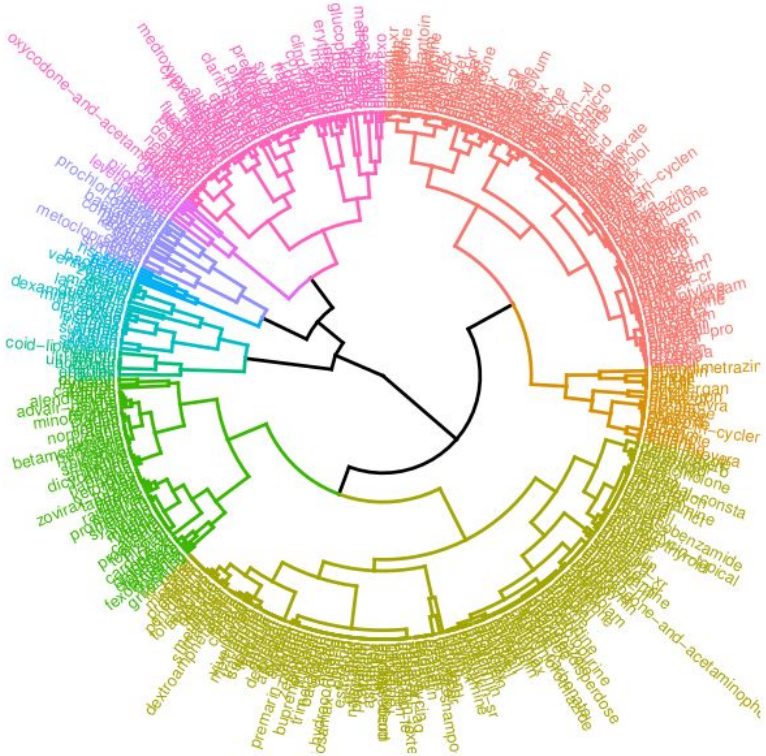
ugr Universidad de Granada



Tratamiento inteligente de datos

DRUGLIB DATASET

Alejandro Campoy Nives
Gema Correa Fernández
Luis Gallego Quero
Jonathan Martín Valera
Andrea Morales Garzón

Índice

1. Nuestros datos
2. Preprocesamiento
   - **Datos no estructurados → Datos estructurados**
3. Modelos exploratorios (EDA)
   - **Análisis estadístico de los datos**
   - **Análisis de sentimientos**
4. Modelos descriptivos
   - **Reglas de asociación y agrupamientos**
5. Modelos predictivos
   - **KNN, SVM, Naive Bayes, Regresión, Árboles de decisión y Random Forest**

NUESTROS DATOS

Nuestros datos

- **urlDrugName** → Nombre del medicamento .
- **rating** → Puntuación del medicamento (1-10).
- **effectiveness** → Clasificación de la efectividad (5).
- **sideEffects** → Clasificación efectos secundarios (5).
- **condition** → Diagnóstico.
- **benefitsReview** → Opinión sobre los beneficios.
- **sideEffectsReview** → Opinión sobre los efectos secundarios.
- **commentsReview** → Comentario general.

PREPROCESAMIENTO

Preprocesamiento I

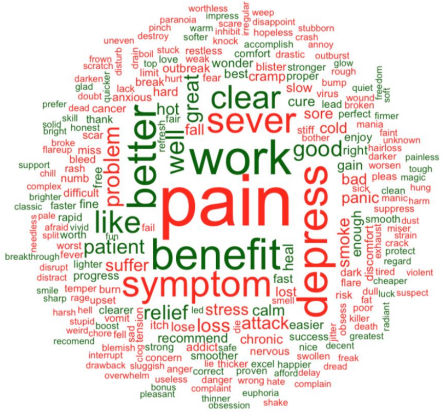
1. Lectura del dataset a usar.
2. Eliminar columnas sin información relevante.
   - ID y comentsReview.
3. Eliminar filas que contienen caracteres raros.
4. Eliminar medicamentos repetidos.
5. Cuantificación de variables.
   - sideEffects → sideEffectsNumber
   - effectiveness → effectivenessNumber
6. Cálculo del rating ponderado.
7. Conversión de rating a variable binaria.

Preprocesamiento II

8. Cambiar el orden para la columna sideEffectsNumber.
9. Representación gráfica de los datos → Diagramas de barras.
10. Creación del Corpus → benefitsReview y sideEffectsReview.
11. Correlación entre las variables.
12. Representación gráfica de las frecuencias del Corpus.
13. Eliminar signos de puntuación.
14. Conversión de mayúsculas a minúsculas.
15. Stopwords.
16. Agrupación de sinónimos.

Preprocesamiento III

17. TF-IDF.
    - Frecuencia de aparición del término (tf).
    - Frecuencia inversa del documento (idf).
18. Stemming.
20. Valores perdidos → Obtención y eliminación.
21. Sparsity.
22. Matriz de términos.
23. Nube de palabras.



MODELOS EXPLORATORIOS

Análisis exploratorio

- Rating

  Min-max [1,10]

  Media: 7

  Mediana: 8

- weightedRating

  Min-max [2,10]

  Media: 7.7

  Mediana: 8

- EffectivenessNumber

  Min-max [1,5]

  Media: 3.93

  Mediana: 4

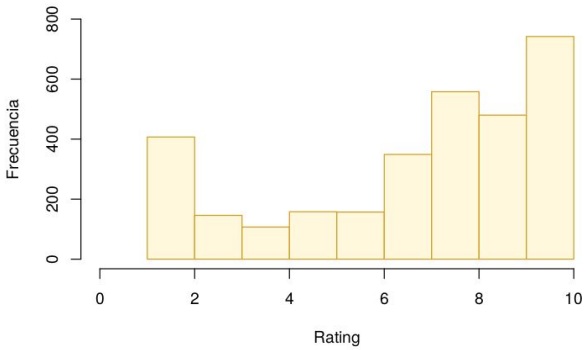
- sideEffectsNumber

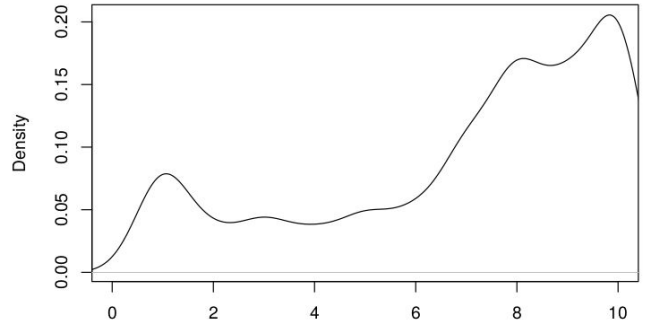
  Min-max [1,5]

