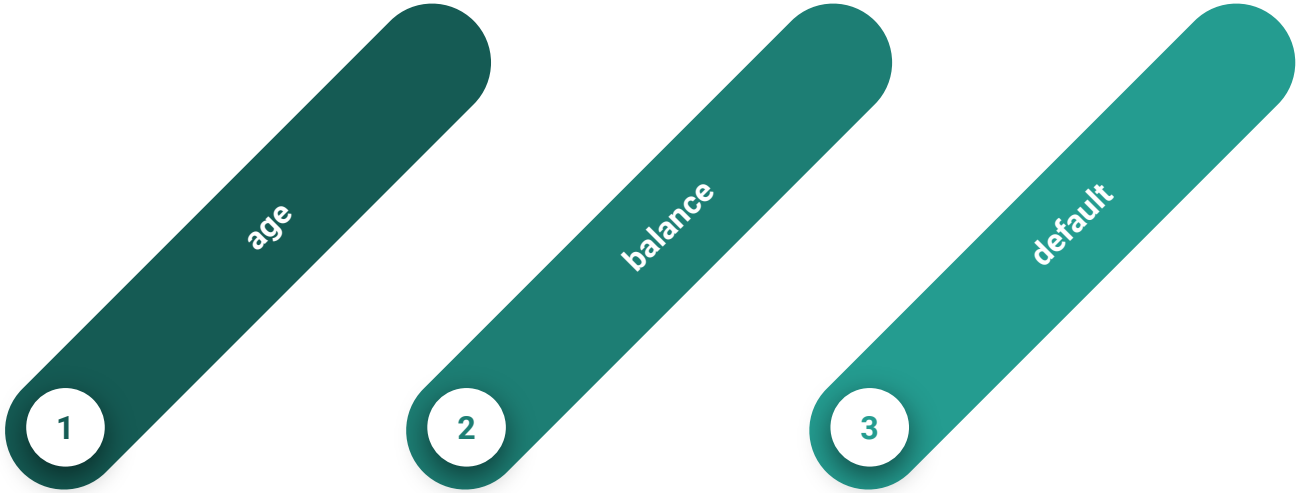
MINERÍA DE DATOS A CONJUNTO DE DATOS BANCARIOS.

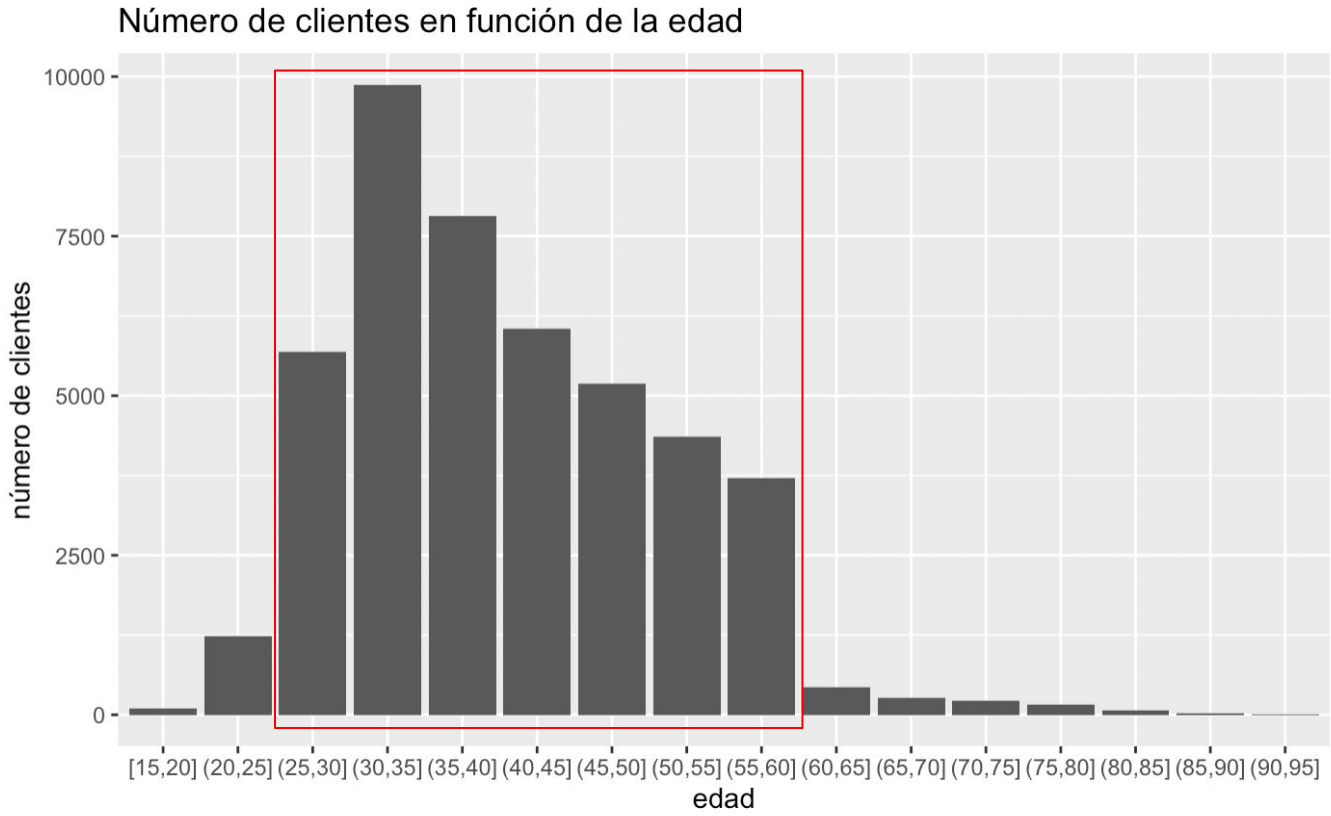
BOUCHAIB, EL ASRI
ENRIQUEZ CAICEDO, BRIAN DANILO
ESTRADA NIETO, JUAN CARLOS
PALMA, JUAN JOSÉ
PREEL, LUCAS

