Universidad de Granada

Máster Universitario en Ingeniería Informática

Tratamiento Inteligente de Datos

STUDENT ALCOHOL CONSUMPTION Dataset

Agüero Torales, Marvin Matías
Estévez Olivas, Aythami

2016-17

Índice de los contenidos

Introducción

Origen del dataset

El dataset utilizado se ha recolectado entre estudiantes portugueses de dos institutos en el curso 2005-2006. Está construido por dos fuentes, una concerniente a los datos académicos de los alumnos, la otra por encuestas realizadas a estos, con preguntas sobre temas demográficos, sociales, personales y emocionales. Finalmente los datos se agruparon en dos datasets, uno con los alumnos de portugués (649 ítems) y otro con alumnos de matemáticas (395 items).

Puesto que estos dos datasets tienen bastantes ítems comunes, en concreto 379 en total, el trabajo se ha llevado a cabo con el dataset de portugués, el más numeroso.

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 de hogar	Binario	"U" Urbano o "R" Rural
famsize	Tamaño de la familia	Binario	“LE3” menor o igual a 3 o “GT3” mayor a 3
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

Atributo	Descripción	Tipo	Valores Posibles
guardian	Tutor	Nominal	“mother” madre, “father” padre u “other” otro
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 del instituto	Binario	“Yes” Sí o “No” No
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 los 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 primer periodo académico	Numérico	0 a 20
G2	Calificación del segundo periodo académico	Numérico	0 a 20
G3	Calificación del periodo académico final	Numérico	0 a 20

Tabla 1: Descripción de los atributos del dataset

Aunque parece haber muchos atributos numéricos la mayoría son valoraciones graduadas entre *muy malo/bajo* y *muy bueno/alto*, aunque se representan con números del 1 al 5, son atributos nominales en realidad. Los únicos atributos realmente numéricos son:

- Edad, de 15 a 22 años
- Notas de los alumnos (G1, G2, G3), de 0 a 20
- Ausencias, de 0 a 93
- Asignaturas suspensas el curso pasado (failures), aunque solo es numérico hasta 3 ya que el 4 representa más de 3 unidades.

Otros trabajos con este dataset

Using Data Mining To Predict Secondary School Student Alcohol Consumption

Este trabajo data del 2016, los autores emplean árboles de decisión (Random Forest en concreto) para predecir el consumo de alcohol de los estudiantes, llegando a un acierto cercano al 92%.

Cabe destacar que (Pagnotta & Hossain, 2016), sintetizan las dos variables relacionadas con el consumo de alcohol, Dalc y Walc, en una sola *Alc*, mediante una media ponderada inventada por ellos mismos en la que dan más peso a la variable Dalc como puede apreciarse en la figura 1:

$$Alc = \frac{Walc \times 2 + Dalc \times 5}{7}$$

Figura 1: Media ponderada del alcohol usada en (Pagnotta & Hossain, 2016)

Using Data Mining to Predict Secondary School Student Performance

Este trabajo corresponde a los autores originales del dataset, donde utilizan diversas técnicas, desde árboles de decisión, a redes neuronales pasando por máquinas de soporte vectorial, incluso regresión, para predecir las notas finales de los estudiantes (correspondiente al atributo G3). Demuestran que se puede predecir con bastante certeza las notas finales de los alumnos si se tienen las notas pasadas de los mismos (los atributo G1 y G2 respectivamente) (Cortez & Silva, 2008).

Análisis exploratorio

Nuestra hipótesis

Comprobar que atributos tienen relación con el consumo de alcohol en los estudiantes de secundaria para realizar posteriormente técnicas de minería de datos que permitan extraer conocimiento de este dataset.

Primeras impresiones

Realizamos un análisis exploratorio de datos utilizando un WorkFlow en Knime según los apuntes de preprocesamiento (Vila, 2014).

Utilizando el nodo "Statistic" de Knime comprobamos que no hay valores perdidos.

Con histograma interactivo hemos comprobado que el consumo medio de alcohol se dispara a partir de los 21 años y es curioso un cambio de tendencia, mientras que los menores de 21 beben más en fines de semana los mayores beben prácticamente igual o incluso más entre diario como se puede ver en la siguiente figura en la que se aprecia en rojo el consumo de alcohol en días de diario (Dalc) y en azul el consumo de alcohol en fin de semana (Walc) respecto a la edad (entre 15 y 22 años):

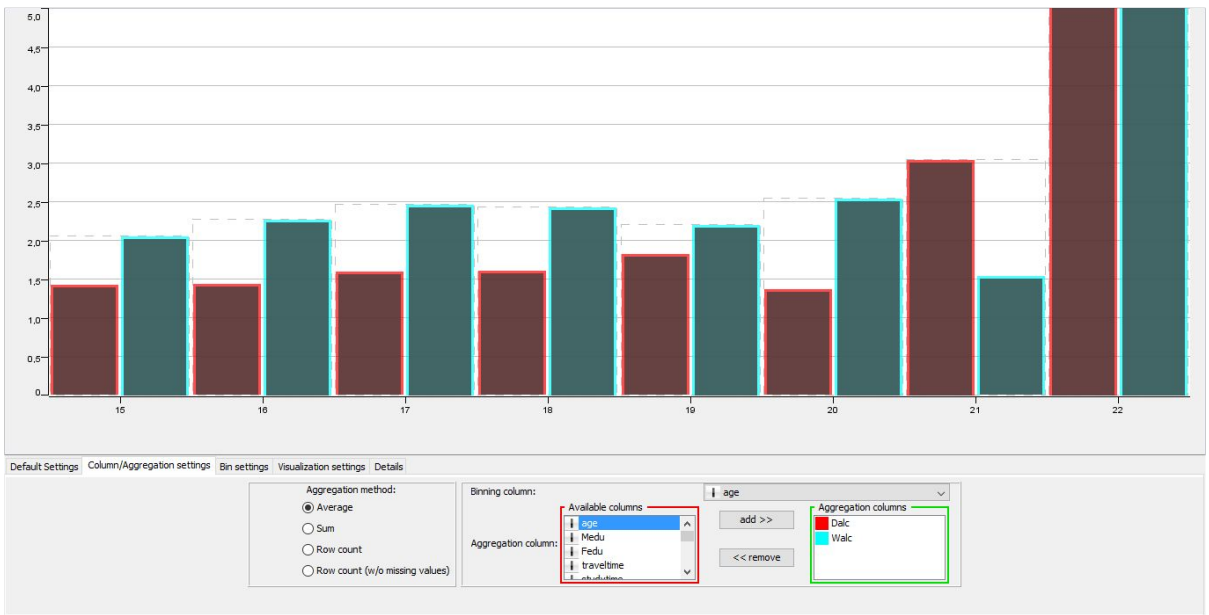


Figura 2: Consumo de alcohol entre diario y en fin de semana respecto a la edad

También se observa un mayor consumo en hombres pero tampoco muy significativo como se aprecia en la siguiente figura:

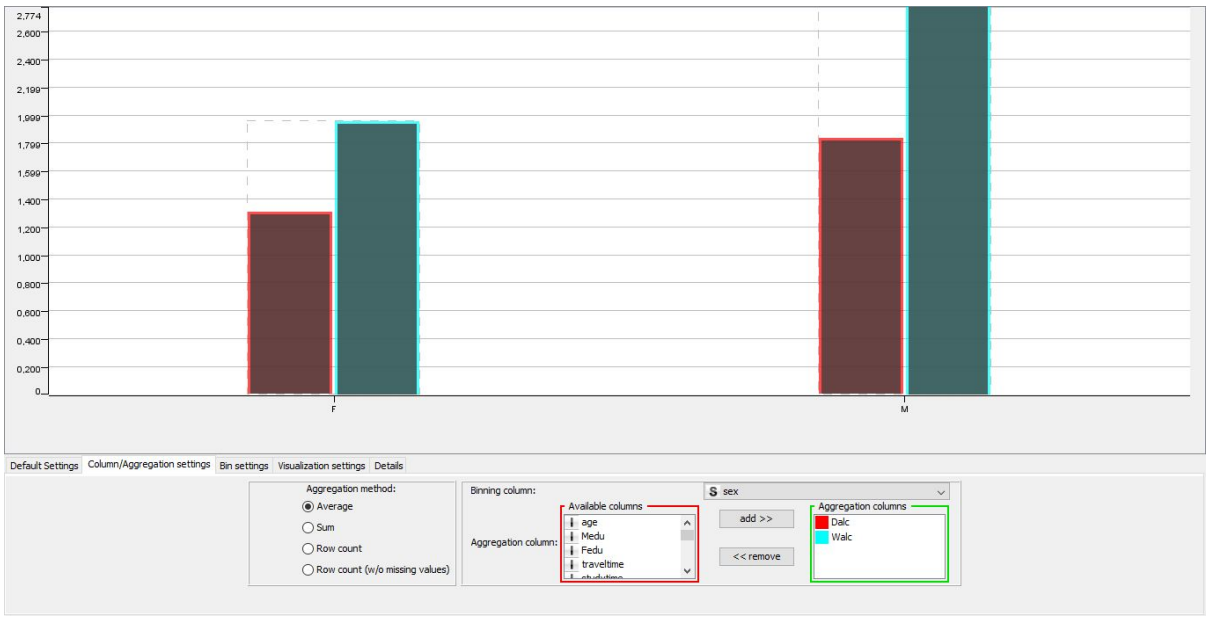


Figura 3: Consumo de alcohol respecto al sexo

Se puede observar que los alumnos que estudian menos de 2 horas a la semana, es decir que su variable studytime tiene un valor de 1, tienen un consumo de alcohol algo superior a los demás como se puede apreciar en la figura 4. Esto unido a que el consumo se incremente en los alumnos que obtienen unas notas finales por debajo de 6 (un 3 en nuestro sistema de calificaciones) (Sistema de calificaciones portugués, n.d.) como se aprecia en la figura 5 en la que aparecen las notas finales agrupadas en intervalos de dos valores, por ejemplo (4, 6]; nos permite extraer que los alumnos con peores rendimientos académicos parecen tender a un mayor alcoholismo.

