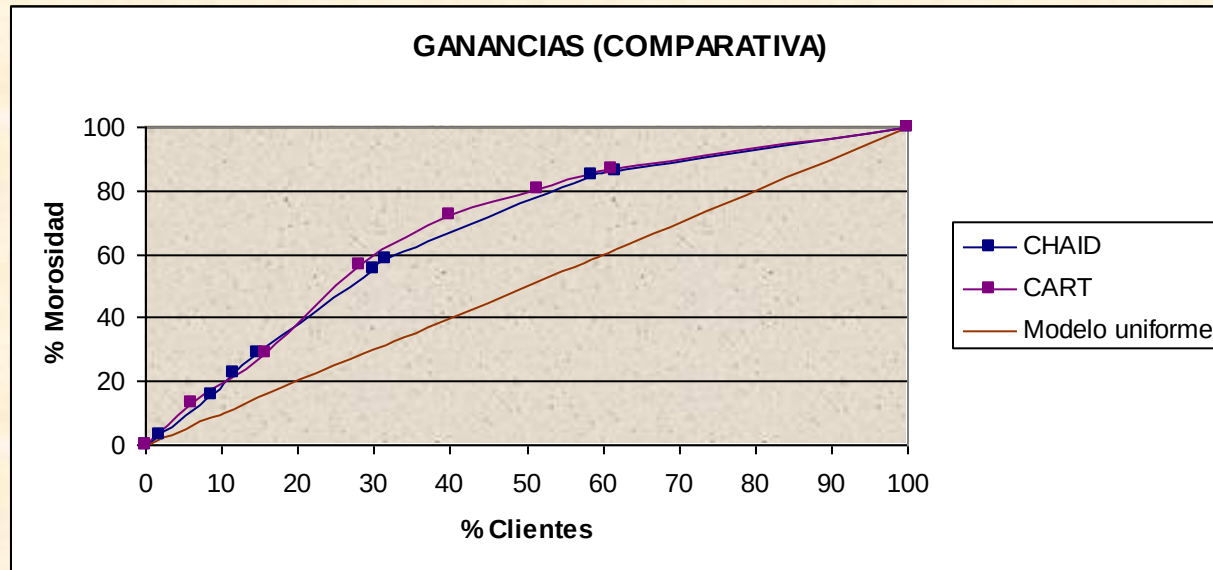
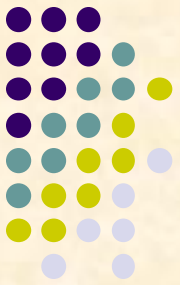


Ganancias CART



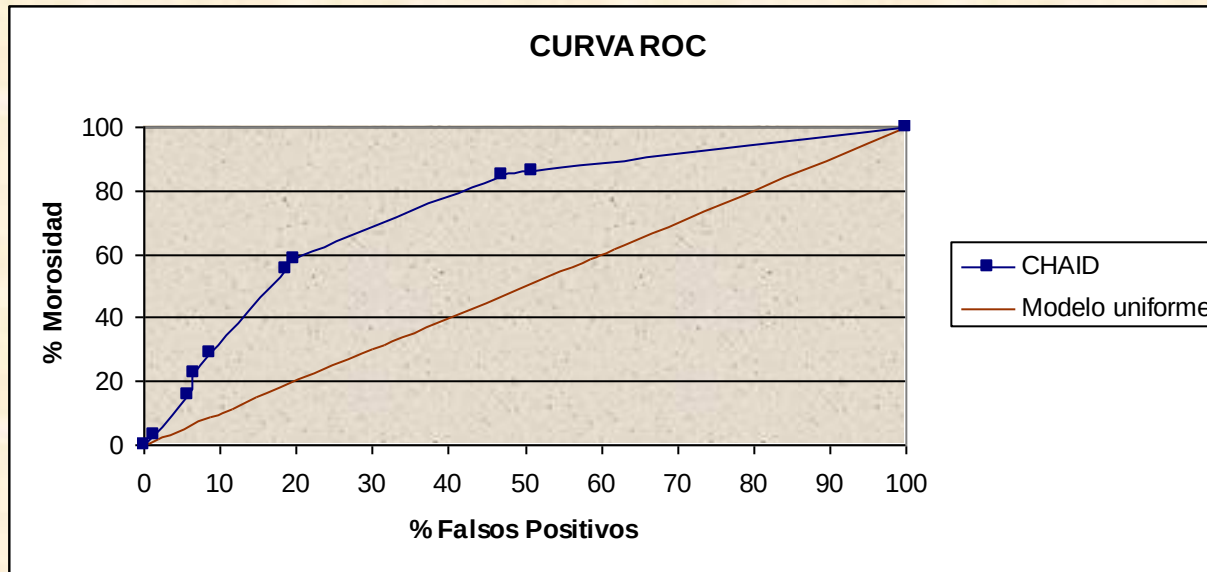
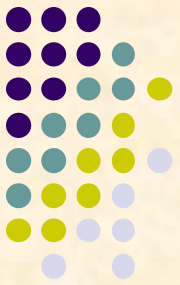
Node	Cases Tgt. Class	% of Node Tgt. Class	% Tgt. Class	Cum % Tgt. Class	Cum % Pop	% Pop	Cases in Node	Cum lift	Lift Pop	Uniforme
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
7	11,00	64,71	13,25	13,25	5,94	5,94	17,00	2,23	2,23	5,94
3	13,00	46,43	15,66	28,92	15,73	9,79	28,00	1,84	1,60	15,73
6	23,00	65,71	27,71	56,63	27,97	12,24	35,00	2,02	2,26	27,97
5	13,00	38,24	15,66	72,29	39,86	11,89	34,00	1,81	1,32	39,86
4	7,00	21,21	8,43	80,72	51,40	11,54	33,00	1,57	0,73	51,40
2	5,00	17,86	6,02	86,75	61,19	9,79	28,00	1,42	0,62	61,19
1	11,00	9,91	13,25	100,00	100,00	38,81	111,00	1,00	0,34	100,00

CHAID vs CART (ganancias)



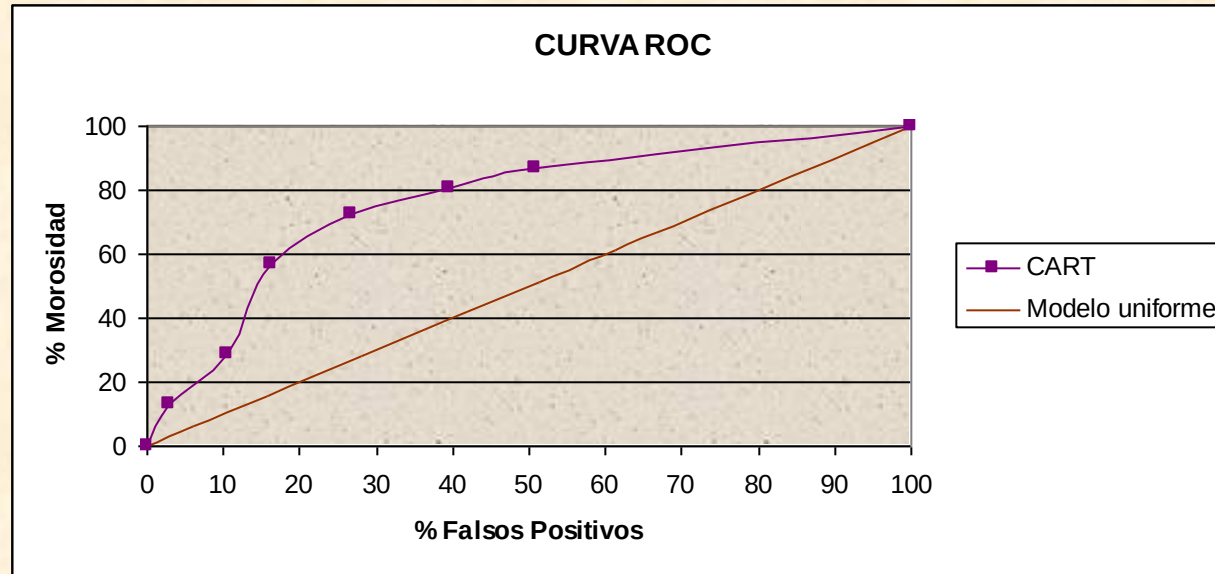
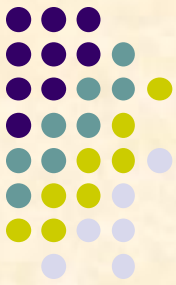
- Los nodos 14, 8, 12, 10 y 13 de CHAID (30% de clientes) contienen el 55% de la morosidad. Lift = 185%
- El 30% de casos con alto riesgo de morosidad están en los nodos 7, 3, 6 y parte del 5 del árbol CART; contienen un 59% de la morosidad. Lift = 197%

