Memòria

TFG-Educational data mining and learning analytics

Joan Carles Vera Hernández
Grau en Enginyeria Informàtica
Treball fi de grau
Consultor: Ramón Caihuelas Quiles

Data: 9 de Gener de 2017

Resum

Aquest projecte consisteix en la construcció d'un sistema capaç de respondre un seguit de qüestions utilitzant algorismes estadístics i de mineria de dades.

Partint d'un arxiu que conté dades reals sobre alumnes matriculats a l'UOC en la seva primera matriculació, indicant si finalment han continuat els seus estudis a l'UOC, s'intenta dona una explicació a les qüestions següents:

- Quines combinacions d'assignatures són les més típiques?
- Quines son "letals"?
- Quines assignatures tenen més incidència en altres quan es matriculen conjuntament?
- Quina relació hi ha entre la matrícula feta i la decisió de tornar-se a matricular el segon semestre?
- Quines assignatures tenen més impacte en la decisió de tornar-se a matricular el segon semestres?

El projecte s'engloba dins del que s'anomena Educational Data Mining and Learning Analytics (EDM & LA), donat que el domini de les dades a analitzar pertanyen a l'àmbit de l'educació i que te com a objectius:

- detectar patrons de comportament en els alumnes
- predir necessitats de suport i atenció pels alumnes
- ajudar al personal docent a millorar el suport als estudiants
- desenvolupar noves ofertes pel pla d'estudis
- controlar la eficiència de les iniciatives adoptades pel que a fa a les noves ofertes i inscripcions.

Per al desenvolupament del treball s'ha utilitzat como a eina el programari R, molt utilitzat a la mineria de dades entre d'altres camps.

Índex de contingut

Índex d'il·lustracions

Índex de taules

1 Introducció

1.1 Data Mining

La mineria de dades o "Data Mining" (DM) és una disciplina utilitzada en l'àmbit de la computació i que té com a objectiu l'anàlisi de les dades contingudes en qualsevol tipus de magatzem de dades per a descobrir relacions, associacions, patrons en aquestes dades. És el que es coneix com "*Knowledge Discovery in Databases*" (KDD).

Per obtenir el 'coneixement' utilitza mètodes estadístics, d'intel·ligència artificial i d'aprenentatge computacional.

El propòsit final és poder mostrar el 'coneixement' obtingut de forma comprensible o bé poder integrar-lo en altres processos a fi de poder millorar-los o optimitzar-los.

1.1.1 Educational data mining

L'Educational data mining (EDM) és l'aplicació de la mineria de dades, aprenentatge computacional, intel·ligència artificial i l'estadística a la informació generada a partir dels entorns educatius. L'objectiu de l'EDM és desenvolupar i millorar els mètodes sobre com les persones aprenen en aquests contexts. El camp està estretament lligat al concepte de "*Learning Analitics*" (LA).

L'EDM extreu el coneixement a partir de repositoris de dades generades per les activitats d'aprenentatge de les persones en els centres educatius i les aprofita per a descobrir informació significativa sobre els diferents tipus d'alumnes i com aprenen.

Aquests anàlisis permeten finalment prendre decisions sobre com oferir i gestionar els recursos educatius i com interactuar amb ells.

Els recents avenços en la tecnologia han donat lloc a un major interès en desenvolupar tècniques per a l'anàlisi de grans quantitats de dades generades pels centres educatius.

L'objectiu és convertir les dades brutes recopilades en informació significativa sobre el procés d'aprenentatge per prendre decisions sobre el

disseny i la trajectòria dels model educatius. Aquest objectiu s'aconsegueix en varies fases:

- Descobrir relacions entre les dades. Els algorismes utilitzats per identificar aquestes relacions son:
  - Classificació
  - Regressió
  - Clustering
  - Anàlisi factorial
  - Anàlisi de xarxes socials
  - Mineria de regles d'associació
  - Mineria de patrons de seqüències
- Validació de les relacions per tal d'evitar el sobreajust.
- Aplicació de les relacions per fer prediccions sobre esdeveniments futurs.
- Integració. Les prediccions es fan servir per donar suport als processos de presa de decisions.

1.1.2 Learning Analytics

L'anàlisi de l'aprenentatge o "*Learning Analytics"* (LA), te com a objectiu bàsic la millora qualitativa, d'efectivitat i eficiència dels processos de l'aprenentatge.

Els avenços tecnològics i amb ells la possibilitat de manejar grans grups de dades units a l'interès d'incrementar el nivell educatiu dels ciutadans ens porten a la creació de disciplines com "*Learning Analytics"*.

LA intenta que l'experiència d'aprenentatge sigui més efectiva, accelerant el desenvolupament de competències i fins i tot incrementant la col·laboració entre estudiantes.

Les dades son l'element fonamental per a dur a terme l'anàlisi de l'aprenentatge, és necessari doncs recopilar, analitzar i presentar les dades sobre la activitat dels estudiants als entorns educacionals virtuals a on es produeix el procés. Les noves tecnologies fan que el volum de dades sigui cada vegada mes gran degut a les capacitats que inclouen els dispositius mòbils com ara càmeres fotogràfiques, sensors de moviments, GPS etc.

Les tecnologies mòbils han trencat amb la dependència dels espais físics permetent una mes gran interacció amb els sistemes d'aprenentatge.

Aquest factors han permès el desenvolupament del LA, amb la seva aplicació es pretén dissenyar un pla d'estudis personalitzat, adaptat i a la vegada poder predir amb l'objectiu final de millorar les capacitats dels estudiants.

Altres factors que han ajudat al l'avenç del LA han estat:

- El creixement del "big data"
- L'interès en augment de l'educació on-line

L'ús del LA permet aconseguir les següents millores en el procés educacional:

- Una millor presa de decisions administratives i una millor assignació de recursos.
- La predicció. Permet identificar els estudiants en risc d'abandonament o el possible fracàs o èxit d'una formació.
- La intervenció, aplicant mesures correctives per donar suport als estudiants que no rendeixen com s'esperava.
- La personalització i adaptació, per proporcionar als estudiants itineraris d'aprenentatge a mida, o materials d'avaluació.
- Donar la capacitat als estudiants per monitoritzar el seu propi progrés.

1.2 Estat actual

Estat de l'art és una expressió anglesa, "State of the art", que s'utilitza per a descriure les últimes novetats relatives a un àmbit concret, especialment en els àmbits del coneixement i l'enginyeria.

L'àmbit concret de l'estudi es basa en la mineria de dades (Data Mining) i mes concretament en el "*Educational Data Mining and Learning Analytics*", podríem dir que seria aplicar les tècniques de mineria de dades a l'àmbit de la educació.

L'estudi en concret es podria definir com l'aplicació de "*Educational Data Mining and Learning Analytics*" per a analitzar la influencia del procés de matriculació en la continuïtat dels estudis.

L'objectiu d'aquest treball és poder determinar a partir de dades real de matriculacions realitzades per alumnes de la Universitat Oberta de Catalunya, com la tria d'assignatures en el procés de matriculació i el resultat final del curs, pot influir en la decisió de continuar o no amb els estudis iniciats.

Volem saber quins altres estudis s'han fet relacionats amb aquest propòsit.

Hi ha varis estudis al respecte que relacionen el procés de matriculació amb la decisió d'abandonaments dels estudis:

Extracting highly positive association rules from students' enrollment data
L'estudi es basa en trobar regles d'associació per a oferir els programes d'estudi mes adients per a cada estudiant. No basant-se en la disponibilitat dels programes sinó en els interessos del estudiant. S'intenta evitar problemes com el poc interès en el curs i baix rendiment acadèmic.
Predicting Academic Success from Student Enrolment Data using Decision Tree Technique
Exemple de com l'EDM s'utilitza per a predir el rendiment dels alumnes i poder prendre mesures correctives per millorar el rendiment als exàmens. Una eina que pretén millorar la qualitat del centres formatius privats per a atreure els estudiants en un ambient de molta competència.
Predicting Students Drop Out: A Case Study
En aquest treball es descriuen els resultats d'un l'estudi EDM dirigit a la predicció de l'abandonament d'estudiants d'Enginyeria elèctrica (EE) després del primer semestre així com la identificació dels factors d'èxit específics pel programa d'EE. Els resultats mostren que utilitzant arbres de decisió s'obtenen precisions entre 75 i 80%.

Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression
Treball fet amb eines de mineria de dades, centrat en l'objectiu d'aconseguir que els estudiants finalitzin els seus estudis, identificant el moment de la matriculació com el focus principal de la investigació en l'educació superior.
Datamining based decision support system for students
Aquest és un sistema de recomanació basat en l'algoritme d'arbre de decisió ID3. El sistema proposa als estudiants, la llista de cursos que són preferibles per a ells en funció de les seves actuacions en cursos anteriors.
DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE
Estudi per a la creació d'un model base per al recolzament a la presa de decisions als sistemes d'educació superior basat en les dades recollides a partir de les enquestes dutes a terme durant el semestre d'estiu a la Facultat d'Econòmiques de la Universitat de Tuzla, curs 2010-2011, entre els estudiants de primer any durant la matriculació i avaluat amb la qualificació en l'examen.
Data Mining and Knowledge Management in Higher Education - Potential Applications.
Aquest projecte presenta un treball de mineria de dades per predir la possibilitat de continuar estudiant dels els estudiants actualment matriculats en una escola comunitària a Silicon Valley. El projecte aplica tècniques de xarxes neuronals i els algoritmes C&RT i C5.0, per a triar la millor predicció seguida d'una anàlisi de clustering utilitzant TwoStep. La predicció gracies a l'EDM dona a la universitat l'oportunitat d'actuar abans que un estudiant abandoni.
Assisting Higher Education in Assessing, Predicting, and Managing Issues Related to Student Success: A Web-based Software using Data Mining and Quality Function Deployment
Desenvolupament d'un programari per ajudar en l'avaluació i la predicció de l'èxit dels estudiants en la educació superior.

El programari utilitza algoritmes de mineria de dades i eines com ara "*Quality Function Deployment"* (QFD) per analitzar i predir temes com ara la matriculació, taxa d'abandonament, temps per a graduar-se, i suggereix com millorar els cursos i els programes.

2 Justificació i objectius del TFG.

La aplicació de la mineria de dades en l'àmbit de l'anàlisi de l'aprenentatge en fonamenta en els següents objectius:

- Descobrir relacions entre les dades.
- Aplicar les relacions per fer prediccions sobre esdeveniments futurs.
- Utilitzar les prediccions per a donar suport a la de presa de decisions.

En el cas concret d'aquest treball basat en aquest elements, s'estudien els patrons de matriculació d'un conjunt d'estudiants a uns estudis oferts per la Universitat Oberta de Catalunya (UOC). Aquest patrons es contrasten amb al resultat de la continuïtat dels estudis dels estudiants.

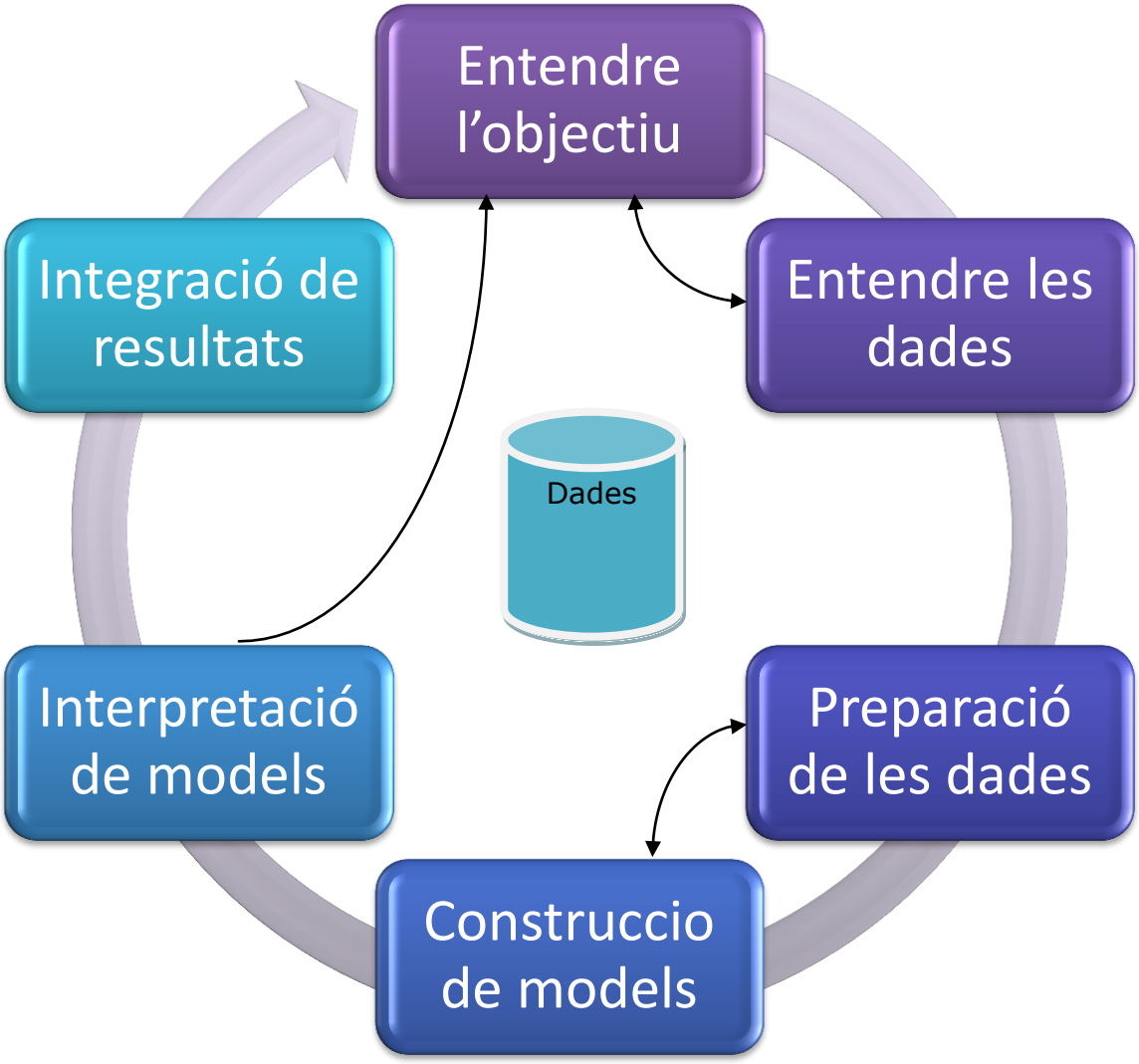
L'objectiu per tant és descobrir si existeix una relació entre la decisió de continuar els estudis i els patrons mencionats.

A partir de la informació descoberta es pot predir durant els procés de matriculació quins casos tenen una alta probabilitat de fracàs.

Els processos de matriculació i seguiment dels estudiants poden ser millorats si les prediccions son integrades en aquest processos per a intentar transformar el possible fracàs en un èxit, fomentant que l'alumne decideixi continuar els seus estudis fins a la graduació, evitant així l'abandonament dels mateixos.

3 Enfocament i mètode seguit.

El enfocament utilitzat està basat en la metodologia CRISP-DM que defineix la següent seqüencia:



Il·lustració 1. Cicle de vida - Metodología CRISP-DM

3.1 Cicle de vida

La metodologia CRISP-DM, contempla les següents fases:

- Entendre l'objectiu
- Entendre les dades
- Preparació de les dades
- Construcció de models
- Avaluació i interpretació del model
- Integració dels resultats en procés
- Observacions finals

3.1.1 Entendre l'objectiu

Aquesta és la primera fase i requereix la comprensió dels objectius del projecte, però no des de la perspectiva tècnica sinó funcional, La finalitat és entendre de la manera més completa el problema que es vol resoldre, això ens permetrà transformar el problema en un problema de mineria de dades.

3.1.2 Entendre les dades

Aquesta fase comprèn el contacte inicial amb les dades amb l'objectiu per a familiaritzar-se amb elles, analitzar la seva qualitat i establir les primeres hipòtesis.

3.1.3 Preparació de les dades

En aquesta fase es preparen les dades per adaptar-les a les tècniques de mineria de dades que es volen utilitzar en fases posteriors.

La preparació de les dades inclou neteja de dades, generació de variables addicionals, integració de diferents orígens de dades i canvis de format.

Les fases de preparació i de modelatge interactuen de forma permanent doncs depenent del model a implementar les dades requeriran un format diferent.

3.1.4 Construcció de models

En aquesta fase es decideix quines son les tècniques de modelatge més apropiades per al projecte.

Els paràmetres utilitzats en la generació del model depenen de les característiques de les dades i de les característiques de precisió que es vol aconseguir amb el model.

3.1.5 Avaluació i interpretació del model

Si el model generat és vàlid en funció dels criteris d'èxit establerts en la fase anterior, es procedeix a l'explotació del model.

Si cap dels models aconsegueix els resultats esperats, s'ha d'alterar algun dels passos anteriors per generar nous models.

3.1.6 Integració dels resultats en procés

En aquesta fase, el coneixement obtingut es transforma en accions dins del procés de negoci, es tractaria de inserir o implementar el coneixement obtingut amb el nostre model dins els sistemes d'informació de la empresa.

Hi ha una iniciativa per part del Data Mining Group, per estandarditzar el llenguatge PMML (Predictive Model Markup Language), de manera que els principals fabricants de bases de dades puguin fer us d'aquest estàndard.