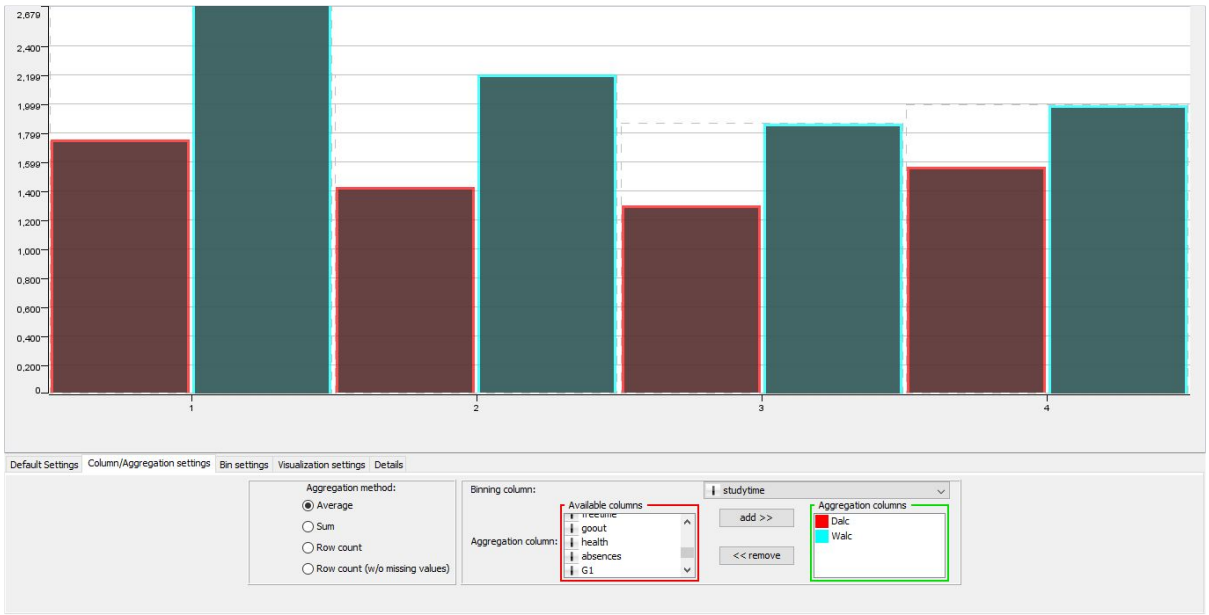


Figura 4: Consumo de alcohol respecto al tiempo de estudio (studytime)

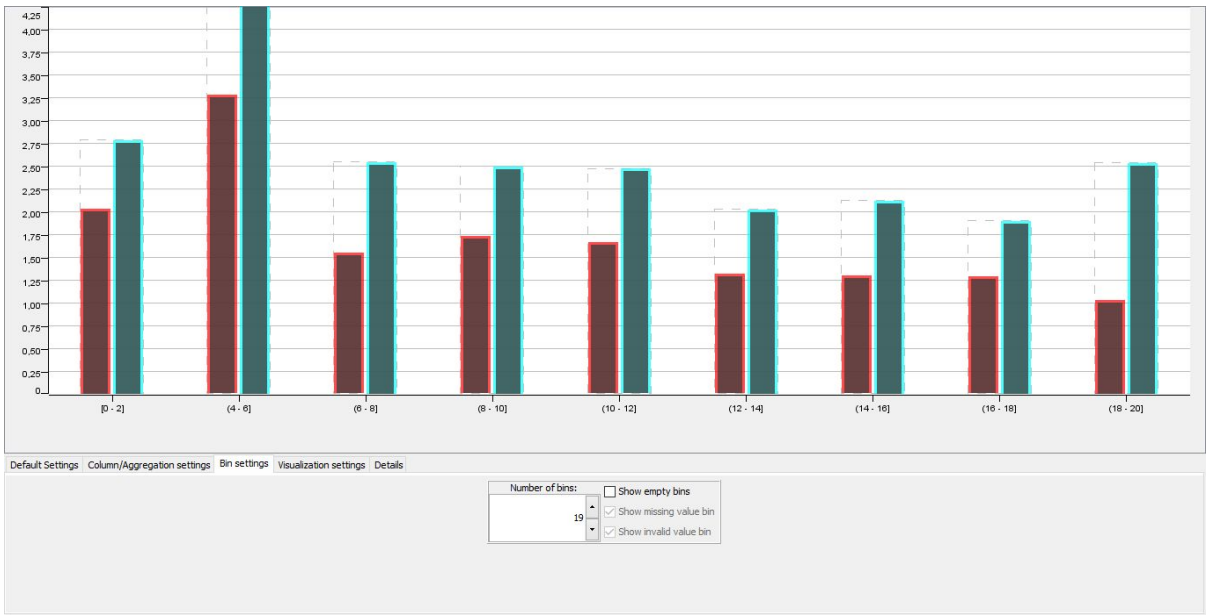


Figura 5: Consumo de alcohol respecto a las notas finales agrupadas en intervalos (G3)

Estudio de la correlación entre atributos

Para comprobar las observaciones realizadas mediante los histogramas pasamos a realizar una correlación lineal entre las variables con el objetivo de encontrar que las variables observadas tienen cierta correlación con el consumo de alcohol y además comprobar si hay algunas variables con una correlación muy alta lo que puede indicar que se derivan unas de otras y se pueden eliminar del dataset al aportar la misma información.

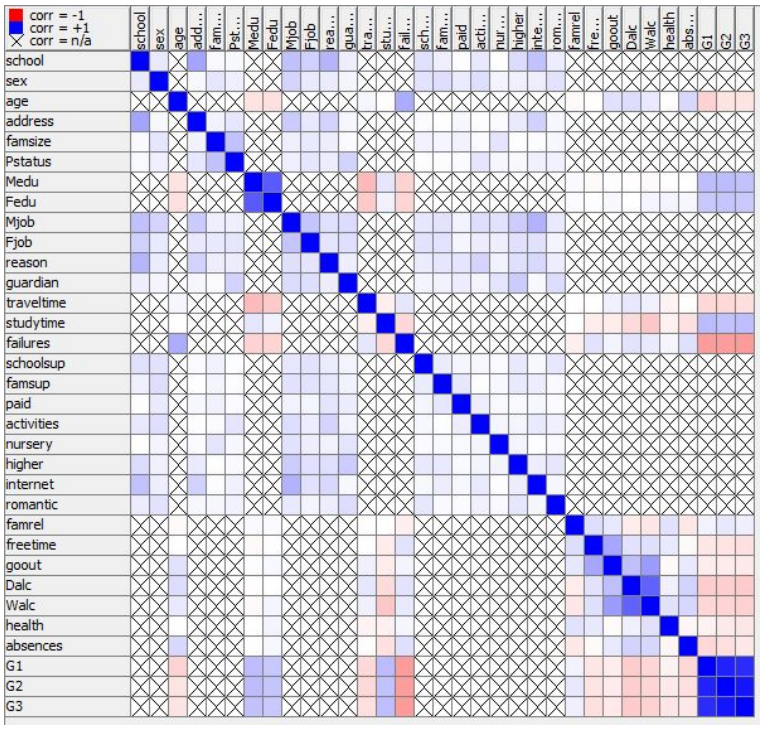


Figura 6: Matriz de correlación lineal entre variables

Como se puede apreciar existen muchas X en esta matriz lo que indica que no existe correlación entre las variables, esto se puede deber a que muchas variables son categóricas y es complicado realizar una correlación entre variables categóricas o entre una variable categórica y otra numérica al no saber cuales son los valores superiores o inferiores.

