Estudis d'Informàtica, Multimèdia i Telecomunicació

TFC Àrea Mineria de dades

Estudi de les possibles causes de l'abandonament d'un determinat pla d'estudis del departament d'Economia de la UOC

Alumne: Antoni Corral Herrerías

ETIS

Consultor: Ramon Caihuelas Quiles

7 de gener de 2013

Agraïments i dedicatòria

Volia agrair la paciència que ha tingut la meva família amb mi al llarg d'aquest sis anys que he dedicat a fer aquests estudis, sobretot en els períodes de cap de setmana i altres festes.

A la meva dona Ana, que ha aguantat els meus canvis d'humor sobretot quan s'apropava el lliurament de les PACs.

Al meu fill Sergio, que m'ha fet companyia en els moments en que necessitava desconnectar de la carrera.

A la meva filla Mònica, que ha estat rigorosa en la supervisió de la meva agenda com a estudiant. Gràcies per la teva comprensió i ajut que m'has brindat en tot moment, especialment per les classes particulars d'anglès.

I també voldria tenir un record pels meus pares i la meva germana als quals m'hagués agradat poder-los ensenyar la feina que he anat fent aquest darrers anys amb un final, espero feliç.

Índex

1. Introducció

1.1. Descripció del TFC

El departament d'economia de la UOC té interès en conèixer el motiu de perquè en els últims 10 anys (20 semestres) hi ha un percentatge força alt, al voltant del 25%, d'alumnes que inicien els seus estudis i no els finalitzen. Per aquest motiu, ha fet una recopilació de dades dels alumnes matriculats, més de divuit mil, al llarg d'aquest període. Les dades recullen els alumnes que han abandonat, el sexe, l'edat quan van fer la 1a matrícula, el semestre que van iniciar els estudis, semestres que porten, assignatures i nº crèdits matriculats en el 1r semestre, el nº d'assignatures / crèdits superats i més dades relacionades amb les 12 assignatures més comuns de matrícula en el 1r semestre d'aquests alumnes.

Aquest treball de fi de carrera utilitzarà tècniques de mineria de dades per estudiar aquesta informació, sense partir de cap premissa prèvia. Utilitzant programari lliure, es treballaran diferents algorismes per intentar trobar un model de negoci que permeti justificar aquest fet.

1.2. Objectius generals i específics

L'objectiu general d'aquest TFC és desenvolupar un cicle de vida comú a qualsevol treball de mineria de dades. A partir dels objectius que ens demanen, es fa una selecció de les dades proporcionades i, si cal, es netegen i/o transformen en dades que donin més valor per a poder extreure coneixement. Donat que no es preveu d'inici quin és el model que ens pot donar més informació, es tracta d'anar experimentant els diferents models que es puguin crear amb aquestes dades i anar avaluant-los fins arribar al que considerem millor.

L'objectiu específic que es pretén és poder trobar una casuística comuna amb els alumnes que abandonen els estudis per tal que el departament d'economia pugui estudiar els resultats d'aquest treball i permeti fer una orientació més acurada als nous alumnes que compleixin amb aquesta regla. D'aquesta manera es podrà optimitzar els recursos posats a disposició de cada estudiant que es matricula per primera vegada.

1.3. Motivació

La motivació personal a l'hora d'escollir l'àrea de Mineria de dades per fer el treball de final de carrera ve donada al descobriment que vaig tenir en el semestre anterior de l'assignatura de Mineria de dades. Jo sempre havia sentit parlar de les bases de dades i les he treballat al llarg de la meva carrera de forma molt mecànica i gairebé prescindint del tipus d'informació que tenien. Només em preocupava si un camp determinat era numèric enter o amb decimals, text d'una sola paraula o cadena de paraules, etc...

Aquesta àrea m'ha fet veure la importància no només del format, sinó del contingut i del valor qualitatiu que té tant dintre dels valors del mateix camp com respecte als valors dels altres camps de la mateixa base de dades, fins i tot amb camps que aparentment no tenen cap relació entre ells.

Per motius de feina, sóc professor de secundària de formació professional de grau mitjà de la branca d'electricitat, estic envoltat de dades relacionades amb el món acadèmic: notes, faltes, dni d'alumnes, telèfons de pares, adreces, expedients, etc ... A més, mentre estava fent aquesta assignatura, a la feina estava implicat en un projecte d'innovació que pretenia millorar la informació que tenen els nostres alumnes a l'hora de fer la preinscripció en el meu institut. Fins i tot vaig fer algun exercici amb una petita mostra de dades que vaig poder extreure d'una enquesta que vam fer als assistents a la jornada de portes obertes que es va fer per les dates de la preinscripció.

Per aquest motiu la temàtica proposada pel tutor Ramon Caihuelas, d'aplicar tècniques de mineria de dades a un fitxer aportat pel departament d'Informàtica de la UOC en el que hi ha enregistrades les dades de matrícula i resultats (del 1r semestre de la seva matrícula) d'estudiants de 20 semestres d'un pla d'estudis d'economia de la pròpia UOC em va semblar interessant i extrapolable a la meva realitat laboral.

1.4. Estat de l'art de la mineria de dades

L'aplicació de tècniques de mineria de dades ha crescut en els darrers anys de forma molt ràpida, es podria dir que paral·lelament al creixement del negoci a Internet. Una de les aplicacions més característiques de la mineria de dades és poder extreure coneixement de la informació, conegut com **Knowledge Discovery in Databases – KDD.** Fa uns anys la informació per fer aquest estudi estava repartida en diferents documents dels quals s'havien d'extreure les dades a estudiar dintre dels diferents programes que portaven la comptabilitat de l'empresa i que calia exportar o mitjançant estudis, observacions o enquestes plantejades per tenir informació sobre el fet en concret. Avui en dia, a més, aquesta informació es pot anar enregistrant de forma instantània amb la utilització d'Internet: tant al realitzar compres, emetre opinions de productes, perfils en xarxes socials, etc ...

Aquest coneixement de la informació pot anar adreçat a crear i/o modificar estratègies de mercat de qualsevol tipus de producte o servei. Abans, potser aquestes estratègies anaven adreçades bàsicament a augmentar la venda de productes que estaven relacionats amb altres que coincidien en la cistella de la compra. En aquest moment, aquestes estratègies poden servir no tan per augmentar, sinó per mantenir aquestes vendes i, en el cas de serveis, millorar el grau de satisfacció dels clients perquè no es vagin a la competència i si cal, oferir-li a cada client serveis a mida.

Aquestes dades que hi ha per Internet també es van actualitzant de forma molt ràpida i la seva informació pot anar canviant de tendència més sovint que abans. Això fa que les tècniques de mineria de dades hagin de donar resposta també cada vegada més ràpid.

1.5. Mineria de dades en l'àmbit acadèmic

Les tècniques de mineria de dades cada vegada més es van estenent en tots els àmbits. En l'àmbit acadèmic es poden trobar diferents exemples d'utilització de tècniques de mineria de dades per a poder preveure o justificar determinats comportaments. Es poden veure alguns casos.

La mineria de dades en el descobriment de perfils de deserció estudiantil en la Universitat de Nariño.

Aquest estudi es va fer a l'any 2009 amb dades del 2006 i pretén determinar si hi ha uns determinats perfils que coincideixen amb estudiants que finalment abandonen els estudis. Al conèixer aquests perfils, la universitat pot crear estratègies per reduir aquest abandonament. Aquest treball s'ha fet amb TariyKDD, un programa de mineria de dades també de codi lliure. (http://sourceforge.net/projects/tariykdd.berlios/)

http://revistas.udenar.edu.co/index.php/USalud/article/view/218

Estratègies intel·ligents aplicables a un sistema educatiu.

Aquesta investigació es centra en l'estudi i desenvolupament d'estratègies que pertanyen a l'àrea de l' intel·ligència artificial que siguin aplicables a sistemes educacional.

http://sedici.unlp.edu.ar/bitstream/handle/10915/20667/Documento_completo.pdf?sequence=1

Models predictius i tècniques de mineria de dades per l' identificació de factors associats al rendiment acadèmic d'alumnes universitaris.

És un projecte que té com a objectiu construir models predictius del rendiment acadèmic dels estudiants de la FACENA de la UNNE. Les dades que s'han tingut en compte són: resultats de test de diagnòstic de coneixements matemàtics previs i les condicions socioeconòmiques dels alumnes de primer any.

http://sedici.unlp.edu.ar/handle/10915/19846

Aplicacions de mineria de dades en l'educació superior

IBM en un fulletó de propaganda del seu software **IBM Business Analytics** parla de varis casos d'aplicacions de mineria de dades en l'educació superior. Cas 1: Creació de tipologies significatives de resultats d'aprenentatge. Cas 2: Planificació e intervencions acadèmiques: predicció de trasllat dels estudiants.

ftp://ftp.software.ibm.com/common/ssi/ecm/es/imw14303eses/IMW14303ESES.PDF

Califòrnia Watch: datamining, corrupció, escoles i terratrèmols.

Aquest projecte encara que no és estrictament educatiu, tracta del coneixement d'un fenomen natural, els terratrèmols, que afecten a tots els edificis i en el cas d'estudi a les escoles. Les dades han estat extretes de documents relacionat amb aquest fenòmens. Al llarg del 2011 es va fer un projecte a Califòrnia traient dades d'un total de 30000 documents que parlaven de terratrèmols i el seu impacte en les zones en les que hi havia centres educatius. Es va repassar tota la informació, contrastant-la amb la d'altres documents i es va construir una base de dades interactiva contenint les escoles de primària de l'estat. Els pares podien fer un seguiment sobre la seguretat de les escoles dels seus fills i això va permetre fer una proposta de seguiment per part de l'estat per adequar les instal·lacions per fer-les més segures.

http://blogs.lanacion.com.ar/data/entrevistas/california-watch-datamining-corrupcion-escuelas-y-terremotos/

Mapa de perills sísmics a prop de Califòrnia http://seismic.apps.cironline.org/

1.6. Programari per fer mineria de dades

Hi ha molts programes per poder treballar algorismes de mineria de dades. Alguns d'ells són privatius i d'altres són de codi lliure. D'aquests últims hi ha diferents estudis de quins són els millors o més utilitzats. De la font http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/ indiquen que els cinc millors són:

Orange. http://orange.biolab.si/

RapidMiner. http://rapid-i.com/content/view/181/190/

Weka. http://www.cs.waikato.ac.nz/~ml/weka/

JHepWork. http://jwork.org/jhepwork/

Knime. http://www.knime.org/

Es poden trobar molts més programes que permeten treballar amb mineria de dades.

http://www.kdnuggets.com/software/suites.html

1.7. Planificació amb fites i temporització

La planificació del TFC va lligada a les dates previstes en el calendari del semestre en que estic fent el TFC i per tant, les agafaré com a referència per a desenvolupar totes les tasques inherents en un projecte de mineria de dades.

Aquestes són:

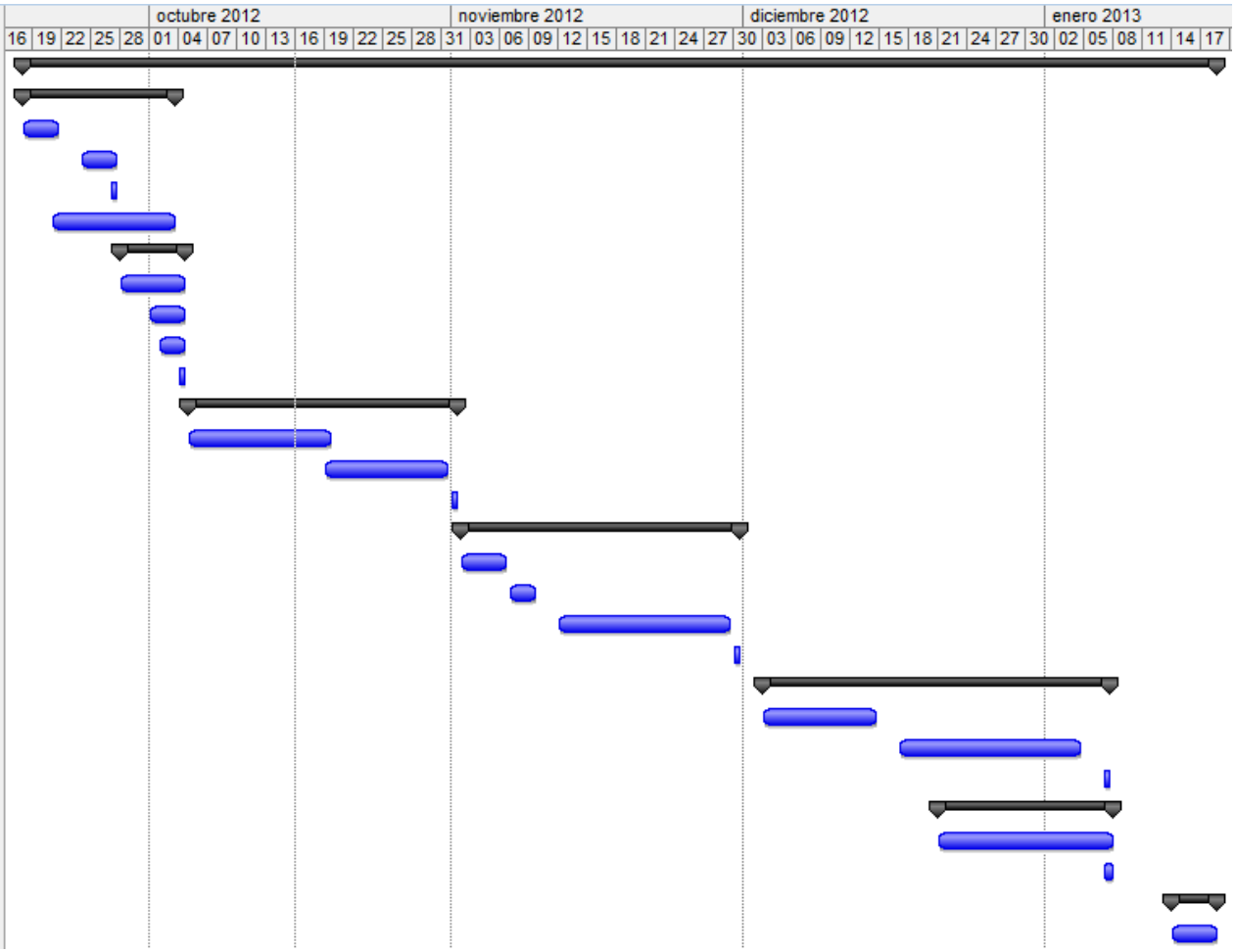
27 setembre	Trobada virtual.
4 octubre	PAC1 (Pla de treball del TFC)
1 novembre	PAC2 (Lliurament de la feina prevista a la fase 2)
30 novembre	PAC3 (Lliurament de la feina prevista a la fase 3)
7 gener	Lliurament final de la memòria i de la presentació virtual.
14-18 gener	Debat virtual

Cal a dir que el pla de treball és només una mostra d'intencions i per tant està incomplert ja que s'anirà modificant a mida que vagi avançant en el coneixement de les dades a treballar.

1.8. Llista de tasques i diagrama de Gantt

	Nombre de tarea	Duración	Comienzo	Fin
1	**Treball final de carrera**	**89 días**	**mar 18/09/12**	**vie 18/01/13**
2	**Inici semestre**	**12 días**	**mar 18/09/12**	**mié 03/10/12**
3	Lectura pla docent	4 días	mar 18/09/12	vie 21/09/12
4	Concreció tema TFC	4 días	lun 24/09/12	jue 27/09/12
5	Trobada virtual	1 día	jue 27/09/12	jue 27/09/12
6	Lectura documents aula TFC	9 días	vie 21/09/12	mié 03/10/12
7	**Fase 1. Pla de treball**	**5 días**	**vie 28/09/12**	**jue 04/10/12**
8	Descripció del TFC	5 días	vie 28/09/12	jue 04/10/12
9	Objectius generals i específics	4 días	lun 01/10/12	jue 04/10/12
10	Planificació TFC	3 días	mar 02/10/12	jue 04/10/12
11	Presentació Pla de treball. PAC1	1 día	jue 04/10/12	jue 04/10/12
12	**Fase 2.**	**20 días**	**vie 05/10/12**	**jue 01/11/12**
13	Anàlisi de les dades	11 días	vie 05/10/12	vie 19/10/12
14	Estratègia segons les dades trobades	9 días	vie 19/10/12	mié 31/10/12
15	Presentació PAC2	1 día	jue 01/11/12	jue 01/11/12
16	**Fase 3.**	**21 días**	**vie 02/11/12**	**vie 30/11/12**
17	Preparació de les dades	3 días	vie 02/11/12	mar 06/11/12
18	Depuració de les dades	3 días	mié 07/11/12	vie 09/11/12
19	Aplicar diferents algorismes MD	14 días	lun 12/11/12	jue 29/11/12
20	Presentació PAC3	1 día	vie 30/11/12	vie 30/11/12
21	**Fase 4.**	**26 días**	**lun 03/12/12**	**lun 07/01/13**
22	Anàlisi dels resultats obtinguts	10 días	lun 03/12/12	vie 14/12/12
23	Conclusions	15 días	lun 17/12/12	vie 04/01/13
24	Finalització memòria TFC	1 día	lun 07/01/13	lun 07/01/13
25	**Fase 5.**	**12 días**	**vie 21/12/12**	**lun 07/01/13**
26	Elaboració presentació virtual	12 días	vie 21/12/12	lun 07/01/13
27	Lliurament presentació virtual	1 día	lun 07/01/13	lun 07/01/13
28	**Fase 5.**	**5 días**	**lun 14/01/13**	**vie 18/01/13**
29	Debat virtual	5 días	lun 14/01/13	vie 18/01/13

	Nombre de tarea	Duración	Comienzo	Fin
1	**Treball final de carrera**	**89 días**	**mar 18/09/12**	**vie 18/01/13**
2	**Inici semestre**	**12 días**	**mar 18/09/12**	**mié 03/10/12**
3	Lectura pla docent	4 días	mar 18/09/12	vie 21/09/12
4	Concreció tema TFC	4 días	lun 24/09/12	jue 27/09/12
5	Trobada virtual	1 día	jue 27/09/12	jue 27/09/12
6	Lectura documents aula TFC	9 días	vie 21/09/12	mié 03/10/12
7	**Fase 1. Pla de treball**	**5 días**	**vie 28/09/12**	**jue 04/10/12**
8	Descripció del TFC	5 días	vie 28/09/12	jue 04/10/12
9	Objectius generals i específics	4 días	lun 01/10/12	jue 04/10/12
10	Planificació TFC	3 días	mar 02/10/12	jue 04/10/12
11	Presentació Pla de treball. PAC1	1 día	jue 04/10/12	jue 04/10/12
12	**Fase 2.**	**20 días**	**vie 05/10/12**	**jue 01/11/12**
13	Anàlisi de les dades	11 días	vie 05/10/12	vie 19/10/12
14	Estratègia segons les dades trobade	9 días	vie 19/10/12	mié 31/10/12
15	Presentació PAC2	1 día	jue 01/11/12	jue 01/11/12
16	**Fase 3.**	**21 días**	**vie 02/11/12**	**vie 30/11/12**
17	Preparació de les dades	3 días	vie 02/11/12	mar 06/11/12
18	Depuració de les dades	3 días	mié 07/11/12	vie 09/11/12
19	Aplicar diferents algorismes MD	14 días	lun 12/11/12	jue 29/11/12
20	Presentació PAC3	1 día	vie 30/11/12	vie 30/11/12
21	**Fase 4.**	**26 días**	**lun 03/12/12**	**lun 07/01/13**
22	Anàlisi dels resultats obtinguts	10 días	lun 03/12/12	vie 14/12/12
23	Conclusions	15 días	lun 17/12/12	vie 04/01/13
24	Finalització memòria TFC	1 día	lun 07/01/13	lun 07/01/13
25	**Fase 5.**	**12 días**	**vie 21/12/12**	**lun 07/01/13**
26	Elaboració presentació virtual	12 días	vie 21/12/12	lun 07/01/13
27	Lliurament presentació virtual	1 día	lun 07/01/13	lun 07/01/13
28	**Fase 5.**	**5 días**	**lun 14/01/13**	**vie 18/01/13**
29	Debat virtual	5 días	lun 14/01/13	vie 18/01/13



2. Mineria de dades

2.1. Cicle de vida d'un projecte de Mineria de dades

Les fases que composen el cicle de vida d'un projecte de mineria de dades és el següent:

1. Definició de la tasca de mineria de dades
2. Origen de les dades
3. Preparació de les dades
4. Mineria de dades
5. Avaluació i interpretació del model
6. Integració dels resultats en el procés

Cal a dir que una vegada fet tot el procés al experimentar-ho poden sorgir noves necessitats i potser caldrà redefinir la tasca inicial, afegir/eliminar més dades i tornar a plantejar altre vegada el procés i així de forma iterativa fins trobar la solució òptima.

Es podem representar amb el següent diagrama:

2.1.1 Definició de la tasca de mineria de dades

En aquest primer punt és on cal precisar quin serà l'objectiu del projecte de mineria de dades. S'ha de definir la tasca principal que es vol treballar, per exemple:

Pot interessar **trobar similituds i agrupar elements semblants**. Per aquesta tasca el model típic que s'utilitzarà és el model d'agregació (***clustering***) procedent de l'anàlisi de dades o de l'aprenentatge automàtic i els models associatius. Normalment es fa servir quan es té poca informació i se'n vol començar a tenir.

També pot interessar **classificar objectes** quan ja es té informació d'aquests i l'existència de grups ja definits. En aquest cas es vol treballar en conèixer les diferències entre aquests grups. Existeixen uns models classificatoris típics com: els arbres de decisió, les xarxes neuronals i les regles de classificació.

A vegades, el que interessa és **predir** el que passarà. En certa manera és una forma de classificar. Aquesta classificació pot ser des d'una de tipus binària, o pertany o no a la classe; o de tipus finit o infinit tenint diferents classes per aquesta classificació.

També pot interessar **descriure** trobant i/o expressant associacions que poden ser significatives entre dos variables i/o **explicar,** quan es pugui, les raons de per què s'ha donat un comportament determinat.

2.1.2 Origen de les dades

Una vegada s'ha definit la tasca a realitzar, s'ha de localitzar les dades. Es pot comptar amb una empresa que tingui un magatzem de dades (**Data Warehouse**) el qual integra totes les dades procedents de les diferents dades que hi ha a l'empresa.
Però això no és sempre així i possiblement el que caldrà fer és anar cercant totes les dades disperses en diferents bases de dades que puguin haver-hi en l'empresa. També s'ha de tenir en compte les bases de dades transaccionals que recullen les operacions diàries i el seu historial.

2.1.3 Preparació de les dades

Una vegada es tenen localitzades les dades, aquestes s'han de preparar per tal que se les puguin aplicar el model triat. Per això cal que les dades tinguin la qualitat suficient: que no tinguin errors, redundàncies, ...; que siguin les necessàries: unes no caldran i altres potser s'hauran d'afegir; també caldrà que estiguin en la forma adequada per poder-se adaptar al format en que les necessita el model escollit.

Dintre d'aquesta preparació caldrà netejar les dades processant-les per tal de, per exemple, completar possibles dades incompletes, eliminar dades redundants o dades incorrectes o inconsistents. També caldria arreglar o unificar criteris per a errors de transcripció, actualitzar dades envellides, variacions en les referències a un mateix concepte per estar en diferents base de dades introduïdes per persones diferents. A més és possible que aquestes dades siguin esbiaixades, és a dir, que pertanyin a un conjunt d'objectes molt determinat.

Una vegada ja estiguin "preparades" igual, encara no estan en la forma més adequada i caldrà algun tipus de transformació: passar dades numèriques a categòriques, per exemple notes numèriques (0, 1, .., 10) a categories (Insuficient, ..., Excel·lent) o a l'inrevés de dades categòriques a numèriques; també es poden fer altres transformacions, simplificant valors, agrupant valors continus, normalitzant dades, afegint etiquetes o expandint atributs.

No sempre tenir el màxim de dades es bo pel model. Pot interessar reduir la dimensionalitat, és a dir, treballar amb menys dades i intentar obtenir els mateixos resultats. Això es pot fer reduint el nombre de registres o reduint el nombre d'atributs que cal tractar.
Així es tenen les dades que interessen i com fa falta. La part de la preparació de les dades moltes vegades té un cost de temps molt gran

Weka té una pestanya, **Preprocess**, des de la que es pot accedir a diferents filtres que permeten preparar les dades carregades al programa. Al llarg del TFC es fan servir alguns.

2.1.4 Mineria de dades

Ara ja es pot escollir un mètode de construcció de models per tal de trobar el model (coneixement) que ha de respondre millor a les característiques implícites dins les dades. Per això es fa un procés de cerca explorant un espai de models possibles.

Cal establir una mecànica general començant per un model, pot ser el model buit, i anar modificant-lo fins trobar un que tingui prou qualitat i es pugui considerar com a definitiu.
Els mètodes de mineria de dades es poden dividir pel que fa al coneixement a priori, pel que fa al tipus de dades i pel que fa al procés de construcció.

2.1.5 Avaluació i interpretació del model

Una vegada es té un model que es creu que és el millor es pot preguntar si aquest és millorable i es podria plantejar tornar a començar de nou a veure si es troba un de millor. Per avaluar-lo es poden tenir dos conjunts de dades que són del mateix conjunt inicial, amb un es fa el model i amb l'altre s''avalua. A vegades, es fa un tercer conjunt anomenat de validació.

Un cop ja es té un model de prou qualitat i que ha estat validat mitjançat el procés d'avaluació, s'ha d'interpretar i extreure'n el significat del coneixement que es mostra.

2.1.6 Integració dels resultats en el procés

Finalment cal integrar els resultats de la mineria de dades en el procés típic del sistema d'informació en que s'aplica.

2.2 Models de Mineria de dades

2.2.1 De classificació: arbres de decisió, xarxes neuronals, regles.

Els arbres de decisió donen una estructura tal que a cada node se li fa una pregunta sobre un atribut determinat: el valor que prengui indica que cal seguir la branca corresponent a l'atribut. Els nodes finals corresponen a conjunt d'exemples que pertanyen a la mateixa classe..

Alguns arbres permeten mètodes de poda que eliminen la generació de subarbres que compliquen la seva comprensió.

Les xarxes neuronals també són bons models classificatoris i predictius. Tenen certes analogies amb la manera en què estan connectades les neurones cerebrals i s'organitzen en forma de molts nodes de procés connectats que donen una o més sortides. Les diverses capes de nodes estan connectades entre si amb més o menys força a través d'uns factors o pesos que indiquen la importància de les sortides produïdes per cada node. En conjunt, el que fan és aprendre a ajustar els valors d'aquests pesos per a ser tan predictives com sigui possible.

Les regles de classificació imposen una sèrie de condicions sobre els valors que prenen els atributs d'entrada per tal d'indicar a quina classe poden pertànyer.

Weka té una pestanya, **Classify**, des de la que es pot aplicar diferents algorismes de classificació: arbres de decisió (trees: J48, Id3, ...), regles de classificació (**rules**: JRip, OneR, Part, ...)

2.2.2 D'agregació (clustering)

És la classificació d'objectes similars en diferents grups o el que és el mateix fer la partició de les dades en diferents subconjunts (clústers). Els criteris per fer l'assignació a un clúster o altre s'estableix a partir de mesures de distància en l'espai d'observacions o que volen reflectir la proximitat de les distribucions de probabilitat conjunta dels atributs que hi ha en les observacions realitzades.

Weka té una pestanya, **Cluster**, que permet aplicar algorismes d'agregació (clustering): (**clusterer**: Cobweb, EM, SimpleKMeans, ...)

2.2.3 Regles d'associació

Les regles d'associació cerquen trobar concurrències prou significatives entre grups de variables. L'únic requeriment que imposen és que s'indiqui el "nivell de suport" que es vol que tinguin a partir de les dades, la proporció de les dades que es vol cobrir amb aquesta regla. Llavors cal trobar grups de variables i combinacions de valors que arribin a tenir aquest grau de suport.

Weka té una pestanya, **Associate**, que permet aplicar algorismes d'associació. (**Associator**: Apriori, PredictiveApriori)

2.2.4 Xarxes bayesianes

Són models gràfics que representen la relació probabilística entre certes variables d'interès.

2.2.5 Models segons objectius proposats

Quan l'objectiu és trobar **similituds** i **agrupar** objectes semblants es poden utilitzar models d'agregació (clustering) i models associatius.

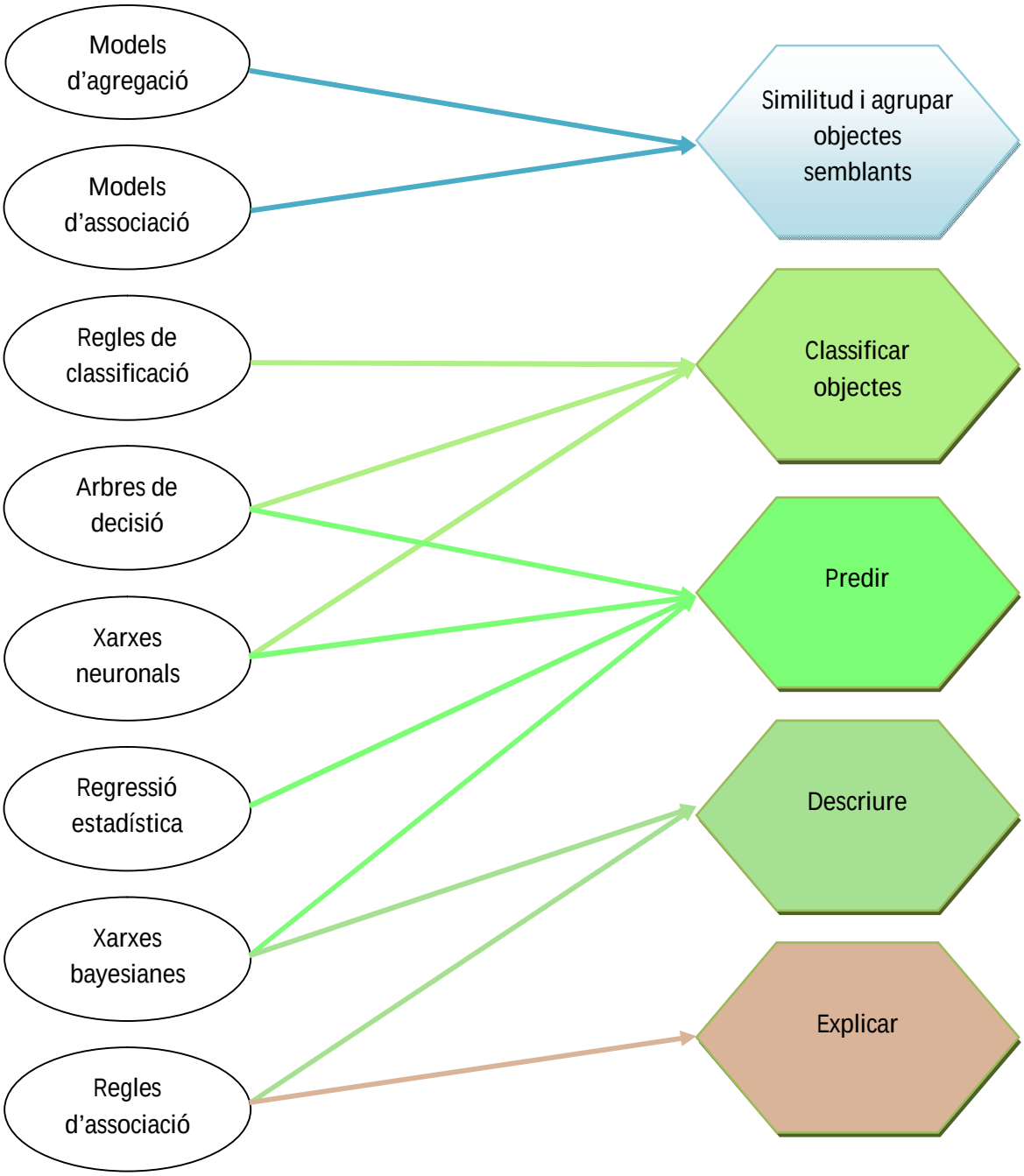
Quan el que es vol és **classificar objectes** es poden utilitzar arbres de decisió, xarxes neuronals i regles de classificació.

Quan es vol **predir** també es poden utilitzar a més dels arbres de decisió i les xarxes neuronals, les xarxes bayesianes i la regressió estadística.

Per **descriure** s'utilitzen les xarxes bayesianes i regles d'associació.

I si el que es vol és **explicar**, es tenen les xarxes bayesianes.

Similitud i agrupar objectes semblants	Classificar objectes	Predir	Descriure	Explicar
•Models d'agregació •Models d'associació	•Arbres de decisió •Xarxes neuronals •Regles de classificació	•Arbres de decisió •Xarxes neuronals •Xarxes bayesianes •Regressió estadística	•Xarxes bayesianes •Regles d'associació	•Xarxes bayesianes



3. Estudi, preparació i transformació de les dades aportades per la UOC

3.1. Pla d'estudis Diplomatura Ciències Empresarials

http://www.uoc.edu/estudis/titulacions/ciencies_empresarials/pla_estudis/index.html

Observant el nom de les assignatures que hi ha en el fitxer adjunt de les dades estudiades, es veu que el pla d'estudis treballat és el de la diplomatura de Ciències Empresarials que actualment ja no es fa a cap universitat. Just les 12 assignatures més matriculades en el primer semestre corresponen a les recomanades en el pla d'estudis per que es facin en el 1r any. Únicament l'assignatura **Matemàtiques II** no surt en aquest llistat i en canvi sí que surt **Anglès III** que en les orientacions es proposa pel 3r semestre.

	Assignatures	Crèdits	Totals
Semestre 1	• Introducció al dret • Introducció a la macroeconomia • Matemàtiques I • Estadística I • Multimèdia i comunicació per a economia i empresa • Anglès I	6 4,5 6 6 4,5 4,5	31,5
Semestre 2	• Introducció a la comptabilitat • Organització i administració d'empreses I • Matemàtiques II • Introducció a la microeconomia • Direcció de la producció I • Anglès II	6 6 6 4,5 4,5 4,5	31,5

3.2. Anàlisi de les dades

Des de la UOC s'han recollit dades en un fitxer **TFC_MD.dat** corresponents a diferents camps relacionats amb la matrícula d'estudiants d'economia. Aquests camps volen recollir informació referent a la matrícula del 1r semestre, els seus resultats i el fet que l'alumne hagi abandonat els estudis, a més d'altres dades relacionades amb l'estudiant.

Juntament amb aquestes dades hi ha un fitxer de text que descriu cadascun d'aquest camps.

Les variables son les següents:

ID: identificador unic per cada estudiant

ABANDONA: abandona després del primer semestre (0 fals, 1 cert)

TITULAT: es titula després del primer semestre (0 fals, 1 cert)

SEXE: 0 dona, 1 home

EDAT: edat en anys en el moment d'entrar

FRANJA: franja d'edat <=24 anys (1), <=27 (2), <=30 (3), <=36 (4), > 36 (5)

SEMESTRE: semestre en el que inicia els estudis

NSEM: numero de semestre relatiu des de que es van iniciar els estudis

NA: numero d'assignatures matriculades el primer semestre

NC: numero de crèdits matriculats

NASUP: numero d'assignatures superades

NCSUP: numero de crèdits superats

PCTAS: percentatge d'assignatures superades (NASUP/NA)

PCTCS: percentatge de crèdits superats (NCSUP/NC)

VIA: via d'accés 1 no cou, 2 cou, 3 estudis inacabats, 4 titulat

NACMAT: numero d'assignatures matriculades del conjunt de les 12 mes comuns del 1r semestre

NACPRE: numero d'assignatures a les quals es presenta del conjunt de les 12 mes comuns del 1r semestre

NACSUP: numero d'assignatures superades del conjunt de les 12 mes comuns del 1r semestre

A1M: 0 si no es matricula de l'assignatura 1, 1 si es matricula

A1S: -1 si no supera l'assignatura 1, 0 si no la matricula o no es presenta, 1 si la supera

A2M A2S: ídem per l'assignatura 2

A3M A3S A4M A4S A5M A5S A6M A6S A7M A7S A8M A8S A9M A9S A10M A10S A11M A11S A12M A12S: ídem per la resta

Llista d'assignatures:

00.010: multimèdia i comunicació

00.002: angles I

01.001: introducció al dret

01.079: introducció a la macroeconomia

01.005: introducció a la comptabilitat

01.003: matemàtiques I

01.006: organització i administració d'empreses I

01.004: estadística I

01.078: introducció a la microeconomia

01.009: direcció de la producció I

00.004: angles III

00.003: angles II

Una vegada rebut aquest fitxer s'ha procedit a veure el contingut d'aquestes dades per a poder fer més endavant una estratègia fent mineria de dades i intentar trobar un model que justifiqui la raó per la que abandonen aquests estudis un 25% dels alumnes matriculats.

Per fer una primera observació, s'ha procedit a estudiar els valors que tenen cadascú d'aquests camps. Per fer-ho, s'ha intentat obrir directament el fitxer TFC_MD.dat amb el programari WEKA (Waikato Enviroment for Knowledge Analysis), que és el que s'utilitzarà posteriorment per analitzar les diferents regles de mineria de dades. Aquest programa té una tota una biblioteca de classes d'aprenentatge en Java i ha estat desenvolupat en la universitat de Waikato, en Nova Zelanda. Es pot descarregar de la web http://www.cs.waikato.ac.nz/ml/weka/.Per a fer aquest TFC s'ha utilitzat la versió 3.6.5.

Al intentar-ho ha donat el següent error: **TFC_DAT not recognised as an 'svm light data files' file**. **Reason: Unable to determine structure a svm light: java.lang.Exception: Error parsing line "ID ABANDONA TITULAT SEXE EDAT FRANJA SEMESTRE NSEM NA NC NASUP NCSUP PCTAS PCTCS VIA NACMAT NACPRE NACSUP A1M A1S A2M A2S A3M A3S A4M A4S A5M A5S A6M A6S A7M A7S A8M A8S A9M A9S A10M A10S A11M A11S A12M A12S' : java.lang.Exception:java.lang.StringIndexOutOfBoundsException:String index out of range:-1**

Al obrir-lo amb l'excel per després convertir-ho a csv, sortia un missatge d'error: **"SYLK: formato de archivo no es válido" mensaje de error al abrir archivo**

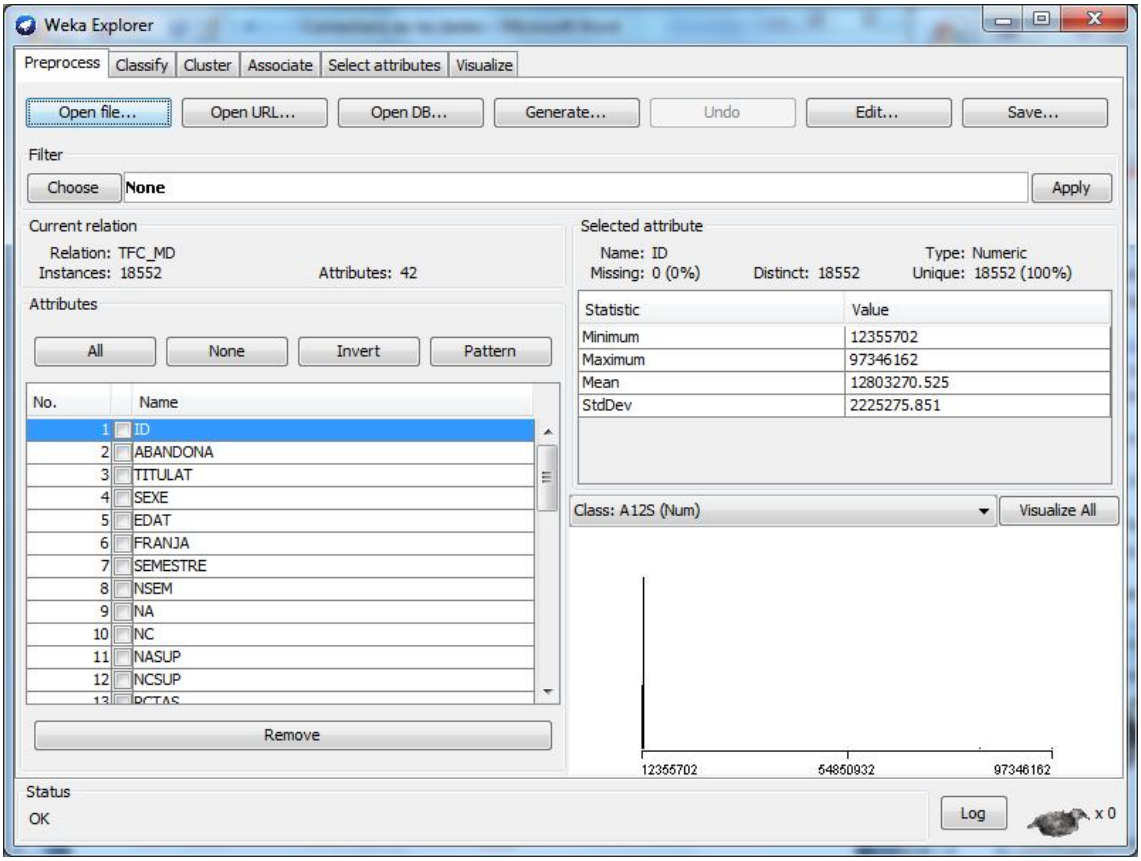
Fent una cerca del problema s'ha trobat la web http://support.microsoft.com/kb/323626/es . El problema es dona si a la primera línea les 2 primeres lletres estan en majúscules i la solució era que obris amb el bloc de notes i li afegeixo un apòstrof al davant de la 1ª línea.