DATASET

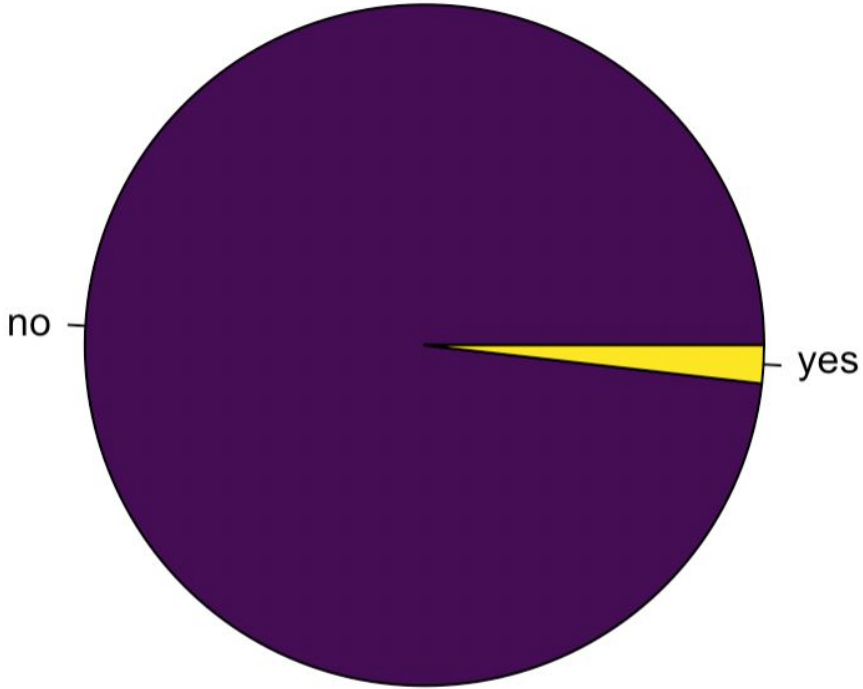
- Campaña de marketing bancario.
- 17 atributos