ROC CHAID



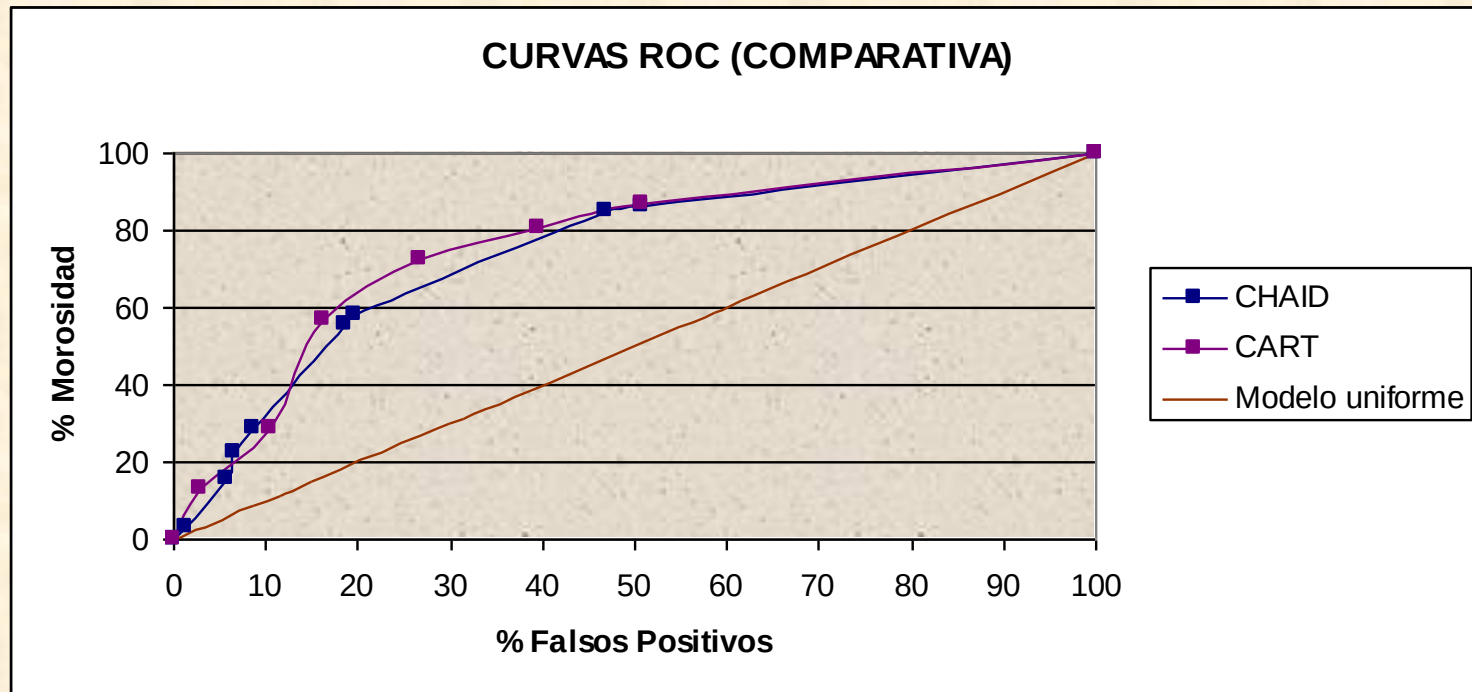
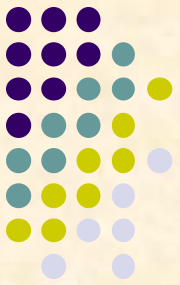
Nodos	Nodo: n	Nodo (%)	Ganancia: n	Ganancia (%)	Respuesta (%)	Indice (%)	Uniforme	FP	FP(%)
0	0	0	0	0	0	0	0	0	0
14	6	1,81	3	2,97	50	163,86	1,81	3	1,30
8	29	8,76	16	15,84	55,17	180,81	8,76	13	5,65
12	38	11,48	23	22,77	60,53	198,36	11,48	15	6,52
10	49	14,8	29	28,71	59,18	193,96	14,8	20	8,70
13	99	29,91	56	55,45	56,57	185,38	29,91	43	18,70
11	104	31,42	59	58,42	56,73	185,92	31,42	45	19,57
9	194	58,61	86	85,15	44,33	145,28	58,61	108	46,96
6	204	61,63	87	86,14	42,65	139,76	61,63	117	50,87
2	331	100	101	100	30,51	100	100	230	100,00

ROC CART

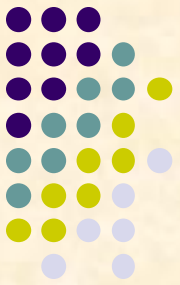


Node	Cases Tgt. Class	% of Node Tgt. Class	% Tgt. Class	Cum % Tgt. Class	Cum % Pop	% Pop	Cases in Node	Cum lift	Lift Pop	Uniforme	FP	Cum FP	Cum % FP
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
7	11,00	64,71	13,25	13,25	5,94	5,94	17,00	2,23	2,23	5,94	6,00	6,00	2,96
3	13,00	46,43	15,66	28,92	15,73	9,79	28,00	1,84	1,60	15,73	15,00	21,00	10,34
6	23,00	65,71	27,71	56,63	27,97	12,24	35,00	2,02	2,26	27,97	12,00	33,00	16,26
5	13,00	38,24	15,66	72,29	39,86	11,89	34,00	1,81	1,32	39,86	21,00	54,00	26,60
4	7,00	21,21	8,43	80,72	51,40	11,54	33,00	1,57	0,73	51,40	26,00	80,00	39,41
2	5,00	17,86	6,02	86,75	61,19	9,79	28,00	1,42	0,62	61,19	23,00	103,00	50,74
1	11,00	9,91	13,25	100,00	100,00	38,81	111,00	1,00	0,34	100,00	100,00	203,00	100,00

CHAID vs CART (curvas ROC)



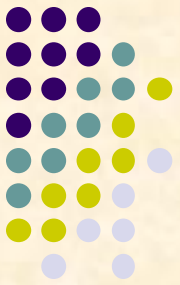
Abandono de clientes (churn)



State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn
KS	128	415	382-4657	no	yes	25	265,10	110	45,07	197,40	99	16,78	244,70	91	11,01	10,00	3	2,70	1	0
OH	107	415	371-7191	no	yes	26	161,60	123	27,47	195,50	103	16,62	254,40	103	11,45	13,70	3	3,70	1	0
NJ	137	415	358-1921	no	no	0	243,40	114	41,38	121,20	110	10,30	162,60	104	7,32	12,20	5	3,29	0	0
OH	84	408	375-9999	yes	no	0	299,40	71	50,90	61,90	88	5,26	196,90	89	8,86	6,60	7	1,78	2	0
OK	75	415	330-6626	yes	no	0	166,70	113	28,34	148,30	122	12,61	186,90	121	8,41	10,10	3	2,73	3	0
AL	118	510	391-8027	yes	no	0	223,40	98	37,98	220,60	101	18,75	203,90	118	9,18	6,30	6	1,70	0	0
MA	121	510	355-9993	no	yes	24	218,20	88	37,09	348,50	108	29,62	212,60	118	9,57	7,50	7	2,03	3	0
MO	147	415	329-9001	yes	no	0	157,00	79	26,69	103,10	94	8,76	211,80	96	9,53	7,10	6	1,92	0	0
LA	117	408	335-4719	no	no	0	184,50	97	31,37	351,60	80	29,89	215,80	90	9,71	8,70	4	2,35	1	0
WV	141	415	330-8173	yes	yes	37	258,60	84	43,96	222,00	111	18,87	326,40	97	14,69	11,20	5	3,02	0	0
IN	65	415	329-6603	no	no	0	129,10	137	21,95	228,50	83	19,42	208,80	111	9,40	12,70	6	3,43	4	1
RI	74	415	344-9403	no	no	0	187,70	127	31,91	163,40	148	13,89	196,00	94	8,82	9,10	5	2,46	0	0
IA	168	408	363-1107	no	no	0	128,80	96	21,90	104,90	71	8,92	141,10	128	6,35	11,20	2	3,02	1	0
MT	95	510	394-8006	no	no	0	156,60	88	26,62	247,60	75	21,05	192,30	115	8,65	12,30	5	3,32	3	0
IA	62	415	366-9238	no	no	0	120,70	70	20,52	307,20	76	26,11	203,00	99	9,14	13,10	6	3,54	4	0
NY	161	415	351-7269	no	no	0	332,90	67	56,59	317,80	97	27,01	160,60	128	7,23	5,40	9	1,46	4	1
ID	85	408	350-8884	no	yes	27	196,40	139	33,39	280,90	90	23,88	89,30	75	4,02	13,80	4	3,73	1	0
VT	93	510	386-2923	no	no	0	190,70	114	32,42	218,20	111	18,55	129,60	121	5,83	8,10	3	2,19	3	0
VA	76	510	356-2992	no	yes	33	189,70	66	32,25	212,80	65	18,09	165,70	108	7,46	10,00	5	2,70	1	0
TX	73	415	373-2782	no	no	0	224,40	90	38,15	159,50	88	13,56	192,80	74	8,68	13,00	2	3,51	1	0
FL	147	415	396-5800	no	no	0	155,10	117	26,37	239,70	93	20,37	208,80	133	9,40	10,60	4	2,86	0	0
CO	77	408	393-7984	no	no	0	62,40	89	10,61	169,90	121	14,44	209,60	64	9,43	5,70	6	1,54	5	1
AZ	130	415	358-1958	no	no	0	183,00	112	31,11	72,90	99	6,20	181,80	78	8,18	9,50	19	2,57	0	0
SC	111	415	350-2565	no	no	0	110,40	103	18,77	137,30	102	11,67	189,60	105	8,53	7,70	6	2,08	2	0
VA	132	510	343-4696	no	no	0	81,10	86	13,79	245,20	72	20,84	237,00	115	10,67	10,30	2	2,78	0	0



- Compañía de Telecomunicaciones (EEUU)
- **Objetivo:** Modelizar el churn



Descripción de los datos

Datos: matriz con 3333 filas y 21 columnas

Filas: cada fila es un cliente. 483 abandonaron (14.5%) y 2850 no lo hicieron (85.5%)