Ara si he pogut obrir el fitxer **TFC_MD.dat** amb l'excel i l'he desat com a **TFC_MD.csv**

Ara al obrir aquest fitxer **TFC_MD.csv** amb el weka, ho posava tot com un sol camp.

He editat aquest fitxer amb el bloc de notes i la separació dels camps estava amb punt i coma i la he substituït amb comes.

Ara ja s'ha pogut obrir les dades amb el weka!



Ara es pot veure el contingut de cada camp i valorar el tipus d'informació que aporta. En l'annex hi ha els valors màxims, mínims, mitjana i una visualització gràfica de les dades..

Les dades són de 10 cursos. Total 20 semestres.

Inici 1998 2n semestre i final 2008 1r semestre. Total: 18552 alumnes

ID Cada estudiant té un ID únic que va des del 12.355.702 al 97.346.162

Aquest camp, en principi no aporta cap dada rellevant, ja que és diferent per a cada alumne.

ABANDONA Hi ha 13865 estudiants que NO abandonen (valor 0) i 4687 que SI (valor 1)

Aquest camp si que interessa, ja que precisament és l'objectiu a relaciona amb altres dades. Podem veure que el **25 %** dels alumnes matriculats abandonen els estudis.

TITULAT Hi ha 18548 estudiants que NO es titulen en el 1r semestre i 4 que SI

Aquest camp, d'inici, no serà interessant ja que l'objectiu no es estudiar els que es titulen sinó els que abandonen els estudis.

SEXE 9200 SÓN DE SEXE 0 (dona) I 9352 SÓN DE SEXE 1 (home)

Observem que els estudiants es divideixen a parts iguals en els dos sexes. Aquest camp pot donar informació.

EDAT Dels 17 als 75 anys en el moment d'entrar

Veiem que l'alumne més jove en tenia 17 anys i el més vell 75 anys quan van fer la matrícula. També pot ser un camp a estudiar.

FRANJA 4310 → 1 (<=24) , 4393 → 2 (<=27) , 3483 → 3 (<=30) , 3968 → 4 (<=36) , 2398 → 5 (>36)

Aquest camp també reflecteix l'edat però de forma discretitzada. S'ha de valorar si utilitzo aquest rangs d'edat o es creen d'altres més adients.

SEMESTRE (d'inici dels estudis) 898 → 1998 2n semestre ... més... 1082 → 2008 1r semestre (hi ha matrícula en tots els semestres)

Alumnes matriculats en cada semestre. Es pot donar el cas que coincideixin molts abandonaments en semestres concrets.

NSEM (nº se semestre relatiu des de que van iniciar estudis)

Indica el número de semestre des de que es va iniciar els estudis. Per exemple els del semestre 20082 tenen el valor 6. Això vol dir que fa 6 semestres des de que està aquest pla d'estudis.

NA (assig. matriculades en el 1r semestre) 546→ 1 , 4427→2, 8430→3, 3653 →4, 951→5, 423→6, 111→ 7, 9 →8, 1→ 9, 1 → 13

Aquest camp recull el número d'assignatures matriculades en el 1r semestre. Aquesta dada pot ser important, ja que un nombre alt d'assignatures pot ajudar a l'abandonament per massa càrrega horària. Podem observar que hi ha 546 alumnes que només s'han matriculat d'una assignatura i com a extrem 1 estudiant que s'ha matriculat de 13 assignatures.

NC (crèdits. matriculats) 401 → 4,5 , 145→6, 1→7,5 , 928→9, 2865→11, més....53 → 37,5 més.. 1→48, 1→69

El mateix que el cas anterior però mesurat en nombre de crèdits. És possible que menys assignatures puguin tenir una càrrega horària més alta pel fet que aquestes siguin de molts crèdits. Podem observar que hi ha 401 alumnes que només s'han matriculat de 4,5 crèdits (segurament 1 assignatura)

NASUP (assignatures superades) 5142→0, 2610→1, 4393→2, 4511→3, 1443→4, 314→5, 114→6, 23→7

Aquest camp recull el número d'assignatures superades en el 1r semestre. Abans havíem vist que un estudiant s'havia matriculat de 13 assignatures i aquí veiem que el màxim d'aprovades és de 7. Encara que pot no ser el mateix estudiant, segur que aquest alumne va suspendre com a mínim 6 assignatures, gairebé la meitat.

NCSUP (crèdits superats) 5142→0, 2065→4,5, 545→6, més..., 1→52,5

El mateix que l'altre camp però mesurant els crèdits superats.

PCTAS (% assignatures superades NASUP / NA) 5142→1, ... 1263→ ≈0,5 ... 8065 → 1

Aquest camp és un camp calculat entre les assignatures aprovades i les matriculades.

PCTCS (% crèdits superades NCSUP / NC) 5142→1, ... més ... 8065 → 1

Aquest camp és un camp calculat entre els crèdits aprovats i els matriculats.

En aquest dos casos s'ha de valorar si n'hi ha prou en valorar el percentatge de superació respecte l'abandonament dels estudis.

VIA 3787 → 1 (NO COU) , 3111 → 2 (COU) , 7609 → 3 (EST. INACABATS), 4045 → 4 (TITULATS)

Un camp a tenir en compte és els estudis previs que tenen els estudiants abans de fer la 1ª matrícula. A priori pot semblar que a menys estudis més risc a l'abandonament.

NACMAT Nº assig. A les que es matricula del conjunt de 12 més comuns 1r semestre

360→0, 1649→1, 5221→2, 7412→3, 2891→4, 692→ 5, 289→6, 37→7, 1→10

Aquí es pot veure la coincidència de la 1ª matrícula amb les 12 assignatures més comunes del 1r semestre. Es pot suposar que aquestes 12 més comunes podrien ser les que es recomana en el pla d'estudis.

NACPRE Nº assig. A les que es presenta del conjunt de 12 més comuns 1r semestre

4673→0, 3156→1, 4451→2, 4404→3, 1464→4, 295→ 5, 98→6, 10→7, 1→8

Donat que no sempre es presenten els alumnes de totes les assignatures matriculades, aquesta data pot influir en l'abandonament dels estudis. Un alumne que no es presenta a cap assignatura potser més endavant ho deixa tot.

NACSUP Nº assig. Superades del conjunt de 12 més comuns 1r semestre

5512→0, 3213→1, 4440→2, 3910→3, 1169→4, 232→ 5, 68→6, 8→7

El fet de que es presenti a un examen, no vol dir que ho aprovi. El resultat del mateix pot influir amb els ànims de continuar estudiant.

(La relació d'assignatures ve amb les dades i s'ha preguntat si l'ordre coincideix amb l'ordre del nom del camp A1, A2, ..., A12. La resposta ha estat que si.)

A1M 4327 → 0 (NO es matricula) , 14225 → 1 (SI es matricula)

00.010: multimèdia i comunicació

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A1S 857 → -1 (NO supera) , 7809 → 0 (NO matricula o NO es presenta) , 9886 → 1 (SI supera)

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat. Realment els que no s'han presentat serien la resta dels que tenen aquest valor (0) menys els que no s'han matriculat.

A2M 14424 → 0 , 4128 → 1

00.002: angles I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A2S 229 → -1 , 16000 → 0 , 2323 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A3M 11906 → 0 , 6646 → 1

01.001: introducció al dret

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A3S 303 → -1 , 14213 → 0 , 4036 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A4M 12670 → 0 , 5882 → 1

01.079: introducció a la macroeconomia

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A4S 412 → -1 , 15101 → 0 , 3039 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A5M 12618 → 0 , 5934 → 1

01.005: introducció a la comptabilitat

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A5S 410 → -1 , 14807 → 0 , 3335 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A6M 14702 → 0 , 3850 → 1

01.003: matemàtiques I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A6S 358 → -1 , 16625 → 0 , 1569 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A7M 14625 → 0 , 3927 → 1

01.006: organització i administració d'empreses I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A7S 120 → -1 , 16037 → 0 , 2395 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A8M 16217 → 0 , 2335 → 1

01.004: estadística I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A8S 238 → -1 , 17440 → 0 , 874 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A9M 17160 → 0 , 1392 → 1

01.078: introducció a la microeconomia

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A9S 66 → -1 , 17707 → 0 , 779 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A10M 17479 → 0 , 1073 → 1

01.009: direcció de la producció I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A10S 75 → -1 , 17925 → 0 , 552 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A11M 17487 → 0 , 1065 → 1

00.004: angles III

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A11S 25 → -1 , 17786 → 0 , 741 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

A12M 17655 → 0 , 897 → 1

00.003: angles II

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

A12S 51 → -1 , 17907 → 0 , 594 → 1

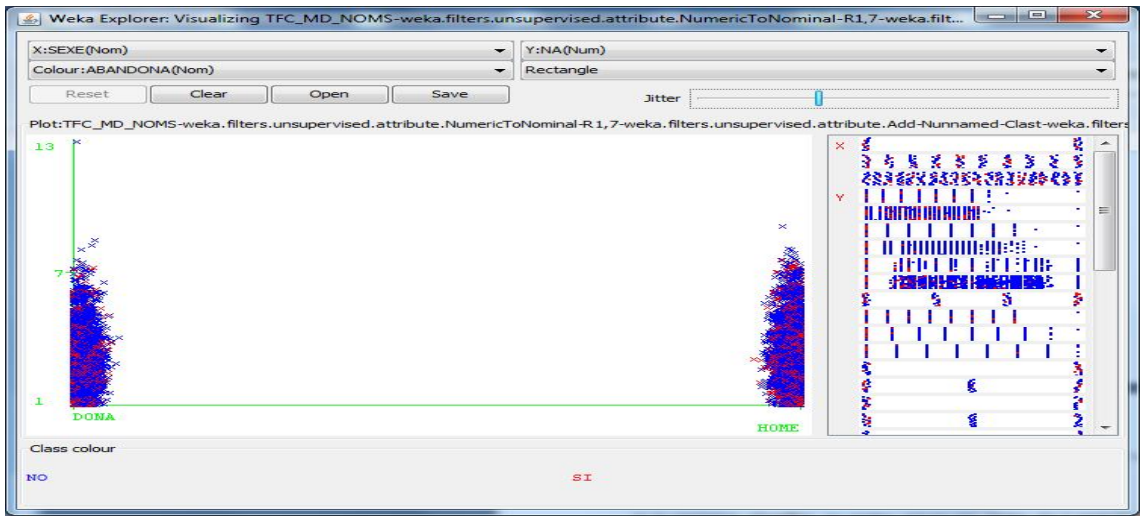
Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

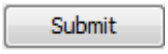
3.3. Visualize. Una eina per veure més informació.

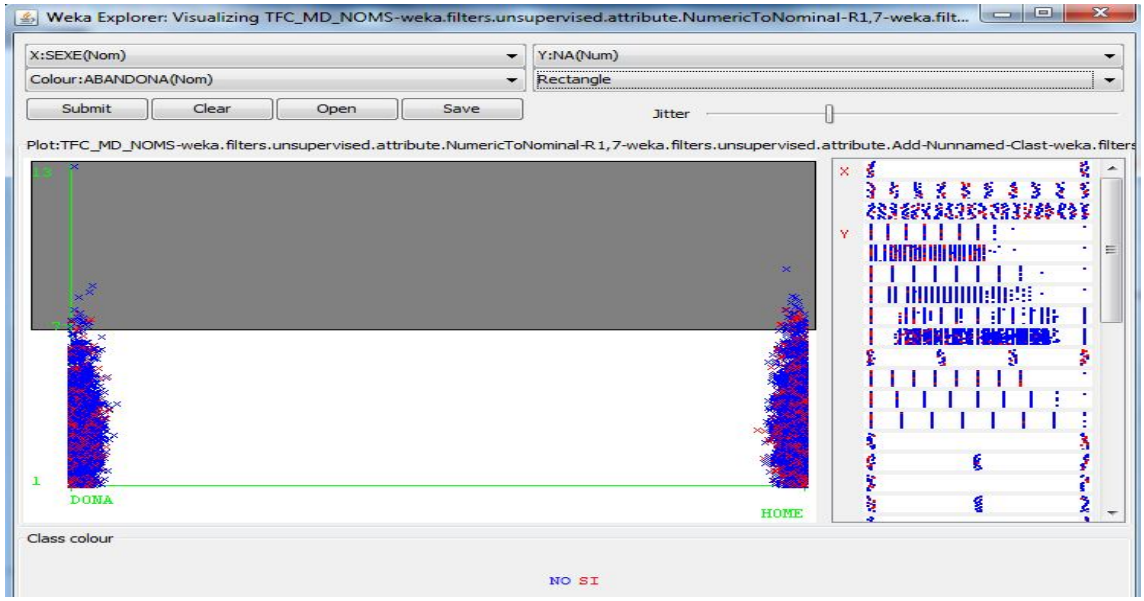
A la pestanya **visualize**, es poden creuar les dades de 2 atributs de forma gràfica posant cadascun d'ells en els eixos **x** i **y**, i a més es pot posar un 3r atribut que diferenciï amb color el seu valor. Alguns creuaments no aporten gaire informació però d'altres si que poden ser interessants.

En els gràfics següents es diferencia amb color blau els que **NO** abandonen i amb vermell els que **SÍ**.

Relacionant **SEXE** i **NA** es pot veure visualment com es reparteixen els homes i les dones segons el número d'assignatures matriculades i a més amb vermell els que **SÍ** han abandonat.



Si es vol es pot seleccionar amb un rectangle, polígon o polilínea una quantitat determinada de dades. Per exemple, amb l'opció rectangle es marca els valors que superen les 7 assignatures matriculades, es clica al botó 

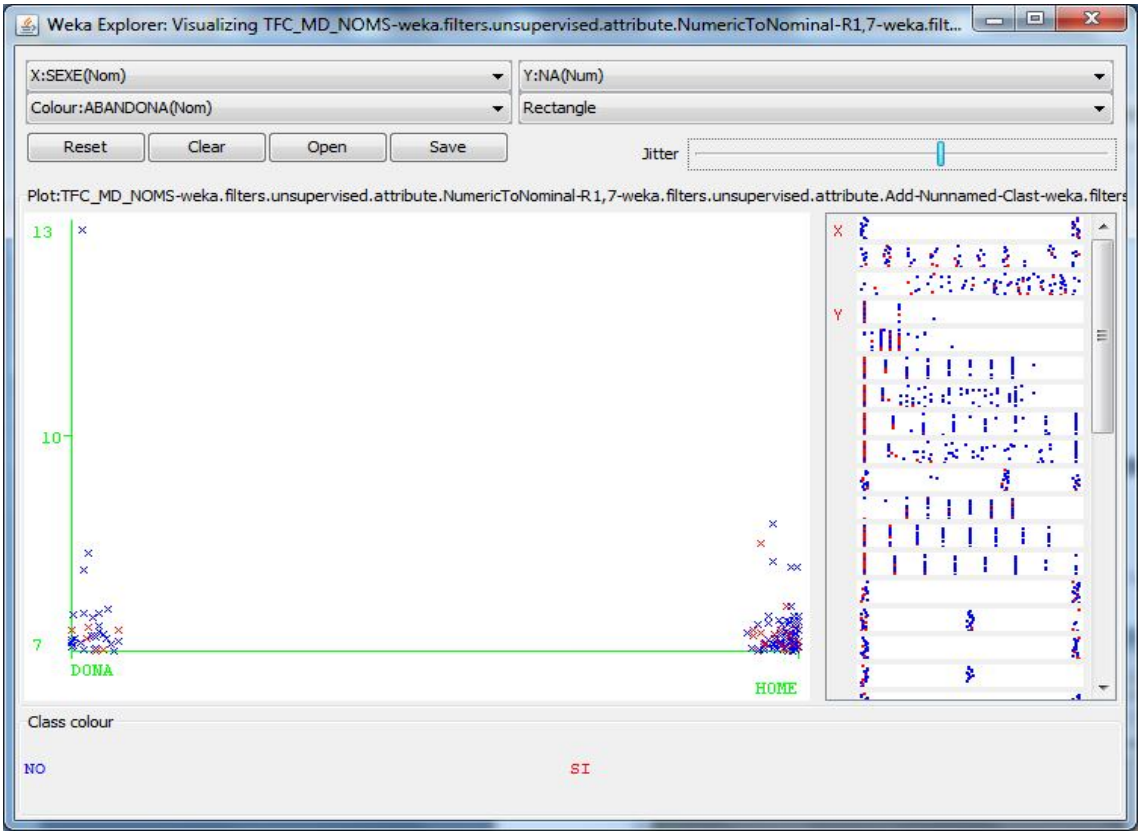


Es pot observar que en les instàncies que s'han matriculat per sobre de 7 assignatures hi ha majoria que **NO** abandonen els estudis. Fins i tot es veu que l'estudiant que de més assignatures s'ha matriculat és una dona que tot tenint 13 assignatures matriculades, **NO** ha abandonat els estudis. Clicant a sobre de la creu s'obre una finestra amb totes les dades d'aquesta instància. En aquest cas en concret aquesta alumne va aprovar el 76,9% d'aquestes 13 assignatures.

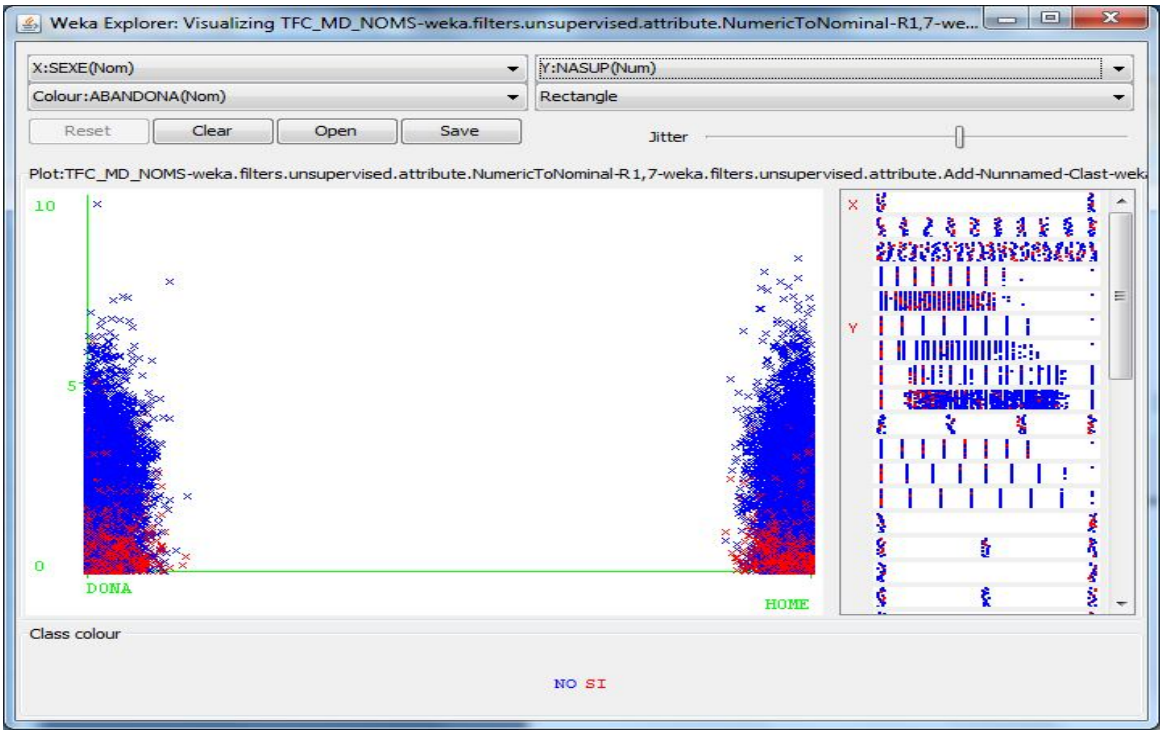
```
Plot : new plot
Instance:  63
    SEXE: DONA
    EDAT:    '(27.5-
29.5]'
SEMESTRE: 20071
      NA: 13.0
      NC: 69.0
   NASUP: 10.0
   NCSUP: 52.5
   PCTAS: 0.769231
   PCTCS: 0.76087
     VIA: NO_COU
  NACMAT: 10.0
```

```
  NACPRE: 8.0
  NACSUP: 7.0
     A1M: SI
     A1S: SI
     A2M: SI
     A2S: SI
     A3M: SI
     A3S: SI
     A4M: SI
     A4S: SI
     A5M: SI
     A5S: SI
     A6M: SI
     A6S: NP_NM
```

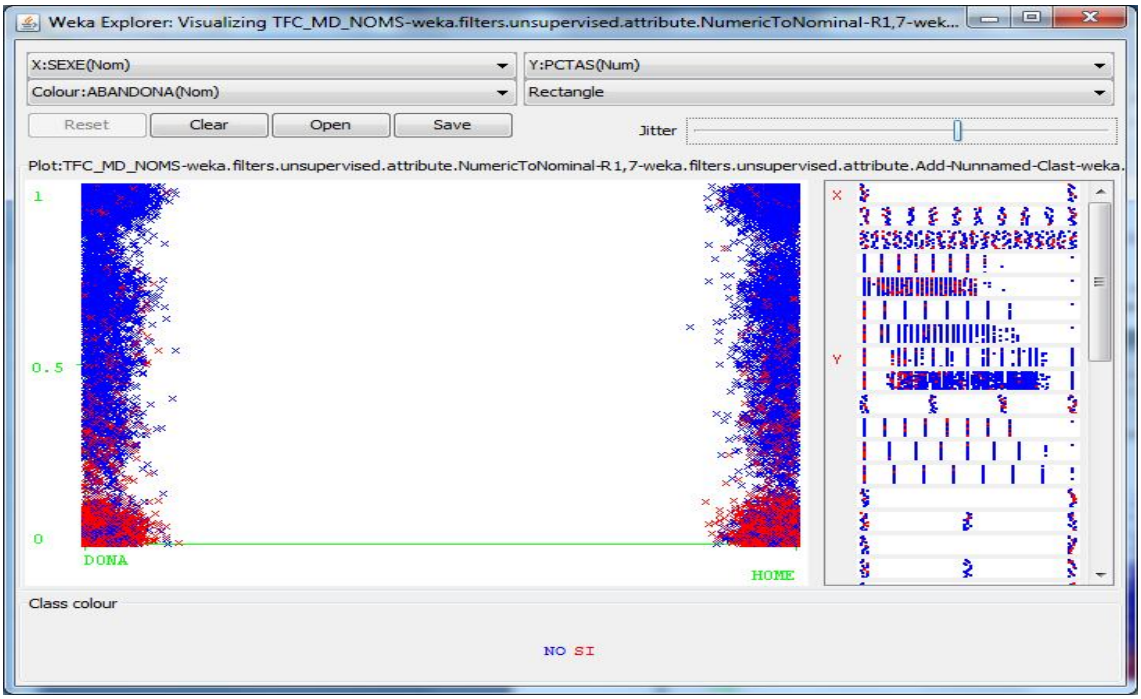
```
     A7M: SI
     A7S: SI
     A8M: SI
     A8S: NP_NM
     A9M: SI
     A9S: SI
    A10M: SI
    A10S: NO
    A11M: NO
    A11S: NP_NM
    A12M: NO
    A12S: NP_NM
ABANDONA: NO
```



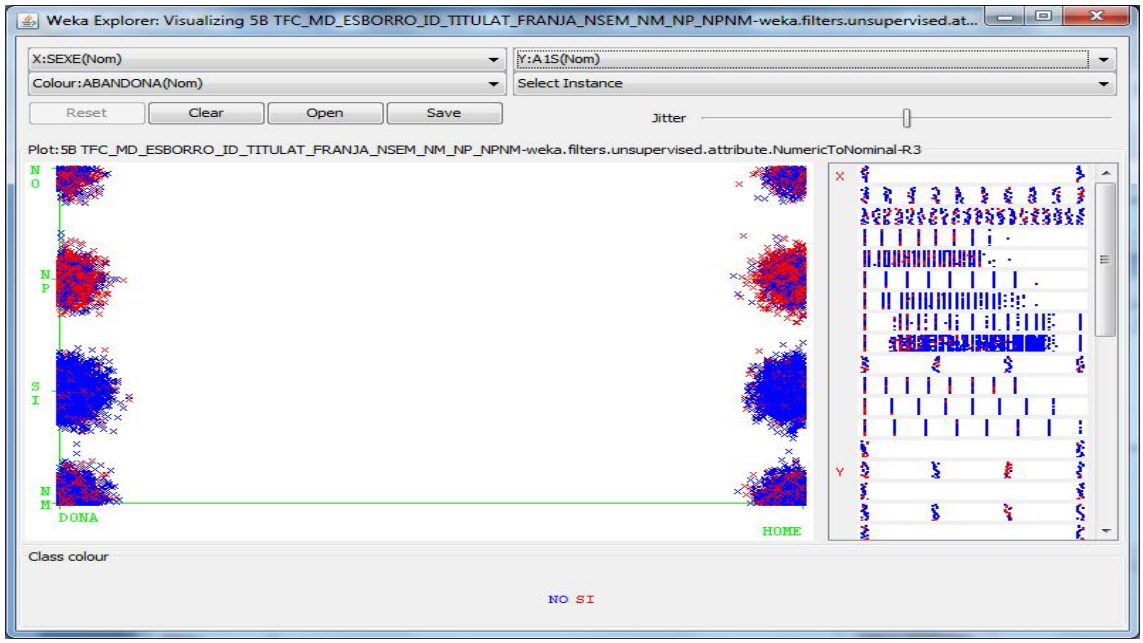
Relacionant **SEXE** i **NASUP** es veu clarament que aquells alumnes que superen menys assignatures, tant homes com dones, són els que **SÍ** abandonen els estudis.



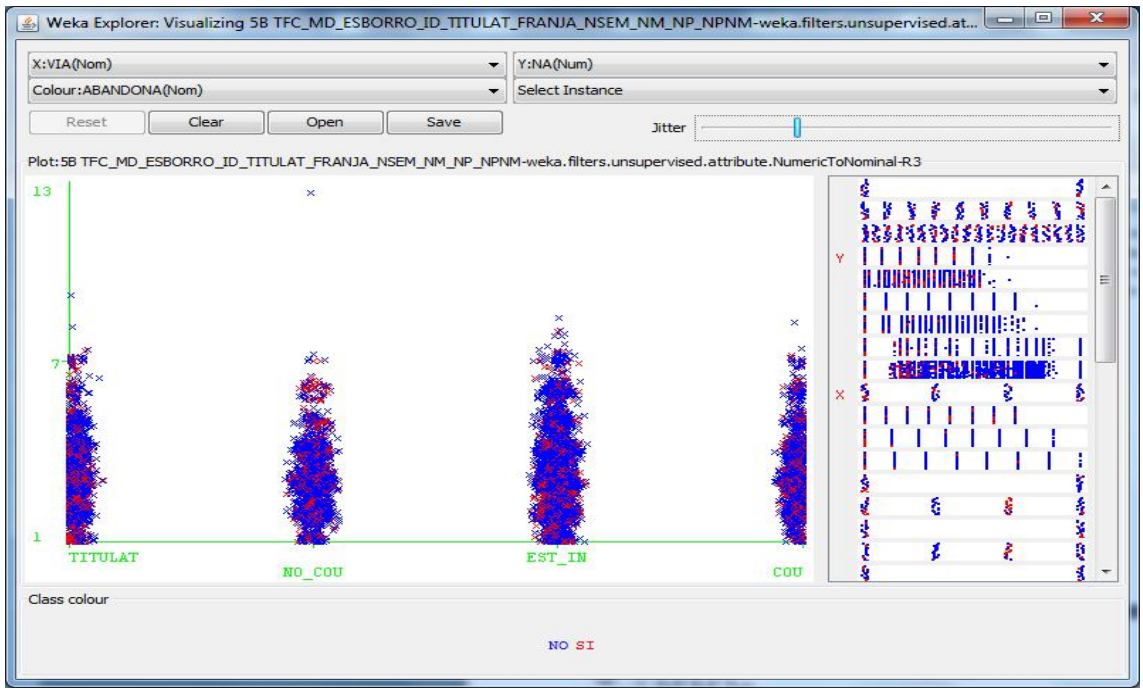
Al relacionar **SEXE** i **PCATS** es veu que encara que mentre més baix sigui el percentatge d'assignatures aprovades més instàncies hi ha que **SÍ** abandonen els estudis, això no ho assegura del tot. A dalt de tot es veu com hi ha alumnes que encara que ho han aprovat tot, al final també abandonen.



Ara si es relaciona **SEXE** i **A1S** es pot veure que aquells alumnes que aproven, majoritàriament continuen els estudis. Els que no es presenten tenen més possibilitats a abandonar els estudis que els que no aproven. Segurament el fet de presentar-se a l'examen, encara que no aprovin, demostra la persistència en els estudis. El fet que no es matriculi d'aquesta assignatura no indica que abandoni els estudis.



Al creuar **VIA** i **NA**, pot veure que l'alumne que s'ha matriculat de 13 assignatures ha accedit als estudis via **NO_COU**.



3.4. Estratègia segons les dades trobades

Aquest primer estudi de les dades s'ha fet de forma separada sense veure la coincidència o no de diferents aspectes que poden influir en l'abandonament dels estudis. Ara el que es tracta és d'anar creant diferents models de mineria de dades per tal de veure les diferents regles que poden justificar l'abandonament dels estudis d'aquests estudiants d'economia.

La meva intenció és provar tots els models i veure quins són els resultats i si van en la línea del que vull trobar, un model que en permeti trobar una regla que "justifiqui" la raó per la que un nombre important d'alumnes deixen els estudis d'economia.

Les dades que m'han proporcionat no tenen cap valor nul i per tant no cal omplir-los amb valors segons diferents mètodes: posar la mitjana, el valor més repetit, ...

Hi ha alguns camps que potser hauré de plantejar si em calen. Cal tenir en compte només el nombre d'assignatures matriculades?, o el nombre de crèdits? Segurament provaré amb els dos tipus d'informació i després valoraré si calen tots dos.

També hi ha camps, per exemple el % de superats, que puc treballar amb ells o amb els valors numèric de superats i matriculats.

En camps que hi ha molta diversitat de valors, per exemple l'edat, he de plantejar rangs de discretització que siguin raonables.

El tipus d'assignatura, de moment no la tractaré de forma diferent. Però una vegada tingui alguns resultats puc intentar trobar alguna analogia entre algunes assignatures, per exemple que pertanyin a la mateixa àrea.

L'estratègia que seguiré és adaptar les dades d'alguns camps perquè em permetin realitzar la mineria de dades. Discretitzar alguns camps, no utilitzar alguns atributs, ... En els comentaris que he anat fent per cada camp, ja he comentat la importància del tipus de dada.

Després aplicaré els diferents models: d'agregació, associatius, arbres de decisió, xarxes neuronals, regles de classificació, xarxes bayesianes. En raó dels resultats que surtin en cada model, aniré seleccionant aquells que millor expliquin l'objectiu específic proposat. Es a dir, es pot justificar l'abandonament d'aquests estudis i amb aquestes justificacions, es poden preveure possibles abandonaments?

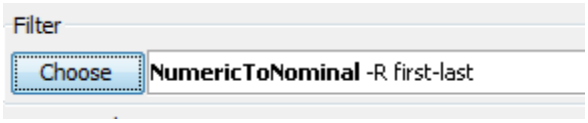
3.5. Preparació i depuració de les dades

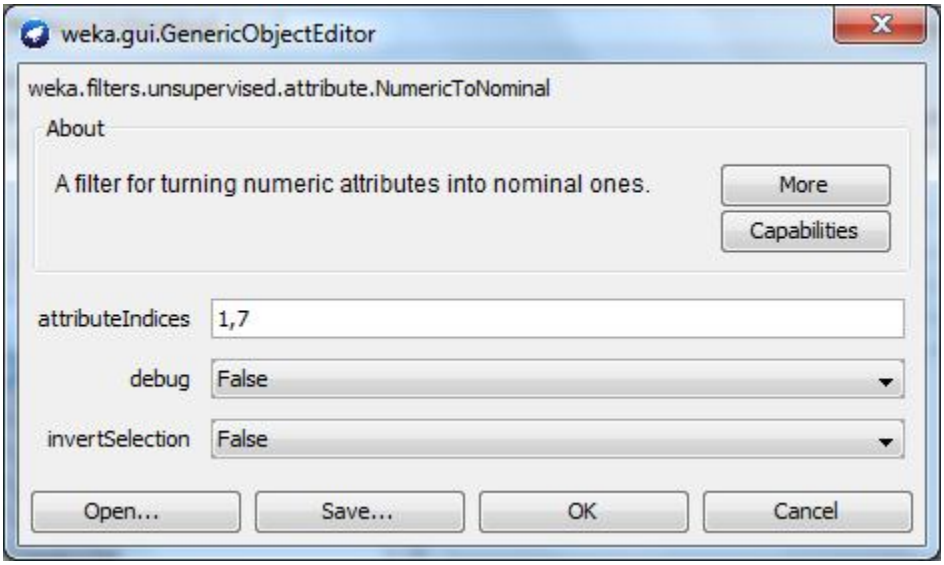
Una vegada ja es pot obrir les dades amb el weka des de el fitxer **TFC_MD.csv** es factible desar una còpia amb el format propi de Weka, **TFC_MD.arff**. Una vegada fet això, s'han de preparar les dades per tal que la informació sigui d'utilitat. Cal veure si tots els camps tenen la informació que els hi pertoca i dintre dels límits coherents que corresponen a cada atribut. També s'ha de veure si hi ha dades amb valor *null* i plantejar-se de quina manera cal omplir aquests valors.

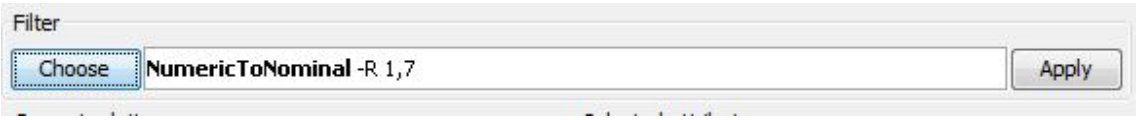
El primer camp **ID**, té un identificador únic per a cada estudiant. En la conversió, weka ha agafat aquest camp com a numèric i donat el significat del camp, això no té sentit. En aquest cas, pot interessar passar aquest camp de numèric a nominal.

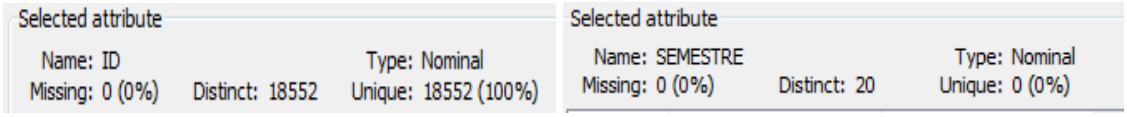
Una vegada s'ha plantejat aquesta reflexió, es poden cercar altres camps que tinguin el mateix plantejament. S'ha trobat el camp **SEMESTRE**, que també l'ha agafat com a numèric i interessa més que sigui de tipus nominal.

Això es fa amb el weka, dintre de la pestanya **Preprocess**, amb el filtre **NumericToNominal**. En aquest cas es vol aplicar als atributs **1** i **7** (**ID** i **SEMESTRE**), per tant es clicarà a sobre del filtre i s'indicarà en **attributeIndices** els valors **1** i **7**. Una vegada fet això, se li donarà al botó **Apply** i es podrà observar com aquests dos camps ara són nominals.









Hi ha altres camps que tenen valor numèric però només existeixen els valors **0** i **1**. Aquests valors representen realment els valors **NO** pel **0** i **SI** pels **1** pels camps **ABANDONA**, **TITULAT**, **A1M, A2M, A3M, A4M, A5M, A6M, A7M, A8M, A9M, A10M, A11M** i **A12M**. Per ser més exacte, els dos primers camp representen els valors **FALS** i **CERT**, però donat el significat similar es procedeix a substituir els valors **0** d'aquests camps pel valor **NO** i els valors **1** pel valor **SI**. Això s'ha fet des de l'Excel amb l'ordre *Reemplazar* i seleccionant prèviament les columnes que representes aquests camps. D'aquesta manera aquests camps seran tractat com a nominals però amb un valor que dona un significat més directe del que representa.

El camp **SEXE**, també té valor numèric **0** i **1** i representa respectivament el significat **DONA** i **HOME**. De la mateixa manera que en el paràgraf anterior, es procedeix a substituir els valors numèrics pel nom que representa.

El camp **VIA**, pot agafar valors numèrics amb **1, 2, 3** i **4** com a valors possibles. Aquests números representen respectivament **NO_COU, COU, ESTUDIS_INACABATS i TITULAT**. Amb el mateix procediment substituïm els valors numèrics amb valors nominals. S'abrevia alguns d'aquests valors perquè siguin més curts. Els valors posats son: **NO_COU, COU, EST_IN i TITULAT**

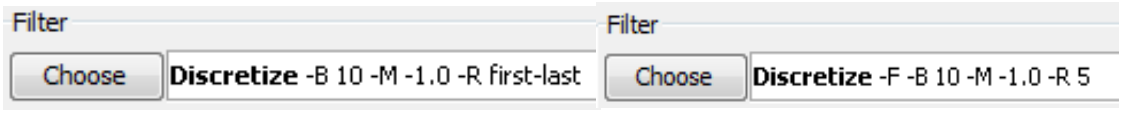
Els camps **A1S, A2S, A3S, A4S, A5S, A6S, A7S, A8S, A9S, A10S, A11S** i **A12S** contenen valors numèrics que poden ser el **-1, 1** i **0**. Indiquen respectivament que **NO** han superat l'assignatura, que **SI** l'han superada i que o bé no s'han matriculat o no s'han presentat **NP_NM**. Es procedeix a substituir els valors numèrics pels valors nominals abans esmentats.