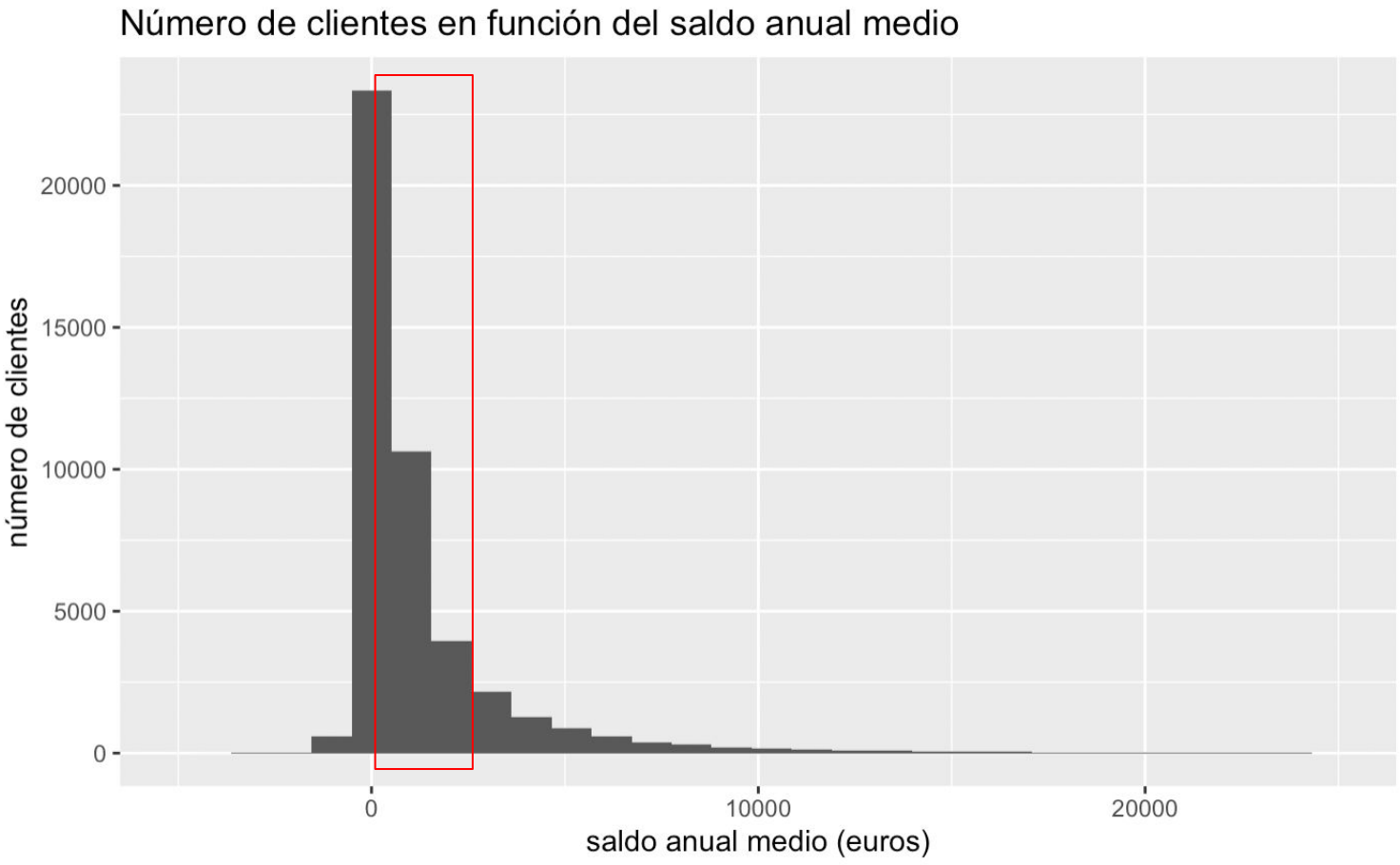
Visualización



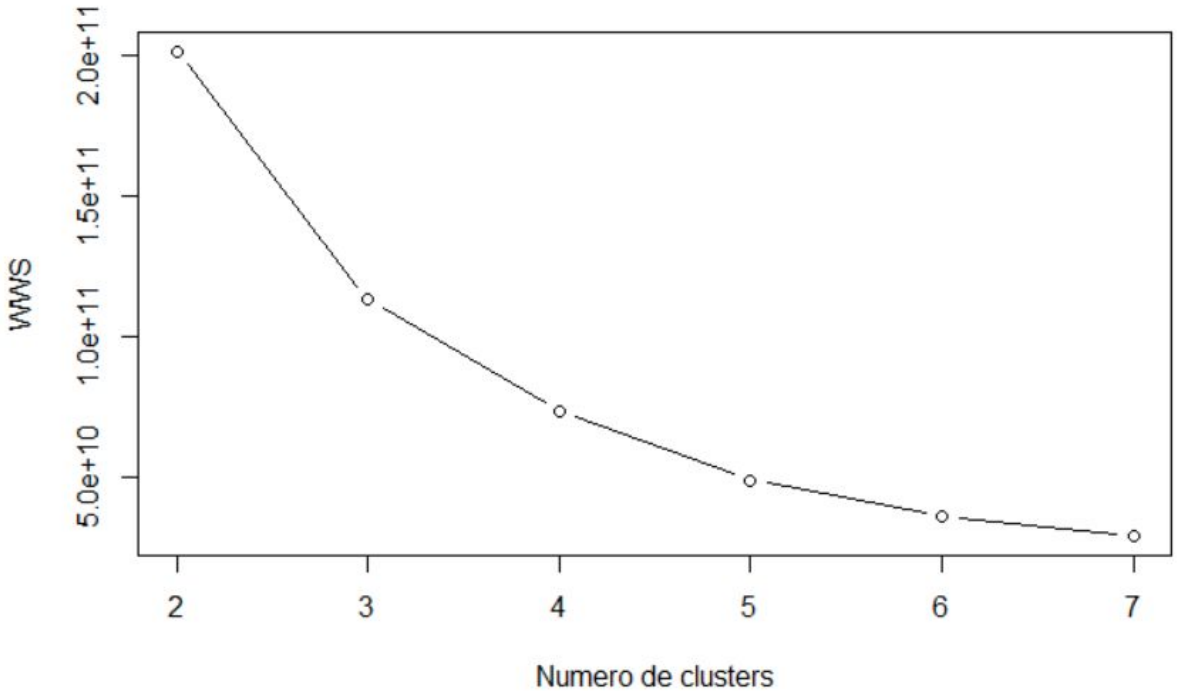

Número de clientes en función de la edad
10000
7500
5000
2500
0
número de clientes
[15,20]
(20,25]
(25,30]
(30,35]
(35,40]
(40,45]
(45,50]
(50,55]
(55,60]
(60,65]
(65,70]
(70,75]