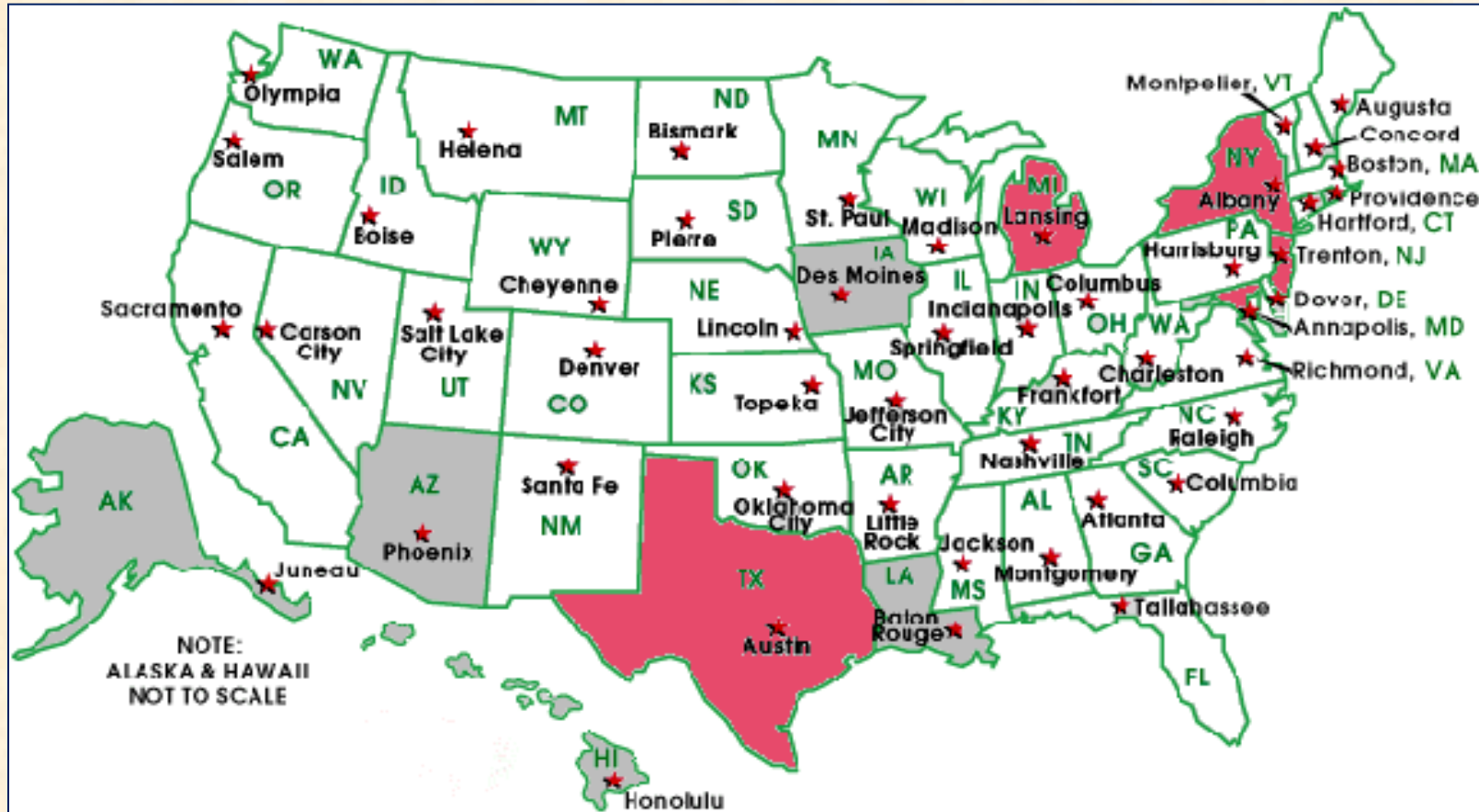
Columnas: cada campo es una variable. La variable criterio es churn. Churn = 1 si hubo abandono y churn = 0 si no lo hubo

Variables explicativas:

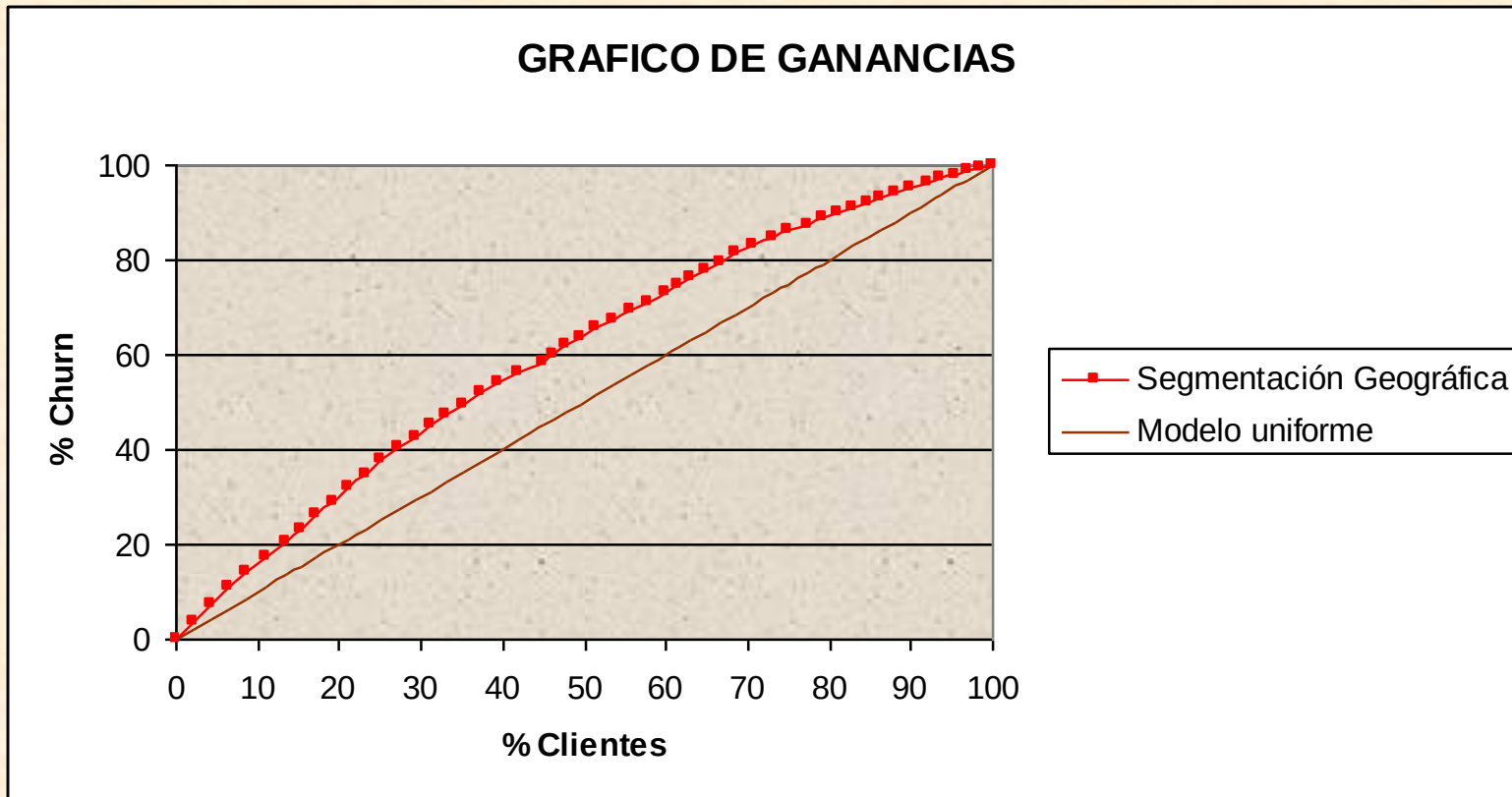
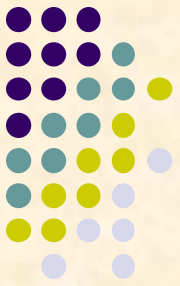
State, Account length, Area code, Phone number, International plan, Voice Mail plan, Number of voicemail messages, Total day minutes, Total day calls, Total day charge, Total evening minutes, Total evening calls, Total evening charge, Total night minutes, Total night calls, Total night charge, Total international minutes, Total international calls, Total international charge, Number of calls to customer service.

Fuente: <http://dataminingconsultant.com>

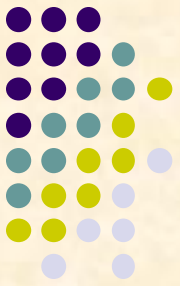
Churn. Segmentación geográfica



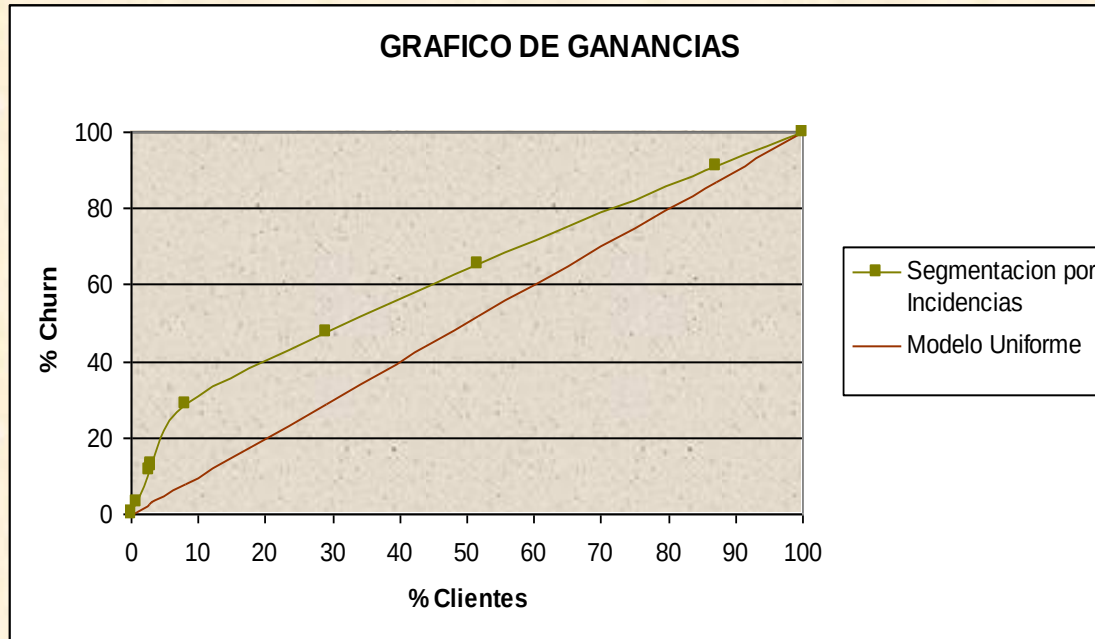
Ganancias. Segmentación geográfica



Churn. Segmentación por incidencias

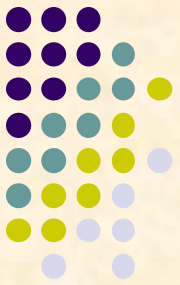


- Las incidencias guardan una alta correlación con el número de llamadas al servicio de atención al cliente

