

Explotación del B2B y segmentación de la clientela

Cluster, Clasificación y Segmentación

Sesión 4. 24/10/2018

A.M.Mayoral (asun.mayoral@umh.es), J.Morales (j.morales@umh.es)

Ejemplo

Adquisición de material de oficina

Objetivo:

- Agrupación de clientes
- Estudio de los ingresos por cliente

- Cliente (500)
- Ventana de ventas
- El cliente sigue activo
- Promedio gasto en marketing para retener cliente
- Número de artículos comprados
- Número total de compras realizadas
- Ingresos por ventas al cliente
- Número empleados cliente

Ejemplo

Pagos a crédito

Objetivo:

- Determinar clientes en situación de riesgo crediticio

- Cliente (10000)
- Ingresos (\$)
- Límite de crédito (\$)
- Clasificación crediticia
- Número de tarjetas crédito
- Edad
- Sexo
- Estudiante (Si, No)
- Casado (Si, No)
- Raza
- Saldo promedio tarjeta crédito (\$)

Ejemplo

Precios de las viviendas en los suburbios de Boston

Objetivo:

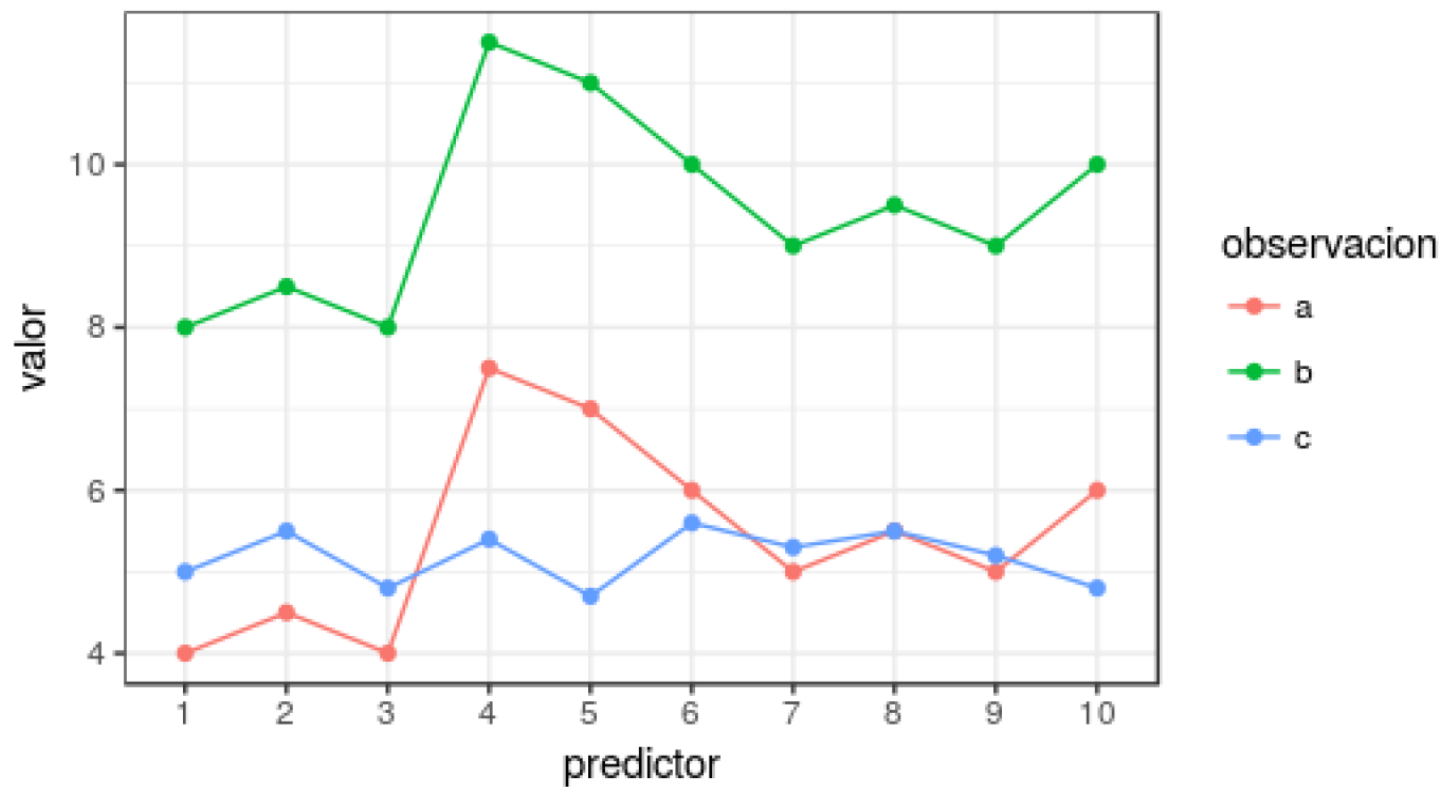
- Estudio valor medio de las residencias ocupadas.

- Viviendas (506)
- Tasa de criminalidad
- % de zona residencial
- % terreno negocios no minoristas
- Limita con el río
- Concentración de NOX
- Número medio de habitaciones
- % viviendas antes 1940
- Media distancias a centros de empleo
- Índice accesibilidad autopistas
- Impuestos a la propiedad
- % estudiante-profesor
- Índice racial
- % menores
- Valor medio de la vivienda

Técnicas de cluster

El término *clustering* hace referencia a un amplio abanico de técnicas *unsupervised* cuya finalidad es encontrar patrones o grupos (*clusters*) dentro de un conjunto de observaciones.

Objetivo: Semenjanza entre sujetos del grupo y diferencias entre grupos.