3.1.7 Observacions finals

Aquest no és un procés lineal, queda clar que un cop avaluat el model ens plantegem si hem aconseguit l'objectiu plenament i tornem a començar el procés. Be per que l'objectiu pot haver canviat be per que no satisfem l'objectiu proposat inicialment.

3.2 Eines utilitzades

Existeixen moltes eines actualment de programari per al desenvolupament de models de mineria de dades. Exemples:

- dVelox de APARA
- KXEN
- KNIME
- Neural Designer
- OpenNN
- Orange

- Powerhouse
- Quiterian
- RapidMiner
- R
- SPSS Clementine
- SAS Enterprise Miner
- STATISTICA Data Miner
- Weka
- KEEL

En aquest treball s'ha utilitzat únicament RStudio. RStudio és un IDE (entorn de desenvolupament integrat) per a R, que és un entorn i llenguatge de programació amb un enfocament a l'anàlisi estadística.

R és una implementació de programari lliure del llenguatge S, i ha esdevingut un dels llenguatges més utilitzats en investigació per la comunitat estadística, mineria de dades, investigació biomèdica, bioinformàtica i matemàtiques financeres.

RStudio s'ha d'instal·lar conjuntament amb el programari estadístic R. Permet editar i executar el codi R, disposa d'eines per al traçat, la depuració del codi i la gestió de l'espai de treball.

Existeix una amplia varietat de biblioteques amb funcionalitats gràfiques i de càlcul disponibles a internet que qualsevol usuari pot descarregar i instal·lar de forma molt simple.

4 Planificació

La planificació d'aquest TFG ve donada per les fites definides al Pla docent de la assignatura.

4.1 Fites principals

Es mostra una taula a continuació que conte les fites definides amb la seva data de inici i final:

Fita	Data Inici	Data Final	Descripció
Inici TFG	21/09/2016	26/09/2016	Lectura pla docent
PAC1	26/09/2016	10/10/2016	Elaboració del pla de treball
PAC2	10/10/2016	07/11/2016	Estat de l'art EDM & LA
PAC3	07/11/2016	19/12/2016	Disseny i implementació
Lliurament	19/12/2016	09/01/2017	Lliurament final de TFG
Debat	16/01/2017	20/01/2017	Tribunal

Taula 1. Fites principals

Aquestes fites s'han de complir obligatòriament i constitueixen en general les dates a on s'ha de lliurar cada una de les parts del projecte.

4.2 Tasques

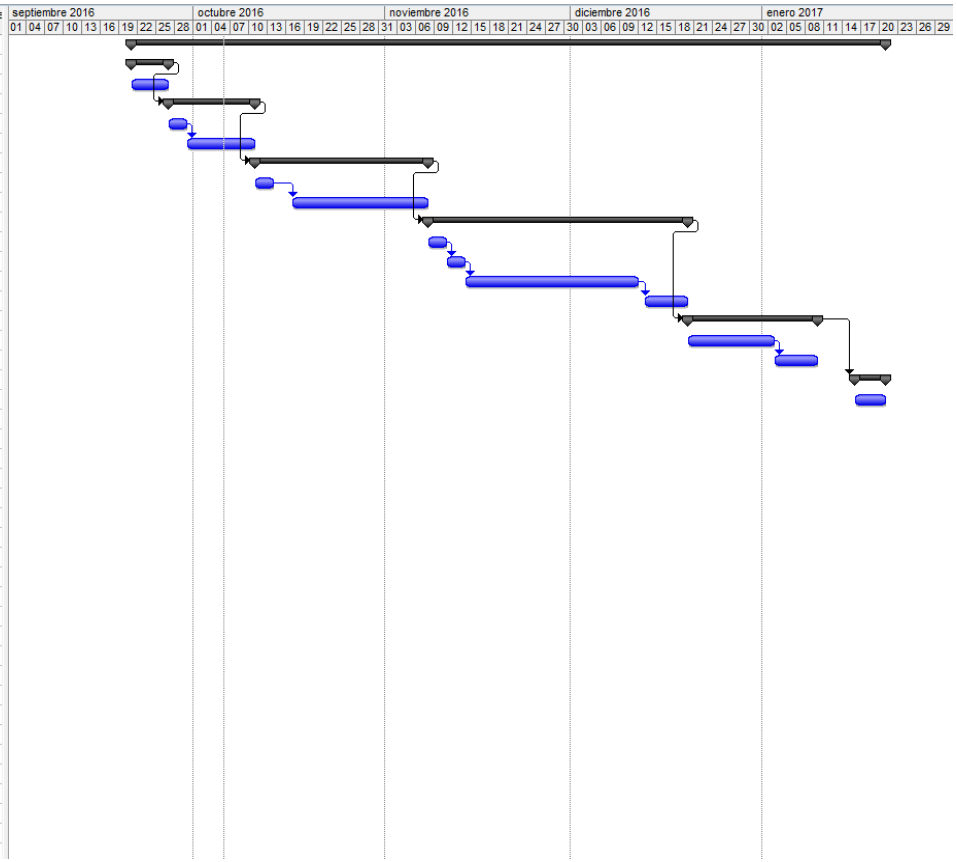
Cada una de les fites està desglossada en tasques, el detall de totes les tasques es mostra a continuació.

Tasca	Jornades
Lectura del Pla docent	4
Repàs materials Mineria de dades	3
Elaboració Pla treball	8
Cerca informació	3
Elaboració document estat de l'art	17
Lectura introducció a R	3
Anàlisi de les qüestions	2
Manipulació dades, neteja, transformació / Construcció del model	21
Obtenir conclusions	6
Elaboració de la memòria	11
Elaboració de la presentació	5
Tribunal	5

Taula 2. Tasques

4.3 Diagrama de Gantt

		Nombre de tarea	Duración	Comienzo	Fin	Predece
1		**TFG - Educational Data Mining and Learning Analitics**	**95 días**	**mié 21/09/16**	**vie 20/01/17**	
2		**Lectura Pla docent**	**4 días**	**mié 21/09/16**	**lun 26/09/16**	
3		Lectura Pla docent	4 días	mié 21/09/16	lun 26/09/16	
4		**PAC1 - Pla de treball**	**11 días**	**mar 27/09/16**	**lun 10/10/16**	**2**
5		Repàs materials Mineria de dades	3 días	mar 27/09/16	jue 29/09/16	
6		Elaboració Pla treball	8 días	vie 30/09/16	lun 10/10/16	5
7		**PAC2 - Estat de l'art de EDM & LA**	**22 días**	**mar 11/10/16**	**lun 07/11/16**	**4**
8		Cerca Informació	3 días	mar 11/10/16	jue 13/10/16	
9		Elaboració document estat de l'art	17 días	lun 17/10/16	lun 07/11/16	8
10		**PAC3 - Diseny i implementació**	**33 días**	**mar 08/11/16**	**lun 19/12/16**	**7**
11		Lectura introducció a R	3 días	mar 08/11/16	jue 10/11/16	
12		Anàlisi problema	2 días	vie 11/11/16	dom 13/11/16	11
13		Manipulació dades, neteja, transformació / creació models	21 días	lun 14/11/16	dom 11/12/16	12
14		Obtenir conclusions	6 días	mar 13/12/16	lun 19/12/16	13
15		**Lliurament**	**16 días**	**mar 20/12/16**	**lun 09/01/17**	**10**
16		Elaboració de la memòria	11 días	mar 20/12/16	lun 02/01/17	
17		Elaboració de la presentació	5 días	mar 03/01/17	lun 09/01/17	16
18		**Tribunal**	**5 días**	**lun 16/01/17**	**vie 20/01/17**	**15**
19		Tribunal	5 días	lun 16/01/17	vie 20/01/17	



Il·lustració 2. Diagrama de Gantt del TFG

5 Productes obtinguts

A continuació es detallen els productes obtinguts durant la realització d'aquest TFG.

- **Pla de Treball**. S'elabora al començament del projecte per a donar una idea de l'abast del mateix, concretament es defineixen els objectius que es volen aconseguir, es dona una planificació temporal de les taques en que es divideix el projecte, es defineixen els requeriments de tota mena necessaris per a dur-lo a terme, tals com maquinari, programari, requeriments de formació, i d'altres. Finalment es descriuen els riscos que amenacen el bon termini del projecte, intentant proposar una possible solució en cas que el risc es faci realitat.
- **Estat de l'art**. Document que descriu el cicle de vida d'un projecte de mineria de dades, les últimes novetats relatives a l'àmbit de l'EDM i LA, així com un seguit de projectes de contingut similar realitzats en altres parts del mon.
- **Anàlisi i Disseny.** En aquest document es dona solució a cada una de les qüestions platejades al projecte. És a on es descriu detalladament cada un dels passos necessaris per a desenvolupar un projecte de mineria de dades. Inclou l'estudi de les dades, processament, aplicació de models, gràfics i taules així com les conclusions extretes del procés. S'inclou el codi font desenvolupat en R i utilitzat en cada un del processos.
- **Memòria.** És aquest document. Recopila tota la informació recollida en els altres documents .
- **Presentació.** Presentació en format Microsoft Powerpoint contenint un resum de tot el projecte.

6 Estudi i tractament de les dades.

6.1 anàlisi de les dades.

Les dades contenen les matriculacions de 3824 alumnes en el seu primer semestre.

A partir de les dades proveïdes s'ha generat un arxiu csv anomenat EDMLA.csv al que s'ha afegit una capçalera per a identificar les variables.

Aquesta capçalera te la següent estructura:

Variable	Tipus	Descripció
sexe	qualitativa	M – Masculí F - Femení
semestre	qualitativa	Identificador del semestre
any	qualitativa	Any naixement alumne
matr	quantitativa	Nombre assignatures matriculades
pres	quantitativa	Nombre assignatures presentades
aprv	quantitativa	Nombre assignatures aprovades
05.554	qualitativa	0 no es matricula, 1 es matricula i no la presenta, 2 es matricula y la suspèn, 3 se matricula y la supera (assignatura 05.554)
05.554d1	qualitativa	1 no la matricula, 0 la matricula
05.554d2	qualitativa	1 la matricula, 0 no la matricula
05.554d3	qualitativa	1 la matricula i no la presenta, 0 al contrari
05.554d4	qualitativa	1 la matricula i la presenta, 0 al contrari
05.554d5	qualitativa	1 la matricula i no la supera, 0 al contrari
05.554d6	qualitativa	1 la matricula i la supera, 0 al contrari
…	qualitativa	Les ultimes 7 variables es repeteixen per a cada una de les assignatures: 05.554 05.555 05.556 05.557 05.558 05.559 05.560 05.561 05.562 05.563 05.564 05.565 05.566 05.567 05.568 05.569 05.570 05.571 05.572 05.573 05.574 05.575 05.576 05.577 05.578 05.579 05.580 05.581 05.582 05.583 05.584 05.585 05.586 05.587 05.588 05.589 05.590 05.591 05.592

		05.593 05.594 05.595 05.596 05.597 05.598 05.599 05.600 05.601 05.604 05.607 05.610 05.611 05.613 05.614 05.615 05.616 05.658
Rematricula	Objectiu	0 no es rematricula el segon semestre, 1 es rematricula el segon semestre

Taula 3. Estructura de l'arxiu de dades inicial

Un primer cop d'ull ja ens diu que tenim 406 variables, que son moltes sens dubte, però també podem apreciar que per a cada assignatura tenim molta informació redundant, que podem eliminar.

6.2 Conversió de les dades

Es realitza un primer procés de conversió de les dades que inclou les següents modificacions:

- S'eliminen de l'arxiu totes les variables dummy que fan un total de 57*6 = 342. Nomes es deixa les variables d'assignatures que tenen un rang de valors de 0 – 3
- La atribut rematricula es modifica segons el criteri
  - 0 - no
  - 1 – si
- L'atribut semestre es discretitza segons en criteri del seu ordre, el primer semestre s'anomena con a '*hivern'* i el segon com a '*estiu'*
  - XXXX1 - estiu
  - XXXX2 - hivern
- L'atribut '*any'* s'elimina i es crea un de nou que és el resultat de discretitzar l'atribut '*edat'* en 4 grups.:
  - <=25 – grup 1
  - >25, <=40 – grup 2
  - >40 , <=55 - grup 3
  - >55 - grup 4

L'script R que realitza aquesta conversió el trobem a l'annex 11.1 Script de conversió del arxiu inicial. Obtenim un arxiu de dades com aquest:

	sexe	semestre	edat	matr	pres	aprv	X05.554	X05.555	X05.556	X05.557	X05.558	X05.559	X05.560	X05.561	X05.562	X05.563	rematricula
1	M	hivern	Grup 3	2	2	2	0	0	0	0	0	0	0	0	0	0	si
2	M	hivern	Grup 2	2	0	0	0	0	0	0	0	0	0	0	1	0	no
3	M	hivern	Grup 3	2	1	1	0	0	3	0	0	0	0	0	0	1	si
4	F	hivern	Grup 3	3	3	3	0	0	0	3	0	0	0	0	0	3	si
5	M	hivern	Grup 3	2	1	1	0	0	0	0	0	0	0	3	1	0	si
6	M	estiu	Grup 3	3	3	3	0	0	0	0	0	0	0	0	0	3	si
7	M	hivern	Grup 3	2	1	1	0	0	0	0	0	0	0	0	0	0	no
8	M	estiu	Grup 3	2	2	2	0	0	0	0	0	0	0	0	0	3	si
9	M	hivern	Grup 3	2	2	2	0	0	0	0	0	0	0	3	0	3	si
10	M	estiu	Grup 3	2	2	2	0	0	0	0	0	0	0	0	0	0	si
14	M	hivern	Grup 3	1	0	0	0	0	0	0	0	1	0	0	0	0	si
15	M	hivern	Grup 3	2	1	1	0	0	0	0	0	0	0	3	0	0	si
16	M	estiu	Grup 3	3	2	2	0	0	0	0	0	0	0	3	1	0	no
17	M	hivern	Grup 3	2	0	0	0	0	0	0	0	0	0	0	0	0	si
18	M	estiu	Grup 3	2	0	0	0	0	0	0	0	0	0	0	0	0	si
19	M	estiu	Grup 3	2	2	1	0	0	0	3	0	0	0	0	2	0	si
20	M	hivern	Grup 3	1	0	0	0	0	0	0	0	0	0	0	0	0	no
21	M	hivern	Grup 2	2	2	2	0	0	0	0	0	0	0	0	0	0	no
22	M	estiu	Grup 2	2	1	0	0	0	0	0	0	0	0	0	0	0	no
23	M	hivern	Grup 3	2	2	1	0	0	0	2	0	0	0	0	0	3	si

Il·lustració 3. Conversió inicial de les dades

6.3 Anàlisi de les dades

El primer pas d'aquest procés és fer un anàlisi de les dades. Això significa tenir un primer contacte amb el conjunt de dades, veure quina és la distribució de dades per cada variable, comprovar que son coherents, que no falten dades i en definitiva, familiaritzar-se amb el joc de dades abans de fer cap aproximació a diferents models. Aquesta primera presa de contacte amb les dades ens dirà de forma aproximada el que podem treure d'elles.

A R disposem d'una llibreria, '*rattle',* aquesta llibreria te com a funcionalitats bàsiques, la descripció estadística de les dades fent diversos anàlisis i gràfics.

Carreguem les dades amb les definicions comentades i demanem a '*rattle*' l'anàlisi d'aquestes dades, que inclou el recull de valors únics per variable:

- Annex 11.2.Taula de valors únics per cada variable

Aquest recull ens permet veure els valors pera cada una de les variables del conjunt, per a variables qualitatives podem veure que no hi ha valors estranys fora de les classes esperades i per a les variables quantitatives podem verificar quines son el valor mínims, màxims, mitjanes etc. Fet que ens dona una idea de la 'salut' de les dades.

Veiem que hi ha un valor en '*any*' que correspon a un alumne nascut a 1911, que es va matricular al 2011 és a dir que tenia 100 anys. És molt improbable, però no impossible. Hi ha altres alumnes matriculats amb poc mes de 60 anys, cosa mes raonable.

Tampoc hi ha res erroni a la resta de variables, donat que contenen valors dins els rangs esperats.

Hi ha variables d'assignatura com X05.583 que nomes prenen els valors 0,1

Això vol dir que aquesta assignatura els alumnes que s'han matriculat no s'han presentat, per tant no hi ha suspensos ni aprovats.

Finament trobem la resta de variables d'assignatura en varis grups:

Grup amb valors 0,1,3. Indica assignatures que si algú s'ha matriculat, o no es presenten o aproven, però ningú has suspès.

Grup amb valors 0,3. Assignatures que els que s'han matriculat tots han aprovat.

Totes les alternatives son raonables.

