



DECSAI

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada

Tratamiento Inteligente de Datos

Guión de Prácticas

Reglas de Asociación

Amparo Vila



FICHEROS DE DATOS

Iris.csv

Bankloan.csv

Preparación de los datos.

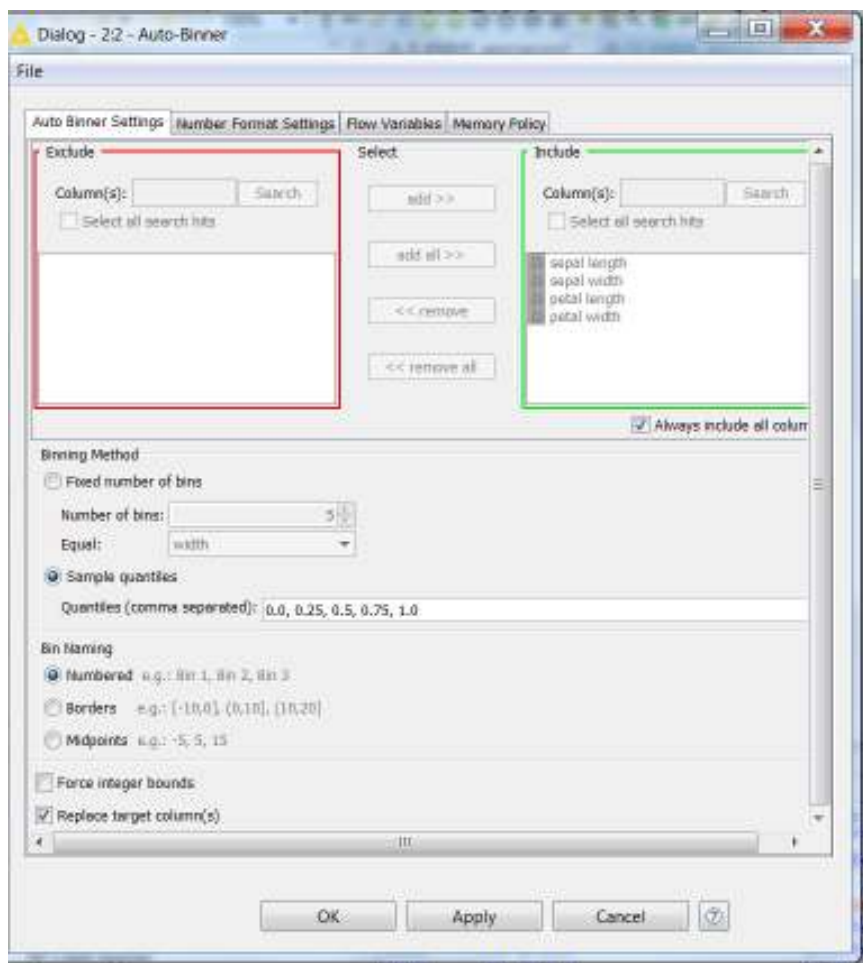
Ejemplo del Iris

Como se ha indicado en la teoría, los datos con los que se trabaja en la extracción de asociación deben ser discretos, realmente nominales, además estos datos deben estar en forma transaccional, de forma que realmente se trabaja con valores de cada atributo, no con atributos completos. Por ello antes de trabajar con reglas de asociación se debe realizar una labor de preparación de datos de forma que:

- Solamente se trabaje con atributos nominales
- Los valores de un atributo queden perfectamente identificados, como valores de este, es decir, que podamos distinguir entre valores atributos que tengan el mismo dominio nominal.

La preparación de los datos supone pues, un proceso de discretización de todos los atributos numéricos, y un proceso de renombrado de dominios que permita reconocer claramente los valores.

Utilizaremos los datos de Iris. Para discretizar los datos se utiliza un nodo de auto-Binner seleccionando los cuartiles como intervalos y generando un valor string que por ahora es “Bin 1”, “Bin 2” etc. Para todos los atributos considerados:
La configuración es:



Una vez ejecutado nos da la tabla:

The screenshot shows a window titled "Binned Data - 2:2 - Auto-Binner". It contains a table with 17 rows (Row0 to Row16) and 6 columns: Row ID, S sepal l..., S sepal ..., S petal l..., S petal ..., and S class. The data is as follows:

Row ID	S sepal l...	S sepal ...	S petal l...	S petal ...	S class
Row0	Bin 1	Bin 4	Bin 1	Bin 1	Iris-setosa
Row1	Bin 1	Bin 2	Bin 1	Bin 1	Iris-setosa
Row2	Bin 1	Bin 3	Bin 1	Bin 1	Iris-setosa
Row3	Bin 1	Bin 3	Bin 1	Bin 1	Iris-setosa
Row4	Bin 1	Bin 4	Bin 1	Bin 1	Iris-setosa
Row5	Bin 2	Bin 4	Bin 2	Bin 2	Iris-setosa
Row6	Bin 1	Bin 4	Bin 1	Bin 1	Iris-setosa
Row7	Bin 1	Bin 4	Bin 1	Bin 1	Iris-setosa
Row8	Bin 1	Bin 2	Bin 1	Bin 1	Iris-setosa
Row9	Bin 1	Bin 3	Bin 1	Bin 1	Iris-setosa
Row10	Bin 2	Bin 4	Bin 1	Bin 1	Iris-setosa
Row11	Bin 1	Bin 4	Bin 1	Bin 1	Iris-setosa
Row12	Bin 1	Bin 2	Bin 1	Bin 1	Iris-setosa
Row13	Bin 1	Bin 2	Bin 1	Bin 1	Iris-setosa
Row14	Bin 2	Bin 4	Bin 1	Bin 1	Iris-setosa
Row15	Bin 2	Bin 4	Bin 1	Bin 2	Iris-setosa
Row16	Bin 2	Bin 4	Bin 1	Bin 2	Iris-setosa

Como vemos ya tenemos los datos discretizados y en forma de cadena de caracteres, pero recordemos que las reglas de asociación tratan con valores de forma que no podremos saber en la salida de las mismas a que atributo se refiere “Bin 1” si no se indica el atributo de dicho valor. En definitiva, tendremos que cambiar los valores de los atributos para que se pueda interpretar adecuadamente la salida.

La idea del cambio es que el valor “Bin 1” del atributo Sepal Length se distinga del valor “Bin 1” del atributo Petal Width, por ejemplo. Para ello utilizaremos nodos del tipo String Replace (Dictionary), para cada uno de los atributos que se desee cambiar. Estos nodos necesitan un fichero adicional .csv, donde se indica cuales son los cambios que se desean hacer en los valores. La configuración para Petal Length, por ejemplo sería la siguiente:

The screenshot shows the configuration window for a 'String Replace (Dictionary)' node. The 'Target Column' is set to 'S petal length'. The 'Dictionary Location' is 'C:\Users\vila\Dropbox\Docencia\Data-Mining\Datos\csv-paraknime\Iris-PL.csv'. The 'Delimiter in Dictionary' is ';'. The 'Append new column' checkbox is unchecked.

Donde el fichero Iris-PL.csv lo hemos creado desde Excel con el siguiente contenido:

	A	B
1	PL-Bajo	Bin 1
2	PL-Medio	Bin 2
3	PL-Alto	Bin 3
4	PL-Muy Alto	Bin 4

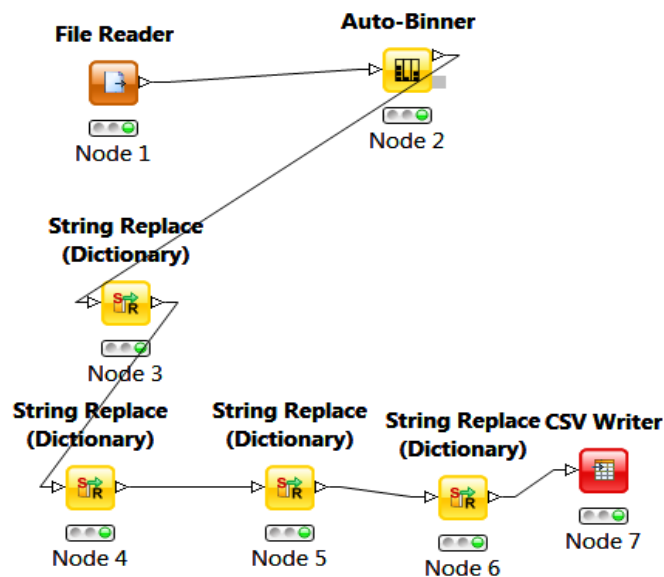
Una vez ejecutado este nodo obtenemos la tabla siguiente:

Input with amended column - 2:3 - String Replace (...)