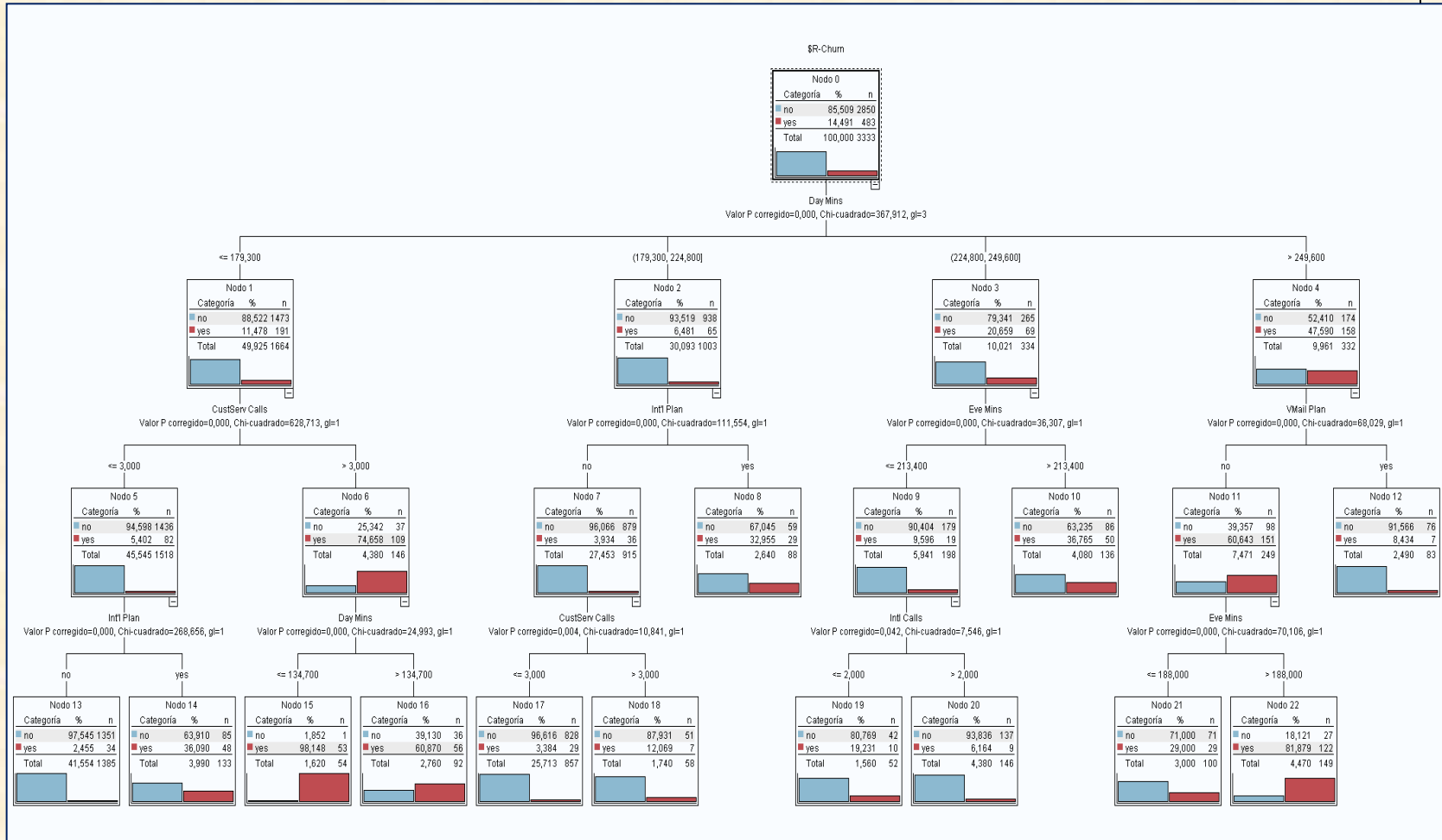


CustServ Calls	no	yes	Total general
9	0	2	2
6	8	14	22
5	26	40	66
7	4	5	9
8	1	1	2
4	90	76	166
0	605	92	697
2	672	87	759
1	1059	122	1181
3	385	44	429
Total general	2850	483	3333

Churn. Segmentación CHAID

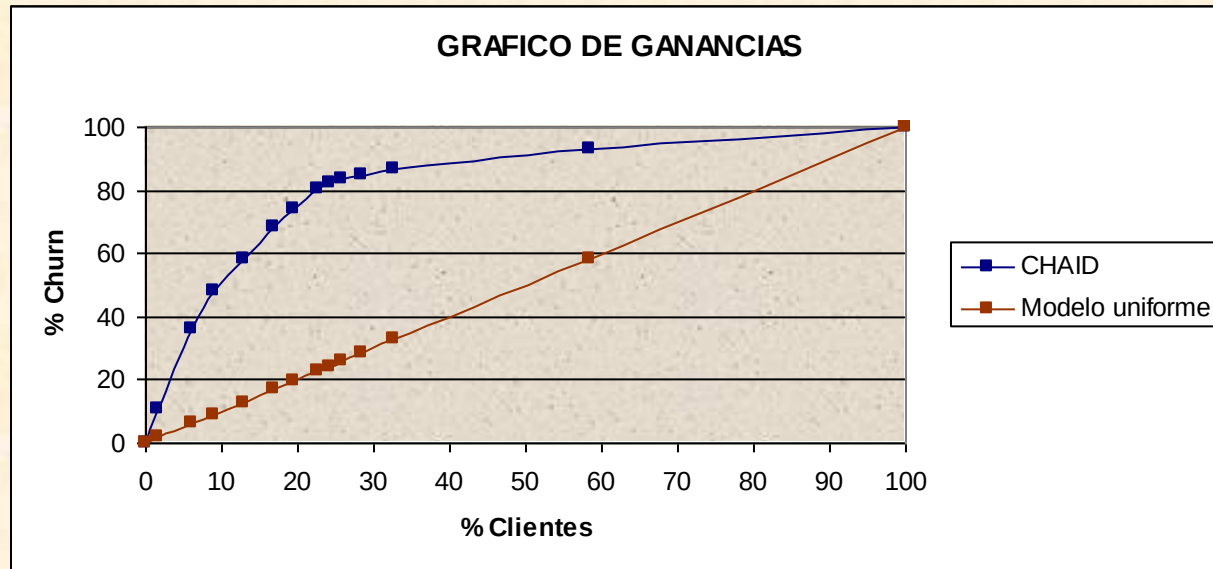
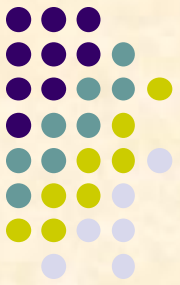


$\alpha_{\text{merge}}=0.001$. $\alpha_{\text{split}}=0.05$. Parental = 100. Filial = 50. Profundidad = 3 niveles



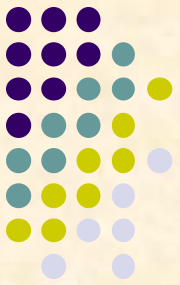
Variables explicativas: State. Account length. International plan. VoiceMail plan. Number of voicemail messages. Total day minutes. Total day calls. Total evening minutes. Total evening calls. Total night minutes. Total night calls. Total international minutes. Total international calls. Number of calls to customer service.

Ganancias CHAID

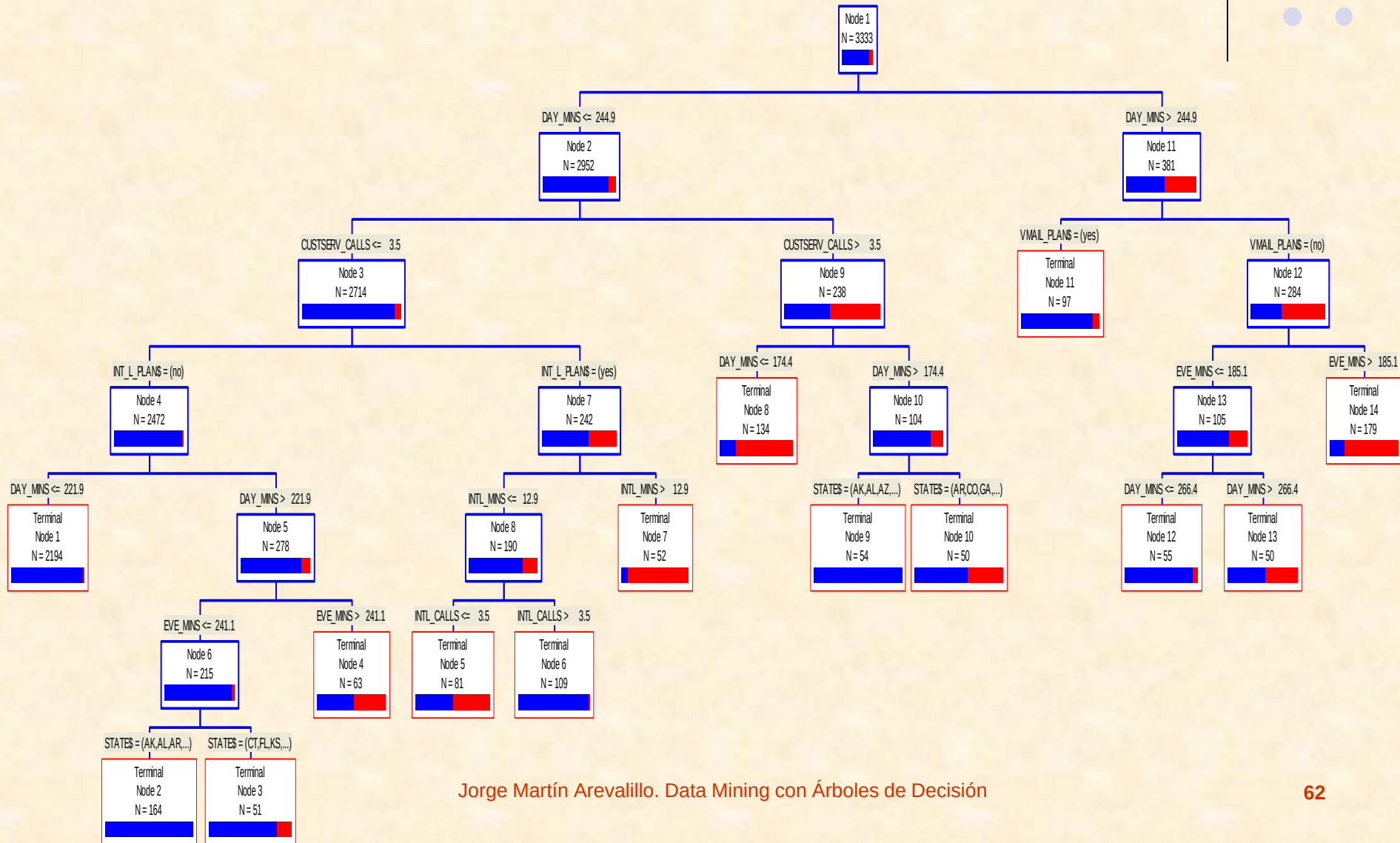


Nodos	Nodo: n	Nodo (%)	Ganancia: n	Ganancia (%)	Respuesta (%)	Indice (%)	Uniforme
0	0	0	0	0	0	0	0
15	54	1,62	53	10,97	98,15	677,28	1,62
22	203	6,09	175	36,23	86,21	594,88	6,09
16	295	8,85	231	47,83	78,31	540,35	8,85
10	431	12,93	281	58,18	65,2	449,9	12,93
14	564	16,92	329	68,12	58,33	402,54	16,92
8	652	19,56	358	74,12	54,91	378,9	19,56
21	752	22,56	387	80,12	51,46	355,13	22,56
19	804	24,12	397	82,19	49,38	340,74	24,12
18	862	25,86	404	83,64	46,87	323,42	25,86
12	945	28,35	411	85,09	43,49	300,12	28,35
20	1091	32,73	420	86,96	38,5	265,65	32,73
17	1948	58,45	449	92,96	23,05	159,05	58,45
13	3333	100	483	100	14,49	100	100