'*Rattle'* ens avisa que les assignatures 05.581, 05.610 i 05.615 no tenen valors diferents de '0' perquè ningú s'ha matriculat d'aquestes assignatures i per aquest motiu les eliminarem. Ens quedem finalment amb 62 variables

Pel que fa a les freqüències dels valors i d'altres mesures estadístiques com valor màxim, mínim, mitja, mitjana, primer i tercer quartil, '*rattle'* també ens ofereix aquesta taula:

- Annex 11.3. Taula de freqüències dels valors de cada variable

Pel que fa a la resta de variables ens mostra les freqüències absolutes dels seus valors:

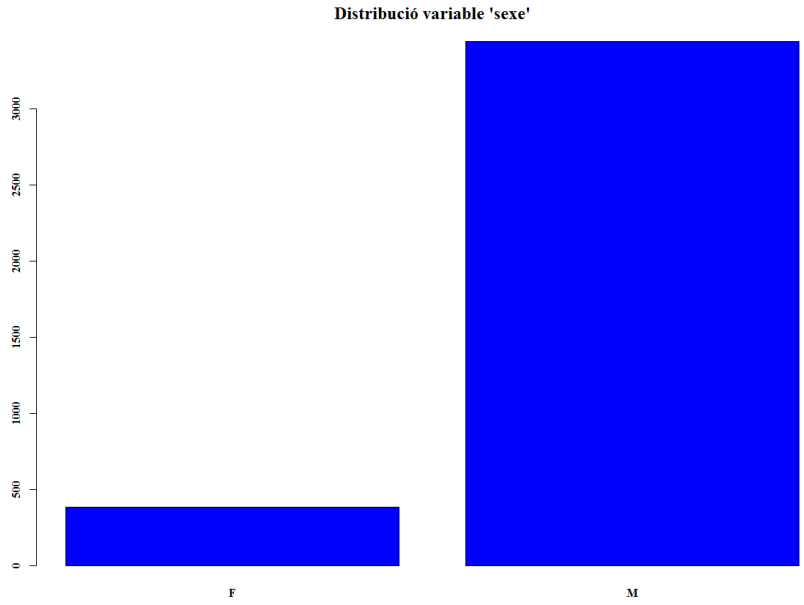
- Annex 11.4. Taula de freqüències dels valors de les assignatures

També podem conèixer per a cada atribut si falten valors entre d'altres paràmetres.

- Annex 11.5. Taula de valors faltants i únics per cada variable

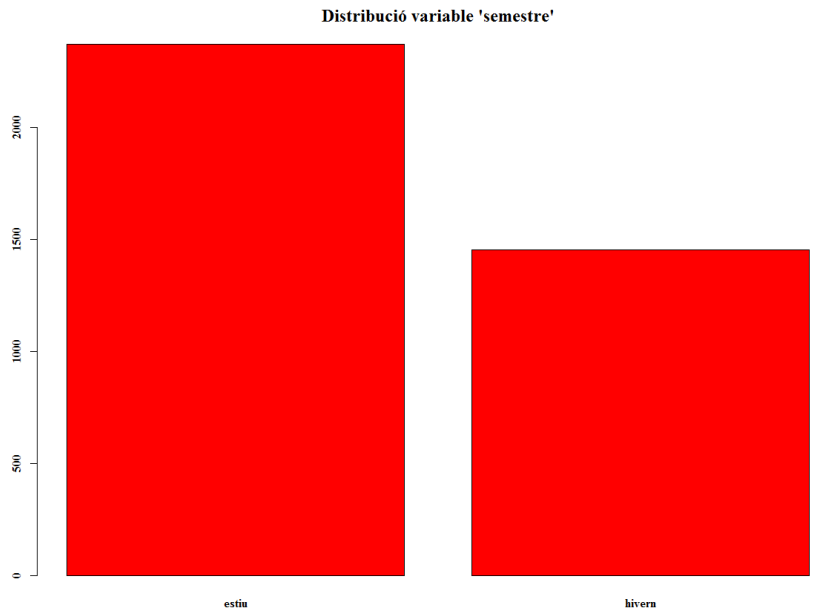
Com a conclusió sabem que les dades estan netes no falten valors ni hi ha valors incorrectes.

Es generen una sèrie de gràfics per a visualitzar de forma mes senzilla aquestes dades:



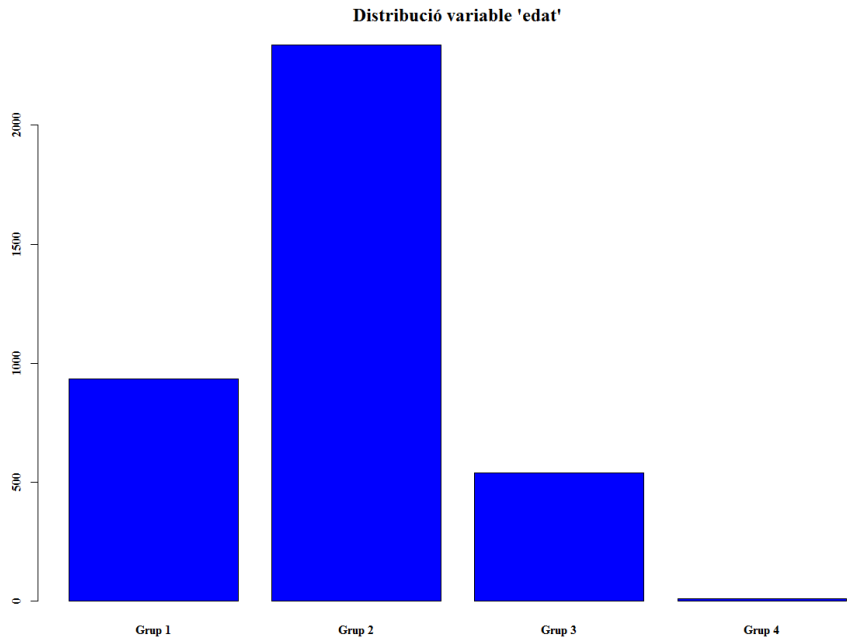
Il·lustració 4. Distribució per '*sexe*'

Crida l'atenció, les distribucions per sexe, un 10% dels registres corresponen a dones i la resta a homes. Sembla que les matriculacions corresponen a uns estudis concrets i és possible que aquest estudis tinguin una majoria d'homes.



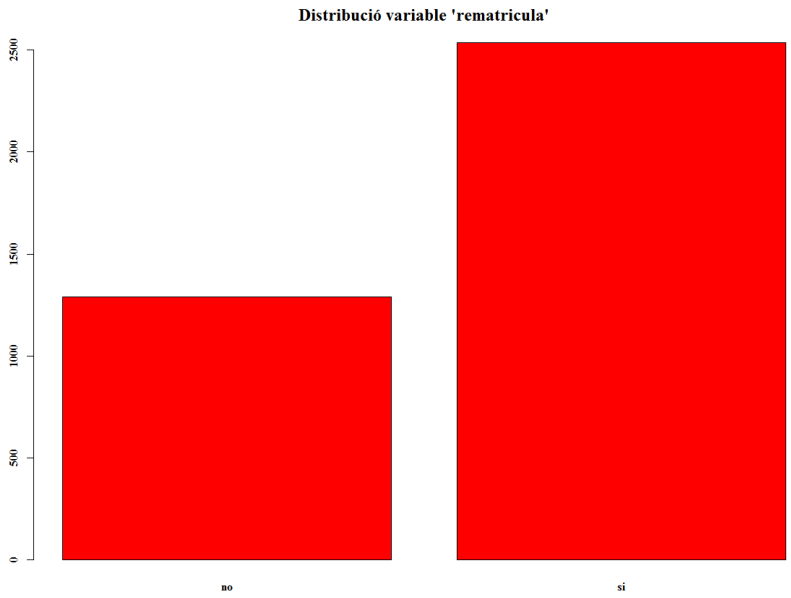
Il·lustració 5. Distribució per '*semestre*'

Hi ha mes matricules a l'estiu que a l'hivern. Un 62% i 38% respectivament.



Il·lustració 6. Distribució per '*edat*'

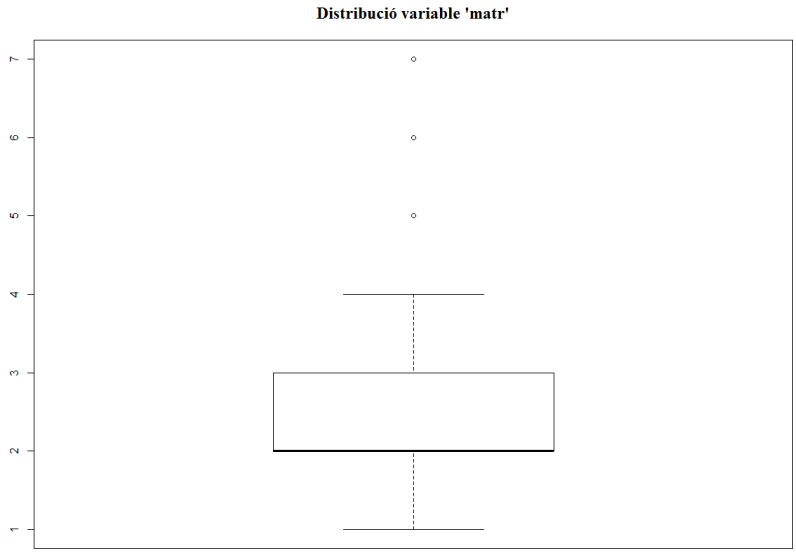
Per edats, el grup 2 és el majoritari, compren alumnes entre 26 i 40 anys. En canvi alumnes de mes de 55 anys son pràcticament inexistents



Il·lustració 7. Distribució per la variable '*rematricula*'

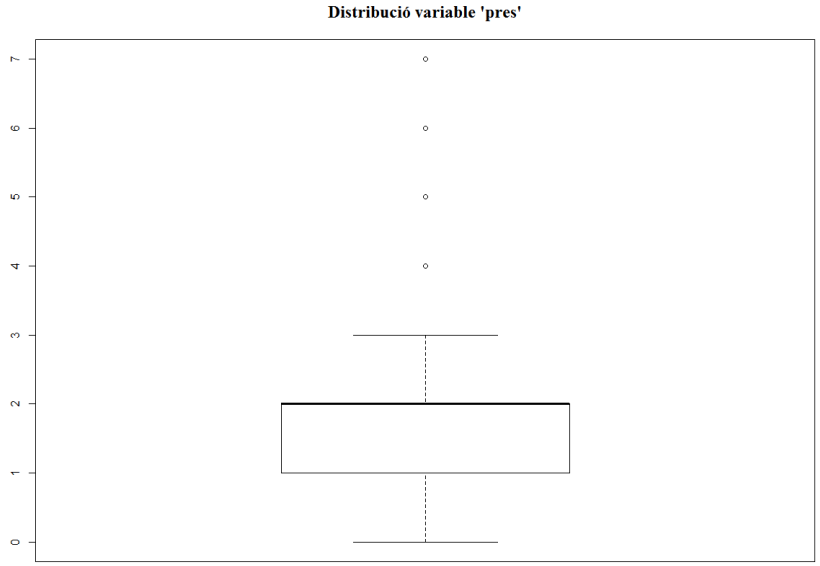
La variable objectiu diu que no s'han tornat a matricular 1289 alumnes dels 3824 matriculats. Això representa el 33,7% del total. Un 66,3 % dels alumnes continua els seus estudis.

Veiem també com es distribueixen les variables quantitatives



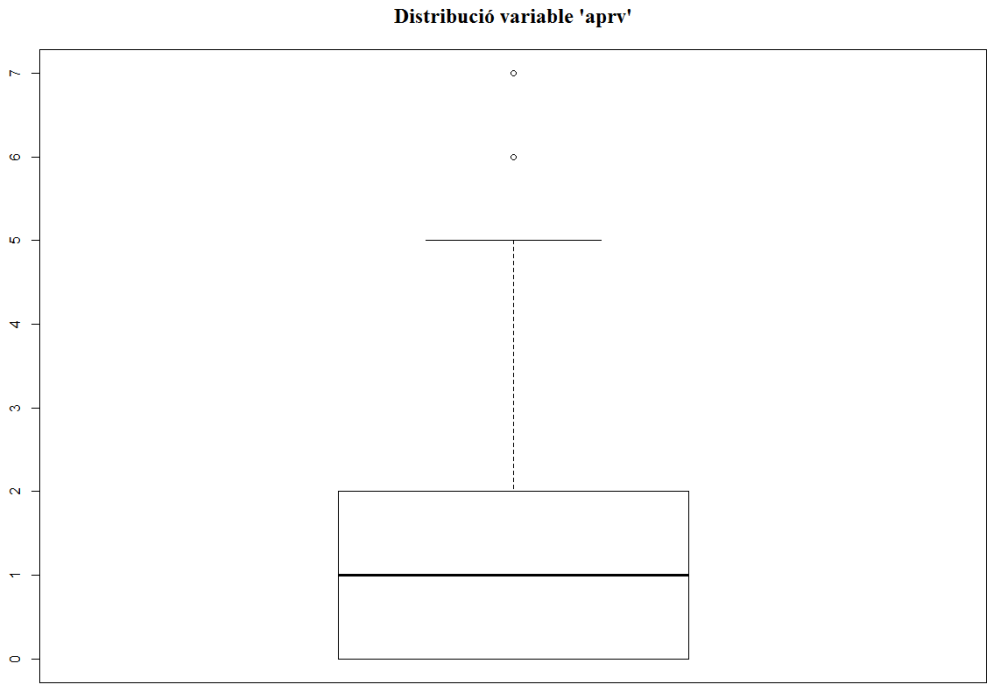
Il·lustració 8. Distribució de la variable '*matr*'

La gran majoria d'alumnes es matriculen entre 1 i 4 assignatures. Els que ho fan de mes de 4 son '*outliers*'.



Il·lustració 9.Distribució de la variable '*pres*'

El 50% dels alumnes es presenten de una o dues assignatures i en general es presenten com a màxim de 3. Els que es presenten de mes de 3 son casos molt poc freqüents.

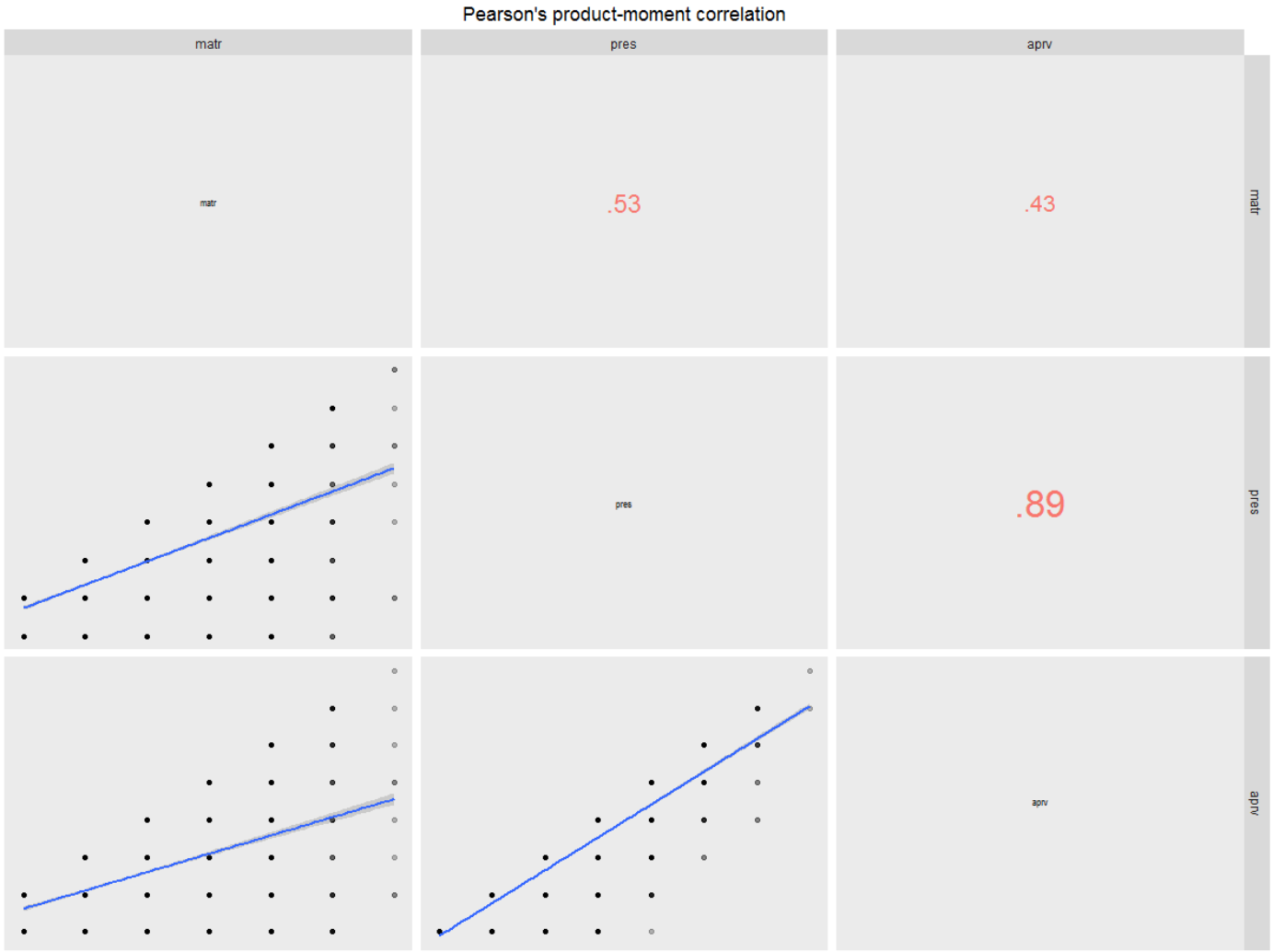


Il·lustració 10.Distribució de la variable '*aprv*'