(75,80]
(80,85]
(85,90]
(90,95]
edad

Clientes que tienen un crédito impagado y los que no



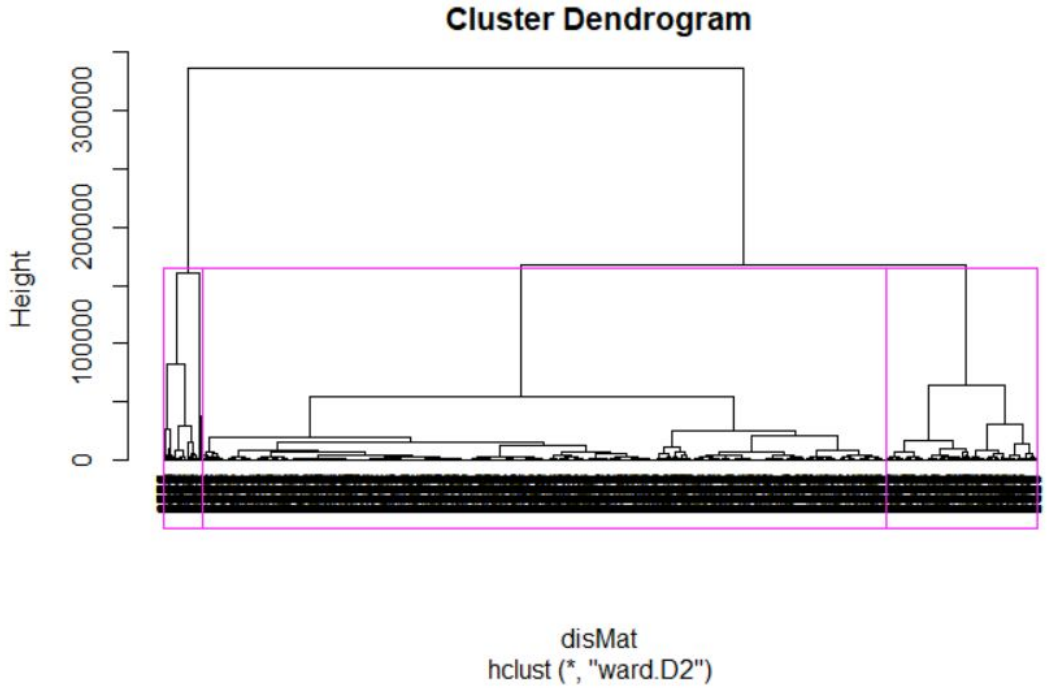

Número de clientes en función del saldo anual medio
número de clientes
20000
15000
10000
5000
0
0
10000
20000
saldo anual medio (euros)