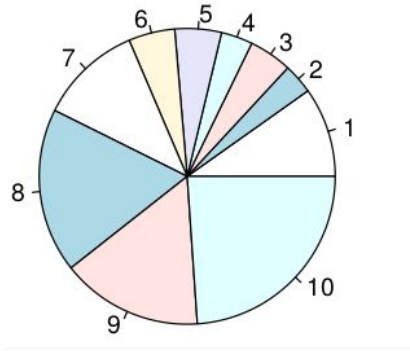
  Media: 2.3

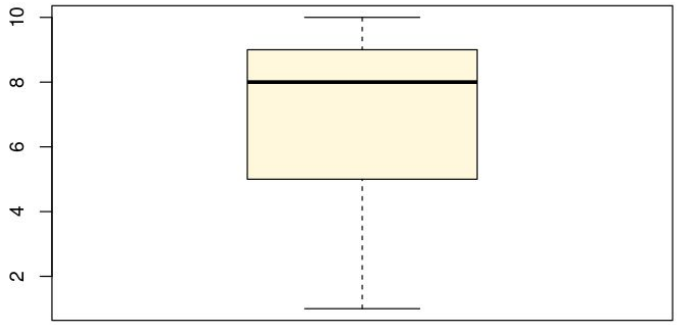
  Mediana: 2

Valoraciones de los usuarios



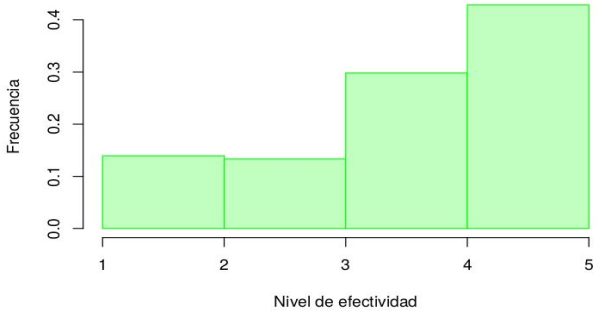


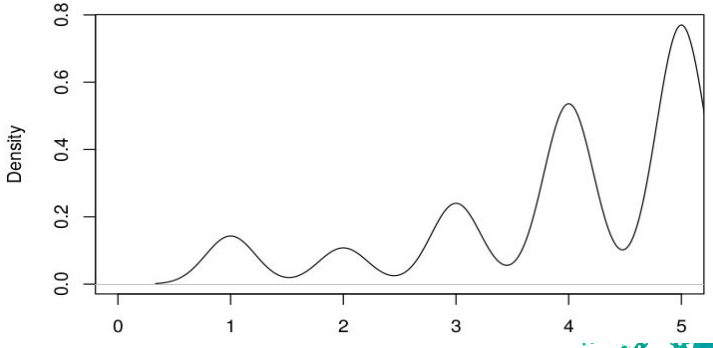


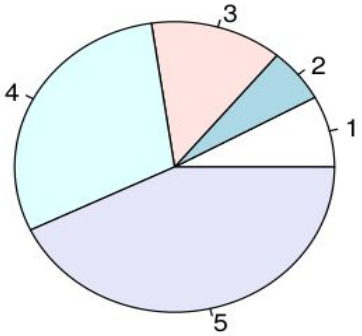


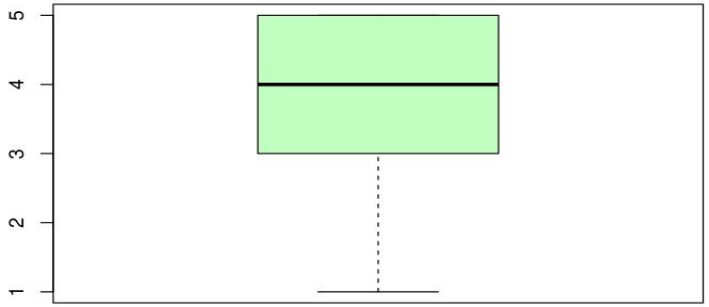
Desviación típica: 2.93

Efectividad



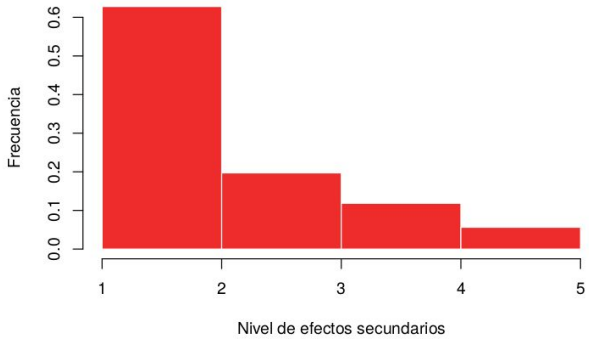


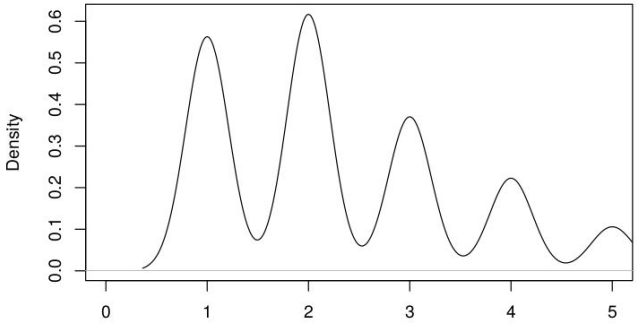


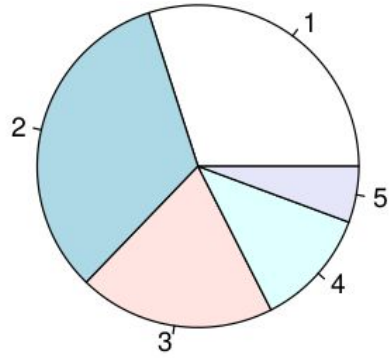


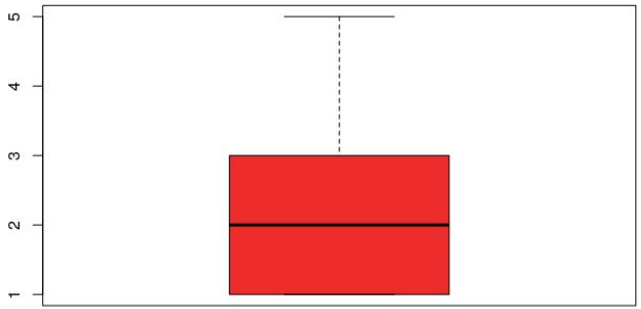
Desviación típica: 1.23

Efectos secundarios



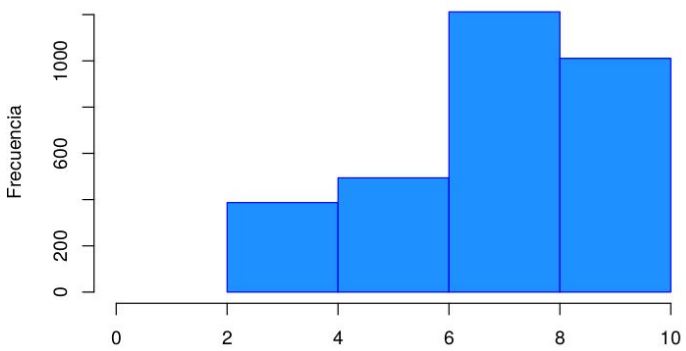


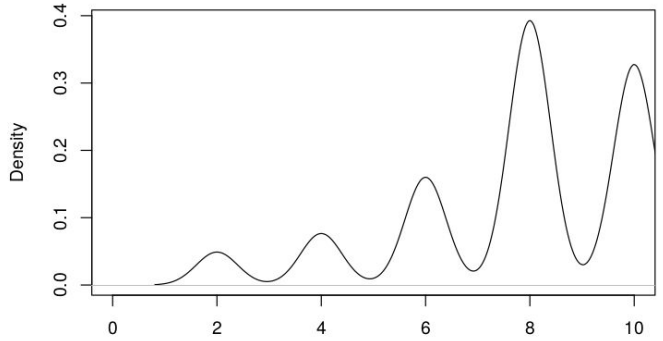


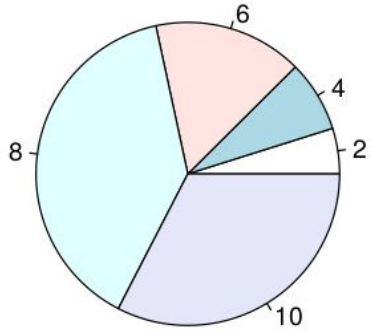


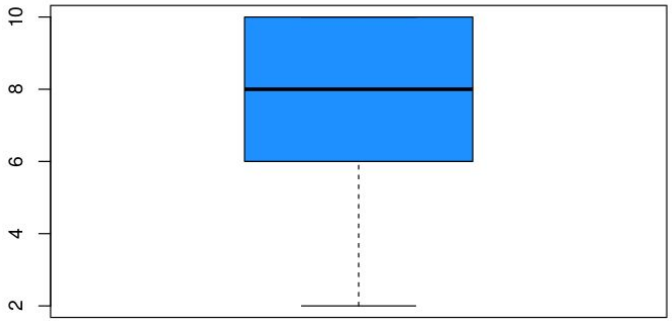
Desviación típica: 1.17

Valoración ponderada (efectividad/efectos secundarios)









Desviación típica: 2.19

Correlación entre variables

```
# Correlación lineal entre rating y effectivenessNumber
##                          rating effectivenessNumber
## rating                1.0000000           0.7498171
## effectivenessNumber 0.7498171           1.0000000
```

```
# Correlación lineal entre rating y sideEffectsNumber
##                          rating sideEffectsNumber
## rating                1.000000          -0.682939
## sideEffectsNumber -0.682939           1.000000
```

```
# Correlación lineal entre efectivenessNumber y weightedRating
##                      effectivenessNumber weightedRating
## effectivenessNumber           1.0000000       0.9237202
## weightedRating                0.9237202       1.0000000
```

```
# Correlación lineal entre sideEffectsNumber y weightedRating
##                      sideEffectsNumber weightedRating
## sideEffectsNumber            1.000000       -0.649161
## weightedRating              -0.649161        1.000000
```

```
# Correlación lineal entre efectivenessNumber y sideEffectsNumber
##                      sideEffectsNumber effectivenessNumber
## sideEffectsNumber            1.0000000          -0.3953789
## effectivenessNumber         -0.3953789           1.0000000
```

Efectividad ↑

Rating ↑

Efectos secundarios ↑

Rating ↓

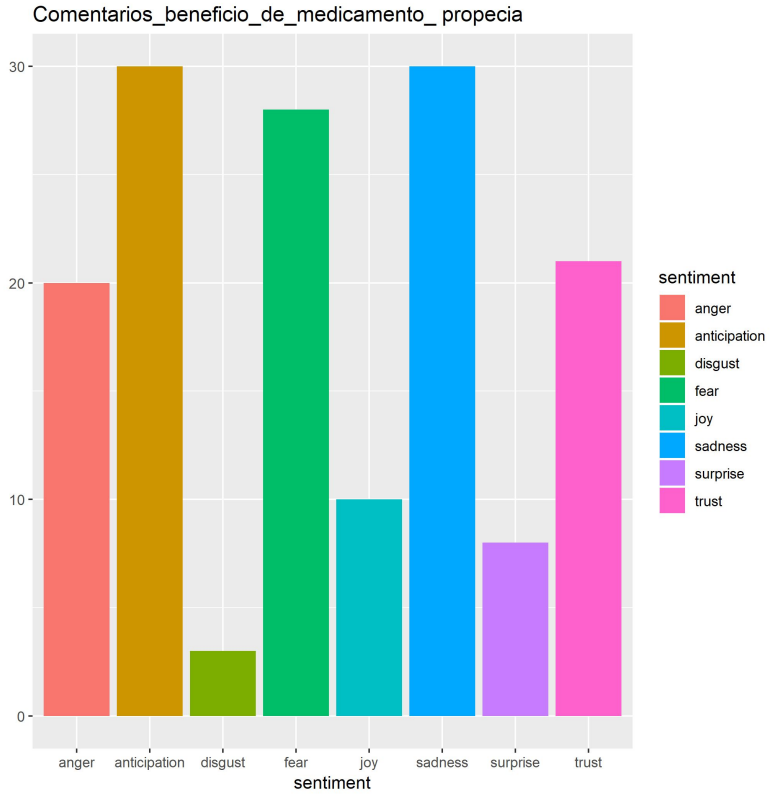
Efectos secundarios ↑

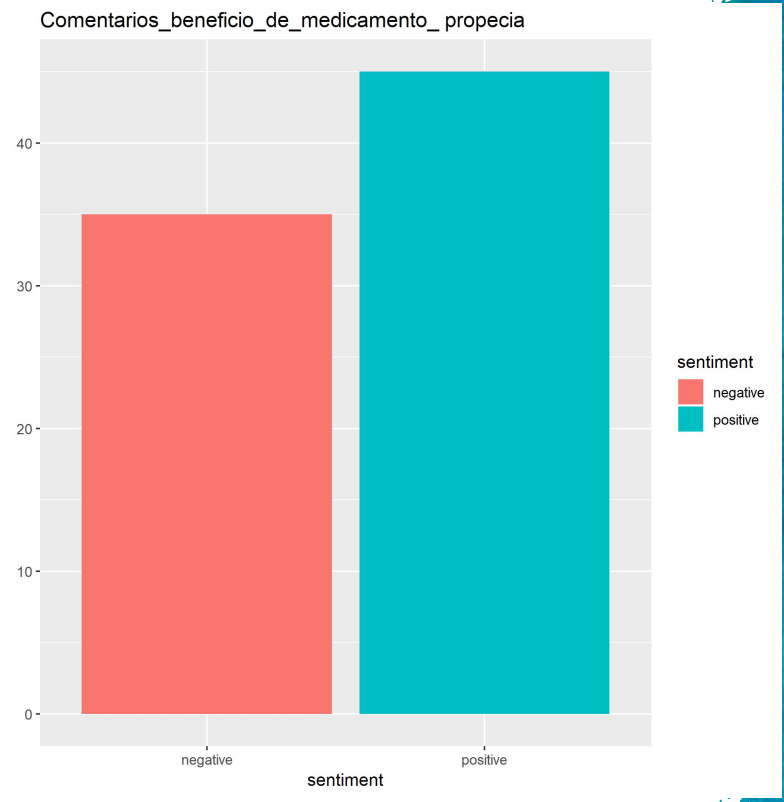
Efectividad ↓

Conclusiones *(análisis exploratorio)*

- Buena valoración de los usuarios sobre los medicamentos (Media 7).
- Alta efectividad de los medicamentos.
- Bajos efectos secundarios.
- Relación efectividad/efectos secundarios es positiva.
- Los medicamentos con mayor tasa de efectos secundarios son menos efectivas.