Las correlaciones más importantes que se observan son:

- Correlación entre las notas de las evaluaciones (G1, G2 y G3).
- Correlación entre la educación de los padres (Medu y Fedu).
- Correlación negativa entre las notas y las asignaturas suspensas el año pasado (failures); positiva entre las notas, el tiempo de estudio (studytime); positiva entre las notas y la educación de los padres.
- Correlación entre la vida social (gout) y el tiempo libre (freetime).
- **Correlación entre el consumo de alcohol y la vida social.**

Destaca esta última correlación que si se observa en un histograma la variable goout y el consumo de alcohol es prácticamente lineal, como se aprecia en la siguiente figura:

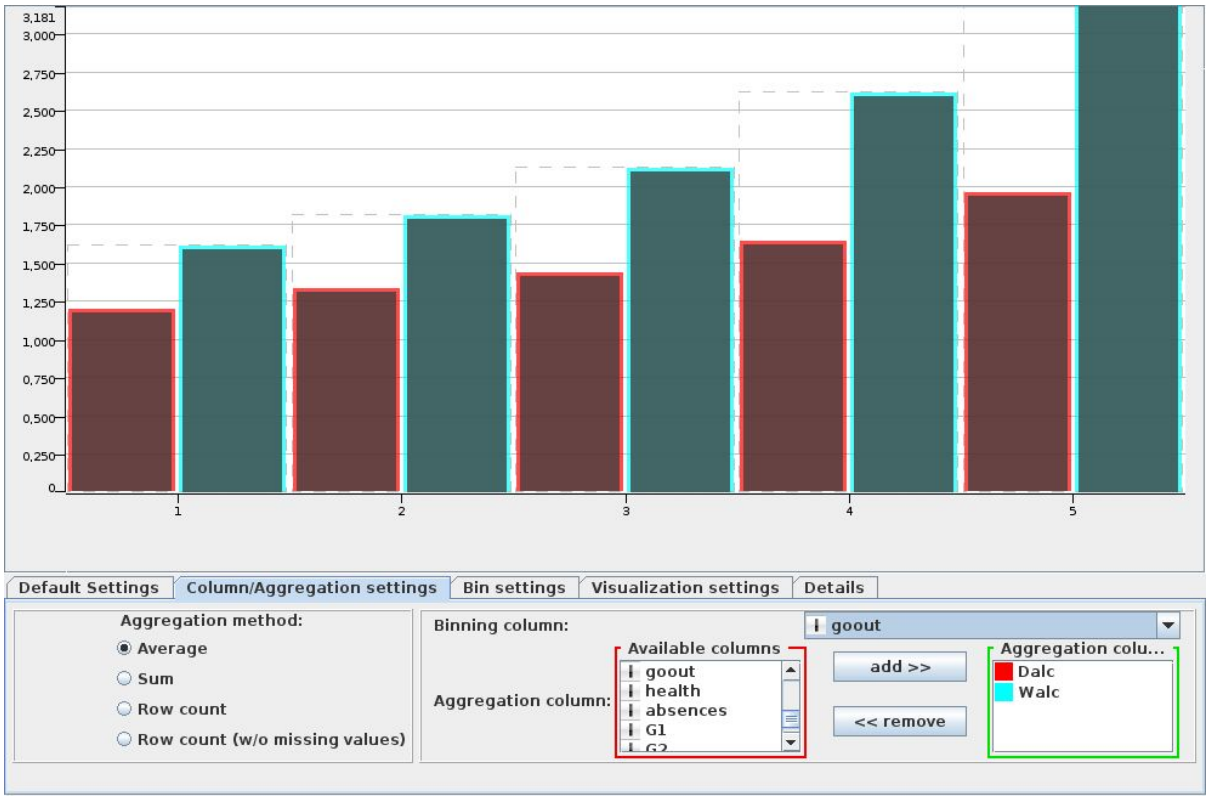


Figura 7: Consumo de alcohol respecto a la vida social (goout)

Conclusiones del análisis exploratorio

No se observan correlaciones destacables entre el consumo de alcohol y las variables mencionadas en las primeras observaciones.

Algunas técnicas exploratorias que funcionan muy bien con otros datasets como el box plot o el scatter plot no funcionan demasiado bien con este dataset, probablemente debido a que apenas contiene variables numéricas realmente, que no sean categóricas representadas como numéricas como es el caso de los consumos de alcohol, y ninguna continua.

Hay que tener cuidado con las conclusiones extraídas con los histogramas únicamente, ya que pueden ser engañosas si no existen suficientes individuos con dichos valores. Un ejemplo de esto lo encontramos con la edad, que como hemos dicho el consumo de alcohol se dispara a partir de los 20 años, pero observando el box plot de la figura 8 se observa que el tercer cuartil se establece a los 18 años y se solo se encuentra un individuo de 22 años considerado como un outlayer, con lo que las conclusiones extraídas no son significativas.

A continuación se puede ver el box plot de este dataset, como se aprecia hay muchos atributos que van entre 1 y 5 al ser variables categóricas representadas como numéricas:

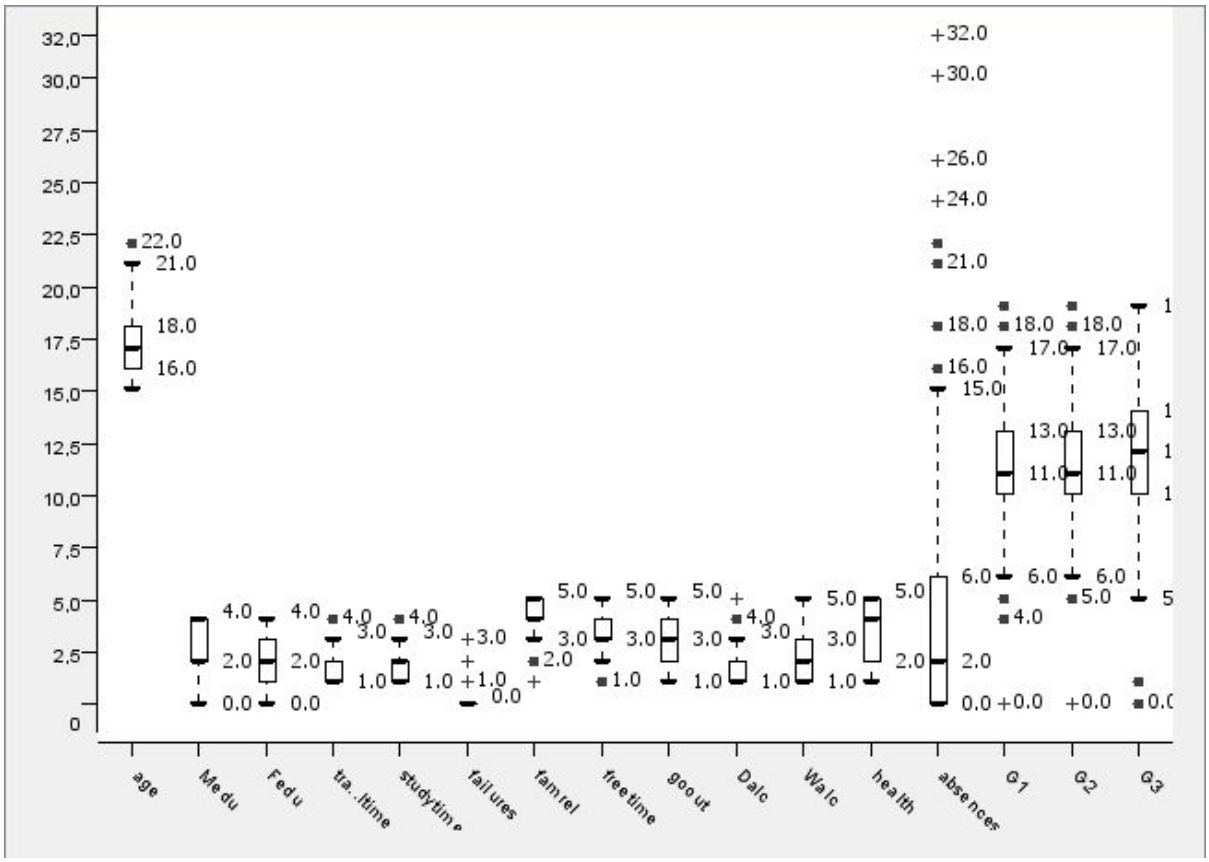


Figura 8: Box plot

Reglas de asociación

¿Por qué reglas de asociación?

Clustering

El clustering necesita de una función de distancia y/o semejanza para extraer la matriz de proximidad, estas funciones trabajan muy bien con variables numéricas o incluso nominales ordinales, pero no tan bien con variables nominales no ordinales (Vila, 2014).

En cualquier caso teniendo en cuenta que este dataset tiene variables muy distintas, unas pocas numéricas, principalmente nominales, algunas binarias, otras ordinales (sobre todo valoraciones personales) y algunas no ordinales (como el de los trabajo padres: Mjob y Fjob), donde la selección de la distancia con variables de tipos tan dispares requeriría la combinación de varias distancias de manera cuidadosa y los resultados obtenidos dependerá mucho de estas decisiones.

Clasificación

Las técnicas de clasificación han sido ya ampliamente utilizadas con este dataset, como se puede comprobar en los dos papers mencionados anteriormente, por ello resulta interesante probar con técnicas distintas y poder cotejar de cierta forma las conclusiones obtenidas.

Regresión

La regresión es interesante para predecir variables continuas en función de otras variables independientes de tipo numérico al menos, ya que constituye un modelo predictivo que busca una relación en forma de función (Vila, 2014).

Teniendo en cuenta que los atributos de este dataset son principalmente nominales y que la variable que se quiere predecir es el consumo de alcohol, que es nominal, no tiene sentido aplicar estas técnicas para llegar a conclusiones respecto al consumo.

Reglas de asociación

Las reglas de asociación son una técnica verdaderamente creada en data mining, que aporta resultados aún donde las demás pueden fallar. Describe dependencias significativas parciales o completas mediante un modelo descriptivo sin necesidad de tener conocimiento previo o hacer suposiciones sobre los datos (Vila, 2014). Puesto que trabaja bien con variables nominales, como las que predominan en este dataset, decidimos extraer reglas para obtener conclusiones respecto al consumo de alcohol.

Preparación de los datos para asociación

Discretización

Para trabajar con reglas de asociación es necesario usar Bases de Datos (BD) transaccionales, para ello hay que discretizar variables creando intervalos para las variables numéricas, utilizando la distribución por cuartiles en general con el Auto-Binner de Knime; por reemplazo de valores como el caso del consumo del alcohol por ejemplo (Dalc, Walc), empleando el String Replace (Dictionary) de Knime o por rangos de valores con el Numeric Binner de Knime, como el caso de la notas (G1, G2, G3). Para esto último, nos basamos en la escala de calificaciones internacional utilizada por la Universidad de Granada (Sistema de calificaciones portugués, 2017).

