STUDENT ALCOHOL CONSUMPTION Dataset

Aythami Estévez Olivas
Marvin Matías Agüero

Introducción

- **Origen del dataset:** Encuestas a estudiantes portugueses de dos institutos en el curso 2005-2006.
  - Dos datasets
    - Alumnos de portugués: 649 items
    - Alumnos de matemáticas: 395 items

Atributos del Dataset

Atributo	Descripción	Tipo	Valores Posibles
school	Instituto Educativo	Binario	"GP" Gabriel Pereira o "MS" Mousinho da Silveira
sex	Sexo	Binario	“M” Masculino o “F” Femenino
age	Edad	Numérico	15 a 22
address	Tipo de ubicación del hogar	Binario	"U" Urbano o "R" Rural
famsize	Tamaño de la familia	Binario	“LE3” menor o igual a 3 o “GT3” mayor a 3

Atributos del Dataset

Atributo	Descripción	Tipo	Valores Posibles
Pstatus	Estado civil de los padres	Binario	“T” Juntos o “A” Separados
Medu	Educación de la madre	Numérico	0, no, 1, educación primaria (4° grado), 2, 5° a 9° grado, 3, educación secundaria o 4, educación superior
Fedu	Educación del padre	Numérico	0, no, 1, educación primaria (4° grado), 2, 5° a 9° grado, 3, educación secundaria o 4, educación superior
Mjob	Trabajo de la madre	Nominal	“teacher” profesor, “health” salud, “services” funcionarios públicos, “at_home” en casa u “other” otros
Fjob	Trabajo del padre	Nominal	“teacher” profesor, “health” salud, “services” funcionarios públicos, “at_home” en casa u “other” otros
reason	Razón por la que eligió este instituto	Nominal	“home” el hogar, “reputation” reputación del instituto, “course” preferencia de cursos u “other” otros
guardian	Tutor	Nominal	“mother” madre, “father” padre u “other” otro

Atributos del Dataset

Atributo	Descripción	Tipo	Valores Posibles
traveltime	Tiempo de viaje de la casa al instituto	Numérico	1 menor a 15 min., 2 15 a 30 min., 3 30 min. a 1 hora o 4 mayor a 1 hora
studytime	Tiempo estudio de semanal	Numérico	1 menor a 2 horas, 2 2 a 5 horas, 3 5 a 10 horas o 4 mayor a 10 horas
failures	Número de suspensos en asignaturas	Numérico	“n” si son menores o iguales a 3 o 4 si son mayores
schoolsup	Soporte educacional fuera del instituto	Binario	“Yes” Sí o “No” No
famsup	Soporte educacional de la familia	Binario	“Yes” Sí o “No” No
paid	Pago de clases extras del curso en cuestión (de Portugués o Matemática)	Binario	“Yes” Sí o “No” No
activities	Actividades extra-curriculares	Binario	“Yes” Sí o “No” No
nursery	Atención recibida en la enfermería	Binario	“Yes” Sí o “No” No

Atributos del Dataset

Atributo	Descripción	Tipo	Valores Posibles
higher	Desea seguir estudios universitarios	Binario	“Yes” Sí o “No” No
internet	Posee acceso a internet en la casa	Binario	“Yes” Sí o “No” No
romantic	Posea una relación amorosa	Binario	“Yes” Sí o “No” No
famrel	Calidad de la relación familiar	Numérico	1, muy malo a 5, excelente
freetime	Tiempo libre después de clases	Numérico	1, muy bajo a 5, muy alto
goout	Salidas con sus amigos	Numérico	1, muy bajo a 5, muy alto
Dalc	Consumo de alcohol en días laborales	Numérico	1, muy bajo a 5, muy alto
Walc	Consumo de alcohol fines de semana	Numérico	1, muy bajo a 5, muy alto
health	Estado de salud	Numérico	1, muy malo a 5, muy bueno
absences	Número de ausencias a clases	Numérico	0 a 93
G1	Calificación del 1° periodo académico	Numérico	0 a 20
G2	Calificación del 2° periodo académico	Numérico	0 a 20
G3	Calificación del periodo académico final	Numérico	0 a 20

Otros trabajos con este dataset

- *Using Data Mining To Predict Secondary School Student Alcohol Consumption*, 2016
  - Emplea árboles de decisión (Random Forest) para predecir el consumo de alcohol de los estudiantes llegando a un acierto cercano al 92%.
  - Sintetiza las dos variables relacionadas con el consumo de alcohol en una mediante una media ponderada.

- *Using Data Mining to Predict Secondary School Student Performance*, 2008.
  - Autores originales del dataset.
  - Utilizan diversas técnicas, desde árboles de decisión, a redes neuronales pasando por máquinas de soporte vectorial para predecir las notas de los estudiantes.
  - Demuestran que se puede predecir con bastante certeza las notas de los alumnos si se tienen las notas pasadas.