Análisis de sentimientos I





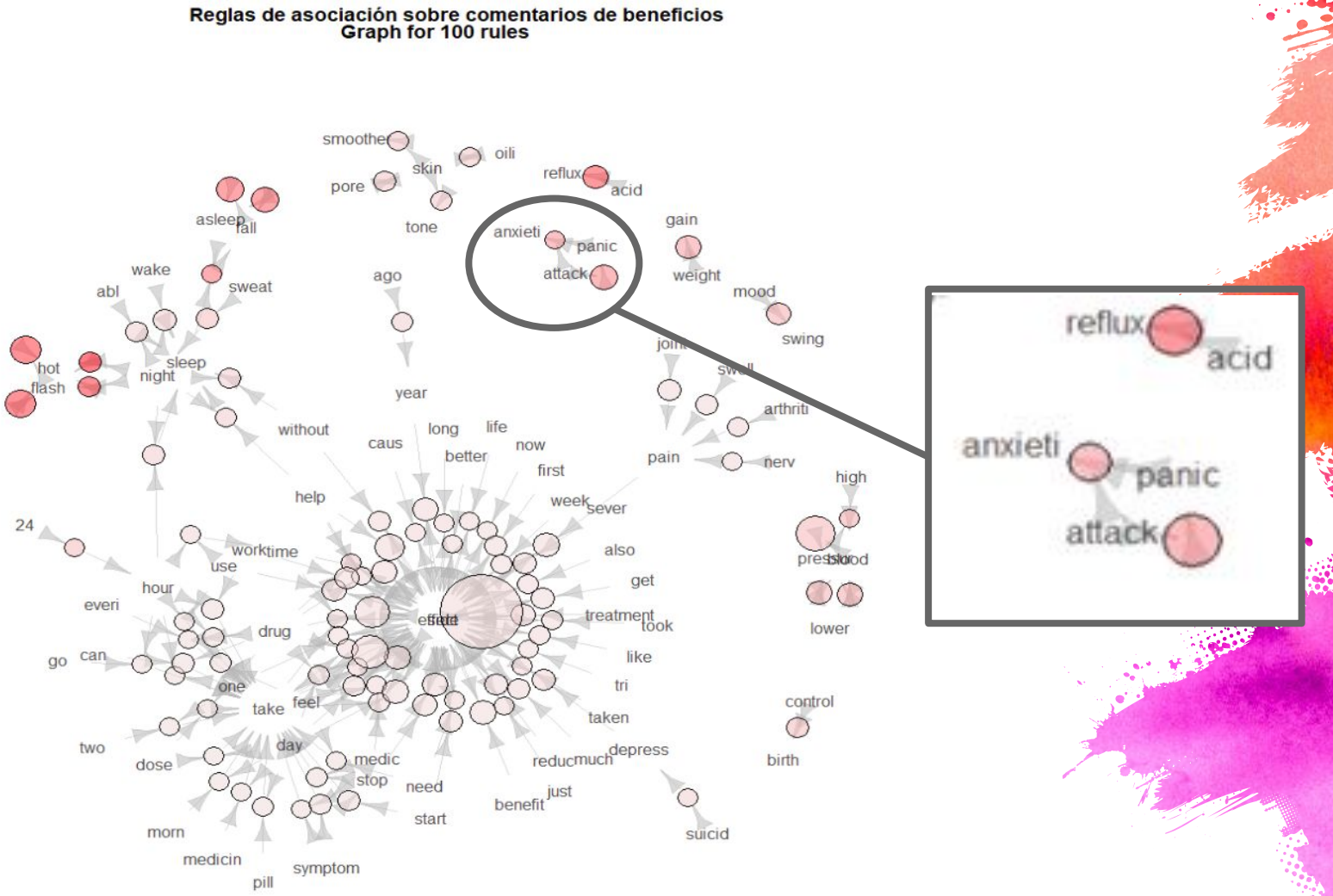
Análisis de sentimientos II

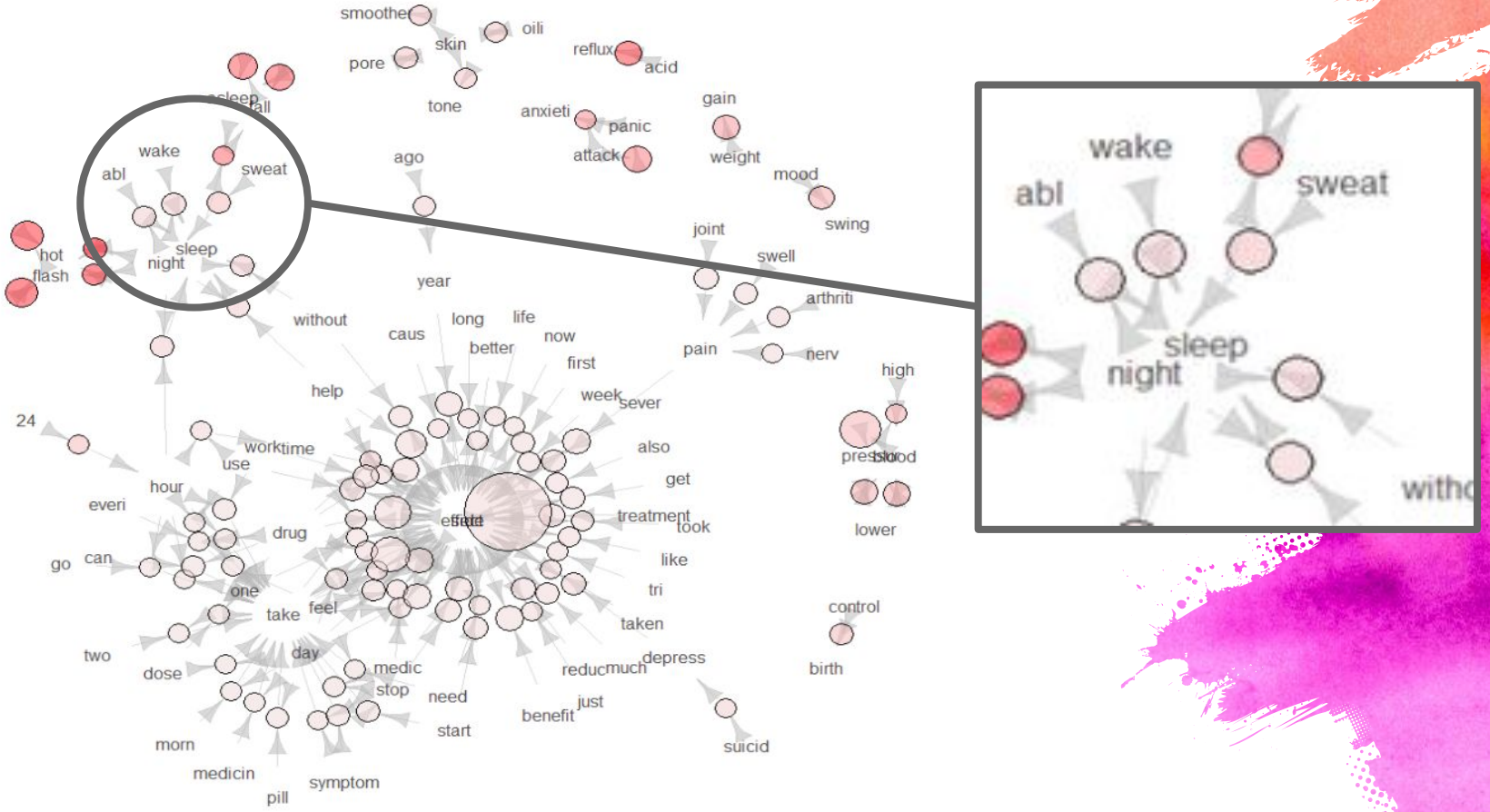
- Mayor presencia de negativismo que positivismo.
- Las personas suelen expresar más los aspectos negativos que los positivos.
- Trabajamos con un dataset altamente subjetivo.
- Las personas hablan sobre los problemas de salud que están experimentando y los efectos secundarios de los medicamentos → tener en cuenta en los resultados
- Todos los gráficos realizados en el apéndice de la documentación

MÉTODOS DESCRIPTIVOS

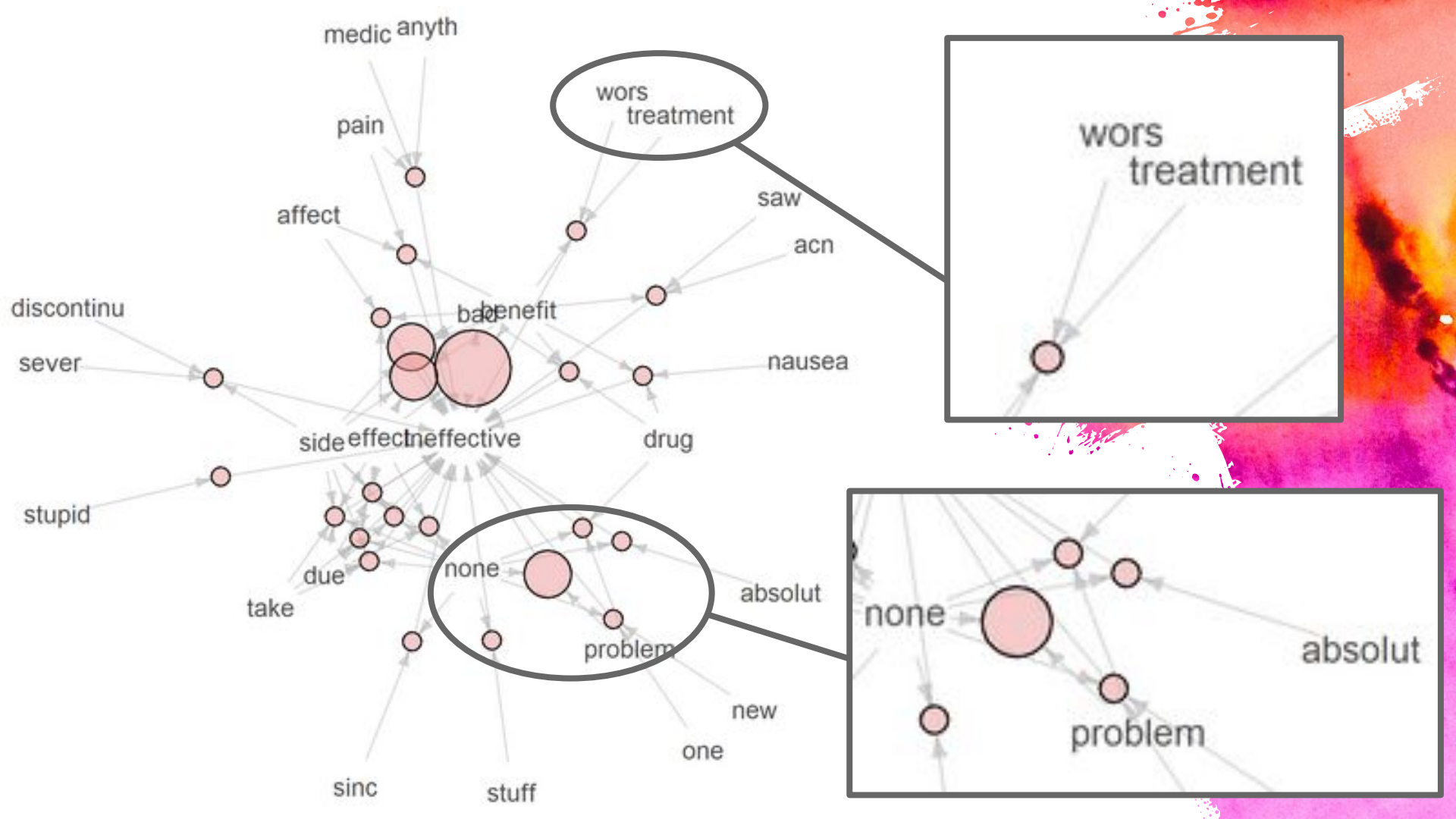
REGLAS DE ASOCIACIÓN

- Adaptación del conjunto de datos
- Algoritmo Apriori
  - ➢ Ajuste de soporte y confianza
- Guiado por el usuario

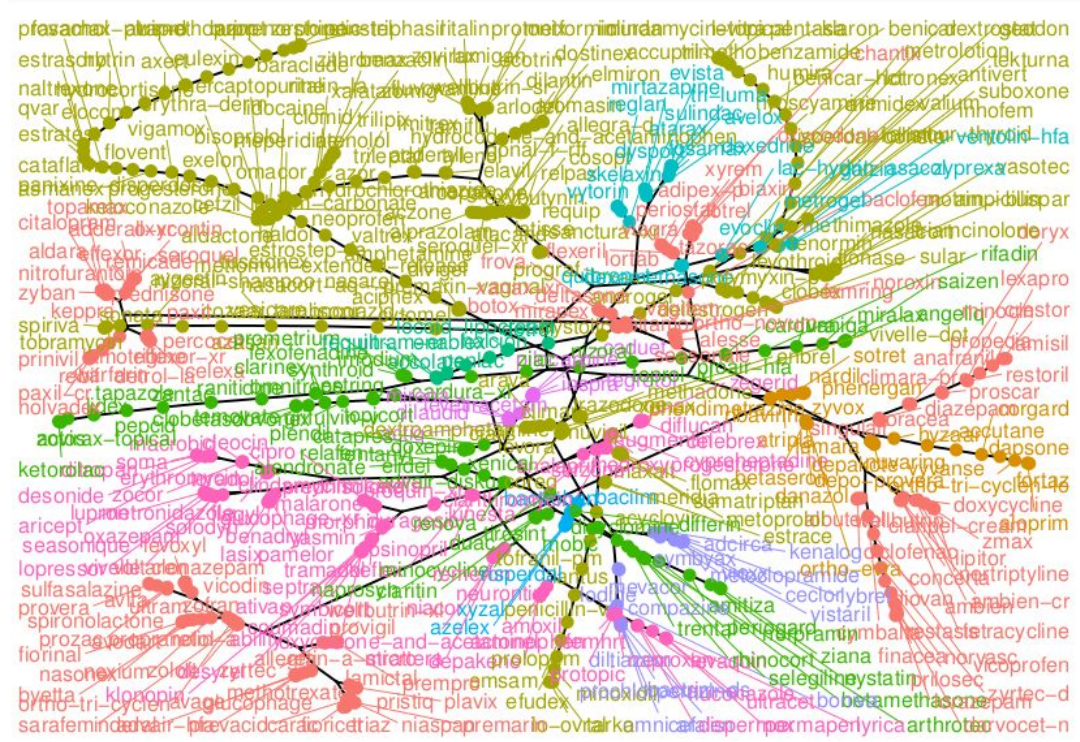

Reglas de asociación sobre comentarios de beneficios
Graph for 100 rules
smoothe
oili
skin
pore
reflux
acid
tone
asleep
fall
anxieti
panic
attack
gain
weight
wake
sweat
abl
ago
mood
swing
hot
flash
sleep
night
year
without
caus
long
life
better
now
first
pain
nerv
arthriti
high
help
week
sever
24
worktime
use
also
get
hour
everi
drug
treatment
took
lower
go
can
like
one
tri
take
feel
control
two
day
taken
depress
dose
medic
reduc
much
birth
stop
need
just
benefit
start
suicid
morn
medicin
symptom
pill