Una vez discretizados todos los atributos se han sustituido los valores por los característicos de una BD transaccional, de la siguiente manera: *PrefijoAtrib-Valor*, como se puede ver en la Figura 9.

Row ID		freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
Row0	ho	FreeT-Medio	G0-Alto	DA-Muy bajo	WA-Muy bajo	HLT-Normal	ABS-Bastantes	G1-Suspenso	G2-Aprobado	G3-Aprobado
Row1	e	FreeT-Medio	G0-Medio	DA-Muy bajo	WA-Muy bajo	HLT-Normal	ABS-Pocas	G1-Suspenso	G2-Aprobado	G3-Aprobado
Row2	ho	FreeT-Medio	G0-Bajo	DA-Bajo	WA-Medio	HLT-Normal	ABS-Bastantes	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row3		FreeT-Bajo	G0-Bajo	DA-Muy bajo	WA-Muy bajo	HLT-Muy b...	ABS-Cero	G1-Notable	G2-Notable	G3-Notable
Row4	ho	FreeT-Medio	G0-Bajo	DA-Muy bajo	WA-Bajo	HLT-Muy b...	ABS-Cero	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row5	e	FreeT-Alto	G0-Bajo	DA-Muy bajo	WA-Bajo	HLT-Muy b...	ABS-Bastantes	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row6	ho	FreeT-Alto	G0-Alto	DA-Muy bajo	WA-Muy bajo	HLT-Normal	ABS-Cero	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row7	ho	FreeT-Muy ...	G0-Alto	DA-Muy bajo	WA-Muy bajo	HLT-Muy ...	ABS-Pocas	G1-Aprobado	G2-Aprobado	G3-Aprobado
Row8	ho	FreeT-Bajo	G0-Bajo	DA-Muy bajo	WA-Muy bajo	HLT-Muy ...	ABS-Cero	G1-Notable	G2-Notable	G3-Sobresa...
Row9	e	FreeT-Muy ...	G0-Muy b...	DA-Muy bajo	WA-Muy bajo	HLT-Muy b...	ABS-Cero	G1-Aprobado	G2-Aprobado	G3-Aprobado

Figura 9. Atributos discretizados.

Nuevo atributo *Alc*

Para obtener mejores resultados se ha creado un atributo nuevo denominado Alc, que sintetiza los atributos Dalc y Walc, en él tiene más peso Dalc al ser más significativo que Walc, puesto que socialmente el consumo de alcohol los fines de semana es de esperarse. Los valores de este nuevo atributo se recogen en la Tabla 2, utilizando una especie de inferencia difusa con variables lingüísticas.

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

Tabla 2: Valores del atributo Alc en función de Dalc y Walc.

Para crear el nuevo atributo se ha usado un Column Aggregator de Knime para concatenar Dalc y Walc en un solo atributo (Alc) y un String Replace (Dictionary). Al crearse el nuevo atributo, se eliminan las variables Dalc y Walc del dataset, porque en las reglas de asociación salen siempre juntos al estar muy correlacionados como se puede ver en la Figura 6.

Extraer reglas de asociación

Knime

Para realizar la extracción de reglas de asociación hemos comenzado usando el nodo Association Rule Learner de Knime, configurandolo de la siguiente manera:

- Con soporte mínimo de 0,12 y confianza mínima de 0,9 se obtienen 328.217 reglas.
- Con soporte mínimo de 0,155 y confianza mínima de 0,7 se obtienen 620.498 reglas.

No pudimos configurar menores valores porque tuvimos problemas de ejecución en Knime, esto debido al excesivo número de reglas a obtener, inclusive en mucho casos no se pudo finalizar la ejecución, aun cuando se ha incrementando la memoria RAM (Random Access Memory) dedicada hasta 5,8 GB. En cuanto a los tiempos de ejecución, han sido bastante elevados, hasta de 2 horas.

Weka

Usando el nodo Apriori de Weka en Knime se pueden obtener las 100 mejores reglas, incluso se puede indicar que sólo se obtengan las reglas concernientes al atributo Alc. Para ello, se lo configuró de la siguiente manera:

- Con soporte mínimo de 0,12 y confianza mínima de 0,7 a 0,9, con Alc como consecuente.

Con el nodo Apriori de Weka, se percibe la reducción de problemas de ejecución, al ser un mejor algoritmo y puesto que filtra las reglas antes de presentarlas. Un punto muy importante, es que se puede seleccionar un consecuente concreto para la extracción de reglas, lo cual es interesante para extraer reglas de Alc solamente.

R

Hemos usado R y el algoritmo Apriori del paquete arules, indicando que extraiga solo reglas que tengan como consecuente al atributo Alc, configurandolo de la siguiente manera:

- Con soporte mínimo de 0,05 y confianza mínima de 0,7 se obtienen 64.815 reglas.

Pudimos observar que existen menos problemas de ejecución que Knime pero también consumos y tiempos de ejecución considerables.

Primeros resultados

De todas las reglas generadas, todas tienen como consecuente Alc = Alc-Muy bajo, como se puede ver en detalle en este apartado.En él se analizarán las reglas obtenidas con la configuración que se explica previamente, en concreto las siguientes gráficas y tablas hacen referencia a las reglas obtenidas con R ya que permite la representación de reglas más variada.

Antes de pasar a analizar las reglas en la siguiente figura se aprecia un histograma con las frecuencias relativas de los ítems más frecuentes del dataset con un soporte mayor o igual a 0,3. Esto da una idea de cuales son los valores que más se repiten y ayuda a entender mejor las reglas ya que hay ciertos valores como paid = Paid-No que están muy repetidos en el dataset por lo que no aportarían mucha información.

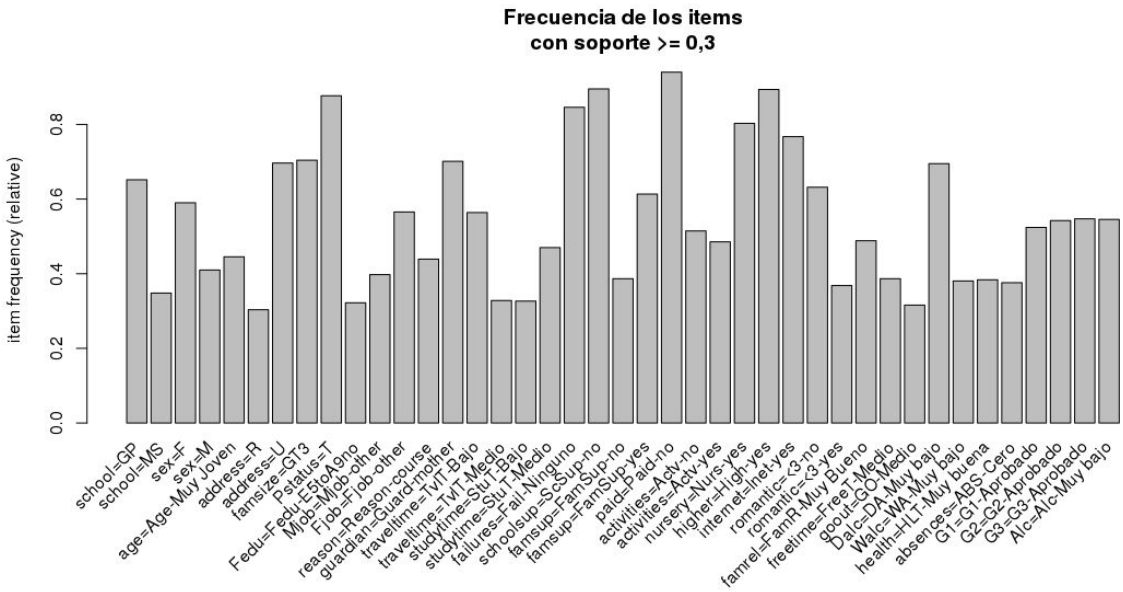
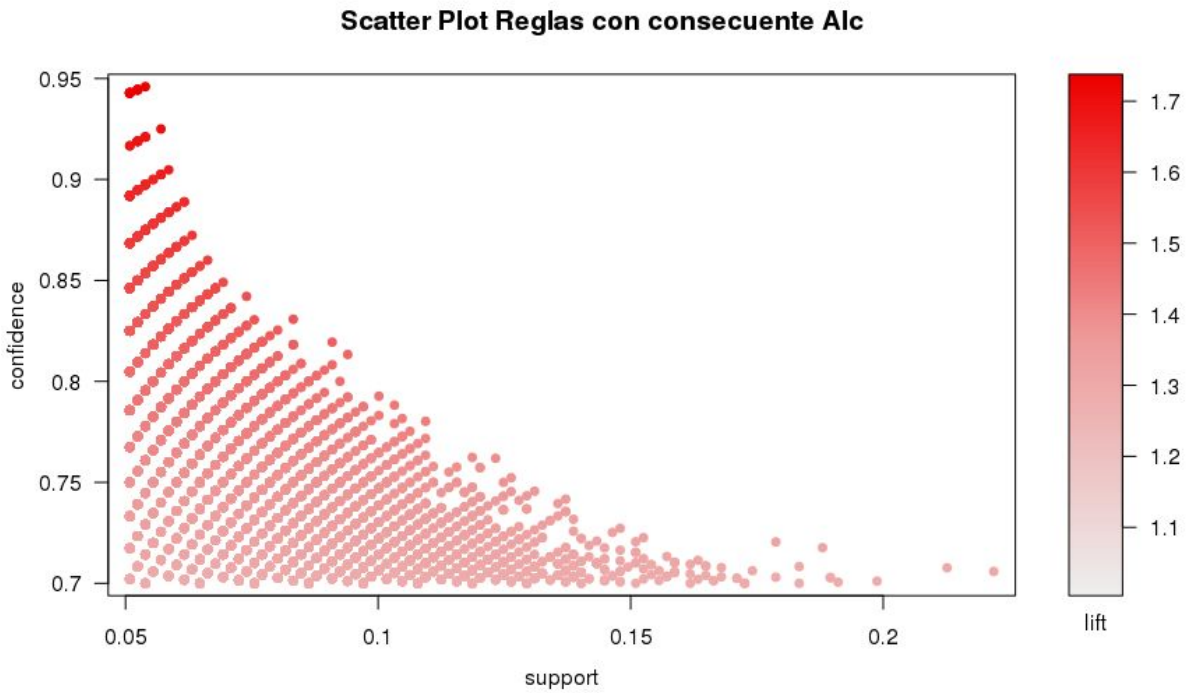


Figura 10: Ítems más frecuentes en el dataset con soporte mayor o igual a 0,3

En la siguiente figura se observa una representación en scatter plot de las reglas obtenidas con R respecto a su confianza (en el eje Y) su soporte (en el eje X) y su lift (medido por un gradiente de color).