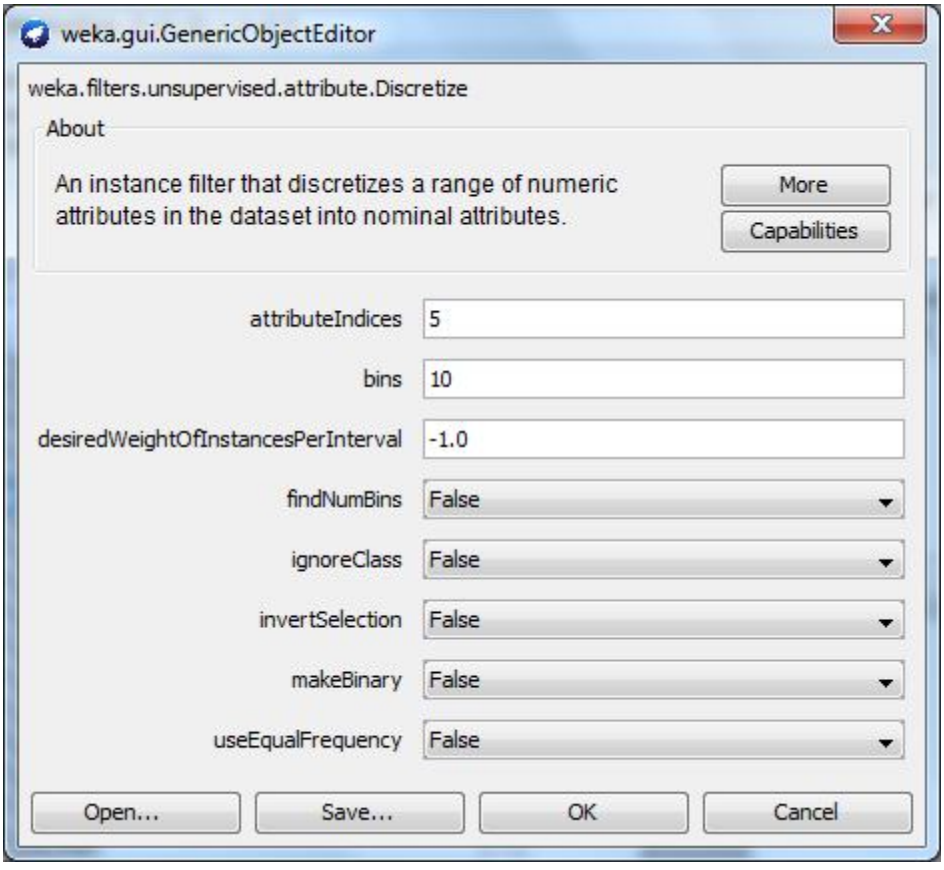
Els camps **NSEM, NA, NC, NASUP, NCSUP, PCTAS, PCTCS, NACMAT, NACPRE** i **NACSUP** que tenen valors numèrics, de moment es deixem com a numèrics. És possible que més endavant pugui interessar discretitzar alguns d'aquests camps.

El camp **FRANJA**, representa 5 franges d'edats dels estudiants. Just el camp anterior, **EDAT**, té l'edat en valor numèric que tenen aquests estudiants. Es pot valorar si interessa fer anàlisi amb 5 franges d'edat o en un número diferent. Per fer-ho es pot discretitzar el camp **EDAT** en 10

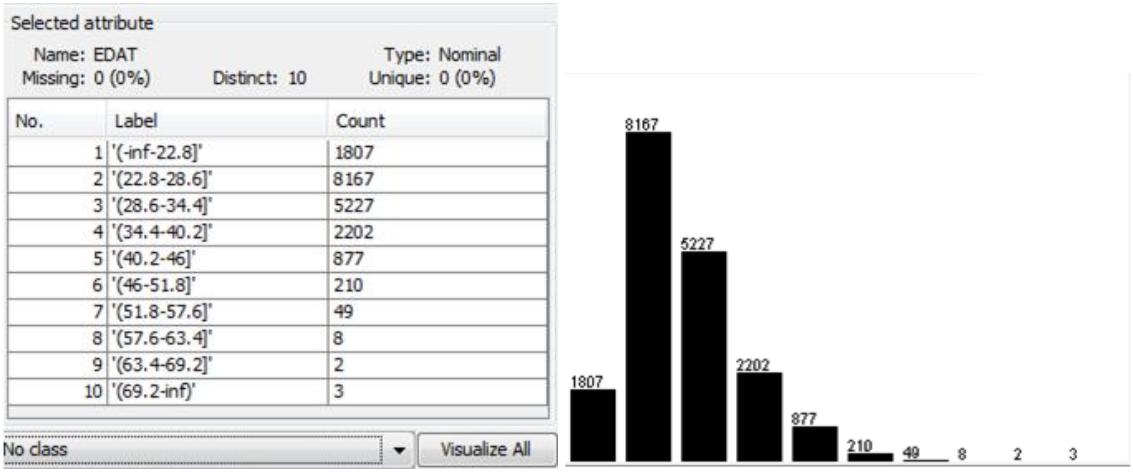
intervals per tal que hi hagi més diversitat de franges. Això es fa amb el filtre **Discretize** que està en la pestanya **Preprocess**

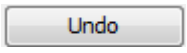
Per defecte aquest filtre discretitzaria tots els camps numèrics. Donat que només interessa fer-ho amb l'atribut **EDAT**, es clica en el filtre i es posa en **attributeIndices** el valor 5 que correspon al camp desitjat. Es deixa el valor en bins a 10, que vol dir que crearà automàticament 10 intervals.



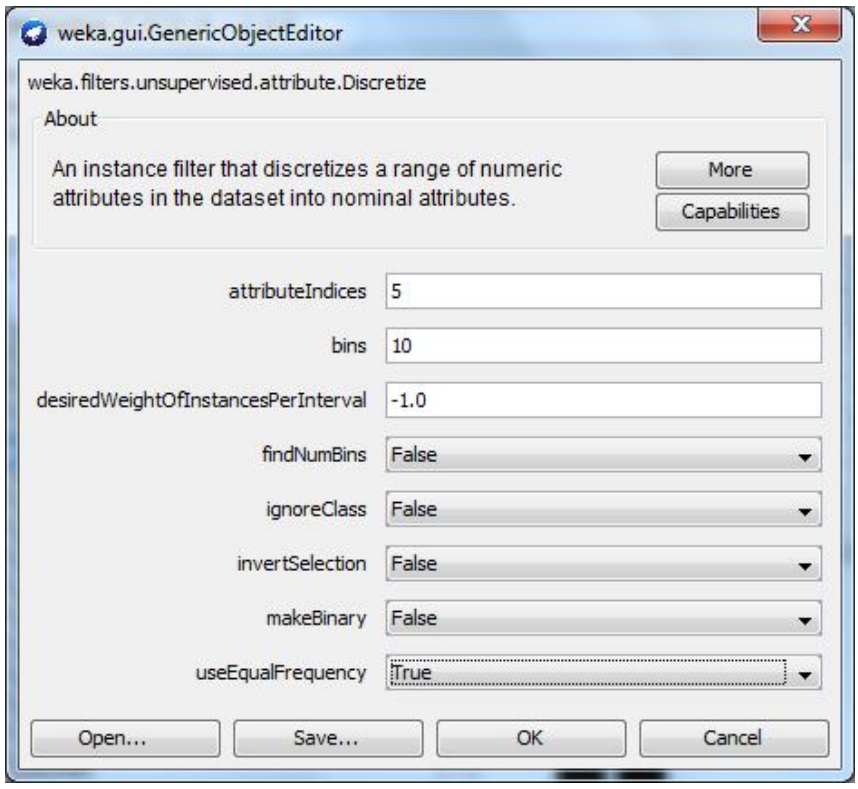


Després de donar-li a **Apply**, s'observa la discretització que ha fet i es veu que els intervals estan molt descompensats. Encara que els 8 intervals que estan al mig estan separats cada 5,8 anys, la quantitat d'estudiants de cada interval té moltes diferències.

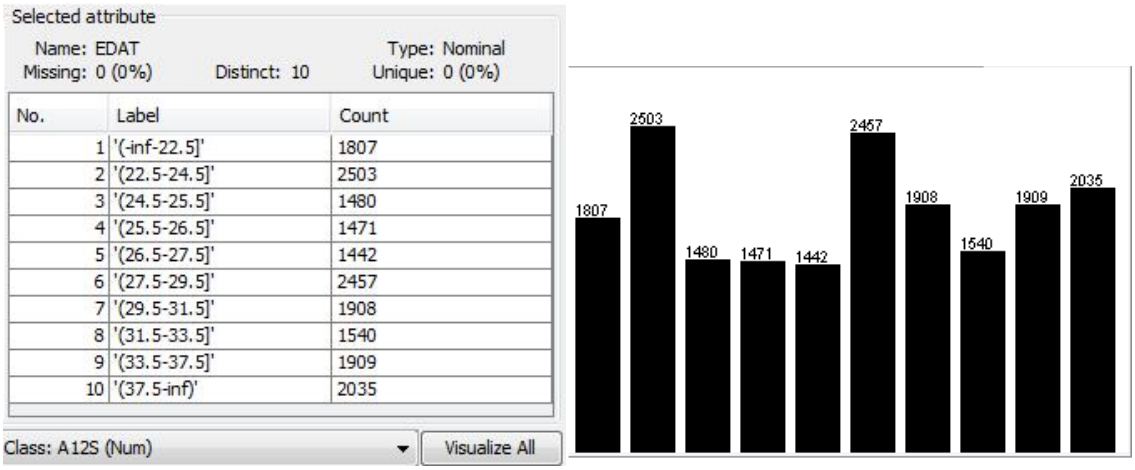


Vist això, es desfà aquesta discretització amb el botó 

i es procedeix a posar a **True** el valor de **useEqualFrequency**. Això fa que la distribució dels intervals es faci mirant de que la quantitat d'estudiants de cada interval sigui aproximadament el mateix.



Ara es pot observar el resultat següent, que està més equilibrat pel que fa a instàncies de cada interval.



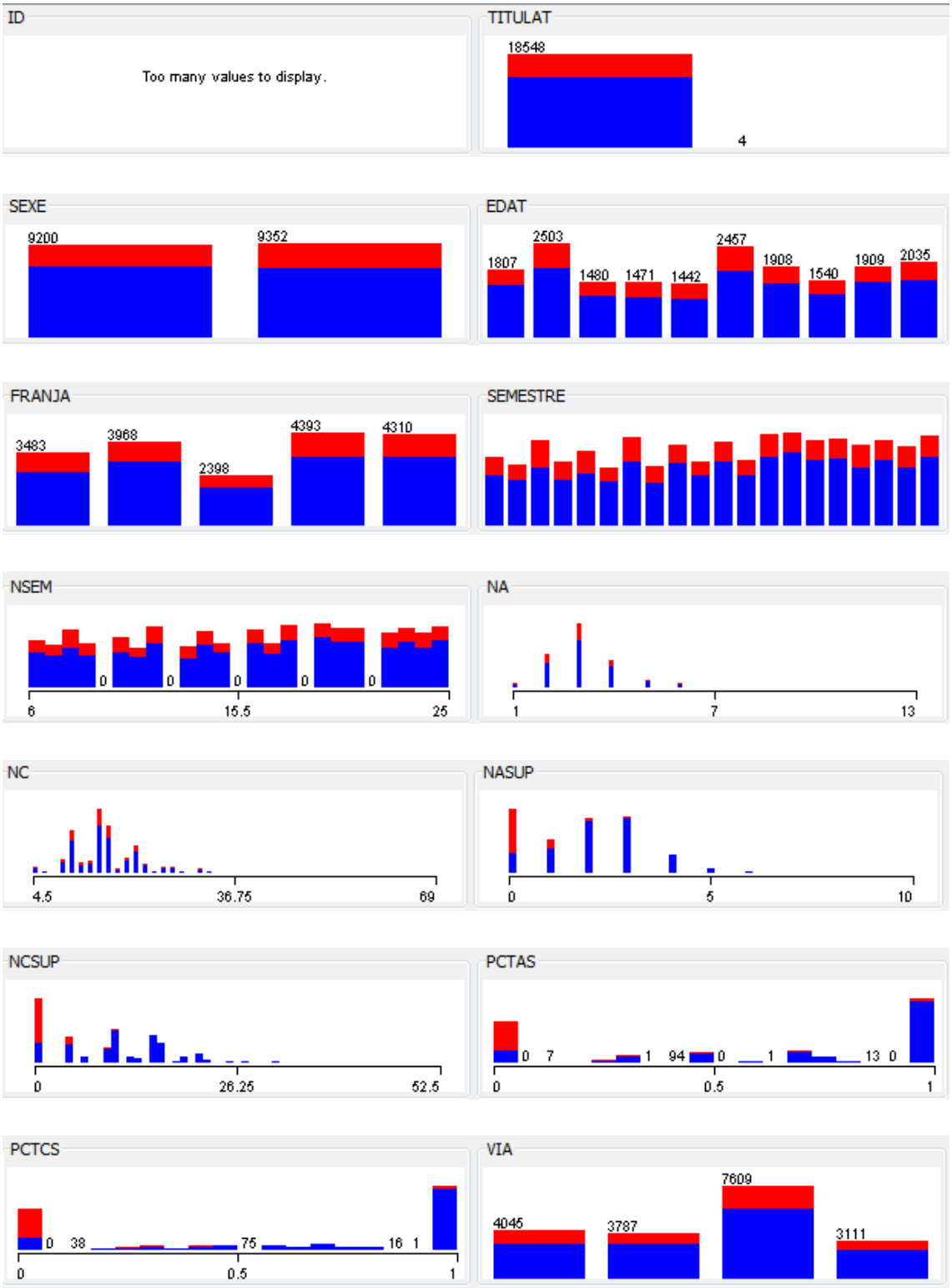
Després de fer tots aquests canvis, ja es té una proposta de valors normalitzats.

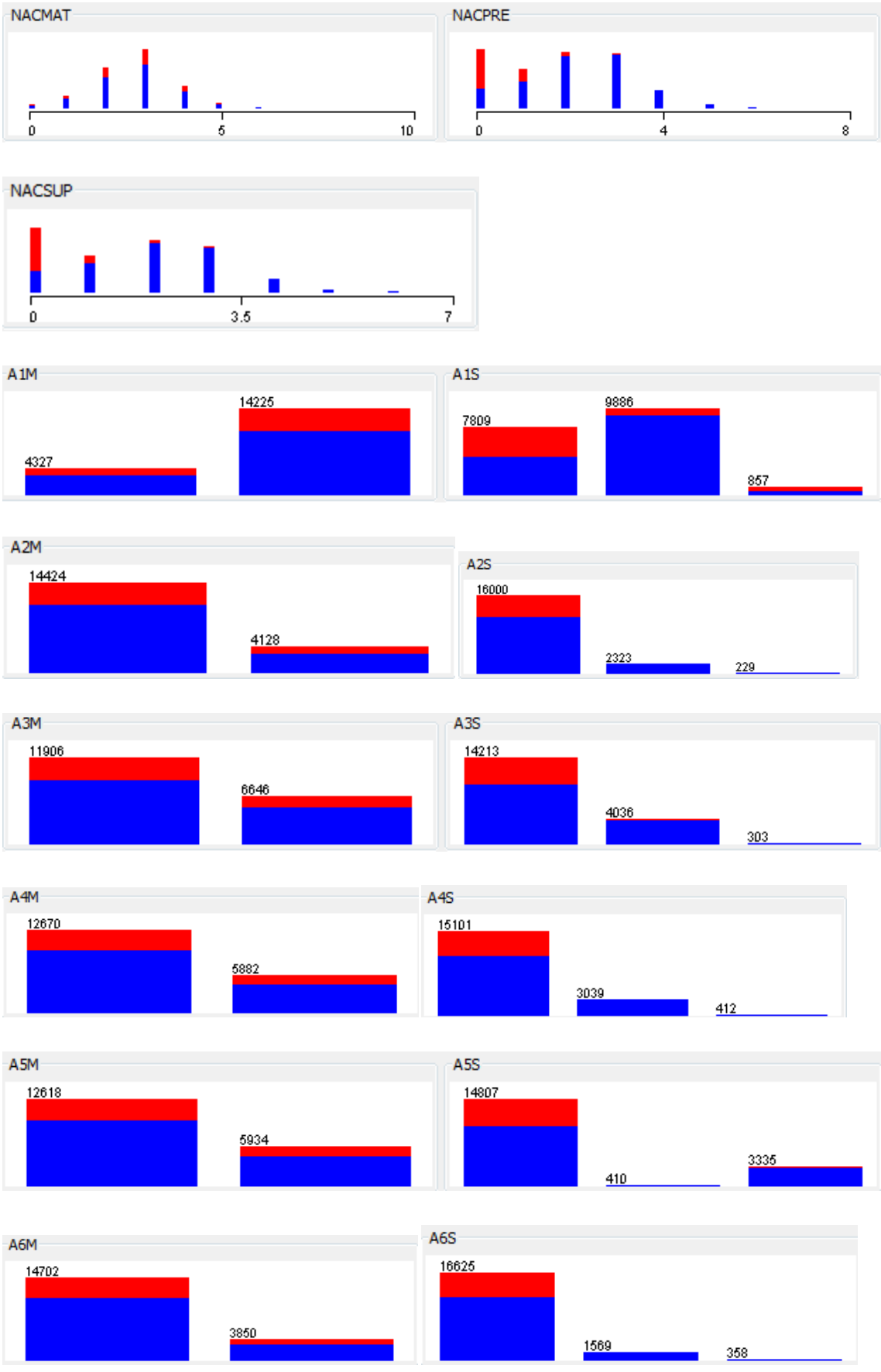
El resum dels canvis fets és el següent:

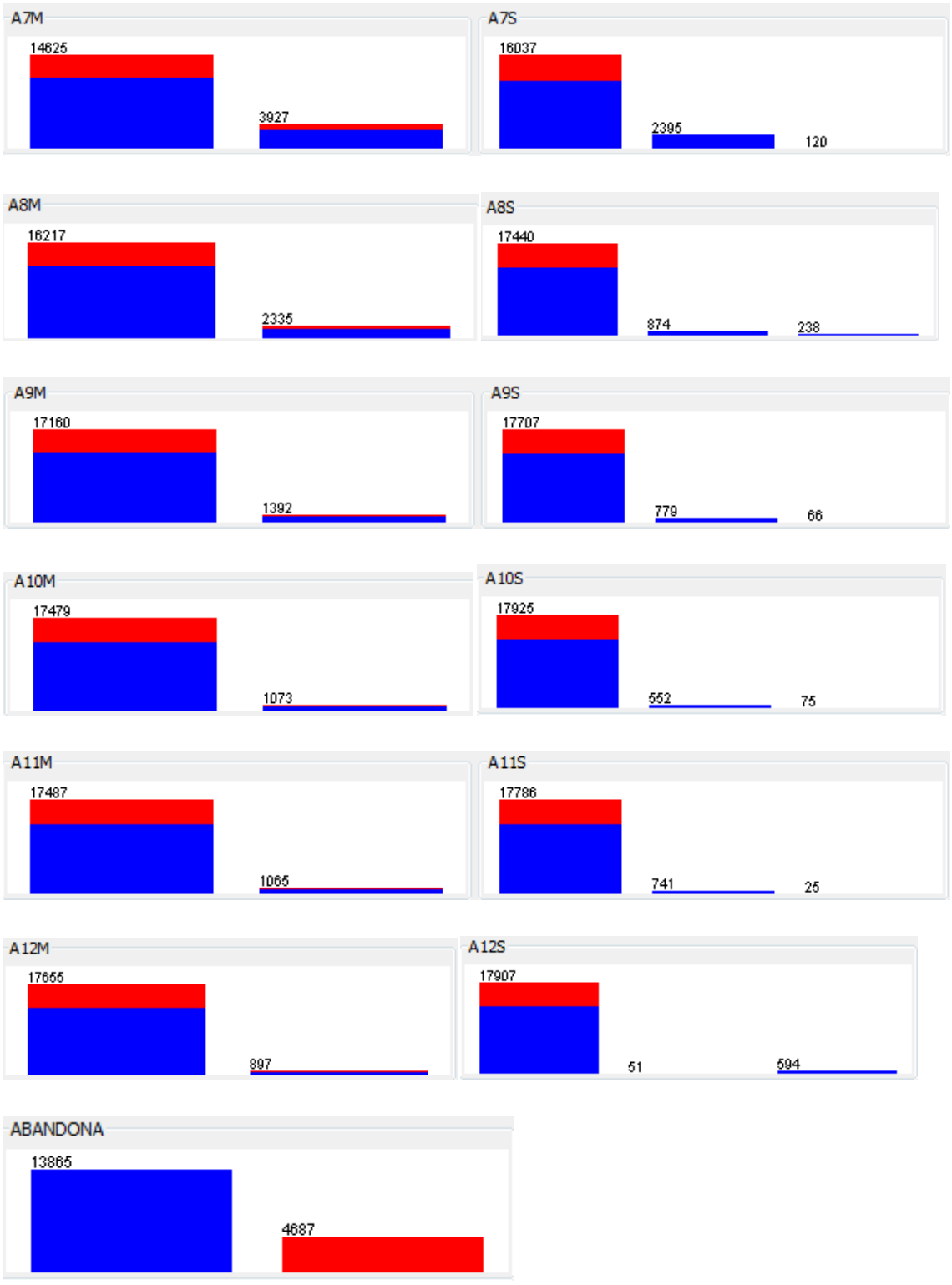
Nº	ATRIBUT	TIPUS	CANVI	NOU TIPUS
1	**ID**	Numèric: valors no repetits	**numericToNominal**	Nominal
2	**ABANDONA**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
3	**TITULAT**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
4	**SEXE**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**HOME, DONA**}
5	**EDAT**	Numèric	**Discretize**	10 intervals d'edat
6	**FRANJA**	Intervals		
7	**SEMESTRE**	Numèric	**numericToNominal**	Nominal
8	**NSEM**	Numèric		
9	**NA**	Numèric		
10	**NC**	Numèric		
11	**NASUP**	Numèric		
12	**NCSUP**	Numèric		
13	**PCTAS**	Numèric		
14	**PCTCS**	Numèric		
15	**VIA**	Numèric: {**1,2,3,4**}	**Reemplazar** (Excel)	Nominal: {**NO_COU, COU, EST_IN, TITULAT**}
16	**NACMAT**	Numèric		
17	**NACPRE**	Numèric		
18	**NACSUP**	Numèric		

19	**A1M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
20	**A1S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
21	**A2M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
22	**A2S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
23	**A3M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
24	**A3S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
25	**A4M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
26	**A4S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
27	**A5M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
28	**A5S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
29	**A6M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
30	**A6S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
31	**A7M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
32	**A7S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
33	**A8M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
34	**A8S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
35	**A9M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
36	**A9S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
37	**A10M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
38	**A10S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
39	**A11M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
40	**A11S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}
41	**A12M**	Numèric: {**0,1**}	**Reemplazar** (Excel)	Nominal: {**NO, SI**}
42	**A12S**	Numèric: {**0,1,-1**}	**Reemplazar** (Excel)	Nominal: {**NP_NM, SI, NO**}

Aquí hi ha el resum de les visualitzacions de les dades després de normalitzar els atributs.




NACMAT
0
5
10
NACPRE
0
4
8
NACSUP
0
3.5
7
A1M
4327
14225
A1S
7809
9886
857
A2M
14424
4128
A2S
16000
2323
229
A3M
11906
6646
A3S
14213
4036
303
A4M
12670
5882
A4S
15101
3039
412
A5M
12618
5934
A5S
14807
410
3335
A6M
14702
3850
A6S
16625
1569
358


A7M
14625
3927
A7S
16037
2395
120
A8M
16217
2335
A8S
17440
874
238
A9M
17160
1392
A9S
17707
779
66
A10M
17479
1073
A10S
17925
552
75
A11M
17487
1065
A11S
17786
741
25
A12M
17655
897
A12S
17907
51
594
ABANDONA
13865
4687

Una vegada es tenen els valors normalitzats i després de repassar les dades , es pot reduir el número d'atributs eliminant aquells que no aportin informació substancial.

D'entrada es veu que es pot eliminar sense cap problema els atributs següents:

ID: hi ha 18852 valors diferents, un per cada alumne. Al ser valors únics no tenen cap transcendència per determinar l'abandonament dels estudis.

TITULAT: Només hi ha 4 alumnes titulats. Aquesta quantitat respecte al total d'estudiants, més de 18000, es menyspreable.

FRANJA / EDAT: En **FRANJA** hi ha 5 intervals d'edat i en **EDAT** 10 intervals, es pot fer anar provant els 2 atributs per si hi ha diferències.

NSEM: Indica el número de semestre relatiu des de que van començar els estudis. La informació que hi ha en **SEMESTRE** que indica el semestre en que van iniciar els estudis té més significat.

Abans d'esborrar aquest atributs, es pot comprovar que no siguin importants per justificar el camp **ABANDONA** que s'ha posat en l'últim lloc ja que weka és el que mira per defecte.

Per mirar quins atributs són més importants, a la pestanya **Select attributes**, es selecciona com atribut avaluador **CfsSubsetEval** i com a mètode de cerca, l'algorisme **BestFirst**.

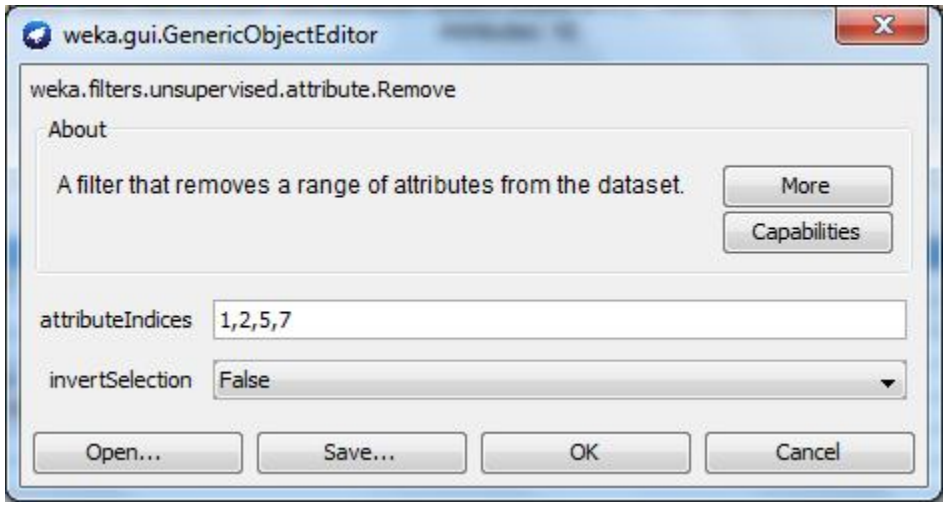
El resultat és el següent:

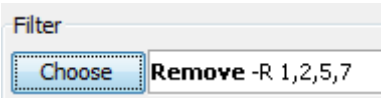
```
Evaluation mode:evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
        Best first.
        Start set: no attributes
        Stale search after 5 node expansions
        Total number of subsets evaluated: 575
        Merit of best subset found:    0.27

Attribute Subset Evaluator (supervised, Class (nominal): 42 ABANDONA):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 10,11,12,13,16,19,21,23,25,27,29,31 : 12
                     NASUP
                     NCSUP
                     PCTAS
                     PCTCS
                     NACPRE
                     A1S
                     A2S
                     A3S
                     A4S
                     A5S
                     A6S
                     A7S
```

A més dels atributs de número d'assignatures (**NASUP**) / crèdits (**NCSUP**) superats i corresponents percentatges (**PCTAS** i **PCTS**) i el número d'assignatures matriculades de les 12 més comunes en el primer any (**NACPRE**), es veu que són atributs importants els corresponents a les 7 assignatures en les que més es matriculen al 1r semestre. Cinc d'elles de les que es recomanen en el pla d'estudis pel 1r semestre, **A5** i **A7** es recomanen al 2n semestre.

A1 00.010: multimèdia i comunicació
A2 00.002: angles I
A3 01.001: introducció al dret
A4 01.079: introducció a la macroeconomia
A5 01.005: introducció a la comptabilitat
A6 01.003: matemàtiques I
A7 01.006: organització i administració d'empreses I

S'observa que cap dels 4 atributs que volia esborrar està dintre dels seleccionats com importants. Per tant, es pot procedir a esborrar-los. Per fer-ho, des de la pestanya **Preprocess** s'aplica el filtre **Remove**, indicant els valors corresponents als atributs a esborrar (**1,2,5,7**) en **attributeIndices**.





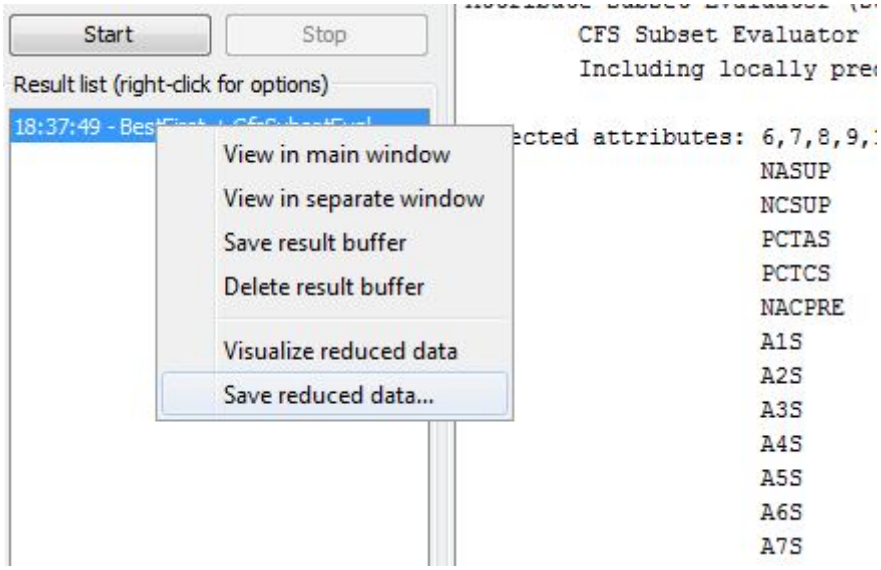
Al donar-li al botó 

, es queda reduït a 38 atributs. Es pot tornar a seleccionar quins són els atributs més importants per justificar l'atribut **ABANDONA** i surten els mateixos.

Si s'observa aquests atributs, es veu que no surt cap que estigui relacionat amb l'alumne abans de fer la matrícula: **SEXE, EDAT, VIA**. Tampoc sembla que importa el **NSEM** en el que es matricula ni el número d'assignatures o crèdits dels que es matricula (**NA, NC, NACMAT**).

Si que tenen importància el número d'assignatures / crèdits superats (**NASUP, NCSUP**) o els seus percentatges, que ve a dir el mateix (**PCTAS, PCTS**), el número d'assignatures a les que es presenta **NACPRE** i si supera o no les 7 primeres assignatures, que coincideix amb les 7 que més s'han matriculat (**A1S, A2S, A3S, A4S, A5S, A6, A7S**). Potser massa evident. El fet de que superi o no les assignatures a les que es presenta, segur que condiciona la continuïtat o no en els estudis d'un alumne.

Si es donés com a vàlid aquests atributs, weka permet desar un fitxer que contingui només aquests atributs. Es pot fer amb el botó dret a sobre del model de selecció que hem fet i s'agafa l'opció **Save reduced data** i demana un nom de fitxer.



Encara que aquests atributs són molt evidents, es pot provar algunes tècniques de mineria de dades per veure quin és el resultat. A priori el fet que el mateix weka proposi aquests atributs, fa pensar que es pot obtenir models força bons.

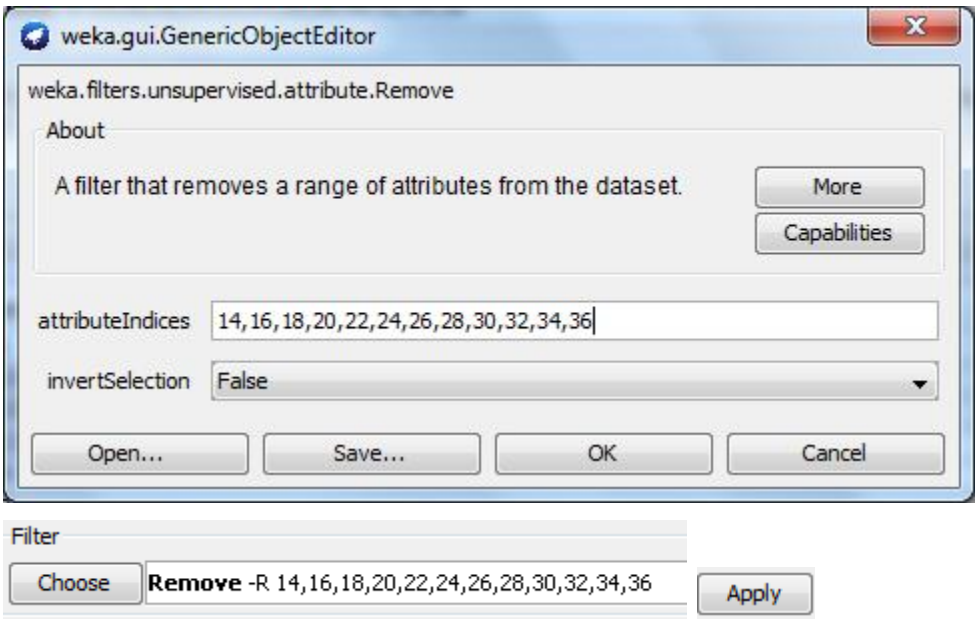
Per curiositat es pot tornar a provar amb aquests 12 atributs si és possible trobar menys atributs que siguin representatius del fet de l'abandonament dels estudis. Després de tornar a fer-ho, dona com a resultat els mateixos 12 atributs. No hi ha cap reducció.

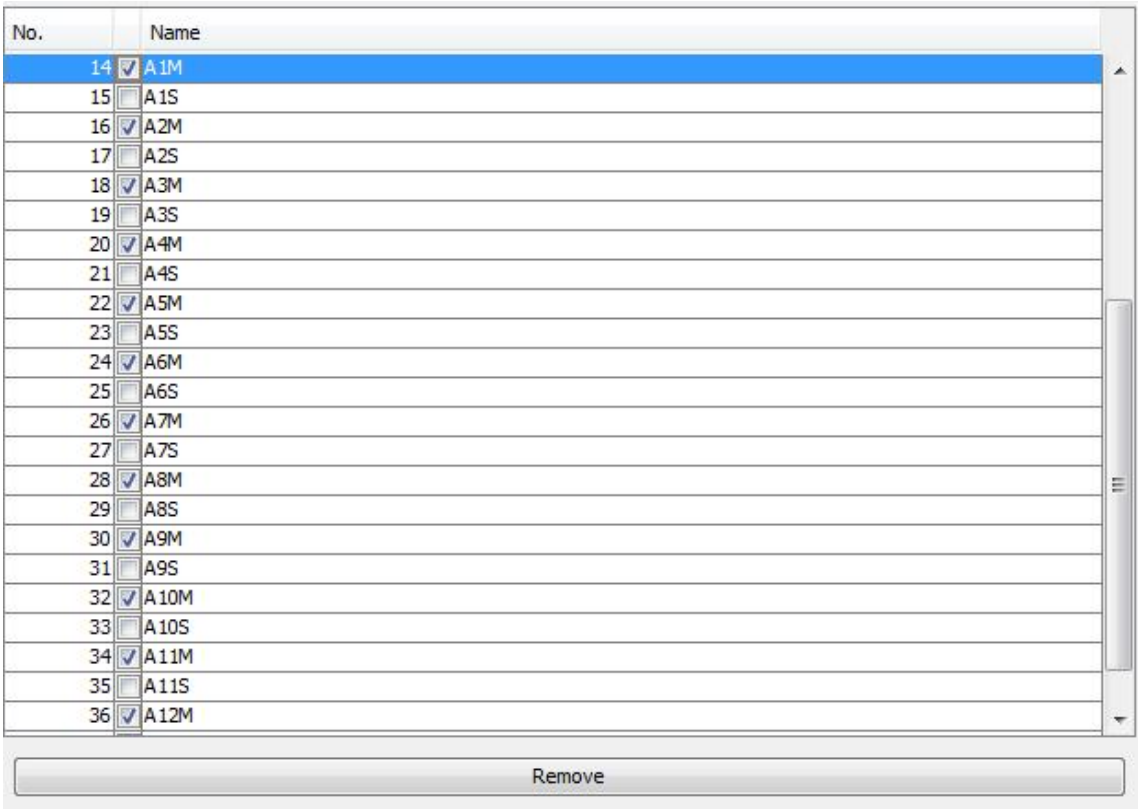
3.6. Reducció d'atributs redundants

Després d'aplicar diferents algorismes, es veu que potser seria millor que la dada que hi ha en els camps **A1S, A2S, ..., A12S**, que indica si un alumne no s'ha presentat o matriculat de les 12 assignatures més comunes en la matrícula del 1r semestre, no tingués el mateix valor: **NP_NM**. Donat que ens els camps **A1M, A2M, ..., A12M** ja indica si s'ha matriculat o no de l'assignatura, es pot tenir en compte per posar **NM** quan no s'ha matriculat i **NP** quan sí s'ha matriculat però no es presenta. S'ha procedit a modificar aquests valors des de l'excel i s'ha tornat a exportar en format csv i a llegir el fitxer des del weka.

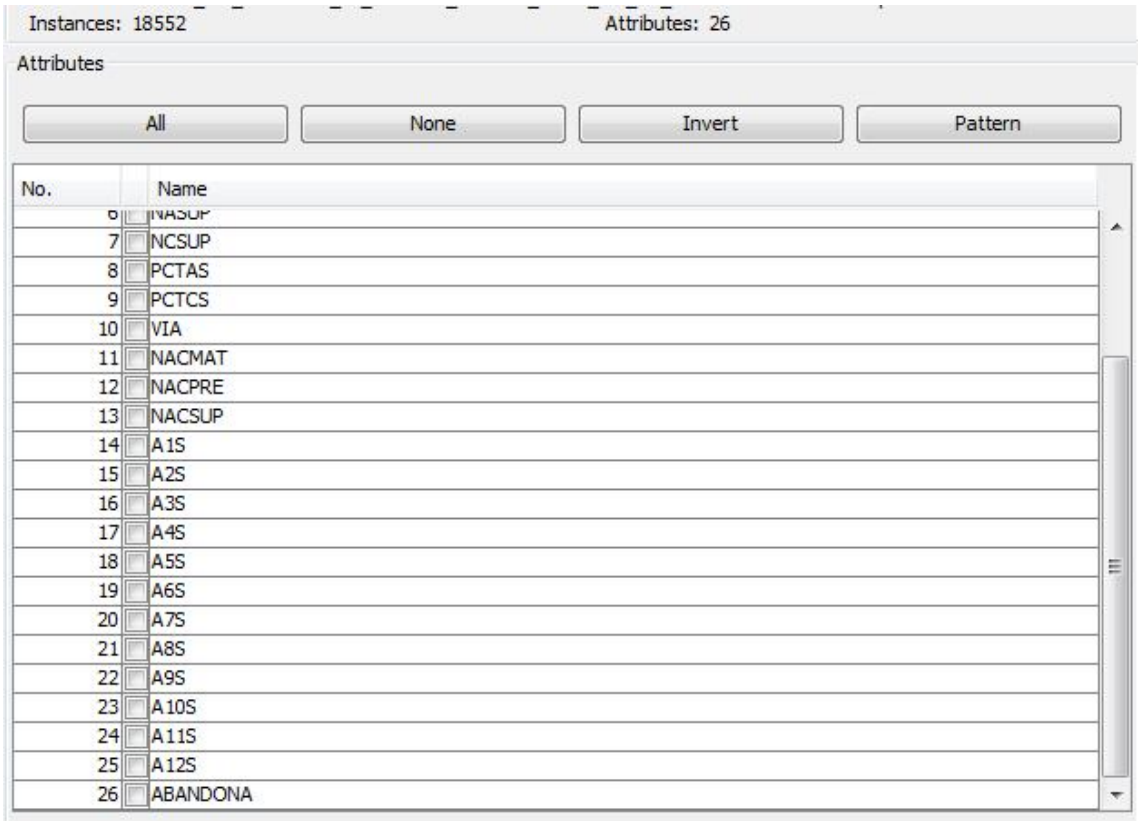
Al separar la informació de si un alumne no es matricula, no es presenta, aprova o no aprova, està redundant la informació que té els camps **A1M, A2M, ..., A12M** que només té els valors de **NO** es matricula (valor NM de **A1S, A2S, ..., A12S**) i **SÍ** es matricula (valors **SÍ, NO, NP** de **A1S, A2S, ..., A12S**). Per tant es pot eliminar els 12 atributs que indiquen si s'han matriculat o no de les 12 assignatures més freqüents del 1r semestre. Ara es passa a tenir unes dades amb 26 atributs.

Per esborrar aquests atributs, es pot fer amb el filtre **Remove** i indicant en **AttributeIndices** els valors corresponents als atributs a esborrar o també es pot fer seleccionant els atributs i clicant al botó **Remove**.