Técnicas de cluster

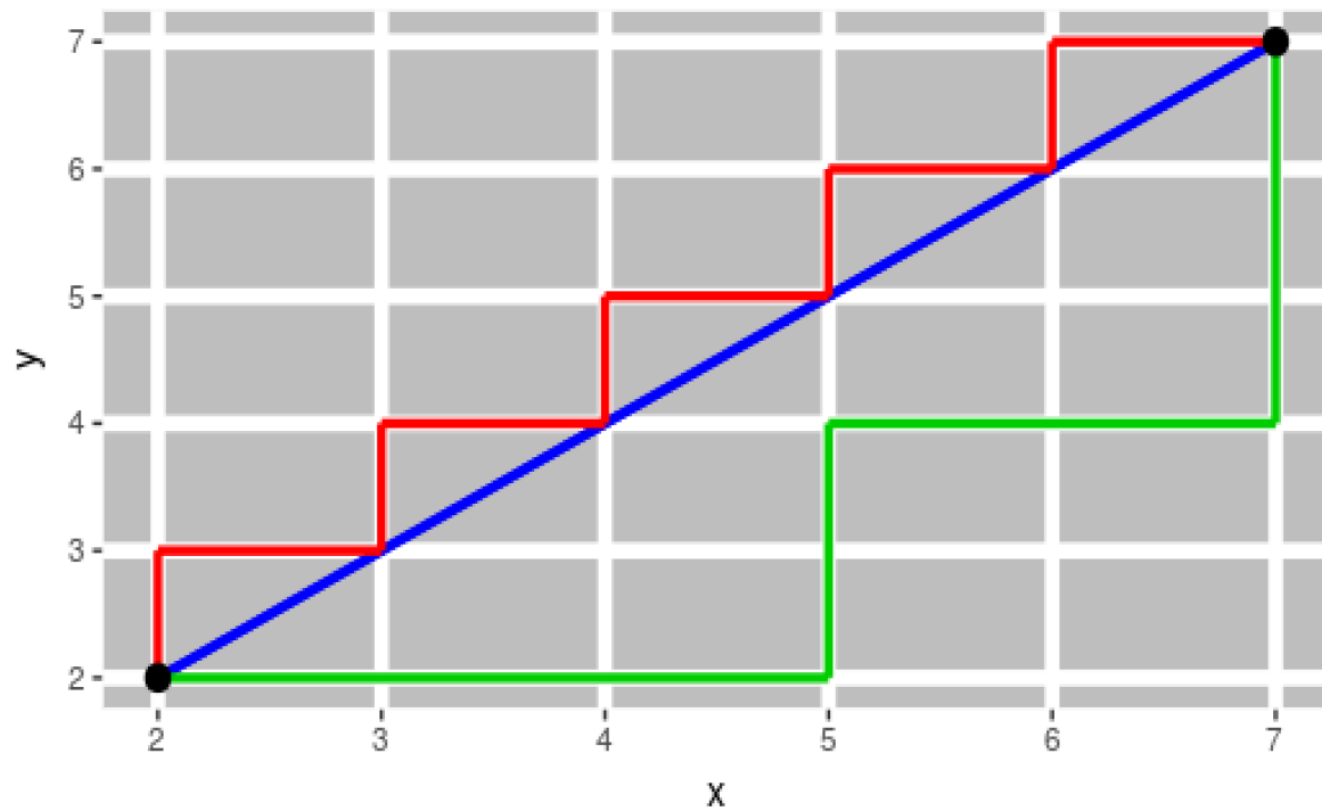
Tres tipos:

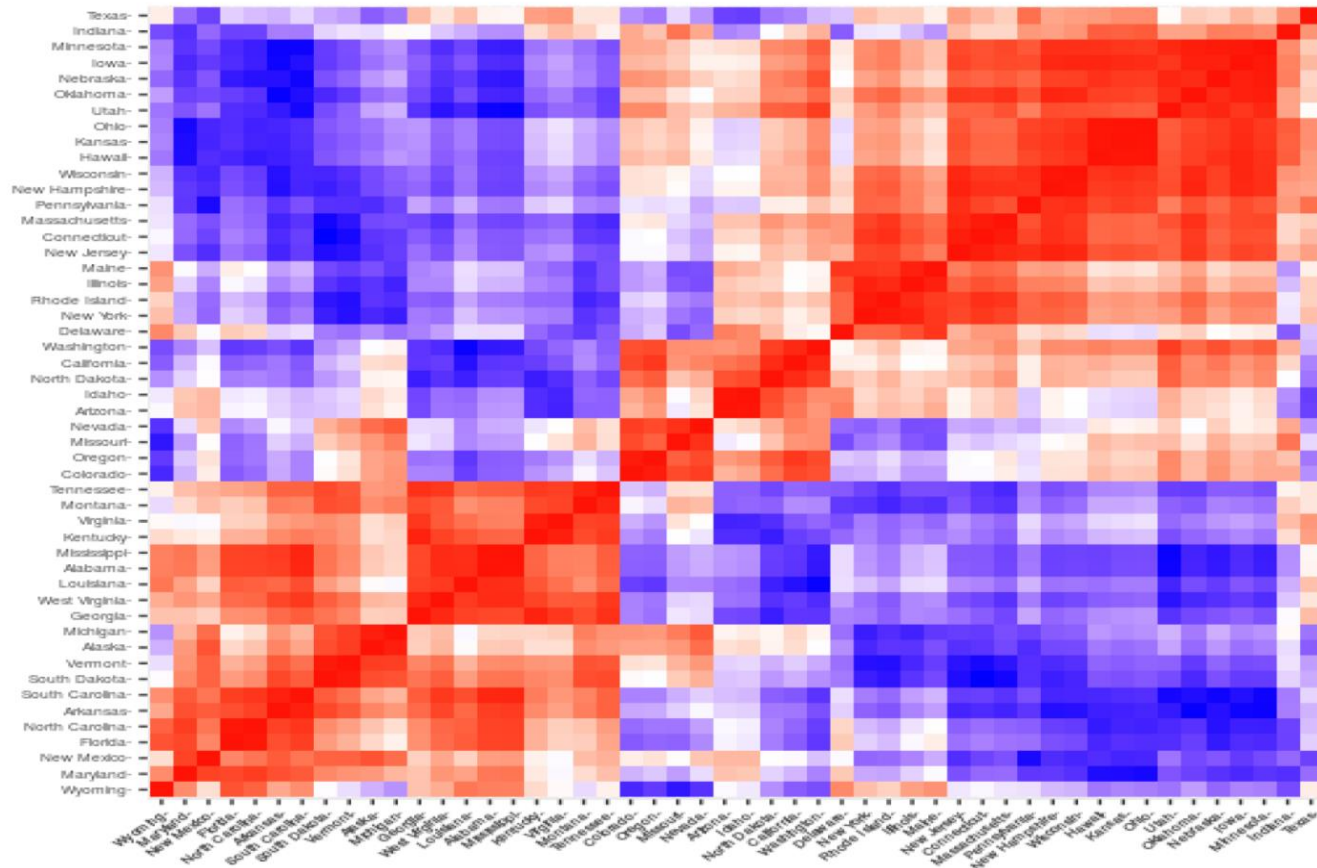
- *Partitioning Clustering (grupos fijos: K-means, PAM, CLARA)*
- *Hierarchical Clustering (grupos sin fijar: AGNES)*
- *Combinados (FUZZY)*

Distancia entre sujetos

Escalado de variables y establecimiento de:

- Distancias
 - Distancia Euclídea
 - Distancia de Manhattan
 - Correlación
 - Jackknife correlation
- Similitudes
 - Simple matching coefficient
 - Índice Jaccard



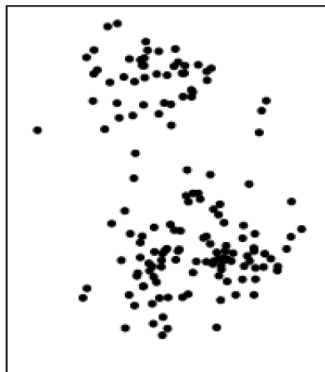


Partitioning Clustering

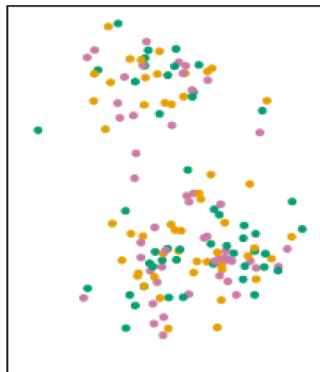
Proceso:

1. Fijamos número de clusters
2. Asignamos inicialmente un sujeto a cada grupo de forma aleatoria y fijamos un centro para el grupo.
3. Valoramos si podemos cambiar un sujeto de grupo
4. Calculamos el error cometido basado en el promedio de las distancia de cada punto al centro de todas ellas.