Figura 11: Reglas de asociación obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05



En la siguiente figura se aprecia otra representación de estas reglas pero esta vez el color mide el tamaño de las reglas entendido como número de antecedentes más el consecuente.

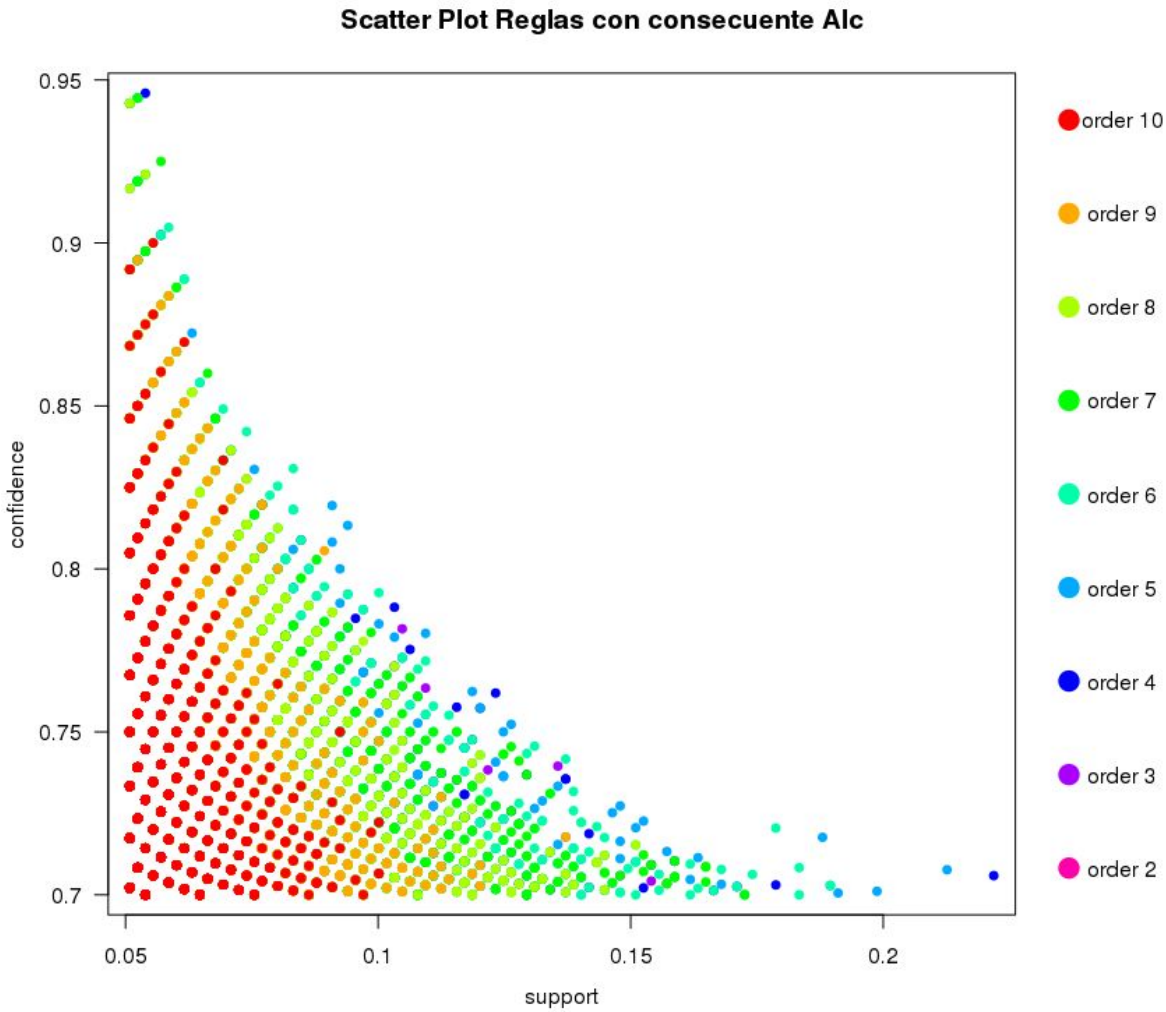


Figura 12: Reglas de asociación obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05 respecto al tamaño (order) de estas.

En esta última figura se muestra una gráfica de grafo en la que se representan las 10 mejores reglas con consecuente Alc obtenidas, entendiendo como mejores reglas aquellas con mejor confianza, ya que como se puede ver en la figura 10 la confianza es casi directamente proporcional al lift por lo que los resultados de ordenar dichas reglas por lift serían prácticamente los mismos.

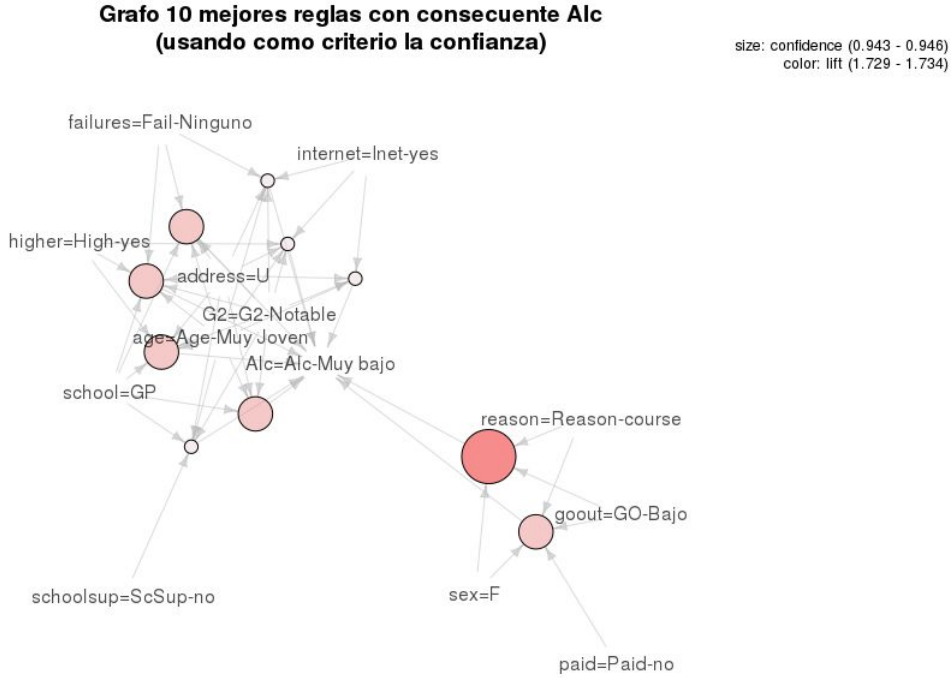


Figura 13: Grafo con las diez mejores reglas de asociación obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05.

En la siguiente tabla se pueden apreciar estas diez mejores reglas en detalle con su soporte, confianza y lift.

	lhs	rhs	support	confidence	lift
[1]	sex=F + reason=Reason-course + goout=GO-Bajo	----> Alc=Alc-Muy bajo	0.05392912	0.9459459	1.734234
[2]	school=GP + age=Age-Muy Joven + address=U + G2=G2-Notable	----> Alc=Alc-Muy bajo	0.05238829	0.9444444	1.731481
[3]	sex=F + reason=Reason-course + paid=Paid-no + goout=GO-Bajo	----> Alc=Alc-Muy bajo	0.05238829	0.9444444	1.731481
[4]	school=GP + age=Age-Muy Joven + address=U + failures=Fail-Ninguno + G2=G2-Notable	----> Alc=Alc-Muy bajo	0.05238829	0.9444444	1.731481
[5]	school=GP + age=Age-Muy Joven +				

	address=U + higher=High-yes + G2=G2-Notable	---->	Alc=Alc-Muy bajo	0.05238829	0.9444444 1.731481
[6]	school=GP + age=Age-Muy Joven + address=U + failures=Fail-Ninguno + higher=High-yes + G2=G2-Notable	---->	Alc=Alc-Muy bajo	0.05238829	0.9444444 1.731481
[7]	age=Age-Muy Joven + address=U + internet=Inet-yes + G2=G2-Notable	---->	Alc=Alc-Muy bajo	0.05084746	0.9428571 1.728571
[8]	school=GP + age=Age-Muy Joven + address=U + schoolsup=ScSup-no + G2=G2-Notable	---->	Alc=Alc-Muy bajo	0.05084746	0.9428571 1.728571
[9]	age=Age-Muy Joven + address=U + failures=Fail-Ninguno + internet=Inet-yes + G2=G2-Notable	---->	Alc=Alc-Muy bajo	0.05084746	0.9428571 1.728571
[10]	age=Age-Muy Joven + address=U + higher=High-yes + internet=Inet-yes + G2=G2-Notable	---->	Alc=Alc-Muy bajo	0.05084746	0.9428571 1.728571

Tabla 3: Diez mejores reglas de asociación extraídas con R con confianza mínima 0,7 y soporte mínimo 0,05

Conclusiones parciales

Como se puede ver en el apartado previo solo se obtienen reglas con consecuente Alc = Alc-Muy bajo, esto se explica si se observa más detenidamente la distribución de valores del atributo Alc que aparece representada en la figura 14. Como se puede ver existe un desequilibrio muy significativo, se observa que más de la mitad de individuos del dataset tiene un valor del atributo Alc = Alc-Muy Bajo. Esto hace que soporte de Alc-Muy bajo sea muy superior al resto de valores y explica porque ninguna regla con soporte superior a 0,05 tiene algún consecuente distinto.