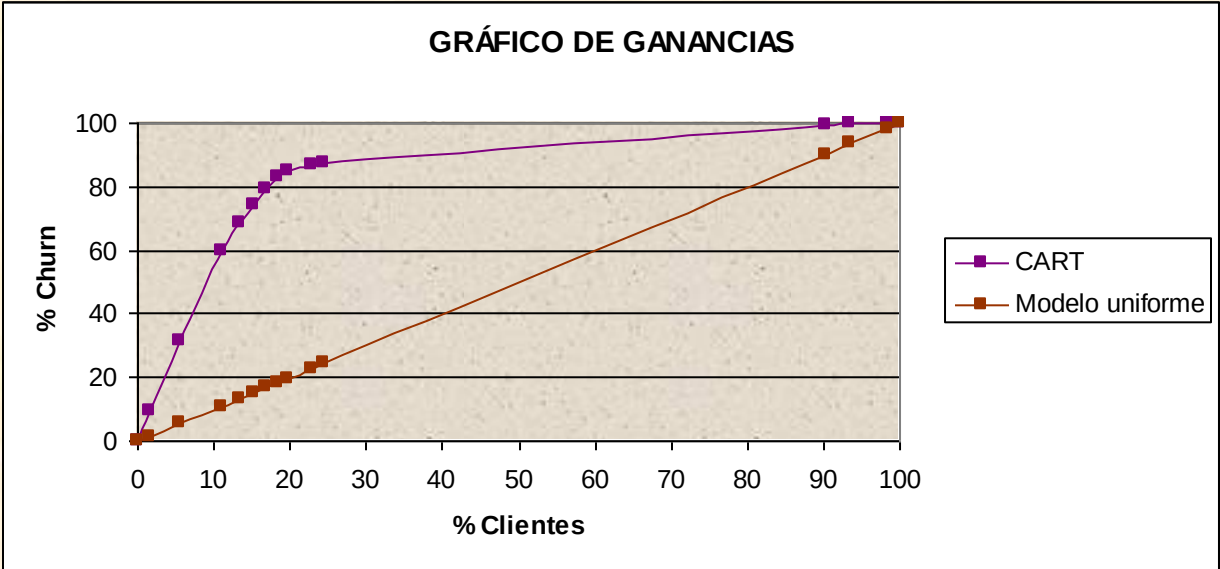
Churn. Segmentación CART



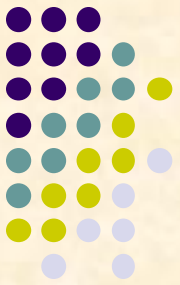
Method = Gini. Parental=100. Filial=50. Profundidad = 10. Priors = EQUAL



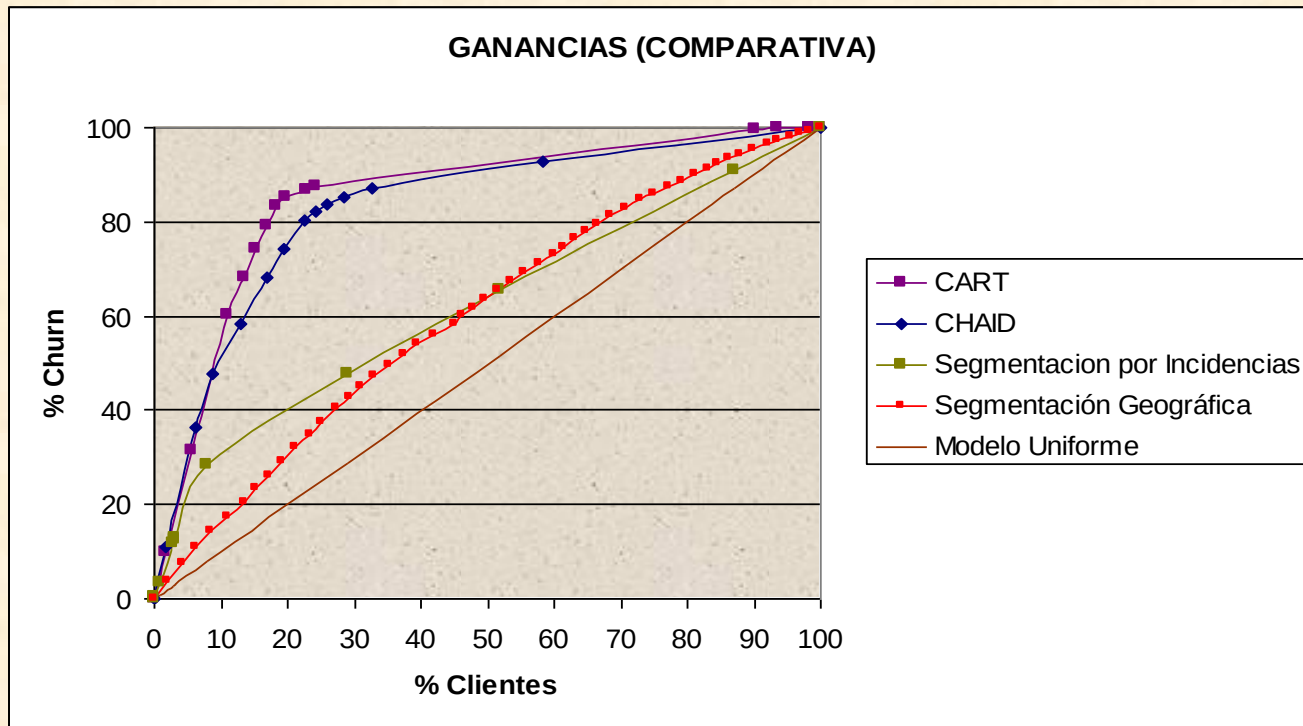
Ganancias CART



Node	Cases Tgt. Class	% of Node Tgt. Class	% Tgt. Class	Cum % Tgt. Class	Cum % Pop	% Pop	Cases in Node	Cum lift	Lift Pop	Uniforme
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
7	47,00	90,38	9,73	9,73	1,56	1,56	52,00	6,24	6,24	1,56
8	105,00	78,36	21,74	31,47	5,58	4,02	134,00	5,64	5,41	5,58
14	138,00	77,09	28,57	60,04	10,95	5,37	179,00	5,48	5,32	10,95
5	40,00	49,38	8,28	68,32	13,38	2,43	81,00	5,11	3,41	13,38
4	29,00	46,03	6,00	74,33	15,27	1,89	63,00	4,87	3,18	15,27
13	23,00	46,00	4,76	79,09	16,77	1,50	50,00	4,72	3,17	16,77
10	20,00	40,00	4,14	83,23	18,27	1,50	50,00	4,56	2,76	18,27
3	9,00	17,65	1,86	85,09	19,80	1,53	51,00	4,30	1,22	19,80
11	8,00	8,25	1,66	86,75	22,71	2,91	97,00	3,82	0,57	22,71
12	4,00	7,27	0,83	87,58	24,36	1,65	55,00	3,59	0,50	24,36
1	58,00	2,64	12,01	99,59	90,19	65,83	2194,00	1,10	0,18	90,19
6	2,00	1,83	0,41	100,00	93,46	3,27	109,00	1,07	0,13	93,46
2	0,00	0,00	0,00	100,00	98,38	4,92	164,00	1,02	0,00	98,38
9	0,00	0,00	0,00	100,00	100,00	1,62	54,00	1,00	0,00	100,00



Ganancias (comparativa)



- Los nodos CHAID 15, 22, 16, 10, 14, 8, 21 son el 22.5% de los clientes. Contienen aproximadamente el 80% del churn
- Con el mismo porcentaje de clientes, los segmentos CART 7, 8, 14, 5, 4, 13, 10, 3 y 11 contienen casi el 87% del churn

Detección de spam



Objetivo: detección de correo basura. Compañía hp

Datos: una matriz con 4601 registros y 58 columnas

Filas: cada fila o registro corresponde a un e-mail.

Los datos contienen 1813 correos clasificados como *spam* (1) y 2788 clasificados como *nonspam* (0)

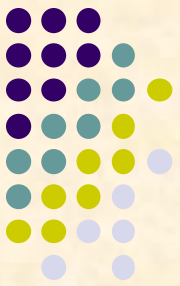
Columnas: cada campo es una variable



VARIABLES 1-48	VARIABLES 49-54	VARIABLE 55	VARIABLE 56	VARIABLE 57	VARIABLE 58
Son palabras. Miden la frecuencia con que aparece cada palabra. El prefijo <i>num</i> es identificador de número	Frecuencia con que aparecen los caracteres: ; ([! \$ y #	CapitalAve: Longitud media de las secuencias de letras mayúsculas	CapitalLong: Longitud de la secuencia de letras mayúsculas más larga	CapitalTotal: Número total de letras mayúsculas en el e-mail	type: Etiqueta de clase (1 para spam y 0 para no spam)

Fuente: <http://www.r-project.org> (librería kernlab)

Detección de spam (cont.)