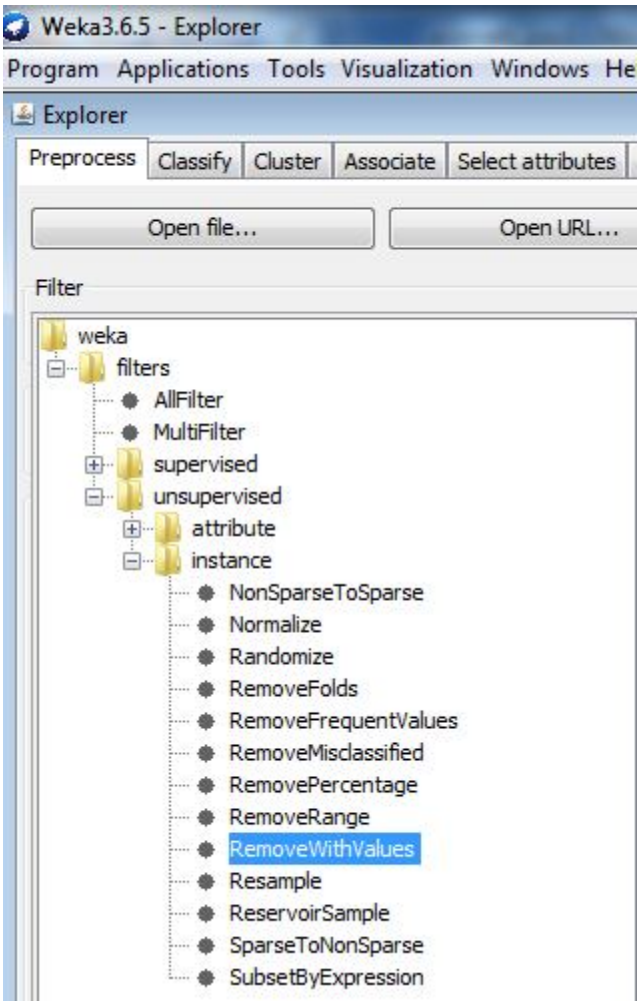
Es pot veure que ara només **hi ha 26 atributs**.

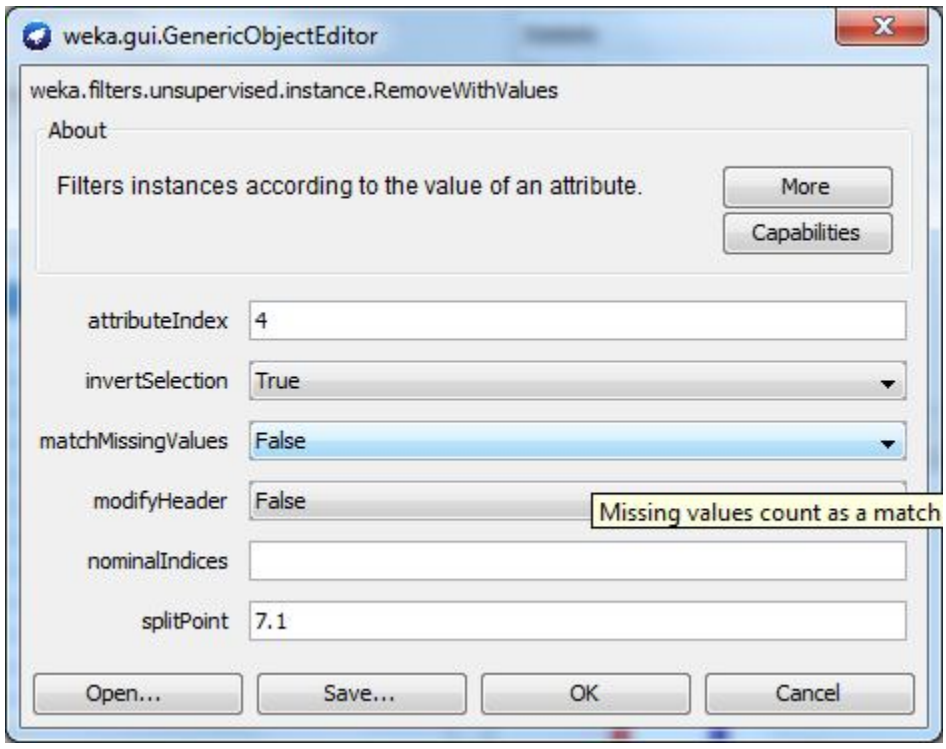


3.7. Reducció de més atributs i instàncies fora de rangs

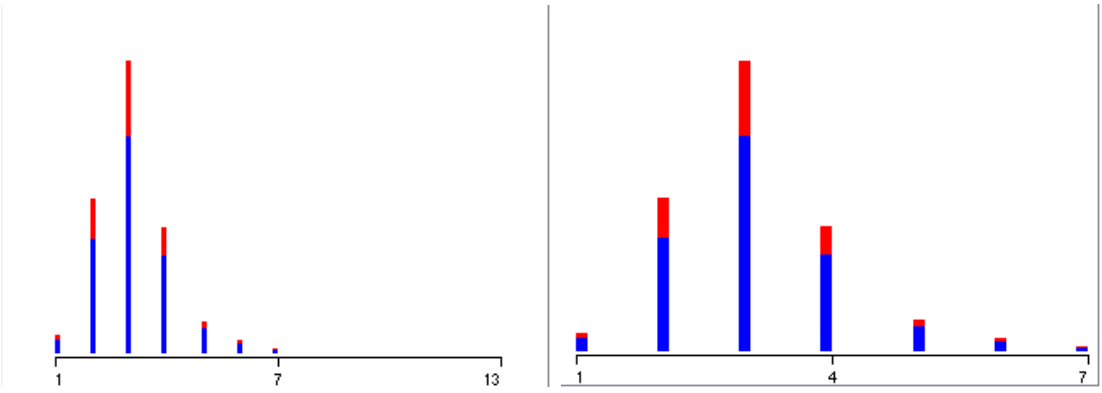
Es pot intentar reduir encara el número d'atributs i/o esborrar les dades d'alguns estudiants de que tinguin alguns valors fora dels rangs normals.

Es pot observar que en l'atribut **NA**, número d'assignatures matriculades, hi ha molt pocs estudiants que es matricules de més de 7 assignatures. Es pot eliminar les instàncies d'aquests estudiants amb el filtre de la pestanya **Preprocess** / **RemoveWithValues**. Donat que es volen els registres que en el camp **NA** hi ha valors de fins a 7 assignatures incloses, s'ha d'indicar en **attributeIndex** el valor 4 que correspon a l'instancia **NA**, en **splitPoint** el valor numèric a partir del que es vol el tall. Es posa **7,1** perquè inclogui el 7. Si es donés ara a **Apply**, s'obtindria els estudiants que s'han matricular de més de 7 assignatures. Justament es vol el contrari per tant s'ha de posar a **True** el valor de **inverSelection**.





Ara es pot veure com es passa de 18552 alumnes amb aquesta distribució d'assignatures matriculades a 18541, 11 menys que representa un percentatge molt baix del total.



Ara encara es podrien eliminar més atributs. Per exemple, donat que dintre de les dades hi ha els atributs relacionats amb les assignatures més comunes en la matrícula del 1r semestre i no hi ha les dades dels crèdits que corresponen a cadascuna, es pot eliminar els atributs **NC, NCSUP** i **PCTS** que estan relacionats amb els crèdits. Ara hi hauria **23 atributs**.

Amb aquestes noves restriccions, es poden repetir els algorismes que s'han realitzat anteriorment per veure si aporten regles més clares.

3.8. Nous atributs que relacionin el seguiment de l'itinerari proposat

Tenint en compte les 12 assignatures més cursades en el 1r semestre i com estan distribuïdes en la proposta d'orientació dels 2 primers semestres, s'han creat 3 atributs nous:

1rSEM → valor numèric del 0 al 1. Indica el percentatge d'assignatures matriculades del 1r semestre. S'ha comptat el número d'assignatures de la proposta del 1r semestre i s'ha dividit per 6.

2nSEM → valor numèric del 0 al 1. Indica el percentatge d'assignatures matriculades del 2n semestre. S'ha comptat el número d'assignatures de la proposta del 2n semestre i s'ha dividit per 6. (Donat que hi ha una assignatura del 2n semestre que no està en les 12 més matriculades, el màxim serà un 0,8)

SEGUEIX_PLA → Indica la proporció que segueix l'alumne del número d'assignatures matriculades respecte l'itinerari recomanat. S'ha dividit NACMAT entre NA.

3.9. Probabilitat ponderada de la matrícula de cada assignatura

S'ha procedit a realitzar un clustering de les 12 assignatures més matriculades en el primer semestre i s'ha calculat la probabilitat de cada assignatura matriculada a que pertany o no en el clúster generat. Aquests valors s'han substituït en lloc dels valors SI de cada assignatura matriculada.

S'han fet diferents mostres que van des de 2 a 10 clústers. Es posa l'exemple del resultat de 5 clústers. En l'annex (10.2) hi ha la resta d'exemples.

Full Data	0	1	2	3	4							
18552	5517	2734	4038	2091	4172	**Clústers**						
=====	=====	=====	=====	=====	=====	0	1	2	3	4	Probabilitat	
SI	SI	SI	NO	SI	SI	0,2974	0,1474	0,0000	0,1127	0,2249	**0,7823**	A1M
NO	NO	NO	NO	SI	NO	0,0000	0,0000	0,0000	0,1127	0,0000	**0,1127**	A2M
NO	NO	NO	NO	SI	NO	0,0000	0,0000	0,0000	0,1127	0,0000	**0,1127**	A3M
NO	SI	NO	NO	NO	NO	0,2974	0,0000	0,0000	0,0000	0,0000	**0,2974**	A4M
NO	NO	NO	SI	SI	NO	0,0000	0,0000	0,2177	0,1127	0,0000	**0,3304**	A5M
NO	NO	NO	NO	NO	NO	0,0000	0,0000	0,0000	0,0000	0,0000	**0,0000**	A6M
NO	NO	SI	NO	NO	NO	0,0000	0,1474	0,0000	0,0000	0,0000	**0,1474**	A7M
NO	NO	NO	NO	NO	NO	0,0000	0,0000	0,0000	0,0000	0,0000	**0,0000**	A8M
NO	NO	NO	NO	NO	NO	0,0000	0,0000	0,0000	0,0000	0,0000	**0,0000**	A9M
NO	NO	NO	NO	NO	NO	0,0000	0,0000	0,0000	0,0000	0,0000	**0,0000**	A10M
NO	NO	NO	NO	NO	NO	0,0000	0,0000	0,0000	0,0000	0,0000	**0,0000**	A11M
NO	NO	NO	NO	NO	NO	0,0000	0,0000	0,0000	0,0000	0,0000	**0,0000**	A12M

4. Models de mineria de dades aplicats.

L'aplicació d'algorismes de mineria de dades s'ha anat realitzant de forma paral·lela al capítol anterior, ja que ha mida que s'anaven experimentat els diferents models, sorgia la necessitat d'anar transformant atributs i/o crear-ne altres de nous.

Es pot dir que l'estratègia inicial ha estat de tipus exploratòria, sense tenir cap objectiu molt en concret a perseguir. El punt de partida era: tinc unes dades, com estan relacionades entre si? Aquest apartat ha servit per provar diferents algorismes, sobretot de classificació i classificació, a més d'intentar preveure quins atributs eren més importants a l'hora d'incidir en l'abandonament dels estudis. Aquesta experimentació esta desenvolupada a l'apartat 10.3 en l'annex.

Una vegada obtinguts resultats, encara dispersos i no concrets, s'ha fet un lliurament d'aquesta informació al client, en aquest cas el departament d'informàtica de la UOC que és qui ens ha lliurat les dades i vol investigar sobre les mateixes. El client és el que en té coneixement del negoci i ha fet comentaris i recomanacions sobre les primeres relacions que començaven a sortir en aquest estudi previ. D'aquesta reunió surt l'enfoc que s'aplica després en els apartats següents: fer primer un estudi sociodemogràfic dels alumnes que deixen els estudis, analitzar només l'informació de la matrícula sense tenir en compte els resultats, analitzar associacions d'assignatures en la mateixa matrícula, seguiment o no de l'itinerari recomanat per la UOC

Ara s'aplicaran diferents models que permetin trobar una justificació als objectius proposats. Tal com s'ha comentat prèviament, s'ha desenvolupat en 5 apartats.

- **Primer**: s'han aplicat algorismes de tots tipus per observar els resultats i anar valorant l'importància de determinats atributs. **Aquesta part és llarga i està posada en l'annex.**
- **Segon**: després de les observacions de l'apartat anterior, s'han aplicat algorismes d'associació només a les dades sociodemogràfiques
- **Tercer**: només en tindrà en compte la matrícula de les 12 assignatures més matriculades del 1r semestre, sense resultats de superació o no de l'assignatura.
- **Quart**: s'ha prescindit del tipus d'assignatura i s'han afegit els nous atributs **1rSEM**, **2nSEM** i **SEGUEIX_PLA** de l'apartat 3.8
- **Cinquè**: S'ha intentat millorar els resultats aplicant algorismes de classificació a les dades que recullen la probabilitat de matrícula de l'apartat 3.9

En aplicar els algorismes de clustering, associació i classificació amb el WEKA, cal ajustar alguns valors previs a l'aplicació de l'algorisme, els més bàsics són per exemple:

Clustering: Cal ajustar el número de clústers

Associació: Cal ajustar els valors de suport i de confiança mínim que han de tenir les regles

Classificació: En els arbres J48 es pot ajustar el factor de confiança que permet podar l'arbre

També cal fixar-se en alguns indicadors per valorar la qualitat del model generat.

Clúster

En el cas de que els clústers es basin en un atribut (en el nostre cas si abandonen o no) cal fixar-se en el percentatge d'instàncies incorrectament classificades.

Associate

Per les regles d'associació Apriori, cal anar ajustant a més del número de regles a cercar, el nivell de suport i el valor de confiança. El nivell de suport és la proporció que hi ha en les dades que compleixen la part de la regla de l'esquerra. El valor de confiança és la probabilitat de trobar la part dreta de la regla condicionada a que també estigui a la part esquerra.

Les regles surten ordenades de major a menor confiança.

Classify

Quan s'executa l'algorisme de classificació, s'obté el resultat corresponent i hi ha diferents indicadors que permeten valorar la qualitat del model generat. Per exemple:

```
Correctly Classified Instances       13725               73.9812 %
Incorrectly Classified Instances      4827               26.0188 %
Kappa statistic                          0.0013
Mean absolute error                      0.3761
Root mean squared error                  0.4394
Relative absolute error                 99.5841 %
Root relative squared error            101.1321 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.984     0.983      0.748     0.984     0.85       0.528    NO
                 0.017     0.016      0.264     0.017     0.031      0.528    SI
Weighted Avg.    0.74      0.739      0.625     0.74      0.643      0.528

=== Confusion Matrix ===

     a     b   <-- classified as
 13647   218 |     a = NO
  4609    78 |     b = SI
```

A més dels percentatges d'instàncies correctament i incorrectament classificades i l'error relatiu i d'altres, tenim aquests valors a tenir en compte:

TP rate: True Positive rate, és la proporció d'exemples classificats com una classe que realment són d'aquesta classe.

FP rate: False Positive rate, és la proporció d'exemples classificats com una classe que realment són d'un altre classe.

Precision: és la proporció dels elements que realment són d'una classe entre els que van ser classificats com aquesta classe.

També és molt visual veure la matriu de confusió que quan tots els elements estan correctament classificats, la diagonal coincideix amb el valor de les instàncies classificades i la resta són zeros.

Weka té 4 opcions per a poder entrenar els algorismes de classificació:

Use training test: Entrena el classificador amb tot el conjunt de dades i per fer servir el test també es fa servir el mateix conjunt de dades.

Supplied test set: S'entrena el classificador amb un conjunt de dades i es testeja amb un altre fitxer diferent.

Crossvalidation: Es divideix el conjunt de dades entre k parts, s'entrena amb k-1 parts i es testeja amb la que queda. Aquest procés es torna a repetir amb les k parts, fent una validació creuada. Per defecte K=10.

Aquesta és l'opció agafada per fer totes les proves de classificació.

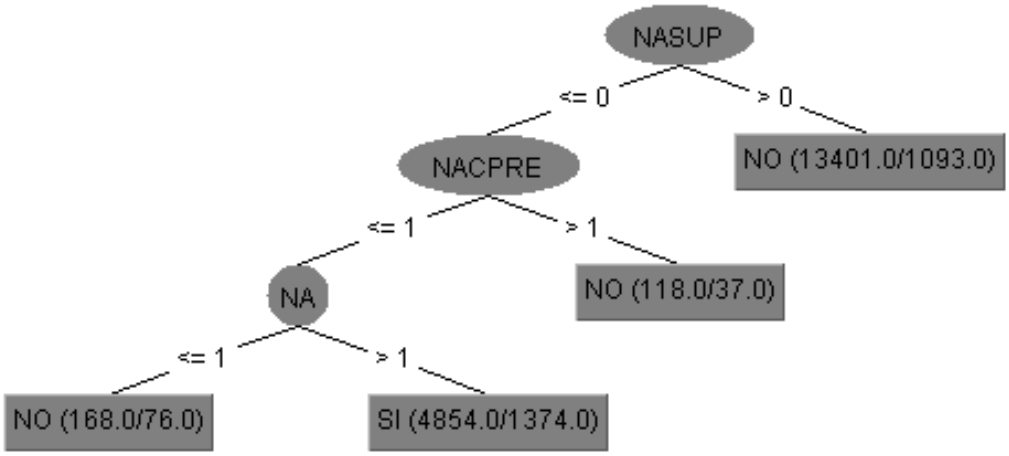
Porcentage split: Es divideix el conjunt de dades en un percentatge per entrenar i la resta per fer el test.

4.1. Mineria de dades amb dades de rendiment acadèmic

Estudi dels atributs relacionats amb els resultats al final del semestre de les assignatures matriculades. **A1S, A2S, A3S, A4S, A5S, A6S, A6S, A7S, A8S, A9S, A10S, A11S, A12S (+ABANDONA).** També es pot afegir **NASUP, NCSUP, PCTAS, PCTCS, NACPRE, NACSUP**.

S'han aplicat diferents algorismes per fer un primer estudi exploratiu que permeti observar l'importància de determinats atributs i començar a esbrinar possibles casuístiques que portin a l'abandonament dels estudis.

Gairebé totes les regles de classificació obtingudes condicionen el resultat a la superació d'assignatures obtenint regles de classificació com aquestes:



Com es pot veure en aquest arbre, és massa evident que si un alumne no supera cap assignatura, té molts números per abandonar els estudis.

Totes les proves estan a l'annex. Apartat 10.3

4.2. Mineria de dades amb dades sociodemogràfiques

Aquí s'ha aplicat només regles d'associació amb les dades sociodemogràfiques ja que les regles de classificació no ha classificat cap dels estudiants que **SI** abandonen. Cal dir que en les dades que disposem només es tenen els atributs **SEXE**, **FRANJA** i **VIA**.

Les regles d'associació obtingudes amb un nivell de confiança mínima del 0,6, només relacionen els estudiants que **NO** abandonen. Llavors es pot treure la conclusió que les dades sociodemogràfiques que disposem no influeixen en l'abandonament dels estudis. Caldria plantejar si es podrien afegir altres dades que sí poguessin influir: treballa, dedicació a la feina i a l'estudi, hobbies que puguin restar temps.

Associació / Apriori

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.1 -S -1.0
-c -1

Instances:    18552
Attributes:   4
              SEXE
              FRANJA
              VIA
              ABANDONA
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (1855 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 18

Best rules found:

  1. SEXE=DONA VIA=EST_IN 3214 ==> ABANDONA=NO 2511    conf:(0.78)
  2. SEXE=DONA FRANJA=MENYS=24 2446 ==> ABANDONA=NO 1878    conf:(0.77)
  3. FRANJA=MENYS=36 3968 ==> ABANDONA=NO 3044    conf:(0.77)
  4. SEXE=DONA VIA=NO_COU 2568 ==> ABANDONA=NO 1958    conf:(0.76)
  5. SEXE=DONA 9200 ==> ABANDONA=NO 7008    conf:(0.76)
  6. VIA=EST_IN 7609 ==> ABANDONA=NO 5774    conf:(0.76)
  7. VIA=NO_COU 3787 ==> ABANDONA=NO 2866    conf:(0.76)
  8. FRANJA=MENYS=24 4310 ==> ABANDONA=NO 3231    conf:(0.75)
  9. VIA=COU 3111 ==> ABANDONA=NO 2326    conf:(0.75)
 10. SEXE=HOME VIA=EST_IN 4395 ==> ABANDONA=NO 3263    conf:(0.74)
 11. FRANJA=MENYS=30 3483 ==> ABANDONA=NO 2556    conf:(0.73)
 12. SEXE=HOME 9352 ==> ABANDONA=NO 6857    conf:(0.73)
 13. FRANJA=MENYS=27 4393 ==> ABANDONA=NO 3209    conf:(0.73)
 14. VIA=TITULAT 4045 ==> ABANDONA=NO 2899    conf:(0.72)
 15. VIA=NO_COU ABANDONA=NO 2866 ==> SEXE=DONA 1958    conf:(0.68)
 16. VIA=NO_COU 3787 ==> SEXE=DONA 2568    conf:(0.68)
```

Baixant encara més el nivell de suport, surten més regles però va relacionant els atributs **SEXE**, **FRANJA**, **VIA** i **ABANDONA** amb les diferents possibilitats de combinacions que té.

També s'ha fet una clusterització amb 2 clústers per veure com agrupa els estudiants amb els que si i no abandonen.

Clusterització / KMeans

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last"
-I 500 -S 10

Instances:18552
Attributes:4
              SEXE
              FRANJA
              VIA
Ignored:
              ABANDONA
Test mode:Classes to clusters evaluation on training data
=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 24933.0
Missing values globally replaced with mean/mode

Clúster centroids:
                         Clúster#
Attribute    Full Data          0          1
               (18552)    (11236)     (7316)
=============================================
SEXE               HOME       DONA       HOME
FRANJA         MENYS=27   MENYS=27   MENYS=36
VIA              EST_IN     EST_IN     EST_IN

Clustered Instances

0      11236 ( 61%)
1       7316 ( 39%)

Class attribute: ABANDONA
Classes to Clusters:

    0    1  <-- assigned to cluster
 8464 5401 | NO
 2772 1915 | SI

Clúster 0 <-- NO
Clúster 1 <-- SI

Incorrectly clustered instances :    8173.0   44.0545 %
```

La clusterització no posa correctament un 44% d'estudiants. Si només es fan 2 clusters es defineix cadascun de la següent manera:

NO abandonen → Dona, de menys de 28 i més de 24 anys amb estudis inacabats.

SI abandonen → Home, de menys de 37 i més de 30 anys amb estudis inacabats.

4.3. Mineria de dades amb dades només assignatures matriculades

En aquest cas es cerquen regles d'associació entre les 12 assignatures de matrícula més comuna en el 1r semestre però només dels alumnes que SI abandonen. Per fer això es fa un filtre **RemoveWithValues** i s'esborren les instàncies dels alumnes que NO abandonen. Una vegada fet això, s'esborra l'atribut ABANDONA ja que en totes les instàncies té el mateix valor.

Aplicant l'algorisme d'associació **Apriori** amb 500 regles es pot veure que surten moltes combinacions de l'assignatura matriculada A1 amb la no matrícula de diferents combinacions d'altres assignatures. Concretament es pot veure la regla 26 que diu:

A1M=SI 3633 ==> ABANDONA=SI 3633 conf:(1)

O sigui dels 4687 estudiants que SI abandonen els estudis, 3633 s'havien matriculat d'aquesta assignatura. S'ha de tenir present que aquesta assignatura, **Multimèdia i comunicació**, és la que té més matrícula.

Per fer altre tipus d'observació, es torna a fer regles d'associació només dels estudiants que SI deixen els estudis però separant les assignatures en dos blocs: el primer bloc de les assignatures proposades segons l'itinerari recomanat en el 1r semestre i l'altre bloc per les recomanades en el 2n semestre. Ara surten altres combinacions d'assignatures:

Assignatures orientació 1r semestre: A1M, A2M, A3M, A4M, A6M i A8M

A1M=SI 3633 ==> ABANDONA=SI 3633 conf:(1) 14225 estudiants matriculats

A3M=SI 1552 ==> ABANDONA=SI 1552 conf:(1) 6646 estudiants matriculats

A4M=SI 1517 ==> ABANDONA=SI 1517 conf:(1) 5882 estudiants matriculats

De totes les regles es pot veure que aquestes 3 assignatures surten matriculades en els alumnes que SI abandonen:

- A1M Multimèdia i comunicació
- A3M Introducció al dret
- A4M Introducció a la macroeconomia

A1M=SI A3M=SI 1171 ==> A8M=NO 1022 conf:(0.87)

A1M=SI A4M=SI 1150 ==> A8M=NO 968 conf:(0.84)

A1M=SI A2M=SI 840 ==> A8M=NO 699 conf:(0.83)

A1M=SI A6M=SI 683 ==> A8M=NO 542 conf:(0.79)

A3M=SI A4M=SI 618 ==> A8M=NO 488 conf:(0.79)

A6M=SI 961 ==> A8M=NO 742 conf:(0.77) 3850 estudiants matriculats

A2M=SI 1073 ==> A6M=NO 784 conf:(0.73) 4128 estudiants matriculats

Assignatures orientació 2n semestre: A5M, A7M, A9M, A10M, A11M i A12M

A5M=SI 1541 ==> ABANDONA=SI 1541 conf:(1) 5934 estudiants matriculats

A7M=SI 1009 ==> A10M=NO 970 conf:(0.96) 3927 estudiants matriculats

- A5M Introducció a la comptabilitat
- A7M Organització i administració d'empreses I

Amb aquestes regles es poden veure assignatures i combinacions d'assignatures que es donen entre els estudiants que abandonen els estudis. Però aquesta informació no porta a poder afirmar que tots els que combinin en la matrícula abandonaran els estudis, ja que també han sortit regles amb les mateixes combinacions que no abandonen els estudis.

4.3.1 Arbres classificadors

Per a poder visualitzar les regles de classificació, el que va molt bé són els arbres de classificació. El problema és que en les proves que s'han realitzat no classifica correctament els alumnes que SI abandonen els estudis i en el cas dels alumnes que no abandonen els classifica correctament però amb un número molt alt de fals positius.

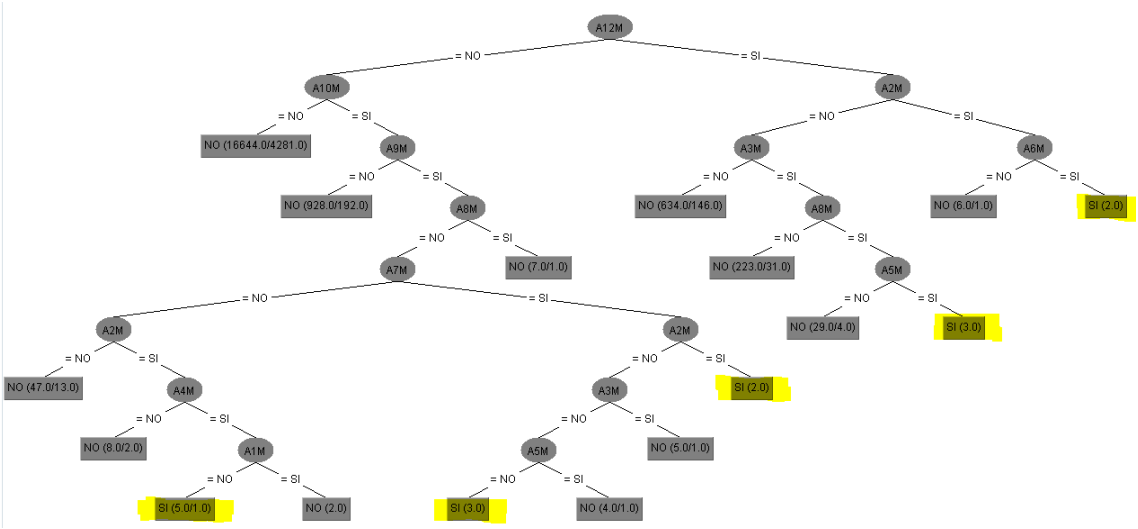
Classify / trees / J48

L'arbre de classificació J48 permet fer podes. Quan es té com a resultat un arbre amb molts nodes i moltes fulles, aquest arbre es pot tornar difícil de llegir i interpretar. Aplicant l'algorisme J48 sobre les dades de matrícula de les 12 assignatures i deixant el factor de confiança que té com a defecte el valor 0,25 crea un arbre d'un sol node. Si es puja el factor de confiança al màxim (1) s'obté una arbre amb 127 nodes i una mida de l'arbre de 273 elements.

Per fer un arbre que pugui ser llegible, s'ha baixat el factor de confiança a 0,47 i s'obté el següent arbre:

```
Number of Leaves  :    17

Size of the tree :    33


Time taken to build model: 0.27seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances       13848               74.6442 %
Incorrectly Classified Instances      4704               25.3558 %
Kappa statistic                         -0.001
Mean absolute error                      0.3775
Root mean squared error                  0.4351
Relative absolute error                 99.9756 %
Root relative squared error            100.1315 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===
```

```
                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                  0.998     0.999      0.747     0.998     0.855      0.504    NO
                  0.001     0.002      0.16      0.001     0.002      0.504    SI
Weighted Avg.     0.746     0.747      0.599     0.746     0.639      0.504

=== Confusion Matrix ===
     a     b   <-- classified as
 13844    21 |     a = NO
  4683     4 |     b = SI
```



D'aquest arbre, si s'observa només els fulls dels que SI abandonen, es podrien crear les següents regles de classificació:

- **Si es matricula** de les assignatures A10M + A9M + A2M + A4M i **no es matricula** de les assignatures A12M + A8M + A7M + A1M l'estudiant SI abandona els estudis. (5 alumnes classificats, 1 malament)
- **Si es matricula** de les assignatures A10M + A9M + A7M i **no es matricula** de les assignatures A12M + A8M + A2M + A3M i A5M l'estudiant SI abandona els estudis. (3 estudiants)
- **Si es matricula** de les assignatures A10M + A9M + A7M + A2M i **no es matricula** de les assignatures A12M i A8M l'estudiant SI abandona els estudis. (2 estudiants)
- **Si es matricula** de les assignatures A12M + A3M + A8M + A5M i **no es matricula** de l'assignatura A2M l'estudiant SI abandona els estudis. (3 estudiants)
- **Si es matricula** de les assignatures A12M + A2M i A6M l'estudiant SI abandona els estudis. (2 estudiants)

També es podrien crear regles de classificació dels alumnes que NO abandonen, el problema és que encara que classifica gairebé tots aquests estudiants, té un índex de fals positiu (FP_rate) tant alt com el de verdader positiu (TP_rate)

Aplicant diferents algorismes no s'aconsegueix pujar del 75% d'instàncies correctament classificades.

4.4. Mineria de dades amb nous atributs d'itinerari recomanat

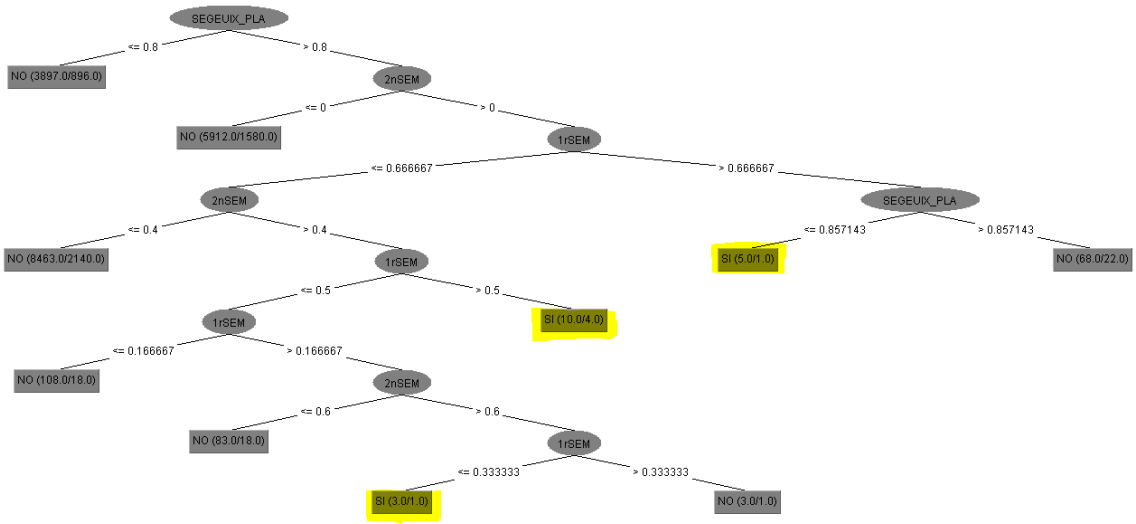
Tal com s'explica en l'apartat 3.8, s'han creat uns nous atributs que recullen el percentatge de seguiment de la proposta d'itinerari recomanat del 1r i 2n semestre del pla d'estudis (**1rSEM** I **2nSEM**) així com el percentatge de les assignatures matriculades respecte a les 12 més matriculades (**SEGUEIX_PLA**).

S'aplica directament un arbre de classificació J48.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 1.0 -M 2

Instances:18552
Attributes:4
              1rSEM
              2nSEM
              SEGEUIX_PLA
              ABANDONA
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

SEGEUIX_PLA <= 0.8: NO (3897.0/896.0)
SEGEUIX_PLA > 0.8
|   2nSEM <= 0: NO (5912.0/1580.0)
|   2nSEM > 0
|   |   1rSEM <= 0.666667
|   |   |   2nSEM <= 0.4: NO (8463.0/2140.0)
|   |   |   2nSEM > 0.4
|   |   |   |   1rSEM <= 0.5
|   |   |   |   |   1rSEM <= 0.166667: NO (108.0/18.0)
|   |   |   |   |   1rSEM > 0.166667
|   |   |   |   |   |   2nSEM <= 0.6: NO (83.0/18.0)
|   |   |   |   |   |   2nSEM > 0.6
|   |   |   |   |   |   |   1rSEM <= 0.333333: SI (3.0/1.0)
|   |   |   |   |   |   |   1rSEM > 0.333333: NO (3.0/1.0)
|   |   |   |   1rSEM > 0.5: SI (10.0/4.0)
|   |   1rSEM > 0.666667
|   |   |   SEGEUIX_PLA <= 0.857143: SI (5.0/1.0)
|   |   |   SEGEUIX_PLA > 0.857143: NO (68.0/22.0)

Number of Leaves  :   10

Size of the tree :    19


Time taken to build model: 0.39seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13863               74.7251 %
Incorrectly Classified Instances      4689               25.2749 %
Kappa statistic                          0.001
Mean absolute error                      0.3774
Root mean squared error                  0.4346
Relative absolute error                 99.9297 %
Root relative squared error            100.0219 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.999      0.747     0.999     0.855      0.506    NO
                 0.001     0.001      0.429     0.001     0.003      0.506    SI
```

```
Weighted Avg.    0.747     0.747      0.667     0.747     0.64       0.506

=== Confusion Matrix ===

     a     b   <-- classified as
 13857     8 |     a = NO
  4681     6 |     b = SI
```

Si el número d'assignatures matriculades és més del 80% de les 12 més comunes, d'elles hi ha entre el 40% i el 60% del 2n semestre i entre el 33,3% i el 50% del 1r semestre, els estudiants SI abandonen.



D'aquest arbre es poden treure les següents regles:

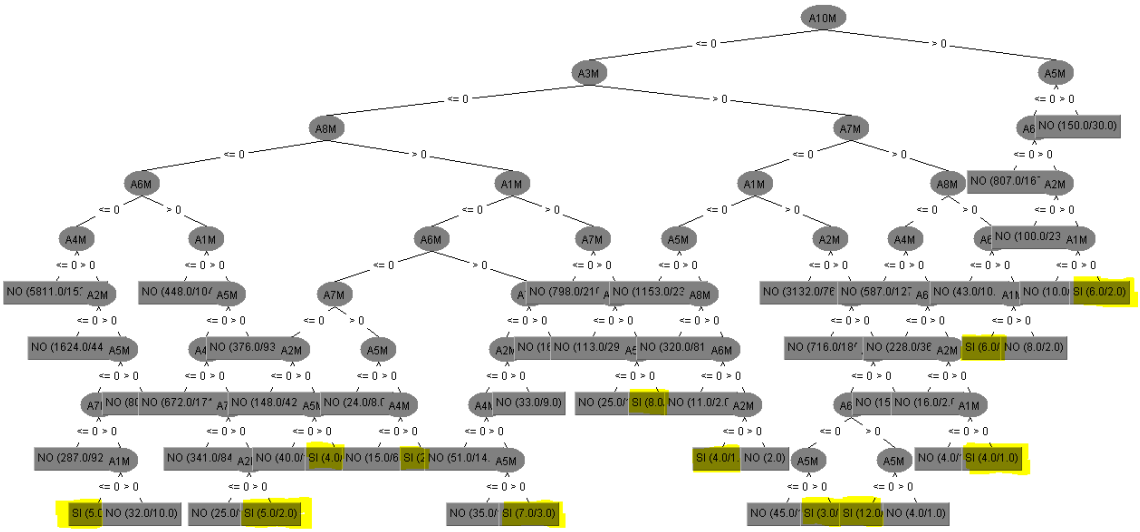
- Si el número d'assignatures matriculades està entre 80% i el 85,7% de les 12 més comunes, d'elles hi ha més d'una del 2n semestre i més del 66% del 1r semestre, els estudiants SI abandonen.
- Si el número d'assignatures matriculades és més del 80% de les 12 més comunes, d'elles hi ha més del 40% del 2n semestre i més del 50% del 1r semestre, els estudiants SI abandonen.
- Si el número d'assignatures matriculades és més del 80% de les 12 més comunes, d'elles hi ha entre el 40% i el 60% del 2n semestre i entre el 33,3% i el 50% del 1r semestre, els estudiants SI abandonen.

4.5. Mineria de dades amb dades probabilitat matrícula assignatures

A l'apartat 3.9 s'ha fet un clustering amb les dades de les 12 assignatures més matriculades en el primer semestre. Amb les agrupacions fetes s'ha calculat la probabilitat que té cada assignatura de pertànyer o no en un clúster. Ara en lloc de tractar l'informació de si està o no matriculat en cada assignatura, es posarà el valor de la probabilitat total de cada assignatura de pertànyer en els diferents clústers creats. Amb aquesta informació directament es farà un arbre de classificació per la facilitat d'interpretació de les seves regles.

Els resultats que es mostren són els generats quan es fa la clusterització amb 10 clústers.