Se nos ocurrió orientar el problema a la extracción de características que determinan el no consumo de alcohol, pero antes decidimos probar con una idea extraída de (Pagnotta & Hossain, 2016), ellos para la mejor predicción del consumo de alcohol con Random Forest llevan a cabo una binarización del atributo englobando el consumo muy bajo y bajo en un valor de no consumo de alcohol y el resto en un valor de consumo de alcohol. Por ello binarizamos la variable Alc de la siguiente manera:

- **Alc-Sí:** englobará Alc-Muy alto, Alc-Alto y Alc-Medio (193 individuos).
- **Alc-No:** englobará Alc-Muy bajo y Alc-Bajo (456 individuos).

A pesar de esto las clases siguen estando desequilibradas como se aprecia, pero esperamos que al reducir los valores del atributo estos aumenten su soporte y extraigamos alguna regla que tenga como consecuente Alc-Si.

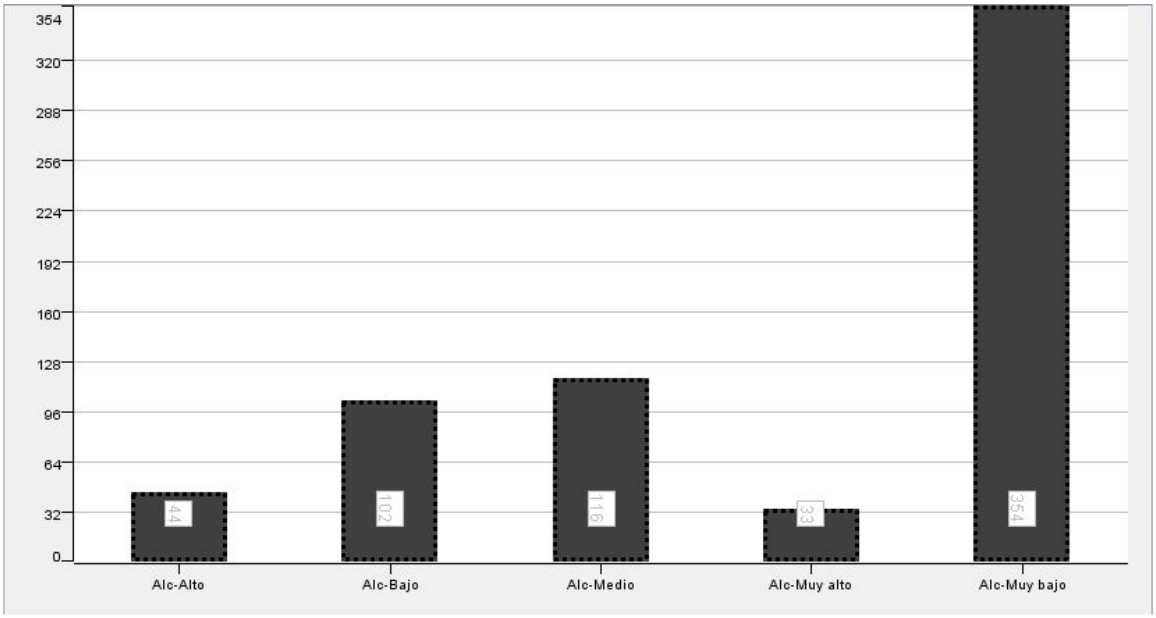


Figura 14: Histograma de distribución de valores del atributo Alc

Extracción de reglas con atributo *Alc* binarizado

Knime

Hemos intentado extraer reglas de Knime, con el nodo Association Rule Learner, configurandolo de la siguiente manera:

- Con soporte mínimo 0,05 y confianza mínima: 0,7 sin llegar a obtener regla ninguna.

Incluso esta ejecución ni siquiera llega a finalizar, ni filtrando con un Value Filter el atributo Alc en Sí o No se obtienen reglas de asociación en tiempos razonables. Podemos decir que se necesita un equipo informático de mayores prestaciones para poder trabajar en Knime y obtener reglas con valores bajos de soporte y confianza, puesto que en un equipo Intel Core i7 de 2.4 GHz con 8GB de RAM no se logró finalizar la ejecución, hasta con RAM dedicada de 6.5 GB.

Weka

Configurando el nodo “Apriori” de Weka en Knime de la siguiente manera, se obtienen las 100 mejores reglas con consecuente Alc:

- Con soporte mínimo 0,05 y confianza mínima de 0,7.

Cabe agregar que solamente filtrando por uno de los valores de Alc, utilizando el nodo Value Filter de Knime antes de la ejecución, se pueden obtener reglas para Alc-Sí y Alc-No, de lo contrario se obtienen reglas asociados a Alc-No solamente.

R

En R, utilizando el algoritmo “Apriori” del paquete arules, se pueden extraer reglas de asociación que solo contengan como consecuente al atributo Alc configurandolo así:

- Con soporte mínimo 0,05 y confianza mínima de 0,7 se obtienen 847.840 reglas

Al igual que en Weka, solamente filtrando por el valor Alc (Sí o No) se pueden obtener reglas para Alc-Sí o Alc-No, pero al contrario que Weka, el filtro se hace posterior al obtención de reglas.

Resultados

Esta vez sí se han obtenido algunas reglas con consecuente Alc-Sí (muy pocas en comparación), al igual que en el caso anterior se analizarán las reglas obtenidas con R.

En la siguiente gráfica podemos observar la misma representación que en la figura 11 con las reglas obtenidas con el atributo Alc binarizado. Lo primero que llama la atención son los doce puntos que destacan en un color mucho más rojo, esto se debe a que tienen un lift mucho mayor, estos puntos representan las reglas que tienen como consecuente Alc-Si. Como se puede ver estas reglas tienen un soporte bastante bajo y una confianza no excesivamente buena (menor de 0,85), se analizarán en detalle a continuación. Además de eso el resto de los puntos que parecen casi grises no tienen un lift tan alto pero tampoco está mal (son todos mayores que 1), también se observa que hay una densidad de puntos mucho mayor que en el caso de la figura 11 debido al aumento de reglas por el incremento del soporte.Si nos fijamos en eje X comprobamos que alcanzan reglas con un soporte muy alto, cercano al 0,7, en comparación a lo obtenido en la figura 11 en la que no se alcanza el 0,25.

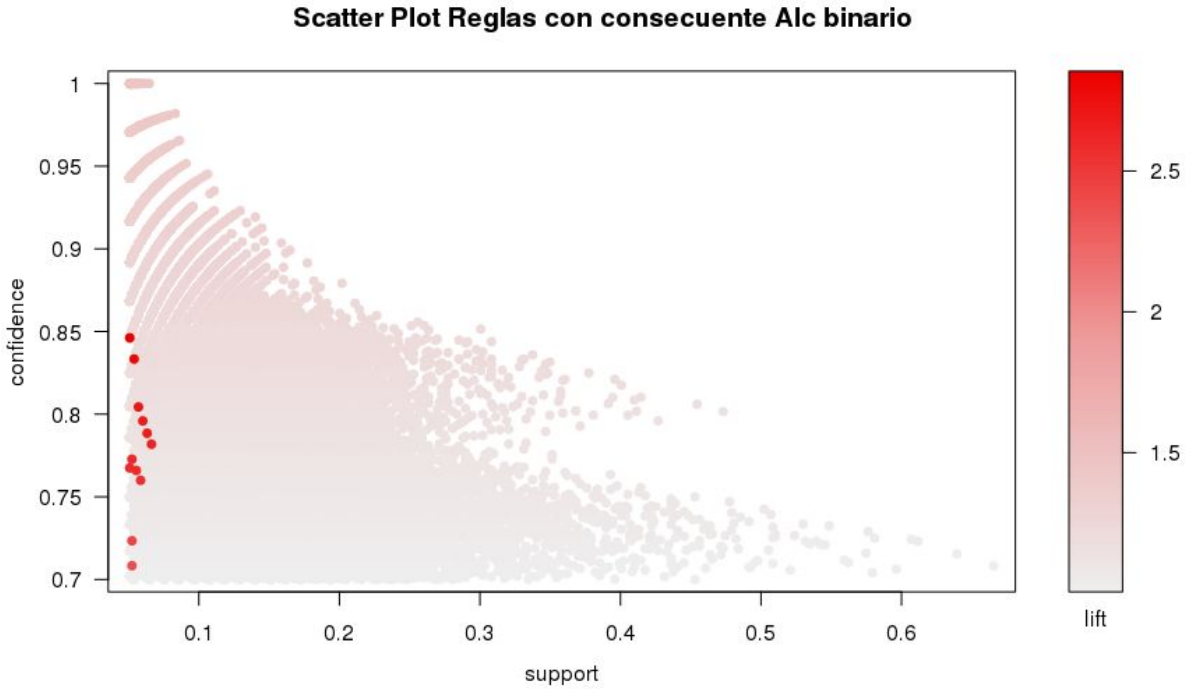


Figura 15: Reglas de asociación con Alc binarizado obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05

En la siguiente figura se representa las reglas respecto a su tamaño como en la figura 12.