Table "default" - Rows: 150					
	Spec	Columns: 5	Properties	Flow Variables	
Row ID	S sepal l...	S sepal ...	S petal l...	S petal ...	S class
Row0	Bin 1	Bin 4	PL-Bajo	Bin 1	Iris-setosa
Row1	Bin 1	Bin 2	PL-Bajo	Bin 1	Iris-setosa
Row2	Bin 1	Bin 3	PL-Bajo	Bin 1	Iris-setosa
Row3	Bin 1	Bin 3	PL-Bajo	Bin 1	Iris-setosa
Row4	Bin 1	Bin 4	PL-Bajo	Bin 1	Iris-setosa
Row5	Bin 2	Bin 4	PL-Medio	Bin 2	Iris-setosa
Row6	Bin 1	Bin 4	PL-Bajo	Bin 1	Iris-setosa
Row7	Bin 1	Bin 4	PL-Bajo	Bin 1	Iris-setosa
Row8	Bin 1	Bin 2	PL-Bajo	Bin 1	Iris-setosa
Row9	Bin 1	Bin 3	PL-Bajo	Bin 1	Iris-setosa
Row10	Bin 2	Bin 4	PL-Bajo	Bin 1	Iris-setosa
Row11	Bin 1	Bin 4	PL-Bajo	Bin 1	Iris-setosa
Row12	Bin 1	Bin 2	PL-Bajo	Bin 1	Iris-setosa

donde como puede verse ya se han cambiado los valores de este atributo.

El flujo total de Knime, incluyendo un nodo de escritura de datos sería:



Donde se escribe en un fichero la siguiente tabla:

Table "default" - Rows: 150					
		Spec - Columns: 5		Properties	Flow Variables
Row ID	\$ sepal l...	\$ sepal ...	\$ petal l...	\$ petal ...	\$ class
Row0	SL-Bajo	SW-Muy A...	PL-Bajo	PW-Bajo	Iris-setosa
Row1	SL-Bajo	SW-Medio	PL-Bajo	PW-Bajo	Iris-setosa
Row2	SL-Bajo	SW-Alto	PL-Bajo	PW-Bajo	Iris-setosa
Row3	SL-Bajo	SW-Alto	PL-Bajo	PW-Bajo	Iris-setosa
Row4	SL-Bajo	SW-Muy A...	PL-Bajo	PW-Bajo	Iris-setosa
Row5	SL-Medio	SW-Muy A...	PL-Medio	PW-Medio	Iris-setosa
Row6	SL-Bajo	SW-Muy A...	PL-Bajo	PW-Bajo	Iris-setosa
Row7	SL-Bajo	SW-Muy A...	PL-Bajo	PW-Bajo	Iris-setosa
Row8	SL-Bajo	SW-Medio	PL-Bajo	PW-Bajo	Iris-setosa
Row9	SL-Bajo	SW-Alto	PL-Bajo	PW-Bajo	Iris-setosa

Con esta preparación de los datos ya estamos en condiciones de hacer un análisis utilizando reglas de asociación.

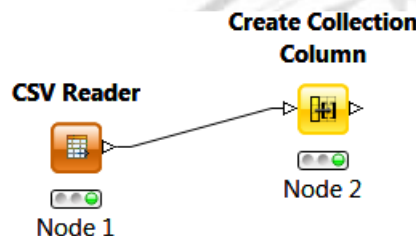
Extracción de reglas de asociación.

Ejemplo del Iris

En Knime se puede trabajar don reglas de asociación desde Weka o bien con los nodos propios de sistema. Vamos a ver las dos opciones

Utilización de los nodos de propios del sistema.

Para utilizar los nodos propios de Knime es necesario construir un atributo que transforme el data set en transaccional, este atributo se denomina de tipo “colección” y que, en definitiva, es un conjunto que se construye para cada fila con los valores de varias columnas. Para nuestro ejemplo



Incluyendo en la configuración las columnas que se quieran utilizar para extraer reglas de asociación (en este caso todas). Se obtiene la siguiente salida:

(...) AggregatedValues
[SL-Bajo,SW-Muy Alto,PL-Bajo,...]
[SL-Bajo,SW-Medio,PL-Bajo,...]
[SL-Bajo,SW-Alto,PL-Bajo,...]
[SL-Bajo,SW-Alto,PL-Bajo,...]
[SL-Bajo,SW-Muy Alto,PL-Bajo,...]
[SL-Medio,SW-Muy Alto,PL-Medio,...]
[SL-Bajo,SW-Muy Alto,PL-Bajo,...]
[SL-Bajo,SW-Muy Alto,PL-Bajo,...]
[SL-Bajo,SW-Medio,PL-Bajo,...]
[SL-Bajo,SW-Alto,PL-Bajo,...]
[SL-Medio,SW-Muy Alto,PL-Bajo,...]
[SL-Bajo,SW-Muy Alto,PL-Bajo,...]
[SL-Bajo,SW-Medio,PL-Bajo,...]
[SL-Bajo,SW-Medio,PL-Bajo,...]
[SL-Medio,SW-Muy Alto,PL-Bajo,...]

Ejecutando ahora el nodo de “Association Rule Learner” , tenermos como configuración:

Options **Flow Variables** Memory Policy

Itemset Mining

Column containing transactions: (...) AggregatedValues

Minimum support (0-1): 0,1

Underlying data structure: ARRAY

Output

Itemset type: FREE

Maximal itemset length: 10

Association Rules

☒ Output association rules

Minimum confidence: 0,9

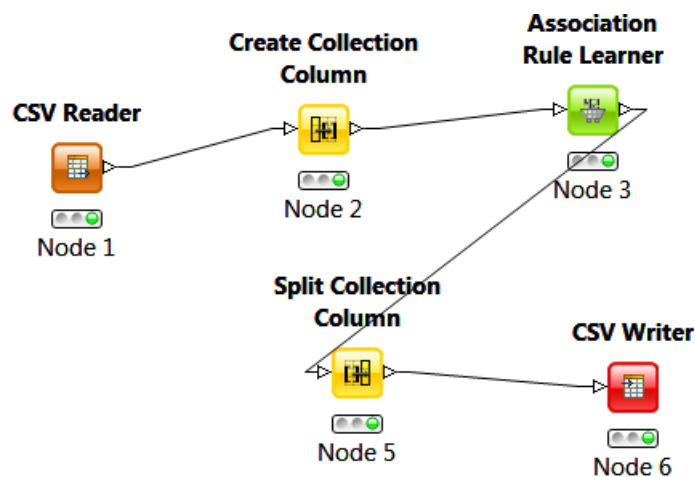
Podemos ver que se elige el tipo de itemset, la medida de bondad podemos pedir como salida reglas de asociación o itemsets etc.. La salida del nodo es la siguiente tabla:

Row ID	D Support	D Confid...	D Lift	S Conse...	S implies	(...) Items
rule0	0.1	1	3	Iris-setosa	<---	[SW-Muy Alto,SL-Bajo,PL-Bajo]
rule1	0.113	1	3	Iris-setosa	<---	[SW-Muy Alto,SL-Bajo]
rule2	0.2	0.968	3.299	PL-Bajo	<---	[SL-Bajo,Iris-setosa,PW-Bajo]
rule3	0.2	0.909	3.326	PW-Bajo	<---	[SL-Bajo,Iris-setosa,PL-Bajo]
rule4	0.2	1	3	Iris-setosa	<---	[SL-Bajo,PW-Bajo,PL-Bajo]
rule5	0.22	0.917	3.125	PL-Bajo	<---	[SL-Bajo,Iris-setosa]
rule6	0.22	1	3	Iris-setosa	<---	[SL-Bajo,PL-Bajo]
rule7	0.207	1	3	Iris-setosa	<---	[SL-Bajo,PW-Bajo]
rule8	0.133	1	3	Iris-setosa	<---	[SW-Muy Alto,PW-Bajo,PL-Bajo]
rule9	0.173	1	3	Iris-setosa	<---	[SW-Muy Alto,PL-Bajo]
rule10	0.153	1	3	Iris-setosa	<---	[SW-Muy Alto,PW-Bajo]
rule11	0.253	0.927	3.16	PL-Bajo	<---	[Iris-setosa,PW-Bajo]
rule12	0.253	1	3	Iris-setosa	<---	[PW-Bajo,PL-Bajo]
rule13	0.293	1	3	Iris-setosa	<---	[PL-Bajo]
rule14	0.273	1	3	Iris-setosa	<---	[PW-Bajo]

Esta tabla se puede guardar como .csv y analizar por ejemplo con Excel, o incluso separar los consecuentes y luego guardarla y realizar análisis posteriores con Excel o SPSS, o incluso con el propio Knime. Para ello utilizaríamos el nodo “Split collection Column”, que nos da la salida:

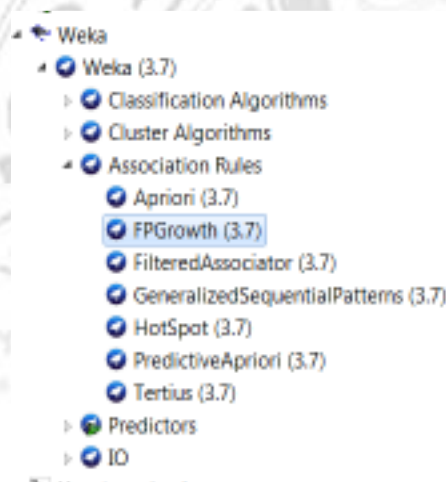
Row ID	D Support	D Confid...	D Lift	S Conse...	S implies	S Split V...	S Split V...	S Split V...
rule0	0.1	1	3	Iris-setosa	<---	SW-Muy A...	SL-Bajo	PL-Bajo
rule1	0.113	1	3	Iris-setosa	<---	SW-Muy A...	SL-Bajo	?
rule2	0.2	0.968	3.299	PL-Bajo	<---	SL-Bajo	Iris-setosa	PW-Bajo
rule3	0.2	0.909	3.326	PW-Bajo	<---	SL-Bajo	Iris-setosa	PL-Bajo
rule4	0.2	1	3	Iris-setosa	<---	SL-Bajo	PW-Bajo	PL-Bajo
rule5	0.22	0.917	3.125	PL-Bajo	<---	SL-Bajo	Iris-setosa	?
rule6	0.22	1	3	Iris-setosa	<---	SL-Bajo	PL-Bajo	?
rule7	0.207	1	3	Iris-setosa	<---	SL-Bajo	PW-Bajo	?
rule8	0.133	1	3	Iris-setosa	<---	SW-Muy A...	PW-Bajo	PL-Bajo
rule9	0.173	1	3	Iris-setosa	<---	SW-Muy A...	PL-Bajo	?
rule10	0.153	1	3	Iris-setosa	<---	SW-Muy A...	PW-Bajo	?
rule11	0.253	0.927	3.16	PL-Bajo	<---	Iris-setosa	PW-Bajo	?

El flujo total para obtener la tabla de salida sería:



Utilización de los nodos de Weka.

En el ámbito de las reglas de asociación e ítem-sets frecuentes Weka ofrece una amplia variedad



si bien casi siempre se enfoca a buscar otra forma de clasificación. Por lo que nos centraremos en el nodo “A priori”, que implementa una forma iterativa del algoritmo “A priori” que reduce el soporte desde una cota superior hasta otra inferior.

Después de leer el fichero creado antes, conectamos el nodo “A priori”, cuya configuración es:

Options | Flow Variables | Memory Policy

About

Class implementing an Apriori-type algorithm. [More](#) [Capabilities](#)

car: False

classIndex: -1

delta: 0.05

lowerBoundMinSupport: 0.1

metricType: Confidence

minMetric: 0.9

numRules: 100

outputItemSets: True

removeAllMissingCols: False

significanceLevel: -1.0

treatZeroAsMissing: False

upperBoundMinSupport: 1.0

verbose: False

Preliminary Attribute check

sepal length: ok

sepal width: ok

petal length: ok

petal width: ok

class: ok

- Un umbral de soporte mínimo del 10% (0.1 representa un 10% en lowerBoundMinSupport)
- Un umbral de soporte máximo de 100%
- Confianza del 90% (0.9 en minMetric).

Observaciones

- [More](#) muestra información adicional sobre el método empleado ([Capabilities](#), restricciones y tipos sobre los que opera).

- En `metricType` podemos escoger otras medidas de evaluación de las reglas (p.ej. lift).
- `OutputItemsetItems` indica si deseamos obtener los ítem sets frecuentes.

Obtenemos la siguiente salida:

Apriori

=====

Minimum support: 0.1 (15 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 19

Large Itemsets L(1):

sepallength=SL-Alto 35

sepallength=SL-Bajo 41

sepallength=SL-Medio 39

sepallength=SL-MuyAlto 35

sepalwidth=SW-Alto 30

sepalwidth=SW-Bajo 47

sepalwidth=SW-Medio 36

sepalwidth=SW-MuyAlto 37

petallength=PL-Alto 41

petallength=PL-Bajo 44

petallength=PL-Medio 31

petallength=PL-MuyAlto 34

petalwidth=PW-Alto 38

petalwidth=PW-Bajo 41

petalwidth=PW-Medio 37

petalwidth=PW-MuyAlto 34

class=Iris-setosa 50

class=Iris-versicolor 50

class=Iris-virginica 50

Size of set of large itemsets L(2): 40

Large Itemsets L(2):

sepallength=SL-Alto sepalwidth=SW-Bajo 16

sepallength=SL-Alto petallength=PL-Alto 20

sepallength=SL-Alto petalwidth=PW-Alto 21

sepallength=SL-Alto class=Iris-versicolor 17

sepallength=SL-Alto class=Iris-virginica 18

sepallength=SL-Bajo sepalwidth=SW-MuyAlto 17

sepallength=SL-Bajo petallength=PL-Bajo 33

sepallength=SL-Bajo petalwidth=PW-Bajo 31
 sepallength=SL-Bajo class=Iris-setosa 36
 sepallength=SL-Medio sepalwidth=SW-Bajo 19
 sepallength=SL-Medio petallength=PL-Medio 19
 sepallength=SL-Medio petalwidth=PW-Medio 21
 sepallength=SL-Medio class=Iris-versicolor 20
 sepallength=SL-MuyAlto petallength=PL-MuyAlto 24
 sepallength=SL-MuyAlto petalwidth=PW-MuyAlto 21
 sepallength=SL-MuyAlto class=Iris-virginica 26
 sepalwidth=SW-Bajo petallength=PL-Alto 20
 sepalwidth=SW-Bajo petallength=PL-Medio 18
 sepalwidth=SW-Bajo petalwidth=PW-Medio 21
 sepalwidth=SW-Bajo class=Iris-versicolor 27
 sepalwidth=SW-Bajo class=Iris-virginica 19
 sepalwidth=SW-Medio petalwidth=PW-Alto 15
 sepalwidth=SW-Medio class=Iris-versicolor 15
 sepalwidth=SW-MuyAlto petallength=PL-Bajo 26
 sepalwidth=SW-MuyAlto petalwidth=PW-Bajo 23
 sepalwidth=SW-MuyAlto class=Iris-setosa 31
 petallength=PL-Alto petalwidth=PW-Alto 28
 petallength=PL-Alto class=Iris-versicolor 25
 petallength=PL-Alto class=Iris-virginica 16
 petallength=PL-Bajo petalwidth=PW-Bajo 38
 petallength=PL-Bajo class=Iris-setosa 44
 petallength=PL-Medio petalwidth=PW-Medio 26
 petallength=PL-Medio class=Iris-versicolor 25
 petallength=PL-MuyAlto petalwidth=PW-MuyAlto 26
 petallength=PL-MuyAlto class=Iris-virginica 34
 petalwidth=PW-Alto class=Iris-versicolor 22
 petalwidth=PW-Alto class=Iris-virginica 16
 petalwidth=PW-Bajo class=Iris-setosa 41
 petalwidth=PW-Medio class=Iris-versicolor 28
 petalwidth=PW-MuyAlto class=Iris-virginica 34

Size of set of large itemsets L(3): 22

Large Itemsets L(3):

sepallength=SL-Alto petallength=PL-Alto petalwidth=PW-Alto 17
 sepallength=SL-Bajo sepalwidth=SW-MuyAlto petallength=PL-Bajo 15
 sepallength=SL-Bajo sepalwidth=SW-MuyAlto class=Iris-setosa 17
 sepallength=SL-Bajo petallength=PL-Bajo petalwidth=PW-Bajo 30
 sepallength=SL-Bajo petallength=PL-Bajo class=Iris-setosa 33
 sepallength=SL-Bajo petalwidth=PW-Bajo class=Iris-setosa 31
 sepallength=SL-Medio petallength=PL-Medio petalwidth=PW-Medio 16
 sepallength=SL-Medio petallength=PL-Medio class=Iris-versicolor 16
 sepallength=SL-Medio petalwidth=PW-Medio class=Iris-versicolor 17
 sepallength=SL-MuyAlto petallength=PL-MuyAlto petalwidth=PW-MuyAlto 19
 sepallength=SL-MuyAlto petallength=PL-MuyAlto class=Iris-virginica 24
 sepallength=SL-MuyAlto petalwidth=PW-MuyAlto class=Iris-virginica 21
 sepalwidth=SW-Bajo petallength=PL-Medio petalwidth=PW-Medio 17

sepalwidth=SW-Bajo petallength=PL-Medio class=Iris-versicolor 18
 sepalwidth=SW-Bajo petalwidth=PW-Medio class=Iris-versicolor 21
 sepalwidth=SW-MuyAlto petallength=PL-Bajo petalwidth=PW-Bajo 20
 sepalwidth=SW-MuyAlto petallength=PL-Bajo class=Iris-setosa 26
 sepalwidth=SW-MuyAlto petalwidth=PW-Bajo class=Iris-setosa 23
 petallength=PL-Alto petalwidth=PW-Alto class=Iris-versicolor 20
 petallength=PL-Bajo petalwidth=PW-Bajo class=Iris-setosa 38
 petallength=PL-Medio petalwidth=PW-Medio class=Iris-versicolor 23
 petallength=PL-MuyAlto petalwidth=PW-MuyAlto class=Iris-virginica 26