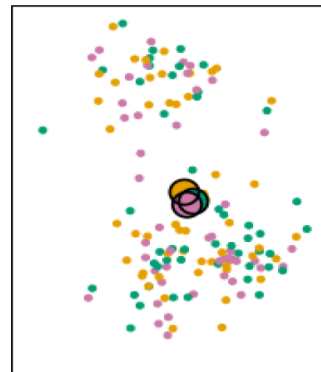
Data



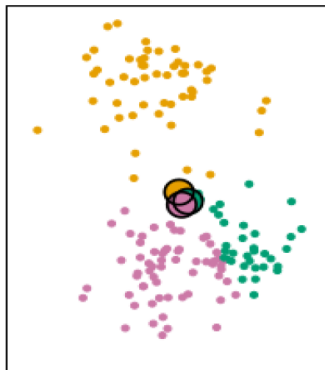
Step 1



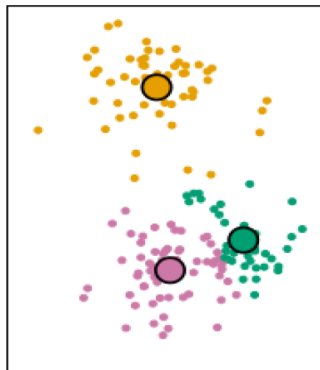
Iteration 1, Step 2a



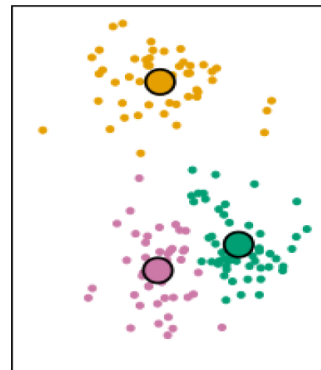
Iteration 1, Step 2b



Iteration 2, Step 2a



Final Results



Desventajas

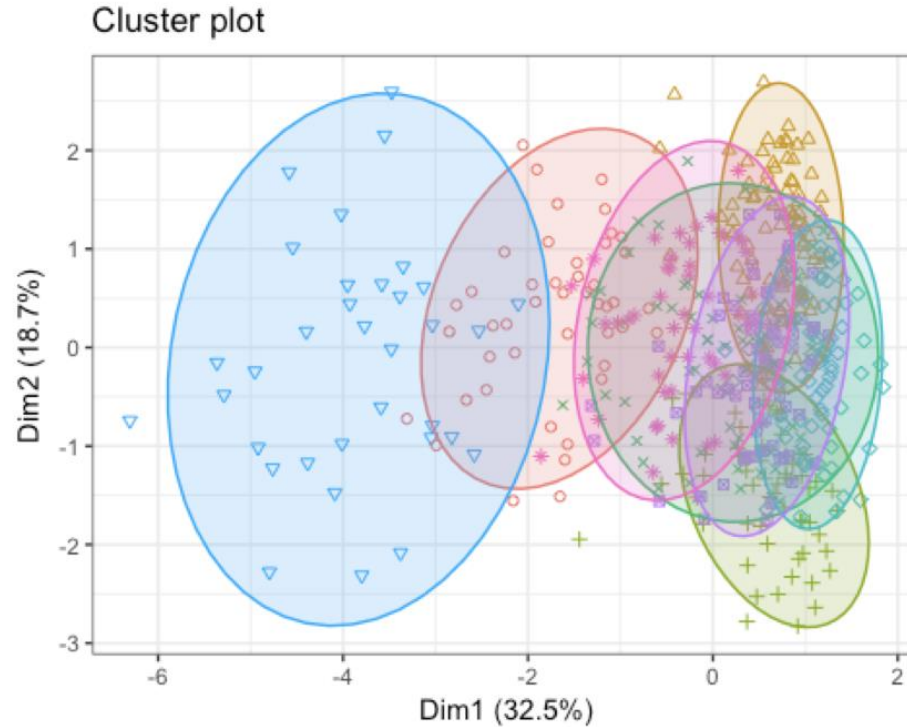
1. Mismo tipo de variables
2. Número de clusters fijo
3. La agrupación varía en función de los centros iniciales
4. Falta de robustez con sujetos alejados

Ventajas

1. Fácil de realizar
2. Implementación rápida en cualquier interfaz de programación

Ejemplo Material de oficina

- 8 grupos
- CLARA



Ejemplo Material de oficina

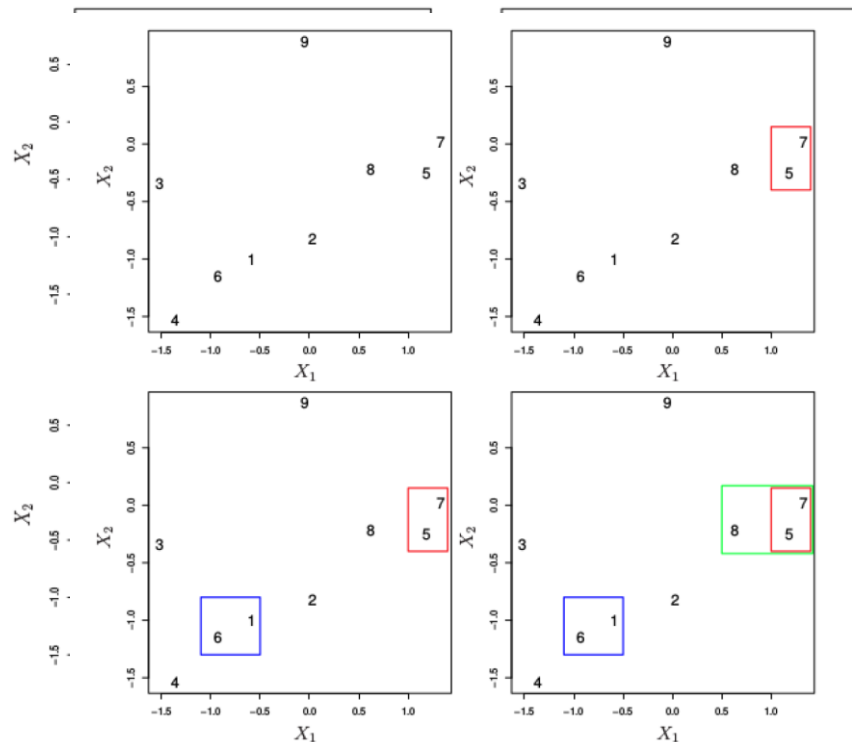
	Ventana	Marketing	Artículos	Beneficio	Compras	Empleados
C1	-1.31	-0.13	1.08	-1.04	-0.27	-0.65
C2	0.63	-0.60	-0.66	-0.81	1.03	0.12
C3	0.63	1.50	-0.07	0.47	-0.93	1.31
C4	0.08	-1.09	0.15	1.00	-0.27	-0.73
C5	0.63	-0.13	-0.62	0.63	1.03	1.01
C6	-2.84	-1.09	2.65	-0.26	-0.27	-0.00
C7	0.63	0.71	-0.61	0.29	-0.27	-0.73
C8	-0.19	0.47	-0.39	-1.07	-0.93	-0.88