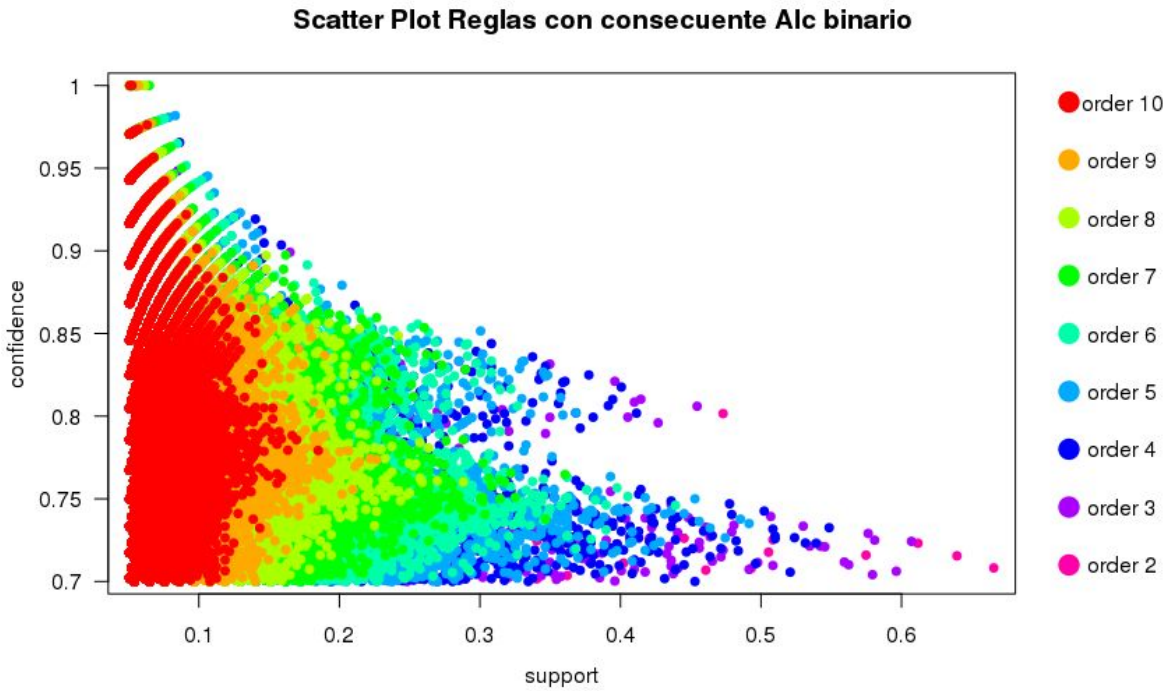


Figura 16: Reglas de asociación con Alc binarizado obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05 respecto al tamaño (order) de estas.

A continuación se representan las diez mejores reglas que tienen como consecuente Alc-No en un gráfico de grafo parecido al de la figura 13. Tras esto se detallan esas diez reglas.

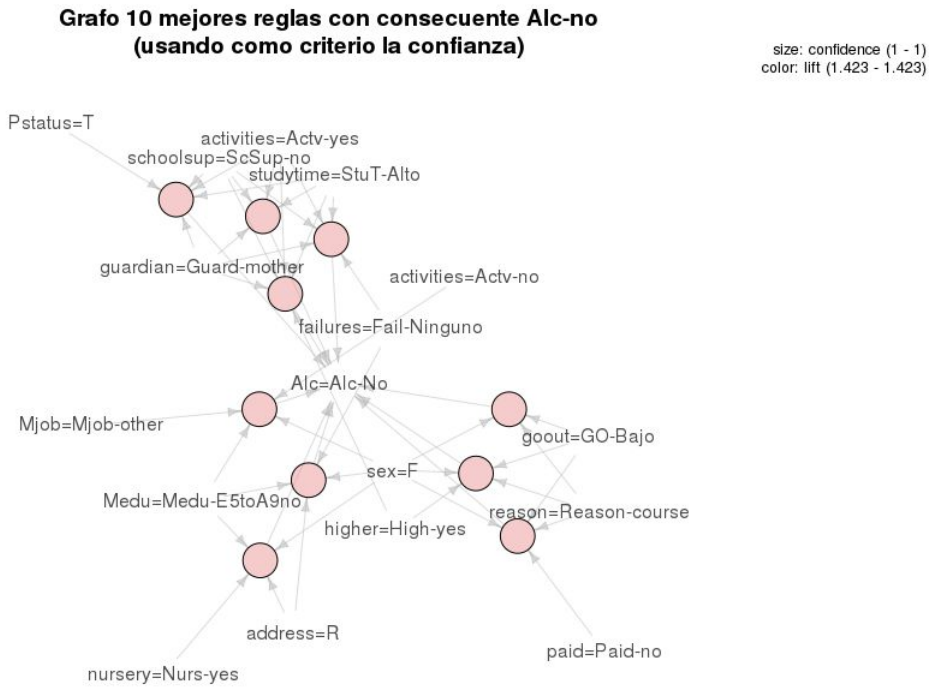


Figura 17: Grafo con las diez mejores reglas de asociación con consecuente Alc-No obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05.

	lhs		rhs	support	confidence	lift
[1]	sex=F + reason=Reason-course + goout=GO-Bajo	---->	Alc=Alc-No	0.05701079	1.0000000	1.4232456
[2]	guardian=Guard-mother + studytime=StuT-Alto + schoolsup=ScSup-no + activities=Actv-yes	---->	Alc=Alc-No	0.05546995	1.0000000	1.4232456
[3]	sex=F + reason=Reason-course + higher=High-yes + goout=GO-Bajo	---->	Alc=Alc-No	0.05238829	1.0000000	1.4232456
[4]	sex=F + reason=Reason-course + paid=Paid-no + goout=GO-Bajo	---->	Alc=Alc-No	0.05546995	1.0000000	1.4232456
[5]	sex=F + address=R + Medu=Medu-E5toA9no + nursery=Nurs-yes	---->	Alc=Alc-No	0.05546995	1.0000000	1.4232456
[6]	sex=F + address=R + Medu=Medu-E5toA9no + failures=Fail-Ninguno	---->	Alc=Alc-No	0.05546995	1.0000000	1.4232456
[7]	sex=F + Medu=Medu-E5toA9no + Mjob=Mjob-other + activities=Actv-no	---->	Alc=Alc-No	0.05238829	1.0000000	1.4232456
[8]	guardian=Guard-mother + studytime=StuT-Alto + failures=Fail-Ninguno + schoolsup=ScSup-no + activities=Actv-yes	---->	Alc=Alc-No	0.05084746	1.0000000	1.4232456
[9]	Pstatus=T + guardian=Guard-mother + studytime=StuT-Alto + schoolsup=ScSup-no + activities=Actv-yes	---->	Alc=Alc-No	0.05238829	1.0000000	1.4232456
[10]	guardian=Guard-mother + studytime=StuT-Alto + schoolsup=ScSup-no + activities=Actv-yes + higher=High-yes	---->	Alc=Alc-No	0.05392912	1.0000000	1.4232456

Tabla 4: Diez mejores reglas de asociación extraídas con R con confianza mínima 0,7 y soporte mínimo 0,05 que tengan como consecuente Alc-No

En esta figura se puede observar otro gráfico de grafo, pero en este caso con las diez mejores reglas que tienen como consecuente Alc-Si. Tras esto se detallan esas reglas, que como se puede comprobar tienen un lift bastante más alto que los de la tabla 4 como ya se había comentado.