Size of set of large itemsets L(4): 6

Large Itemsets L(4):

sepallength=SL-Bajo sepalwidth=SW-MuyAlto petallength=PL-Bajo class=Iris-setosa 15
 sepallength=SL-Bajo petallength=PL-Bajo petalwidth=PW-Bajo class=Iris-setosa 30
 sepallength=SL-Medio petallength=PL-Medio petalwidth=PW-Medio class=Iris-versicolor 15
 sepallength=SL-MuyAlto petallength=PL-MuyAlto petalwidth=PW-MuyAlto class=Iris-virginica 19
 sepalwidth=SW-Bajo petallength=PL-Medio petalwidth=PW-Medio class=Iris-versicolor 17
 sepalwidth=SW-MuyAlto petallength=PL-Bajo petalwidth=PW-Bajo class=Iris-setosa 20

Best rules found:

1. petallength=PL-Bajo 44 ==> class=Iris-setosa 44 <conf:(1)> lift:(3) lev:(0.2) [29] conv:(29.33)
2. petalwidth=PW-Bajo 41 ==> class=Iris-setosa 41 <conf:(1)> lift:(3) lev:(0.18) [27] conv:(27.33)
3. petallength=PL-Bajo petalwidth=PW-Bajo 38 ==> class=Iris-setosa 38 <conf:(1)> lift:(3) lev:(0.17) [25] conv:(25.33)
4. petallength=PL-MuyAlto 34 ==> class=Iris-virginica 34 <conf:(1)> lift:(3) lev:(0.15) [22] conv:(22.67)
5. petalwidth=PW-MuyAlto 34 ==> class=Iris-virginica 34 <conf:(1)> lift:(3) lev:(0.15) [22] conv:(22.67)
6. sepallength=SL-Bajo petallength=PL-Bajo 33 ==> class=Iris-setosa 33 <conf:(1)> lift:(3) lev:(0.15) [22] conv:(22)
7. sepallength=SL-Bajo petalwidth=PW-Bajo 31 ==> class=Iris-setosa 31 <conf:(1)> lift:(3) lev:(0.14) [20] conv:(20.67)
8. sepallength=SL-Bajo petallength=PL-Bajo petalwidth=PW-Bajo 30 ==> class=Iris-setosa 30 <conf:(1)> lift:(3) lev:(0.13) [20] conv:(20)
9. sepalwidth=SW-MuyAlto petallength=PL-Bajo 26 ==> class=Iris-setosa 26 <conf:(1)> lift:(3) lev:(0.12) [17] conv:(17.33)
10. petallength=PL-MuyAlto petalwidth=PW-MuyAlto 26 ==> class=Iris-virginica 26 <conf:(1)> lift:(3) lev:(0.12) [17] conv:(17.33)
11. sepallength=SL-MuyAlto petallength=PL-MuyAlto 24 ==> class=Iris-virginica 24 <conf:(1)> lift:(3) lev:(0.11) [16] conv:(16)

12. sepalwidth=SW-MuyAlto petalwidth=PW-Bajo 23 ==> class=Iris-setosa 23
 <conf:(1)> lift:(3) lev:(0.1) [15] conv:(15.33)

13. sepallength=SL-MuyAlto petalwidth=PW-MuyAlto 21 ==> class=Iris-virginica 21
 <conf:(1)> lift:(3) lev:(0.09) [14] conv:(14)

14. sepalwidth=SW-Bajo petalwidth=PW-Medio 21 ==> class=Iris-versicolor 21
 <conf:(1)> lift:(3) lev:(0.09) [14] conv:(14)

15. sepalwidth=SW-MuyAlto petallength=PL-Bajo petalwidth=PW-Bajo 20 ==>
 class=Iris-setosa 20 <conf:(1)> lift:(3) lev:(0.09) [13] conv:(13.33)

16. sepallength=SL-MuyAlto petallength=PL-MuyAlto petalwidth=PW-MuyAlto 19
 ==> class=Iris-virginica 19 <conf:(1)> lift:(3) lev:(0.08) [12] conv:(12.67)

17. sepalwidth=SW-Bajo petallength=PL-Medio 18 ==> class=Iris-versicolor 18
 <conf:(1)> lift:(3) lev:(0.08) [12] conv:(12)

18. sepallength=SL-Bajo sepalwidth=SW-MuyAlto 17 ==> class=Iris-setosa 17
 <conf:(1)> lift:(3) lev:(0.08) [11] conv:(11.33)

19. sepalwidth=SW-Bajo petallength=PL-Medio petalwidth=PW-Medio 17 ==>
 class=Iris-versicolor 17 <conf:(1)> lift:(3) lev:(0.08) [11] conv:(11.33)

20. sepallength=SL-Bajo sepalwidth=SW-MuyAlto petallength=PL-Bajo 15 ==>
 class=Iris-setosa 15 <conf:(1)> lift:(3) lev:(0.07) [10] conv:(10)

21. sepallength=SL-Bajo petalwidth=PW-Bajo 31 ==> petallength=PL-Bajo 30
 <conf:(0.97)> lift:(3.3) lev:(0.14) [20] conv:(10.95)

22. sepallength=SL-Bajo petalwidth=PW-Bajo class=Iris-setosa 31 ==>
 petallength=PL-Bajo 30 <conf:(0.97)> lift:(3.3) lev:(0.14) [20] conv:(10.95)

23. sepallength=SL-Bajo petalwidth=PW-Bajo 31 ==> petallength=PL-Bajo class=Iris-
 setosa 30 <conf:(0.97)> lift:(3.3) lev:(0.14) [20] conv:(10.95)

24. sepalwidth=SW-Bajo petallength=PL-Medio 18 ==> petalwidth=PW-Medio 17
 <conf:(0.94)> lift:(3.83) lev:(0.08) [12] conv:(6.78)

25. sepalwidth=SW-Bajo petallength=PL-Medio class=Iris-versicolor 18 ==>
 petalwidth=PW-Medio 17 <conf:(0.94)> lift:(3.83) lev:(0.08) [12] conv:(6.78)

26. sepalwidth=SW-Bajo petallength=PL-Medio 18 ==> petalwidth=PW-Medio
 class=Iris-versicolor 17 <conf:(0.94)> lift:(5.06) lev:(0.09) [13] conv:(7.32)

27. sepallength=SL-Medio petallength=PL-Medio class=Iris-versicolor 16 ==>
 petalwidth=PW-Medio 15 <conf:(0.94)> lift:(3.8) lev:(0.07) [11] conv:(6.03)

28. sepallength=SL-Medio petallength=PL-Medio petalwidth=PW-Medio 16 ==>
 class=Iris-versicolor 15 <conf:(0.94)> lift:(2.81) lev:(0.06) [9] conv:(5.33)

29. petalwidth=PW-Bajo 41 ==> petallength=PL-Bajo 38 <conf:(0.93)> lift:(3.16)
 lev:(0.17) [25] conv:(7.24)

30. petalwidth=PW-Bajo class=Iris-setosa 41 ==> petallength=PL-Bajo 38
 <conf:(0.93)> lift:(3.16) lev:(0.17) [25] conv:(7.24)

31. petalwidth=PW-Bajo 41 ==> petallength=PL-Bajo class=Iris-setosa 38
 <conf:(0.93)> lift:(3.16) lev:(0.17) [25] conv:(7.24)

32. sepallength=SL-MuyAlto class=Iris-virginica 26 ==> petallength=PL-MuyAlto 24
 <conf:(0.92)> lift:(4.07) lev:(0.12) [18] conv:(6.7)

33. petallength=PL-Medio class=Iris-versicolor 25 ==> petalwidth=PW-Medio 23
 <conf:(0.92)> lift:(3.73) lev:(0.11) [16] conv:(6.28)

34. sepallength=SL-Bajo class=Iris-setosa 36 ==> petallength=PL-Bajo 33
 <conf:(0.92)> lift:(3.13) lev:(0.15) [22] conv:(6.36)

35. sepallength=SL-Bajo petallength=PL-Bajo 33 ==> petalwidth=PW-Bajo 30
 <conf:(0.91)> lift:(3.33) lev:(0.14) [20] conv:(6)