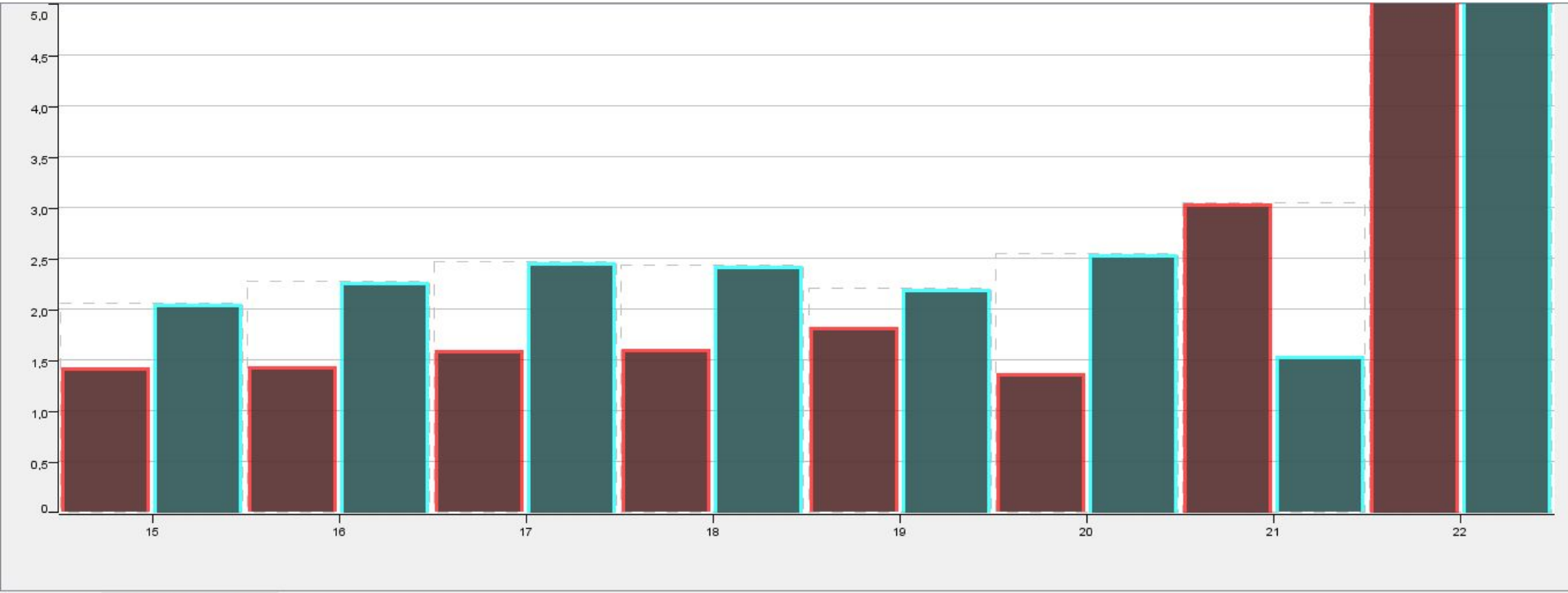
- Ver apartado de referencias para signatura completa de estos trabajos.

Análisis exploratorio

Primeras Impresiones

- **Nuestra hipótesis**: Comprobar que atributos tienen relación con el consumo de alcohol.

- **Observaciones iniciales:**
  - **Edad:** Consumo elevado en mayores de 20 años.
  - **Sexo:** Algo más de consumo en hombres, no muy significativo.
  - **Estudio:** Los alumnos que estudian menos de 2 horas al día y/o los que obtienen notas finales por debajo de 3, tienen un consumo más elevado.



Default Settings | Column/Aggregation settings | Bin settings | Visualization settings | Details

Aggregation method:

- (•) Average
- ( ) Sum
- ( ) Row count
- ( ) Row count (w/o missing values)

Binning column: age

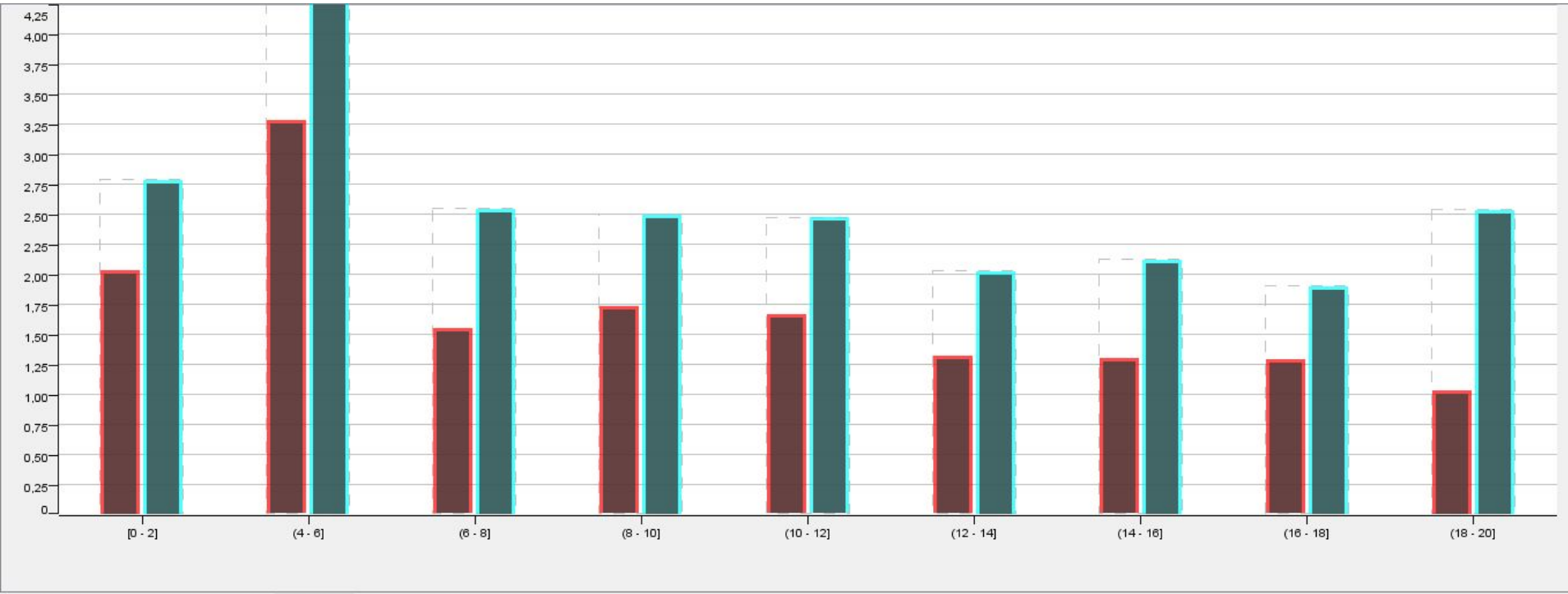
Aggregation column:

Available columns
- age
- Medu
- Fedu
- traveltime
- studytime

add >>

<< remove

Aggregation columns
- Dalc
- Walc



Default Settings | Column/Aggregation settings | Bin settings | Visualization settings | Details