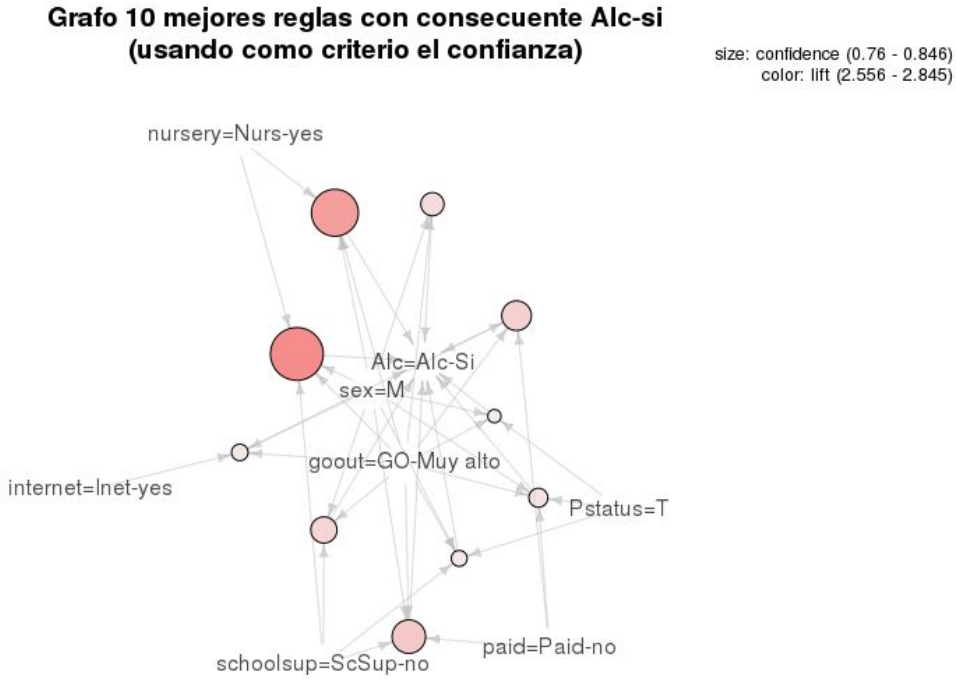


Figura 18: Grafo con las diez mejores reglas de asociación con consecuente Alc-Si obtenidas con R con confianza confianza mínima 0,7 y soporte mínimo 0,05.

	lhs		rhs	support	confidence	lift
[1]	sex=M + schoolsup=ScSup-no + nursery=Nurs-yes + goout=GO-Muy alto	---->	Alc=Alc-Si	0.05084746	0.8461538	2.845357
[2]	sex=M + nursery=Nurs-yes + goout=GO-Muy alto	---->	Alc=Alc-Si	0.05392912	0.8333333	2.802245
[3]	sex=M + schoolsup=ScSup-no + paid=Paid-no + goout=GO-Muy alto	---->	Alc=Alc-Si	0.05701079	0.8043478	2.704776
[4]	sex=M + paid=Paid-no + goout=GO-Muy alto	---->	Alc=Alc-Si	0.06009245	0.7959184	2.676430
[5]	sex=M + schoolsup=ScSup-no + goout=GO-Muy alto	---->	Alc=Alc-Si	0.06317411	0.7884615	

```
2.651355
[6]  sex=M +
     goout=GO-Muy alto      ----> Alc=Alc-Si 0.06625578  0.7818182
2.629016
[7]  sex=M +
     Pstatus=T +
     paid=Paid-no +
     goout=GO-Muy alto      ----> Alc=Alc-Si 0.05238829  0.7727273
2.598446
[8]  sex=M +
     internet=Inet-yes +
     goout=GO-Muy alto      ----> Alc=Alc-Si 0.05084746  0.7674419
2.580672
[9]  sex=M +
     Pstatus=T +
     schoolsup=ScSup-no +
     goout=GO-Muy alto      ----> Alc=Alc-Si 0.05546995  0.7659574
2.575681
[10] sex=M +
     Pstatus=T +
     goout=GO-Muy alto      ----> Alc=Alc-Si 0.05855162  0.7600000
2.555648
```

Tabla 5: Diez mejores reglas de asociación extraídas con R con confianza mínima 0,7 y soporte mínimo 0,05 que tengan como consecuente Alc-Si

Conclusiones finales

A nivel técnico sobre las herramientas empleadas, hemos llegado a la conclusión de que R es mucho más potente que Knime y consume menos recursos. Esto se puede contemplar cuando intentamos ejecutar trabajos algo intensivos como es la extracción de un número significativo de reglas de asociación, también se ve en la mayor flexibilidad que aporta R a la hora de tratar o representar los datos, por último se puede observar la diferencia comparando los tiempos de ejecución de estas actividades intensivas en R o Knime, aunque en R lleven cierto tiempo en Knime estos tiempos se incrementan hasta el punto que no es posible finalizar algunas operaciones en tiempos más o menos razonables.

Respecto a los datos hemos elaborado unos perfiles del consumidor y el no consumidor de alcohol usando las reglas extraídas de R y Weka. Para ello hemos seleccionado los valores de los atributos más discriminantes para determinar el consumo de alcohol de un individuo y hemos obviado algunos por no existir mucha variabilidad en sus valores a lo largo del dataset, lo que hace que aparezcan en ambos gráficos de grafos, como es internet=Inet-yes, paid = Paid-No, schoolsup=ScSup-no.

Perfil del no consumidor de alcohol:

- Mujeres.
- Interesadas por sus estudios, esta percepción se obtiene de varios atributos como que quieren ir a la universidad, sacan buenas notas, no van a particulares o no les ha quedado ninguna asignatura.

- Poca vida social.
- Familia con más de tres miembros.

Perfil del consumidor de alcohol:

- Hombres.
- Mucha vida social.
- Sus padres viven juntos.
- No se ha encontrado relación directa entre el consumo de alcohol y el desempeño académico al contrario de lo que podría parecer sin analizar en detalle los datos.

Este perfil del consumidor de alcohol tiene como factores principales la vida social elevada y el sexo masculino, lo cual cuadra con un estudio de los principales factores que determinan la predicción del consumo de alcohol con árboles de decisión elaborado por (Pagnotta & Hossain, 2016). En el crean la siguiente tabla en la que se representan los atributos y su porcentaje de impacto sobre el atributo Alc.

Attribute	**Percentage**
Male	25.35%
Social	21.13%
More Free time	9.39%
Less study time	8.45%
Mother less educational quality	7.98%
Good Health	7.04%
No Higher education	4.23%
No family support	3.76%
Small family	3.76%
High travel time	1.88%
Less activities	1.88%
No support school	1.88%
Father work	1.88%
Internet connectivity	1.41%

Tabla 6: Variables con mayor impacto en el atributo Alc y su porcentaje de impacto. (Pagnotta & Hossain, 2016)

Bibliografía

P. Cortez & A. Silva. (2008). 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, EUROSIS, ISBN 978-9077381-39-7.

Documentación de Knime (n.d). Recuperado el 13 de Enero de 2017, desde https://www.knime.org/nodeguide

Documentación de R (n.d)

F. Pagnotta & M. A. Hossain. (2016). Using Data Mining To Predict Secondary School Student Alcohol Consumption. *Department of Computer Science,University of Camerino, Advanced Database*.

Sistema de calificaciones portugués (n.d.). Recuperado el 13 de Enero de 2017, desde http://internacional.ugr.es/pages/conversion-calificaciones/tablaconversioncalificaciones/!

M. A. Vila. (2014). 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.*

Anexos

Migrar reglas de asociación a un gestor de BD transaccional

Como pudimos ver en esta memoria, trabajar con reglas de asociación puede generar miles de registros y volcarlos a un CSV (Comma Separated Values) o planilla electrónica parece razonable, de todas formas se hace difícil poder sacar conclusiones con tantos registros. *OpenOffice Database* permite realizar una conexión directa a una fuente de datos del tipo texto, CSV en este caso, lo que nos permite poder filtrar registros usando SQL, en este caso utilizamos la sentencia siguiente por ejemplo:

```
select * from "nombre_tabla_relacional" where consequent like '%Alc%' order
by Confidence DESC, Lift DESC, Support DESC
```

Lo primero que se debe hacer es instalar OpenOffice en su versión 4.1.3, ir al menú para conectar a una base de datos existente, seleccionar tipo texto, ingresar la ubicación del archivo, elegir el tipo (CSV, .txt) e ingresar el delimitador de campos (,). Con cada CSV existente en la ubicación seleccionada, se creará como una tabla relacional en OpenOffice Database.

En la Figura 19, se observan las reglas de asociación extraídas con Knime y exportadas como CSV, migradas al gestor, eventualmente también se podría migrar las extraídas con R.

row ID	Support	Confidence	Lift	Consequent	im...	Items	Split Value 1
rule10180	0,16	0,86	1,22	Alc-No	<---	[ScSup-no Guard-mother F GT3 Inet-yes Fail-Nin	ScSup-no
rule10231	0,16	0,86	1,22	Alc-No	<---	[ScSup-no <3-no T F Nurs-yes Fail-Ninguno Paid	ScSup-no
rule6349	0,16	0,86	1,22	Alc-No	<---	[Guard-mother F GT3 GP Fail-Ninguno Paid-no]	Guard-mother
rule3880	0,16	0,85	1,21	Alc-No	<---	[<3-no F GT3 Nurs-yes Fail-Ninguno]	<3-no
rule3956	0,16	0,85	1,21	Alc-No	<---	[F Reason-course High-yes Nurs-yes Paid-no]	F
rule3790	0,16	0,85	1,21	Alc-No	<---	[ScSup-no Guard-mother U F GT3]	ScSup-no
rule2629	0,16	0,85	1,21	Alc-No	<---	[High-yes Nurs-yes Inet-yes GO-Medio]	High-yes
rule5889	0,16	0,85	1,21	Alc-No	<---	[ScSup-no T Nurs-yes GO-Medio Fail-Ninguno]	ScSup-no
rule3915	0,16	0,85	1,21	Alc-No	<---	[ScSup-no F GT3 FamR-Muy Bueno Paid-no]	ScSup-no
rule10193	0,16	0,84	1,20	Alc-No	<---	[ScSup-no T F GT3 Nurs-yes Inet-yes Fail-Ningun	ScSup-no

Record 1 of 40 * (10)

```
SELECT * FROM "Association_rules_alc_0.7conf_0.155supp" WHERE "Consequent" LIKE '%Alc%' AND "Lift" > 1.0
```

Figura 19. Reglas de asociación extraídas desde knime en CSV migradas a OpenOffice Database.

Repositorio

Puede encontrar este trabajo completo en GitHub https://github.com/AythaE/Estudiantes-TID