36. sepallength=SL-Bajo petallength=PL-Bajo class=Iris-setosa 33 ==>
 petalwidth=PW-Bajo 30 <conf:(0.91)> lift:(3.33) lev:(0.14) [20] conv:(6)

37. sepallength=SL-Bajo petallength=PL-Bajo 33 ==> petalwidth=PW-Bajo class=Iris-setosa 30 <conf:(0.91)> lift:(3.33) lev:(0.14) [20] conv:(6)
 38. petalwidth=PW-Alto class=Iris-versicolor 22 ==> petallength=PL-Alto 20 <conf:(0.91)> lift:(3.33) lev:(0.09) [13] conv:(5.33)
 39. sepallength=SL-MuyAlto petalwidth=PW-MuyAlto 21 ==> petallength=PL-MuyAlto 19 <conf:(0.9)> lift:(3.99) lev:(0.09) [14] conv:(5.41)
 40. sepallength=SL-MuyAlto petalwidth=PW-MuyAlto class=Iris-virginica 21 ==> petallength=PL-MuyAlto 19 <conf:(0.9)> lift:(3.99) lev:(0.09) [14] conv:(5.41)
 41. sepallength=SL-MuyAlto petalwidth=PW-MuyAlto 21 ==> petallength=PL-MuyAlto class=Iris-virginica 19 <conf:(0.9)> lift:(3.99) lev:(0.09) [14] conv:(5.41)

Donde puede verse el mecanismo de generación de ítem-sets frecuentes, la propiedad a priori y el hecho de que el análisis de reglas con una salida de este tipo es complejo.

La mayor ventaja de esta forma de Weka es que permite extraer reglas de asociación considerando un atributo consecuente que sería el que se pretende predecir como una clase. Para ello:

- Se selecciona la opción car a true
- Se selecciona a 5 el parámetro classindex para predecir el atributo “class”, (de todas formas si se deja a -1 también funcionaría para predecir porque -1 supone que el que hay que predecir es el último atributo

Obtenemos la salida:

```
Apriori
=====

Minimum support: 0.1 (15 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18


Generated sets of large itemsets:

Size of set of large itemsets L(1): 19
Size of set of large itemsets L(2): 15
Size of set of large itemsets L(3): 6

Best rules found:

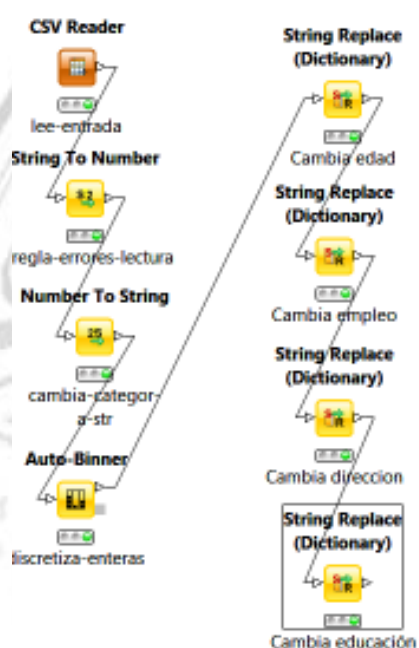
1. petallength=PL-Bajo 44 ==> class=Iris-setosa 44 conf:(1)
2. petalwidth=PW-Bajo 41 ==> class=Iris-setosa 41 conf:(1)
3. petallength=PL-Bajo petalwidth=PW-Bajo 38 ==> class=Iris-setosa 38 conf:(1)
4. petallength=PL-MuyAlto 34 ==> class=Iris-virginica 34 conf:(1)
5. petalwidth=PW-MuyAlto 34 ==> class=Iris-virginica 34 conf:(1)
6. sepallength=SL-Bajo petallength=PL-Bajo 33 ==> class=Iris-setosa 33 conf:(1)
7. sepallength=SL-Bajo petalwidth=PW-Bajo 31 ==> class=Iris-setosa 31 conf:(1)
8. sepallength=SL-Bajo petallength=PL-Bajo petalwidth=PW-Bajo 30 ==> class=Iris-setosa 30 conf:(1)
9. sepalwidth=SW-MuyAlto petallength=PL-Bajo 26 ==> class=Iris-setosa 26 conf:(1)
10. petallength=PL-MuyAlto petalwidth=PW-MuyAlto 26 ==> class=Iris-virginica 26 conf:(1)
11. sepallength=SL-MuyAlto petallength=PL-MuyAlto 24 ==> class=Iris-virginica 24 conf:(1)
12. sepalwidth=SW-MuyAlto petalwidth=PW-Bajo 23 ==> class=Iris-setosa 23 conf:(1)
13. sepallength=SL-MuyAlto petalwidth=PW-MuyAlto 21 ==> class=Iris-virginica 21 conf:(1)
14. sepalwidth=SW-Bajo petalwidth=PW-Medio 21 ==> class=Iris-versicolor 21 conf:(1)
15. sepalwidth=SW-MuyAlto petallength=PL-Bajo petalwidth=PW-Bajo 20 ==> class=Iris-setosa 20 conf:(1)
16. sepallength=SL-MuyAlto petallength=PL-MuyAlto petalwidth=PW-MuyAlto 19 ==> class=Iris-virginica 19 conf:(1)
17. sepalwidth=SW-Bajo petallength=PL-Medio 18 ==> class=Iris-versicolor 18 conf:(1)
18. sepallength=SL-Bajo sepalwidth=SW-MuyAlto 17 ==> class=Iris-setosa 17 conf:(1)
19. sepalwidth=SW-Bajo petallength=PL-Medio petalwidth=PW-Medio 17 ==> class=Iris-versicolor 17 conf:(1)
20. sepallength=SL-Bajo sepalwidth=SW-MuyAlto petallength=PL-Bajo 15 ==> class=Iris-setosa 15 conf:(1)
21. sepallength=SL-Medio petallength=PL-Medio petalwidth=PW-Medio 16 ==> class=Iris-versicolor 15 conf:(0.94)
```


Nuevamente nos encontramos con la dificultad del análisis de sólo textos cuando se tiene una salida de reglas de asociación. Basta simplemente ordenar la tabla que se obtiene como salida del nodo de Knime para encontrar más fácilmente las asociaciones que tienen la clase como consecuente:

Row ID	D Support	D Confid...	D Lift	S  Co...	S implies	S Split V...	S Split V...	S Split V...
rule0	0.1	1	3	Iris-setosa	<---	SW-Muy A...	SL-Bajo	PL-Bajo
rule1	0.113	1	3	Iris-setosa	<---	SW-Muy A...	SL-Bajo	?
rule4	0.2	1	3	Iris-setosa	<---	SL-Bajo	PW-Bajo	PL-Bajo
rule6	0.22	1	3	Iris-setosa	<---	SL-Bajo	PL-Bajo	?
rule7	0.207	1	3	Iris-setosa	<---	SL-Bajo	PW-Bajo	?
rule8	0.133	1	3	Iris-setosa	<---	SW-Muy A...	PW-Bajo	PL-Bajo
rule9	0.173	1	3	Iris-setosa	<---	SW-Muy A...	PL-Bajo	?
rule10	0.153	1	3	Iris-setosa	<---	SW-Muy A...	PW-Bajo	?
rule12	0.253	1	3	Iris-setosa	<---	PW-Bajo	PL-Bajo	?
rule13	0.293	1	3	Iris-setosa	<---	PL-Bajo	?	?
rule14	0.273	1	3	Iris-setosa	<---	PW-Bajo	?	?
rule16	0.1	0.938	2.812	Iris-versic...	<---	SL-Medio	PL-Medio	PW-Medio
rule18	0.113	1	3	Iris-versic...	<---	PL-Medio	PW-Medio	SW-Bajo
rule20	0.12	1	3	Iris-versic...	<---	PL-Medio	SW-Bajo	?
rule21	0.14	1	3	Iris-versic...	<---	PW-Medio	SW-Bajo	?
rule23	0.127	1	3	Iris-virginica	<---	PW-Muy Alto	PL-Muy Alto	SL-Muy Alto
rule25	0.16	1	3	Iris-virginica	<---	PL-Muy Alto	SL-Muy Alto	?
rule26	0.14	1	3	Iris-virginica	<---	PW-Muy Alto	SL-Muy Alto	?
rule28	0.173	1	3	Iris-virginica	<---	PW-Muy Alto	PL-Muy Alto	?
rule29	0.227	1	3	Iris-virginica	<---	PL-Muy Alto	?	?
rule30	0.227	1	3	Iris-virginica	<---	PW-Muy Alto	?	?