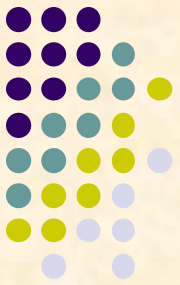
Variables explicativas:

make	address	all	num3d	our	over	remove
internet	order	mail	receive	will	people	report
addresses	free	business	email	you	credit	your
font	num000	money	hp	hpl	george	num650
lab	labs	telnet	num857	data	num415	num85
technology	num1999	parts	pm	direct	cs	meeting
original	project	re	edu	table	conference	charSemicolon
charRoundbracket	charSquarebracket	charExclamation	charDollar	charHash	capitalAve	
capitalLong	capitalTotal					

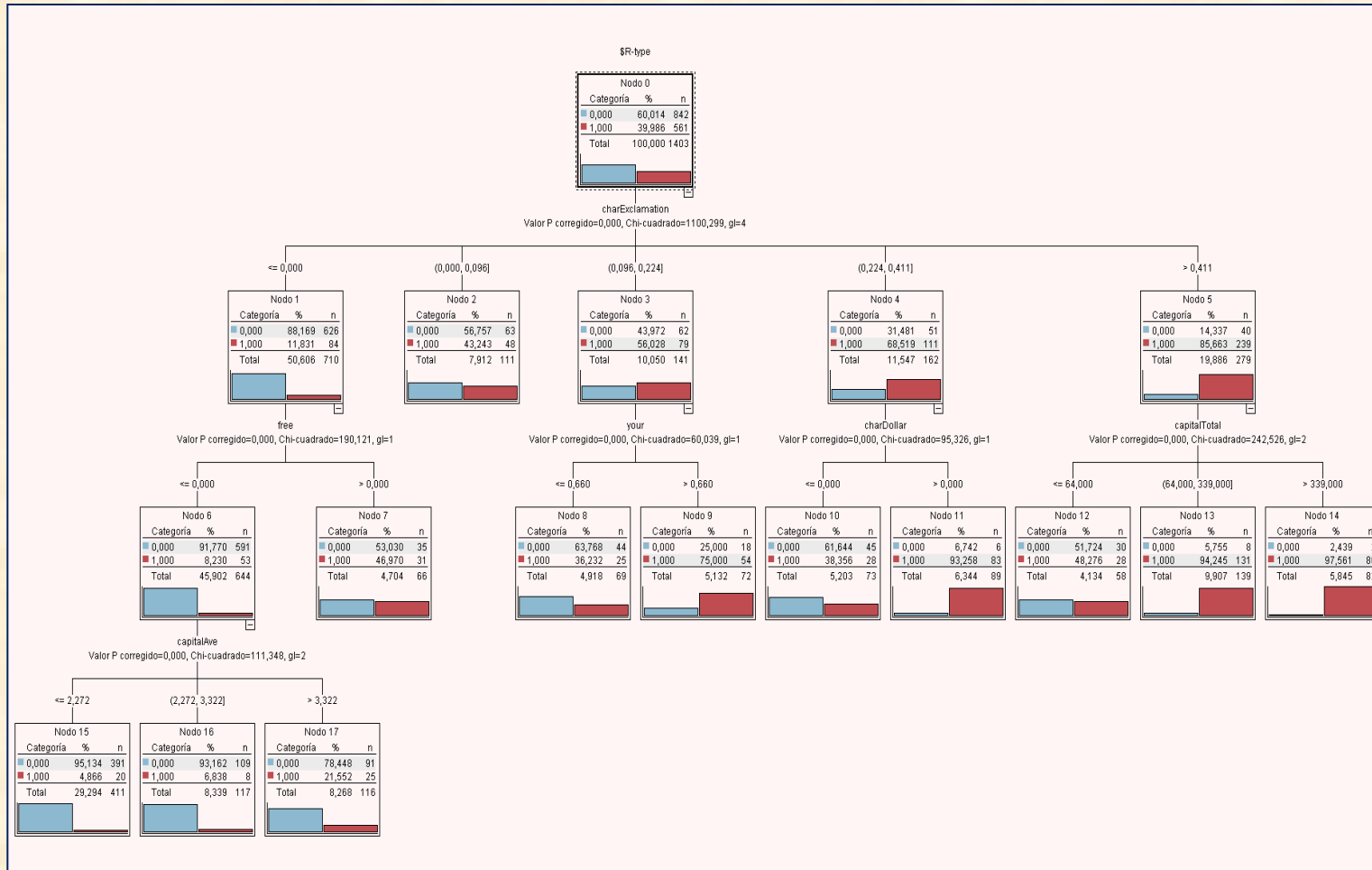
Propuesta: utilizar un árbol de decisión para segmentar la base de datos. Encontrar un modelo de árbol que permita identificar el correo basura. Buscar perfiles de e-mails en la base de datos

Partición: 70% (training). 30% (test)

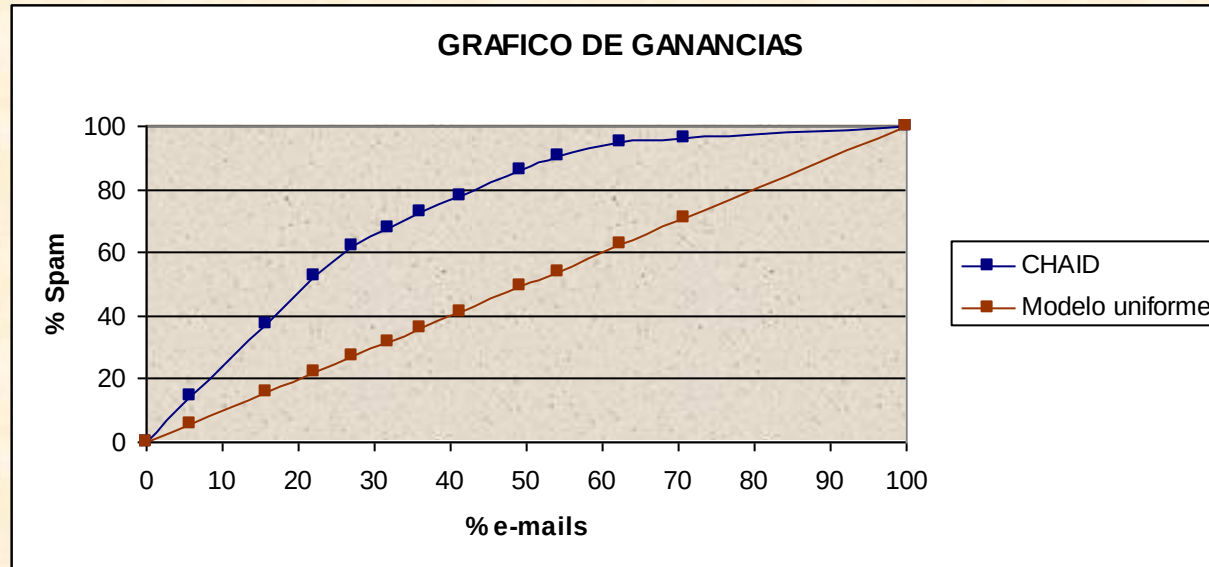
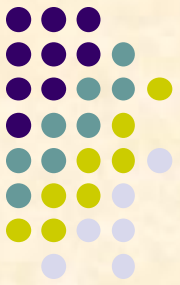
Spam. Segmentación CHAID



$\alpha_{\text{merge}}=0.001$. $\alpha_{\text{split}}=0.05$. Parental = 300. Filial = 150. Profundidad = 3 niveles