Hierarchical Clustering

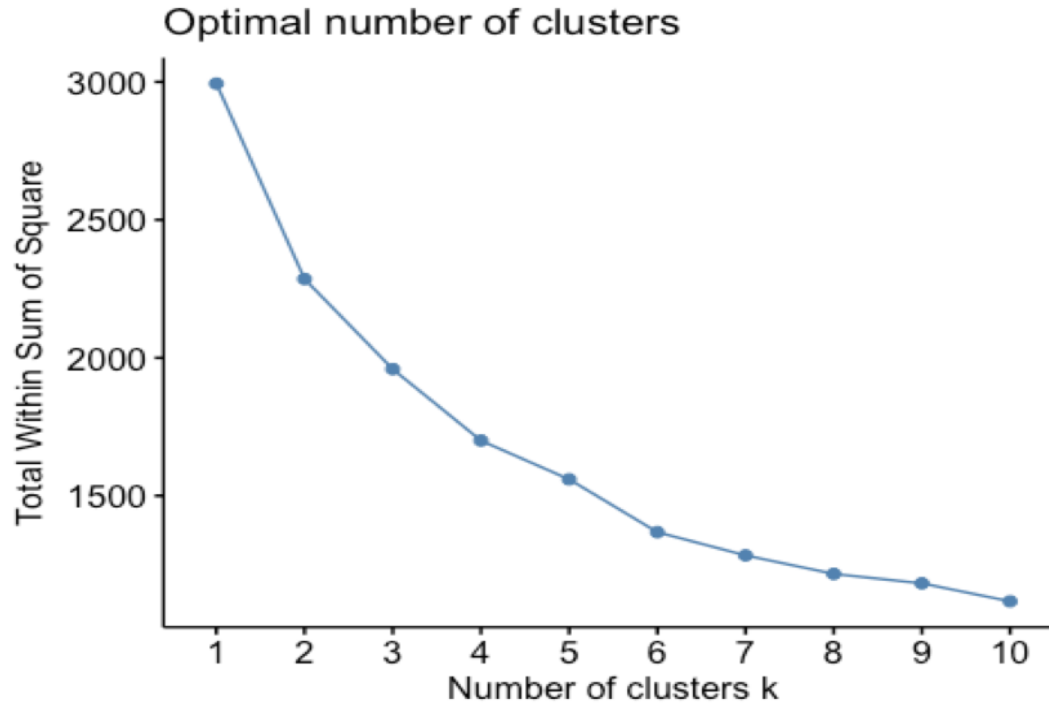
Proceso AGNES:

1. Fijar medida de distancia o similitud entre sujetos
2. Fijar distancia entre grupos (Máxima, Mínima, Promedio, Centroide, Ward)
3. Paso 0: Todo sujeto es un cluster.
4. Paso 1: Con la medida definida en dos se agrupan los dos grupos más cercanos.
5. Se sigue agrupando hasta que sólo queda un grupo.