Ejemplo de Bankloan.csv

Vamos ahora a trabajar con el otro data set, intentando encontrar asociaciones entre algunas variables personales de los clientes. Concretamente nos vamos a centrar en : nivel de educación, que es nominal, edad, años en el empleo y años en la dirección. Estas tres últimas deben ser discretizadas y cambiaremos también el nivel de educación para que sus valores queden más claros. El flujo concreto sería



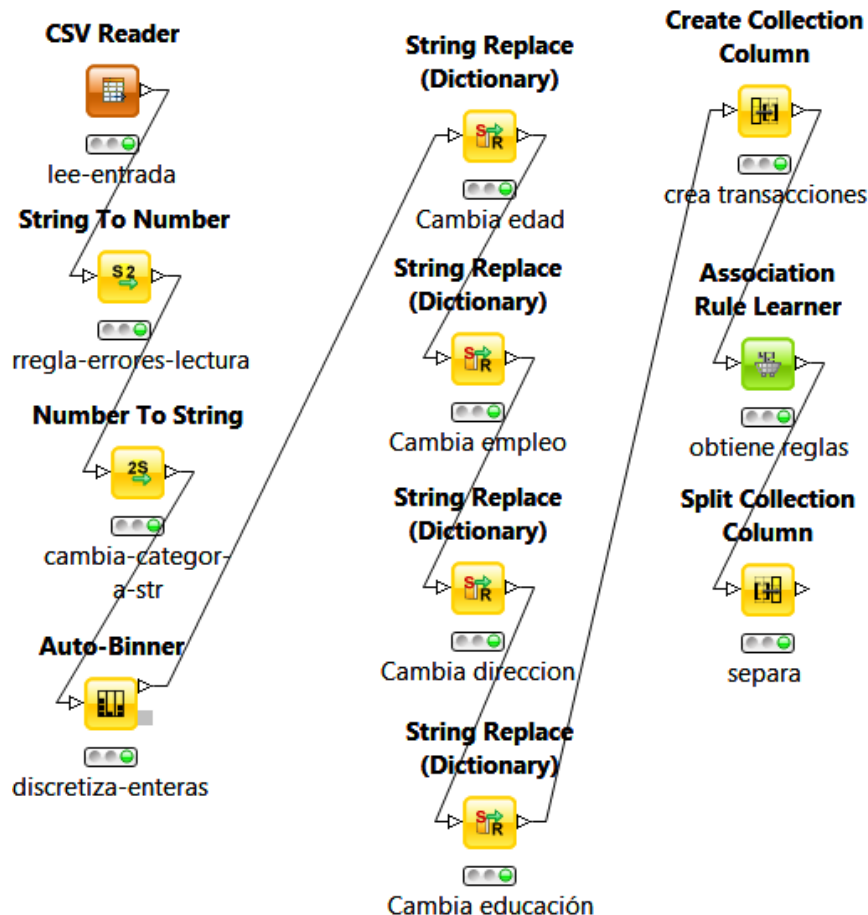
Lo que nos da como salida:

Table "default" - Rows: 5000												
Spec - Columns: 16 Properties Flow Variables												
Row ID	i age	S ed	i employ	i address	D income	D debtinc	D cred...	D othdebt	S default	S age [8...	S emplo...	S adres...
Row0	41	Grado	17	12	35.9	11.9	0.504	3.768	0	Maduro	Emp-Much...	Dir-Bastant...
Row1	30	Secundaria	13	8	46.7	17.88	1.353	6.997	0	Joven	Emp-Bast...	Dir-Bastant...
Row2	40	Secundaria	15	14	61.8	10.64	3.439	3.137	0	Maduro	Emp-Much...	Dir-Muchos
Row3	41	Secundaria	15	14	72	29.67	4.166	17.197	0	Maduro	Emp-Much...	Dir-Muchos
Row4	57	Secundaria	7	37	25.6	15.86	1.498	2.562	0	Mayor	Emp-Algu...	Dir-Muchos
Row5	45	Secundaria	0	13	28.1	4.28	0.925	0.278	0	Mayor	Emp-Pocos	Dir-Muchos
Row6	36	Secundaria	1	3	19.6	12.82	1.211	1.302	1	Maduro	Emp-Pocos	Dir-Pocos
Row7	39	Secundaria	20	9	80.5	12.32	1.855	8.063	0	Maduro	Emp-Much...	Dir-Bastant...
Row8	43	Secundaria	12	11	68.7	6.82	1.429	3.256	0	Mayor	Emp-Bast...	Dir-Bastant...
Row9	34	Grado	7	12	33.8	10.71	1.423	2.197	0	Joven	Emp-Algu...	Dir-Bastant...
Row10	26	Secundaria	1	2	22.2	7.27	0.581	1.033	0	Muy joven	Emp-Pocos	Dir-Pocos
Row11	37	Bachiller	17	10	78.3	25.44	7.091	12.828	1	Maduro	Emp-Much...	Dir-Bastant...
Row12	44	Secundaria	8	15	77.8	10.55	3.611	4.596	0	Mayor	Emp-Bast...	Dir-Muchos
Row13	36	Bachiller	8	1	48.1	5.28	0.737	1.803	1	Maduro	Emp-Bast...	Dir-Pocos

Con esta salida podemos generar el atributo colección, tomando: ed, age[binned] etc..., lo que nos da:

(...) AggregatedValues
[Grado, Maduro, Emp-Muchos,...]
[Secundaria, Joven, Emp-Bastantes,...]
[Secundaria, Maduro, Emp-Muchos,...]
[Secundaria, Maduro, Emp-Muchos,...]
[Secundaria, Mayor, Emp-Algunos,...]
[Secundaria, Mayor, Emp-Pocos,...]
[Secundaria, Maduro, Emp-Pocos,...]
[Secundaria, Maduro, Emp-Muchos,...]
[Secundaria, Mayor, Emp-Bastantes,...]
[Grado, Joven, Emp-Algunos,...]
[Secundaria, Muy joven, Emp-Pocos,...]
[Bachiller, Maduro, Emp-Muchos,...]
[Secundaria, Mayor, Emp-Bastantes,...]
[Bachiller, Maduro, Emp-Bastantes,...]

A partir de la cual ya podemos obtener las reglas de asociación mediante el nodo correspondiente, y separa los antecedentes agregados, para obtener algunas reglas de asociación, en este caso el soporte mínimo ha de ser de 0.01 y la confianza mínima de 0.6. El flujo completo sería:



Lo que nos dá como salida:

Row ID	D Confid...	D Lift	S Conse...	S implies	S Split V...	S Split V...	S Split V...
rule0	0.649	1.203	Secundaria	<---	Emp-Much...	Dir-Algunos	Maduro
rule1	0.602	2.557	Dir-Muchos	<---	Secundaria	Mayor	Emp-Algu...
rule2	0.67	2.355	Emp-Pocos	<---	Dir-Pocos	Master	?
rule3	0.829	2.913	Emp-Pocos	<---	Muy joven	Dir-Pocos	Grado
rule4	0.636	2.085	Dir-Pocos	<---	Muy joven	Grado	Emp-Pocos
rule5	0.724	2.726	Muy joven	<---	Dir-Pocos	Grado	Emp-Pocos
rule6	0.797	1.477	Secundaria	<---	Muy joven	Dir-Algunos	Emp-Bast...
rule7	0.765	3.221	Mayor	<---	Emp-Much...	Grado	?
rule8	0.636	2.393	Muy joven	<---	Dir-Algunos	Bachiller	Emp-Pocos
rule9	0.93	3.269	Emp-Pocos	<---	Muy joven	Master	?
rule10	0.862	1.596	Secundaria	<---	Muy joven	Dir-Pocos	Emp-Bast...
rule11	0.623	1.155	Secundaria	<---	Dir-Muchos	Emp-Much...	Maduro
rule12	0.707	2.485	Emp-Pocos	<---	Muy joven	Grado	?
rule13	0.629	2.212	Emp-Pocos	<---	Muy joven	Dir-Pocos	Bachiller
rule14	0.765	2.882	Muy joven	<---	Dir-Pocos	Bachiller	Emp-Pocos
rule15	0.626	1.159	Secundaria	<---	Dir-Pocos	Emp-Much...	?
rule16	0.644	1.192	Secundaria	<---	Emp-Much...	Dir-Algunos	?
rule17	0.833	1.544	Secundaria	<---	Emp-Much...	Joven	?
rule18	0.624	1.157	Secundaria	<---	Emp-Much...	Dir-Bastan...	?
rule19	0.681	1.261	Secundaria	<---	Muy joven	Dir-Pocos	Emp-Algu...

Ejercicio.

Trabajar con los datos de ingresos, y los tres tipos de deudas para encontrar asociaciones significativas entre ello y el impago. Utilizar los enfoque de reglas de asociación puras y los que proporciona Weka para predecir en su versión de a priori. Habrá que discretizar adecuadamente cambiando los valores de los atributos por medio de diccionarios .

Estudio de reglas de asociación con R-Studio.

En R-studio se calculan las reglas de asociación con el paquete **Arules** este paquete permite el calculo de reglas e itemsets partiendo de datasets cuyos valores de atributos son categóricos.