```
Number of Leaves  :   52

Size of the tree :    103


Time taken to build model: 2.15seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13826               74.5257 %
Incorrectly Classified Instances      4726               25.4743 %
Kappa statistic                          0.0004
Mean absolute error                      0.377
Root mean squared error                  0.436
Relative absolute error                 99.8324 %
Root relative squared error            100.3315 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.996     0.995      0.747     0.996     0.854      0.515    NO
                 0.005     0.004      0.265     0.005     0.009      0.515    SI
Weighted Avg.    0.745     0.745      0.626     0.745     0.64       0.515

=== Confusion Matrix ===

     a     b   <-- classified as
 13804    61 |     a = NO
  4665    22 |     b = SI
```



Les regles de classificació d'aquest arbre són les següents:

- **Si es matricula** de les assignatures A1M + A2M + A4M + A7M i **no es matricula** de les assignatures A5M + A6M + A8M + A3M i A10M, l'estudiant SI abandona els estudis. (5 alumnes classificats, 1 malament)
- **Si es matricula** de les assignatures A2M + A7M + A4M + A1M + A6M i **no es matricula** de les assignatures A5M + A8M + A3M i A10M, l'estudiant SI abandona els estudis. (5 alumnes classificats, 2 malament)
- **Si es matricula** de les **assignatures** A5M + A2M + A8M i **no es matricula** de les assignatures A7M + A6M + A1M + A3M i A10M, l'estudiant SI abandona els estudis. (4 alumnes classificats, 1 malament)
- **Si es matricula** de les assignatures A5M + A4M + A7M + A8M i **no es matricula** de les assignatures A6M + A1M + A3M i A10M, l'estudiant SI abandona els estudis. (2 alumnes classificats)
- **Si es matricula** de les assignatures A5M + A4M + A6M + A8M i **no es matricula** de les assignatures A2M + A7M + A1M + A3M i A10M, l'estudiant SI abandona els estudis. (7 alumnes classificats, 3 malament)
- **Si es matricula** de les assignatures A5M + A2M + A7M + A1M + A8M i **no es matricula** de les assignatures A3M i A10M, l'estudiant SI abandona els estudis. (8 alumnes classificats, 3 malament)
- **Si es matricula** de les assignatures A6M + A8M + A5M + A3M i **no es matricula** de les assignatures A2M + A1M + A7M i A10M, l'estudiant SI abandona els estudis. (4 alumnes classificats, 1 malament)
- **Si es matricula** de les assignatures A5M + A8M + A2M + A1M + A3M i **no es matricula** de les assignatures A6M + A4M + A7M i A10M, l'estudiant SI abandona els estudis. (3 alumnes classificats, 1 malament)
- **Si es matricula** de les assignatures A6M + A8M + A2M + A1M + A3M i **no es matricula** de les assignatures A5M + A4M + A7M i A10M, l'estudiant SI abandona els estudis. (12 alumnes classificats, 5 malament)

- **Si es matricula** de les assignatures A1M + A2M + A6M + A4M + A7M + A3M i **no es matricula** de les assignatures A8M i A10M, l'estudiant SI abandona els estudis. (4 alumnes classificats, 1 malament)
- **Si es matricula** de les assignatures A6M + A8M + A7M + A3M i **no es matricula** de les assignatures A1M i A10M, l'estudiant SI abandona els estudis. (6 alumnes classificats, 2 malament)
- **Si es matricula** de les assignatures A1M + A2M + A6M + A10M i **no es matricula** de l'assignatura A5M, l'estudiant SI abandona els estudis. (6 alumnes classificats, 2 malament)

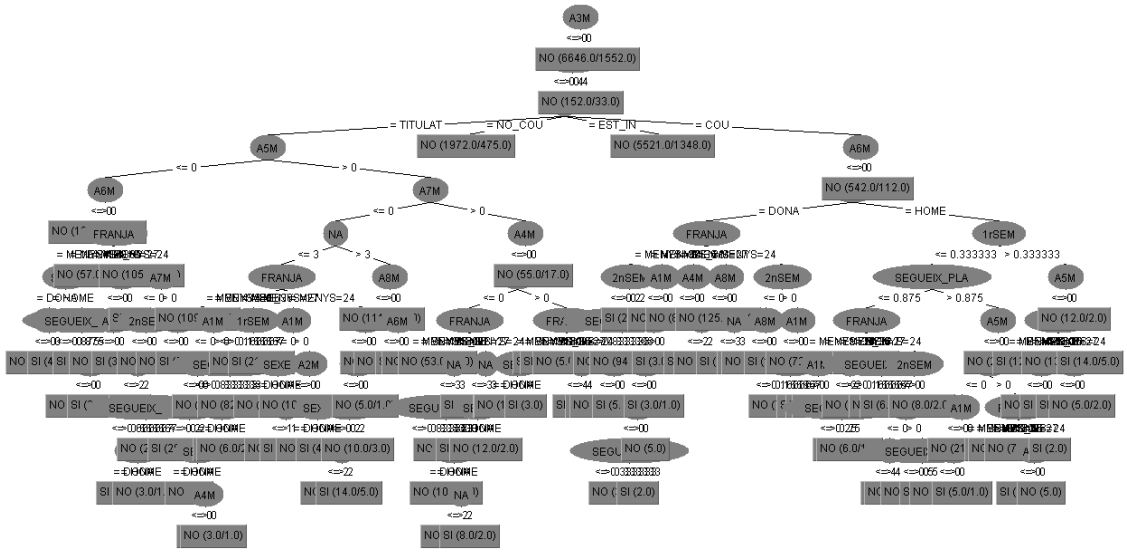
Ara per finalitzar he afegit a aquest valors, els camps de dades sociodemogràfiques i els del seguiment de l'itinerari recomanat. Posant el factor de confiança al màxim, baixa una mica el percentatge de classificació correcta però classifica correctament a 622 estudiants. L'arbre que dona però, és molt gran.

```
Number of Leaves  :   1398

Size of the tree :    2449


Time taken to build model: 39.23seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12880               69.4265 %
Incorrectly Classified Instances      5672               30.5735 %
Kappa statistic                          0.0203
Mean absolute error                      0.3752
Root mean squared error                  0.4764
Relative absolute error                 99.3501 %
Root relative squared error            109.636  %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.884     0.867      0.751     0.884     0.812      0.524    NO
                 0.133     0.116      0.279     0.133     0.18       0.524    SI
Weighted Avg.    0.694     0.677      0.632     0.694     0.652      0.524

=== Confusion Matrix ===

     a     b   <-- classified as
 12258  1607 |     a = NO
  4065   622 |     b = SI
```

Per fer un arbre més petit, amb poda, el màxim que s'ha pogut reduir és amb un factor de confiança de 0,38 .

```
Number of Leaves  :    119

Size of the tree :    211


Time taken to build model: 5.04seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13764               74.1915 %
Incorrectly Classified Instances      4788               25.8085 %
Kappa statistic                          0.0009
Mean absolute error                      0.3777
Root mean squared error                  0.4384
Relative absolute error                100.003  %
Root relative squared error            100.8922 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.989     0.988      0.747     0.989     0.851      0.513    NO
                 0.012     0.011      0.263     0.012     0.023      0.513    SI
Weighted Avg.    0.742     0.741      0.625     0.742     0.642      0.513

=== Confusion Matrix ===

     a     b   <-- classified as
 13708   157 |     a = NO
  4631    56 |     b = SI
```

Encara que no es veu gairebé, es pot donar una idea de com està estructurat.



5. Conclusions

5.1. Conclusions generals

Després de fer totes aquestes proves amb diferents models per observar el tipus d'informació que donava cada model de mineria de dades.

Els models de clusterització donen informació de grups (clústers) a tenir en compte però que no aporten informació revelant que pugui permetre predir el resultat futur d'un estudiant sobre la continuïtat o no en els seus estudis.

Els models d'associació han donat associacions d'atributs quan els hem restringit a un tipus d'instàncies determinades, per exemple dintre dels estudiants que han abandonat els estudis.

Els models que més informació poden donar per predir situacions futures són els models de **classificació**, ja que mitjançant, per exemple d'arbres de decisió, es poden anar generant regles que indiquen el camí que han recorregut per arribar a una determinada situació, com és el cas de veure els alumnes que abandonen els estudis. El problema és que amb les dades que es disposaven i combinacions que s'ha fet de les mateixes, no s'ha arribat a un percentatge de classificació superior al 75%. Si que s'han arribat a percentatges del 86% però quan es tenien en compte el resultat acadèmic de l'estudiant i per tant era difícilment traslladable al comportament d'un estudiant a l'hora de fer la matrícula.

5.2. Conclusions específiques

Amb el treball realitzat aplicant diferents models de mineria de dades, es pot treure les següents conclusions:

- Les dades sociodemogràfiques aportades sembla **no** incideixen en l'abandonament dels estudis.
- Caldria recollir més informació de l'estudiant prèvia a la matrícula per ampliar la informació sociodemogràfica: treballa (nº d'hores), dedicació setmanal a l'estudi i altres dades que puguin influir en el seguiment dels estudis.
- De les regles aportades en l'estudi del seguiment de l'itinerari (4.4) es pot veure que els estudiants poden tenir problemes quan agafen més de la meitat d'assignatures recomanades en el 1r semestre o bé si agafen menys però es matriculen d'aproximadament la meitat de les recomanades en el 2n semestre.

- Encara que el percentatge de classificació no ha estat molt alt, si fem cas dels resultats de classificació tenint en compte la probabilitat de matrícula d'assignatures (4.5), cal vigilar que l'alumne no es matriculi en el mateix semestre de la següent combinació d'assignatures: (en blau assignatures de 1r trimestre i en vermell de 2n semestre)
    - A1M + A2M + A4M + A7M
    - A1M + A2M + A4M + A6M + A7M
    - A2M + A5M + A8M
    - A4M + A5M + A7M + A8M
    - A4M + A5M + A6M + A8M
    - A1M + A2M + A5M + A7M + A8M
    - A3M + A5M + A6M + A8M
    - A1M + A2M + A3M + A5M + A8M
    - A1M + A2M + A3M + A6M + A8M
    - A1M + A2M + A3M + A4M + A6M + A7M
    - A3M + A6M + A7M + A8M
    - A1M + A2M + A6M + A10M
- Per evitar aquestes combinacions i a més fer una proposta més realista, caldria fer un itinerari recomanat de 3 assignatures per semestre que coincideix amb la mitjana de tots els estudiants.
- En finalitzar cada semestre cal recollir les dades de matrícula i resultats per tal d'incloure'ls en el projecte.

6. Aplicabilitat del model

Les regles de combinació d'assignatures no recomanades per realitzar-les en el mateix semestre de la primera matrícula d'un estudiant de la diplomatura de Ciències Empresarials de la UOC s'han d'introduir en una base de dades disponible en el moment de formalitzar la proposta de matrícula.

Al anar seleccionant les assignatures escollides per l'estudiant per incorporar-les en la proposta de matrícula, l'aplicatiu anirà indicat la viabilitat de la seva proposta, d'aquesta manera l'estudiant serà conscient del percentatge previst sobre la continuïtat dels seus estudis amb la combinació d'assignatures de la matrícula actual.

Donat que l'estudiant pot fer cas o no de les recomanacions de l'aplicatiu i posterior revisió / recomanació del tutor, LA UOC una vegada coneguda la matrícula de cada semestre, crearia estratègies per intentar esmenar els perills d'aquells alumnes que segons el model tenen possibilitats d'abandonar els seus estudis. Aquestes poden ser, per exemple, un seguiment automatitzat al llarg del semestre d'aquelles assignatures que tenen un pes més important en el possible abandonament dels estudis.

7. Gestió de cicle de vida del model

Aquest model ha estat creat amb les dades que s'han proporcionat des de la UOC i que comprenen els 20 semestres que van des de l'any 1998 fins el 2008. S'ha de tenir present que el pla d'estudis de la diplomatura de Ciències Empresarials actualment està en extinció. Finalitzen els exàmens extraordinaris a l'any 2014 i per tant ara no es tenen dades de primera matrícula. Si el pla d'estudis estigués vigent, caldria anar incorporant les noves dades que es vagin generant i tornar a aplicar el model per veure si continua sent vàlid. Si amb les dades actuals es produís una desviació, caldria generar altre model que compleixi amb els objectius marcats.

L'experiència d'aquest estudi es podria traslladar als actuals estudis de grau, per exemple, d'Administració i direcció d'empreses que imparteix la UOC . Es pot partir observant assignatures equivalents que previsiblement poden tenir el mateix comportament i integrar la resta més comunes en la primera matrícula.

També fora bo anar observant el comportament dels estudiants de la UOC que utilitzen actualment eines de connexió a la xarxa fora d'un espai fix com pot ser la feina o casa seva. Això introdueix un nou element a tenir en compte que permet a un estudiant estar al dia del que succeeix a la seva aula virtual en un espai que no convida a treballar de forma regular i pot afegir més dispersió en la feina diària i constant que exigeixen uns estudis a distància.

8. Línies de treball futures i accions a fer

Al llarg del treball s'ha detectat els següents problemes que caldria solucionar:

Les dades sociodemogràfiques segur que influeixen en el seguiment correcte d'uns estudis tant presencials com a distància. Caldria investigar quines dades es poden obtenir dels estudiants que s'incorporen a uns estudis que reflecteixen el seu grau de compromís i que es podrien afegir a aquest estudi.

Els itineraris del pla d'estudis de la diplomatura de Ciències empresarials estan distribuïts en semestres tenint cadascun d'ells una mitjana de 6 assignatures. Donat que en l'estudi realitzat s'han trobat aparellaments d'assignatures que poden influir de forma negativa en el seguiment dels estudis i observant que en les dades aportades la mitjana d'assignatures matriculades en el primer semestre és de tres, es podria confeccionar un itinerari recomanat que reflecteixi la realitat de la matrícula dels estudiants de la UOC . Es a dir que en cada semestre realment es fessin la meitat de la proposta actual i per tant s'evitessin els aparellaments no desitjats.

Aplicant diferents algorismes de classificació no s'ha pogut arribar a percentatges de classificació correctes superiors al 75%. Caldria fer més combinacions d'atributs per tal que es permeti millorar aquest percentatge.

9. Bibliografia / webgrafia

Manual de Weka. Diego García Morate
http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf

Técnicas de análisis de datos. Aplicaciones prácticas usando Microsoft Excel y Weka. José Manuel Molina López y Jesús García Herrero
http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf

Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de Weka (Waikato Environment for Knowledge Analysis). Paula Andrea Vizcaino Garzon.
http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf

Materials Mineria de dades UOC

Preparació de dades. Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165719

Classificació: Arbres de decisió. Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165720

Classificació: xarxes neuronals. Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165721

Agregració (clustering). Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165722

Regles d'associació. Luis Carlos Molina Félix
Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165723

Xarxes bayesianes. Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165724

Avaluació de models. Luis Carlos Molina Félix, Ramon Sangüesa i Solé
http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165725

WEKA Manual for versión 3-6-5. The University of WAIKATO.
R. Bouckaert, E.Frank, M.Hall, R.Kirkby, P.Reutemann, A.Seewald, D.Scuse. June28 2011.
http://www.capri-model.org/docs/WekaManual-3-6-5.pdf

10. Annexos

10.1 Valors màxims, mínims, mitjana i gràfic

Les dades són de 10 cursos. Total 20 semestres.

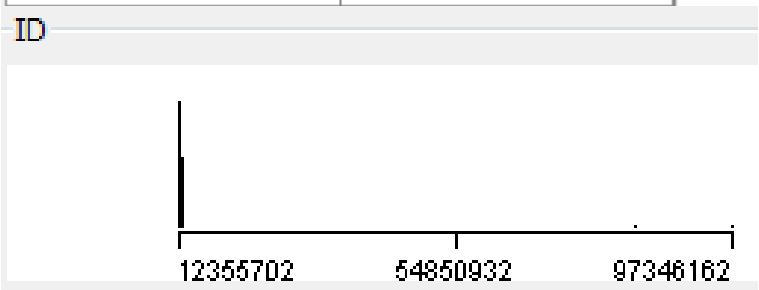
Inici 1998 2n semestre i final 2008 1r semestre. Total: 18552 alumnes

ID **Cada estudiant té un ID únic que va des del 12.355.702 al 97.346.162**

Aquest camp, en principi no aporta cap dada rellevant, ja que és diferent per a cada alumne.

Name: ID Type: Numeric
Missing: 0 (0%) Distinct: 18552 Unique: 18552 (100%)

Statistic	Value
Minimum	12355702
Maximum	97346162
Mean	12803270.525
StdDev	2225275.851

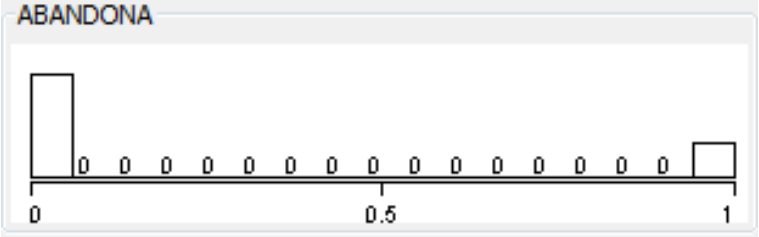


ABANDONA **Hi ha 13865 estudiants que NO abandonen (valor 0) i 4687 que SI (valor 1)**

Aquest camp si que interessa, ja que precisament és l'objectiu a relaciona amb altres dades. Podem veure que el **33,8 %** dels alumnes matriculats abandonen els estudis.

Name: ABANDONA Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.253
StdDev	0.435

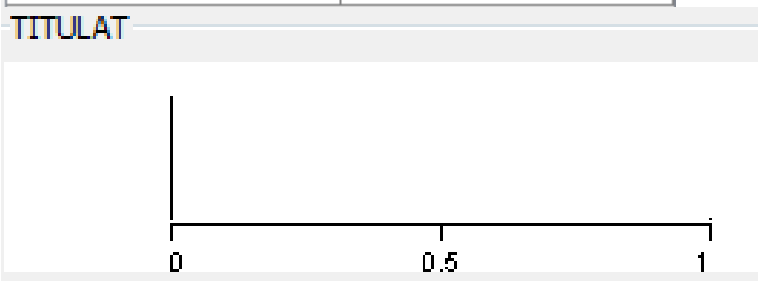


TITULAT **Hi ha 18548 estudiants que NO es titulen en el 1r semestre i 4 que SI**

Aquest camp, d'inici, no serà interessant ja que l'objectiu no es estudiar els que es titulen sinó els que abandonen els estudis.

Name: TITULAT Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0
StdDev	0.015

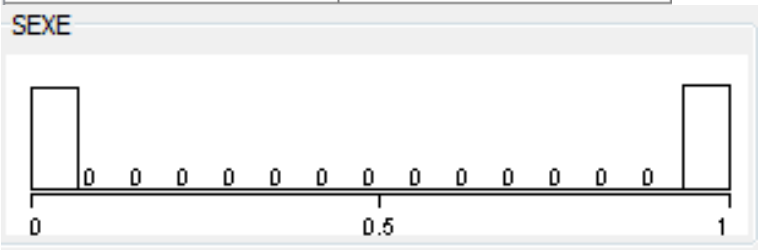


SEXE **9200 SÓN DE SEXE 0 (dona) I 9352 SÓN DE SEXE 1 (home)**

Observem que els estudiants es divideixen a parts iguals en els dos sexes. Aquest camp pot donar informació.

Name: SEXE Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.504
StdDev	0.5

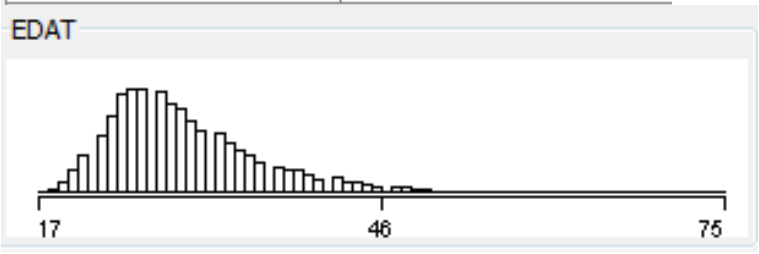


EDAT Dels 17 als 75 anys en el moment d'entrar

Veiem que l'alumne més jove en tenia 17 anys i el més vell 75 anys quan van fer la matrícula. També pot ser un camp a estudiar.

Name: EDAT Type: Numeric
Missing: 0 (0%) Distinct: 50 Unique: 6 (0%)

Statistic	Value
Minimum	17
Maximum	75
Mean	29.245
StdDev	6.241

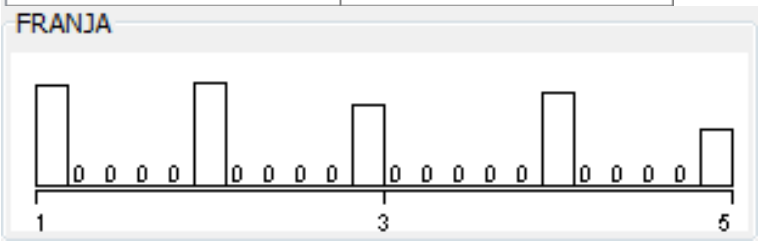


FRANJA 4310 → 1 (<=24) , 4393 → 2 (<=27) , 3483 → 3 (<=30) , 3968 → 4 (<=36) , 2398 → 5 (>36)

Aquest camp també reflecteix l'edat però de forma discretitzada. S'ha de valorar si utilitzo aquest rangs d'edat o es creen d'altres més adients.

Name: FRANJA Type: Numeric
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	2.771
StdDev	1.358

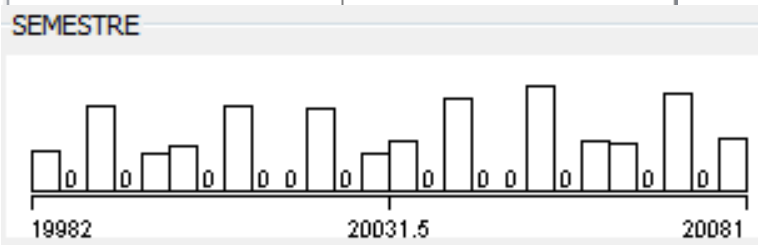


SEMESTRE (d'inici dels estudis) 898 → 1998 2n semestre ... més... 1082 → 2008 1r semestre (hi ha matrícula en tots els semestres)

Alumnes matriculats en cada semestre. Es pot donar el cas que coincideixin molts abandonaments en semestres concrets.

Name: SEMESTRE Type: Numeric
Missing: 0 (0%) Distinct: 20 Unique: 0 (0%)

Statistic	Value
Minimum	19982
Maximum	20081
Mean	20033.763
StdDev	29.135

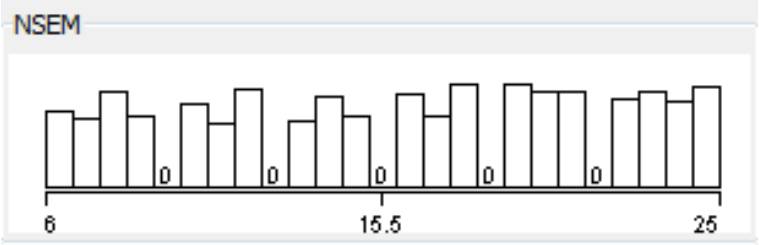


NSEM (nº se semestre relatiu des de que van iniciar estudis)

Indica el número de semestre des de que es va iniciar els estudis. Per exemple els del semestre 20082 tenen el valor 6. Això vol dir que fa 6 semestres des de que està aquest pla d'estudis.

Name: NSEM Type: Numeric
Missing: 0 (0%) Distinct: 20 Unique: 0 (0%)

Statistic	Value
Minimum	6
Maximum	25
Mean	15.977
StdDev	5.752

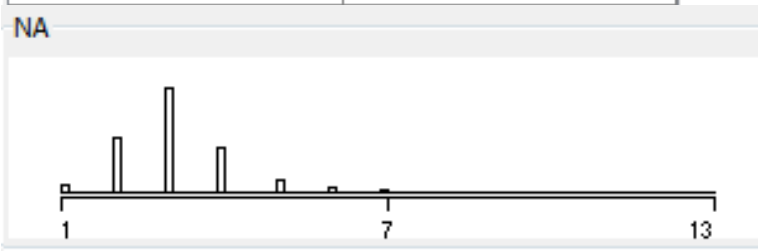


NA (assig. matriculades en el 1r semestre) 546→ 1 , 4427→2, 8430→3, 3653 →4, 951→5, 423→6, 111→ 7, 9 →8, 1→ 9, 1 → 13

Aquest camp recull el número d'assignatures matriculades en el 1r semestre. Aquesta dada pot ser important, ja que un nombre alt d'assignatures pot ajudar a l'abandonament per massa càrrega horària. Podem observar que hi ha 546 alumnes que només s'han matriculat d'una assignatura i com a extrem 1 estudiant que s'ha matriculat de 13 assignatures.

Name: NA Type: Numeric
Missing: 0 (0%) Distinct: 10 Unique: 2 (0%)

Statistic	Value
Minimum	1
Maximum	13
Mean	3.098
StdDev	1.034

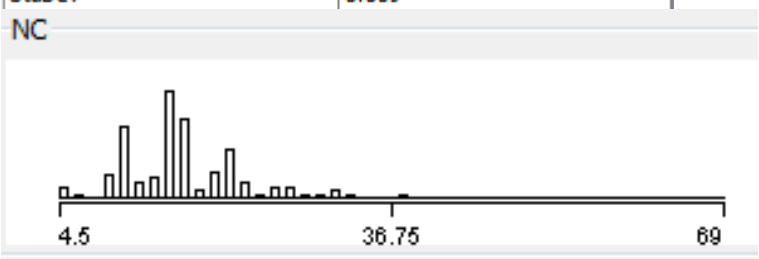


NC (crèdits. matriculats) 401 → 4,5 , 145→6, 1→7,5 , 928→9, 2865→11, més....53 → 37,5 més.. 1→48, 1→69

El mateix que el cas anterior però mesurat en nombre de crèdits. És possible que menys assignatures puguin tenir una càrrega horària més alta pel fet que aquestes siguin de molts crèdits. Podem observar que hi ha 401 alumnes que només s'han matriculat de 4,5 crèdits (segurament 1 assignatura)

Name: NC Type: Numeric
Missing: 0 (0%) Distinct: 39 Unique: 12 (0%)

Statistic	Value
Minimum	4.5
Maximum	69
Mean	16.071
StdDev	5.559



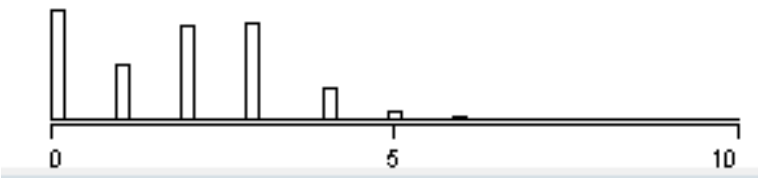
NASUP (assignatures superades) 5142→0, 2610→1, 4393→2, 4511→3, 1443→4, 314→5, 114→6, 23→7

Aquest camp recull el número d'assignatures superades en el 1r semestre. Abans havíem vist que un estudiant s'havia matriculat de 13 assignatures i aquí veiem que el màxim d'aprovades és de 7. Encara que pot no ser el mateix estudiant, segur que aquest alumne va suspendre com a mínim 6 assignatures, gairebé la meitat.

Name: NASUP Type: Numeric
Missing: 0 (0%) Distinct: 10 Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	10
Mean	1.786
StdDev	1.43

NASUP



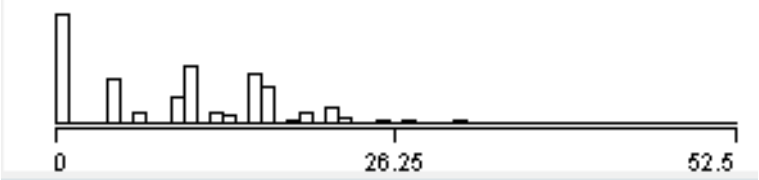
NCSUP (crèdits superats) 5142→0, 2065→4,5, 545→6, més..., 1→52,5

El mateix que l'altre camp però mesurant els crèdits superats.

Name: NCSUP Type: Numeric
Missing: 0 (0%) Distinct: 29 Unique: 5 (0%)

Statistic	Value
Minimum	0
Maximum	52.5
Mean	9.173
StdDev	7.486

NCSUP

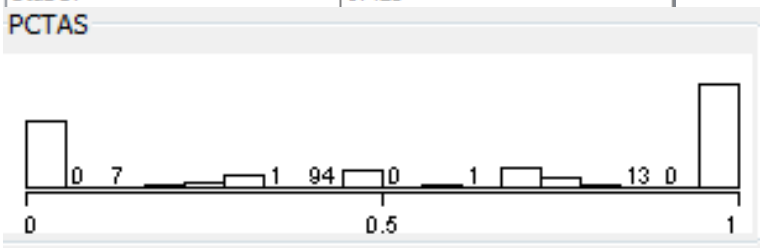


PCTAS (% assignatures superades NASUP / NA) 5142→1, ... 1263→ ≈0,5 ... 8065 → 1

Aquest camp és un camp calculat entre les assignatures aprovades i les matriculades.

Name: PCTAS Type: Numeric
Missing: 0 (0%) Distinct: 24 Unique: 4 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.59
StdDev	0.423



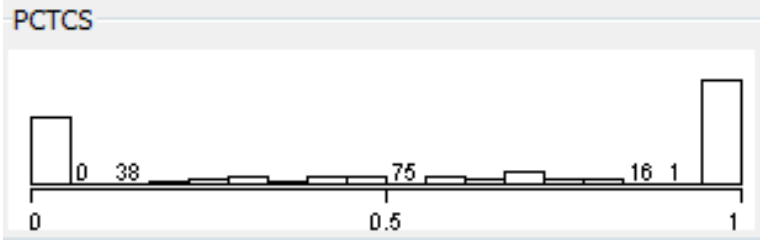
PCTCS (% crèdits superades NCSUP / NC) 5142→1, ... més ... 8065 → 1

Aquest camp és un camp calculat entre els crèdits aprovats i els matriculats.

En aquest dos casos s'ha de valorar si n'hi ha prou en valorar el percentatge de superació respecte l'abandonament dels estudis.

Name: PCTCS Type: Numeric
Missing: 0 (0%) Distinct: 117 Unique: 25 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.586
StdDev	0.425

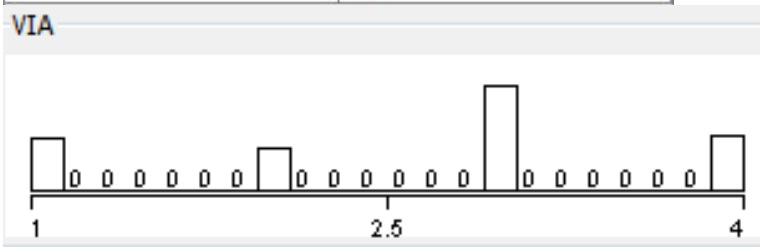


VIA 3787 → 1 (NO COU) , 3111 → 2 (COU) , 7609 → 3 (EST. INACABATS), 4045 → 4 (TITULATS)

Un camp a tenir en compte és els estudis previs que tenen els estudiants abans de fer la 1ª matrícula. A priori pot semblar que a menys estudis més risc a l'abandonament.

Name: VIA Type: Numeric
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	4
Mean	2.642
StdDev	1.036



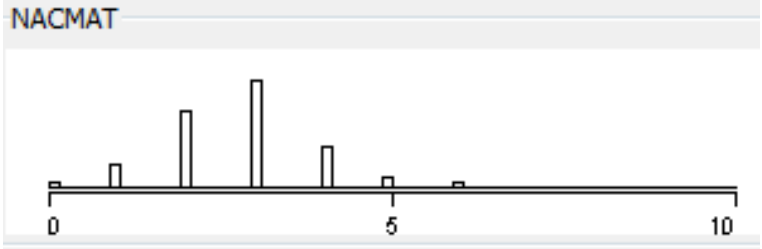
NACMAT Nº assig. A les que es matricula del conjunt de 12 més comuns 1r semestre

360→0, 1649→1, 5221→2, 7412→3, 2891→4, 692→ 5, 289→6, 37→7, 1→10

Aquí es pot veure la coincidència de la 1ª matrícula amb les 12 assignatures més comunes del 1r semestre. Es pot suposar que aquestes 12 més comunes podrien ser les que es recomana en el pla d'estudis.

Name: NACMAT Type: Numeric
Missing: 0 (0%) Distinct: 9 Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	10
Mean	2.768
StdDev	1.113



NACPRE **Nº assig. A les que es presenta del conjunt de 12 més comuns 1r semestre**

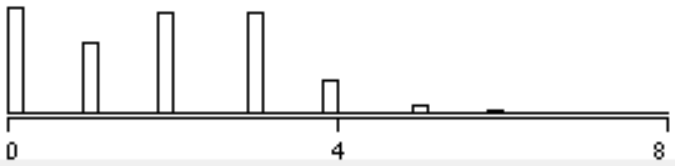
4673→0, 3156→1, 4451→2, 4404→3, 1464→4, 295→ 5, 98→6, 10→7, 1→8

Donat que no sempre es presenten els alumnes de totes les assignatures matriculades, aquesta data pot influir en l'abandonament dels estudis. Un alumne que no es presenta a cap assignatura potser més endavant ho deixa tot.

Name: NACPRE Type: Numeric
Missing: 0 (0%) Distinct: 9 Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	8
Mean	1.793
StdDev	1.39

NACPRE



NACSUP **Nº assig. Superades del conjunt de 12 més comuns 1r semestre**

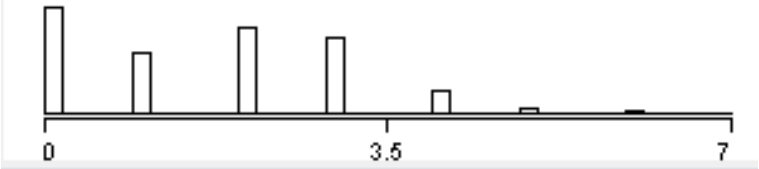
5512→0, 3213→1, 4440→2, 3910→3, 1169→4, 232→ 5, 68→6, 8→7

El fet de que es presenti a un examen, no vol dir que ho aprovi. El resultat del mateix pot influir amb els ànims de continuar estudiant.

Name: NACSUP Type: Numeric
Missing: 0 (0%) Distinct: 8 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	7
Mean	1.624
StdDev	1.366

NACSUP

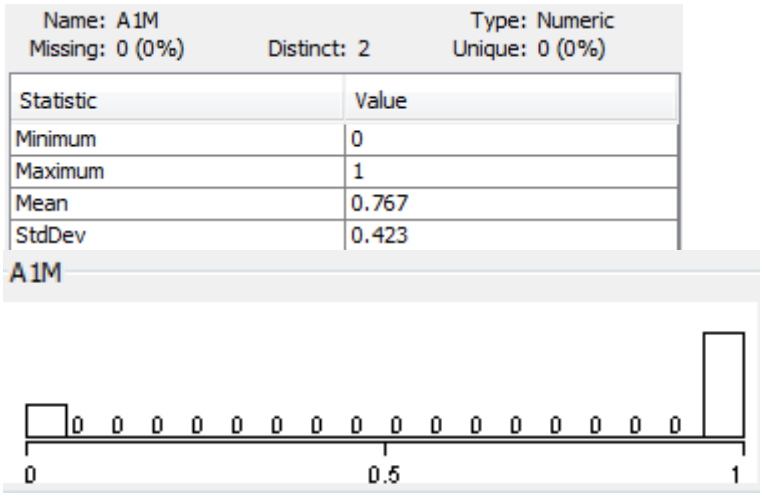


(La relació d'assignatures ve amb les dades i s'ha preguntat si l'ordre coincideix amb l'ordre del nom del camp A1, A2, ..., A12. La resposta ha estat que si.)

A1M 4327 → 0 (NO es matricula) , 14225 → 1 (SI es matricula)

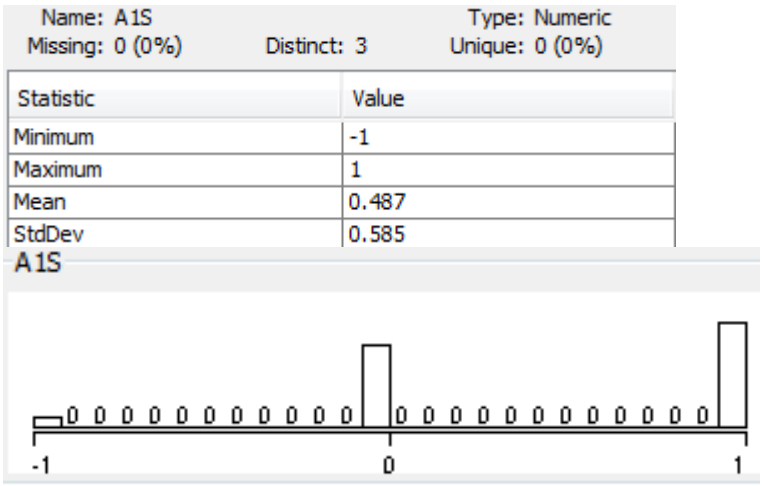
00.010: multimèdia i comunicació

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.



A1S 857 → -1 (NO supera) , 7809 → 0 (NO matricula o NO es presenta) , 9886 → 1 (SI supera)

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat. Realment els que no s'han presentat serien la resta dels que tenen aquest valor (0) menys els que no s'han matriculat.



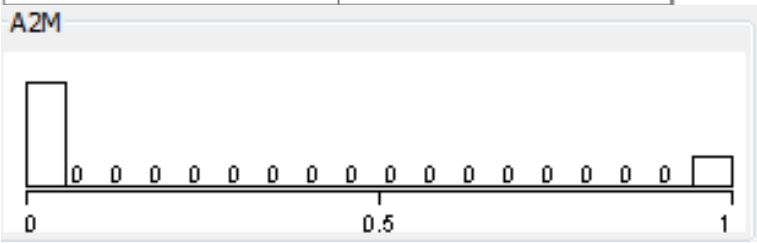
A2M 14424 → 0 , 4128 → 1

00.002: angles I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A2M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.223
StdDev	0.416

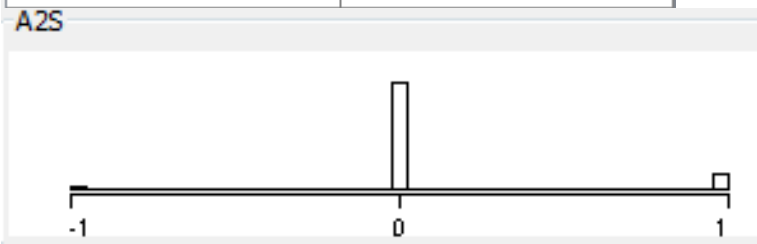


A2S 229 → -1 , 16000 → 0 , 2323 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A2S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.113
StdDev	0.353



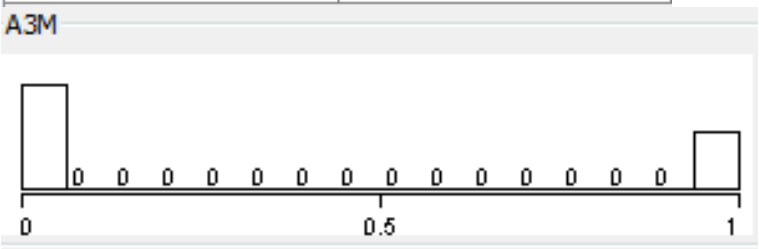
A3M 11906 → 0 , 6646 → 1

01.001: introducció al dret

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A3M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.358
StdDev	0.479

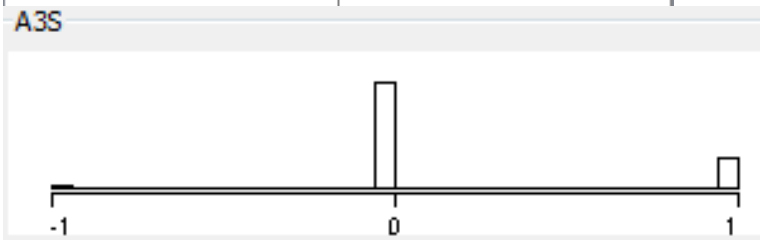


A3S 303 → -1 , 14213 → 0 , 4036 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A3S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.201
StdDev	0.44



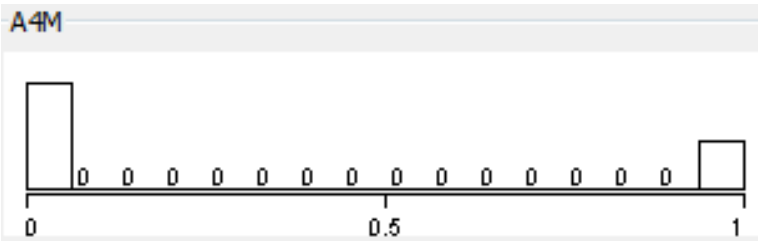
A4M 12670 → 0 , 5882 → 1

01.079: introducció a la macroeconomia

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A4M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.317
StdDev	0.465

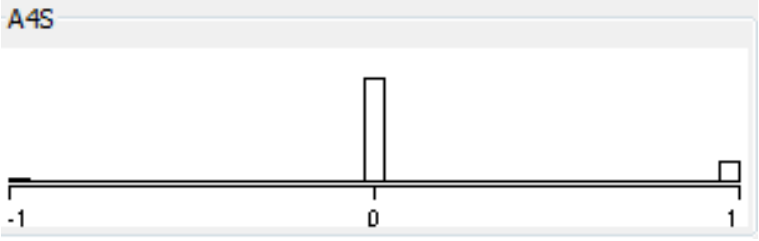


A4S 412 → -1 , 15101 → 0 , 3039 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A4S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.142
StdDev	0.407



A5M 12618 → 0 , 5934 → 1

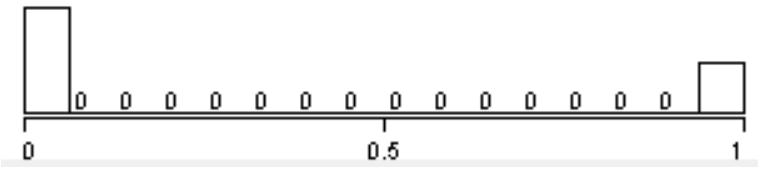
01.005: introducció a la comptabilitat

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A5M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.32
StdDev	0.466

A5M



A5S 410 → -1 , 14807 → 0 , 3335 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A5S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.158
StdDev	0.421

A5S

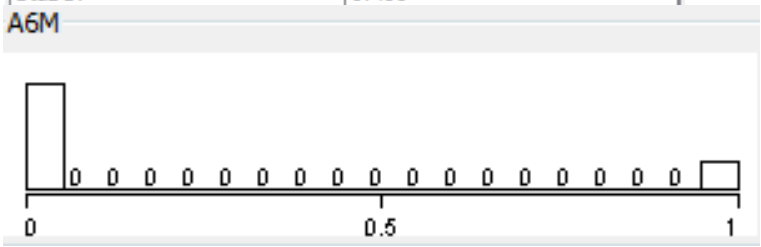
A6M 14702 → 0 , 3850 → 1

01.003: matemàtiques I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A6M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.208
StdDev	0.406

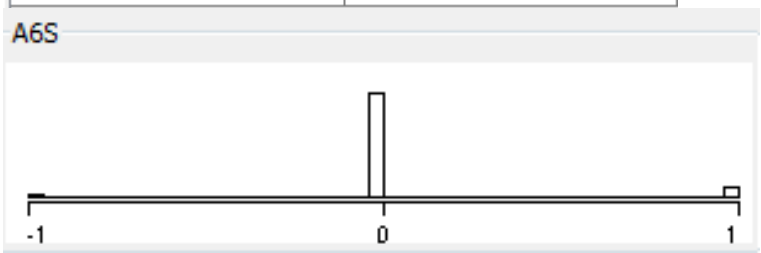


A6S 358 → -1 , 16625 → 0 , 1569 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A6S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.065
StdDev	0.316



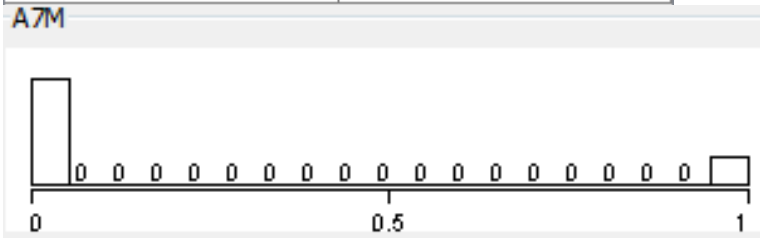
A7M 14625 → 0 , 3927 → 1

01.006: organització i administració d'empreses I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A7M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.212
StdDev	0.409

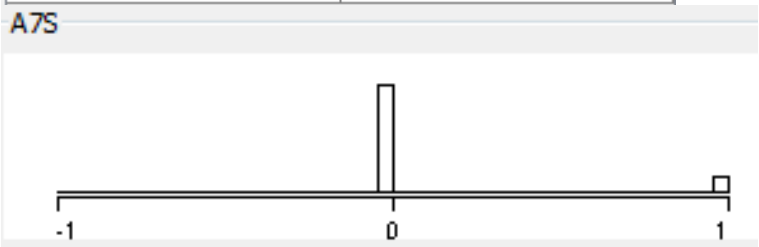


A7S 120 → -1 , 16037 → 0 , 2395 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A7S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.123
StdDev	0.347



A8M 16217 → 0 , 2335 → 1

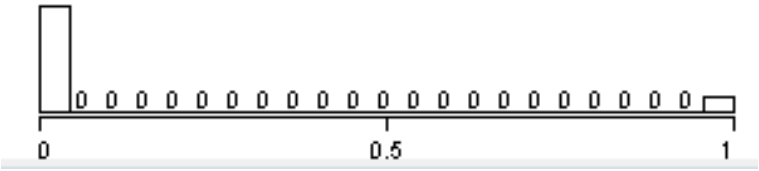
01.004: estadística I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A8M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.126
StdDev	0.332

A8M



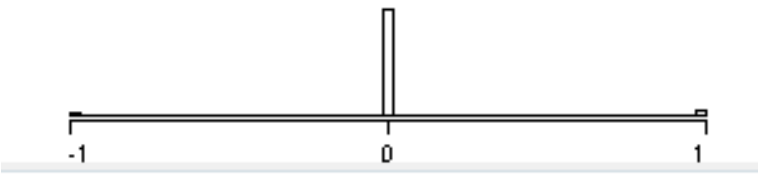
A8S 238 → -1 , 17440 → 0 , 874 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A8S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.034
StdDev	0.242

A8S



A9M 17160 → 0 , 1392 → 1

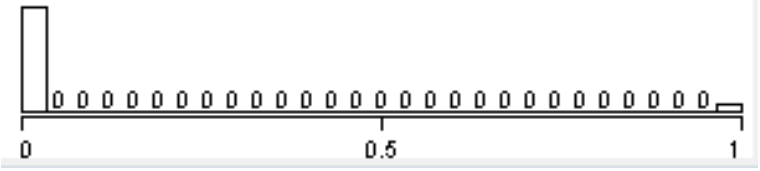
01.078: introducció a la microeconomia

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A9M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.075
StdDev	0.263

A9M



A9S 66 → -1 , 17707 → 0 , 779 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A9S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.038
StdDev	0.21

A9S

A10M 17479 → 0 , 1073 → 1

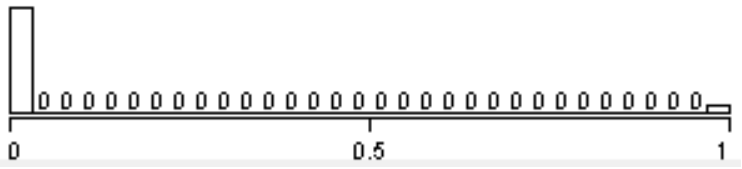
01.009: direcció de la producció I

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A10M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.058
StdDev	0.233

A10M



A10S 75 → -1 , 17925 → 0 , 552 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A10S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.026
StdDev	0.182

A10S

A11M 17487 → 0 , 1065 → 1

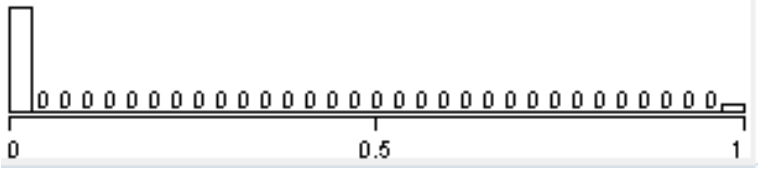
00.004: angles III

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A11M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.057
StdDev	0.233

A11M



A11S 25 → -1 , 17786 → 0 , 741 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A11S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.039
StdDev	0.2

A11S

A12M 17655 → 0 , 897 → 1

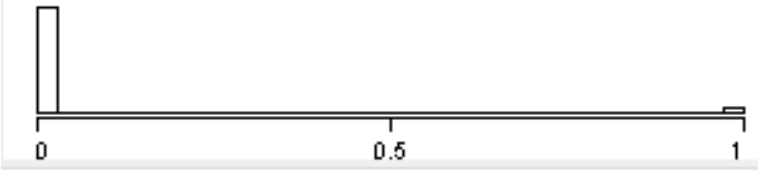
00.003: angles II

Aquí podem veure el nombre d'estudiants que es matriculen d'aquesta assignatura.

Name: A12M Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.048
StdDev	0.215

A12M



A12S 51 → -1 , 17907 → 0 , 594 → 1

Aquí podem veure els que no aproven, aproven i la suma dels que o bé no s'han matriculat o bé no s'han presentat.

Name: A12S Type: Numeric
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	-1
Maximum	1
Mean	0.029
StdDev	0.184

A12S

10.2 Probabilitat de matrícula d'assignatures segons clusterització

S'ha aplicat una clusterització **kmeans** amb **2 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1			
	18552	14625	3927			
=====	=====	=====	=====	0	1	Probabilitat
A1M	SI	SI	SI	0,788	0,212	1 A1M
A2M	NO	NO	NO	0	0	0 A2M
A3M	NO	NO	NO	0	0	0 A3M
A4M	NO	NO	NO	0	0	0 A4M
A5M	NO	NO	NO	0	0	0 A5M
A6M	NO	NO	NO	0	0	0 A6M
A7M	NO	NO	SI	0	0,212	0,212 A7M
A8M	NO	NO	NO	0	0	0 A8M
A9M	NO	NO	NO	0	0	0 A9M
A10M	NO	NO	NO	0	0	0 A10M
A11M	NO	NO	NO	0	0	0 A11M
A12M	NO	NO	NO	0	0	0 A12M

S'ha aplicat una clusterització **kmeans** amb **3 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2				
	18552	10173	3927	4452				
=====	=====	=====	=====	=====	0	1	2	Probabilitat
A1M	SI	SI	SI	SI	0,548	0,212	0,24	1 A1M
A2M	NO	NO	NO	NO	0	0	0	0 A2M
A3M	NO	NO	NO	NO	0	0	0	0 A3M
A4M	NO	NO	NO	NO	0	0	0	0 A4M
A5M	NO	NO	NO	SI	0	0	0,24	0,24 A5M
A6M	NO	NO	NO	NO	0	0	0	0 A6M
A7M	NO	NO	SI	NO	0	0,212	0	0,212 A7M
A8M	NO	NO	NO	NO	0	0	0	0 A8M
A9M	NO	NO	NO	NO	0	0	0	0 A9M
A10M	NO	NO	NO	NO	0	0	0	0 A10M
A11M	NO	NO	NO	NO	0	0	0	0 A11M
A12M	NO	NO	NO	NO	0	0	0	0 A12M

S'ha aplicat una clusterització **kmeans** amb **4 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3						
	18552	7183	2671	5334	3364						
=====	=====	=====	=====	=====	=====	0	1	2	3	Probabilitat	
A1M	SI	SI	SI	NO	SI	0,387	0,144	0	0,181	0,712	A1M
A2M	NO	NO	NO	NO	SI	0	0	0	0,181	0,181	A2M
A3M	NO	NO	NO	NO	SI	0	0	0	0,181	0,181	A3M
A4M	NO	SI	NO	NO	NO	0,387	0	0	0	0,387	A4M
A5M	NO	NO	NO	SI	NO	0	0	0,288	0	0,288	A5M
A6M	NO	NO	NO	NO	NO	0	0	0	0	0	A6M
A7M	NO	NO	SI	NO	NO	0	0,144	0	0	0,144	A7M
A8M	NO	NO	NO	NO	NO	0	0	0	0	0	A8M
A9M	NO	NO	NO	NO	NO	0	0	0	0	0	A9M
A10M	NO	NO	NO	NO	NO	0	0	0	0	0	A10M
A11M	NO	NO	NO	NO	NO	0	0	0	0	0	A11M
A12M	NO	NO	NO	NO	NO	0	0	0	0	0	A12M

S'ha aplicat una clusterització **kmeans** amb **5 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3	4						
	18552	5517	2734	4038	2091	4172						
=====	=====	=====	=====	=====	=====	=====	0	1	2	3	4	Probabilitat	
A1M	SI	SI	SI	NO	SI	SI	0,297	0,147	0	0,113	0,225	0,782	A1M
A2M	NO	NO	NO	NO	SI	NO	0	0	0	0,113	0	0,113	A2M
A3M	NO	NO	NO	NO	SI	NO	0	0	0	0,113	0	0,113	A3M
A4M	NO	SI	NO	NO	NO	NO	0,297	0	0	0	0	0,297	A4M
A5M	NO	NO	NO	SI	SI	NO	0	0	0,218	0,113	0	0,33	A5M
A6M	NO	NO	NO	NO	NO	NO	0	0	0	0	0	0	A6M
A7M	NO	NO	SI	NO	NO	NO	0	0,147	0	0	0	0,147	A7M
A8M	NO	NO	NO	NO	NO	NO	0	0	0	0	0	0	A8M
A9M	NO	NO	NO	NO	NO	NO	0	0	0	0	0	0	A9M
A10M	NO	NO	NO	NO	NO	NO	0	0	0	0	0	0	A10M
A11M	NO	NO	NO	NO	NO	NO	0	0	0	0	0	0	A11M
A12M	NO	NO	NO	NO	NO	NO	0	0	0	0	0	0	A12M

S'ha aplicat una clusterització **kmeans** amb **6 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3	4	5
	18552	4113	3380	3019	2523	3230	2287
=====	=====	=====	=====	=====	=====	=====	=====
A1M	SI	SI	SI	NO	SI	SI	NO
A2M	NO	NO	NO	NO	SI	NO	NO
A3M	NO	NO	NO	NO	SI	NO	SI
A4M	NO	SI	NO	NO	NO	NO	SI
A5M	NO	NO	NO	SI	SI	NO	NO
A6M	NO	SI	NO	NO	NO	NO	NO
A7M	NO	NO	SI	NO	NO	NO	NO
A8M	NO	NO	NO	NO	NO	NO	NO
A9M	NO	NO	NO	NO	NO	NO	NO
A10M	NO	NO	NO	NO	NO	NO	NO
A11M	NO	NO	NO	NO	NO	NO	NO
A12M	NO	NO	NO	NO	NO	NO	NO

0	1	2	3	4	5	Probabilitat	
0,222	0,182	0	0,136	0,174	0	0,714	A1M
0	0	0	0,136	0	0	0,136	A2M
0	0	0	0,136	0	0,123	0,259	A3M
0,222	0	0	0	0	0,123	0,345	A4M
0	0	0,163	0,136	0	0	0,299	A5M
0,222	0	0	0	0	0	0,222	A6M
0	0,182	0	0	0	0	0,182	A7M
0	0	0	0	0	0	0	A8M
0	0	0	0	0	0	0	A9M
0	0	0	0	0	0	0	A10M
0	0	0	0	0	0	0	A11M
0	0	0	0	0	0	0	A12M

S'ha aplicat una clusterització **kmeans** amb **7 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3	4	5	6
	18552	4113	3366	3019	2510	3230	2180	134
=====	=====	=====	=====	=====	=====	=====	=====	=====
A1M	SI	SI	SI	NO	SI	SI	NO	NO
A2M	NO	NO	NO	NO	SI	NO	NO	NO
A3M	NO	NO	NO	NO	SI	NO	SI	SI
A4M	NO	SI	NO	NO	NO	NO	SI	SI
A5M	NO	NO	NO	SI	SI	NO	NO	NO
A6M	NO	SI	NO	NO	NO	NO	NO	NO
A7M	NO	NO	SI	NO	NO	NO	NO	NO
A8M	NO	NO	NO	NO	NO	NO	NO	NO
A9M	NO	NO	NO	NO	NO	NO	NO	NO
A10M	NO	NO	NO	NO	NO	NO	NO	SI
A11M	NO	NO	NO	NO	NO	NO	NO	NO
A12M	NO	NO	NO	NO	NO	NO	NO	NO

0	1	2	3	4	5	6	Probabilitat	
0,222	0,181	0	0,135	0,174	0	0	0,713	A1M
0	0	0	0,135	0	0	0	0,135	A2M
0	0	0	0,135	0	0,118	0,007	0,26	A3M
0,222	0	0	0	0	0,118	0,007	0,346	A4M
0	0	0,163	0,135	0	0	0	0,298	A5M
0,222	0	0	0	0	0	0	0,222	A6M
0	0,181	0	0	0	0	0	0,181	A7M
0	0	0	0	0	0	0	0	A8M
0	0	0	0	0	0	0	0	A9M
0	0	0	0	0	0	0,007	0,007	A10M
0	0	0	0	0	0	0	0	A11M
0	0	0	0	0	0	0	0	A12M

S'ha aplicat una clusterització **kmeans** amb **8 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3	4	5	6	7
	18552	3650	3209	3019	2257	3230	2180	134	873
=====	=====	=====	=====	=====	=====	=====	=====	=====	=====
A1M	SI	SI	SI	NO	SI	SI	NO	NO	SI
A2M	NO	NO	NO	NO	SI	NO	NO	NO	NO
A3M	NO	NO	NO	NO	SI	NO	SI	SI	NO
A4M	NO	SI	NO	NO	NO	NO	SI	SI	SI
A5M	NO	NO	NO	SI	SI	NO	NO	NO	SI
A6M	NO	SI	NO	NO	NO	NO	NO	NO	NO
A7M	NO	NO	SI	NO	NO	NO	NO	NO	NO
A8M	NO	NO	NO	NO	NO	NO	NO	NO	NO
A9M	NO	NO	NO	NO	NO	NO	NO	NO	NO
A10M	NO	NO	NO	NO	NO	NO	NO	SI	NO
A11M	NO	NO	NO	NO	NO	NO	NO	NO	NO
A12M	NO	NO	NO	NO	NO	NO	NO	NO	NO

0	1	2	3	4	5	6	7	Probabilitat	
0,197	0,173	0	0,122	0,174	0	0	0,047	0,713	A1M
0	0	0	0,122	0	0	0	0	0,122	A2M
0	0	0	0,122	0	0,118	0,007	0	0,246	A3M
0,197	0	0	0	0	0,118	0,007	0,047	0,369	A4M
0	0	0,163	0,122	0	0	0	0,047	0,331	A5M
0,197	0	0	0	0	0	0	0	0,197	A6M
0	0,173	0	0	0	0	0	0	0,173	A7M
0	0	0	0	0	0	0	0	0	A8M
0	0	0	0	0	0	0	0	0	A9M
0	0	0	0	0	0	0,007	0	0,007	A10M
0	0	0	0	0	0	0	0	0	A11M
0	0	0	0	0	0	0	0	0	A12M

S'ha aplicat una clusterització **kmeans** amb **9 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3	4	5	6	7	8
	18552	3424	3149	3019	2192	3230	2040	134	873	491
=====	=====	=====	=====	=====	=====	=====	=====	=====	=====	=====
A1M	SI	SI	SI	NO	SI	SI	NO	NO	SI	NO
A2M	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO
A3M	NO	NO	NO	NO	SI	NO	SI	SI	NO	NO
A4M	NO	SI	NO	NO	NO	NO	SI	SI	SI	NO
A5M	NO	NO	NO	SI	SI	NO	NO	NO	SI	NO
A6M	NO	SI	NO	NO	NO	NO	NO	NO	NO	SI
A7M	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
A8M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A9M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A10M	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO
A11M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A12M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO

0	1	2	3	4	5	6	7	8	Probabilitat	
0,185	0,17	0	0,118	0,174	0	0	0,047	0	0,694	A1M
0	0	0	0,118	0	0	0	0	0	0,118	A2M
0	0	0	0,118	0	0,11	0,007	0	0	0,235	A3M
0,185	0	0	0	0	0,11	0,007	0,047	0	0,349	A4M
0	0	0,163	0,118	0	0	0	0,047	0	0,328	A5M
0,185	0	0	0	0	0	0	0	0,026	0,211	A6M
0	0,17	0	0	0	0	0	0	0	0,17	A7M
0	0	0	0	0	0	0	0	0	0	A8M
0	0	0	0	0	0	0	0	0	0	A9M
0	0	0	0	0	0	0,007	0	0	0,007	A10M
0	0	0	0	0	0	0	0	0	0	A11M
0	0	0	0	0	0	0	0	0	0	A12M

S'ha aplicat una clusterització **kmeans** amb **10 clústers** dels atributs de la matrícula de les 12 assignatures i després s'ha calculat la probabilitat de matrícula de cada assignatura.

Attribute	Full Data	0	1	2	3	4	5	6	7	8	9
	18552	3240	3123	3019	2163	3230	1909	134	873	491	370
A1M	SI	SI	SI	NO	SI	SI	NO	NO	SI	NO	SI
A2M	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO
A3M	NO	NO	NO	NO	SI	NO	SI	SI	NO	NO	NO
A4M	NO	SI	NO	NO	NO	NO	SI	SI	SI	NO	SI
A5M	NO	NO	NO	SI	SI	NO	NO	NO	SI	NO	NO
A6M	NO	SI	NO	NO	NO	NO	NO	NO	NO	SI	NO
A7M	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO
A8M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI
A9M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A10M	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO
A11M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A12M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO

0	1	2	3	4	5	6	7	8	9	Probabilitat	
0,175	0,168	0	0,117	0,174	0	0	0,047	0	0,02	0,701	A1M
0	0	0	0,117	0	0	0	0	0	0	0,117	A2M
0	0	0	0,117	0	0,103	0,007	0	0	0	0,227	A3M
0,175	0	0	0	0	0,103	0,007	0,047	0	0,02	0,352	A4M
0	0	0,163	0,117	0	0	0	0,047	0	0	0,326	A5M
0,175	0	0	0	0	0	0	0	0,026	0	0,201	A6M
0	0,168	0	0	0	0	0	0	0	0	0,168	A7M
0	0	0	0	0	0	0	0	0	0,02	0,02	A8M
0	0	0	0	0	0	0	0	0	0	0	A9M
0	0	0	0	0	0	0,007	0	0	0	0,007	A10M
0	0	0	0	0	0	0	0	0	0	0	A11M
0	0	0	0	0	0	0	0	0	0	0	A12M

10.3 Mineria de dades per fer un primer estudi

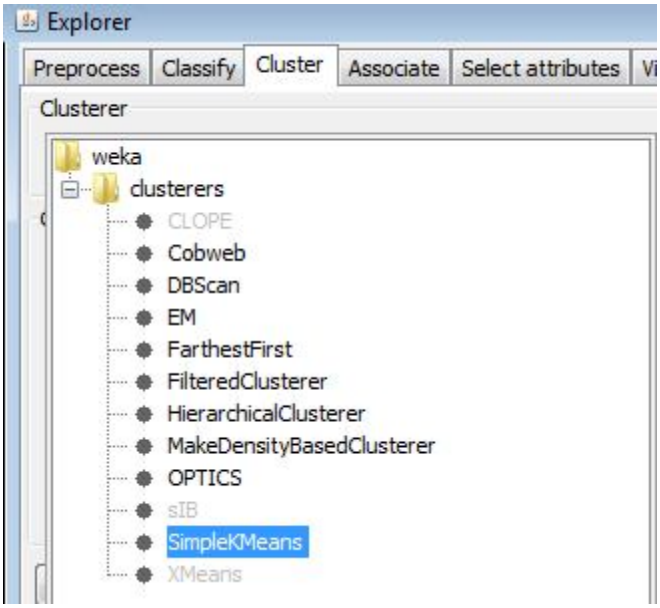
El primer que es farà es intentar trobar similituds i agrupar registres semblants. Això es farà amb **models d'agregació** (clustering) (4.1) i **models associatius** (4.2). Això permetrà observar l' importància d'alguns camps i els valors a tenir en compte. Es farà amb tot el conjunt de dades i atributs i en raó dels resultats també es repetiran ignorants alguns atributs o amb alguna mostra de dades reduïdes o després d'eliminar atributs que al llarg del procés es vagin desestimant per l'estudi.

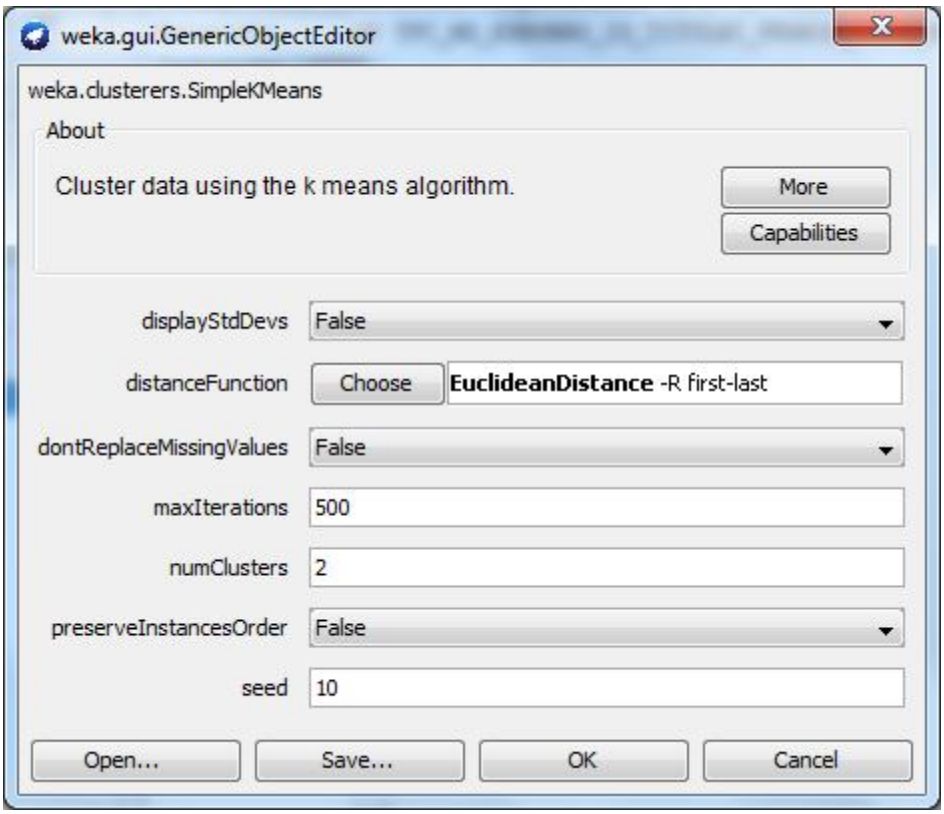
Amb el mateix criteri s'aplicaran **models de classificació** (4.3), bàsicament regles de classificació i també arbres de decisió que ens permetran a més de classificar predir situacions futures.

10.3.1 Clusterització

L'algorisme **SimpleKMeans** agafa k elements aleatòriament, sent k el número de clústers que es volen, i després cadascuna de les instàncies és assignada al centre del clúster més a prop. Després es torna a calcular el centroide de totes les instàncies i aquest es pres com a nou centre dels corresponents clústers. Aquest procés es va repetint de forma iterativa fins que els punts centrals s'han estabilitzat.

Amb l'algorisme **SimpleKMeans** es farà un primer model d'agregació amb els valors per defecte. En aquest cas només genera 2 clústers i el resultat no es molt bo ja que la suma d'errors al quadrat és molt elevada.



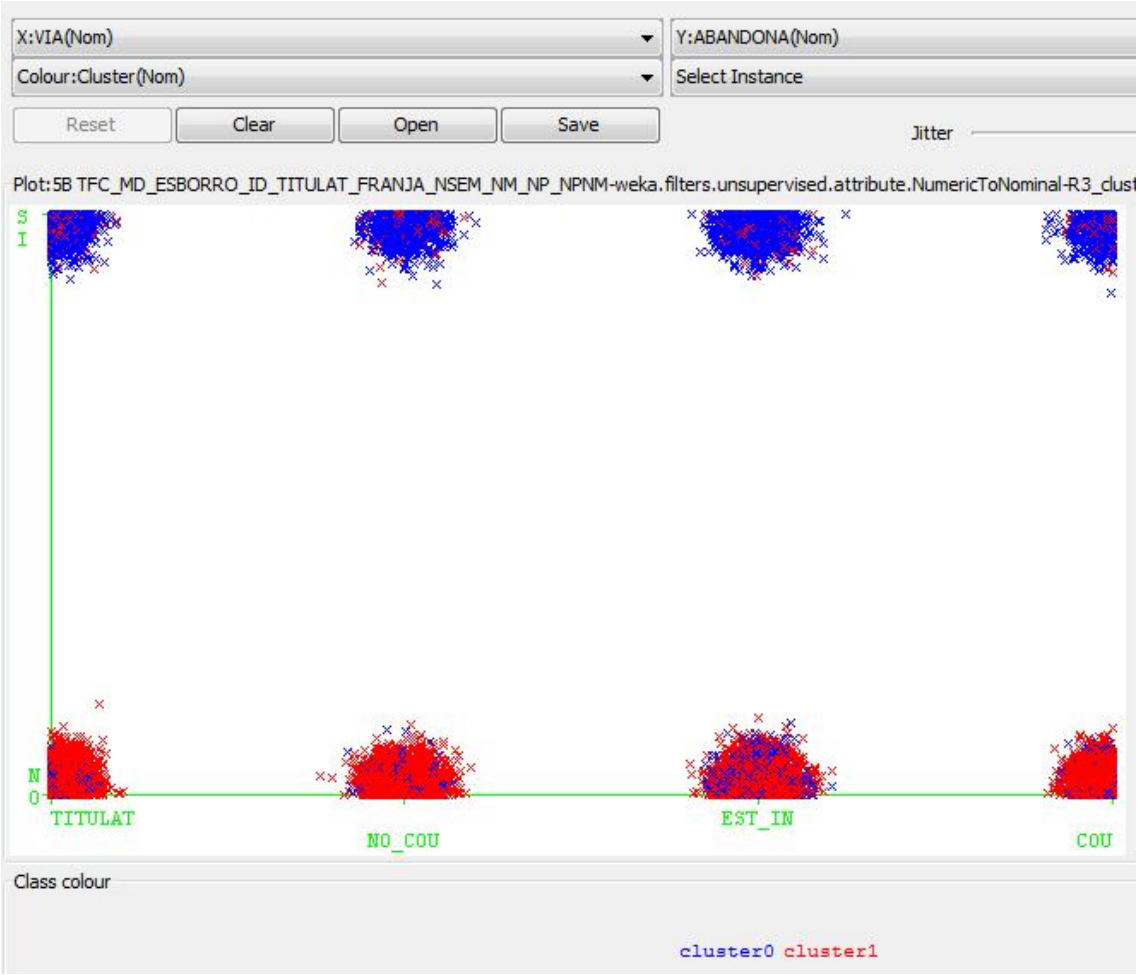