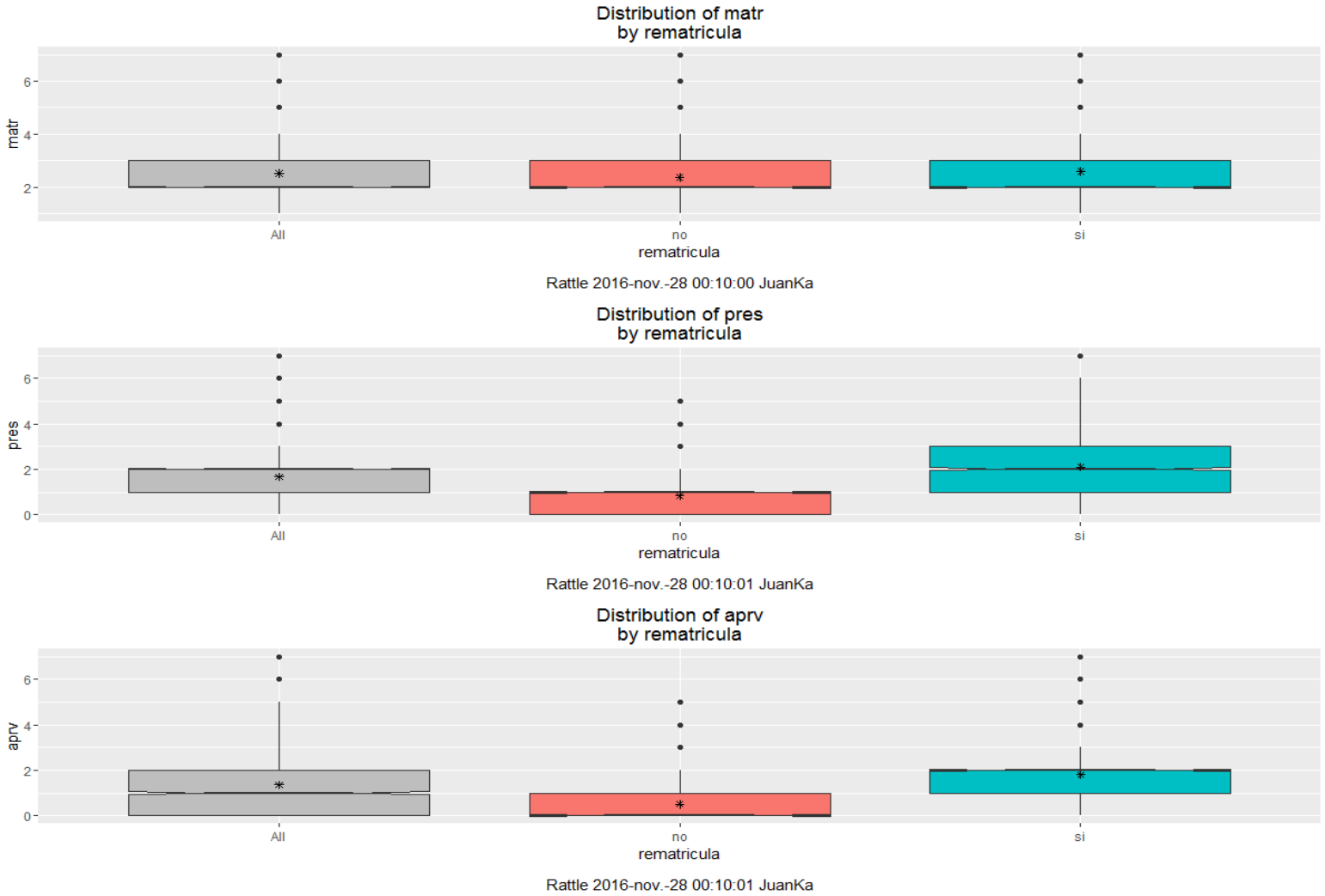
El 75% dels alumnes aproven com a màxim dues assignatures.

Podem veure també que el màxim d'assignatures matriculades és 7 i també el màxim d'aprovades. És a dir algun alumne s'ha matriculat de 7 assignatures i les ha aprovat totes.

També podem tenir gràfics bàsics que comparen les variables numèriques per parelles i calculant les rectes de regressió mostren possibles correlacions entre variables. Com veurem al gràfic hi ha una forta relació entre la variable '*presentades*' i '*aprovades*' indicant-nos que com mes assignatures ens presentem mes assignatures aprovem, raonament molt lògic.

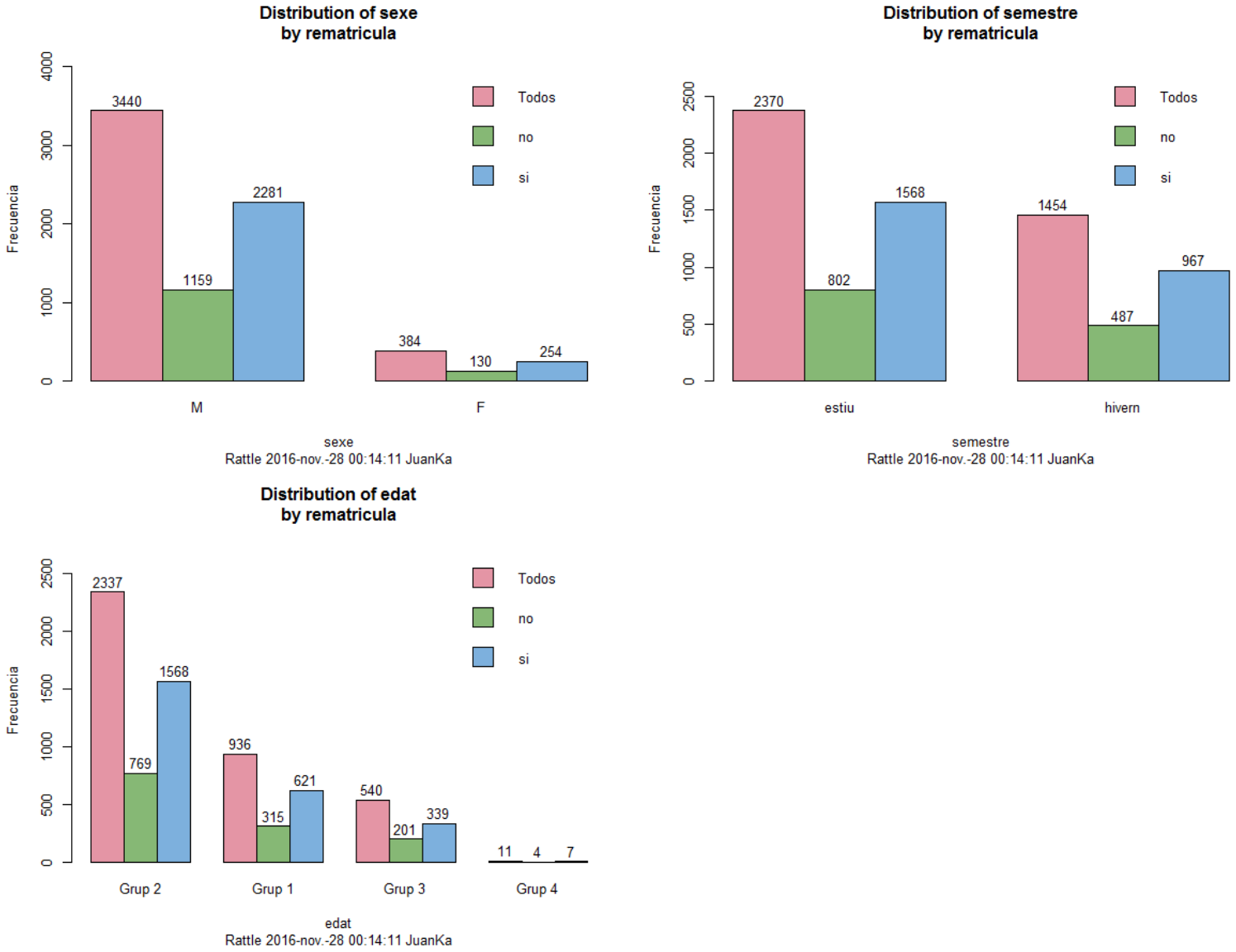


Il·lustració 11. Correlació de Pearson variables quantitatives



Il·lustració 12. Distribució de les variables quantitatives en funció de la variable objectiu

La distribució de variables qualitatives respecte a l'objectiu mostra el que ja hem vist, contra mes aprovades, mes alt és l'índex de rematriculació.



Il·lustració 13. Distribució de les variables qualitatives en funció de la variable objectiu.

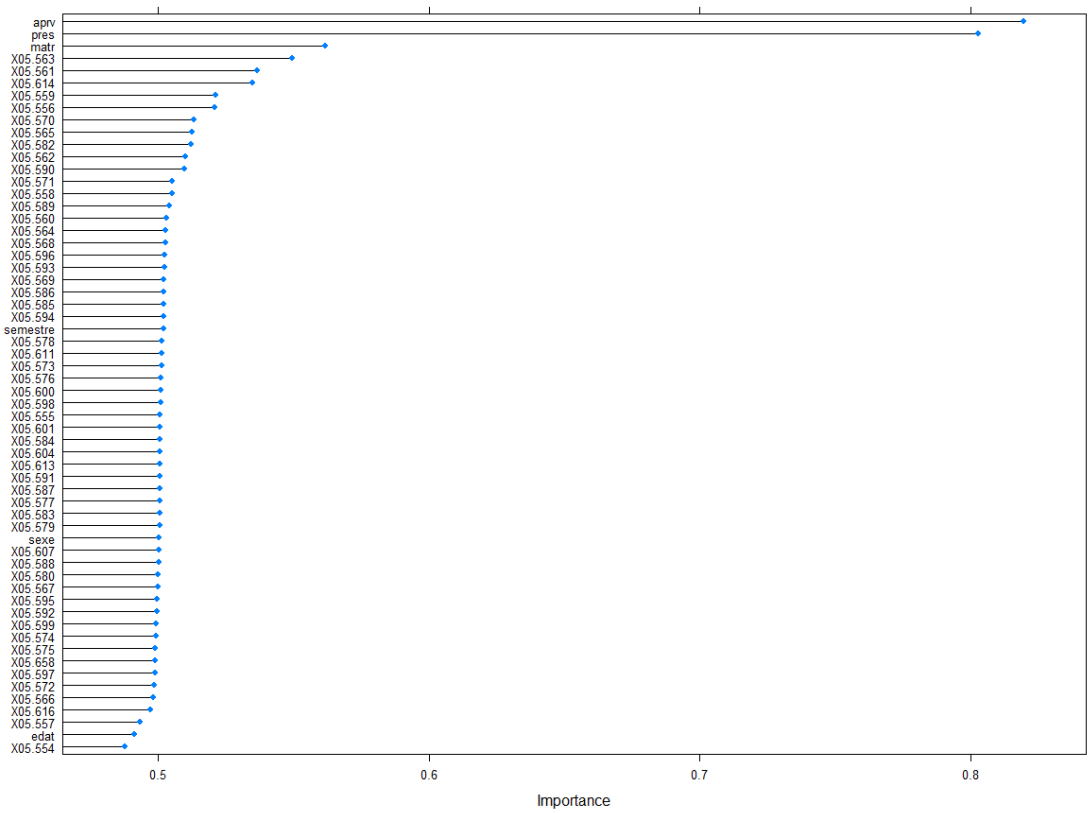
La distribució de les variables qualitatives respecte a l'objectiu ens diu clarament que la distribució és pràcticament igual en cada grup de cada variable, per tant la influencia d'aquestes variables en la rematriculació és nul·la.

6.4 Variables importants

Hem vist que el conjunt de dades conte moltes variables, la gran majoria (54) son assignatures i unes poques son sociodemogràfiques. Voldríem saber si hi ha variables que tenen mes influencia en la decisió de tornar-se a matricular o en general quin pes té cada variable en la decisió.

Un dels mètodes que podem utilitzar a tal fi és el '*lvq*' (Learning Vector Quantization), que explicat de forma simple permet establir el grau d'importància de cada variable pel que fa a la variable objectiu.

Parametritzem el procés indicant quina és la variable objectiu dins el conjunt de totes les variables que volem analitzar. Annex 11.7. Variables importants



Il·lustració 14. Gràfica de les variables més importants.

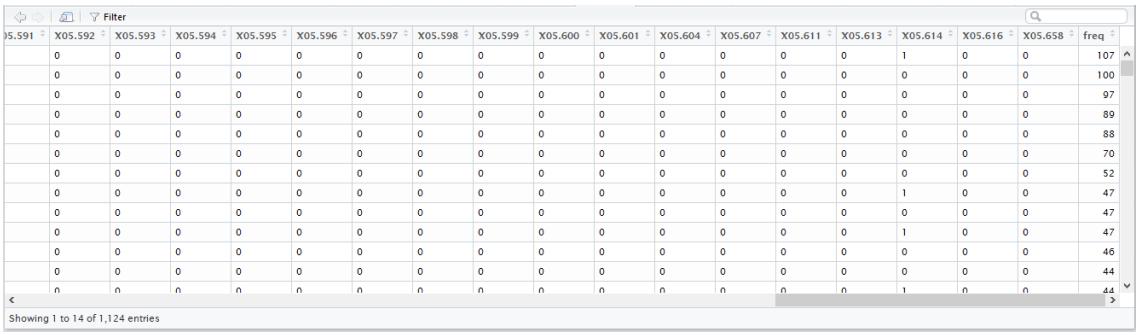
Els atributs '*aprv*' i '*pres*' son els que mes importància tenen, quedant les assignatures gairebé totes amb un pes similar. Edat, sexe i semestre tenen poc pes en el fet de tornar a matricular-se.

```
ROC curve variable importance

  only 20 most important variables shown (out of 60)

        Importance
aprv        0.8196
pres        0.8030
matr        0.5616
X05.563     0.5494
X05.561     0.5365
X05.614     0.5347
X05.559     0.5213
X05.556     0.5208
X05.570     0.5131
X05.565     0.5125
X05.582     0.5122
X05.562     0.5098
X05.590     0.5097
X05.571     0.5052
X05.558     0.5050
X05.589     0.5041
X05.560     0.5029
X05.564     0.5027
X05.568     0.5026
X05.596     0.5024
```

Taula 4. 20 Variables més importants i el seu valor d'importància

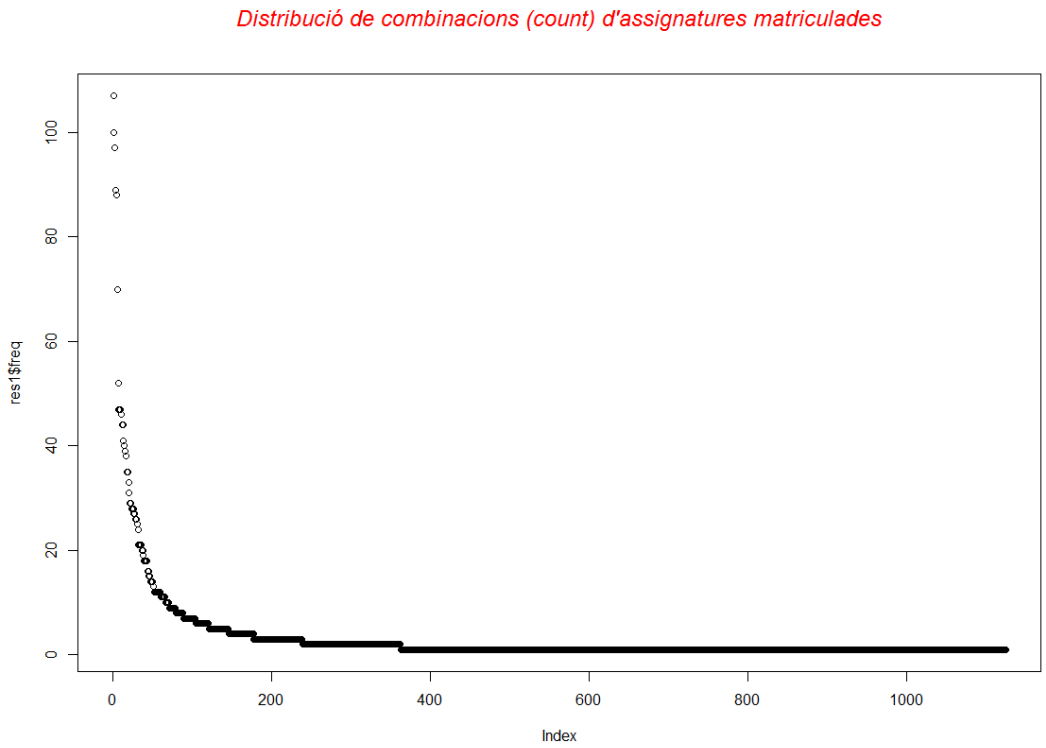
6.5 Dispersió de les dades

S'ha fet un anàlisi per veure les diferents combinacions de les assignatures. Per exemple és crea la matriu d'assignatures matriculades i amb la llibreria '*plyr*' es compten les combinacions diferents de totes les matricules. El que seria un '*distinct*' per les combinacions d'assignatures matriculades. Annex 11.9. Dispersió de les dades. Assignatures matriculades.



Il·lustració 15. Diferents combinacions de les assignatures matriculades i freqüència.