Number of bins: 19

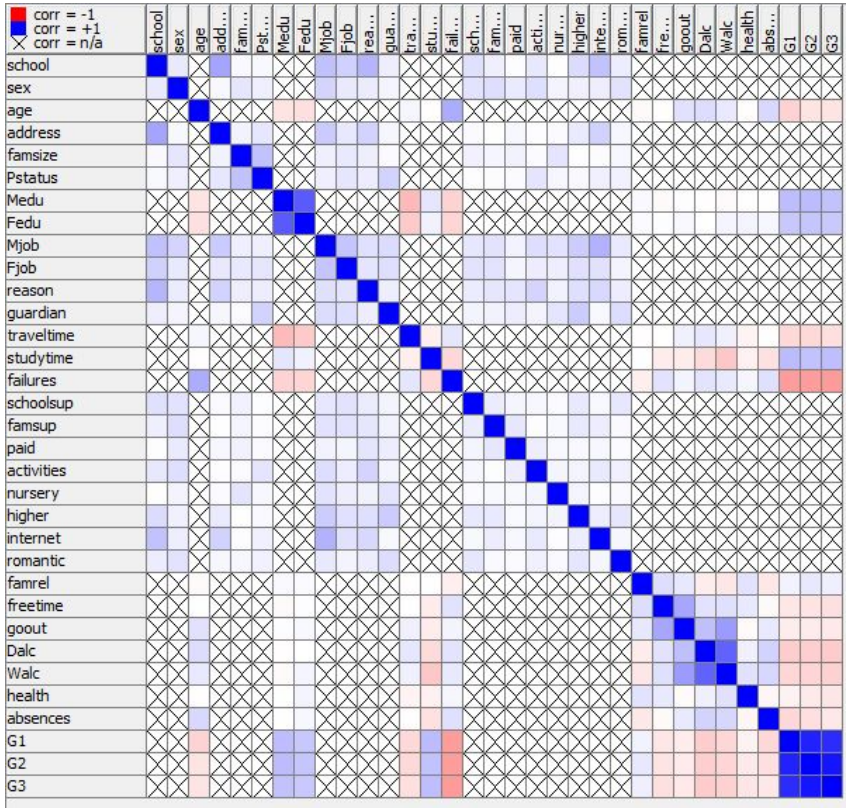
Show empty bins

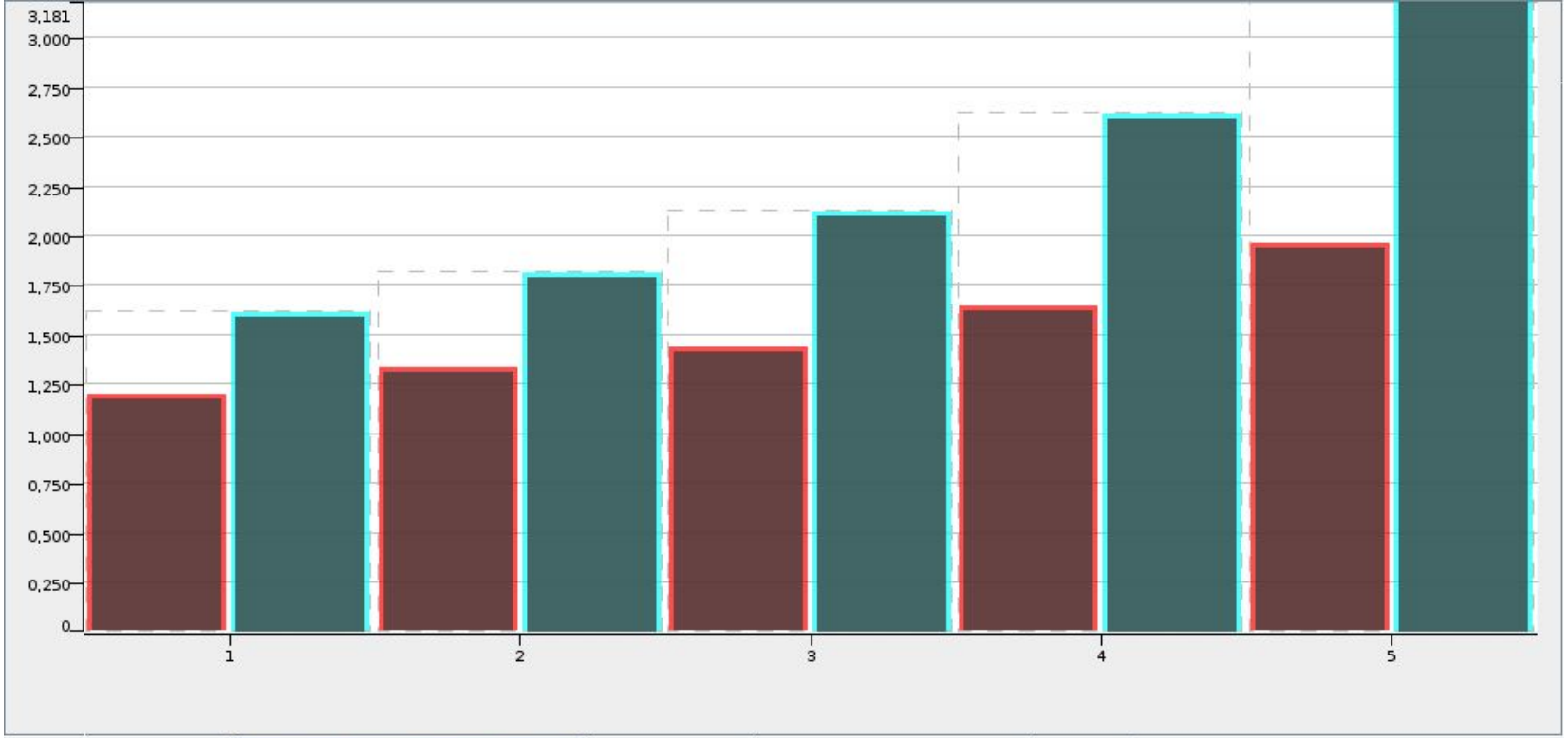
Show missing value bin

Show invalid value bin

Prosiguiendo las observaciones

- Buscando correlaciones entre variables
  - Correlación entre las notas de las evaluaciones.
  - Correlación entre la educación de los padres.
  - Correlación entre las notas y las asignaturas suspensas, el tiempo de estudio, la educación de los padres.
  - Correlación entre la vida social y el tiempo libre.
  - **Correlación entre el consumo de alcohol y la vida social.**





Default Settings | Column/Aggregation settings | Bin settings | Visualization settings | Details

Aggregation method:

- (•) Average
- ( ) Sum
- ( ) Row count
- ( ) Row count (w/o missing values)

Binning column: goout

Aggregation column:

Available columns
- goout
- health
- absences
- G1
- G2

add >>

<< remove

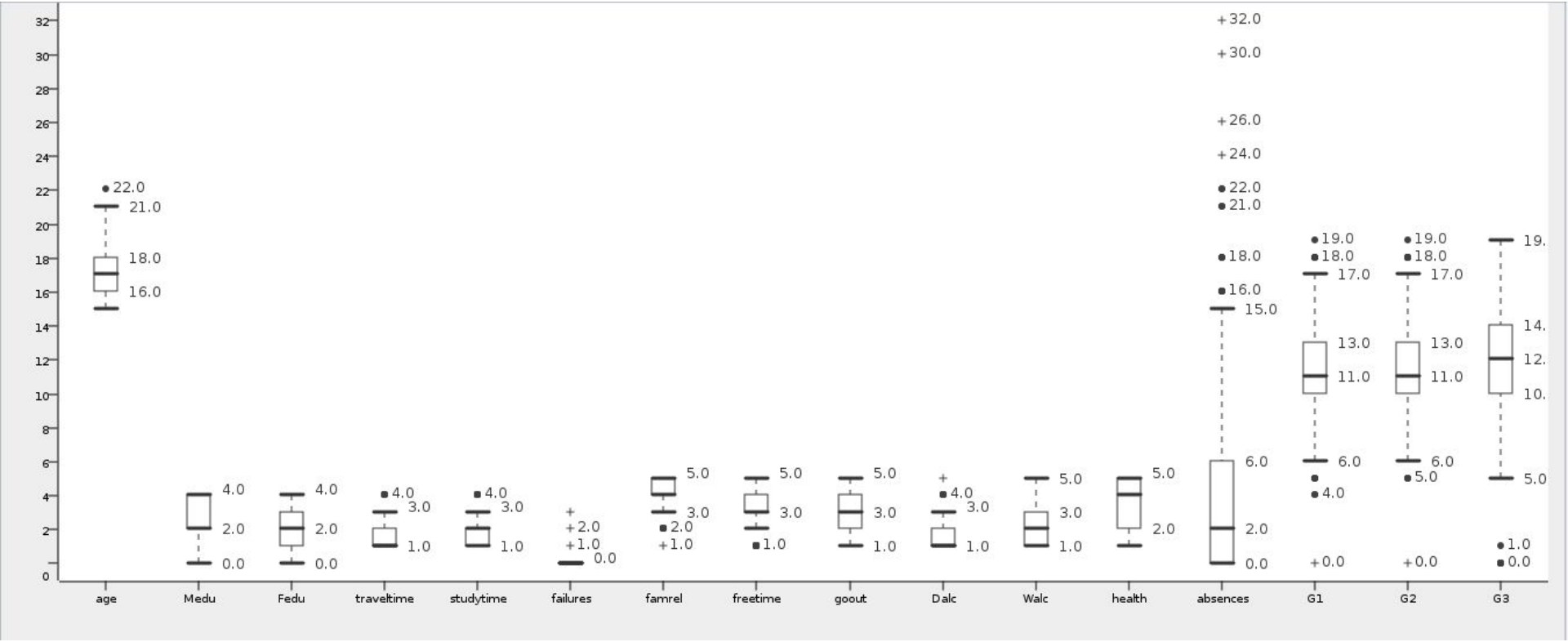
Aggregation colu...
- Dalc
- Walc

Conclusiones del análisis exploratorio

- No se observan correlaciones que apoyen las primeras observaciones.

- Al no tener prácticamente variables numéricas y ninguna continua algunos métodos como Scatter plot o Box plot no funcionan muy bien con este dataset.

- Hay que tener cuidado con los histogramas:
  - Considerar los frecuencias de cada valor. Ej. Edad => consumo muy alto en edades > 20 pero muy pocos individuos de más de 20 años.
  - Puede arrojar conclusiones poco significativas.


32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
22.0
21.0
18.0
16.0
4.0
2.0
0.0
4.0
2.0
0.0
4.0
3.0
1.0
4.0
3.0
1.0
3.0
2.0
1.0
0.0
5.0
3.0
2.0
1.0
5.0
3.0
1.0
5.0
3.0
1.0
4.0
3.0
1.0
5.0
3.0
1.0
5.0
2.0
32.0
30.0
26.0
24.0
22.0
21.0
18.0
16.0
15.0
6.0
2.0
0.0
19.0
18.0
17.0
13.0
11.0
6.0
4.0
0.0
19.0
18.0
17.0
13.0
11.0
6.0
5.0
0.0
19.
14.
12.
10.
5.0
1.0
0.0
age
Medu
Fedu
traveltime
studytime
failures
famrel
freetime
goout
Dalc
Walc
health
absences
G1
G2
G3

Reglas de Asociación

¿Por qué reglas de asociación?

- Clustering:
  - Variables de distintos tipos => Función de distancia mediante combinación de varias y compleja.
  - Resultados muy dependientes de esta función.
- Clasificación:
  - Ampliamente utilizada en este dataset, interesante probar alguna técnica nueva.
- Regresión:
  - Interesante para predecir variables continuas. Modelo predictivo.
  - Atributos a predecir en este dataset no son numéricos ni tampoco lo son la mayoría de las variables del dataset.
- Reglas de asociación:
  - Puede aportar resultados donde las demás técnicas fallen.
  - Describe dependencias significativas sin conocimiento previo o realizar suposiciones sobre los datos. Modelo descriptivo.
  - Trabajan bien con variables nominales como las de este dataset

Preparación de los datos para asociación

- Para trabajar con reglas de asociación es necesario usar bases de datos transaccionales.
- Para ello hay que discretizar variables creando intervalos para las variables numéricas. Para ellas se ha usado la distribución por cuartiles en general.
- Una vez discretizados se han sustituido los valores por los característicos de una BD transaccional: *PrefijoAtrib-Valor*