Hierarchical Clustering

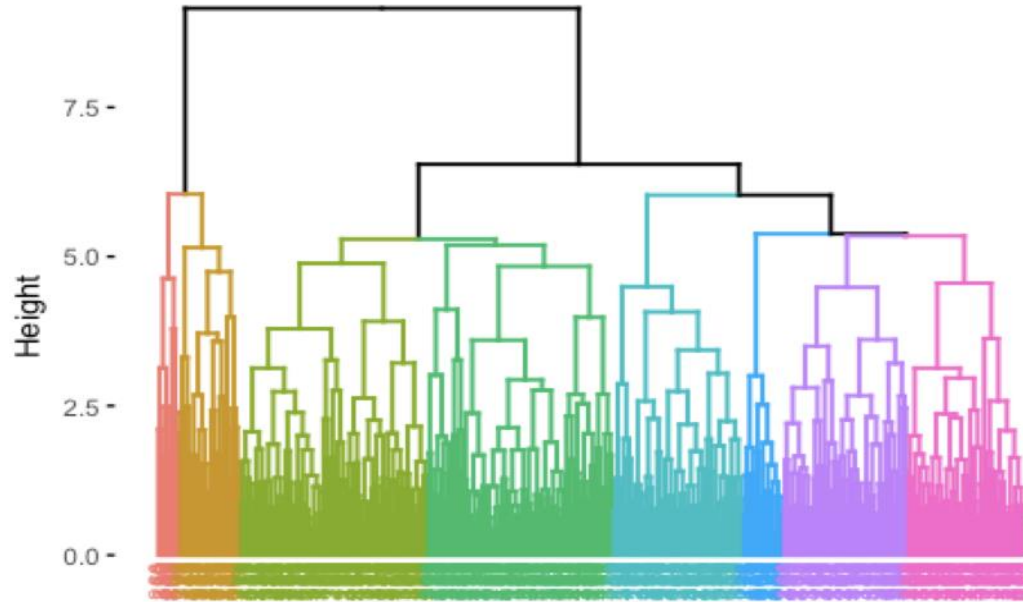


Ejemplo Material de oficina

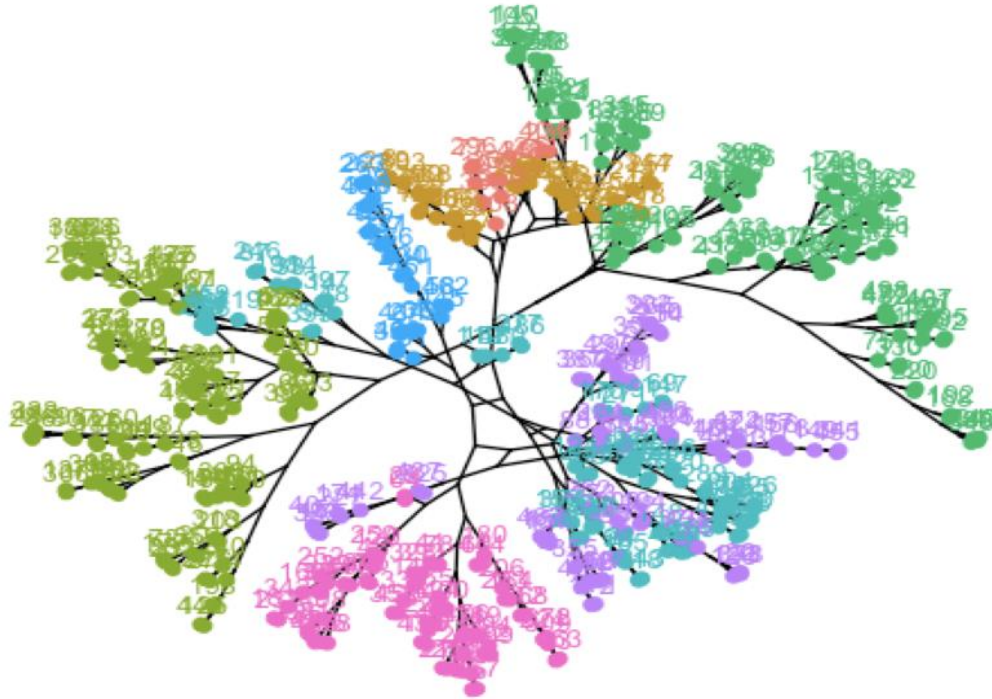


Ejemplo Material de oficina

Cluster Dendrogram



Ejemplo Material de oficina

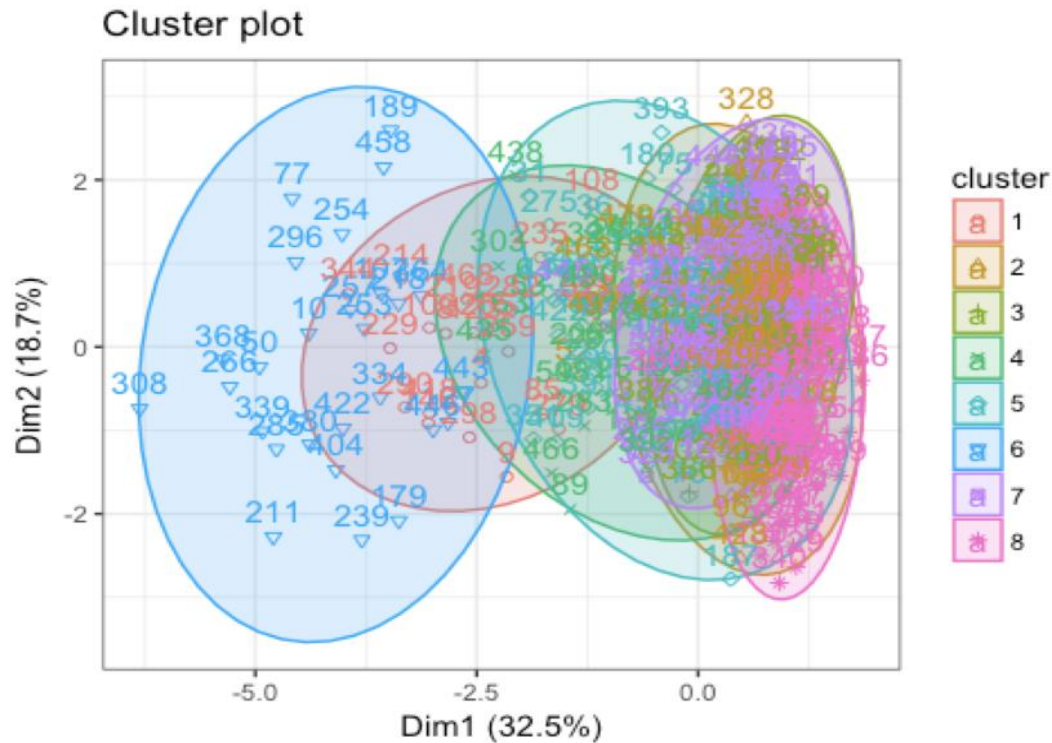


Combinado (Fuzzy)

Proceso FANNY:

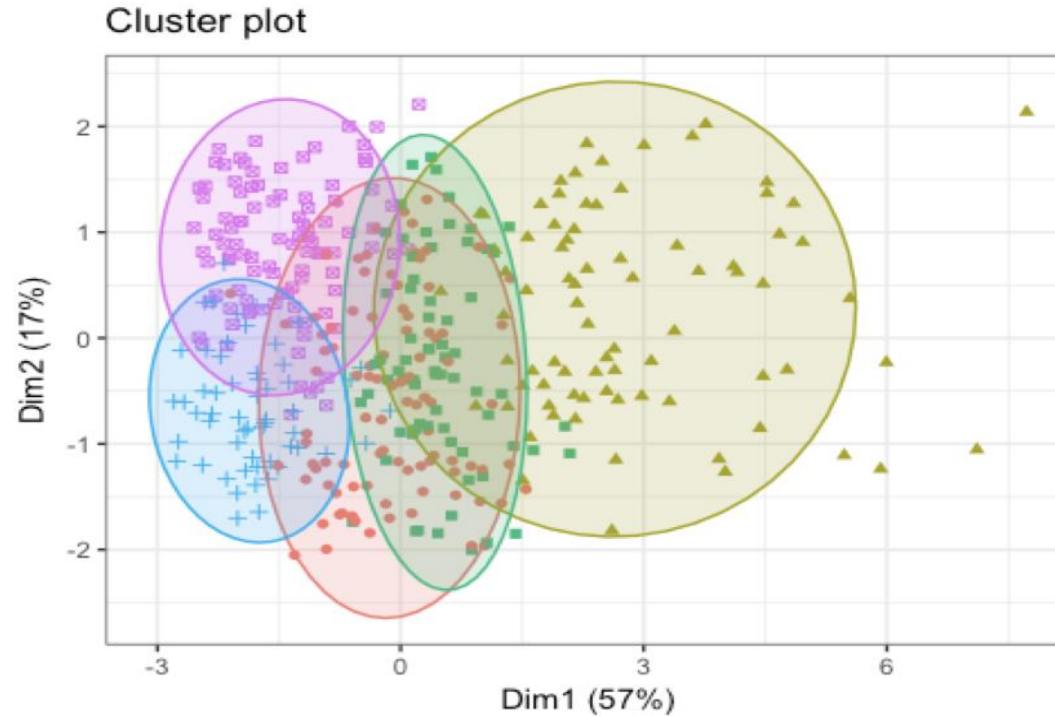
1. Aplicar Hierarchical clustering
2. Calcular centro de cada cluster
3. Aplicar K-means clustering