Algoritmos de Clustering - Kmeans

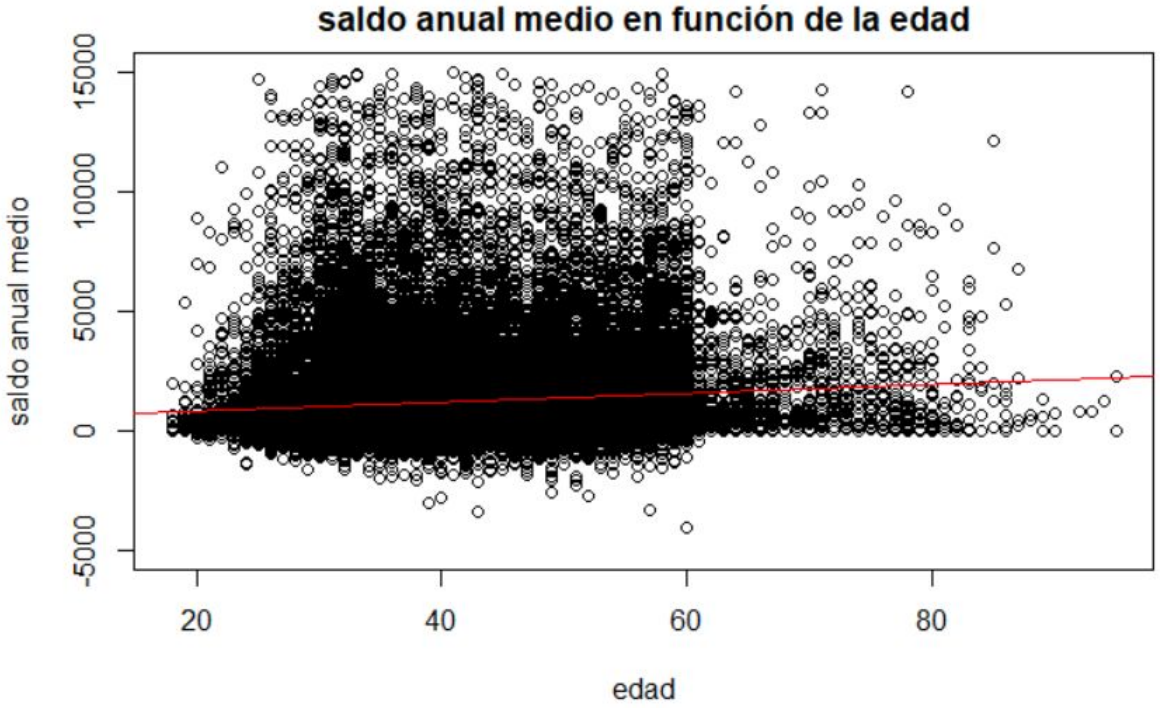


- k = 2 a 7
- Within cluster Sum of Square

Algoritmos de Clustering - Jerárquico



Regresión Simple y Multivariable

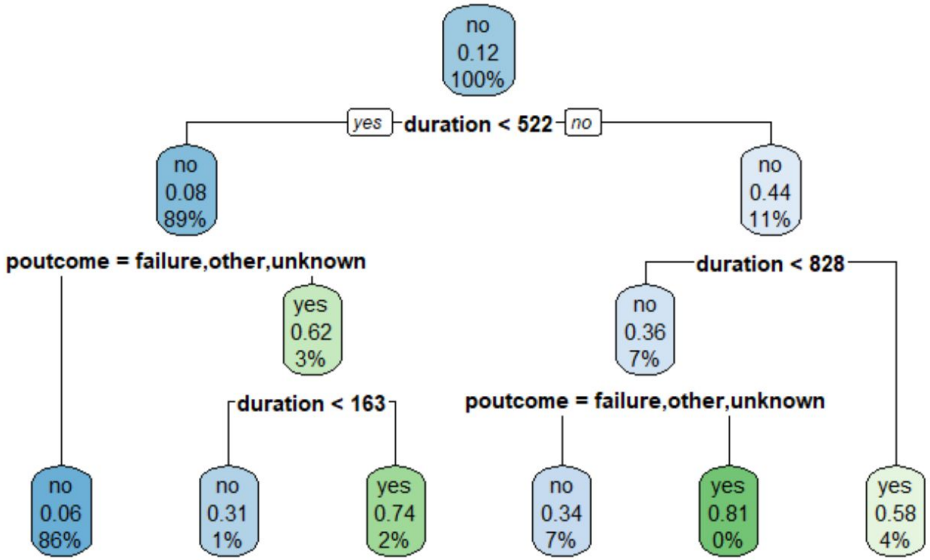


R_square

Simple: 0.0098

Multivariable: 0.06

Clasificación




```
library(rpart)
mytree <- rpart(y ~ .,
data = df_train,
method ="class")
mytree
library(rpart.plot)
rpart.plot(mytree)
```