Row ID		freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
Row0	no	FreeT-Medio	GO-Alto	DA-Muy bajo	WA-Muy bajo	HLT-Normal	ABS-Bastantes	G1-Suspenso	G2-Aprobado	G3-Aprobado
Row1	e	FreeT-Medio	GO-Medio	DA-Muy bajo	WA-Muy bajo	HLT-Normal	ABS-Pocas	G1-Suspenso	G2-Aprobado	G3-Aprobado
Row2	no	FreeT-Medio	GO-Bajo	DA-Bajo	WA-Medio	HLT-Normal	ABS-Bastantes	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row3		FreeT-Bajo	GO-Bajo	DA-Muy bajo	WA-Muy bajo	HLT-Muy b...	ABS-Cero	G1-Notable	G2-Notable	G3-Notable
Row4	no	FreeT-Medio	GO-Bajo	DA-Muy bajo	WA-Bajo	HLT-Muy b...	ABS-Cero	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row5	e	FreeT-Alto	GO-Bajo	DA-Muy bajo	WA-Bajo	HLT-Muy b...	ABS-Bastantes	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row6	no	FreeT-Alto	GO-Alto	DA-Muy bajo	WA-Muy bajo	HLT-Normal	ABS-Cero	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row7	no	FreeT-Muy ...	GO-Alto	DA-Muy bajo	WA-Muy bajo	HLT-Muy ...	ABS-Pocas	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row8	no	FreeT-Bajo	GO-Bajo	DA-Muy bajo	WA-Muy bajo	HLT-Muy ...	ABS-Cero	G1-Notable	G2-Notable	G3-Sobresa...
Row9	e	FreeT-Muy ...	GO-Muy b...	DA-Muy bajo	WA-Muy bajo	HLT-Muy b...	ABS-Cero	G1-Aprobado	G2-Aprobado	G3-Aprobado

Preparación de los datos para asociación

- Para obtener mejores resultados se ha creado un atributo **Alc** que sintetize Dalc y Walc. En él pesa más Dalc al ser más significativo.
- Los valores de este nuevo atributo se recogen en la siguiente tabla:

Dalc \ Walc	Muy bajo	Bajo	Medio	Alto	Muy alto
Muy bajo	Muy bajo	Muy bajo	Bajo	Medio	Alto
Bajo	Bajo	Bajo	Medio	Medio	Alto
Medio	Bajo	Medio	Medio	Alto	Muy alto
Alto	Medio	Alto	Alto	Alto	Muy alto
Muy Alto	Alto	Alto	Muy alto	Muy alto	Muy alto

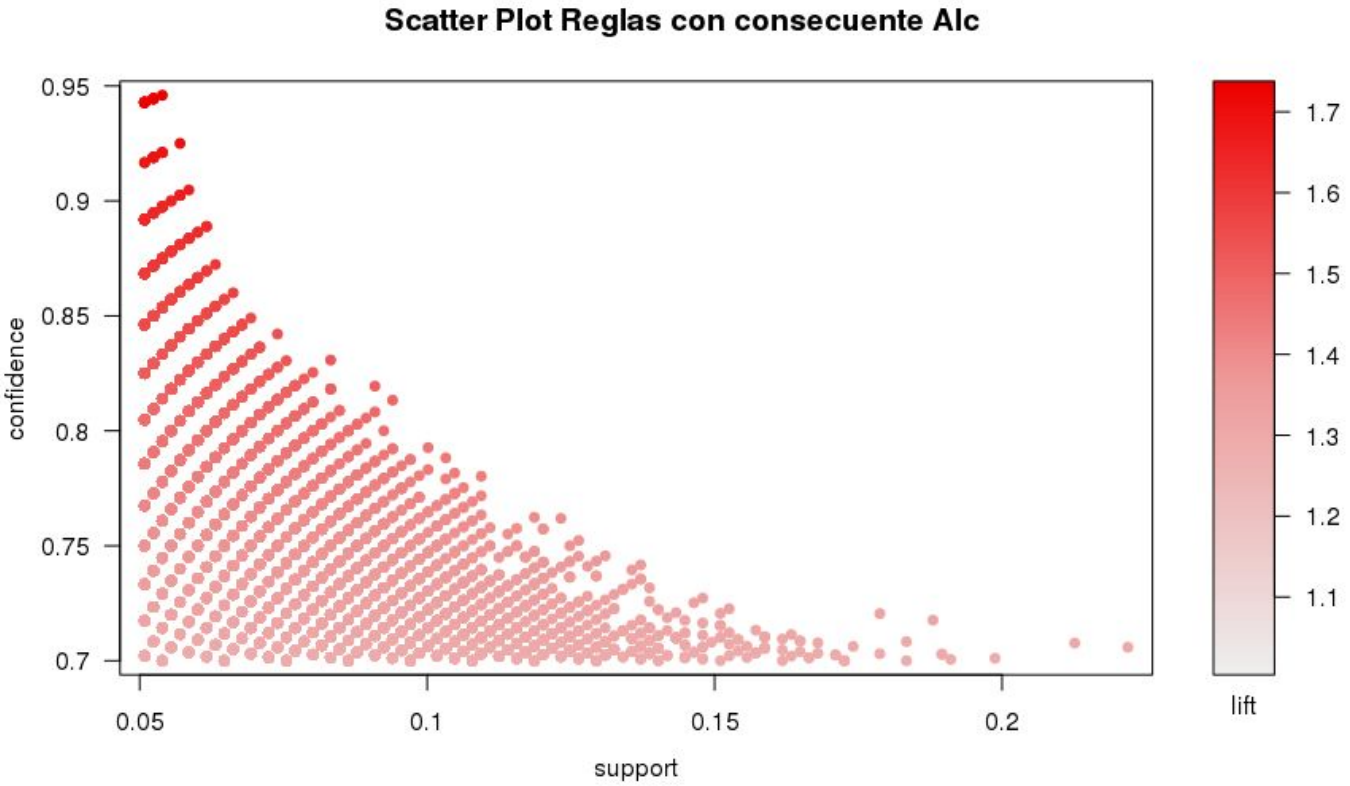
Extracción Reglas

- Hemos comenzado usando el nodo “Association Rule Learner” de **Knime** para extraer reglas de asociación.
  - Con support mínimo de 0,12 y confianza mínima de 0,9 salen 328.217 reglas.
  - Con support mínimo de 0,12 y confianza mínima de 0,7 salen 620.498 reglas.
- Problemas de ejecución debido al excesivo número de reglas:
  - Elevado consumo de RAM: hasta 5,8 GB.
  - Tiempos de ejecución elevados: hasta 2 h.
- Usando el nodo “Apriori” de **Weka** en Knime se obtienen las 100 mejores reglas.
  - Con support mínimo 0,12 y confianza mínima: 0,9 y 0,7.
  - Reducción de problemas de ejecución al ser mejor algoritmo y filtrar las reglas.
  - Se puede seleccionar sólo las reglas con un consecuente concreto, lo cual es interesante para extraer reglas de Alc