Estudio de reglas de asociación con Iris

Para trabajar con este dataset hay que categorizar los datos de las cuatro primeras variables que son continuas. Resto se puede hacer de dos maneras:

- Utilizar el fichero asociacion. iris.csv que se generó en Knime , importándolo como fichero de texto. Se incluye como material prácticas, El fichero **asociación-iris.R** trabaja con esta opción
- Transformar el fichero iris mediante comandos de R El fichero **asociación-iris-nuevo.R** muestra como se hace esto último.

Hay que indicar que ambos ficheros salvo muy al principio son iguales, que calculan reglas de asociación mediante el algoritmo “a priori”, y ordenan , filtran y listan dichas reglas de forma que se vean lo más amigablemente posible.

Probablemente las mejores opciones de R en relación con las reglas de asociación sean las herramientas de visualización que pasamos a describir:

Visualización de las reglas de asociación en R

La visualización se hace por medio del paquete **arulesViz** que obviamente es de los mismo autores que el anterior **Arules**. La función general de llamada es:

```
plot(x, method = NULL, measure = "support", shading = "lift",  
     interactive = FALSE, data = NULL, control = NULL, ...)
```

donde **x** es un conjunto de reglas o itemsets. Son los argumentos los que nos permiten obtener los distintos gráficos. Concretamente, los argumentos más importantes son:

- **method**

Una cadena con valores "scatterplot", "two-key plot", "matrix", "matrix3D", "mosaic", "doubledecker", "graph", "paracoord" o "grouped", "iplots" y sirve para seleccionar el método de visualización

- **measure**

Es la medida de interés (p.e., "support", "confidence", "lift", "order") que se usa en la visualización. Algunos métodos de visualización utilizan una medida, otros

(p.e., scatterplot), utiliza un vector de dos medidas (ejes X e Y). En algunos dibujos (p.e., graphs) se puede utilizar NA para suprimir una medida.

- **shading**

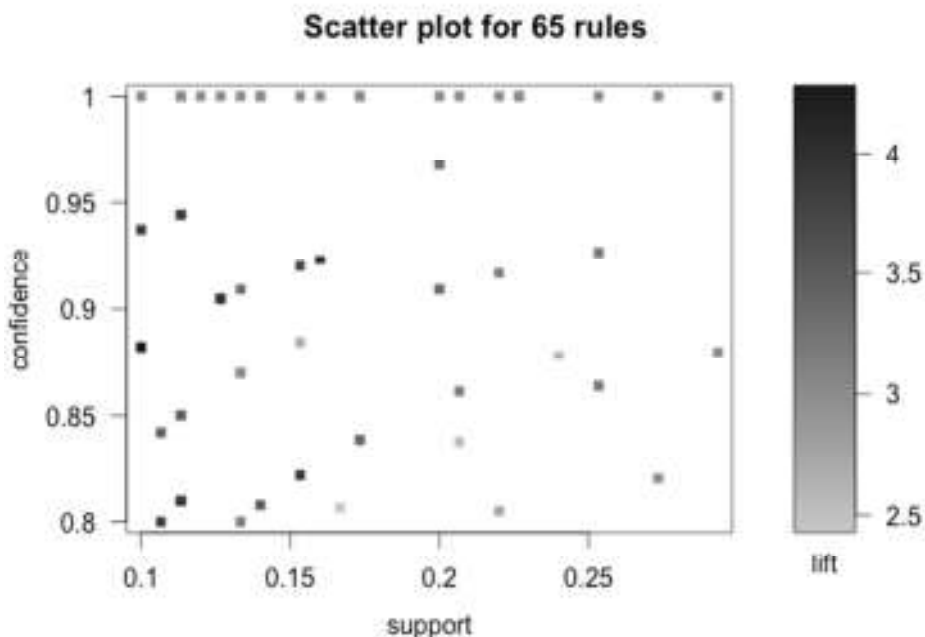
Es una medida de interés para colorear puntos, nodos, flechas etc. (p.e., "support", "confidence", "lift"). Por defecto es "lift". Se puede usar NA para suprimir el shading..

El argumento más importante es "method":

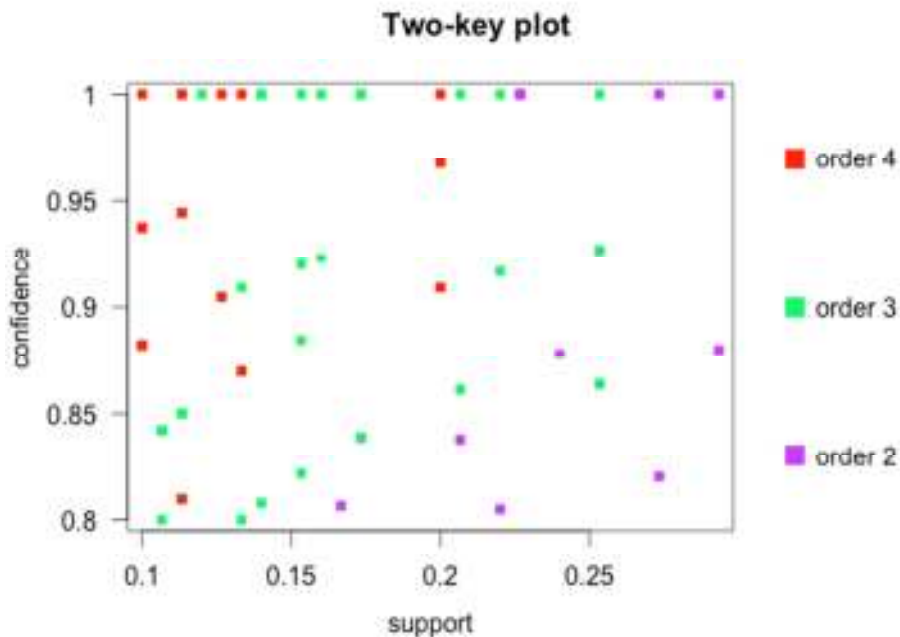
1. "scatterplot", "two-key plot".- Este método de visualización dibuja un diagrama de dispersión de dos dimensiones con diferentes medidas de interés (parámetro " measure") en los ejes y una tercera medida (parámetro " shading") está representado por el color de los puntos. Hay un valor especial para el shading llamado " order" que produce una parcela de dos claves donde el color de los puntos representa la longitud (order) de la regla .

Ejemplos:

```
plot(sale1,measure=c("support","confidence"))
```



```
plot(sale1, shading="order", control=list(main = "Two-key plot",col=rainbow(5)))
```



2. "matrix", "matrix3D".-Organiza las reglas de asociación como una matriz con los itemsets en los antecedentes en un eje y los itemsets en los consecuentes en el otro. La medida de interés se visualiza ya sea por un color (más oscuro significa un valor más alto para la medida) o como la altura de una barra (método "matrix3D").

En este caso la salida nos da los antecedentes y los consecuentes pero según un código numérico que se adjunta como salida en la consola.

Ejemplos:

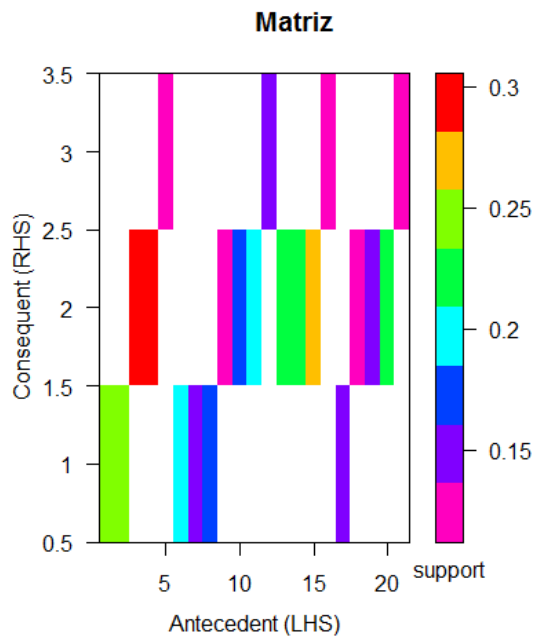
```
plot(sale2, method="matrix", control=list(main = "Matriz", col=rainbow(8)))
```