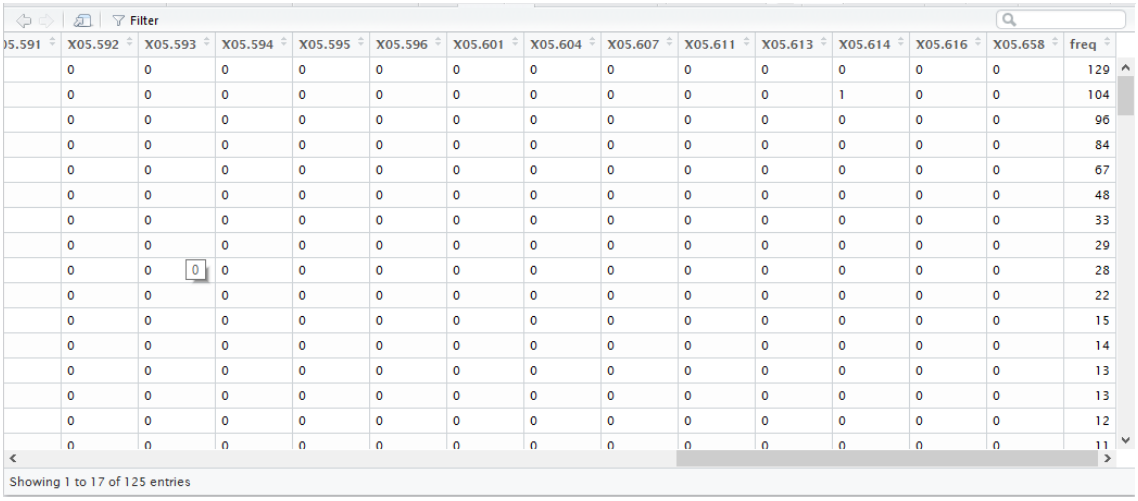
Es genera una taula amb 1124 files, o el que és el mateix 1124 combinacions diferents de entre 3824 matriculacions. Ja dona la idea de que difícilment existeixen patrons comuns amb tanta diversitat i freqüències tan baixes.



Il·lustració 16. Gràfic de la distribució de les diferents combinacions de matricules.

La frequencia mes alta és de 107, que vol dir que hi ha 107 matricules identiques que nomes representa un 2,8% de totes les matricules. Aquest fet denota una gran dispersió en les dades.

Fem el mateix per les assignatures suspeses. Annex 11.10. Dispersió de les dades. Assignatures suspeses.

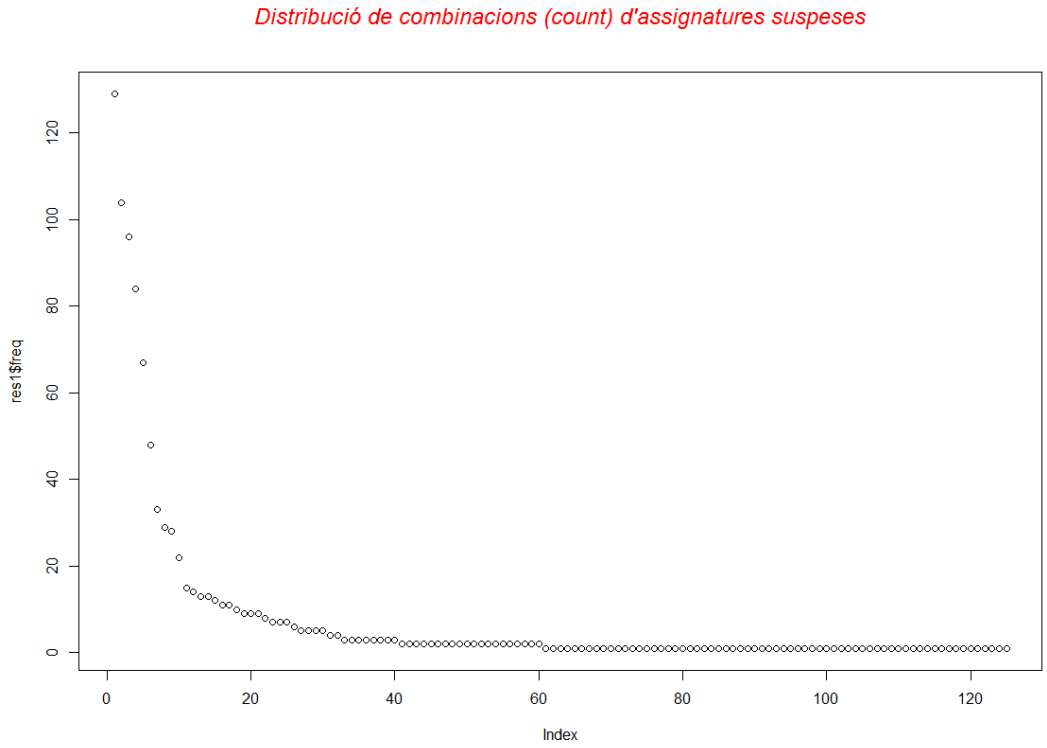


X05.591	X05.592	X05.593	X05.594	X05.595	X05.596	X05.601	X05.604	X05.607	X05.611	X05.613	X05.614	X05.616	X05.658	freq
	0	0	0	0	0	0	0	0	0	0	0	0	0	129
	0	0	0	0	0	0	0	0	0	0	1	0	0	104
	0	0	0	0	0	0	0	0	0	0	0	0	0	96
	0	0	0	0	0	0	0	0	0	0	0	0	0	84
	0	0	0	0	0	0	0	0	0	0	0	0	0	67
	0	0	0	0	0	0	0	0	0	0	0	0	0	48
	0	0	0	0	0	0	0	0	0	0	0	0	0	33
	0	0	0	0	0	0	0	0	0	0	0	0	0	29
	0	0	0	0	0	0	0	0	0	0	0	0	0	28
	0	0	0	0	0	0	0	0	0	0	0	0	0	22
	0	0	0	0	0	0	0	0	0	0	0	0	0	15
	0	0	0	0	0	0	0	0	0	0	0	0	0	14
	0	0	0	0	0	0	0	0	0	0	0	0	0	13
	0	0	0	0	0	0	0	0	0	0	0	0	0	13
	0	0	0	0	0	0	0	0	0	0	0	0	0	12
	0	0	0	0	0	0	0	0	0	0	0	0	0	11

Showing 1 to 17 of 125 entries

Il·lustració 17. Diferents combinacions de les assignatures suspeses i freqüència.

Hi ha 127 combinacions diferents d'assignatures suspeses i la frequencia mes alta és de 129 casos que representa un 3,3% de les matricules. La majoria de combinacions presenten frequencies mol baixes. És el mateix cas, la dispersió és molt gran i no será facil trobar patrons comuns.



Il·lustració 18. Gràfica de la distribució de les diferents combinacions de suspensos.

Amb aquest analisi ja veiem que els grups d'assignatures matriculades son tan variats que será dificil trobar patrons de comportament frequents o comuns.

6.6 Construcció de models

6.6.1 Pregunta 1. Quines combinacions d'assignatures són les més típiques?

El plantejament que es demana són les matricules que es repeteixen més freqüentment entre els alumnes.
La resposta ve donada per l'ús d'un model associatiu, en concret el MBA, (Market Basket Analysis).

Definició:

Si $I = \{i_1, i_2, \ldots i_n\}$ un conjunt de n atributs binaris anomenats items.

Si $T = \{t_1, t_2, \ldots t_n\}$ és un conjunt de transaccions emmagatzemades en la nostra base de dades.

Cada transacció en T té un ID (identificador) únic i conté un subconjunt d'items de I .Una regla es defineix como una implicació de la forma:

$X \Rightarrow Y$ on: $X, Y \subseteq I$ i $X \cap Y = \emptyset$

Els conjunts d'ítems X i I es denominen respectivament '*antecedent*' (o part esquerra) i '*conseqüent*' (o part dreta) de la regla.
El '*suport*' d'un conjunt d'ítems X en una base de dades T es defineix com la proporció de transaccions a la base de dades que conté aquest conjunt d'ítems:

$$sop(X) = \frac{|X|}{|T|}$$

La *'confiança'* de una regla es defineix com:

$$conf(X \Rightarrow Y) = \frac{sop(X \cup Y)}{sop(X)} = \frac{|X \cup Y|}{|X|}$$

L'indicador '*Lift*' = 1 indica que aquest conjunt apareix una quantitat de vegades acord a l'esperat sota condicions d'independència. Un valor de lift> 1 indica els productes es troben en el conjunt més vegades del normal. Un valor de lift <1 indica que els productes formen part del mateix conjunt menys vegades del que és normal.

En realitat una assignatura no deixa de ser un article que un estudiant compra al centre d'estudis. Pel que en resum el que es desprèn d'una matricula és que és una cistella de la compra.

Si apliquem un algoritme que ens permeti determinar tals combinacions haurem resolt la qüestió. L'algoritme o model que sembla indicat per a tal cas és el d'associació, en concret i amb la possibilitat de disposar d'ell des de R tenim el *'apriori'*.

Havent estudiat com funciona l'algorisme, és important destacar que la prioritat és l'estructura de dades amb què treballa.

Apriori treballa amb dos tipus d'arxiu, el simple i el Basket.

- Estructura simple:

Cada línia inclou només 1 article i ha d'anar associat de l'identificador que s'utilitzarà per agrupar els ítems, numero de tiquet etc.

ID	Article
Tiquet 1	Article 1
Tiquet 1	Article 9
Tiquet 1	Article 3
Tiquet 2	Article 4
Tiquet 2	Article 1
Tiquet 3	Article 8
Tiquet 3	Article 6
Tiquet 3	Article 3

Taula 5. Exemple estructura apriori simple

- Estructura Basket

Cada línia representa un tiquet i conte tots els seus article sense especificar ordre.

Tiquets
Article 1 Article 9 Article 3
Article 4 Article 1
Article 8 Article 6 Article 3

Taula 6. Exemple estructura apriori basket

Un altre possibilitat és disposar d'una matriu de dades a on les columnes son les assignatures i les files representen matriculacions, per a identificar si l'assignatura s'ha matriculat o no escriurem un '0' o un '1' a la columna corresponent. Es una format menys eficient que els anteriors donat que ocupa molt mes espai, però es un fet irrellevant per a aquest problema.

Per a obtenir el format adient nomes cal transformar els valors de cada assignatura, 0,1,2,3, per el següent:

- 0 -> 0

- 1,2,3 -> 1

Es a dir, qualsevol valor diferent de '0' serà transformat per 1. Annex 11.11. Pregunta 1. Transformació de dades per a aplicar l'algorisme '*apriori*'.

El que obtenim es un arxiu de transaccions estructura 'basket' que conte una línia d'assignatures corresponent a les assignatures matriculades. Es veu al exemple:

```
X05_578 X05_594
X05_562 X05_614
X05_556 X05_563
X05_557 X05_563 X05_565
X05_561 X05_562
X05_563 X05_565 X05_614
X05_582 X05_586
X05_563 X05_571
X05_561 X05_563
X05_582 X05_601
X05_554 X05_561
X05_554 X05_556 X05_557 X05_559 X05_561 X05_562 X05_614
X05_554 X05_562
X05_559
X05_561 X05_570
X05_561 X05_562 X05_614
X05_567 X05_611
X05_565 X05_573
X05_557 X05_562
X05_565
X05_582 X05_604
X05_565 X05_573
X05_557 X05_563
X05_614
X05_582 X05_593 X05_595 X05_596 X05_600
```

Taula 7. Arxiu de transaccions per a apriori

En aquest punt ja podem analitzar:

```
summary(datos)
```

```
transactions as itemMatrix in sparse format with
 3824 rows (elements/itemsets/transactions) and
 54 columns (items) and a density of 0.04664981

most frequent items:
X05_562 X05_559 X05_557 X05_554 X05_561 (Other)
   1288     967     937     918     836    4687
```

```
element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7
 452 1782 1002  392  142   45    9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   2.519   3.000   7.000
```

Taula 8. Pregunta 1. Sumari de transaccions

Interpretem aquestes dades de la següent manera: Tenim 3824 files, tants com té l'arxiu original. 3824 matricules. 54 columnes que corresponen a les 57 assignatures menys les 3 que no han estat matriculades per cap alumne. I que són: X05.581, X05.610 i X05.615, (han estat eliminades).

L'informe ens diu que la densitat és de 0,04664981. La densitat indica el nombre de cel·les de la matriu que contenen valor, que són 3824 * 54 * 0,04664981 = 9633 assignatures. Les Assignatures mes freqüents, és a dir que han estat mes vegades triades pels alumnes a l'hora de matricular-se són:

Assignatura	Freqüència absoluta	Freqüència relativa
X05.562	1288	0,3368200837
X05.559	967	0,2528765690
X05.557	937	0,2450313808
X05.554	918	0,2400627615
X05.561	836	0,2186192469

Taula 9. Pregunta 1. Assignatures mes freqüents.

Por contingut de cada matrícula

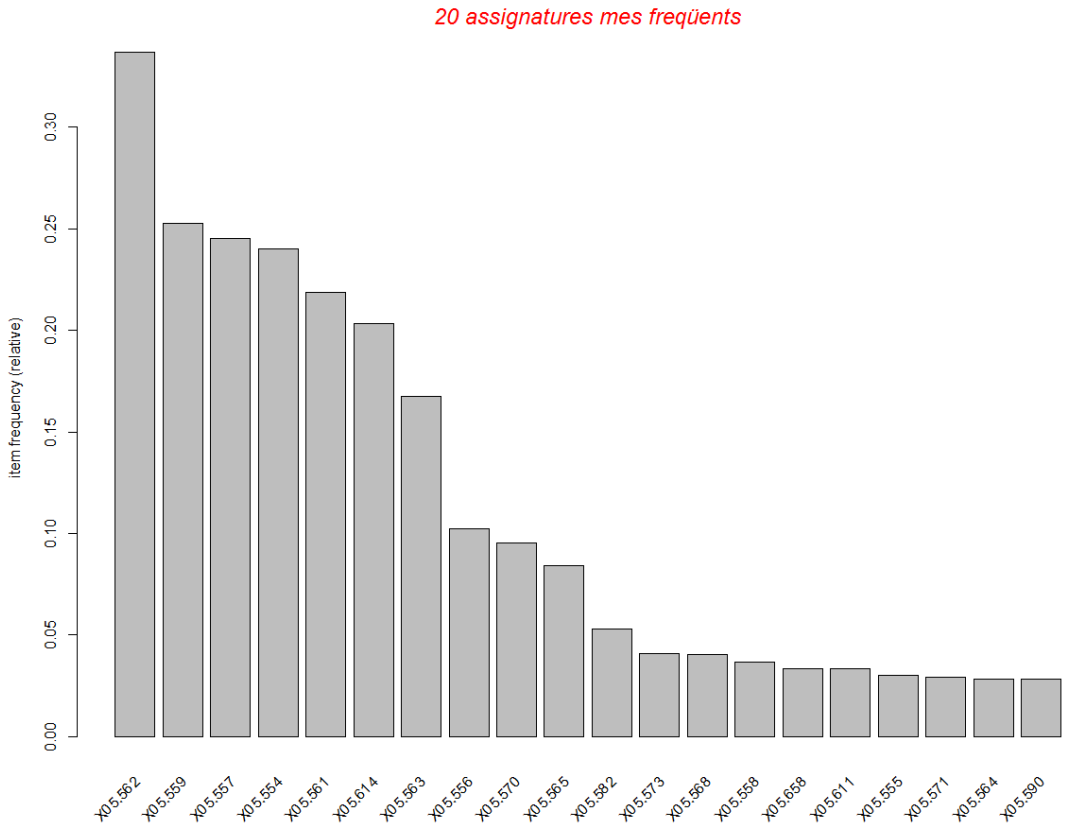
Numero de assignatures per matricula	Freqüència
1	452
2	1782
3	1002
4	392
5	142
6	45
7	9

Taula 10. Pregunta 1. Distribució freqüències per numero d'assignatures matriculades

Annex 11.12. Pregunta 1. Taula de assignatures mes matriculades ordenades per freqüència.

Podem visualitzar la gràfica ordenada amb les 20 primeres:

```
itemFrequencyPlot (dades, topN = 20)
```

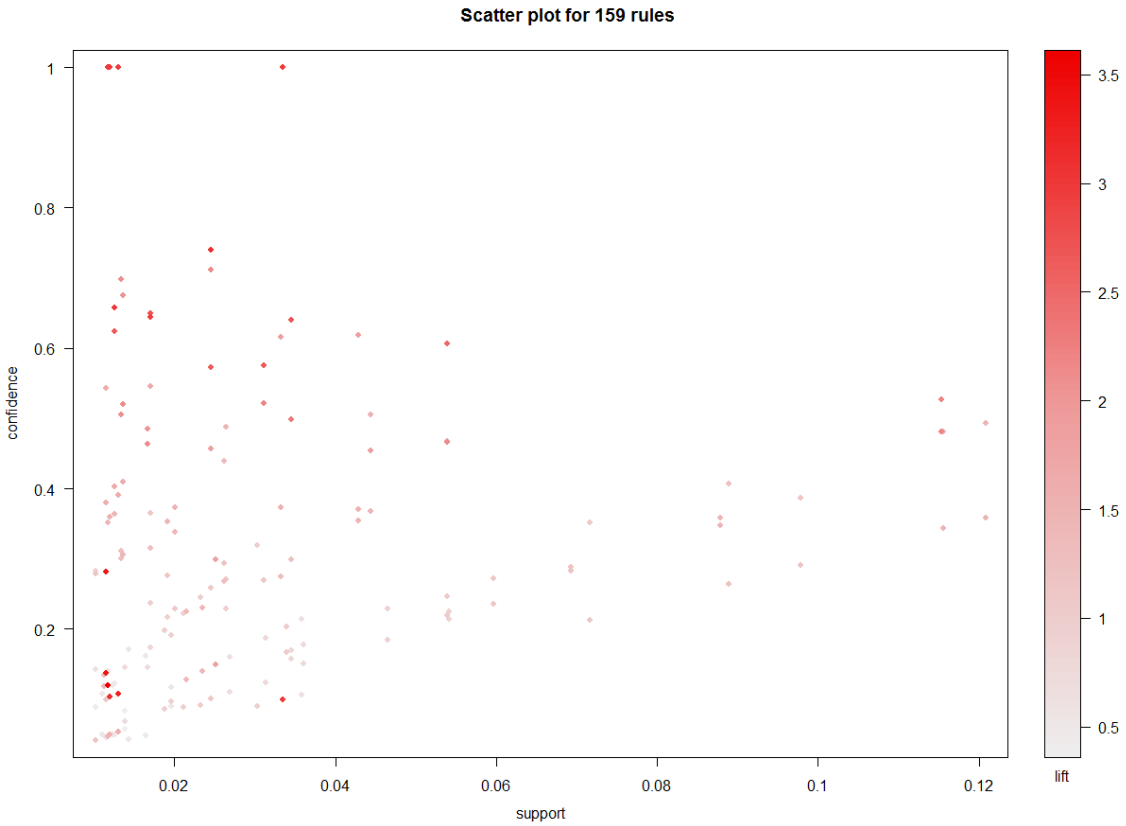


Il·lustració 19. 20 Assignatures mes matriculades ordenades per freqüència.