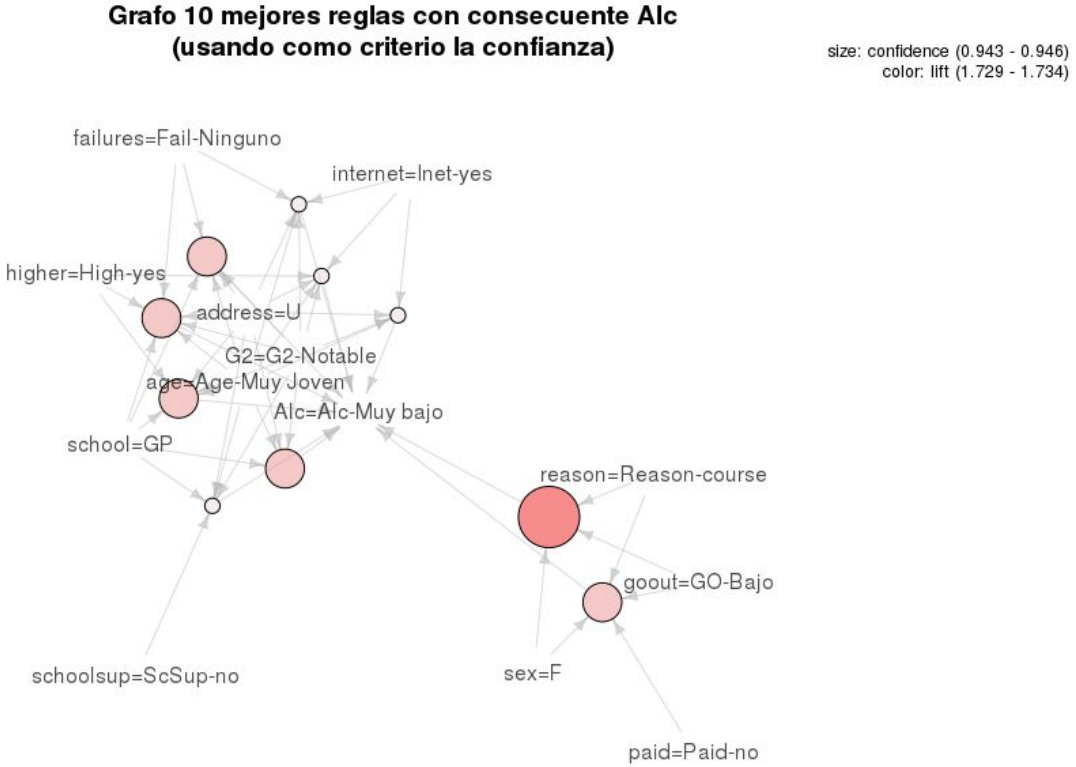
Extracción Reglas

- Usando **R** y el algoritmo “Apriori” del paquete *arules* y extrayendo solo reglas que tengan como consecuente Alc.
  - Con support mínimo de 0,05 y confianza mínima de 0,7 salen 64.815 reglas.
- Menos problemas de ejecución que Knime pero también consumos y tiempos considerables
- De todas las reglas generadas, todas tienen como consecuente Alc = *Alc-Muy bajo*

Resultados

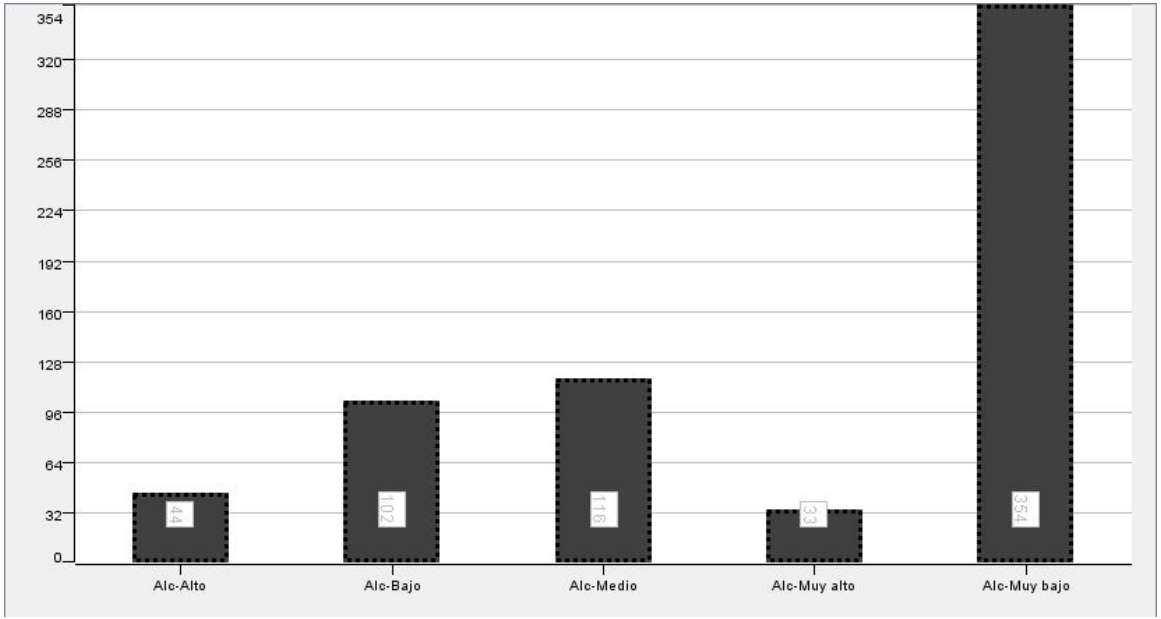

Scatter Plot Reglas con consecuente Alc
confidence
0.95
0.9
0.85
0.8
0.75
0.7
0.05
0.1
0.15
0.2
support
lift
1.7
1.6
1.5
1.4
1.3
1.2
1.1

Resultados


Grafo 10 mejores reglas con consecuente Alc
(usando como criterio la confianza)
size: confidence (0.943 - 0.946)
color: lift (1.729 - 1.734)
failures=Fail-Ninguno
internet=Inet-yes
higher=High-yes
address=U
G2=G2-Notable
age=Age-Muy Joven
Alc=Alc-Muy bajo
school=GP
reason=Reason-course
goout=GO-Bajo
schoolsup=ScSup-no
sex=F
paid=Paid-no

Conclusiones parciales

- Solo se obtienen reglas de Alc = *Alc-Muy bajo* que si se observa más detenidamente se debe al desequilibrio de esta clase que se puede observar en la siguiente gráfica.