Ejemplo Material de oficina

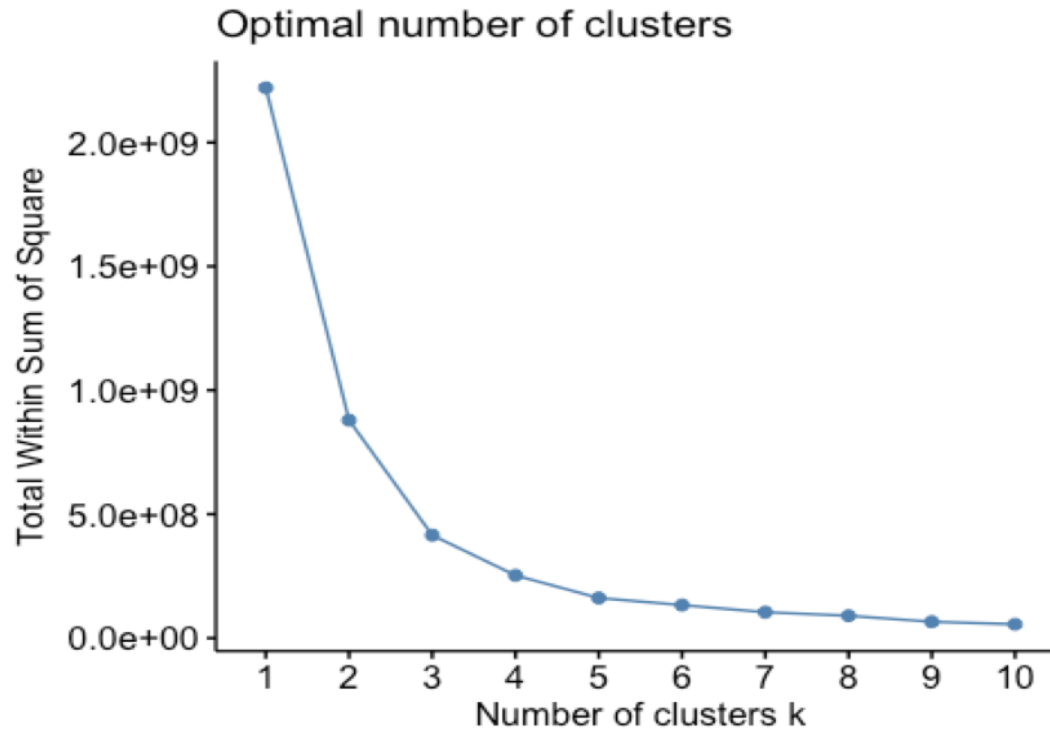


Ejemplo Estudio crédito

CLARA



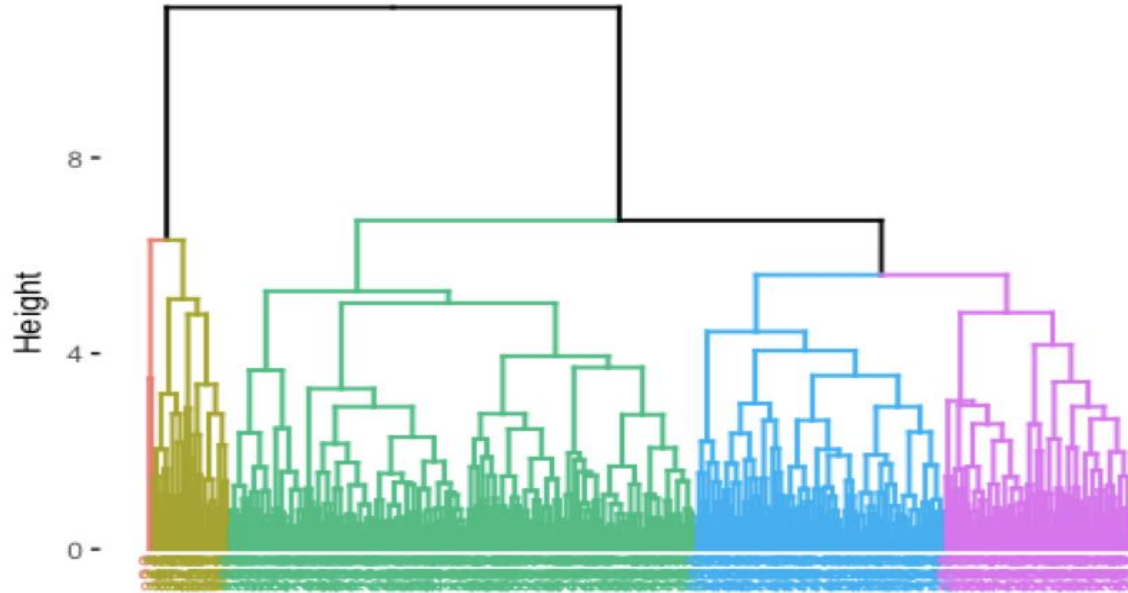
AGNES



Ejemplo Estudio crédito

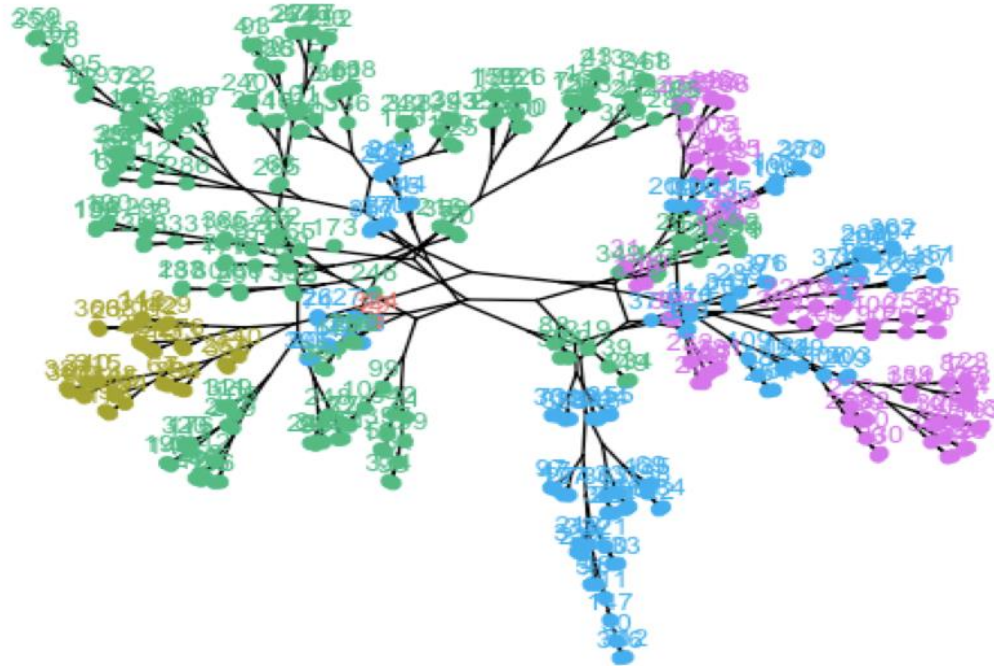
AGNES

Cluster Dendrogram



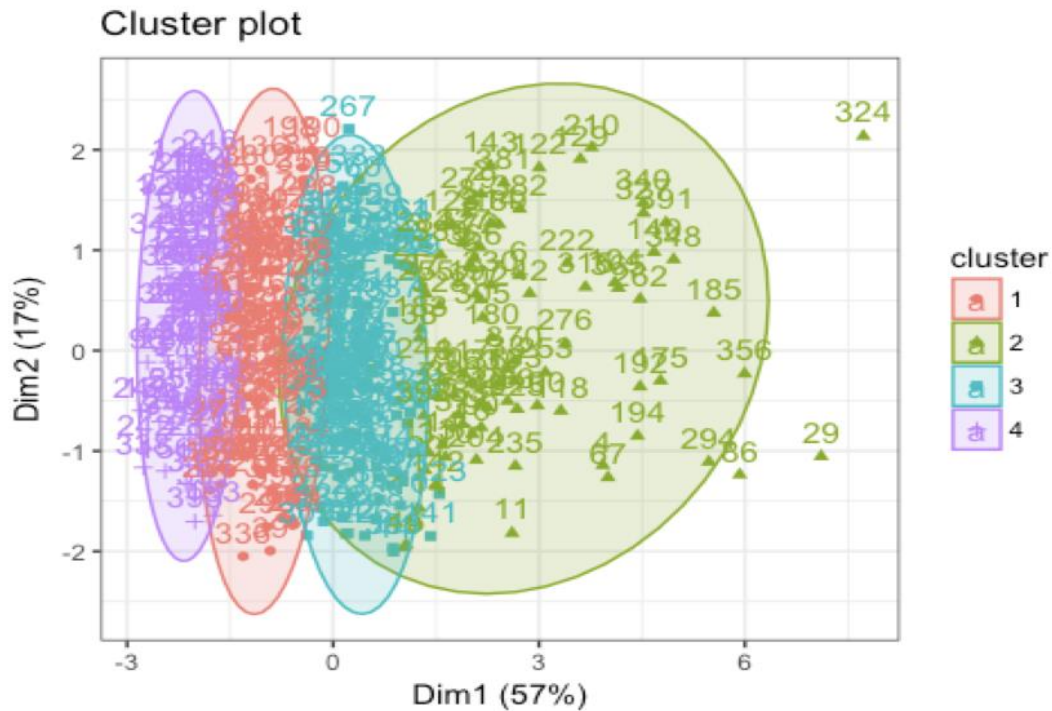
Ejemplo Estudio crédito

AGNES



Ejemplo Estudio crédito

FANNY

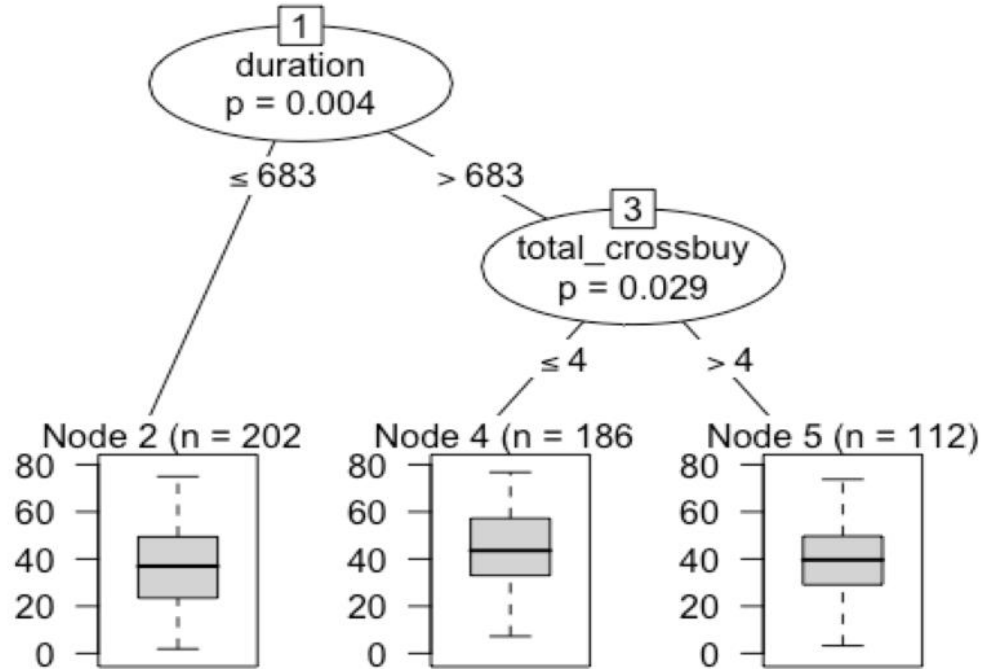


Árboles de clasificación

Magnitudes (predictores) que interaccionan de forma compleja para explicar el comportamiento de una cantidad de interés.

Objetivo: Segmentar el espacio de los predictores en regiones simples, dentro de las cuales es más sencillo manejar las interacciones.

Árboles de clasificación



Árboles de clasificación

Desventajas:

- Simplifican demasiado asociaciones muy complejas