Aquest anàlisi és relatiu a les assignatures de forma individual, és a dir en el 33,7% de les matricules hi ha l'assignatura X05.562.
També ens interessa conèixer si varies assignatures es repeteixen de manera freqüent a les matricules. Per a trobar aquestes associacions apliquem l'algorisme 'apiori'.

Annex 11.13. Pregunta 1. Obtenció de regles d'associació 'apriori'.
Per a obtenir regles es necessari baixar el llindar de suport i confiança mínims a 0,01 respectivament..
Hem obtingut 159 regles que podem inspeccionar així:

```
plot(reglas)
inspect(reglas)
```



Il·lustració 20. Pregunta 1. Gràfic de regles 'apriori'

Les regles estan ordenades pel valor de suport. Observem les 40 primeres

lhs		rhs	support	confidence	lift
{X05.557}	=>	{X05.562}	0,12081590	0,4930630	1,4638764
{X05.562}	=>	{X05.557}	0,12081590	0,3586957	1,4638764
{X05.554}	=>	{X05.562}	0,11558577	0,4814815	1,4294916
{X05.562}	=>	{X05.554}	0,11558577	0,3431677	1,4294916
{X05.561}	=>	{X05.554}	0,11532427	0,5275120	2,1973919
{X05.554}	=>	{X05.561}	0,11532427	0,4803922	2,1973919
{X05.559}	=>	{X05.562}	0,09780335	0,3867632	1,1482783
{X05.562}	=>	{X05.559}	0,09780335	0,2903727	1,1482783
{X05.561}	=>	{X05.562}	0,08891213	0,4066986	1,2074653
{X05.562}	=>	{X05.561}	0,08891213	0,2639752	1,2074653
{X05.557}	=>	{X05.559}	0,08786611	0,3585912	1,4180485
{X05.559}	=>	{X05.557}	0,08786611	0,3474664	1,4180485
{X05.614}	=>	{X05.562}	0,07165272	0,3521851	1,0456178
{X05.562}	=>	{X05.614}	0,07165272	0,2127329	1,0456178
{X05.554}	=>	{X05.557}	0,06929916	0,2886710	1,1780982
{X05.557}	=>	{X05.554}	0,06929916	0,2828175	1,1780982
{X05.561}	=>	{X05.559}	0,05962343	0,2727273	1,0784996
{X05.559}	=>	{X05.561}	0,05962343	0,2357808	1,0784996

{X05.554}	=>	{X05.559}	0,05413180	0,2254902	0,8917006
{X05.559}	=>	{X05.554}	0,05413180	0,2140641	0,8917006
{X05.561}	=>	{X05.557}	0,05387029	0,2464115	1,0056324
{X05.557}	=>	{X05.561}	0,05387029	0,2198506	1,0056324
{X05.554,X05.561}	=>	{X05.562}	0,05387029	0,4671202	1,3868537
{X05.561,X05.562}	=>	{X05.554}	0,05387029	0,6058824	2,5238498
{X05.554,X05.562}	=>	{X05.561}	0,05387029	0,4660633	2,1318496
{X05.614}	=>	{X05.559}	0,04654812	0,2287918	0,9047567
{X05.559}	=>	{X05.614}	0,04654812	0,1840745	0,9047567
{X05.557,X05.559}	=>	{X05.562}	0,04445607	0,5059524	1,5021443
{X05.557,X05.562}	=>	{X05.559}	0,04445607	0,3679654	1,4551185
{X05.559,X05.562}	=>	{X05.557}	0,04445607	0,4545455	1,8550500
{X05.554,X05.557}	=>	{X05.562}	0,04288703	0,6188679	1,8373843
{X05.554,X05.562}	=>	{X05.557}	0,04288703	0,3710407	1,5142580
{X05.557,X05.562}	=>	{X05.554}	0,04288703	0,3549784	1,4786898
{X05.614}	=>	{X05.554}	0,03608787	0,1773779	0,7388813
{X05.554}	=>	{X05.614}	0,03608787	0,1503268	0,7388813
{X05.563}	=>	{X05.562}	0,03582636	0,2137285	0,6345481
{X05.562}	=>	{X05.563}	0,03582636	0,1063665	0,6345481
{X05.614}	=>	{X05.561}	0,03451883	0,1696658	0,7760790
{X05.561}	=>	{X05.614}	0,03451883	0,1578947	0,7760790
{X05.554,X05.561}	=>	{X05.557}	0,03451883	0,2993197	1,2215567

Taula 11. Pregunta 1. 40 regles ordenades per 'support'

Les regles obtingudes no tenen valors de suport i confiança molt elevats, així com el lift. Estem davant d'una situació que denota molta dispersió a les matricules. Es a dir hi ha assignatures matriculades amb una freqüència relativament alta, però a l'hora de matricular-se els estudiants seleccionen combinacions molt variades d'assignatures.

Així podem concloure que la combinació més típica és l'assignatura {X05_557} amb {X05_562} amb una probabilitat del 12,08% i una confiança del 49,3%

Segons la taula també hi ha la regla inversa, quan un alumne es matricula de {X05_562} també ho fa de {X05_557} amb la mateixa probabilitat que la seva inversa, l'única diferència és que quan això passa (matricular-se de {X05_562}) només el 35,87% de les vegades concorre amb {X05_557}.

Farem finalment una poda de les regles per a obtenir regles independents. El sistema mostra regles que inclouen altres regles. Amb la poda eliminem les regles dependents, deixant les més genèriques.

Annex 11.14. Poda de les regles.

El conjunt queda reduït a 44 regles i totes de longitud 2.

lhs		rhs	support	confidence	lift
{X05.557}	=>	{X05.562}	0,1208159	0,4930630	1,4638764
{X05.554}	=>	{X05.562}	0,1155858	0,4814815	1,4294916
{X05.561}	=>	{X05.554}	0,1153243	0,5275120	2,1973919
{X05.559}	=>	{X05.562}	0,0978034	0,3867632	1,1482783
{X05.561}	=>	{X05.562}	0,0889121	0,4066986	1,2074653
{X05.557}	=>	{X05.559}	0,0878661	0,3585913	1,4180485
{X05.614}	=>	{X05.562}	0,0716527	0,3521851	1,0456178
{X05.554}	=>	{X05.557}	0,0692992	0,2886710	1,1780982
{X05.561}	=>	{X05.559}	0,0596234	0,2727273	1,0784996
{X05.554}	=>	{X05.559}	0,0541318	0,2254902	0,8917006
{X05.561}	=>	{X05.557}	0,0538703	0,2464115	1,0056324
{X05.614}	=>	{X05.559}	0,0465481	0,2287918	0,9047567
{X05.614}	=>	{X05.554}	0,0360879	0,1773779	0,7388813
{X05.563}	=>	{X05.562}	0,0358264	0,2137286	0,6345481
{X05.614}	=>	{X05.561}	0,0345188	0,1696658	0,7760790
{X05.563}	=>	{X05.614}	0,0339958	0,2028081	0,9968358
{X05.658}	=>	{X05.562}	0,0334728	1,0000000	2,9689441
{X05.563}	=>	{X05.559}	0,0313808	0,1872075	0,7403117
{X05.570}	=>	{X05.562}	0,0303347	0,3186813	0,9461470
{X05.563}	=>	{X05.557}	0,0269352	0,1606864	0,6557790
{X05.565}	=>	{X05.563}	0,0251046	0,2990654	1,7841282
{X05.570}	=>	{X05.557}	0,0245816	0,2582418	1,0539130
{X05.556}	=>	{X05.563}	0,0235356	0,2301790	1,3731741
{X05.570}	=>	{X05.559}	0,0232741	0,2445055	0,9668966
{X05.570}	=>	{X05.563}	0,0214435	0,2252747	1,3439166
{X05.570}	=>	{X05.554}	0,0211820	0,2225275	0,9269554
{X05.556}	=>	{X05.614}	0,0196130	0,1918159	0,9428070
{X05.563}	=>	{X05.561}	0,0196130	0,1170047	0,5351984
{X05.570}	=>	{X05.561}	0,0188285	0,1978022	0,9047794
{X05.556}	=>	{X05.562}	0,0164749	0,1611253	0,4783721
{X05.565}	=>	{X05.562}	0,0143829	0,1713396	0,5086976
{X05.570}	=>	{X05.614}	0,0138598	0,1456044	0,7156699
{X05.563}	=>	{X05.554}	0,0138598	0,0826833	0,3444237
{X05.658}	=>	{X05.557}	0,0130753	0,3906250	1,5941836
{X05.556}	=>	{X05.559}	0,0125523	0,1227622	0,4854627
{X05.556}	=>	{X05.557}	0,0122908	0,1202046	0,4905682
{X05.658}	=>	{X05.554}	0,0120293	0,3593750	1,4970044
{X05.658}	=>	{X05.559}	0,0117678	0,3515625	1,3902534
{X05.565}	=>	{X05.557}	0,0117678	0,1401869	0,5721182
{X05.573}	=>	{X05.565}	0,0115063	0,2820513	3,3600128

{X05.565}	=>	{X05.559}	0,0115063	0,1370716	0,5420496
{X05.565}	=>	{X05.570}	0,0112448	0,1339563	1,4072781
{X05.556}	=>	{X05.561}	0,0109833	0,1074168	0,4913423
{X05.558}	=>	{X05.557}	0,0101987	0,2785714	1,1368806

Taula 12. Pregunta 1. Regles podades.

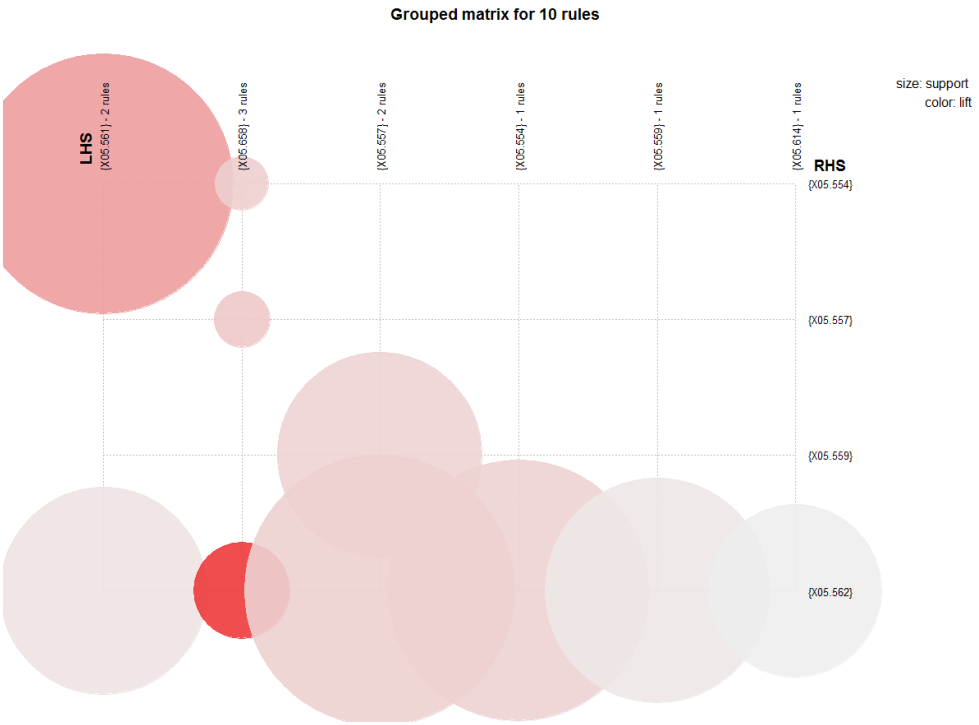
Ordenarem les regles per '*confiança'* i mirem les 5 primeres: Codi.

lhs		rhs	support	confidence	lift
{X05.658}	=>	{X05.562}	0,03347280	1,000000	2,968944
{X05.561}	=>	{X05.554}	0,11532427	0,527512	2,197392
{X05.557}	=>	{X05.562}	0,12081590	0,493063	1,463876
{X05.554}	=>	{X05.562}	0,11558577	0,481482	1,429492
{X05.561}	=>	{X05.562}	0,08891213	0,406699	1,207465

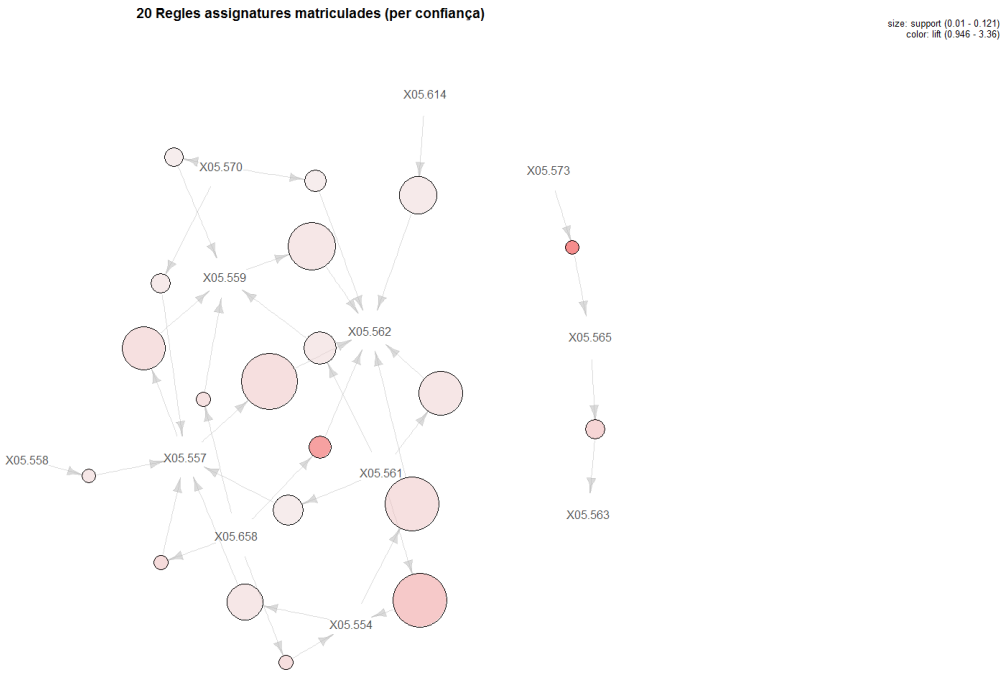
Taula 13. Pregunta 1. Regles podades ordenades per '*confiança'*

Podem dir que aquestes son les regles mes interesants. Veiem que quan un alumne es matricula de X05.658 la confiança es del 100%, és a dir **SEMPRE** es matricula de X05.562.

També els alumnes que es matriculen de X05.561 es matriculen de X05.554, però nomes la meitat de les vegades.



Il·lustració 21. Pregunta 1. Gràfic de les regles podades'.



Il·lustració 22. Pregunta 1. Gràfic de les 20 regles amb més confiança'.

Les 20 regles amb mes confiança mostren com hi ha tres nuclis d'assignatures que actuen principalment com a conseqüent i son:

- X05.562
- X05.557
- X05.559

Un altre mètode de classificació, K-means, és un mètode de classificació per a agrupar elements, que divideix el conjunt de n observacions en k grups de tal forma que cada observació pertany al grup el valor mitjà del qual és més proper.

S'ha fet una agrupació per a aquest conjunt de dades amb k = 8. Annex 11.15. Pregunta 1. K-means

El sistema ha generat 8 grups que es poden visualitzar a la següent taula:

Cluster	Freq.	Assignatures del Cluster
1	436	X05.563,X05.583,X05.590
2	914	X05.559,X05.562,X05.588,X05.658
3	209	X05.555,X05.565,X05.569,X05.585
4	90	X05.564,X05.570,X05.573,X05.587
5	728	X05.554,X05.561
6	95	X05.557,X05.572

7	1189	X05.567,X05.575,X05.576,X05.577,X05.578,X05.580,X05.582,X05.584, X05.589,X05.592,X05.593,X05.594,X05.595,X05.596,X05.597,X05.598, X05.599,X05.600,X05.614,X05.616
8	163	X05.556,X05.558,X05.560,X05.566,X05.568,X05.571,X05.574,X05.579, X05.586,X05.591,X05.601,X05.604,X05.607,X05.611,X05.613

Taula 14. Pregunta1. Taula K-means

Els clústers generat indiquen que el grup 7 és el que matricules agrupa (1189) però no vol dir que aquestes matricules continguin totes les assignatures incloses al grup sinó que les matricules contenen al menys una assignatura del clúster.