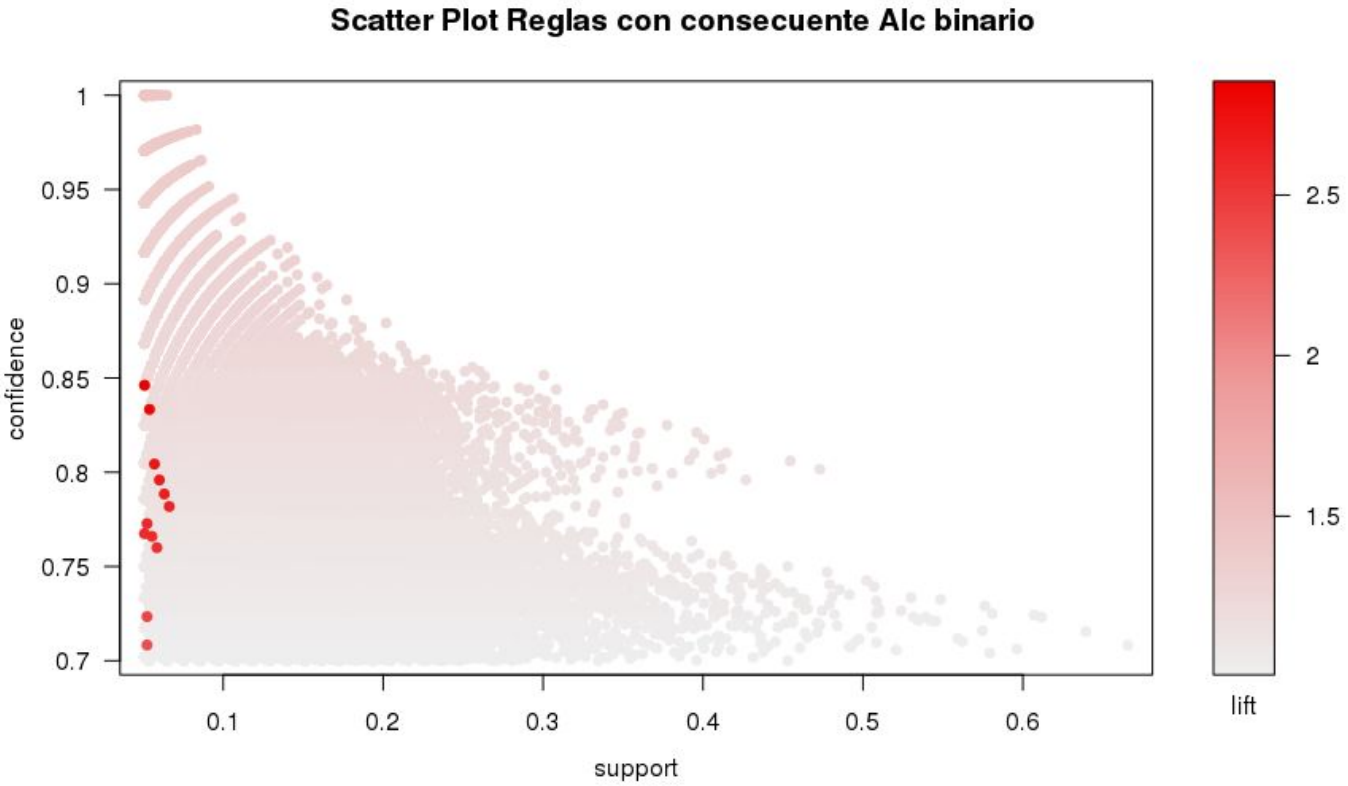


- **Solución:** Binarizar el atributo Alc considerando:
  - ***Alc-Si*:** englobará *Alc-Muy alto*, *Alc-Alto* y *Alc-Medio* (193).
  - ***Alc-No*:** englobará *Alc-Muy bajo* y *Alc-Bajo (456).*