Se realizan predicciones:

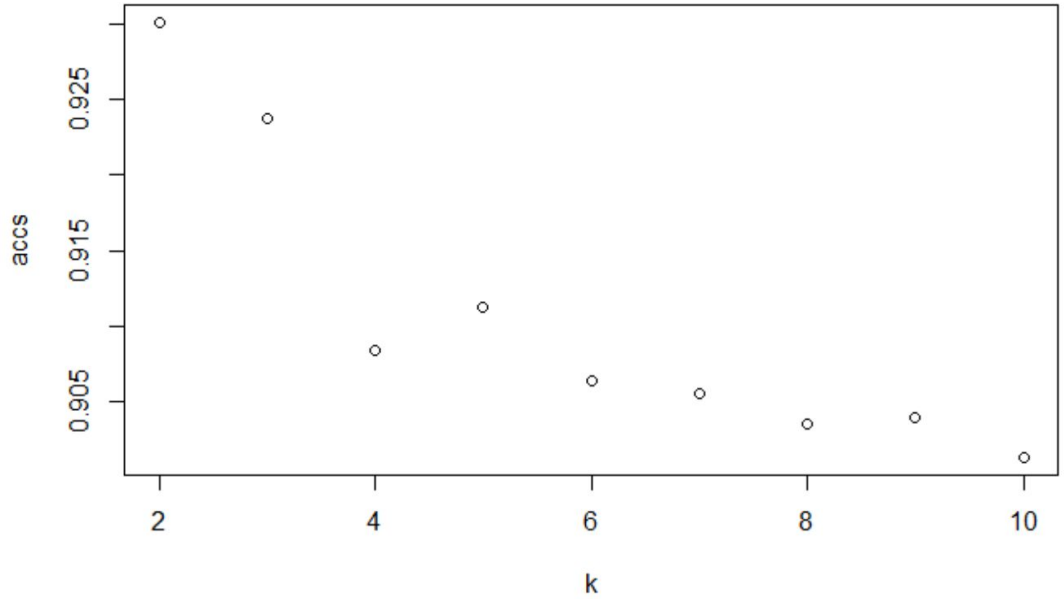
```
preds <- predict(mytree, newdata = df_test, type = "class")
```

Se clasifican por lo tanto de la siguiente manera en la matriz de predicciones:

	Predicciones	
	No	Si
No	3896	104
Si	348	173

KNN (K-Nearest-Neighbors)

- Conversión
- Valor inicial *k*
- Exploración de valores
- Gestión de recursos

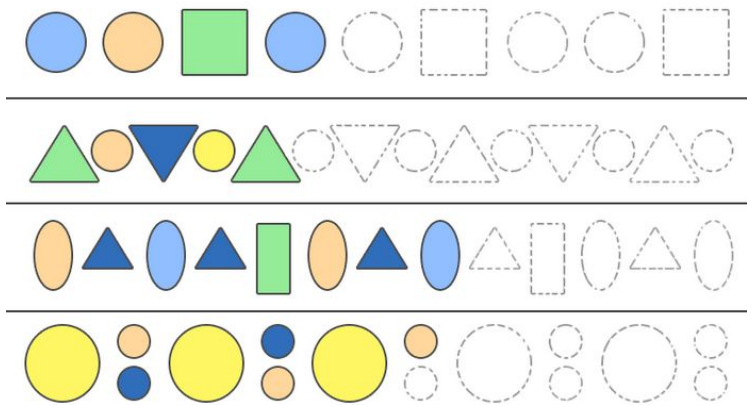


BIGML

- Machine Learning.
- Accesible.
- Reduce barreras de entrada.

Association Discovery

- No supervisado.
- Correlaciones interesantes.
- Patrones.
- Antecedente-consecuente.



Métricas - Soporte

Proporción de transacciones de un conjunto de ítems.