Observant els grups amb menys assignatures veiem el grup 5 que correspon amb una de les regles recentment esmentada al igual que la regla vista amb confiança 100% la veiem en el clúster 2, però barrejada amb altres assignatures.

En resum el clustering amb K-means amb un conjunt de dades dispers agrupa les dades de forma molt generalitzada i poca utilitat. Caldria generar molts mes grups per a obtenir resultats mes precisos.

6.6.2 Pregunta 2. Quines son 'letals'?

Estem davant d'un cas similar a l'anterior, només que igual que abans hem construït el model amb les assignatures matriculades, en aquest punt ho farem amb les assignatures suspeses. Construïm llavors un arxiu de transaccions amb les assignatures que s'han suspès de cada matriculació. Es consideren com suspeses les assignatures que tenen un '2' a la seva columna. Per tant transformarem:

- 2 -> 1
- 0,1,3 –> 0

Donat que probablement algunes assignatures no s'han suspès mai o alguns alumnes ho han aprovat tot, es generen files i/o columnes amb valor '0', eliminem files i columnes amb aquest contingut. En qualsevol cas no afecte als resultats .

Annex 11.16. Pregunta 2. Conversió de les dades.

Interpretem aquestes dades de la següent manera:
Tenim 958 files, si el total és 3824-958 = 2866 Matricules que no han suspès cap assignatura
37 columnes que corresponen a les 57 assignatures menys les 20 que no han estat suspeses per cap alumne.

L'informe ens diu que la densitat és de 0,03275405. La densitat indica el nombre de cel·les de la matriu que contenen valor, que són 958 * 37 * 0,03275405 = 1161 assignatures suspeses. Les assignatures més letals, és a dir que han estat mes vegades suspeses pels alumnes són:

Assignatura	Freqüència absoluta	Freqüència relativa
X05.554	191	19,9 %
X05.562	139	14,5 %
X05.557	137	14,3 %
X05.614	130	13,6 %
X05.559	97	10,1 %

Taula 15. Pregunta 2. 5 assignatures mes suspeses

Por contingut de cada matricula

Numero de suspensos por matricula	Freqüència
1	781
2	152
3	24
4	1

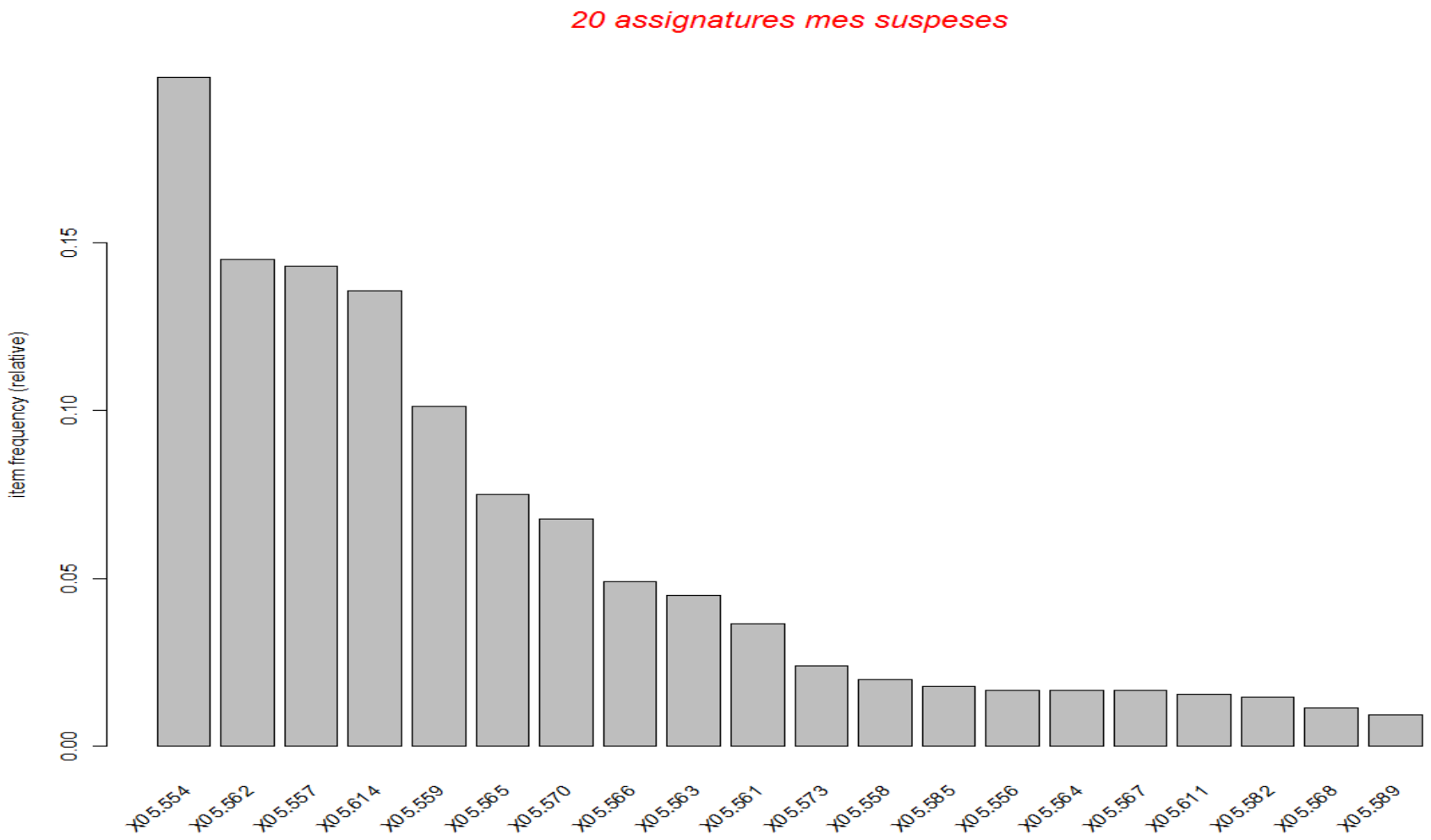
Taula 16. Pregunta 2. Distribució freqüències per numero d'assignatures suspeses.

El màxim nombre de suspesos en una matricula és de 4 assignatures.

Annex 11.17. Pregunta 2. Taula d'assignatures mes suspeses ordenades per freqüència.

I també una gràfica de les 20 assignatures mes suspeses

```
itemFrequencyPlot(trx,topN=20)
```



Il·lustració 23. Pregunta 2. 20 assignatures mes suspeses ordenades per freqüència

Podem cercar si existeixen combinacions letals, com en el cas anterior, tornem a aplicar l'algoritme 'apiori'.
Cerquem regles amb un suport i confiança tots dos de l'1% i regles amb mes de 1 item, ja que les d'1 item ja les hem analitzat prèviament. Les ordenem per suport.

Annex 11.18. Pregunta 2. Obtenció de regles d'associació 'apriori'

Hem obtingut 8 regles totes de longitud 2

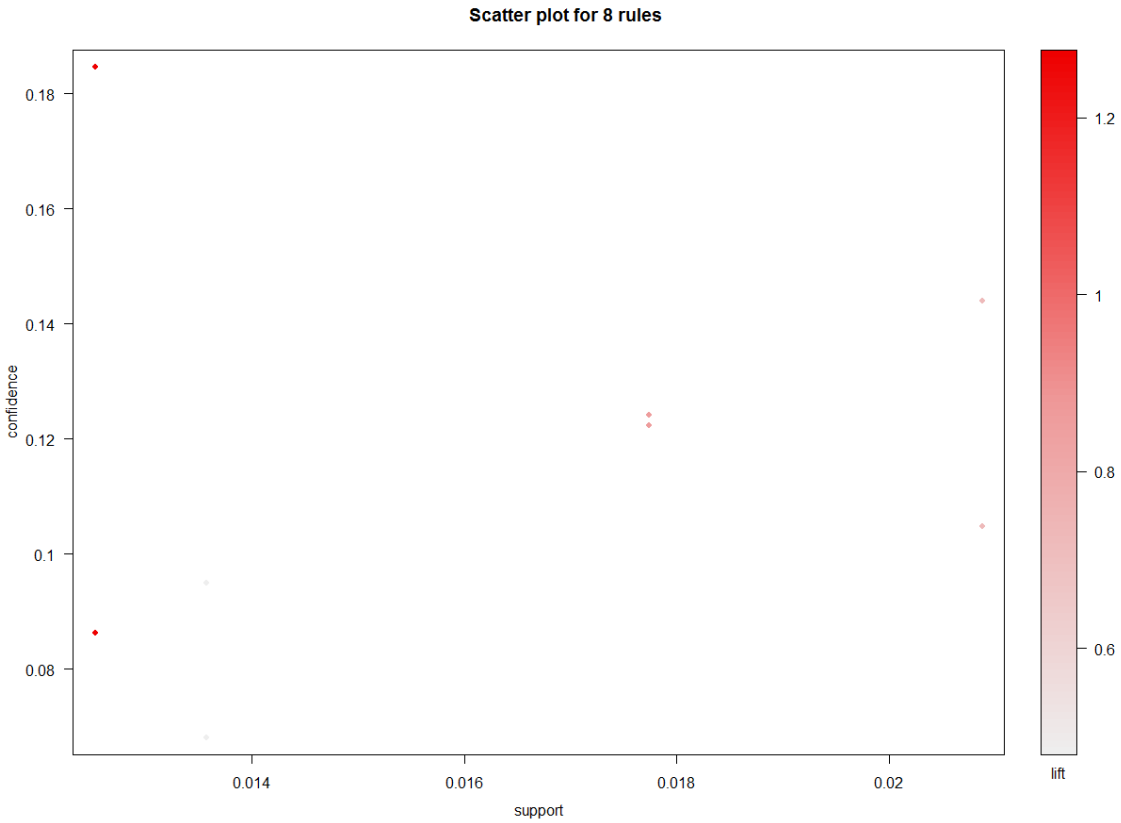
lhs		Rhs	Support	Confidence	lift
{X05,562}	=>	{X05,554}	0,02087683	0,1438849	0,7216844
{X05,554}	=>	{X05,562}	0,02087683	0,1047120	0,7216844
{X05,557}	=>	{X05,562}	0,01774530	0,1240876	0,8552224
{X05,562}	=>	{X05,557}	0,01774530	0,1223022	0,8552224
{X05,557}	=>	{X05,554}	0,01356994	0,0948905	0,4759430
{X05,554}	=>	{X05,557}	0,01356994	0,0680628	0,4759430
{X05,570}	=>	{X05,562}	0,01252610	0,1846154	1,2723852
{X05,562}	=>	{X05,570}	0,01252610	0,0863309	1,2723852

Taula 17. Pregunta 2. Regles ordenades per 'support'

La combinació més típica de suspensos {X05_562} i {X05_554}, cosa que ocorre un 2,08% de les matrícules. Quan algú suspèn {X05_562} el 14.39% de les vegades suspèn també {X05_554}.

No obstant això, les regles que es mostren en realitat són regles 'febles' ja que mostren una probabilitat de poc més del 2%, és a dir, són molt poc freqüents, tenen nivells molt baixos de confiança i lift<1.

```
plot(reglas)
```



Il·lustració 24. Pregunta 2. Gràfic de regles 'apriori'.

Farem també el K-Means amb 10 clústers. Veure codi

La taula en aquest cas és aquesta:

Cluster	Freq.	Assignatures del Cluster
1	14	X05.568,X05.611
2	42	X05.563,X05.572
3	60	X05.562,X05.567,X05.570,X05.589
4	605	X05.554,X05.555,X05.557,X05.559,X05.560,X05.561,X05.566,X05.569, X05.571,X05.575,X05.578,X05.582,X05.586,X05.587,X05.590,X05.593, X05.601,X05.607,X05.613,X05.616
5	26	X05.556,X05.564,X05.604
6	124	X05.614
7	23	X05.558,X05.573
8	64	X05.565,X05.585,X05.591

Taula 18. Pregunta1. Taula K-means

El clúster amb mes matricules és el 4 (605) i inclou l'assignatura mes suspesa (X05.554), però conté moltes assignatures, el que ens impedeix trobar patrons, donat que sabem per

l'anàlisi inicial de les dades que els alumnes suspenen 4 assignatures com a màxim, tot i que la mitjana de suspensos és 1.

Com a la pregunta 1, els clústers agrupen les dades de forma molt generalitzada i poca utilitat. Caldria generar molts mes grups per a obtenir millors resultats.

6.6.3 Pregunta 3. Quines assignatures tenen més incidència en altres quan es matriculen conjuntament?

D'aquest enunciat deduïm que es vol conèixer cóm el fet de matricular-se d'una assignatura pot tenir relació o incidència en altres assignatures. Fent un plantejament genèric podem desenvolupar un sistema que pugui determinar o descobrir influencia entre fets vinculats a la matricula.

En el procés de matriculació l'alumne decideix optar per un conjunt d'assignatures que finalment aprova, suspèn o fins i tot ni arriba a presentar-se a l'examen. Per tant tenim 5 fets que poden produir-se durant el curs:

- No matricular-se d'assignatures (0)
- No presentar-se a l'examen (1)
- Suspendre l'assignatura (2)
- Aprovar l'assignatura (3)
- Matricular-se d'una assignatura (1,2 o 3)

El plantejament per a aquesta pregunta seria construir un model que permeti relacionar aquest fets. És a dir poder analitzar que passa quan un estudiant suspèn la assignatura A_i, relacionat amb el fet d'aprovar A_j.

Hi ha relació entre el fets quan un estudiant es matricula d'una assignatura A_i respecte de aprovar A_j? I amb suspendre A_k, amb $i \neq j, i \neq k$

Sota aquests requeriment i amb les restriccions de que es comparin sempre 2 fets i com afecta a parelles d'assignatures, bàsicament pel cost computacional que suposaria un plantejament mes ampli, es dissenya un procés capaç de modelar els criteris comentats.

Es pretén generar una estructura de dades que mostraria a partir de les dades originals, el següent:

X05.554/X05.561	X05.554/X05.562	X05.554/X05.563	X05.554/X05.564	X05.554/X05.565	X05.554/X05.566	X05.554/X05.567	X05.554/X05.568	X05.554/X05.570
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Il·lustració 25. Pregunta 3. Estructura d'arxiu a generar

El model necessita conèixer dos del 5 fets esmentats. A continuació compara matricula a matricula i assignatura a assignatura si es compleix el fet o no per a cada parella d'assignatures.

- Sempre es compara entre assignatures diferents per que algunes combinacions no tindrien sentit, com per exemple la relació suspendre A1 amb aprovar A1, o be no matricular-se d'A1 i suspendre o aprovar A1.

El resultat de la comparació es materialitza com un '0' o un '1' que s'assignarà a la columna corresponent de la taula a generar. Això significa que en tindrem tantes columnes a la taula com combinacions de parelles d'assignatures. Per a 3 assignatures hi hauria les sis columnes A1/A2, A1/A3, A2/A1, A2/A3, A3/A1, A3/A2. Per tant per a n assignatures tindrem $n^2 - n$ columnes, en aquest cas n = 57, $columnes = 57^2 - 57 = 3192$.

Un cop generat l'arxiu s'eliminen les columnes i files que no son certes en cap del seus valors. Això redueix molt la grandària de l'arxiu.

Annex 11.20. Pregunta 3. Transformació de les dades anàlisi Matricular-se -> Aprovar.

Finalment es fa un anàlisi amb el model associatiu que representa l'MBA, i es raona el resultats obtinguts. Associem els fets '*matricular-se*' i '*aprovar*' és a dir mirarem les parelles d'assignatures A_i, A_j a on $A_i = \{1,2,3\}, A_j = \{3\}, i \neq j$

Calculem les transaccions a partir de l'arxiu generat.

Annex 11.21. Pregunta 3. Genera transaccions 'apriori' anàlisi Matricular-se-> Aprovar.

El numero de matricules és 2386 i el de columnes (assignatures) és de 1029 un cop reduïdes les que no aporten informació al model.

La densitat 0.004079515 ens indica que tenim $2386 * 1029 * 0.004079515 = 10016$ casos certs.

La distribució d'elements arriba fins a 42 casos certs en una matricula. I hi ha 953 matricules amb 2 casos certs.