Reglas de asociación sobre comentarios de beneficios
Graph for 100 rules
smoothe
oili
skin
pore
reflux
acid
tone
anxieti
panic
attack
gain
weight
mood
swing
ago
wake
abl
sweat
sleep
night
hot
flash
year
joint
swell
arthriti
pain
nerv
high
without
caus
long
life
now
better
first
help
week
sever
24
worktime
use
also
get
hour
everi
drug
treatment
took
lower
go
can
like
one
tri
take
feel
control
two
taken
day
depress
dose
medic
reduc
much
birth
stop
need
just
benefit
start
suicid
morn
medicin
symptom
pill

```
 [1] "slow progress left ventricular dysfunct overt heart failur alon agent manag hypertens mangag congest heart
failur Highly-Effective"
 [2] "although type birth control con pros help cramp also effect prevent pregnanc along use condom well Highly-
Effective"
 [3] "use cramp bad leav ball bed least 2 day ponstel doesnt take pain away complet take edg much normal activ
possibl definit miracl medic Highly-Effective"
 [4] "acid reflux went away month just day drug heartburn start soon stop take began treatment 6 month pass stop
take heartburn came back seem wors even doctor said tri anoth 6 month treatment exact thing happen went three year
ask wasnt cure reflux doctor quit frank told wasnt cure treatment symptom told probabl rest life Marginally-
Effective"
 [5] "think lyrica start help pain sideeffect just sever continu Marginally-Effective"
 [6] "take propecia year start 20 year age hair continu thin notic signific benefit Ineffective"
 [7] "mood notic improv energi experi better sleep digest Highly-Effective"
 [8] "although drug origin prescrib depress help sleepless therefor continu take alon still occas problem fall
asleep find can combin melatonin valerian 12 year havent increas elavil dosag Considerably-Effective"
 [9] "simpli just work fast without nasti side effect ssri medicin wont go long stori panic attack sought help mani
year fact just work suppos without annoy side effect taken daili stop panic attack start think pattern lead find
can also use need along daili treatment Highly-Effective"
[10] "none noth help allergi just dryer pain version allergi new allergiesirrit develop top origin one stuff danger
side effect mani peopl experienc list packag class action lawsuit need happen Ineffective"
```


medic
anyth
pain
affect
wors
treatment
saw
acn
discontinu
bad
benefit
sever
nausea
side
effect
neffective
drug
stupid
due
take
none
absolut
problem
new
one
sinc
stuff
wors
treatment
none
absolut
problem

CLUSTERING

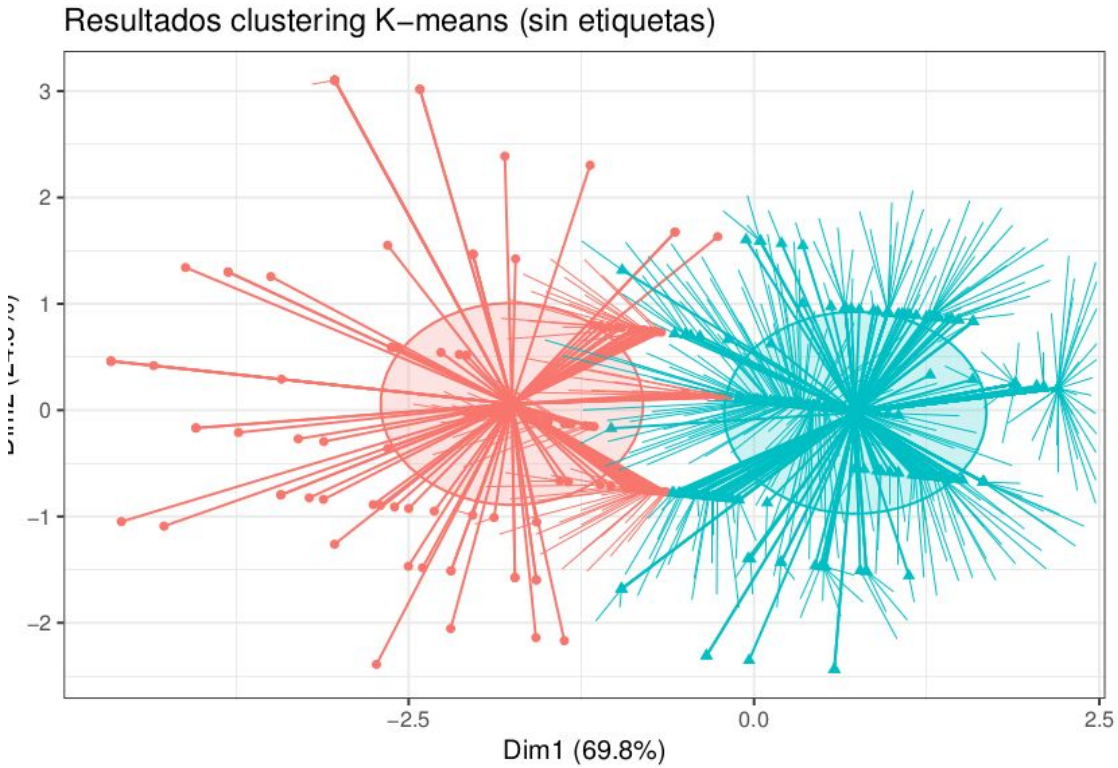


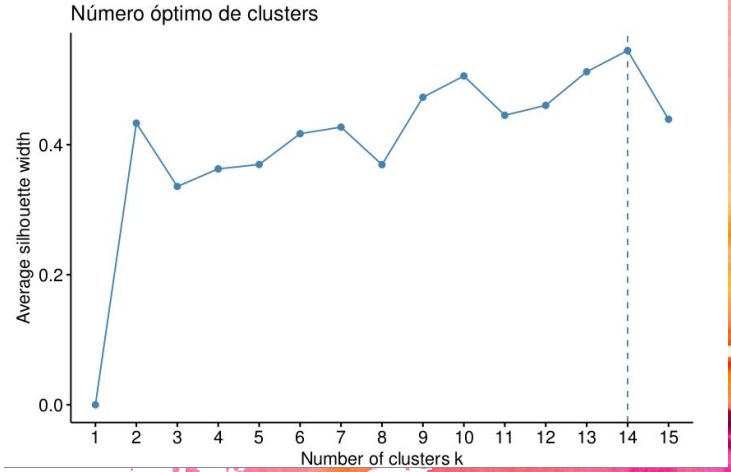
Métodos Clustering

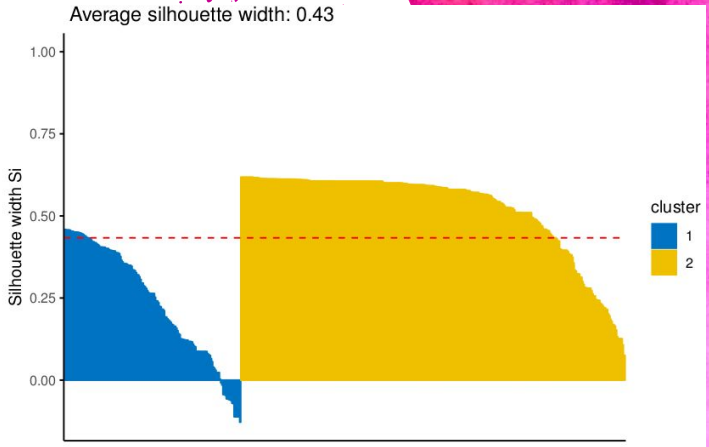
- K-means
- K-medoids
- Clustering Difuso
- Density Based

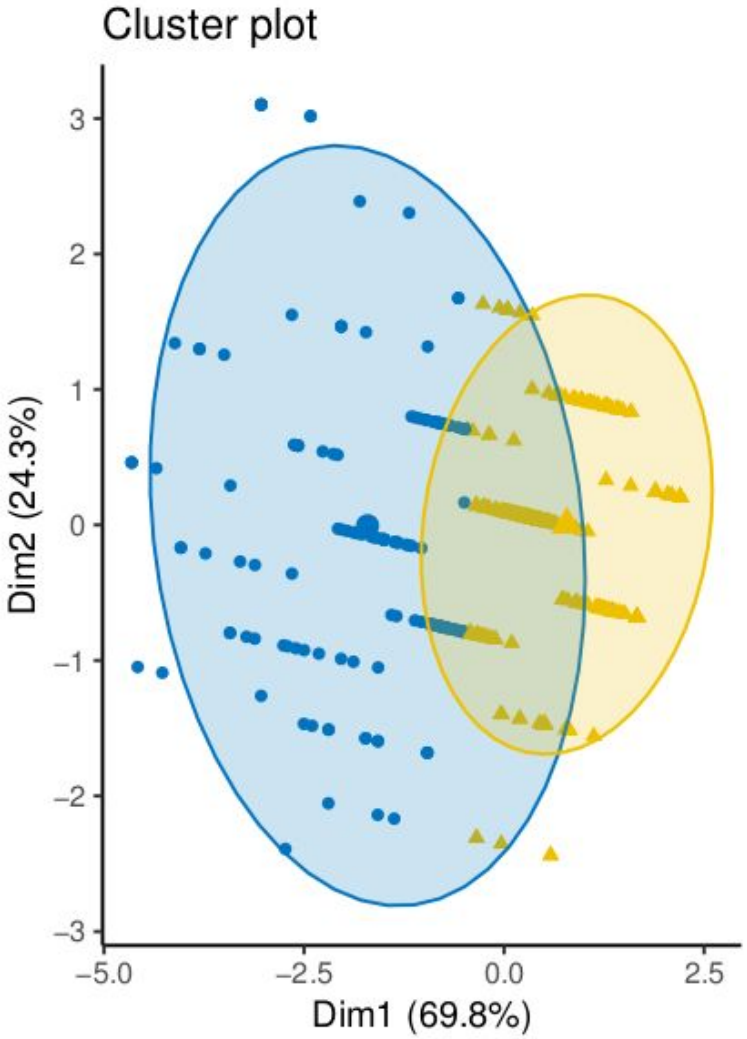
- Agrupamiento jerárquico
- Agrupamiento jerárquico y k-means