```
Test mode:evaluate on training data
=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 142068.1871903053
Missing values globally replaced with mean/mode

Clúster centroids:
                              Clúster#
Attribute        Full Data           0           1
                   (18552)      (5796)     (12756)
=================================================
SEXE                 HOME        HOME        DONA
EDAT          (22.5-24.5] (22.5-24.5] (27.5-29.5]
SEMESTRE            20051       20062       20012
NA                 3.0976      3.1586      3.0698
NC                 16.071      16.424     15.9106
NASUP               1.786      0.2189       2.498
NCSUP              9.1728      1.1071     12.8377
PCTAS              0.5903      0.0701      0.8267
PCTCS              0.5857       0.068      0.8209
VIA                EST_IN      EST_IN      EST_IN
NACMAT             2.7681      2.7864      2.7598
NACPRE             1.7932      0.3799      2.4353
```

```
NACSUP                  1.6237       0.1891       2.2756
A1M                         SI           SI           SI
A1S                         SI           NP           SI
A2M                         NO           NO           NO
A2S                         NM           NM           NM
A3M                         NO           NO           NO
A3S                         NM           NM           NM
A4M                         NO           NO           NO
A4S                         NM           NM           NM
A5M                         NO           NO           NO
A5S                         NM           NM           NM
A6M                         NO           NO           NO
A6S                         NM           NM           NM
A7M                         NO           NO           NO
A7S                         NM           NM           NM
A8M                         NO           NO           NO
A8S                         NM           NM           NM
A9M                         NO           NO           NO
A9S                         NM           NM           NM
A10M                        NO           NO           NO
A10S                        NM           NM           NM
A11M                        NO           NO           NO
A11S                        NM           NM           NM
A12M                        NO           NO           NO
A12S                        NM           NM           NM
ABANDONA                    NO           SI           NO

Clustered Instances

0       5796 ( 31%)
1      12756 ( 69%)
```

Si es fa cas d'aquests resultats, weka ha agrupat 5796 estudiants majoritàriament homes que **SI** abandonen els estudis, que molts d'ells es van matricular al 2n semestre del 2006, d'edat entre 23 i 24 anys. La mitjana d'assignatures matriculades és de 3 i el percentatge d'assignatures superades és del 7% (0,21 assignatura). Molts provenen d'estudis inacabats i la majoria no va aprovar cap assignatura.

Per altre banda, 12756 estudiants majoritàriament dones que **NO** abandonen els estudis i molts es van matricular al 2n semestre del 2001, d'edat entre 28 i 29 anys. La mitjana d'assignatures matriculades també és de 3 però el percentatge d'assignatures superades puja al 82% (2,5 assignatures). També molts provenen d'estudis inacabats i com a mínim van aprovar la **A1M**



Es pot observar amb la visualització de l'encreuament de la procedència dels estudis i l'abandonament dels estudis que hi ha registres que estan assignats a clusters que no corresponen amb l'assignació que s'ha fet de cluster 0 → **SI** abandonen i cluster 1 → **NO** abandonen.

Ara es repetirà l'algorisme però fent més clusters que permetin que els estudiants que pertanyin a cadascun d'ells siguin més similars. En aquest cas es selecciona 20 clusters pel fet que hi ha 20 semestres en les dades treballades.

Es pot veure que dels 20 clusters hi ha 4 en els que en l'atribut **ABANDONA** te com a majoritari el **SI**. Analitzant aquests 4 clusters tenim que:

Clúster 0: 1170 estudiants, home, 23-24 anys, 2n semestre 2006, 3,5 assignatures amb 3% aprovades procedent d'estudis inacabats.

Clúster 9: 323 estudiants, home, 28-29 anys, 2n semestre 2001, 4 assignatures amb 18% aprovades procedent d'estudiants titulats.

Clúster 10: 488 estudiants, dona, 23-24 anys, 1r semestre 2008, 2,7 assignatures amb 18% aprovades procedent d'estudis inacabats.

Clúster 16: 1771 estudiants, dona, 28-29 anys, 2n semestre 1992, 2,9 assignatures amb 10% aprovades procedent d'estudis inacabats.

Attribute	Full	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	18552	1170	1674	506	360	2727	863	1515	1527	274	323	488	344	218	418	508	192	1771	346	731	777
SEXE	HOME	HOME	HOME	HOME	DONA	DONA	DONA	HOME	DONA	HOME	HOME	DONA	HOME	HOME	HOME	DONA	HOME	DONA	HOME	HOME	DONA
EDAT	(22.5-24.5]	(22.5-24.5]	(37.5-inf)	(22.5-24.5]	(22.5-24.5]	(22.5-24.5]	(-inf-22.5]	(27.5-29.5]	(29.5-31.5]	(31.5-33.5]	(27.5-29.5]	(22.5-24.5]	(37.5-inf)	(22.5-24.5]	(-inf-22.5]	(31.5-33.5]	(33.5-37.5]	(27.5-29.5]	(26.5-27.5]	(37.5-inf)	(25.5-26.5]
SEMESTRI	20051	20062	20012	20062	20031	20012	20051	20071	20041	20081	20012	20081	19982	20061	20071	20032	20081	19932	20032	20022	20071
NA	30.376	35.043	3.043	26.067	35.316	25.556	30.742	34.779	28.677	32.226	41.796	27.766	26.451	35.321	34.282	36.634	43.438	29.304	34.979	35.937	2.583
NC	16.071	177.667	160.188	146.307	186.536	127.036	163.169	176.762	152.633	178.577	217.415	153.324	13.652	184.679	180.634	185.463	227.188	156.143	177.283	180.903	137.477
NASUP	1.786	0.1162	23.238	21.067	32.041	1.834	22.433	29.116	24.165	2.708	0.1796	0.1865	0.833	0.4725	0.4928	32.835	34.115	0.1005	1.981	31.431	13.263
NCSUP	31.728	0.5423	121.595	119.585	168.803	32.497	119.687	147.554	127.613	149.781	0.8684	0.359	43.162	2.367	24.364	164.311	177.813	0.4718	96.216	157.715	70.051
PCTAS	0.5903	0.0338	0.7753	0.8385	0.9158	0.7709	0.7478	0.847	0.8578	0.8579	0.0403	0.0654	0.3274	0.1325	0.1397	0.9074	0.7393	0.0342	0.5861	0.8881	0.5109
PCTCS	0.5857	0.0313	0.7712	0.8417	0.9138	0.7606	0.75	0.8446	0.852	0.8579	0.0371	0.062	0.3254	0.1277	0.1354	0.902	0.7372	0.0302	0.5688	0.8857	0.5087
VIA	EST_IN	EST_IN	EST_IN	EST_IN	NO_COU	EST_IN	NO_COU	TITULAT	EST_IN	COU	TITULAT	EST_IN	EST_IN	EST_IN	EST_IN	TITULAT	TITULAT	EST_IN	EST_IN	EST_IN	NO_COU
NACMAT	27.681	33.752	28.548	2.243	34.684	20.264	27.372	33.367	27.472	29.672	40.712	22.932	13.665	30.642	31.633	35.365	41.034	27.631	32.326	34.311	14.736
NACPRE	17.932	0.2812	23.453	19.348	33.031	16.964	22.375	30.172	25.095	26.241	0.3437	0.2932	0.572	0.6055	0.7153	33.878	3.5	0.2603	21.723	32.339	0.7328
NACSUP	16.237	0.1137	21.977	1.83	31.561	15.541	20.313	28.568	23.287	25.109	0.1672	0.1578	0.3951	0.3807	0.4641	32.264	3.25	0.0999	18.584	30.807	0.6396
A1M	SI	SI	SI	NO	SI	SI	NO	SI	SI	NO	SI	NO	NO	NO	NO	SI	NO	SI	SI	SI	NO
A1S	SI	NP	SI	NM	SI	SI	NM	SI	SI	NM	NP	NM	NM	NM	NM	SI	NM	NP	SI	SI	NM
A2M	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO
A2S	NM	NM	NM	NM	SI	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	SI	NM	NM
A3M	NO	NO	NO	NO	SI	NO	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO
A3S	NM	NM	NM	NM	SI	NM	SI	SI	NM	NM	NM	NM	NM	NM	NM	NM	SI	NM	NM	NM	NM
A4M	NO	SI	NO	NO	NO	NO	SI	SI	NO	NO	SI	NO	NO	NO	SI	SI	SI	NO	NO	SI	NO
A4S	NM	NP	NM	NM	NM	NM	SI	SI	NM	NM	NP	NM	NM	NM	NP	SI	SI	NM	NM	SI	NM
A5M	NO	NO	NO	SI	SI	NO	NO	NO	SI	NO	SI	SI	NO	NO	NO	NO	SI	SI	NO	SI	NO
A5S	NM	NM	NM	SI	SI	NM	NM	NM	SI	NM	NP	NP	NM	NM	NM	NM	SI	NP	NM	SI	NM
A6M	NO	SI	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO	SI	NO	NO	NO	NO	SI	NO	NO
A6S	NM	NP	NM	NM	NM	NM	NM	NM	NM	SI	NM	NM	NM	NP	NM	NM	NM	NM	NP	NM	NM
A7M	NO	NO	SI	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A7S	NM	NM	SI	NM	NM	NM	NM	NM	NM	SI	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
A8M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO	SI	SI	NO	NO	NO	NO	NO
A8S	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NP	NM	NM	NM	NP	SI	NM	NM	NM	NM	NM
A9M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO	SI	NO	NO	NO	NO	NO
A9S	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NP	NM	NM	SI	NM	NM	NM	NM
A10M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A10S	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
A11M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A11S	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
A12M	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
A12S	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
ABANDON	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	SI	SI	NO	NO	NO	NO	NO	SI	NO	NO	NO

Mentre més clústers es facin, més similars seran els estudiants que en formaran part.

No obstant el mètode d'agregació, amb aquestes dades no dona prou informació que permeti detectar alguns valors de determinats atributs que són rellevants per tal que un alumne abandoni els estudis que ha iniciat.

S'ha tornat a repetir l'algorisme **SimpleKMeans** ignorant atributs relacionats amb crèdits i amb si s'han matriculat o no a les 12 assignatures més comunes, ja que aquesta informació ja surt en els corresponents atributs de si l'han superada o no i si no s'han matriculat o no presentat.

Els resultats són:

Clúster 0: 1650 estudiants, home, 26 anys, 2n semestre 2006, 3,3 assignatures amb 2,7% aprovades procedent d'estudis inacabats.

Clúster 9: 542 estudiants, dona, 28anys, 1r semestre 2001, 3,8 assignatures amb 6% aprovades procedent d'estudiants titulats.

Clúster 10: 594 estudiants, dona, 23-24 anys, 1r semestre 2006, 2,7 assignatures amb 7% aprovades procedent d'estudis inacabats.

Clúster 16: 1210 estudiants, dona, 34-36 anys, 2n semestre 1992, 2,9 assignatures amb 0% aprovades procedent d'estudis inacabats.

En tots aquests 4 clústers la majoria d'alumnes que s'han matriculat d'alguna de les 12 assignatures més comunes en el 1r semestre no s'han presentat.

El mètode d'agregació no ajuda a veure clarament algun atribut no evident perquè justifiqui l'abandonament dels estudis. Es clar que si no es presenta un alumne a les assignatures de les que s'ha matriculat, té molts números per deixar els estudis.

10.3.2 Associació

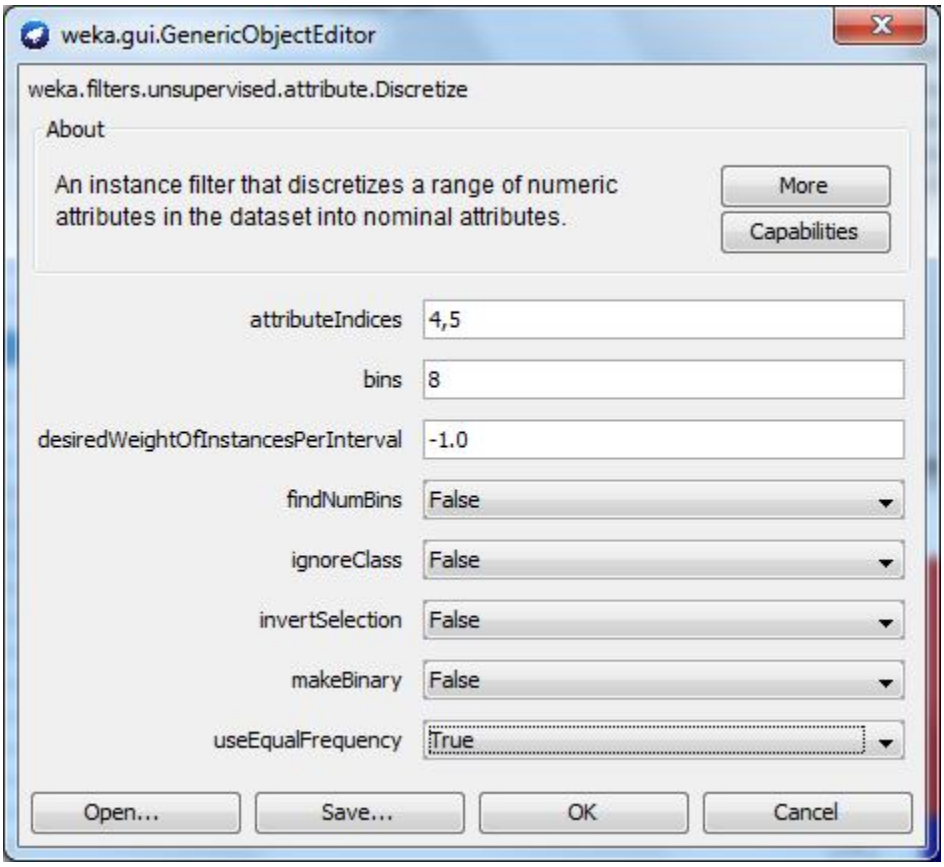
Associate / Apriori

L'algorisme **Apriori**, va agafant totes les combinacions d'atributs que compleixen el suport mínim indicat. Després es van generant les regles que compleixen els valors mínims de confiança.

A la pestanya **Associate** hi ha algorismes que permeten crear models d'associació. Molts d'ells només ho permeten si els atributs són **nominals**, per exemple l'algorisme **Apriori**.

Amb la penúltima reducció d'atributs en la que es van quedar 26 atributs, es discretitzaran els atributs numèrics**, NA, NASUP, PCTAS, NACMAT, NACPRE** i **NACSUP**. S'aplicarà per exemple els 10 intervals que hi per defecte en el filtre **Discretize** de la pestanya **Preprocess**.

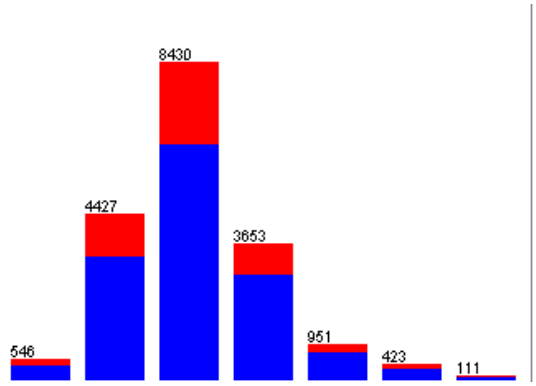
Aquesta discretització es farà per separat. Pels atributs **NA** (valors del 1 al 7) i **NASUP** (valors dels 0 al 7) es fan 8 intervals i d'aquesta manera cada interval m'agafa un únic número enter dels possibles. En el cas de **NA** només farà 7 intervals.



Selected attribute

Name: NA Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count
1	'(-inf-1.5]'	546
2	'(1.5-2.5]'	4427
3	'(2.5-3.5]'	8430
4	'(3.5-4.5]'	3653
5	'(4.5-5.5]'	951
6	'(5.5-6.5]'	423
7	'(6.5-inf)'	111

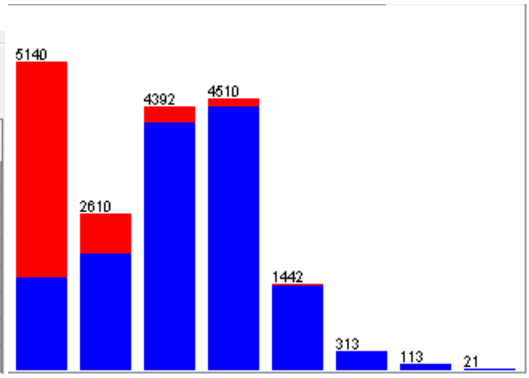


Es pot veure que el número d'assignatures matriculades no altera la proporció d'alumnes que abandonen els estudis (color vermell)

Selected attribute

Name: NASUP Type: Nominal
Missing: 0 (0%) Distinct: 8 Unique: 0 (0%)

No.	Label	Count
1	'(-inf-0.5]'	5140
2	'(0.5-1.5]'	2610
3	'(1.5-2.5]'	4392
4	'(2.5-3.5]'	4510
5	'(3.5-4.5]'	1442
6	'(4.5-5.5]'	313
7	'(5.5-6.5]'	113
8	'(6.5-inf)'	21



Aquí es veu que a mida que augmenta el número d'assignatures aprovades, va disminuint el percentatge d'alumnes que abandonen els estudis (color vermell)

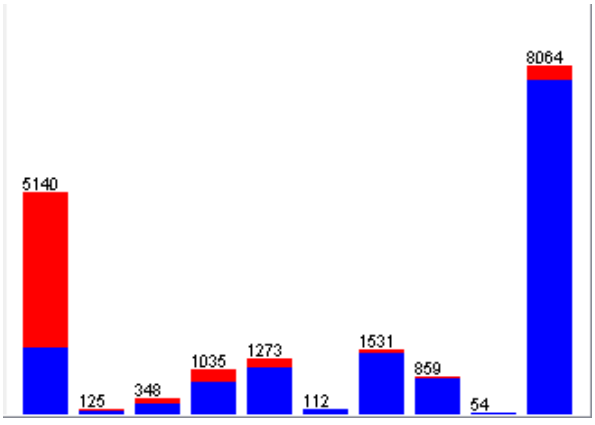
El mateix es fa pels atributs **NACMAT, NACPRE** i **NACSUP** ja que els valors màxims en aquest casos també és de 7.

Per l'atribut **PCTAS** es fan 10 intervals perquè coincideixi cada interval amb un 10%.

Selected attribute

Name: PCTAS Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count
1	'(-inf-0.1]'	5140
2	'(0.1-0.2]'	125
3	'(0.2-0.3]'	348
4	'(0.3-0.4]'	1035
5	'(0.4-0.5]'	1273
6	'(0.5-0.6]'	112
7	'(0.6-0.7]'	1531
8	'(0.7-0.8]'	859
9	'(0.8-0.9]'	54
10	'(0.9-inf)'	8064



Ara que no es té cap atribut numèric, s'aplica l'algorisme **Apriori** de la pestanya **Associate**