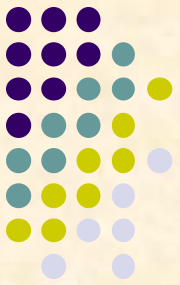


Ganancias CHAID

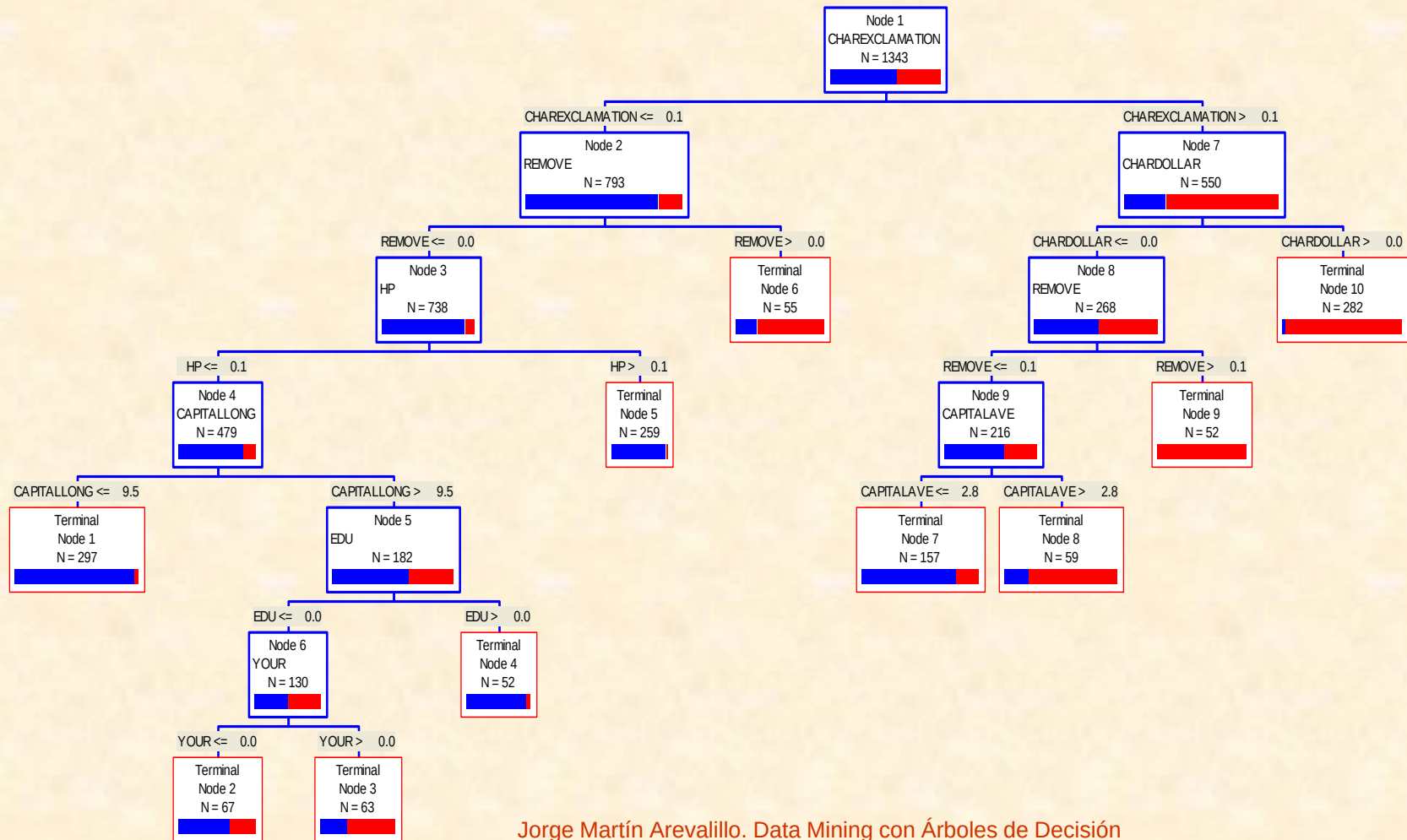


Nodos	Nodo: n	Nodo (%)	Ganancia: n	Ganancia (%)	Respuesta (%)	Indice (%)	Uniforme
0	0	0	0	0	0	0	0
14	82	5,84	80	14,26	97,56	243,99	5,84
13	221	15,75	211	37,61	95,48	238,77	15,75
11	310	22,1	294	52,41	94,84	237,18	22,1
9	382	27,23	348	62,03	91,1	227,83	27,23
7	448	31,93	379	67,56	84,6	211,57	31,93
12	506	36,07	407	72,55	80,43	201,16	36,07
10	579	41,27	435	77,54	75,13	187,89	41,27
2	690	49,18	483	86,1	70	175,06	49,18
8	759	54,1	508	90,55	66,93	167,39	54,1
17	875	62,37	533	95,01	60,91	152,34	62,37
16	992	70,71	541	96,43	54,54	136,39	70,71
15	1403	100	561	100	39,99	100	100

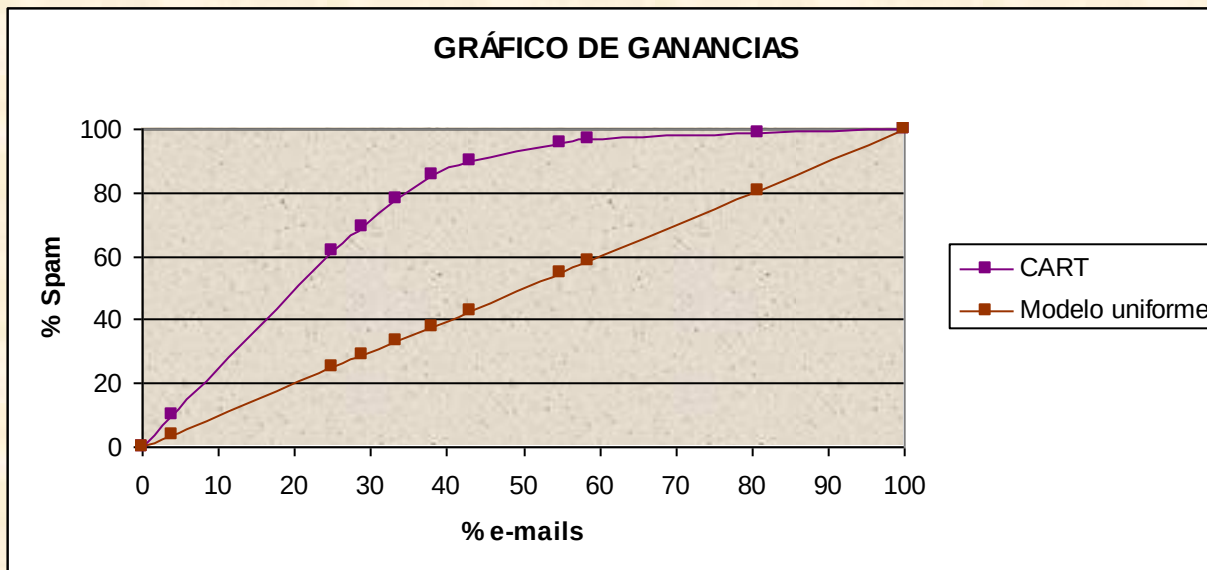
Spam. Segmentación CART



Method = Entropy. Parental = 200. Filial = 100. Profundidad = AUTO. Priors = EQUAL

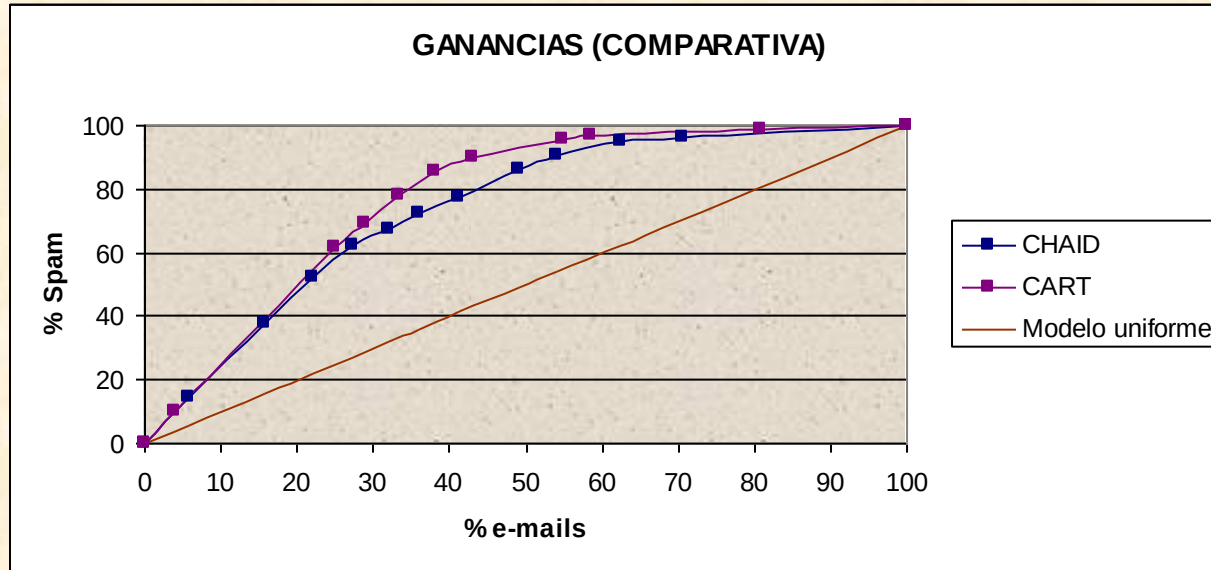
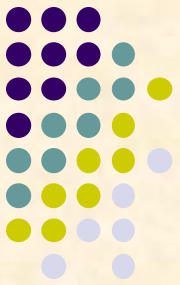


Ganancias CART



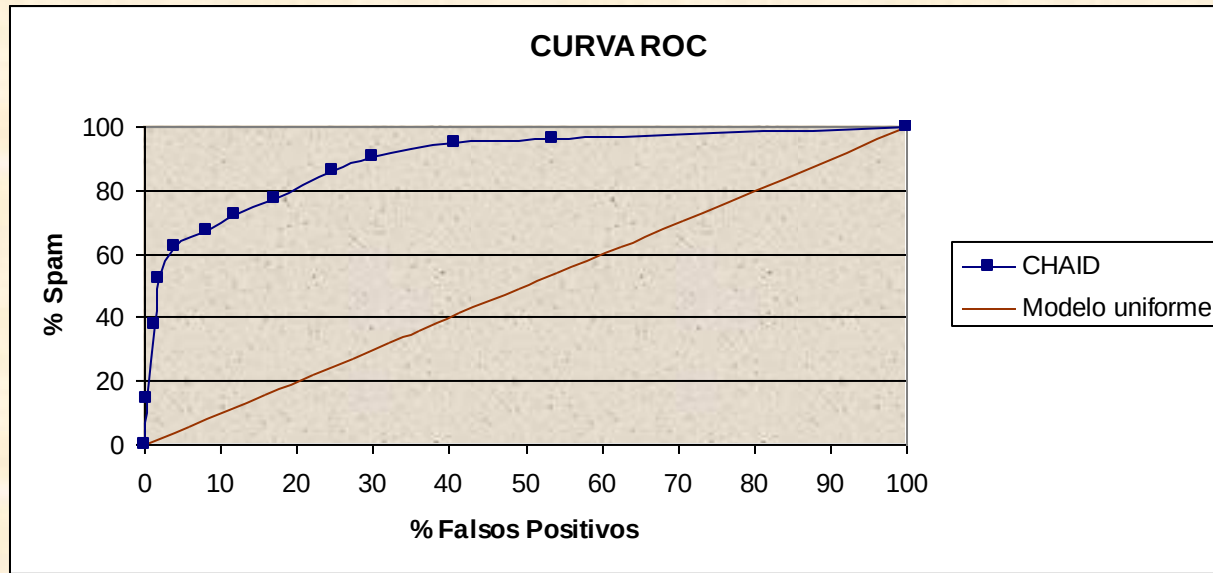
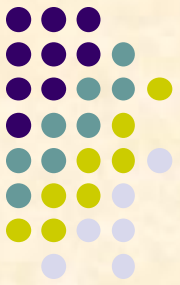
Node	Cases Tgt. Class	% of Node Tgt. Class	% Tgt. Class	Cum % Tgt. Class	Cum % Pop	% Pop	Cases in Node	Cum lift	Lift Pop	Uniforme
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
9	52	100,00	9,85	9,85	3,87	3,87	52	2,54	2,54	3,87
10	273	96,81	51,70	61,55	24,87	21,00	282	2,48	2,46	24,87
6	41	74,55	7,77	69,32	28,97	4,10	55	2,39	1,90	28,97
8	46	77,97	8,71	78,03	33,36	4,39	59	2,34	1,98	33,36
3	40	63,49	7,58	85,61	38,05	4,69	63	2,25	1,61	38,05
2	23	34,33	4,36	89,96	43,04	4,99	67	2,09	0,87	43,04
7	31	19,75	5,87	95,83	54,73	11,69	157	1,75	0,50	54,73
4	4	7,69	0,76	96,59	58,60	3,87	52	1,65	0,20	58,60
1	13	4,38	2,46	99,05	80,71	22,11	297	1,23	0,11	80,71
5	5	1,93	0,95	100,00	100,00	19,29	259	1,00	0,05	100,00