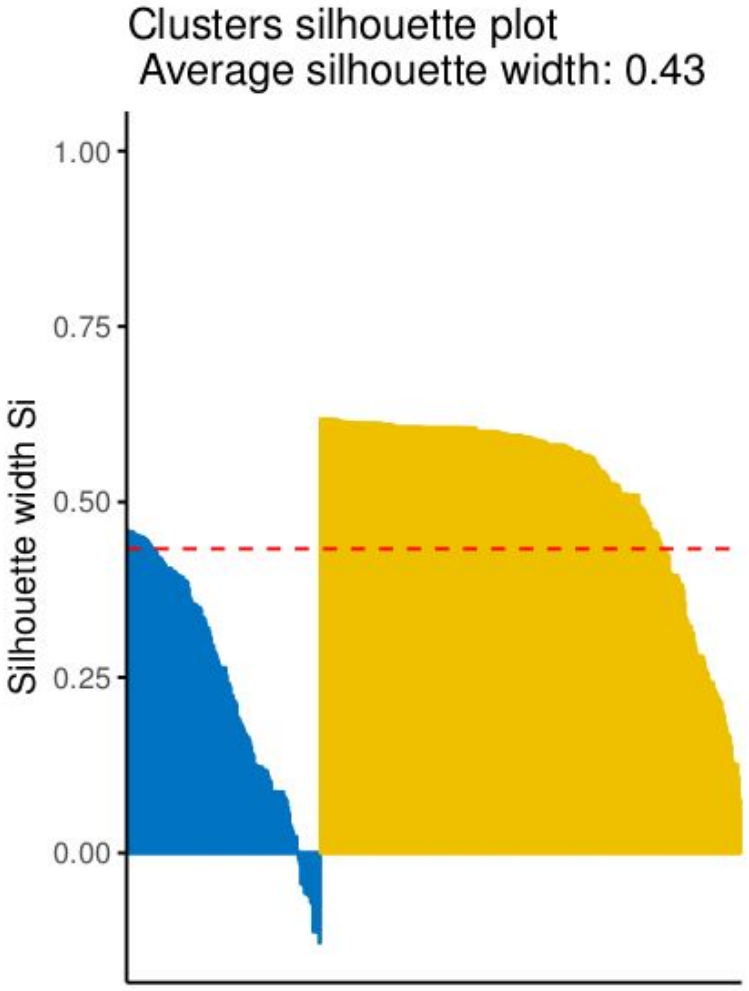
K-means


Resultados clustering K−means (sin etiquetas)
3
2
1
0
−1
−2
−2.5
0.0
2.5
Dim1 (69.8%)

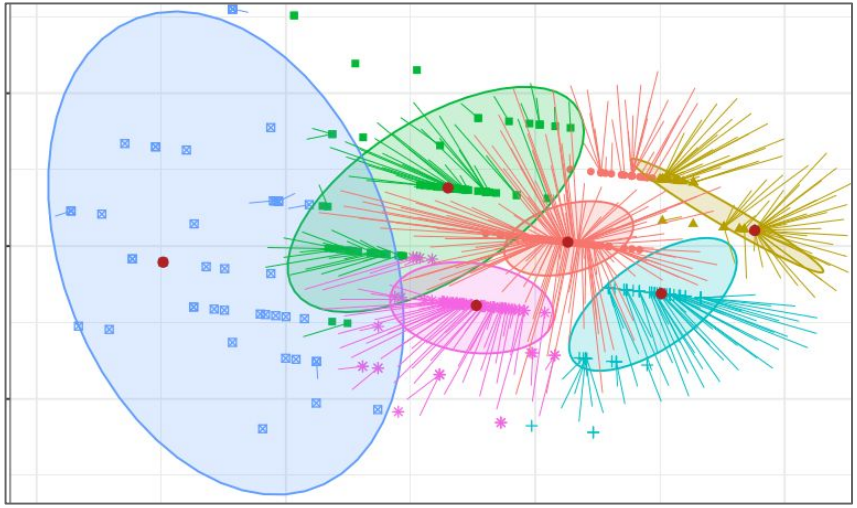

Número óptimo de clusters
Average silhouette width
0.0
0.2
0.4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of clusters k


Average silhouette width: 0.43
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2

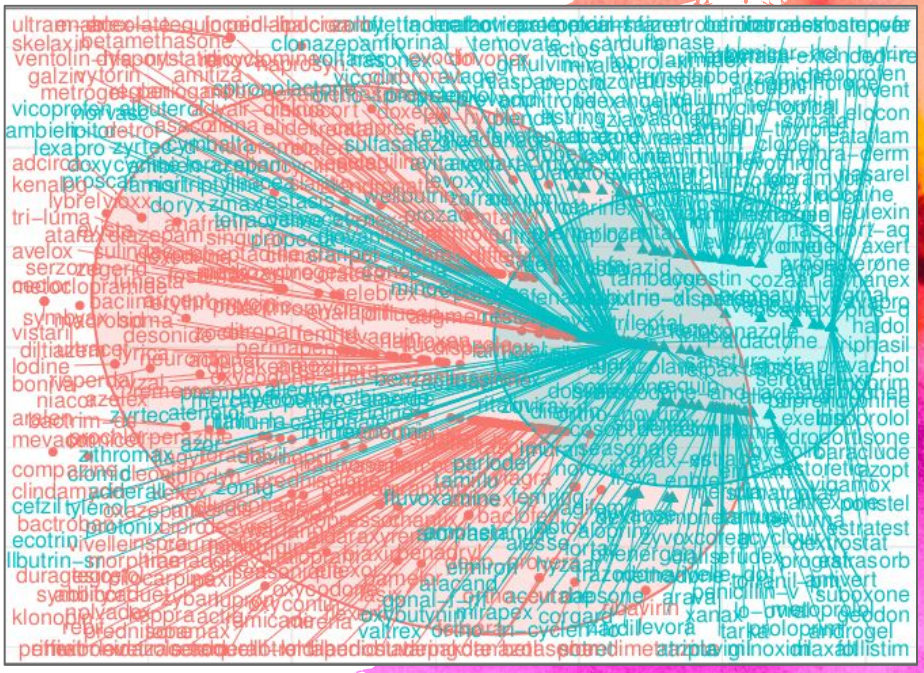

Cluster plot
Dim2 (24.3%)
Dim1 (69.8%)
3
2
1
0
−1
−2
−3
−5.0
−2.5
0.0
2.5

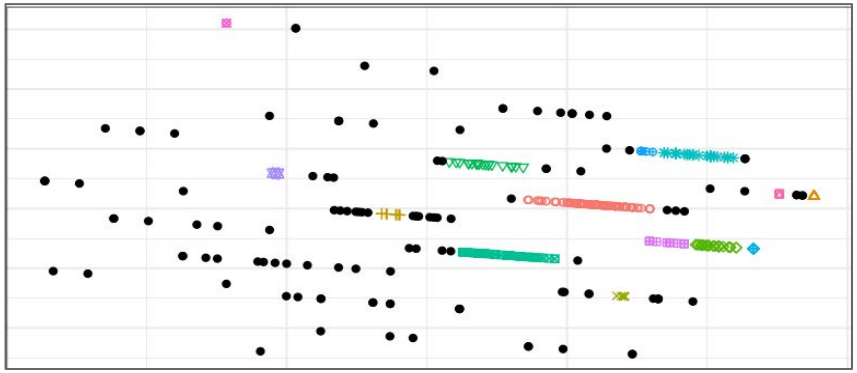

Clusters silhouette plot
Average silhouette width: 0.43
Silhouette width Si
1.00
0.75
0.50
0.25
0.00