Salida por consola:

```
Itemsets in Antecedent (LHS)
[1] "{petal.width=PW-Muy Alto}"
[2] "{petal.length=PL-Muy Alto}"
[3] "{petal.width=PW-Bajo}"
[4] "{petal.length=PL-Bajo}"
[5] "{sepal.width=SW-Bajo,petal.length=PL-Medio}"
[6] "{petal.length=PL-Muy Alto,petal.width=PW-Muy Alto}"
[7] "{sepal.length=SL-Muy Alto,petal.width=PW-Muy Alto}"
[8] "{sepal.length=SL-Muy Alto,petal.length=PL-Muy Alto}"
[9] "{sepal.length=SL-Bajo,sepal.width=SW-Muy Alto}"
[10] "{sepal.width=SW-Muy Alto,petal.width=PW-Bajo}"
[11] "{sepal.width=SW-Muy Alto,petal.length=PL-Bajo}"
[12] "{sepal.width=SW-Bajo,petal.width=PW-Medio}"
[13] "{sepal.length=SL-Bajo,petal.width=PW-Bajo}"
[14] "{sepal.length=SL-Bajo,petal.length=PL-Bajo}"
[15] "{petal.length=PL-Bajo,petal.width=PW-Bajo}"
[16] "{sepal.width=SW-Bajo,petal.length=PL-Medio,petal.width=PW-Medio}"
[17] "{sepal.length=SL-Muy Alto,petal.length=PL-Muy Alto,petal.width=PW-Muy Alto}"
```

```
[18] "{sepal.length=SL-Bajo,sepal.width=SW-Muy Alto,petal.length=PL-Ba
jo}"
[19] "{sepal.width=SW-Muy Alto,petal.length=PL-Bajo,petal.width=PW-Baj
o}"
[20] "{sepal.length=SL-Bajo,petal.length=PL-Bajo,petal.width=PW-Bajo}"
[21] "{sepal.length=SL-Medio,petal.length=PL-Medio,petal.width=PW-Medi
o}"
Itemsets in Consequent (RHS)
[1] "{class=Iris-virginica}" "{class=Iris-setosa}"
[3] "{class=Iris-versicolor}"
```

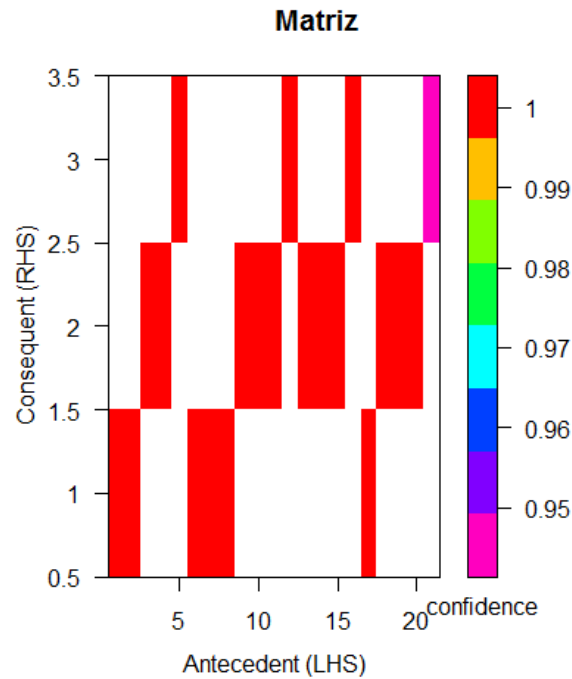
Salida dibujo



Si hacemos ahora:

```
plot(sale2, method="matrix",measure="confidence" control=list(main =
"Matriz",col=rainbow(8)))
```

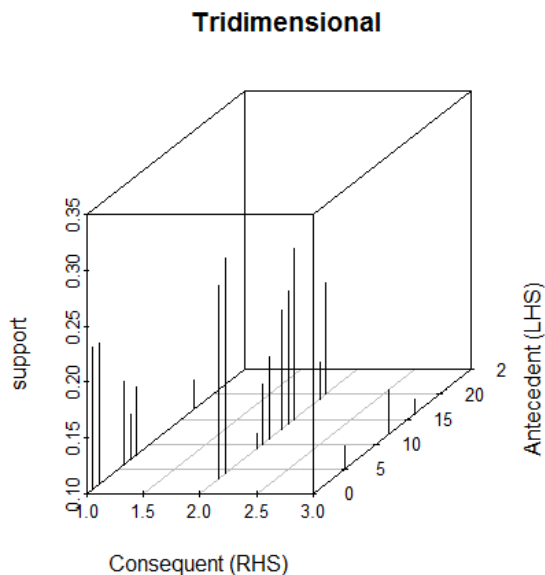
La salida de consola será la misma pero el dibujo es:



Recordemos en en sale2 tiene casi todas las reglas con confianza 1

```
plot(sale2,method="matrix3d",shading="order",control=list(main="Tridimensional"))
```

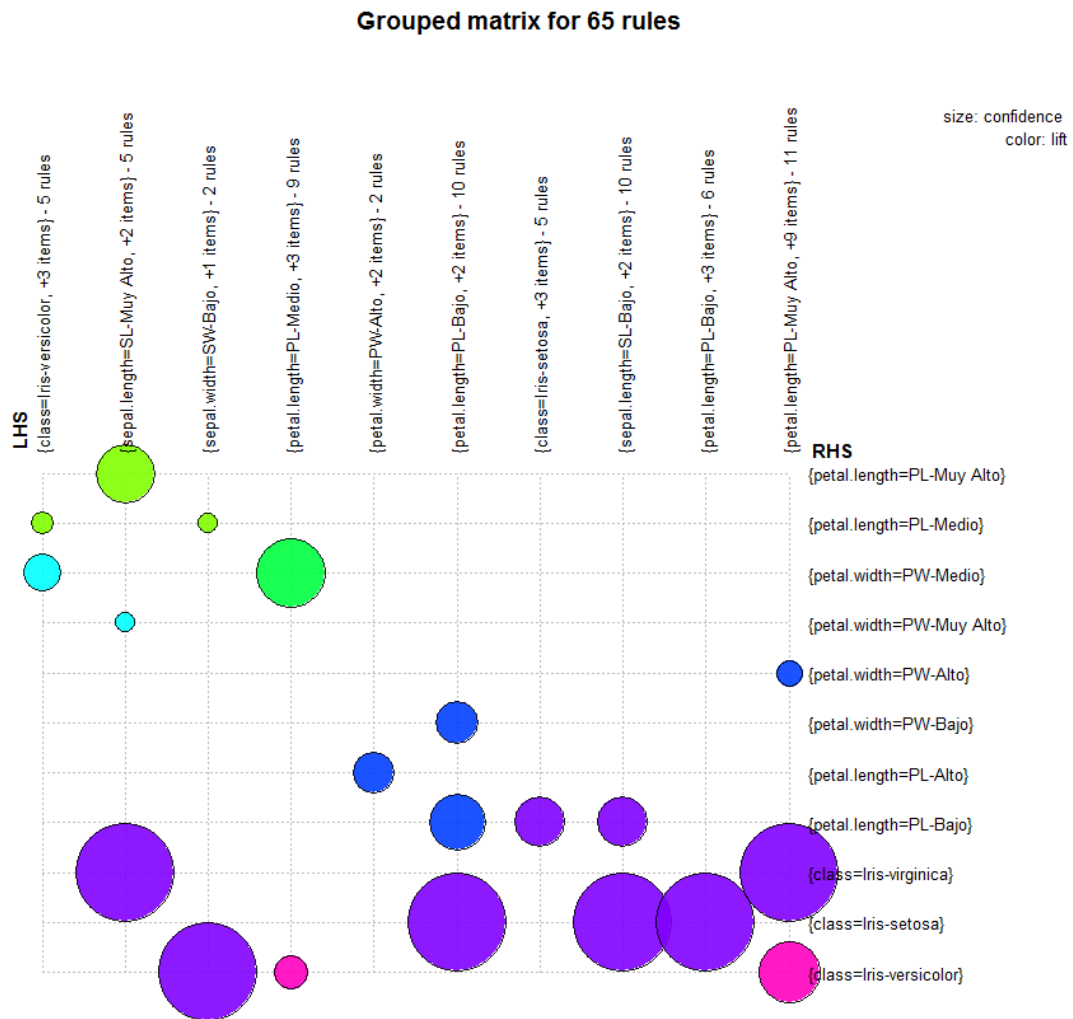
Nos da la misma lista de consola y el dibujo:



3.- "grouped" Los antecedentes (columnas) en la matriz se agrupan mediante una técnica de agrupamiento. Los grupos se representan como bolas en la matriz.

Ejemplo:

plot(sale1, method="grouped",measure="confidence", control=list(main = "Cluster",k=10,col=rainbow(8))) da como salida:



4.- “graph”. Representa las reglas como un grafo de itemsets.

Ejemplo:

plot(sale2[1:10], method="graph",measure="confidence", control=list(main = "Grafo"))

Se obtiene:



Ejercicio.

Replicar los resultados de Bankloan en Knime para los datos personales utilizando los datos discretizados ya obtenidos. Trabajar con los datos de ingresos, y los tres tipos de deudas para encontrar asociaciones significativas entre ello y el impago, transformando los datos mediante R, salvar toda la información discretizada en un fichero csv .