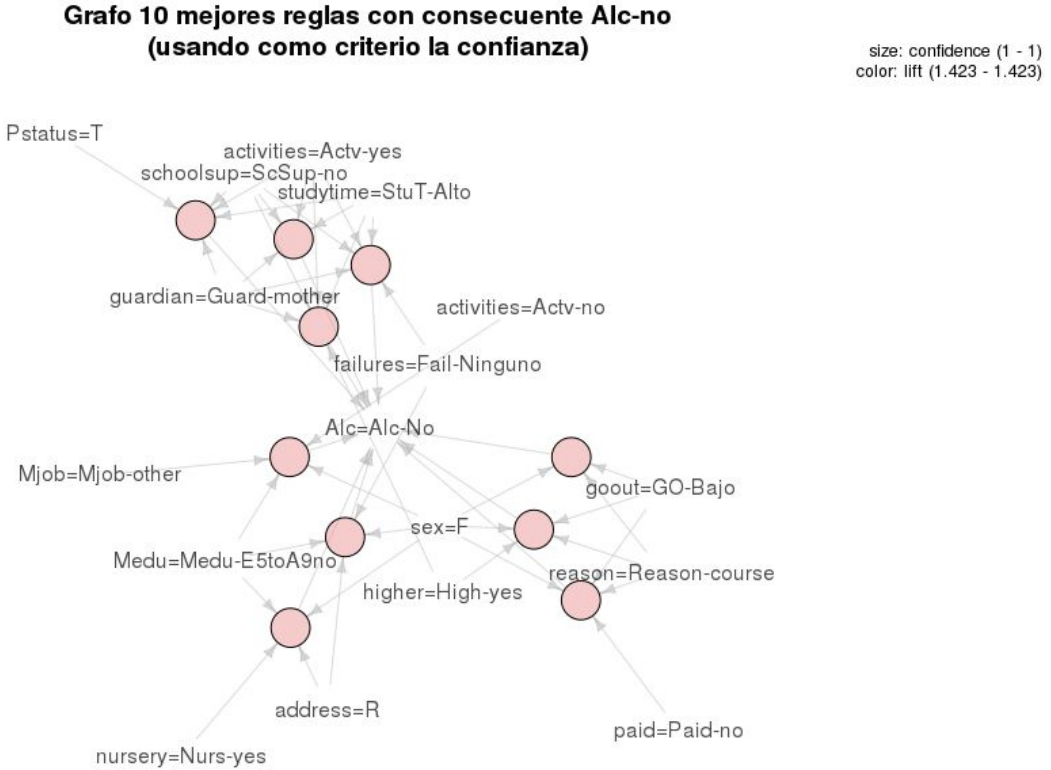
Extracción Reglas con Alc binarizado

- Usando **Knime** ni siquiera llega a finalizar porque ni filtrando se obtienen reglas en tiempos razonables.
- Usando el nodo “Apriori” de **Weka** en Knime se obtienen las 100 mejores reglas.
  - Con support mínimo 0,05 y confianza mínima: 0,7.
  - Filtrando por el valor Alc se pueden obtener reglas para *Alc-No* y *Alc-Si*.
- Usando **R** y el algoritmo “Apriori” del paquete *arules* y extrayendo solo reglas que tengan como consecuente Alc.:
  - Con support mínimo 0,05 y confianza mínima: 0,7 se obtienen 847.840 reglas
  - Filtrando por el valor Alc se pueden obtener reglas para *Alc-No* y *Alc-Si*.

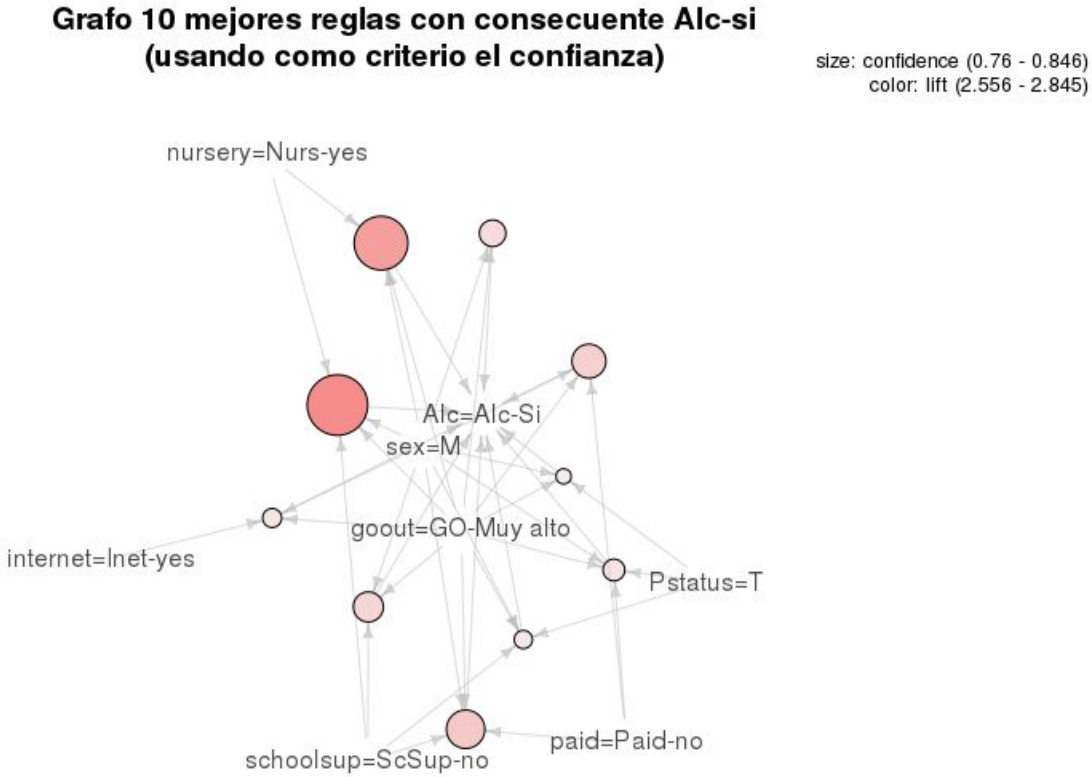
Resultados


Scatter Plot Reglas con consecuente Alc binario
confidence
1
0.95
0.9
0.85
0.8
0.75
0.7
0.1
0.2
0.3
0.4
0.5
0.6
support
2.5
2
1.5
lift

Resultados


Grafo 10 mejores reglas con consecuente Alc-no
(usando como criterio la confianza)
size: confidence (1 - 1)
color: lift (1.423 - 1.423)
Pstatus=T
activities=Actv-yes
schoolsup=ScSup-no
studytime=StuT-Alto
guardian=Guard-mother
activities=Actv-no
failures=Fail-Ninguno
Alc=Alc-No
Mjob=Mjob-other
goout=GO-Bajo
sex=F
Medu=Medu-E5toA9no
reason=Reason-course
higher=High-yes
address=R
paid=Paid-no
nursery=Nurs-yes

Resultados


Grafo 10 mejores reglas con consecuente Alc-si
(usando como criterio el confianza)
size: confidence (0.76 - 0.846)
color: lift (2.556 - 2.845)
nursery=Nurs-yes
Alc=Alc-Si
sex=M
goout=GO-Muy alto
internet=Inet-yes
Pstatus=T
paid=Paid-no
schoolsup=ScSup-no

Conclusiones

- R es más potente que Knime y consume menos recursos.
- Perfil del **no consumidor de alcohol**:
  - Mujeres
  - Interesadas por sus estudios: quieren ir a la universidad, sacan buenas notas, no van a particulares...
  - Poca vida social
  - Familia con más de tres miembros
- Perfil del **consumidor de alcohol**:
  - Hombre
  - Mucha vida social
  - Sus padres viven juntos
  - No se ha encontrado relación directa entre el consumo de alcohol y el desempeño académico

Referencias

- P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance.* In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
- F. Pagnotta, H. Mohammad Amran. *Using Data Mining To Predict Secondary School Student Alcohol Consumption*. Department of Computer Science,University of Camerino, February, 2016
- Sistema de calificaciones portugués. http://internacional.ugr.es/pages/conversion-calificaciones/tablaconversioncalificaciones/!
- Documentación de Knime https://www.knime.org/nodeguide
- Documentación de R
- M. Amparo Vila. *Apuntes de la asignatura Tratamiento Inteligente de Datos*, Máster Universitario en Ingeniería Informática. Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada

Fin

¿Preguntas?

Puede encontrar este trabajo en GitHub:

https://github.com/AythaE/Estudiantes-TID