K-medoids



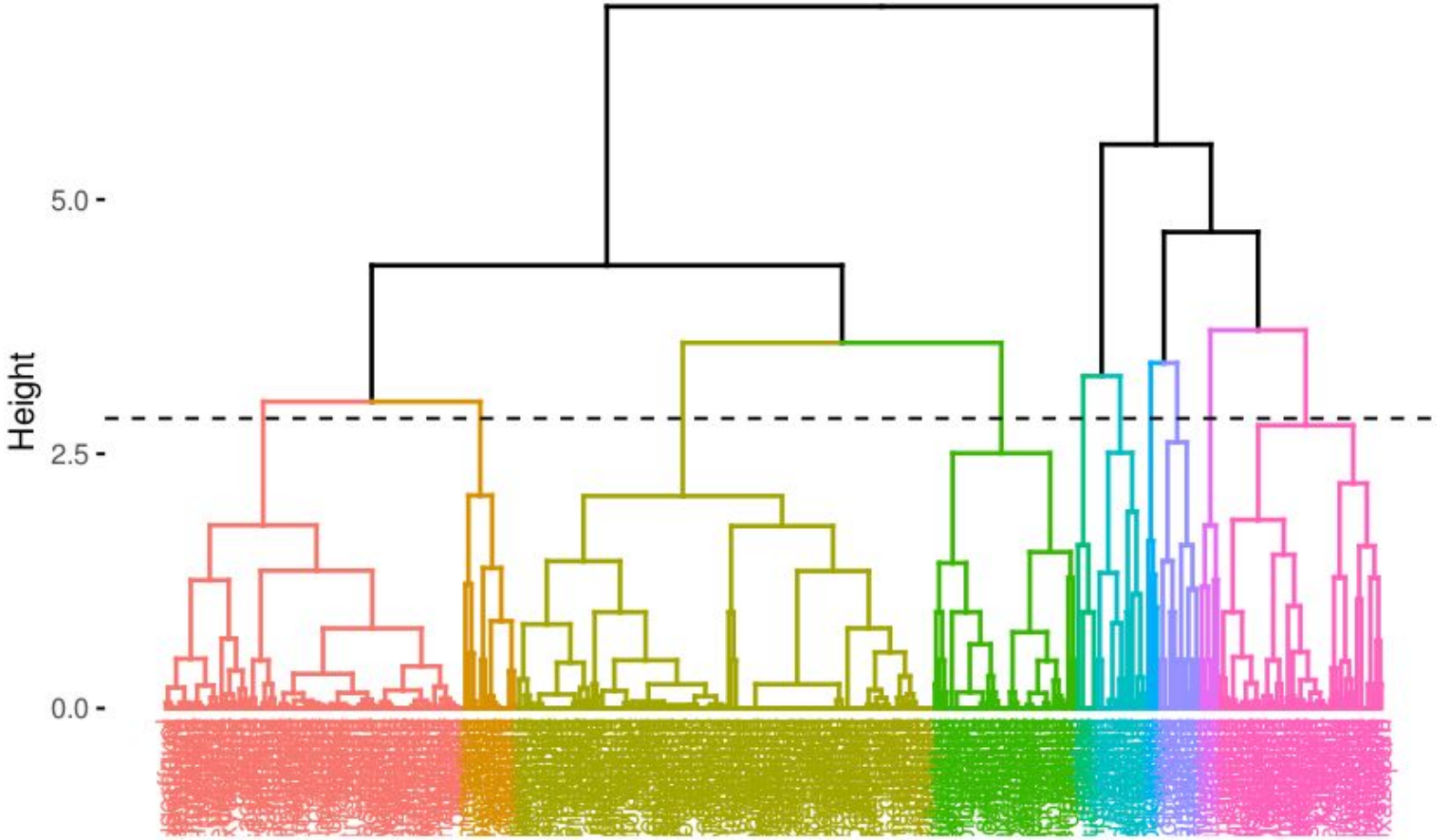
Fuzzy Clustering



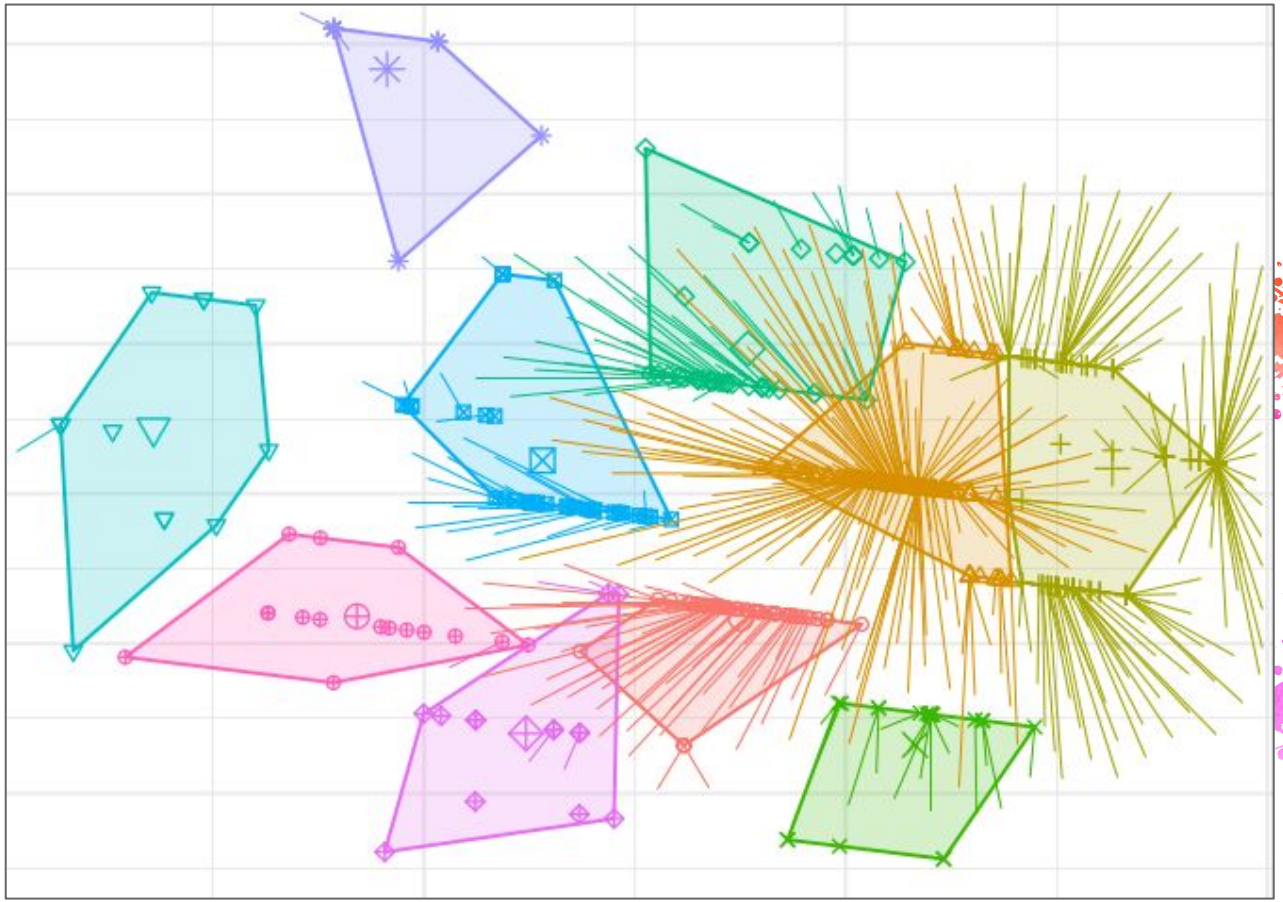


Density Based

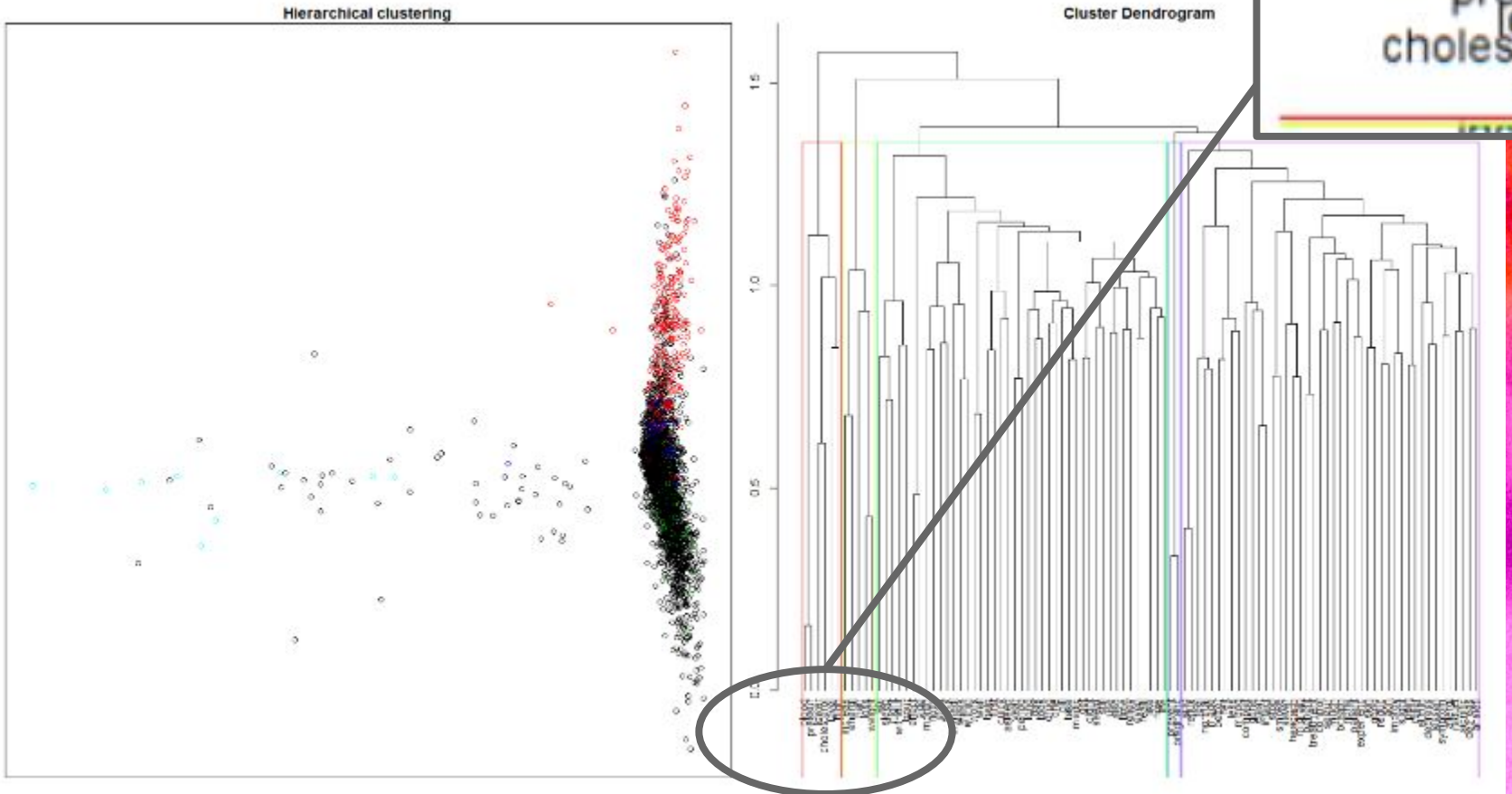
Agrupamiento Jerárquico



Agrupamiento Jerárquico + K-means



Y con los textos...



MODELOS PREDICTIVOS

NAIVE BAYES I

Matriz binaria determinando cada término en presente o ausente en cada documento. Dependiendo de su peso.

```
--------------------------------
**MATRIZ DE CONFUSIÓN TEST**
     real
pred   1   2   3   4   5
   1  44  13   7  12   6
   2   3   3   4   7   1
   3   6   8  21  29  27
   4  14  32  80 150 148
   5  14  20  45 111 229
Error de Test:  56.7698259187621  %
```

```
--------------------------------
**MATRIZ DE CONFUSIÓN TEST**
     real
pred   1   2   3   4   5
   1  25  14  16  26  21
   2   3   5   7  10  13
   3   4   9  17  24  20
   4  18  17  31  58  62
   5  31  31  86 191 295
Error de Test:  61.3152804642166  %
```

```
-------------------------------------
**MATRIZ DE CONFUSIÓN TRAIN**
     real
pred   1   2   3   4   5
   1 196  20  13  26  34
   2   1  55   1   7   9
   3   3  23 198  41  55
   4  25  60 131 665 359
   5  22  28  72 187 873
Error de Train:  35.985824742268  %
-------------------------------------
```

```
-------------------------------------
**MATRIZ DE CONFUSIÓN TRAIN**
     real
pred    1    2    3    4    5
   1  103    2   20   54   61
   2    1   58    6   21   37
   3   13   11  145   27   55
   4   25   33   48  360   91
   5  105   82  196  464 1086
Error de Train:  43.5567010309278  %
-------------------------------------
```

NAIVE BAYES II

```
--------------------------------
**MATRIZ DE CONFUSIÓN TEST**
    real
pred   0   1
   0 163 215
   1  77 579
Error de Test:  28.2398452611219  %
```

```
--------------------------------
**MATRIZ DE CONFUSIÓN TEST**
    real
pred   0   1
   0 116 108
   1 124 686
Error de Test:  22.4371373307544  %
```

```
-----------------------------------
**MATRIZ DE CONFUSIÓN TRAIN**
    real
pred    0    1
   0  531  587
   1  129 1857
Error de Train:  23.0670103092784  %
-----------------------------------
```

```
-----------------------------------
**MATRIZ DE CONFUSIÓN TRAIN**
    real
pred    0    1
   0  346  283
   1  314 2161
Error de Train:  19.2332474226804  %
-----------------------------------
```

KNN

- Problemas para trabajar con valores discretos.
- Trabajamos con texto.
- Queremos predecir ratingLabel, el valor preprocesado de rating.
- Es una técnica puramente predictiva, no nos da más información.
- Importante realizar primero la matriz de términos y después separarla en conjunto de entrenamiento y prueba.
- Importante establecer un valor de k adecuado.