CHAID vs CART (ganancias)



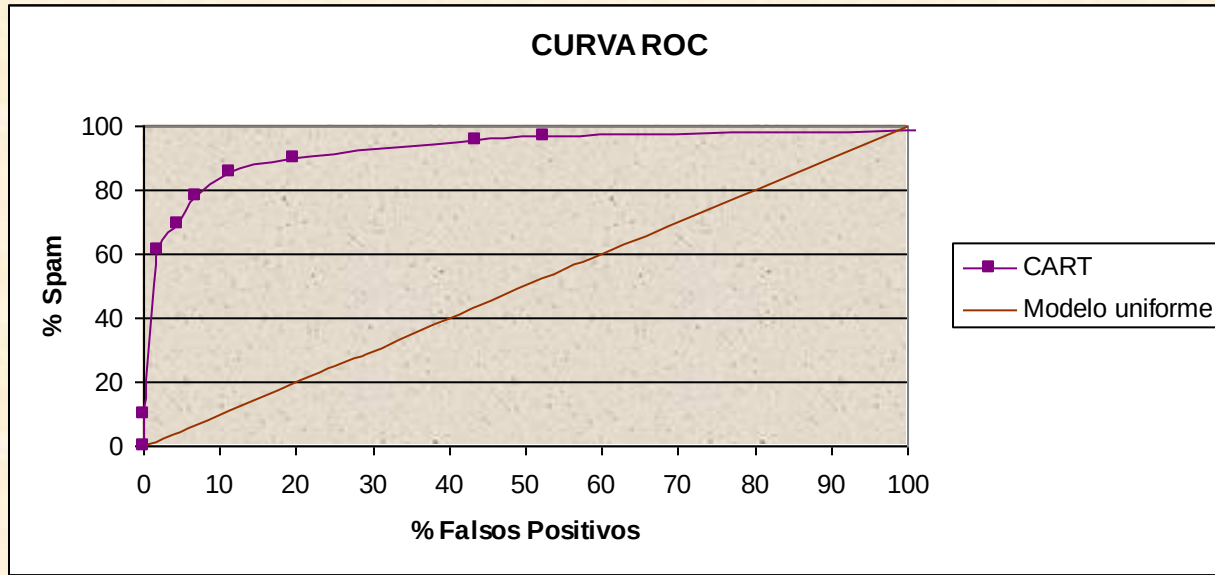
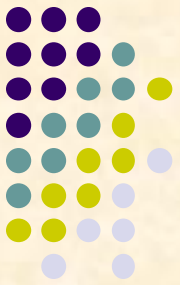
- Los nodos 9, 10, 6, 8, 3, 2 de CART (43% de los e-mails) contienen el 90% del spam. Lift = 209%
- Para capturar el 90% del spam con CHAID, se requiere el 54% de los e-mails (nodos 14, 13, 11, 9, 7, 12, 10, 2 y 8). En este caso Lift = 167.4%

ROC CHAID



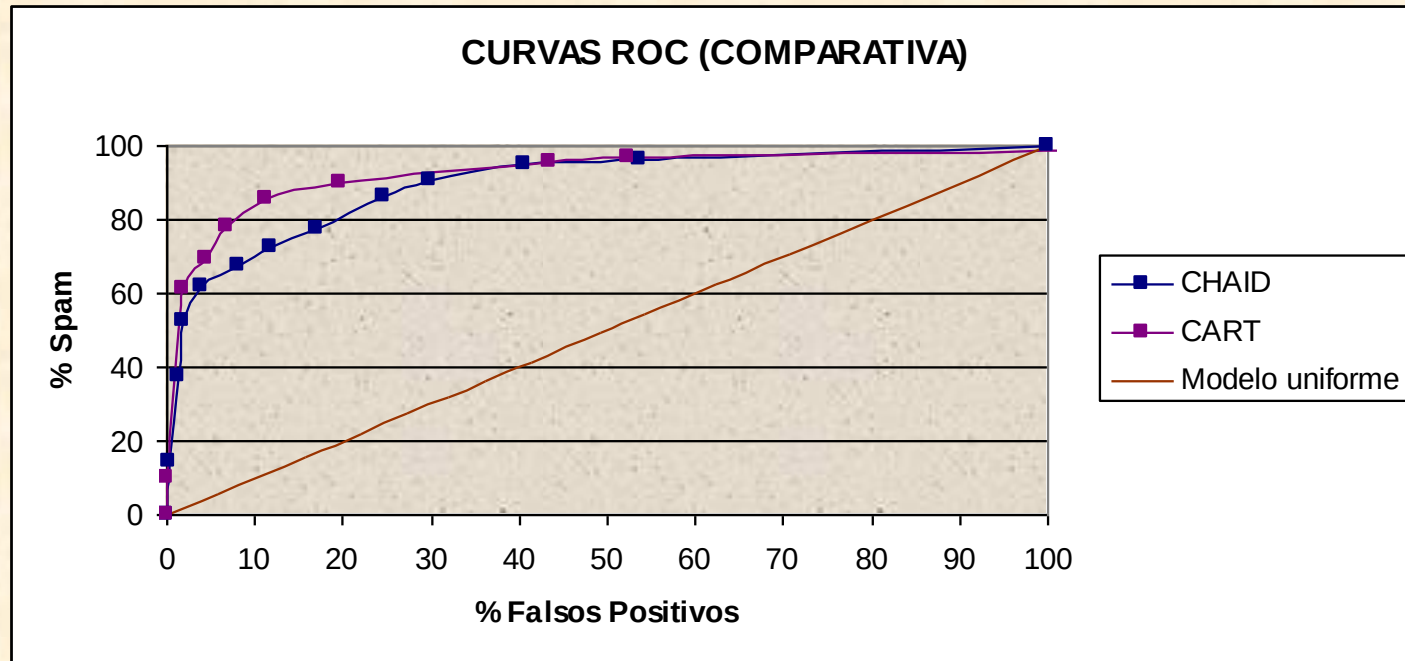
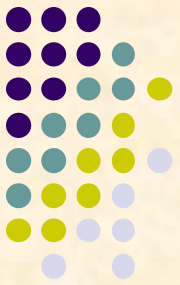
Nodos	Nodo: n	Nodo (%)	Ganancia: n	Ganancia (%)	Respuesta (%)	Indice (%)	Uniforme	FP	FP(%)
0	0	0	0	0	0	0	0	0	0
14	82	5,84	80	14,26	97,56	243,99	5,84	2	0,24
13	221	15,75	211	37,61	95,48	238,77	15,75	10	1,19
11	310	22,1	294	52,41	94,84	237,18	22,1	16	1,90
9	382	27,23	348	62,03	91,1	227,83	27,23	34	4,04
7	448	31,93	379	67,56	84,6	211,57	31,93	69	8,19
12	506	36,07	407	72,55	80,43	201,16	36,07	99	11,76
10	579	41,27	435	77,54	75,13	187,89	41,27	144	17,10
2	690	49,18	483	86,1	70	175,06	49,18	207	24,58
8	759	54,1	508	90,55	66,93	167,39	54,1	251	29,81
17	875	62,37	533	95,01	60,91	152,34	62,37	342	40,62
16	992	70,71	541	96,43	54,54	136,39	70,71	451	53,56
15	1403	100	561	100	39,99	100	100	842	100,00

ROC CART



Node	Cases Tgt. Class	% of Node Tgt. Class	% Tgt. Class	Cum % Tgt. Class	Cum % Pop	% Pop	Cases in Node	Cum lift	Lift Pop	Uniforme	FP	Cum FP	Cum % FP
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
9	52	100,00	9,85	9,85	3,87	3,87	52	2,54	2,54	3,87	0,00	0,00	0,00
10	273	96,81	51,70	61,55	24,87	21,00	282	2,48	2,46	24,87	9,00	9,00	1,70
6	41	74,55	7,77	69,32	28,97	4,10	55	2,39	1,90	28,97	14,00	23,00	4,36
8	46	77,97	8,71	78,03	33,36	4,39	59	2,34	1,98	33,36	13,00	36,00	6,82
3	40	63,49	7,58	85,61	38,05	4,69	63	2,25	1,61	38,05	23,00	59,00	11,17
2	23	34,33	4,36	89,96	43,04	4,99	67	2,09	0,87	43,04	44,00	103,00	19,51
7	31	19,75	5,87	95,83	54,73	11,69	157	1,75	0,50	54,73	126,00	229,00	43,37
4	4	7,69	0,76	96,59	58,60	3,87	52	1,65	0,20	58,60	48,00	277,00	52,46
1	13	4,38	2,46	99,05	80,71	22,11	297	1,23	0,11	80,71	284,00	561,00	106,25
5	5	1,93	0,95	100,00	100,00	19,29	259	1,00	0,05	100,00	254,00	815,00	154,36

CHAID vs CART (curvas ROC)



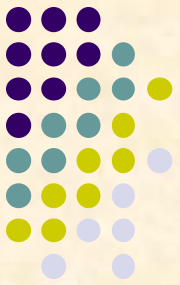
- La segmentación CHAID captura un 90% del spam con aproximadamente un 30% de FP. Por otro lado, CART identifica el mismo porcentaje con un 20% de FP

Referencias bibliográficas



- ✓ Arevalillo J.M. (2008). *Clasificación y Regresión con Árboles de Decisión*. Curso de doctorado “Algoritmos de Segmentación”. Dpto. Estadística, Investigación Operativa y Cálculo Numérico. UNED
- ✓ Breiman L. Friedman J. Olshen R. Stone C. (1984). *Classification and Regression Trees*. Wadsworth
- ✓ Hastie T. Tibshirani R. Friedman J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer
- ✓ Kass G. (1980). *An exploratory technique for investigating large quantities of categorical data*. Applied Statistics 29 (2), 119-127

Datos de contacto



Jorge Martín Arevalillo

Dpto. Estadística, Investigación Operativa y Cálculo Numérico

Universidad Nacional Educación a Distancia

Paseo Senda del Rey nº 9. 28040 Madrid

✓ Teléfono: 91 398 72 64

✓ E-mail: jmartin@ccia.uned.es