- Perdemos información si trabajamos sólo con variables de tipo continuo.
- Se favorece el uso de predictores numéricos frente a predictores categóricos.

Árboles de clasificación

Ventajas:

- Fáciles de interpretar
- Fáciles de representar
- No hace falta depurar los datos
- Poco sensibles a observaciones alejados
- Útiles en la exploración inicial de información

Árboles de clasificación

Tipos:

- Árboles clásicos
- Random forest
- C5.0

Árboles de clasificación

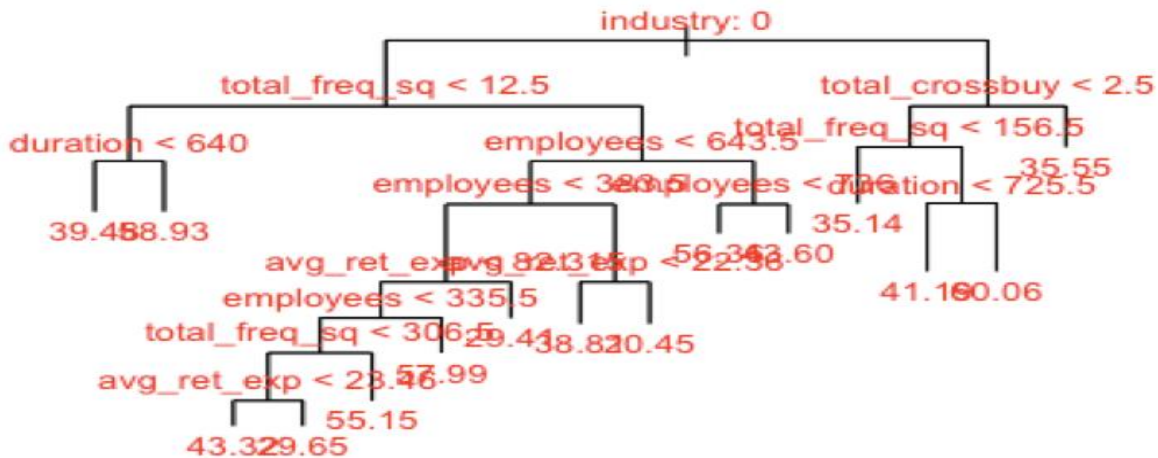
Proceso:

- Dividir el conjunto de datos en entrenamiento y predicción
- Ajustar un primer modelo con los datos de entrenamiento
- Encontrar el mejor modelo
- Predecir el comportamiento del árbol mediante la muestra de predicción aplicada al árbol óptimo.

Ejemplo Material de oficina

Árbol inicial:

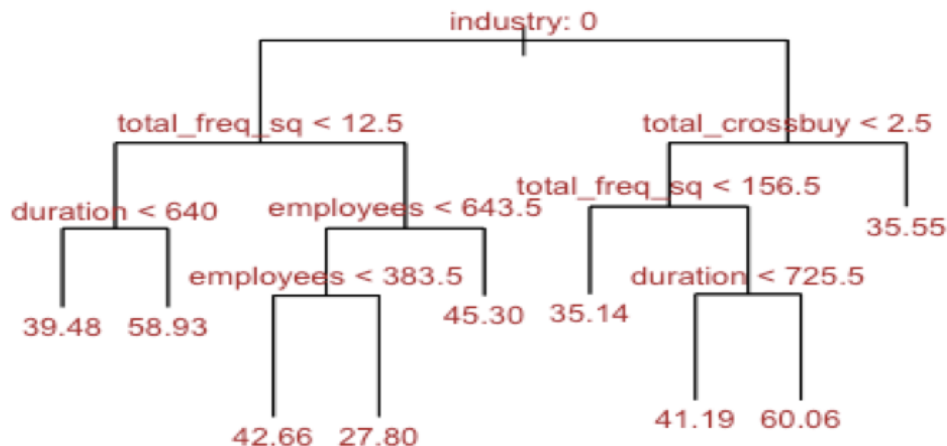
Beneficio



Ejemplo Material de oficina

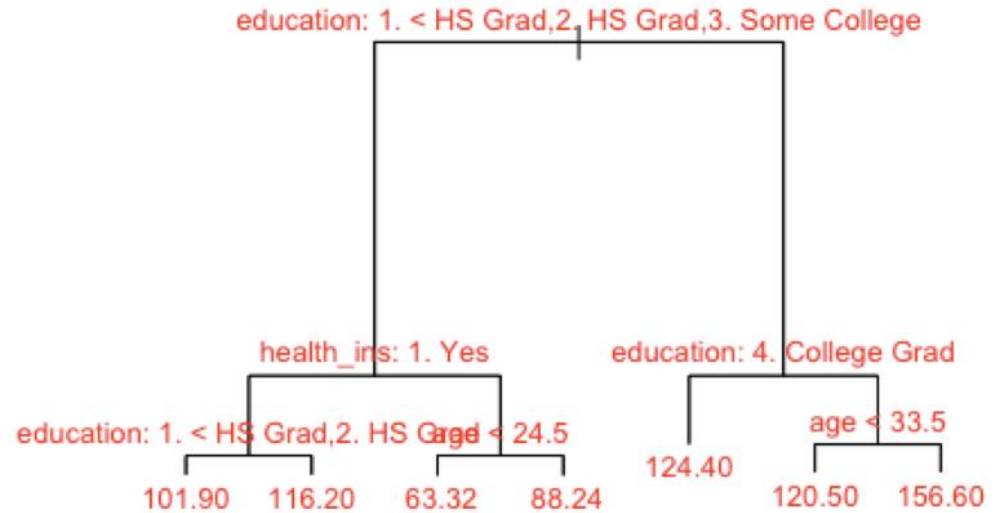
Árbol óptimo:

Beneficio



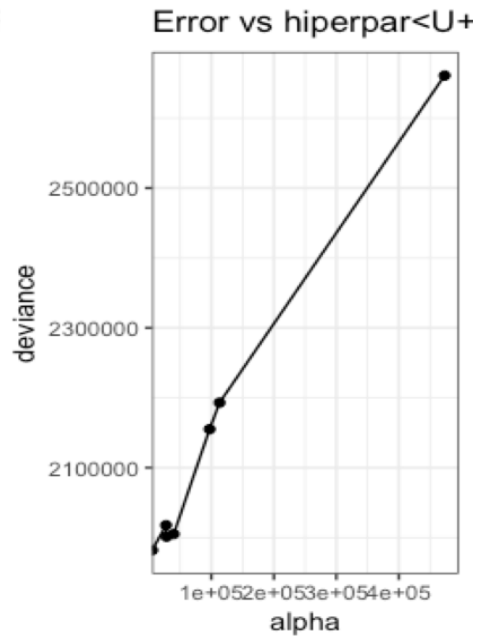
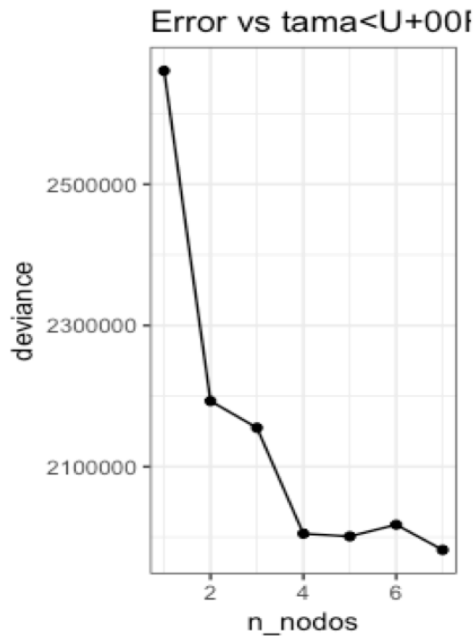
Ejemplo Sueldos

Árbol inicial:
Sueldo



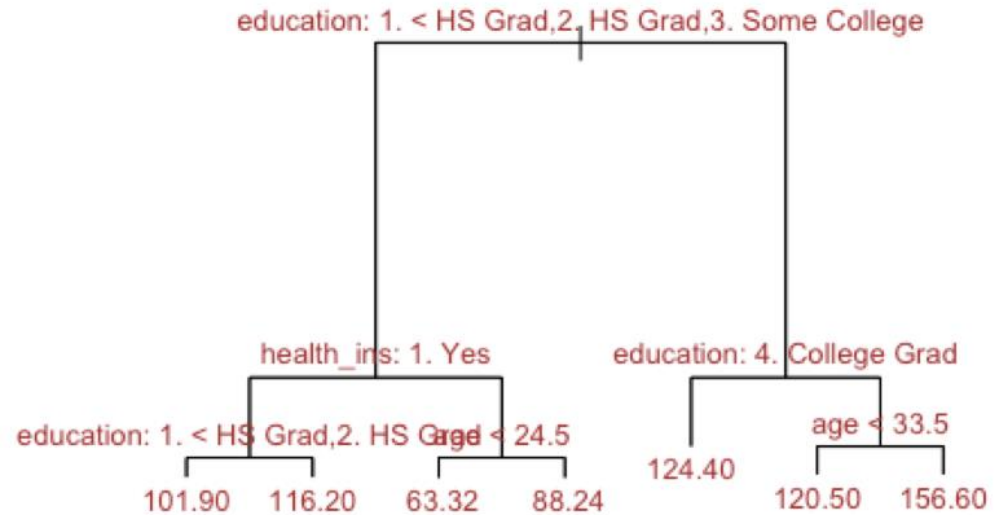
Ejemplo Sueldos

Selección del número de ramas



Ejemplo Sueldos

Árbol óptimo: Sueldo



Ejemplo Sueldos (Split, n, deviance, resp)

```
## 1) root 1500 2651000 112.50
##      2) education: 1. < HS Grad,2. HS Grad,3. Some College 927 911100 98.49
## 4) health_ins: 1. Yes 573 487600 107.20
##      8) education: 1. < HS Grad,2. HS Grad 361 284600 101.90 *
##      9) education: 3. Some College 212 175500 116.20 *
##      5) health_ins: 2. No 354 310500 84.44
## 10) age < 24.5 54 17650 63.32 *
## 11) age > 24.5 300 264400 88.24 *
##      3) education: 4. College Grad,5. Advanced Degree 573 1267000 135.10
## 6) education: 4. College Grad 343 517200 124.40 *
## 7) education: 5. Advanced Degree 230 652300 151.00
## 14) age < 33.5 36 17790 120.50 *
## 15) age > 33.5 194 594800 156.60 *
```

Árboles de clasificación.

Proceso Bagging (Random Forests):

- Calcular multitud de árboles de clasificación formando un bosque
- Se trabaja con una predicción media de todos ellos

Proceso Boosting (C5.0)

- Ir ajustando secuencialmente modelos sencillos de forma que cada uno aprende del anterior.

Ejemplo Precios vivienda (precio)