```
Apriori
=======
Minimum support: 0.85 (15760 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 6

Best rules found:
 1. A10S=NM 17474 ==> A12S=NM 16639    conf:(0.95)
 2. A9S=NM 17153 ==> A12S=NM 16323    conf:(0.95)
 3. A11S=NM 17481 ==> A12S=NM 16584    conf:(0.95)
 4. A9S=NM 17153 ==> A11S=NM 16186    conf:(0.94)
 5. A10S=NM 17474 ==> A11S=NM 16480    conf:(0.94)
 6. A12S=NM 17644 ==> A10S=NM 16639    conf:(0.94)
 7. A9S=NM 17153 ==> A10S=NM 16172    conf:(0.94)
 8. A11S=NM 17481 ==> A10S=NM 16480    conf:(0.94)
 9. A12S=NM 17644 ==> A11S=NM 16584    conf:(0.94)
10. A11S=NM 17481 ==> A9S=NM 16186    conf:(0.93)
```

S'ha deixat per defecte que faci 10 regles i es pot veure que no aporten informació pel que es vol cercar, l'abandonament dels estudis. Només aporta regles que emparellen assignatures de les quals no es matriculen els estudiants. Per exemple la 1ª regla indica que dels 17474 alumnes que no es matriculen hi ha 16639 que tampoc ho fan 16639 estudiants.

Veient aquests resultats, s'eliminaran els atributs A9S, A10, A11S i A12S. Ara els resultats són:

```
Apriori
=======
Minimum support: 0.35 (6489 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:
Size of set of large itemsets L(1): 15
Size of set of large itemsets L(2): 58
Size of set of large itemsets L(3): 57
Size of set of large itemsets L(4): 11

Best rules found:
 1. PCTAS='(0.9-inf)' 8064 ==> ABANDONA=NO 7739    conf:(0.96)
 2. PCTAS='(0.9-inf)' A8S=NM 7397 ==> ABANDONA=NO 7084    conf:(0.96)
 3. PCTAS='(0.9-inf)' A6S=NM 6846 ==> ABANDONA=NO 6535    conf:(0.95)
 4. PCTAS='(0.9-inf)' 8064 ==> A8S=NM 7397    conf:(0.92)
 5. PCTAS='(0.9-inf)' ABANDONA=NO 7739 ==> A8S=NM 7084    conf:(0.92)
 6. A1S=SI A6S=NM 8136 ==> A8S=NM 7433    conf:(0.91)
 7. A1S=SI A6S=NM ABANDONA=NO 7413 ==> A8S=NM 6771    conf:(0.91)
 8. A1S=SI 9877 ==> ABANDONA=NO 9016    conf:(0.91)
 9. NACMAT='(2.5-3.5]' 7412 ==> NA='(2.5-3.5]' 6763    conf:(0.91)
10. A1S=SI A7S=NM 7683 ==> ABANDONA=NO 7009    conf:(0.91)
```

Les 5 primeres regles parlen d'alumnes que aproven més 90% de les assignatures a les que s'han presentat i que **NO** abandonen, cosa previsible. Les regles 6, 7 i 10 relaciona alumnes que SI aproven l'assignatura A1 amb la no matriculació de les assignatures A6 i A7 i que **NO** abandonen.

La regla 8 si que diu alguna cosa més concreta, de 9877 alumnes que **SI** aproven l'assignatura A1, 9016 **NO** abandonen I la regla 9 diu que de 7412 alumnes que es matriculen de 3 assignatures de les més comunes del 1r semestre, 6763 no s'han matriculat de cap més.

Ara s'esborraran els atributs que en el moment de la matrícula no es poden conèixer. Es queda amb el SEXE, EDAT, SEMESTRE, NA, VIA, NACMAT i ABANDONA.

```
Apriori
=======
Minimum support: 0.1 (1854 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:
Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 38
Size of set of large itemsets L(3): 17
Size of set of large itemsets L(4): 2

Best rules found:
 1. NACMAT='(3.5-4.5]' 2887 ==> NA='(3.5-4.5]' 2692    conf:(0.93)
 2. NACMAT='(3.5-4.5]' ABANDONA=NO 2200 ==> NA='(3.5-4.5]' 2045    conf:(0.93)
 3. SEXE=DONA NACMAT='(2.5-3.5]' 3727 ==> NA='(2.5-3.5]' 3450    conf:(0.93)
 4. SEXE=DONA NACMAT='(2.5-3.5]' ABANDONA=NO 2804 => NA='(2.5-3.5]' 2585    conf:(0.92)
 5. NACMAT='(2.5-3.5]' 7412 ==> NA='(2.5-3.5]' 6763    conf:(0.91)
 6. NACMAT='(2.5-3.5]' ABANDONA=NO 5469 ==> NA='(2.5-3.5]' 4963    conf:(0.91)
```

Associate / PredictiveApriori

Ara es provarà amb l'algorisme **PredictiveApriori** que combina el suport i la confiança en una sola mida i fa augmentar el llindar del suport, el suport mínim va augmentant fins arribar al llindar màxim. Ha donat un total de 100 regles d'associació. S'analitzarà només les que les que associen atributs amb l'atribut **ABANDONA**. Queden 12 regles.

PredictiveApriori
==========
Best rules found:

16. EDAT=(37.5-inf) VIA=COU NACMAT='(3.5-4.5]' **ABANDONA=NO** 63 ==> NA='(3.5-4.5]' 63 acc:(0.99471)
20. EDAT=(33.5-37.5] NACMAT='(3.5-4.5]' **ABANDONA=SI** 55 ==> NA='(3.5-4.5]' 55 acc:(0.99456)
22. SEMESTRE=20001 VIA=NO_COU NACMAT='(2.5-3.5]' **ABANDONA=NO** 51 ==> NA='(2.5-3.5]' 51 acc:(0.99446)
45. SEMESTRE=20022 VIA=NO_COU NACMAT='(2.5-3.5]' **ABANDONA=NO** 81 ==> NA='(2.5-3.5]' 80 acc:(0.99376)
49. SEMESTRE=19991 NACMAT='(3.5-4.5]' **ABANDONA=NO** 78 ==> NA='(3.5-4.5]' 77 acc:(0.99356)
58. EDAT=(24.5-25.5] SEMESTRE=20012 NACMAT='(2.5-3.5]' **ABANDONA=NO** 32 ==> NA='(2.5-3.5]' 32 acc:(0.99334)
65. SEXE=HOME EDAT=(33.5-37.5] SEMESTRE=20042 VIA=EST_IN 28 ==> **ABANDONA=NO** 28 acc:(0.99281)
69. EDAT=(33.5-37.5] SEMESTRE=20022 VIA=COU 24 ==> **ABANDONA=NO** 24 acc:(0.992)
83. SEMESTRE=19991 NA='(1.5-2.5]' **ABANDONA=SI** 58 ==> NACMAT='(1.5-2.5]' 57 acc:(0.99096)
91. EDAT=(27.5-29.5] NACMAT='(4.5-5.5]' **ABANDONA=SI** 19 ==> NA='(4.5-5.5]' 19 acc:(0.99025)
94. EDAT=(33.5-37.5] NA='(4.5-5.5]' VIA=TITULAT 18 ==> **ABANDONA=NO** 18 acc:(0.98972)
99. SEXE=DONA EDAT=(37.5-inf) SEMESTRE=20011 17 ==> **ABANDONA=NO** 17 acc:(0.98909)

Amb aquestes associacions no es veu regla que indueixi a un possible comportament futur de tipologia d'estudiant que segons les dades estudiades es pugui predir la seva continuïtat en els estudis.

10.3.3 Classificació

Weka te diferents models de classificació. Es provaran els més comuns i dels 4 tipus possibles d'opcions que hi ha per testejar la validesa del model escollit, utilitzarem a vegades el **Use training set** que entrena el model amb totes les dades i llavors el torna aplicar amb les mateixes dades i altres vegades la validació creuada que té el **Cross-validation** que divideix les dades en un número de particions (**folds**), per defecte 10, i construeix el classificador amb n-1 parts i es proba amb la que queda. Això es va repetint amb cadascuna de les 10 particions.

Regles de classificació

A la sortida que dona s'ha de fixar en el **Detailed accuracy by class**

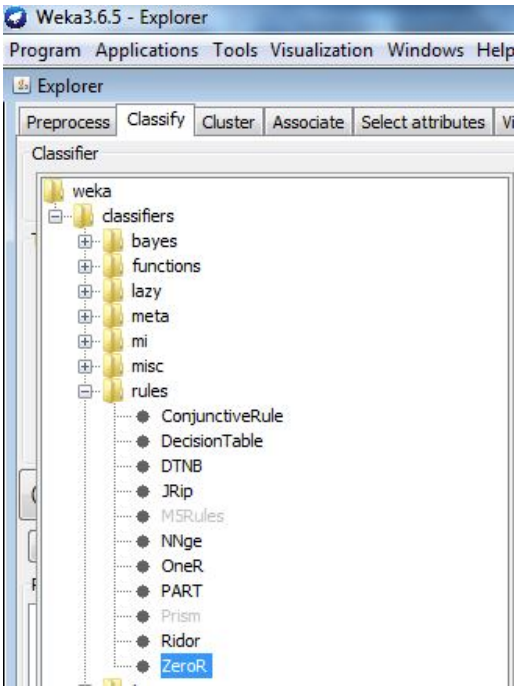
En **TP rate** (true positive) indica la proporció d'instàncies que són classificades correctament.

En **FP rate** (false positive) indica la proporció d'instàncies que estan mal classificades, realment pertanyen a altre classe.

A **Precision** indica la proporció que indica els elements ben classificats

Classify / rules / ZeroR

Es pot escollir, per exemple, l'opció per crear regles que està a la pestanya **Classify** , dintre de la carpeta **rules**. Dintre d'aquest apartat es tria **ZeroR** que classifica a totes les instàncies amb la classe majoritària



```
Test mode:evaluate on training data
=== Classifier model (full training set) ===

ZeroR predicts class value: NO

Time taken to build model: 0.01seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       13865               74.7359 %
Incorrectly Classified Instances      4687               25.2641 %
Kappa statistic                          0
Mean absolute error                      0.3776
Root mean squared error                  0.4345
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.747     1         0.855      0.5      NO
                 0         0          0         0         0          0.5      SI
Weighted Avg.    0.747     0.747      0.559     0.747     0.639      0.5

=== Confusion Matrix ===

     a     b   <-- classified as
 13865     0 |     a = NO
  4687     0 |     b = SI
```

Es pot observar que només ha classificat de forma correcta els estudiants que **NO** abandonen. A més hi ha un factor 1 de falsos positius. La predicció coincideix amb la classe majoritària dels que **NO** abandonen..

Classify / rules / ConjunctiveRule

Si es classifica amb l'algorisme **ConjunctiveRule**, elabora una regla que diu que els alumnes que percentualment aproven més d'un 30% dels crèdits matriculats, **NO** abandonen. Aquesta regla classifica un 83% dels estudiants però té un % d'error molt alt.

```
Test mode:evaluate on training data
=== Classifier model (full training set) ===

Single conjunctive rule learner:
--------------------------------
(PCTCS > 0.300685) => ABANDONA = NO

Class distributions:
Covered by the rule:
NO    SI
0.941506    0.058494
```

```
Not covered by the rule:
NO    SI
0.364728    0.635272

Time taken to build model: 0.3seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15560               83.8724 %
Incorrectly Classified Instances      2992               16.1276 %
Kappa statistic                          0.6156
Mean absolute error                      0.2289
Root mean squared error                  0.338
Relative absolute error                 60.6084 %
Root relative squared error             77.7773 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.835     0.151      0.942     0.835     0.886      0.842    NO
                 0.849     0.165      0.635     0.849     0.727      0.842    SI
Weighted Avg.    0.839     0.155      0.865     0.839     0.845      0.842

=== Confusion Matrix ===

     a     b   <-- classified as
 11583  2282 |     a = NO
   710  3977 |     b = SI
```

Classify / rules / JRip

Amb l'algorisme **JRip** s'obtenen 5 regles diferents. Classifica correctament un 86% dels alumnes.

Les 3 primeres regles justifiquen el fet de que **SÍ** abandonen els estudis els alumnes que:

- Aproven menys d'un 27% dels crèdits matriculats i menys del 14% dels crèdits matriculats i a més no es presenten a cap de les assignatures matriculades de les 12 més comuns en el 1r semestre (Això és vàlid per 4358 alumnes i no per 1221)
- Aproven menys del 14% dels crèdits matriculats i com a molt aprova 4,5 crèdits (és el valor mínim de crèdits d'una assignatura) i es presenta com a molt a 1 assignatura de les 12 més comuns en el 1r semestre i a més no aprova l'assignatura 1. (Això vol dir que si s'ha presentat a aquesta assignatura i l'ha suspès) . (Això és vàlid per 529 alumnes i no per 161)

- Només aproven l'assignatura 1 que és a la única que es presenta i que té 4,5 crèdits encara que aquesta assignatura representa menys del 30% dels crèdits matriculats. . (Això és vàlid per 554alumnes i no per 276)

Les 2 últimes regles són una única regla combinada que justifiquen a la vegada que **sí** abandonen 39 alumnes i **no** abandonen 13072 per aquells que només aproven 1 assignatura de 4,5 crèdits representant aquest fet menys d'un 14% de les assignatures matriculades i que no supera l'assignatura 4 i no es presenta o no matricula de les assignatures 1 i 3.

```
Test mode:evaluate on training data
=== Classifier model (full training set) ===

JRIP rules:
===========

(PCTCS <= 0.272727) and (NACPRE <= 0) and (PCTAS <= 0.142857) =>
ABANDONA=SI (4358.0/1221.0)
(NCSUP <= 4.5) and (PCTAS <= 0.142857) and (A1S = NO) and (NACPRE <=
1) => ABANDONA=SI (529.0/161.0)
(NCSUP <= 4.5) and (NACPRE <= 1) and (PCTCS <= 0.3) and (A1S = SI) =>
ABANDONA=SI (554.0/276.0)
(NASUP <= 1) and (NCSUP <= 4.5) and (PCTAS <= 0.142857) and (A3S =
NP_NM) and (A4S = NO) and (A1S = NP_NM) => ABANDONA=SI (39.0/15.0)
 => ABANDONA=NO (13072.0/880.0)

Number of Rules : 5
Time taken to build model: 1.63seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15999               86.2387 %
Incorrectly Classified Instances      2553               13.7613 %
Kappa statistic                          0.6549
Mean absolute error                      0.2112
Root mean squared error                  0.325
Relative absolute error                 55.9349 %
Root relative squared error             74.7909 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.879     0.188      0.933     0.879     0.905      0.851    NO
                 0.812     0.121      0.695     0.812     0.749      0.851    SI
Weighted Avg.    0.862     0.171      0.873     0.862     0.866      0.851

=== Confusion Matrix ===
     a     b   <-- classified as
 12192  1673 |     a = NO
   880  3807 |     b = SI
```

Classify / rules / OneR

L'algorisme **OneR** dona un resultat molt trivial ja que selecciona l'atribut que millor explica la classe que volem avaluar, en aquest car **ABANDONA**. Els alumnes que no superen cap assignatura en el 1r trimestre (<0,5 assignatures no pot ser) **SÍ** abandona els estudis i els que aproven com a mínim 1, **NO** abandonen els estudis. Aquesta classificació és vàlida per 15911 alumnes dels 18552 que hi ha en el fitxer de dades.

```
=== Classifier model (full training set) ===

NASUP:
      < 0.5 -> SI
      >= 0.5      -> NO
(15911/18552 instances correct)


Time taken to build model: 0.15seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15911               85.7643 %
Incorrectly Classified Instances      2641               14.2357 %
Kappa statistic                          0.6348
Mean absolute error                      0.1424
Root mean squared error                  0.3773
Relative absolute error                 37.6963 %
Root relative squared error             86.8305 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.888     0.233      0.918     0.888     0.903      0.828    NO
                 0.767     0.112      0.699     0.767     0.731      0.828    SI
Weighted Avg.    0.858     0.202      0.863     0.858     0.86       0.828

=== Confusion Matrix ===

     a     b   <-- classified as
 12317  1548 |     a = NO
  1093  3594 |     b = SI
```

Classify / rules / PART

L'algorisme **PART** dona un número molt alt de regles, 68 en total. Classifica correctament un total de 16008 alumnes però té condicions molt diversificades.

```
=== Classifier model (full training set) ===

PART decision list
------------------

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 2 AND NASUP > 3: NO (1896.0/21.0)
```

```
NASUP > 0 AND NCSUP > 4.5 AND NASUP > 2 AND NACPRE > 1 AND A1S = NP_NM AND A6S = SI: NO
(220.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 2 AND NACPRE > 1 AND A1S = NP_NM AND
A6S = NP_NM AND NACPRE <= 2: NO (148.0/1.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 2 AND NACPRE > 1 AND A1S = NP_NM AND
NACPRE <= 3 AND A4S = SI AND NCSUP <= 15: NO (128.0/2.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 2 AND PCTAS > 0.5 AND NACPRE > 1 AND
NACPRE <= 3 AND PCTAS > 0.875: NO (3092.0/79.0)

NASUP > 0 AND NCSUP > 4.5 AND NACPRE > 3: NO (341.0/4.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 1 AND PCTAS > 0.4 AND A6S = SI AND
A5S = NP_NM: NO (303.0/8.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 1 AND PCTCS > 0.4 AND A6S = NP_NM AND
A2S = NP_NM AND A3S = SI AND A4S = SI AND NCSUP <= 15: NO (153.0/4.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 1 AND PCTCS > 0.4 AND A6S = NP_NM AND
A2S = NP_NM AND A3S = SI AND A4S = NP_NM AND NASUP <= 2 AND PCTCS > 0.708333: NO
(519.0/23.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 1 AND PCTCS > 0.4 AND A6S = NP_NM AND A2S = NP_NM
AND A3S = NP_NM AND A4S = NP_NM AND A7S = SI AND A1S = SI: NO (484.0/29.0)

NASUP > 0 AND NCSUP > 4.5 AND NASUP > 1 AND PCTCS > 0.4 AND A6S = NP_NM AND A2S = NP_NM
AND A3S = NP_NM AND A7S = NP_NM AND A1S = SI: NO (1588.0/102.0)

NASUP > 0 AND NCSUP > 4.5 AND NACPRE > 2 AND NASUP > 1 AND PCTAS > 0.625 AND A4S =
NP_NM: NO (345.0/11.0)

NASUP > 0 AND NCSUP > 4.5 AND A2S = NP_NM AND A4S = NO: NO (85.0/4.0)

NASUP > 0 AND NCSUP > 4.5 AND A4S = SI AND A5S = NP_NM AND A7S = NP_NM AND NASUP <= 2:
NO (131.0/8.0)

NASUP > 0 AND NCSUP > 4.5 AND A4S = SI AND A3S = NP_NM: NO (85.0/3.0)

NASUP > 0 AND NCSUP > 4.5 AND A4S = NP_NM AND A3S = NP_NM AND NACPRE <= 2 AND
A6S = NP_NM AND A2S = NP_NM AND A1S = NP_NM AND NASUP > 1 AND NCSUP <= 15 AND A7S =
NP_NM: NO (302.0/12.0)

NASUP > 0 AND NCSUP > 4.5 AND A4S = NP_NM AND A3S = SI AND NASUP <= 2 AND A6S = NP_NM
AND A7S = NP_NM AND A5S = NP_NM AND A2S = NP_NM AND A1S = NP_NM AND PCTAS > 0.285714: NO
(142.0/12.0)

NASUP > 0 AND NCSUP > 4.5 AND A3S = SI AND A7S = NP_NM AND A4S = NP_NM AND NASUP <= 2
AND A6S = NP_NM AND NACPRE > 1 AND A5S = NP_NM AND A2S = NP_NM AND A1S = SI: NO
(129.0/17.0)

NASUP > 0 AND NCSUP > 4.5 AND A3S = SI AND A7S = NP_NM AND A4S = NP_NM AND NASUP <= 2
AND NACPRE > 1: NO (100.0/8.0)

NASUP > 0 AND NCSUP > 4.5 AND A3S = NP_NM AND A7S = SI AND A6S = NP_NM AND A1S = NP_NM
AND PCTAS > 0.666667: NO (91.0/5.0)

NASUP > 0 AND NCSUP > 4.5 AND A3S = SI AND PCTAS > 0.4 AND A4S = NP_NM: NO (43.0)

NASUP > 0 AND NCSUP > 4.5 AND A3S = NP_NM AND A2S = SI AND A6S = NP_NM AND NACPRE > 1
AND A1S = SI: NO (330.0/36.0)

NASUP > 0 AND NCSUP > 4.5 AND A3S = NP_NM AND A2S = SI: NO (116.0/12.0)

NASUP > 0 AND NCSUP > 4.5 AND A2S = NP_NM AND A3S = NP_NM AND A6S = SI AND NACPRE <= 2:
NO (48.0/1.0)

NASUP > 0 AND NCSUP > 4.5 AND A2S = NP_NM AND A3S = NP_NM AND A6S = NP_NM AND A7S = SI
AND A1S = NP_NM AND NASUP <= 1 AND A5S = NP_NM: NO (38.0/4.0)

NASUP > 0 AND NCSUP > 4.5 AND A2S = NP_NM AND A3S = NP_NM AND A6S = NO: NO (23.0/2.0)

NASUP > 0 AND NCSUP > 4.5 AND A2S = NP_NM AND A6S = NP_NM AND A3S = NO: NO (22.0)
```

```
NASUP > 0 AND A1S = NP_NM AND A3S = NP_NM AND A7S = NP_NM: NO (607.0/97.0)

NASUP <= 0 AND A3S = NO AND A1S = NP_NM: NO (50.0/9.0)

NASUP <= 0 AND A7S = NP_NM AND A6S = NP_NM AND A5S = NP_NM AND NACPRE <= 1 AND A2S =
NP_NM AND A4S = NP_NM AND NACPRE <= 0: SI (4357.0/1221.0)

NASUP > 0 AND NACPRE > 1 AND A6S = NO AND NACPRE <= 3: NO (40.0/2.0)

NASUP > 0 AND NACPRE > 1 AND A7S = SI AND A3S = NP_NM AND A2S = NP_NM AND A1S = NP_NM
AND NACPRE <= 2 AND PCTAS > 0.4: NO (37.0/4.0)

NASUP > 0 AND NACPRE > 1 AND A7S = SI AND A3S = NP_NM AND A1S = NO: NO (17.0/1.0)

NASUP > 0 AND NACPRE > 1 AND A6S = NP_NM AND PCTAS > 0.428571 AND A7S = NP_NM AND A2S =
NP_NM AND NACPRE <= 2 AND A5S = NP_NM: NO (87.0/10.0)

NASUP > 0 AND A7S = NO: NO (53.0/6.0)

NASUP > 0 AND PCTCS > 0.30137 AND A2S = NO: NO (29.0/2.0)

NASUP > 0 AND PCTCS > 0.30137 AND A6S = NP_NM AND A4S = SI AND PCTCS <= 0.75: NO (7.0)

NASUP > 0 AND PCTCS > 0.30137 AND A6S = NP_NM AND A1S = NO AND PCTAS <= 0.571429: NO
(47.0/6.0)

NASUP > 0 AND PCTCS > 0.32 AND A2S = NP_NM AND A6S = NP_NM AND A4S = NP_NM AND
A5S = NP_NM AND PCTAS > 0.833333: NO (272.0/62.0)

NASUP > 0 AND NACPRE > 1 AND NASUP <= 1 AND A6S = NP_NM AND NACPRE > 2: NO (80.0/13.0)

NASUP <= 0 AND A7S = NP_NM AND A6S = NO AND A4S = NP_NM: NO (30.0/8.0)

NASUP <= 0 AND A7S = NP_NM AND A5S = NP_NM AND A1S = NO AND A4S = NP_NM: SI
(547.0/170.0)

PCTCS > 0.32 AND A5S = NO: NO (54.0/11.0)

A7S = SI AND A1S = NP_NM AND A3S = NP_NM AND NASUP > 1 AND NCSUP > 11: NO (7.0/1.0)

A3S = NO AND PCTAS > 0.222222: NO (29.0/3.0)

A7S = SI AND A1S = NP_NM AND A3S = NP_NM AND NASUP > 1 AND PCTCS <= 0.666667 AND PCTAS
<= 0.571429: NO (3.0/1.0)

A7S = SI AND NACPRE > 1: NO (7.0)

PCTCS > 0.307692 AND A6S = NP_NM AND A4S = NP_NM AND A5S = NP_NM AND A1S = SI: NO
(401.0/127.0)

A7S = NO: NO (27.0/8.0)

A5S = NO AND A4S = NP_NM AND A3S = NP_NM AND A2S = NP_NM AND NACPRE > 1 AND
A1S = SI AND PCTAS > 0.285714: NO (25.0/5.0)

NACPRE > 1 AND A3S = NP_NM AND A5S = NP_NM AND A6S = NP_NM AND A2S = NP_NM AND A1S = SI:
NO (47.0/13.0)

A1S = NO AND A2S = SI: NO (11.0/3.0)

A1S = NO AND A2S = NP_NM AND A4S = NP_NM AND A3S = NP_NM AND NACPRE <= 1: NO (6.0/1.0)

A5S = NO: NO (52.0/20.0)

A1S = NO AND A4S = NP_NM AND NASUP <= 1 AND PCTCS > 0.181818: NO (8.0/2.0)

A7S = SI AND PCTCS <= 0.666667: SI (3.0/1.0)

A7S = NP_NM AND A6S = NP_NM AND A5S = SI AND NASUP <= 2 AND NCSUP <= 11: NO (3.0/1.0)

A7S = NP_NM AND A6S = NP_NM AND A5S = SI AND NASUP <= 2: SI (3.0/1.0)

A7S = NP_NM AND A6S = NP_NM AND A5S = NP_NM AND NACPRE > 1 AND NACPRE <= 2 AND A3S =
NP_NM AND NASUP > 0 AND PCTAS > 0.222222: NO (26.0/9.0)
```

```
A7S = NP_NM AND A6S = NP_NM AND A5S = NP_NM AND A4S = NO AND A1S = NP_NM: SI (30.0/13.0)

A7S = NP_NM AND NASUP <= 0 AND A6S = NP_NM: NO (52.0/20.0)

A7S = NP_NM AND A3S = SI AND A5S = NP_NM AND NASUP <= 2 AND PCTAS > 0.166667 AND PCTAS >
0.222222 AND PCTCS > 0.272727: NO (4.0/1.0)

A7S = NP_NM AND A3S = NP_NM AND A5S = NP_NM AND A1S = SI AND A6S = NP_NM: SI
(557.0/277.0)

A7S = NP_NM AND A3S = NP_NM AND A5S = NP_NM AND A6S = NO AND NASUP > 0: NO (7.0/2.0)

A7S = NP_NM AND A1S = NP_NM AND A5S = NP_NM AND NACPRE <= 2 AND NASUP > 0 AND PCTAS >
0.166667: SI (7.0/2.0)

NACPRE > 1 AND A1S = NP_NM AND A5S = NP_NM: SI (7.0/2.0)

A1S = NP_NM: NO (6.0)

: SI (8.0/1.0)

Number of Rules  :    68


Time taken to build model: 4.43seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       16008               86.2872 %
Incorrectly Classified Instances      2544               13.7128 %
Kappa statistic                          0.657
Mean absolute error                      0.2016
Root mean squared error                  0.3175
Relative absolute error                 53.3821 %
Root relative squared error             73.0643 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.878     0.183      0.934     0.878     0.905      0.899    NO
                 0.817     0.122      0.694     0.817     0.751      0.899    SI
Weighted Avg.    0.863     0.167      0.874     0.863     0.866      0.899

=== Confusion Matrix ===

     a     b   <-- classified as
 12177  1688 |     a = NO
   856  3831 |     b = SI
```

Arbres de classificació

Altres tipus de classificadors són els arbres. Es pot trobar diferents algorismes en la pestanya **Classify** dintre de la carpeta **trees**. A més de la sortida en format text, alguns poden visualitzar gràficament l'arbre de decisió representat.

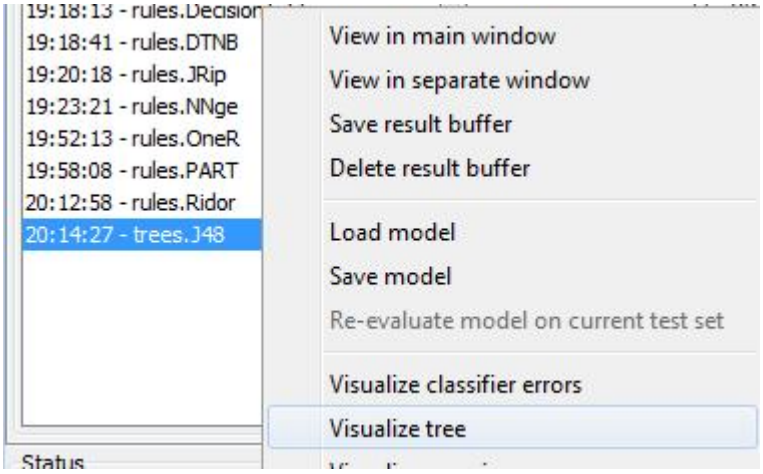
Classify / trees / J48

Algorisme J48. Aquest és un algorisme que genera una estructura de regles a partir de subconjunts extrets del conjunt total de dades d'entrenament. Es pot veure que la rel de l'arbre és el número d'assignatures aprovades. El resultat és el següent:

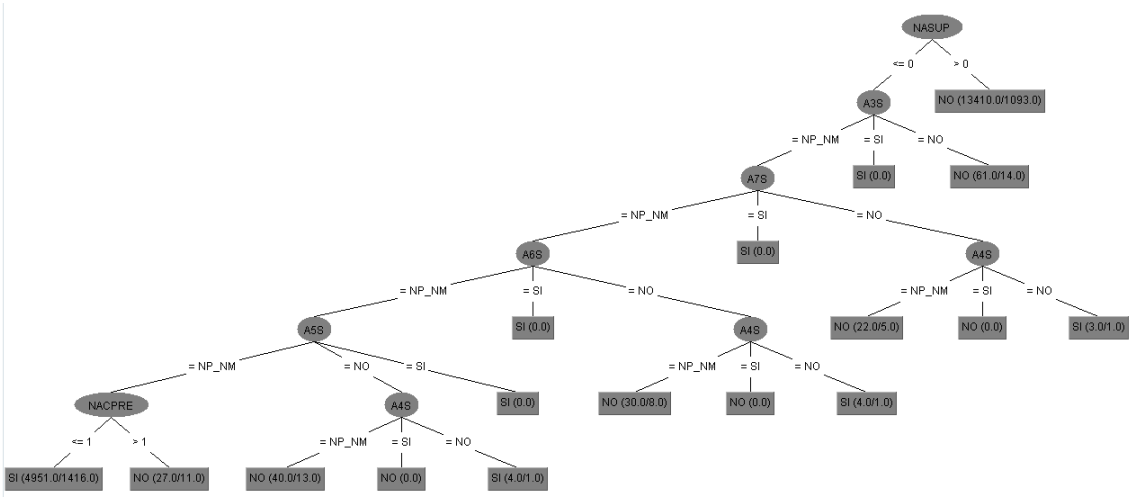
```
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

NASUP <= 0
|   A3S = NP_NM
|   |   A7S = NP_NM
|   |   |   A6S = NP_NM
|   |   |   |   A5S = NP_NM
|   |   |   |   |   NACPRE <= 1: SI (4951.0/1416.0)
|   |   |   |   |   NACPRE > 1: NO (27.0/11.0)
|   |   |   |   A5S = NO
|   |   |   |   |   A4S = NP_NM: NO (40.0/13.0)
|   |   |   |   |   A4S = SI: NO (0.0)
|   |   |   |   |   A4S = NO: SI (4.0/1.0)
|   |   |   |   A5S = SI: SI (0.0)
|   |   |   A6S = SI: SI (0.0)
|   |   |   A6S = NO
|   |   |   |   A4S = NP_NM: NO (30.0/8.0)
|   |   |   |   A4S = SI: NO (0.0)
|   |   |   |   A4S = NO: SI (4.0/1.0)
|   |   A7S = SI: SI (0.0)
|   |   A7S = NO
|   |   |   A4S = NP_NM: NO (22.0/5.0)
|   |   |   A4S = SI: NO (0.0)
|   |   |   A4S = NO: SI (3.0/1.0)
|   A3S = SI: SI (0.0)
|   A3S = NO: NO (61.0/14.0)
NASUP > 0: NO (13410.0/1093.0)

Number of Leaves  :         17

Size of the tree :  26


Time taken to build model: 0.99seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15989               86.1848 %
Incorrectly Classified Instances      2563               13.8152 %
Kappa statistic                          0.6411
Mean absolute error                      0.2213
Root mean squared error                  0.3327
Relative absolute error                 58.6045 %
Root relative squared error             76.5549 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.898    0.244      0.916     0.898     0.907      0.831    NO
                 0.756    0.102      0.714     0.756     0.734      0.831    SI
Weighted Avg.    0.862    0.208      0.865     0.862     0.863      0.831
```