ID	Leche	Pan	Mantequilla	Cerveza
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Regla: {Leche, Pan} => {Mantequilla}

X = {Leche, Pan}

Y = {Mantequilla}

D = todos los datos.

$sop(X) = |X| \div |D|$

$sop(\{Leche, Pan\}) = 2 \div 5 = 0.4$

Es decir, el soporte es del 40% (2 de cada 5 transacciones).

Métricas - Confianza

Probabilidad del consecuente cuando está presente el antecedente.

ID	Leche	Pan	Mantequilla	Cerveza
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Regla: {Leche, Pan} => {Mantequilla}

X = {Leche, Pan}

Y = {Mantequilla}

conf(X => Y) = sop(X U Y) ÷ sop(X)

conf({Leche, Pan} => {Mantequilla})

= sop({Leche, Pan} U {Mantequilla}) ÷ sop({Leche, Pan})

= 0.2 ÷ 0.4 = 0.5

El 50% de las reglas que contienen 'leche' y 'pan' en el antecedente también tienen 'mantequilla' en el consecuente.

Métricas - Lift

Proporción entre el soporte observado y el soporte esperado si X e Y fueran independientes.

ID	Leche	Pan	Mantequilla	Cerveza
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Regla: {Leche, Pan} => {Mantequilla}

X = {Leche, Pan}

Y = {Mantequilla}

lift(X => Y) = sop(X U Y) ÷ (sop(X) x sop(Y))

lift({Leche, Pan} => {Mantequilla})

= sop({Leche, Pan} U {Mantequilla}) ÷ sop({Leche, Pan}) x sop(Mantequilla)

= 0.2 ÷ (0.4 x 0.4) = 1.25

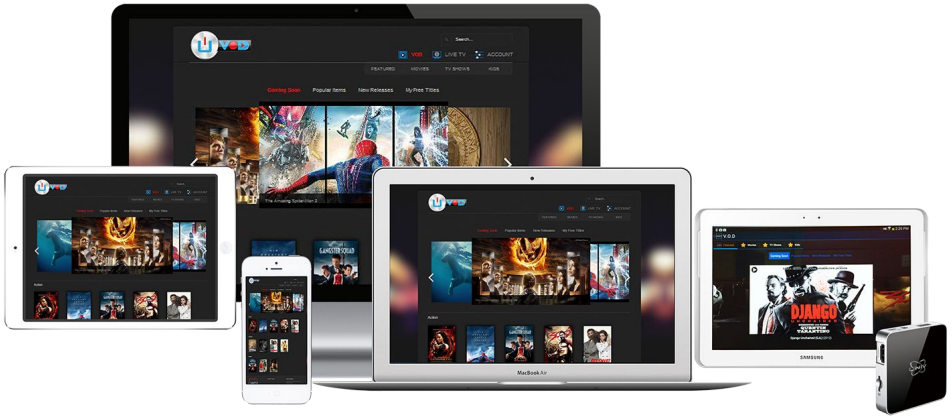
Lift = 1: independientes.

Lift > 1: son dependientes.

Lift < 1: se sustituyen, efecto negativo.

Casos de uso de las reglas de asociación

- Medicina
- Comercio



- User experience
- Entretenimiento

Usos de BIGML

Identifica co relaciones entre los datos y esto permite revisar comportamientos con un objetivo puntual y marcado, esto se realiza a través de un “sí” y un “entonces” es decir, sí sucede algo entonces se toma esta determinación, el comportamiento lo podemos observar de la siguiente manera:

fecha	cliente	cuenta	canal	tipo	Código postal	Monto de compra
10/06/2020	Rob	089	online	a	41006	2.500
03/06/2020	Ash	070	debit	a	42007	1.200
04/08/2020	Eric	064	debit	b	41221	450
10/08/2020	Rob	089	online	b	41006	700
11/01/2020	Rob	089	online	a	47023	1.350
12/25/2020	Ash	070	debit	a	42007	500
12/28/2020	Ash	070	online	b	42007	720

Como configurar AD para BIGML

Configuraciones básicas

- Número máximo de asociaciones
- Max items in antecedent
- Search Strategy
- Search for complementary items
- Search for missing items

Tipos de datos

- Numeric
- Categorical
- Date time
- Text
- Ítems

Configuraciones avanzadas

- Measure
- Minimun significance
- Consequent
- Discretization
- Sampling