KNN - BenefitsReview I

K=1

Predictions	Actual 0	Actual 1
0	137	277
1	88	429

[1] 60.79484

K=3

Predictions	Actual 0	Actual 1
0	138	281
1	87	425

[1] 60.47261

K=5

Predictions	Actual 0	Actual 1
0	139	278
1	86	428

[1] 60.90226

K=10

Predictions	Actual 0	Actual 1
0	162	363
1	63	343

[1] 54.24275

K=15

Predictions	Actual 0	Actual 1
0	180	441
1	45	265

[1] 47.79807

K=30

Predictions	Actual 0	Actual 1
0	196	596
1	11	128

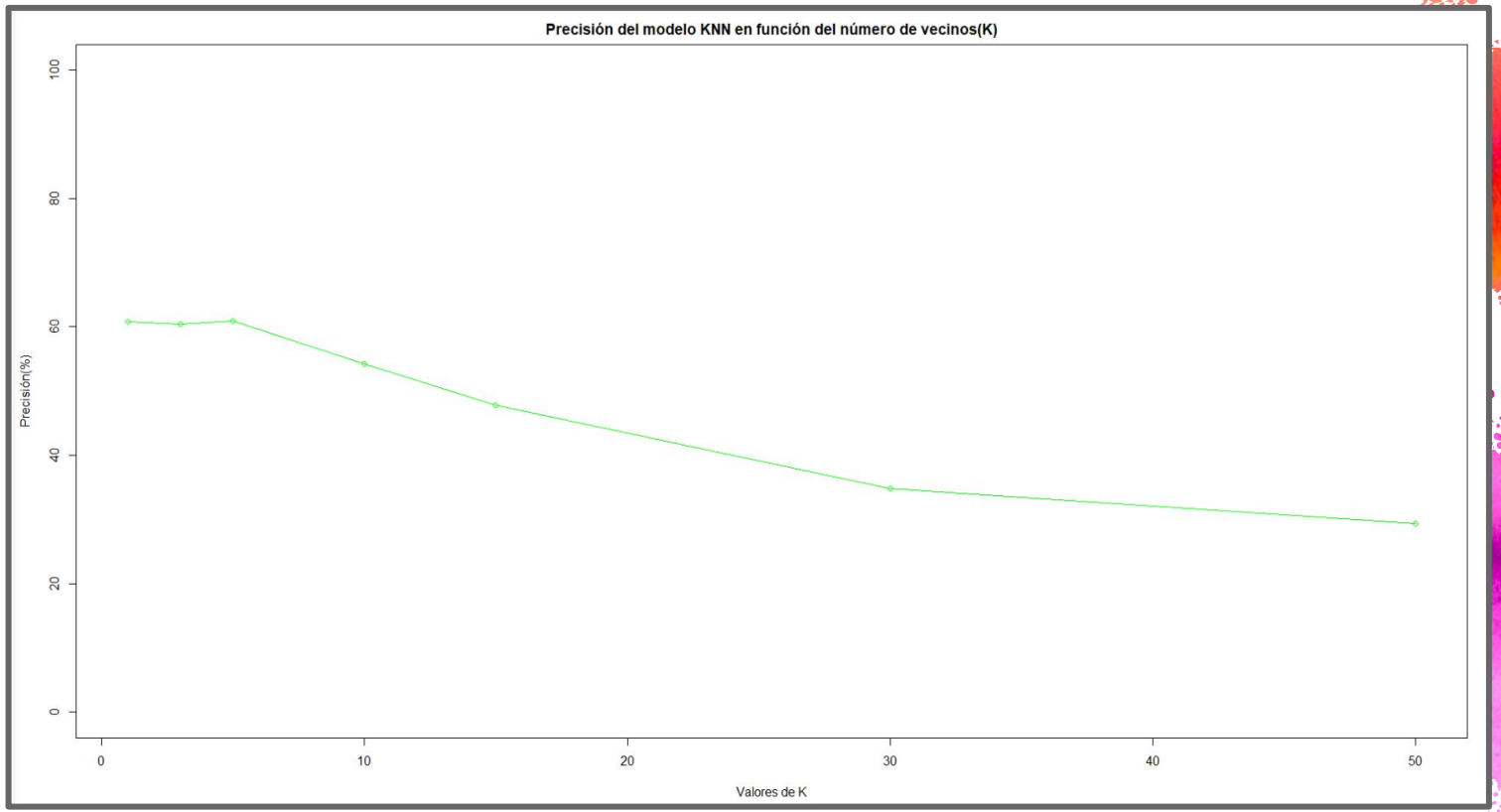
[1] 34.80129

K=50

Predictions	Actual 0	Actual 1
0	199	650
1	8	74

[1] 29.32331

KNN - BenefitsReview II



KNN - SideEffects

K=1

```
           Actual
Predictions   0   1
          0  38  54
          1 158 681
[1] 77.22879
```

K=3

```
           Actual
Predictions   0   1
          0  18  17
          1 178 718
[1] 79.05478
```

K=5

```
           Actual
Predictions   0   1
          0  11   6
          1 185 729
[1] 79.48443
```

K=15

```
           Actual
Predictions   0   1
          0   4   0
          1 192 735
[1] 79.37701
```

SVM

- Análisis no muy detallado dado que es computacionalmente costoso.
- No funciona correctamente con valores discretos.
- Tratamos de clasificar correctamente por el texto.
- Utilizamos un kernel de tipo radial (no lineal).
- **Problema**: La nube de puntos de los distintos comentarios está demasiado aglomerada.
- **Conclusión**: SVM no funciona bien a no ser que se tenga la misma estructura de palabras en el train y en el test. Esto hace que no pueda generar un buen modelo.

SVM

	SVM_LABEL <fctr>	SVM_PROB <dbl>
1	0	0.7585066
2	0	0.7585066
3	0	0.7585066
4	0	0.7585300
5	0	0.7585066
6	0	0.7585066
7	0	0.7585066
8	0	0.7585066
9	0	0.7585066
10	0	0.7585066
11	0	0.7585066
12	0	0.7585066
13	0	0.7585066
14	0	0.7585066
15	0	0.7585066
16	0	0.7585300
17	0	0.7585066
18	0	0.7585066
19	0	0.7585066
20	0	0.7585066
21	0	0.7585066
22	0	0.7585066
23	0	0.7585066
24	0	0.7585066
25	0	0.7585066
26	0	0.7585066
27	0	0.7585066
28	0	0.7585066
29	0	0.7585066
30	0	0.7585066
31	0	0.7585066
32	0	0.7585066
33	0	0.7585066
34	0	0.7585066
35	0	0.7585066
36	0	0.7585066
37	0	0.7585066
38	0	0.7585066
39	0	0.7585300
40	0	0.7585066
41	0	0.7585066
42	0	0.7585066
43	0	0.7585300
44	0	0.7585066
45	0	0.7585066
46	0	0.7585066
47	0	0.7585066
48	0	0.7585300

1-48 of 1,034 rows

REGRESIÓN

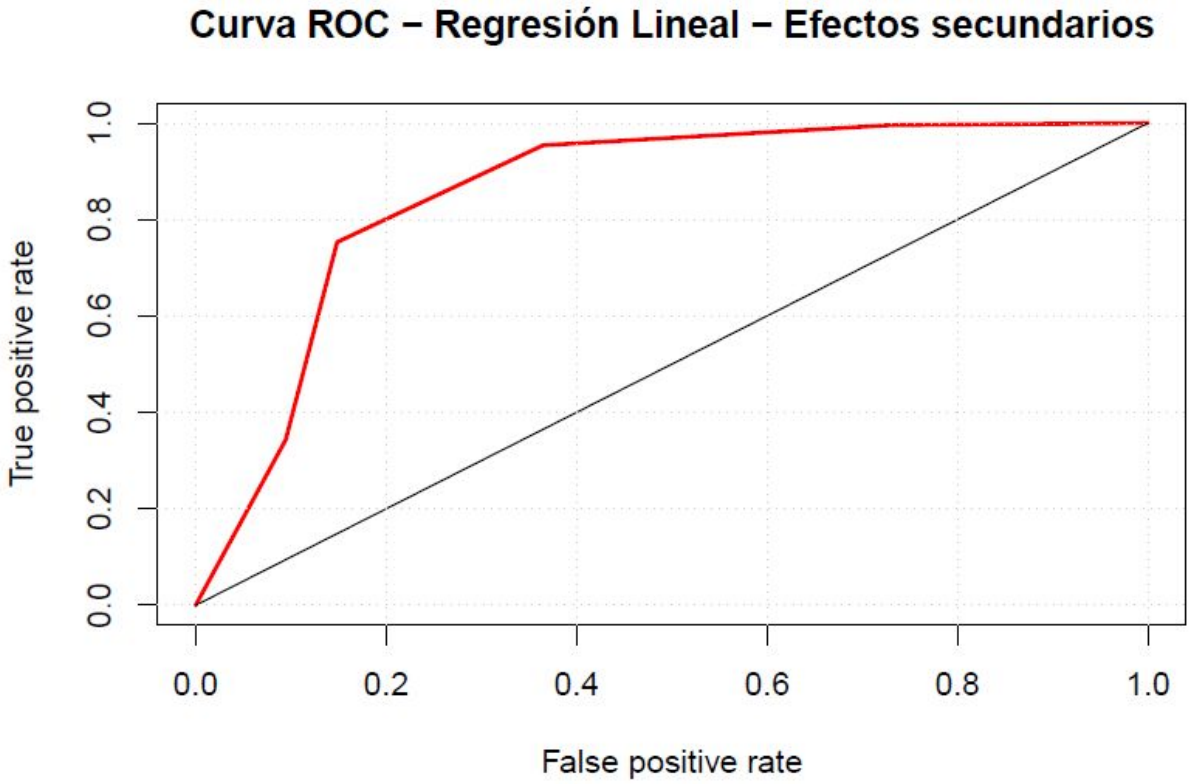
Técnicas	Error Test (%)	Error Train (%)
Regresión Lineal Simple (ratingLabel ~ sideEffectsInverse)	12.1405	12.350
Regresión Lineal Simple (ratingLabel ~ effectivenessNumber)	12.14058	10.1593
Regresión Lineal Múltiple (atributos numéricos)	9.265176	8.366534
Regresión Logística Simple (ratingLabel ~ effectivenessNumber)	12.14058	10.15936
Regresión Logística Simple (ratingLabel ~ sideEffectsInverse)	12.14058	12.3506
Regresión Logística Multivariable (atributos numéricos)	9.265176	8.366534
Ridge Regression (atributos numéricos)	10.46724	-
Regresión Polinomial (atributos numéricos)	8.945687	7.768924

Regresión Lineal I

```
##
## Call:
## lm(formula = ratingLabel ~ sideEffectsInverse, data = data_train_procesado)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0164 -0.0164 -0.0164  0.1615  0.6953
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.12683    0.05284    2.40   0.0167 *
## sideEffectsInverse  0.17792    0.01311   13.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3328 on 500 degrees of freedom
## Multiple R-squared:  0.2691, Adjusted R-squared:  0.2676
## F-statistic: 184.1 on 1 and 500 DF,  p-value: < 2.2e-16
```

```
##      real
## pred   0   1
##    0  47  11
##    1  27 228

## $Etrain
## [1] 12.3506
##
## $Etest
## [1] 12.14058
```

Regresión Lineal II



Regresión Lineal III

```
## Start:  AIC=-1296.57
## ratingLabel ~ sideEffectsInverse + effectivenessNumber
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               37.481 -1296.6
## - sideEffectsInverse     1    10.270 47.751 -1177.0
## - effectivenessNumber    1    17.903 55.384 -1102.6

##
## Call:
## lm(formula = ratingLabel ~ sideEffectsInverse + effectivenessNumber,
##     data = data_train_procesado)
##
## Coefficients:
##         (Intercept)   sideEffectsInverse  effectivenessNumber
##              -0.3362               0.1311               0.1628
```

```
##      real
## pred   0   1
##    0  50   5
##    1  24 234

## $Etrain
## [1] 8.366534
##
## $Etest
## [1] 9.265176
```

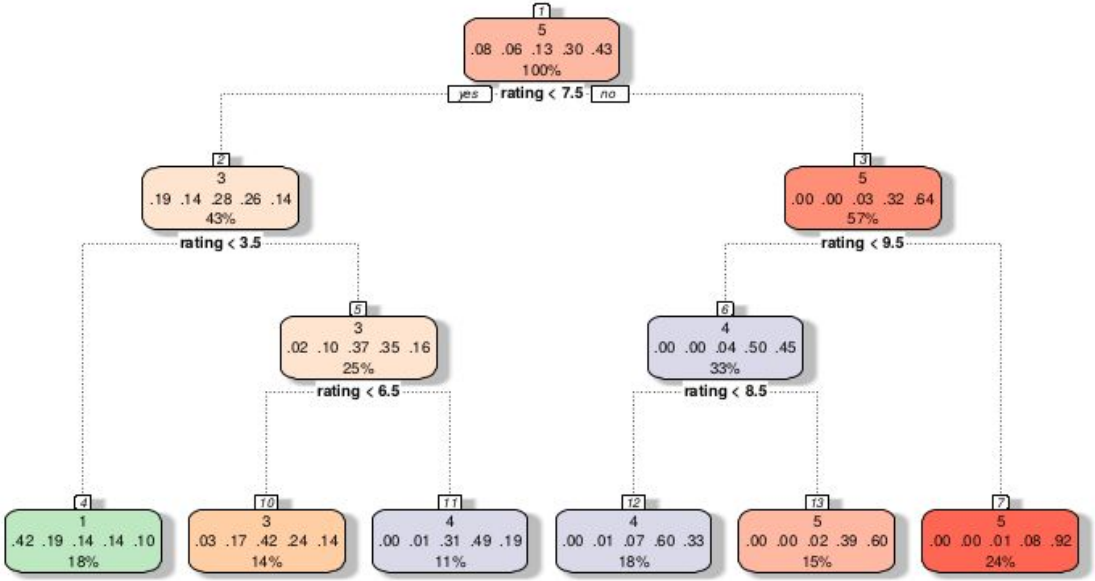
Regresión Polinomial

```
##
## Call:
## lm(formula = ratingLabel ~ poly(sideEffectsInverse, 2) + effectivenessNumber,
##     data = data_train_procesado)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06537 -0.06628  0.05724  0.09101  1.13669
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.19607    0.04170   4.702 3.34e-06 ***
## poly(sideEffectsInverse, 2)1  3.37481    0.27214  12.401  < 2e-16 ***
## poly(sideEffectsInverse, 2)2 -1.82668    0.26304  -6.945 1.19e-11 ***
## effectivenessNumber            0.15638    0.01012  15.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.262 on 498 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.5463
## F-statistic: 202.1 on 3 and 498 DF,  p-value: < 2.2e-16
```