```
=== Confusion Matrix ===

    a     b   <-- classified as
12446  1419 |     a = NO
 1144  3543 |     b = SI
```

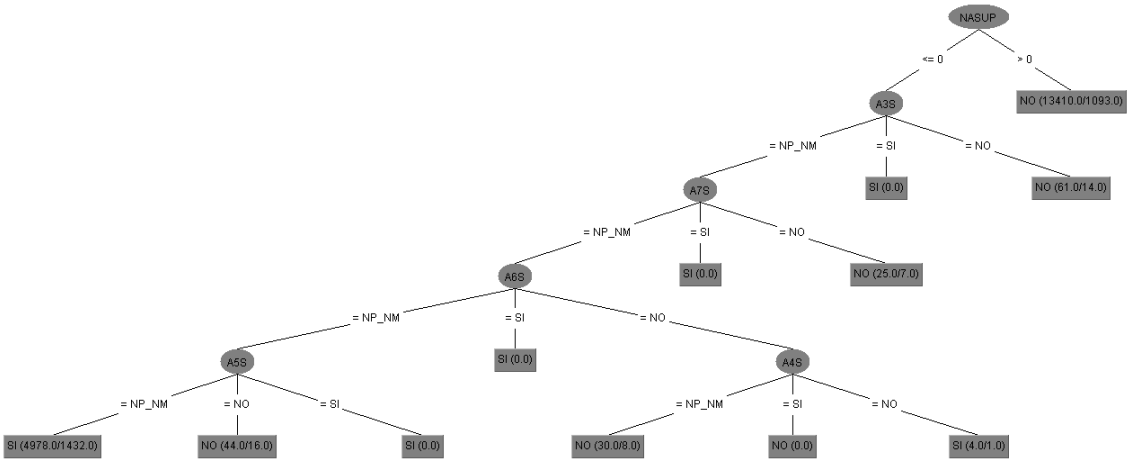
Amb el botó dret a sobre l'algorisme es selecciona l'opció **Visualize tree**



I s'obté el següent arbre:



Si es redueix molt el factor de confiança (per defecte està a **confidenceFactor**= 0,25), per exemple a 0,01, s'obté aquest altre arbre que està podat:



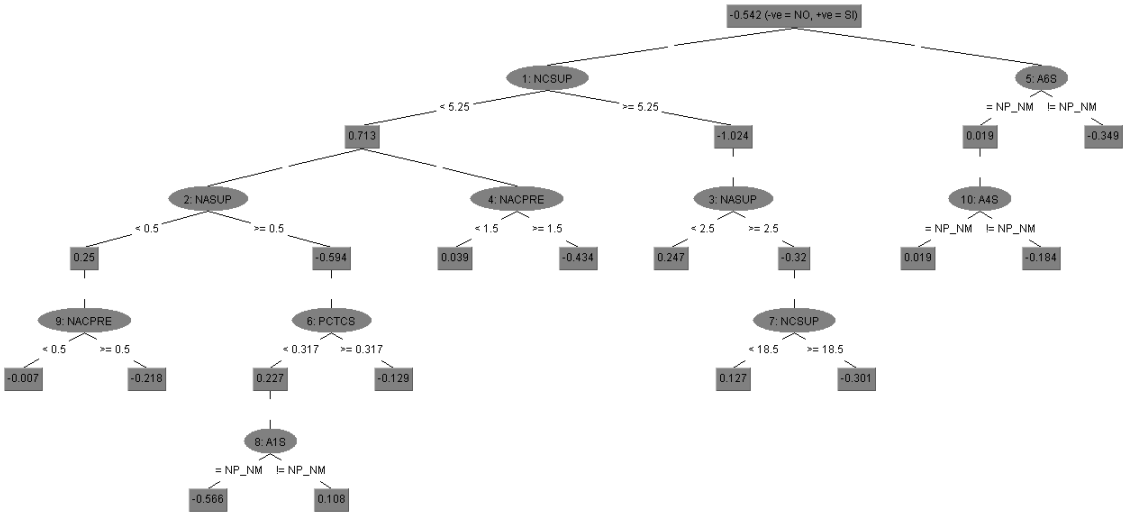

```
=== Summary ===
Correctly Classified Instances       15981               86.1417 %
Incorrectly Classified Instances      2571               13.8583 %
Kappa statistic                          0.6405
Mean absolute error                      0.2217
Root mean squared error                  0.333
Relative absolute error                 58.7098 %
Root relative squared error             76.6237 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.897     0.243      0.916     0.897     0.906      0.83     NO
                 0.757     0.103      0.712     0.757     0.734      0.83     SI
Weighted Avg.    0.861     0.208      0.865     0.861     0.863      0.83

=== Confusion Matrix ===
     a     b   <-- classified as
 12432  1433 |     a = NO
  1138  3549 |     b = SI
```

Classify / trees / ADTree

Algorisme ADTree

```
Test mode:evaluate on training data
=== Classifier model (full training set) ===
Alternating decision tree:
: -0.542
| (1)NCSUP < 5.25: 0.713
| | (2)NASUP < 0.5: 0.25
| | | (9)NACPRE < 0.5: -0.007
| | | (9)NACPRE >= 0.5: -0.218
| | (2)NASUP >= 0.5: -0.594
| | | (6)PCTCS < 0.317: 0.227
| | | | (8)A1S = NP_NM: -0.566
| | | | (8)A1S != NP_NM: 0.108
| | | (6)PCTCS >= 0.317: -0.129
| | (4)NACPRE < 1.5: 0.039
| | (4)NACPRE >= 1.5: -0.434
| (1)NCSUP >= 5.25: -1.024
| | (3)NASUP < 2.5: 0.247
```

```
| | (3)NASUP >= 2.5: -0.32
| | | (7)NCSUP < 18.5: 0.127
| | | (7)NCSUP >= 18.5: -0.301
| (5)A6S = NP_NM: 0.019
| | (10)A4S = NP_NM: 0.019
| | (10)A4S != NP_NM: -0.184
| (5)A6S != NP_NM: -0.349
Legend: -ve = NO, +ve = SI
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21

Time taken to build model: 2.44seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15962               86.0392 %
Incorrectly Classified Instances      2590               13.9608 %
Kappa statistic                          0.6385
Mean absolute error                      0.2877
Root mean squared error                  0.3397
Relative absolute error                 76.1708 %
Root relative squared error             78.1664 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895    0.242      0.916     0.895     0.906      0.894    NO
                 0.758    0.105      0.709     0.758     0.733      0.894    SI
Weighted Avg.    0.86     0.207      0.864     0.86      0.862      0.894

=== Confusion Matrix ===

     a     b   <-- classified as
 12410  1455 |     a = NO
  1135  3552 |     b = SI
```



Nova passada de models amb les dades només amb 23 atributs.

A partir d'ara el conjunt de dades seran el resultant d'haver eliminat atributs redundants o que no aportaven cap informació important. Tindrem 23 atributs.

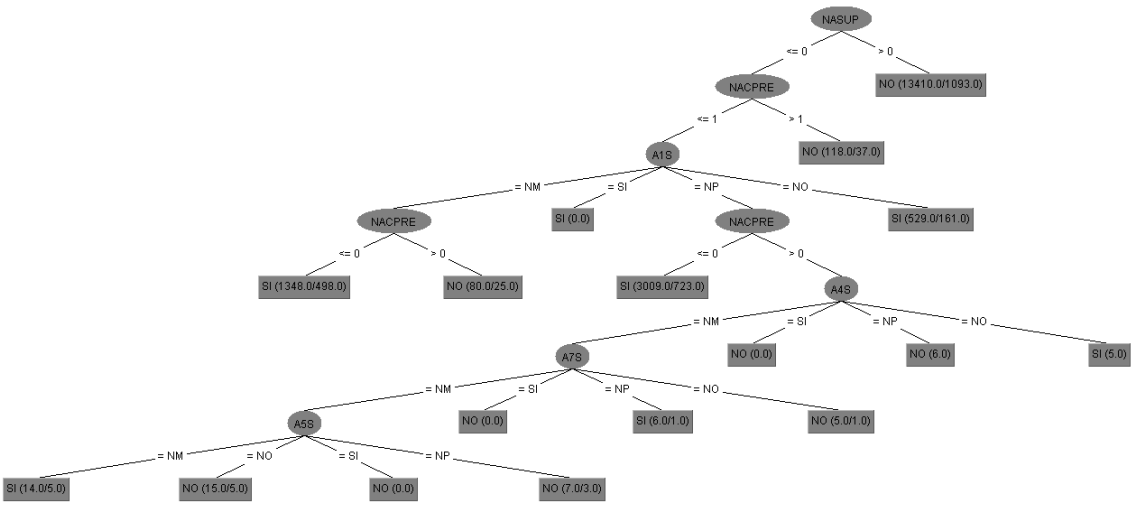
Al quedar-se amb 23 atributs, si es seleccionen els atributs més importants, queden 9 atributs més l'atribut referent ABANDONA. Repetint alguns algorismes de classificació amb aquests 10 atributs, obtenim aquests resultats.

Arbre J48

```
Scheme:weka.classifiers.trees.J48 -C 0.15 -M 2

Instances:18552
Attributes:10
              NASUP
              PCTAS
              NACPRE
              A1S
              A2S
              A3S
              A4S
              A5S
              A7S
              ABANDONA
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

NASUP <= 0
|   NACPRE <= 1
|   |   A1S = NM
|   |   |   NACPRE <= 0: SI (1348.0/498.0)
|   |   |   NACPRE > 0: NO (80.0/25.0)
|   |   A1S = SI: SI (0.0)
|   |   A1S = NP
|   |   |   NACPRE <= 0: SI (3009.0/723.0)
|   |   |   NACPRE > 0
|   |   |   |   A4S = NM
|   |   |   |   |   A7S = NM
|   |   |   |   |   |   A5S = NM: SI (14.0/5.0)
|   |   |   |   |   |   A5S = NO: NO (15.0/5.0)
|   |   |   |   |   |   A5S = SI: NO (0.0)
|   |   |   |   |   |   A5S = NP: NO (7.0/3.0)
|   |   |   |   |   A7S = SI: NO (0.0)
|   |   |   |   |   A7S = NP: SI (6.0/1.0)
|   |   |   |   |   A7S = NO: NO (5.0/1.0)
|   |   |   |   A4S = SI: NO (0.0)
|   |   |   |   A4S = NP: NO (6.0)
|   |   |   |   A4S = NO: SI (5.0)
|   |   A1S = NO: SI (529.0/161.0)
|   NACPRE > 1: NO (118.0/37.0)
NASUP > 0: NO (13410.0/1093.0)

Number of Leaves  :   17

Size of the tree :    25


Time taken to build model: 0.7seconds

=== Stratified cross-validation ===
=== Summary ===
```

```
Correctly Classified Instances       15957               86.0123 %
Incorrectly Classified Instances      2595               13.9877 %
Kappa statistic                          0.637
Mean absolute error                      0.2192
Root mean squared error                  0.3317
Relative absolute error                 58.0568 %
Root relative squared error             76.3395 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.246      0.915     0.896     0.905      0.839    NO
                 0.754     0.104      0.71      0.754     0.731      0.839    SI
Weighted Avg.    0.86      0.21       0.863     0.86      0.861      0.839

=== Confusion Matrix ===

     a     b   <-- classified as
 12423  1442 |     a = NO
  1153  3534 |     b = SI
```

El **confidenceFactor** que per defecte està a 0,25, s'ha baixat al 0,15 per tal que fes poda ja que si no sortia un arbre molt espès.



Es pot veure però, que no surt cap atribut que ja es conegui abans d'iniciar-se el semestre. Ja s'ha vist abans que si un alumne no supera cap assignatura té moltes possibilitats de deixar els estudis. Per això es torna a aplicar alguns models però ara amb les dades amb 23 atributs.

Algorismes de classificació.

Rules

Aplicant primer els algorismes de regles, es pot veure que l'algorisme **ZeroR**, el **ConjunctiveRule** i el **OneR** dona més o menys els mateixos resultats que quan es feia amb els atributs més representatius.

Rules JRip

En canvi amb el **JRip**, ja introdueix en les regles atributs que no sortien dintre dels més representatius com per exemple: **SEXE, VIA** i **EDAT**. També surt en les regles el número d'assignatures matriculades **NACMAT**.

Totes les regles que ha fet, justifiquen el fet que **SÍ** abandonen els estudis menys la última que també justifica els que **NO** abandonen:

- No aproven cap assignatura. (Regla previsible)
- Les dones que no tenen COU, s'han matriculat de 3 o més assignatures, s'han presentat a l'assignatura 1 i la han aprovada. (Cal fixar-se que el fet de **A1S=SI** implica que no hi ha possibilitat que **NASUP** i **NACPRE < 1**)
- Alumnes titulats que s'han matriculat de 3 o més assignatures, s'han presentat a l'assignatura 1 i la han aprovada.
- Alumnes entre 29,5 i 31,5 anys que s'han matriculat de 3 o més assignatures, s'han presentat a l'assignatura 1 i la han aprovada. (27 alumnes **SÍ** abandonen i 13194 **NO** abandonen)

```
Test mode:evaluate on training data

=== Classifier model (full training set) ===

JRIP rules:
===========

(NASUP <= 0) => ABANDONA=SI (5142.0/1548.0)
(NASUP <= 1) and (A1S = SI) and (NACMAT >= 3) and (NACPRE <= 1) and (SEXE = DONA) and
(VIA = NO_COU) => ABANDONA=SI (47.0/10.0)
(NASUP <= 1) and (A1S = SI) and (NACMAT >= 3) and (NACPRE <= 1) and (VIA = TITULAT) =>
ABANDONA=SI (142.0/47.0)
(NASUP <= 1) and (A1S = SI) and (NACPRE <= 1) and (NACMAT >= 3) and (EDAT = (29.5-31.5])
=> ABANDONA=SI (27.0/9.0)
 => ABANDONA=NO (13194.0/943.0)

Number of Rules : 5

Time taken to build model: 6.1seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15995               86.2171 %
Incorrectly Classified Instances      2557               13.7829 %
Kappa statistic                          0.6515
Mean absolute error                      0.2159
Root mean squared error                  0.3286
Relative absolute error                 57.1766 %
Root relative squared error             75.6165 %
Total Number of Instances            18552
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.884     0.201      0.929     0.884     0.906      0.842    NO
                 0.799     0.116      0.699     0.799     0.745      0.842    SI
Weighted Avg.    0.862     0.18       0.87      0.862     0.865      0.842

=== Confusion Matrix ===

     a     b   <-- classified as
 12251  1614 |     a = NO
   943  3744 |     b = SI
```

Rules PART

L'algorisme PART, amb els 23 atributs, ens dona moltes regles. No es copia el resultat, però es pot veure que classifica correctament el 90% dels alumnes

```
Number of Rules  :    645


Time taken to build model: 34.83seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       16811               90.6156 %
Incorrectly Classified Instances      1741                9.3844 %
Kappa statistic                          0.7463
Mean absolute error                      0.1481
Root mean squared error                  0.2722
Relative absolute error                 39.2278 %
Root relative squared error             62.6332 %
Total Number of Instances            18552

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.948     0.217      0.928     0.948     0.938      0.94     NO
                 0.783     0.052      0.835     0.783     0.808      0.94     SI
Weighted Avg.    0.906     0.175      0.905     0.906     0.905      0.94

=== Confusion Matrix ===

     a     b   <-- classified as
 13139   726 |     a = NO
  1015  3672 |     b = SI
```

Arbre J48

Ara es pot fer un arbre de classificació J48 i surt un arbre amb molts nodes i ramificacions. No es copia tot el contingut.

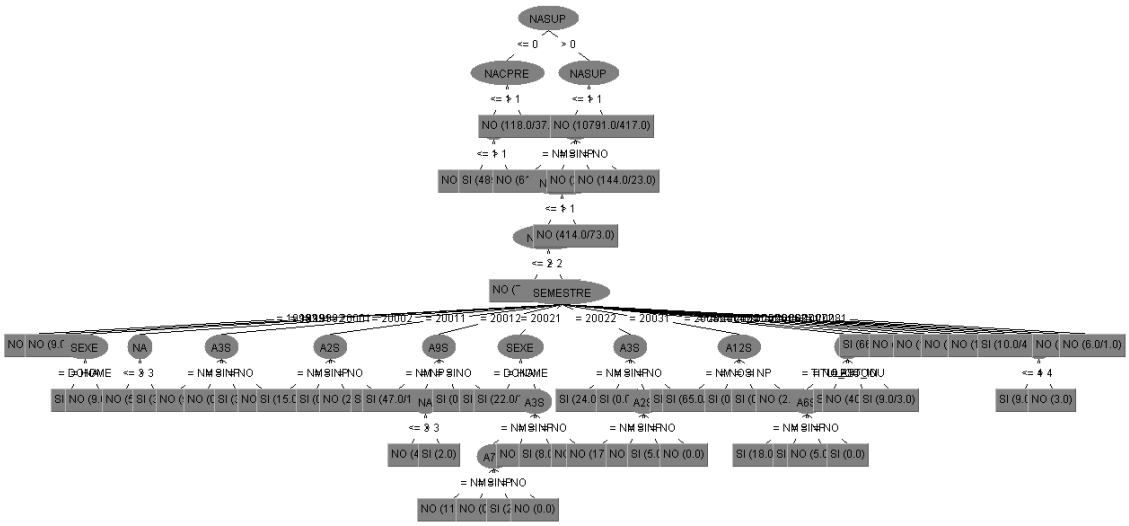
```
Number of Leaves  :   352

Size of the tree :    465


Time taken to build model: 2.07seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       15996               86.2737 %
```

```
Incorrectly Classified Instances      2545               13.7263 %
Kappa statistic                          0.6407
Mean absolute error                      0.1973
Root mean squared error                  0.3248
Relative absolute error                 52.2197 %
Root relative squared error             74.7288 %
Total Number of Instances            18541

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.902     0.254      0.913     0.902     0.908      0.873    NO
                 0.746     0.098      0.721     0.746     0.733      0.873    SI
Weighted Avg.    0.863     0.215      0.864     0.863     0.863      0.873

=== Confusion Matrix ===

     a     b   <-- classified as
 12502  1353 |     a = NO
  1192  3494 |     b = SI
```

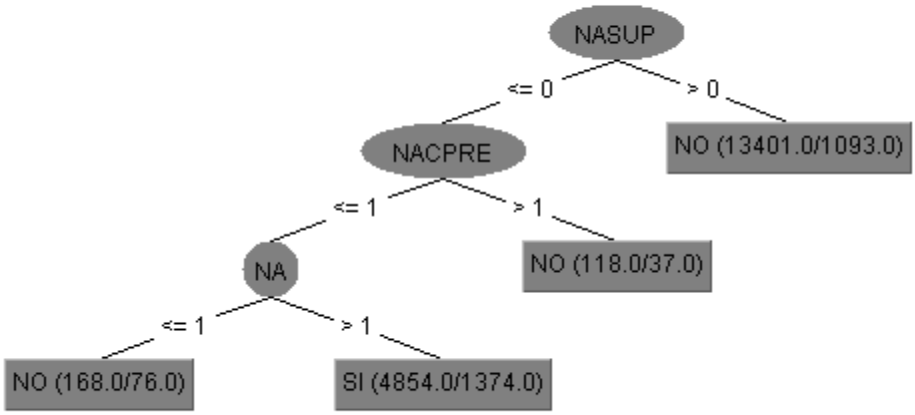
Si es baixa el factor de confiança, **ConfidenceFactor,** a 0,15 es realitza podes en l'arbre i s'óbté:

```
Number of Leaves  :    64
Size of the tree :    87

Time taken to build model: 1.54seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16024               86.4247 %
Incorrectly Classified Instances      2517               13.5753 %
Kappa statistic                          0.654
Mean absolute error                      0.2034
Root mean squared error                  0.3212
Relative absolute error                 53.8598 %
Root relative squared error             73.8999 %
Total Number of Instances            18541

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.889     0.21       0.926     0.889     0.907      0.876    NO
                 0.79      0.111      0.707     0.79      0.746      0.876    SI
Weighted Avg.    0.864     0.185      0.871     0.864     0.867      0.876

=== Confusion Matrix ===

     a     b   <-- classified as
 12322  1533 |     a = NO
   984  3702 |     b = SI
```

La visualització gràfica de l'arbre però, encara està molt atapeïda.

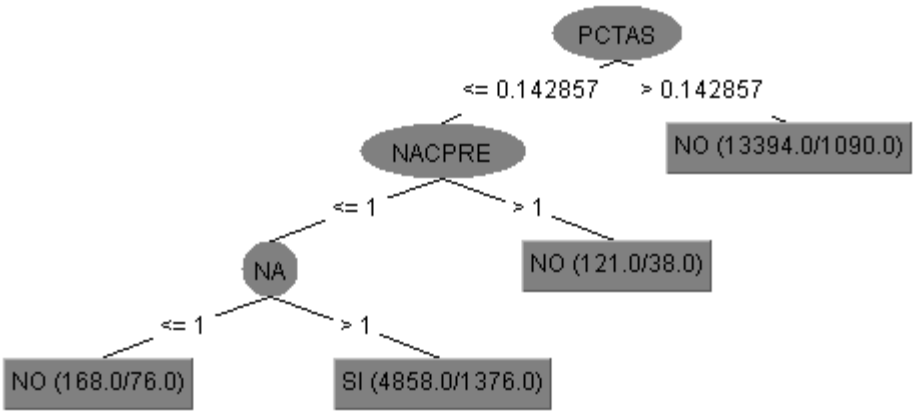


Fins que no es baixa el factor de confiança a 0,1, l'arbre es poda fins arribar a:

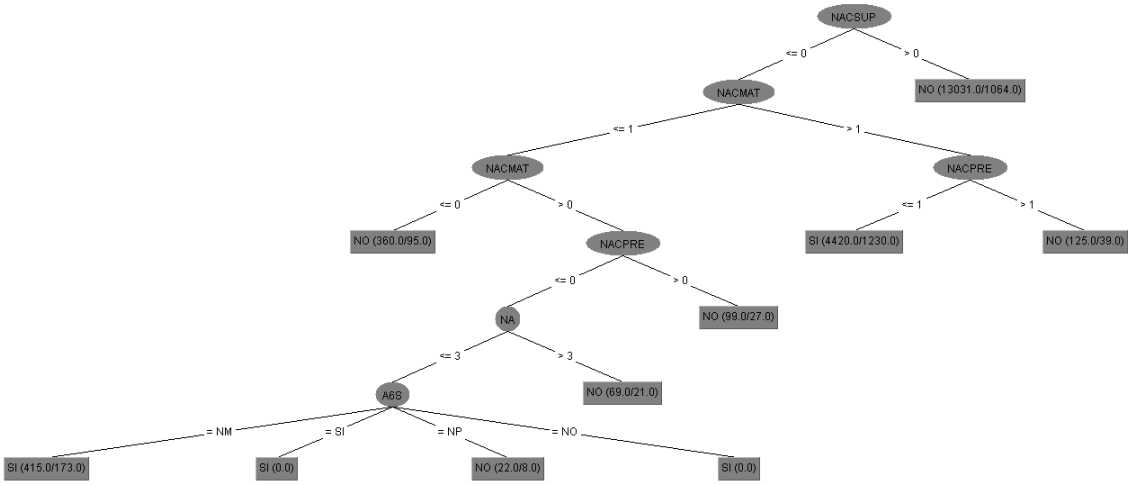


Ara es pot interpretar que és massa evident que si un alumne no supera cap assignatura, té molts números per abandonar els estudis. S'anirà eliminant els atributs que són evidents i es torna a aplicar el mateix arbre.

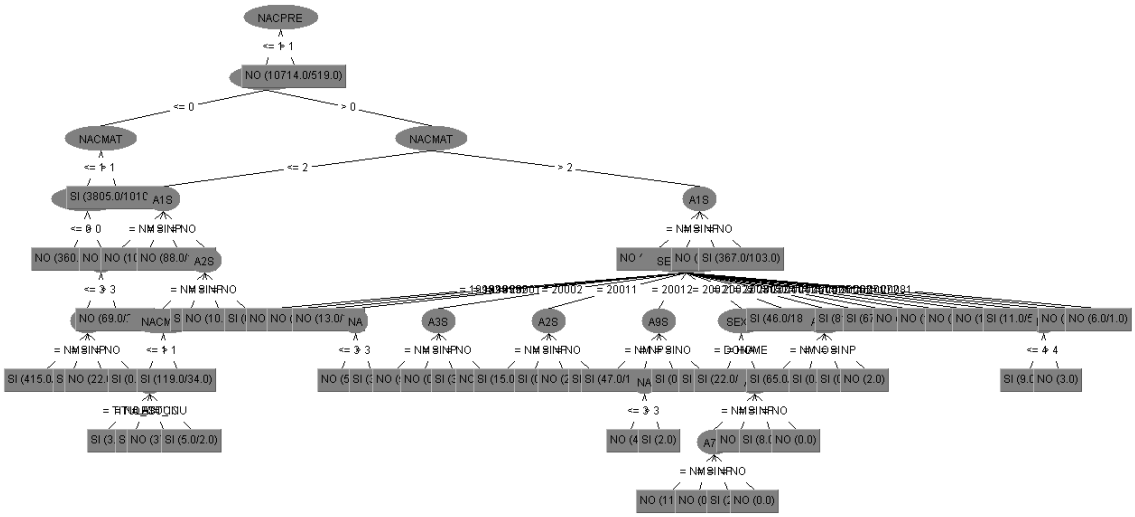
Atribut **NASUP** fora.



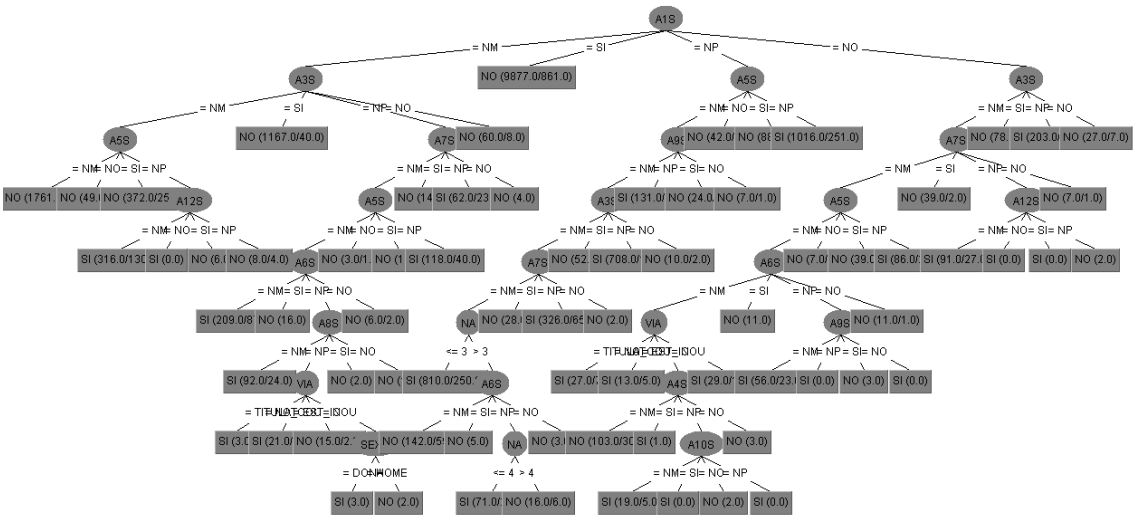
Atribut **PCTAS** fora.



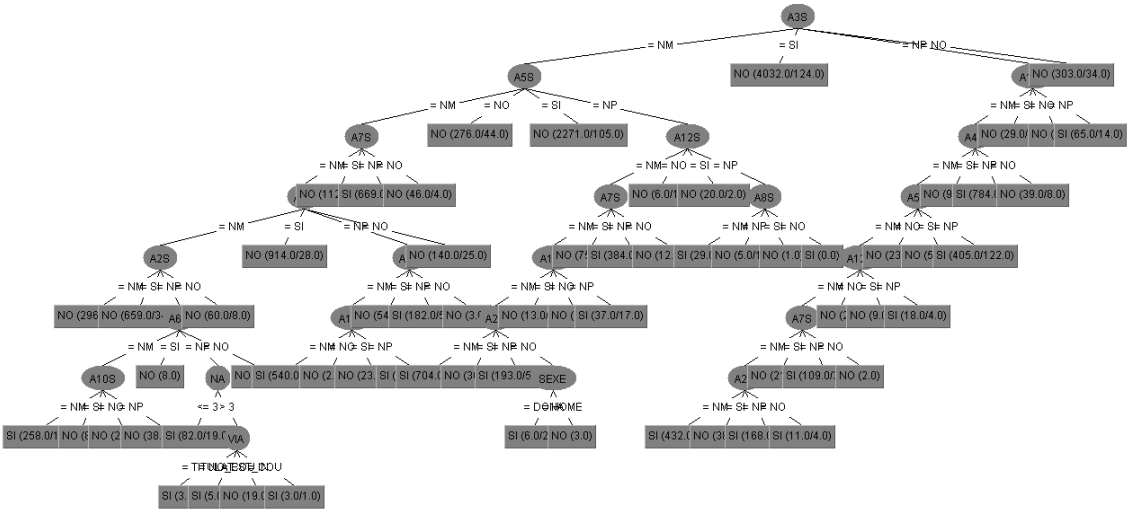
Atribut **NACSUP** fora.



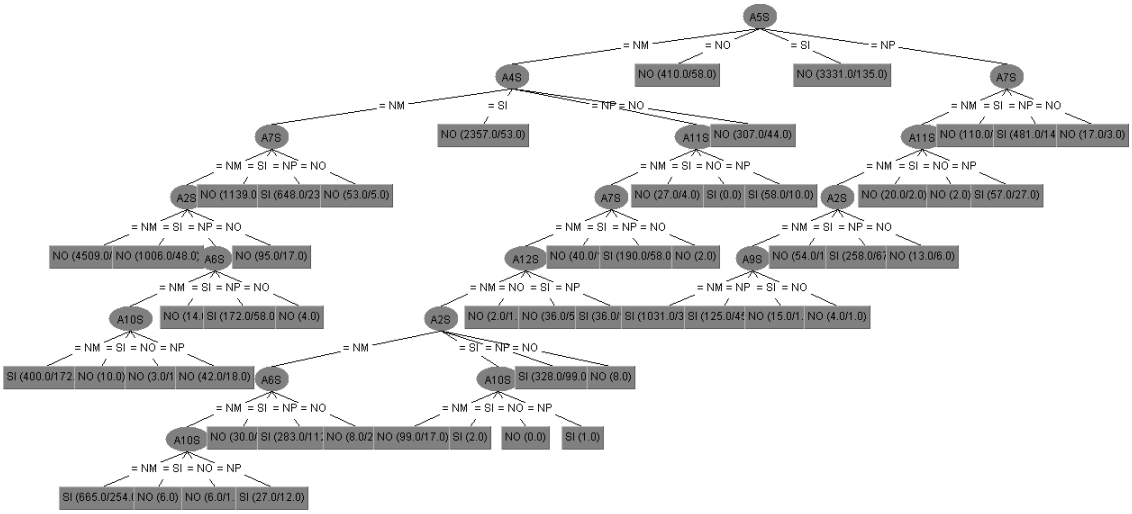
Atribut **NACPRE** fora.



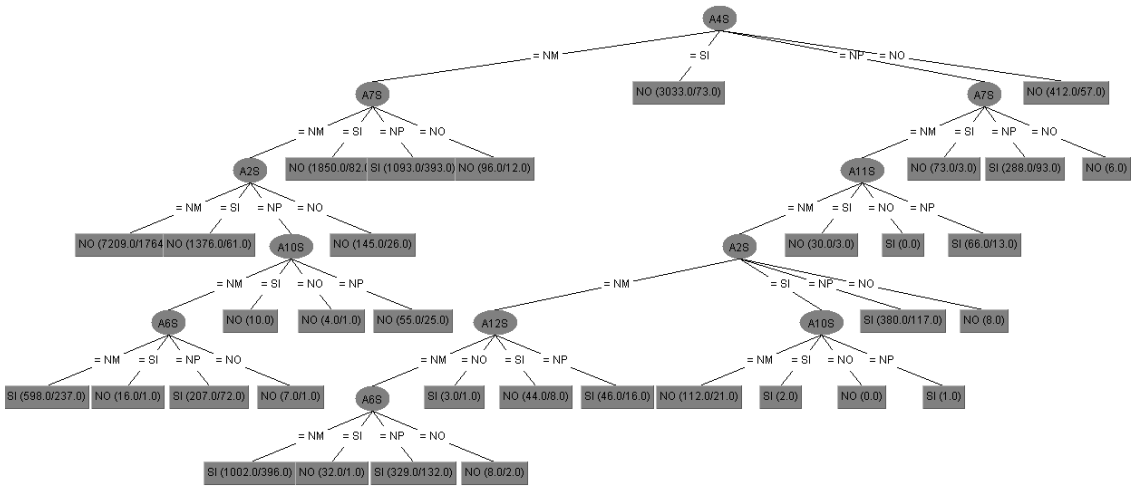
Atribut **A1S** fora.



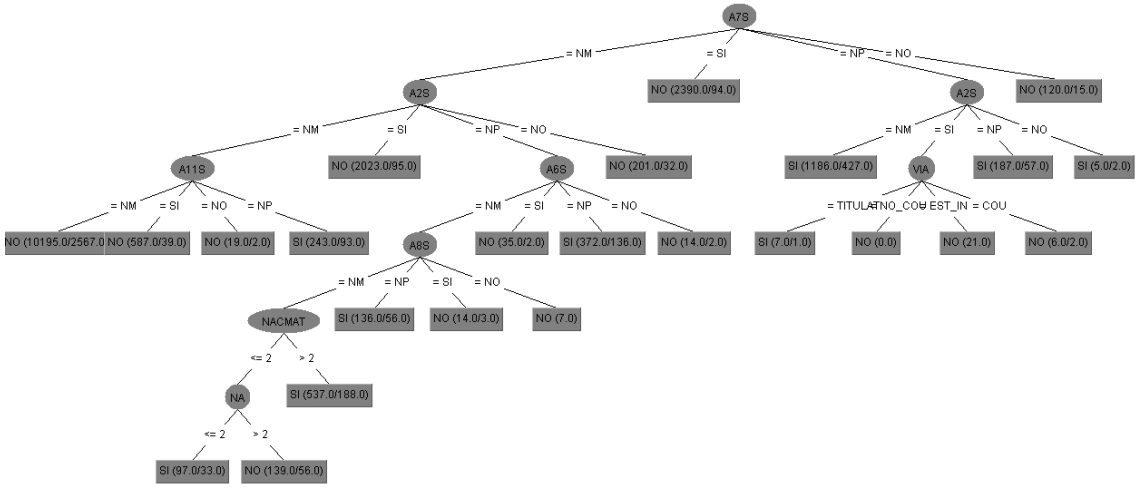
Atribut **A3S** fora.



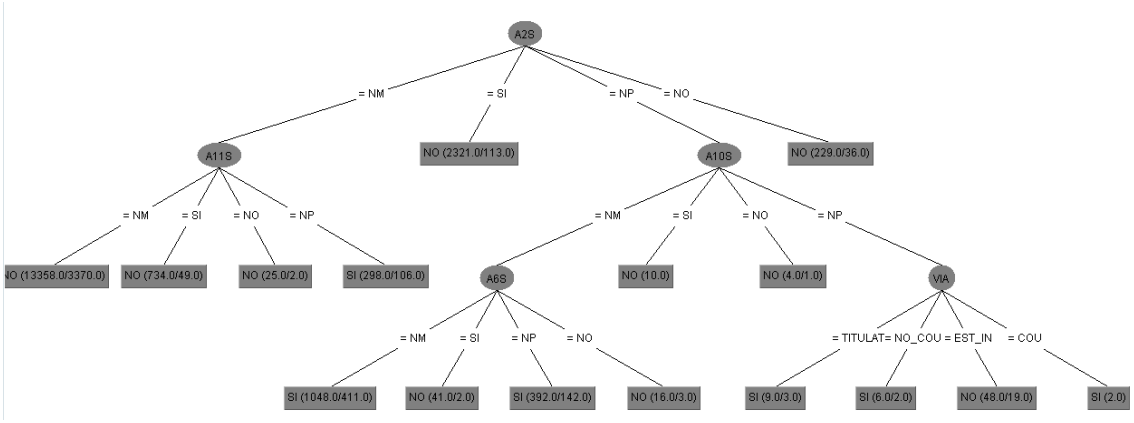
Atribut **A5S** fora.



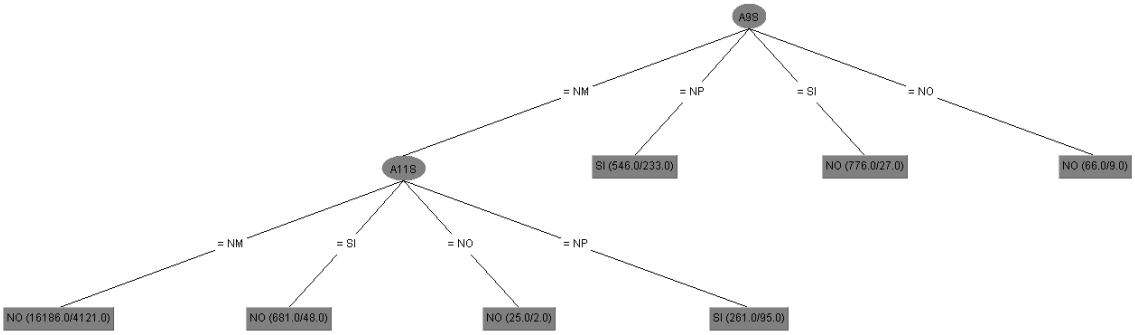
Atribut **A4S** fora.



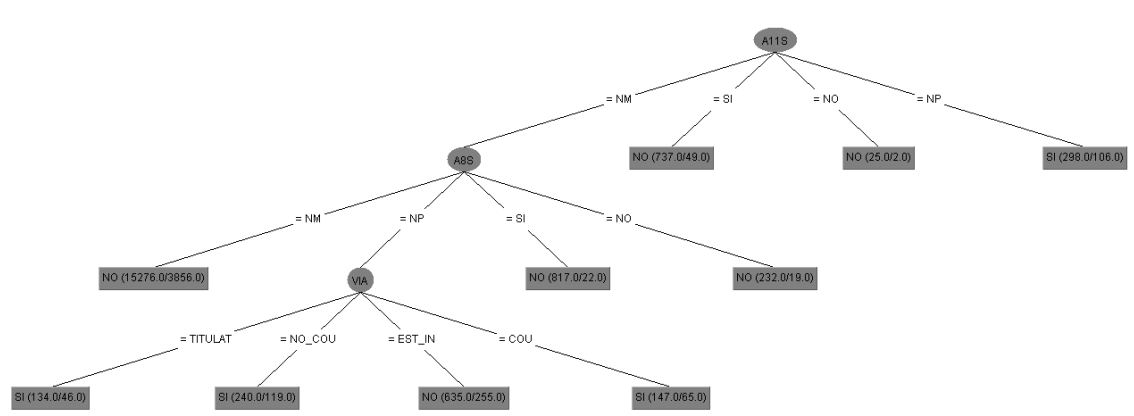
Atribut **A7S** fora



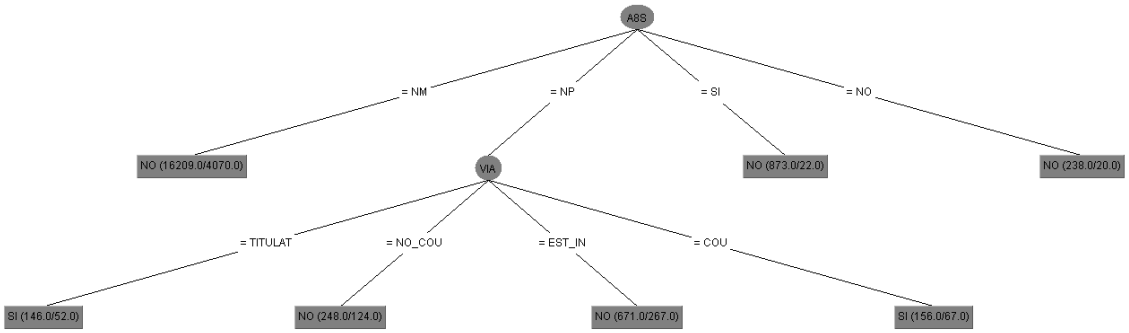
Atribut **A2S** fora



Atribut **A9S** fora



Atribut **A11S** fora



Canvi d'estratègia amb els atributs treballats

Ara en lloc de tenir en compte els resultats de les assignatures superades, tant de la matrícula de l'estudiant com de les més comunes en el 1r semestre, només es tindrà en compte l'informació que es pot generar en el moment de la matrícula.

Aquesta proposta està desenvolupada en l'apartat 4.3.