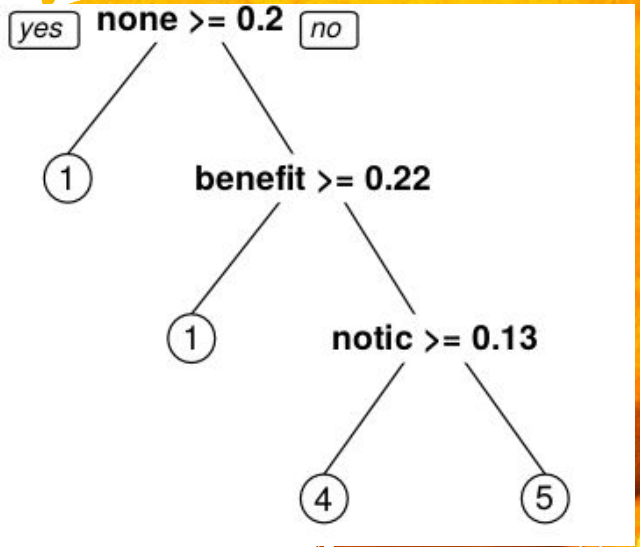
```
##      real
## pred   0   1
##    0  52   6
##    1  22 233

## $Etrain
## [1] 7.768924
##
## $Etest
## [1] 8.945687
```

ÁRBOLES DE DECISIÓN

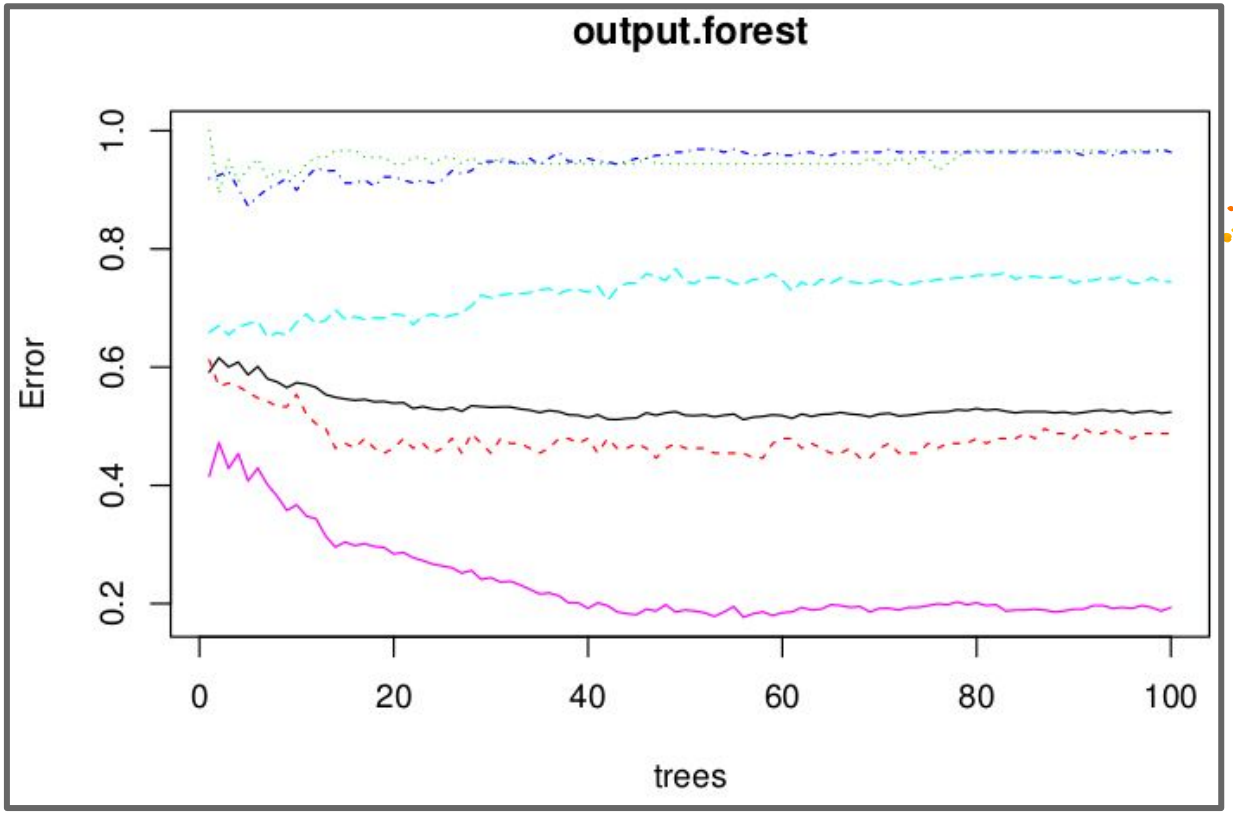
Árbol de decisión efectivenessNumber$rating



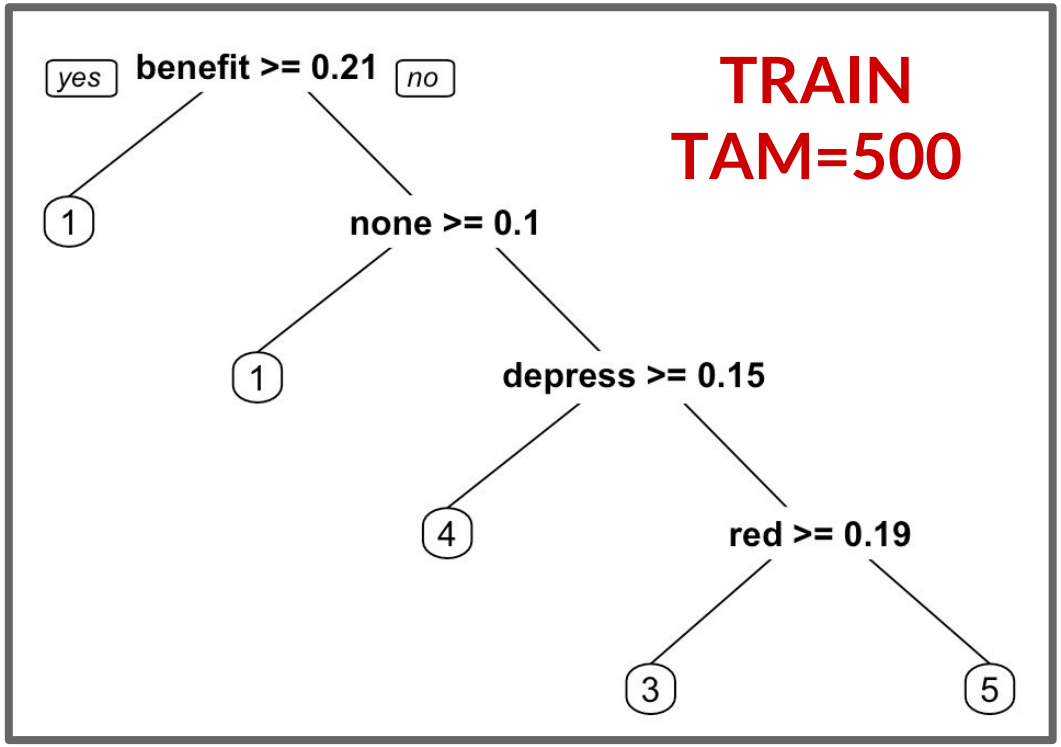
Árbol de decisión rating$benefitsReview

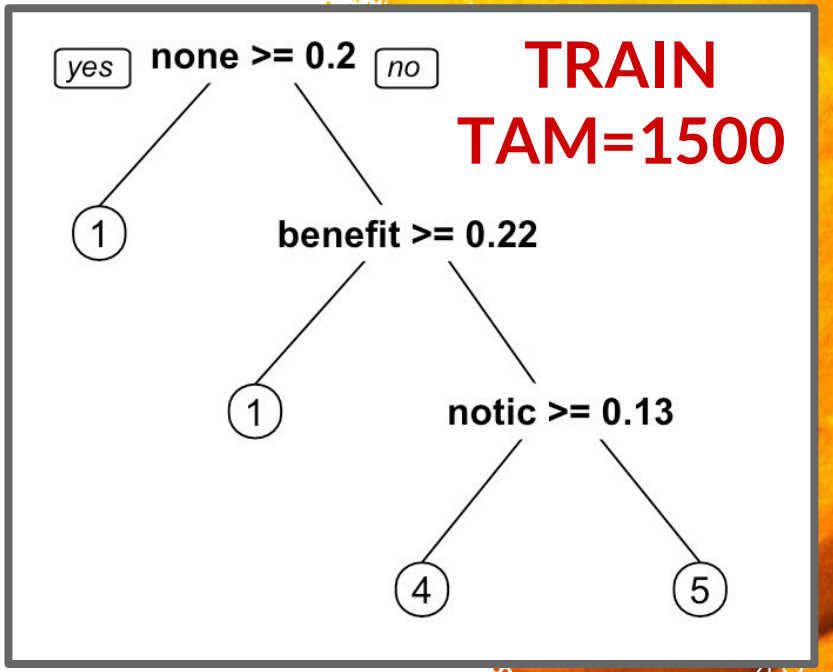


RANDOM FOREST



RANDOM FOREST- Texto





CONCLUSIONES

- Se parte de una base de datos no estructurada, imprescindible realizar un **preprocesamiento** a los datos.
- El **análisis exploratorio de la información** obtiene que los usuarios tienen buenas opiniones sobre los medicamentos, ya que en general son bastantes efectivos y no tienen efectos secundarios severos.
- El **análisis de sentimientos** concluye que aunque puntúen con una buena valoración al medicamento, los comentarios no son del todo positivos, ya que los usuarios relatan más los posibles efectos negativos que los positivos.
- Las **técnicas de tipo descriptivo** funcionan bien. En el caso de ***clustering***, muchos de los datos se solapan, dando lugar a resultados un poco confusos, pero, a medida que se aplican versiones/extensiones de dicha técnica, se comienzan a comprender mejor los datos y sus dependencias.
- Las **reglas de asociación** obtienen buenos resultados y coherentes. Es una de las técnicas que más información aporta.

- Las **técnicas de clasificación *(árboles de decisión)*** han generado unas aceptables predicciones, las cuales se han mejorado utilizando ***randomForest***. Aunque, se considera que estas técnicas no resultan adecuadas a nuestro problema: gran volumen de datos.

- **SVM** no funciona bien con texto a no ser que tengas las mismas palabras tanto en el test como en el train, cosa que no ocurre en nuestro caso.

- El **modelo de regresión** obtuvo unas predicciones de etiquetas muy buenas, obteniendo con ***regresión logística multivariable*** el mínimo error fuera de la muestra, permite estimar de manera aceptable la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa.

Importante...

- El papel que juega la naturaleza de los datos con los que trabajamos, es crucial en este tipo de problemas.
- Debemos tener en cuenta la componente subjetiva de nuestros datos.
- La predicción con datos derivados de un texto es mucho más difícil que la obtenida con variables numéricas.

Más información en: https://github.com/Gecofer/MII_TID_1819


¡GRACIAS!