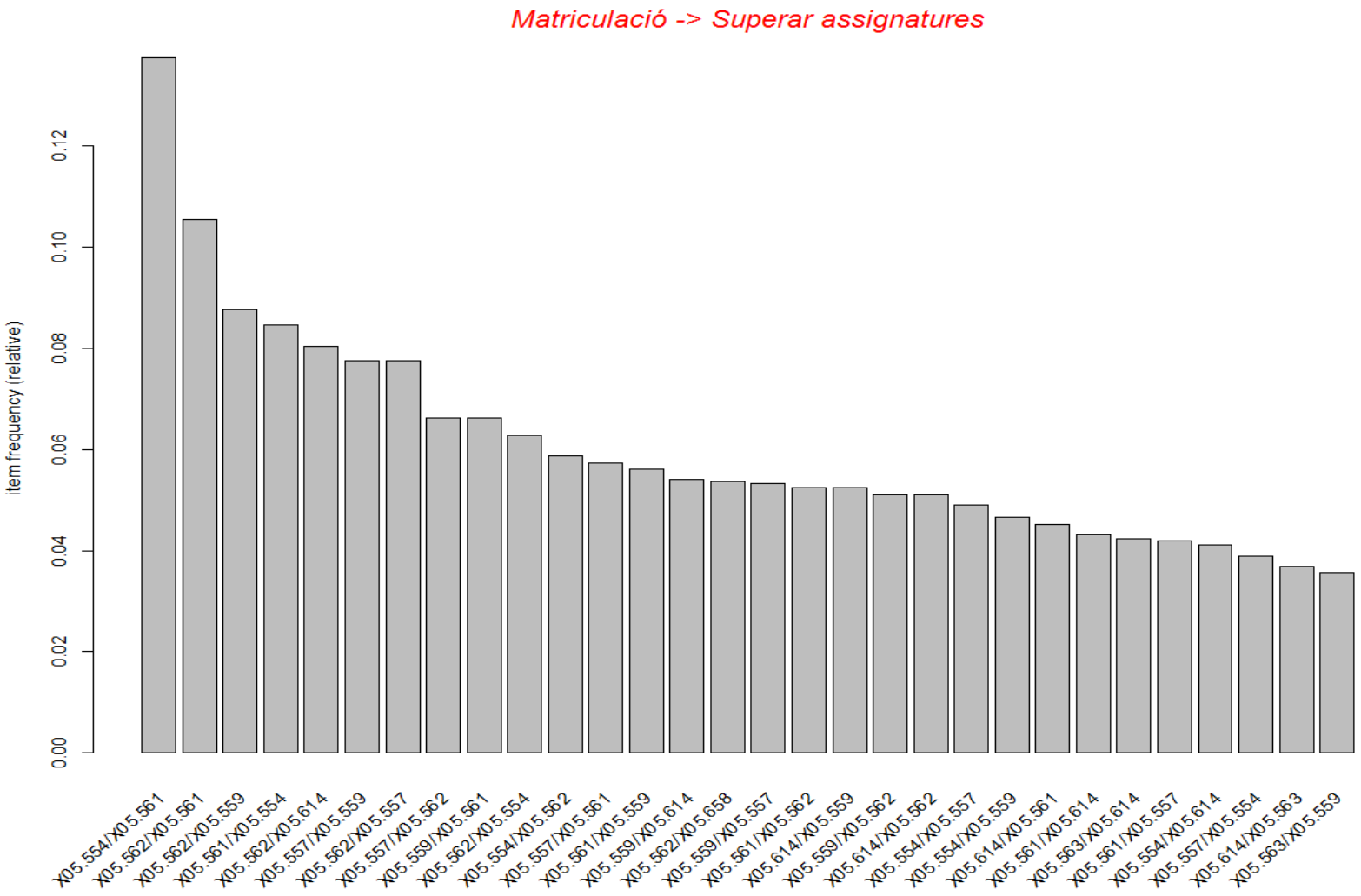
La informació essencial que cerquem son els items mes freqüents i s'ha d'interpretar com que:

Fet Antecedent	Assignatura	Fet Conseqüent	Assignatura	Freqüència	Freq. Rel.
Matricular-se	X05.554	Aprova	X05.561	328	13,7 %
Matricular-se	X05.562	Aprova	X05.561	252	10.6 %
Matricular-se	X05.562	Aprova	X05.559	209	8.7 %
Matricular-se	X05.561	Aprova	X05.554	202	8.5 %
Matricular-se	X05.562	Aprova	X05.614	192	8 %

Taula 19. Pregunta 3. Items mes freqüents anàlisi Matricular-se -> Aprovar

Com a cas mes rellevant podem afirmar que un 13,7 % del alumnes s'han matriculat de la assignatura X05.554 i tots aproven la assignatura X05.561. Un 10.6 % s'ha matriculat de X05.562 també aproven X05.561

Gràficament els 30 items mes freqüents.



Il·lustració 26. Pregunta 3. 40 transaccions mes freqüents anàlisi Matricular-se -> Aprovar

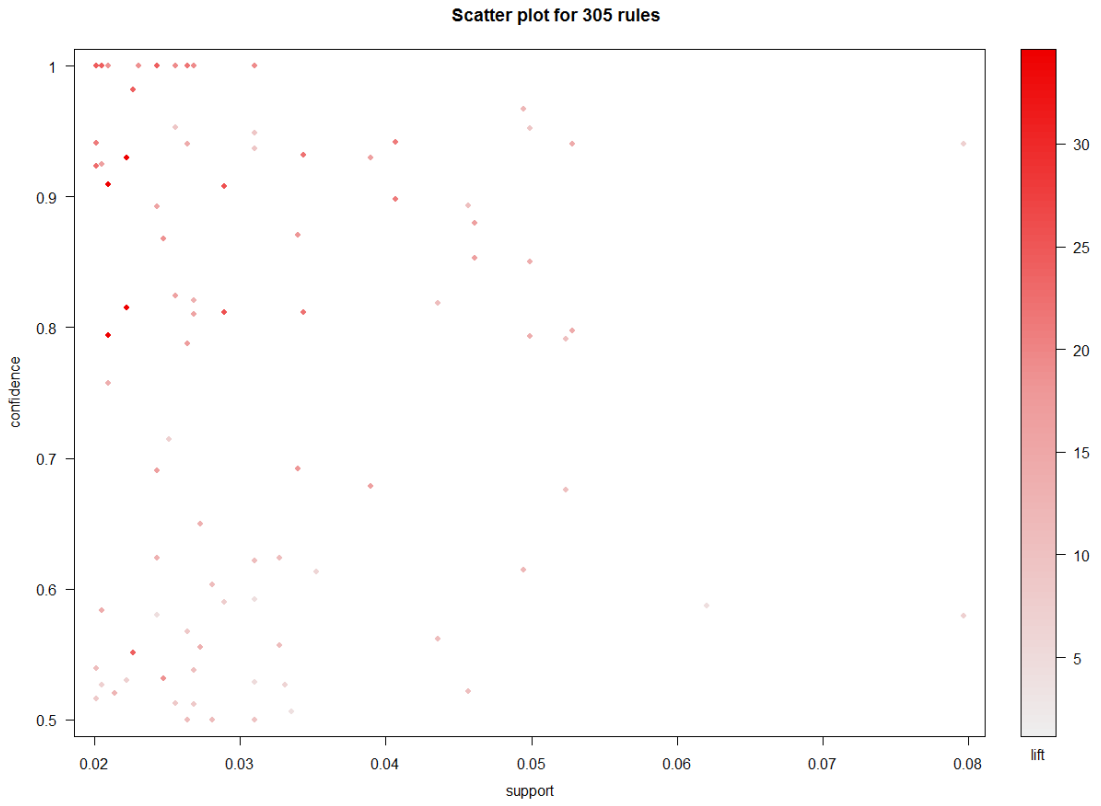
Aquest procés ens ha generat 'regles' per ell mateix, cada element és en sí una regla de un item com ja hem vist. És a dir que cada item implica la associació de dues assignatures . Per tant, sota aquest punt de vista no podem analitzar les regles amb confiança i lift, sinó nomes amb suport.

Disposem de la possibilitat de generar regles d'associació per a determinar si existeixen relacions entre aquests elements.

Annex 11.22. Pregunta 3. Generació regles '*apriori*' anàlisi Matricular-se-> Aprovar.

Obtenim un conjunt de mes 300 regles amb suport > 0.02 i confiança > 0.5

```
plot(reglas)
```



Il·lustració 27. Pregunta 3. Gràfic de les regles del anàlisi Matricular-se -> Aprovar.

Mostrem les primeres 100 ordenades per confiança.

```
reglas <-sort(reglas, by="conf", decreasing=TRUE) # ordena reglas
```

```
inspect(head(reglas,100))
```

Annex 11.23. Pregunta 3. 100 regles amb mes confiança de l'anàlisi Matricular-se-> Aprovar.

Es rellevant que les primeres 100 regles tinguin una confiança de 100% que vol dir que les regles es compleixen sempre.

Hi ha regles de 2 i 3 items, tot i que en tenim regles de fins a 6 elements. Si analitzem els valors de confiança son molt alts, en algun cas del 100% i els valors de lift son significativament alts, indicant que la freqüència de les regles és molt mes alta que l'esperada. El fet que el suport màxim és del 8% tot i que no és un valor destacable, si tenim en compte el numero d'assignatures i la dispersió de les matricules realitzades podem dir que si que hi ha relacions significatives entre els fets de matricular-se i aprovar. Trobem com era d'esperar, però, hi ha relacions inverses. La regla amb support mes gran:

{X05.561/X05.554} => {X05.554/X05.561}

Aquesta regla explica que si ens matriculem de la assignatura X05.561 i aprovem X05.554 llavors també ens matriculem de X05.554 i la aprovem amb una confiança del 94%. També tenim la inversa:

{X05.554/X05.561} =>{X05.561/X05.554}

Que es compleix amb una confiança del 58%. En realitat aquestes regles expressarien el fet de que X0.554 i X0.561 son assignatures de les que els estudiants es matriculen de forma conjunta i aproven de forma conjunta.

Es una mica rebuscat intentar treure informació rellevant a partir d'aquestes associacions per que cada antecedent i conseqüent son en sí ja una regla calculada que té sentit de forma individual. Però en qualsevol cas aquestes regles indiquen fets que estan succeint de forma conjunta.

Finalment farem poda del resultats per a eliminar regles que son dependents i estan incloses dins d'altres regles.

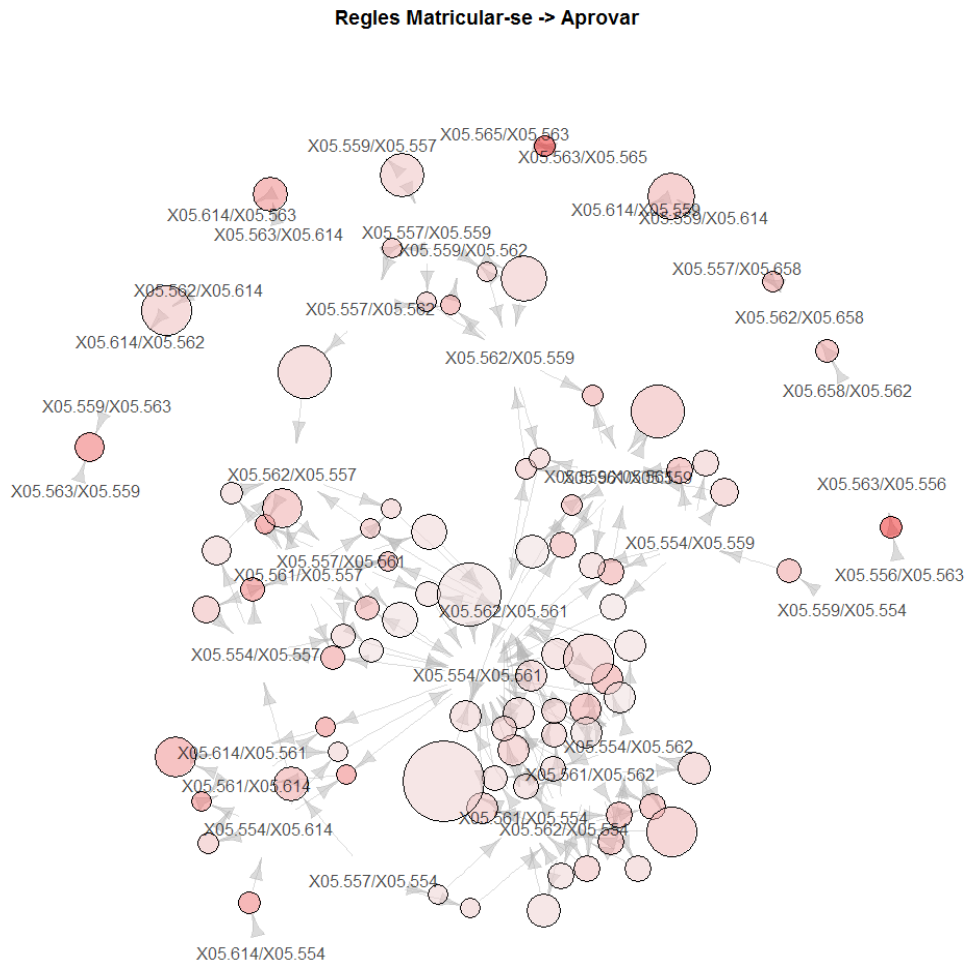
Annex 11.24. Pregunta 3. Poda regles de l'anàlisi Matricular-se -> Aprovar

Obtenim 84 regles.

lhs		rhs	support	Conf.	lift
{X05.557/X05.658}	=>	{X05.562/X05.658}	0.02095557	1,000000	18,64063
{X05.658/X05.562}	=>	{X05.562/X05.658}	0.02305113	1,000000	18,64063
{X05.554/X05.614,X05.614/X05.561}	=>	{X05.561/X05.614}	0.02011735	1,000000	23,16505
{X05.554/X05.561,X05.554/X05.614}	=>	{X05.561/X05.614}	0.02011735	1,000000	23,16505
{X05.554/X05.614,X05.614/X05.561}	=>	{X05.554/X05.561}	0.02011735	1,000000	7,27439
{X05.554/X05.561,X05.561/X05.614}	=>	{X05.614/X05.561}	0.02011735	1,000000	22,09259
{X05.554/X05.557,X05.557/X05.561}	=>	{X05.561/X05.557}	0.02430847	1,000000	23,86000
{X05.554/X05.561,X05.561/X05.557}	=>	{X05.554/X05.557}	0.02430847	1,000000	20,39316
{X05.557/X05.561,X05.562/X05.557}	=>	{X05.561/X05.557}	0.02053646	1,000000	23,86000
{X05.561/X05.557,X05.562/X05.561}	=>	{X05.557/X05.561}	0.02053646	1,000000	17,41606
{X05.554/X05.561,X05.561/X05.557}	=>	{X05.557/X05.561}	0.02430847	1,000000	17,41606
{X05.561/X05.557,X05.562/X05.561}	=>	{X05.562/X05.557}	0.02053646	1,000000	12,89730
{X05.554/X05.559,X05.559/X05.561}	=>	{X05.561/X05.559}	0.02640402	1,000000	17,80597
{X05.554/X05.559,X05.554/X05.561}	=>	{X05.561/X05.559}	0.02640402	1,000000	17,80597
{X05.554/X05.559,X05.559/X05.561}	=>	{X05.554/X05.561}	0.02640402	1,000000	7,27439
{X05.557/X05.559,X05.559/X05.562}	=>	{X05.557/X05.562}	0.02011735	1,000000	15,10127
{X05.557/X05.559,X05.559/X05.562}	=>	{X05.562/X05.559}	0.02011735	1,000000	11,41627
{X05.557/X05.562,X05.562/X05.559}	=>	{X05.559/X05.562}	0.02011735	1,000000	19,55738
{X05.554/X05.557,X05.557/X05.561}	=>	{X05.554/X05.561}	0.02430847	1,000000	7,27439
{X05.559/X05.561,X05.562/X05.559}	=>	{X05.561/X05.559}	0.02095557	1,000000	17,80597
{X05.561/X05.559,X05.562/X05.561}	=>	{X05.559/X05.561}	0.02095557	1,000000	15,10127
{X05.554/X05.561,X05.561/X05.559}	=>	{X05.559/X05.561}	0.02640402	1,000000	15,10127
{X05.561/X05.559,X05.562/X05.561}	=>	{X05.562/X05.559}	0.02095557	1,000000	11,41627
{X05.557/X05.561,X05.562/X05.557}	=>	{X05.562/X05.561}	0.02053646	1,000000	9,46825
{X05.561/X05.562,X05.562/X05.554}	=>	{X05.554/X05.562}	0.02682313	1,000000	17,04286
{X05.561/X05.554,X05.561/X05.562}	=>	{X05.554/X05.562}	0.02682313	1,000000	17,04286
{X05.554/X05.562,X05.562/X05.561}	=>	{X05.561/X05.562}	0.03101425	1,000000	19,08800
{X05.554/X05.561,X05.561/X05.562}	=>	{X05.554/X05.562}	0.03101425	1,000000	17,04286
{X05.561/X05.562,X05.562/X05.554}	=>	{X05.561/X05.554}	0.02682313	1,000000	11,81188
{X05.554/X05.561,X05.561/X05.562}	=>	{X05.562/X05.561}	0.03101425	1,000000	9,46825
{X05.557/X05.559,X05.557/X05.562}	=>	{X05.562/X05.559}	0.02011735	1,000000	11,41627
{X05.559/X05.561,X05.562/X05.559}	=>	{X05.562/X05.561}	0.02095557	1,000000	9,46825
{X05.554/X05.562,X05.561/X05.554}	=>	{X05.562/X05.554}	0.02682313	1,000000	15,90667
{X05.554/X05.562,X05.562/X05.561}	=>	{X05.554/X05.561}	0.03101425	1,000000	7,27439
{X05.562/X05.554,X05.562/X05.561}	=>	{X05.561/X05.554}	0.03101425	1,000000	11,81188
{X05.554/X05.561,X05.562/X05.554}	=>	{X05.561/X05.554}	0.03101425	1,000000	11,81188
{X05.562/X05.554,X05.562/X05.561}	=>	{X05.554/X05.561}	0.03101425	1,000000	7,27439
{X05.561/X05.554,X05.562/X05.561}	=>	{X05.554/X05.561}	0.03101425	1,000000	7,27439
{X05.614/X05.554}	=>	{X05.554/X05.614}	0.02263202	0,981818	23,90427
{X05.614/X05.562}	=>	{X05.562/X05.614}	0.04945516	0,967213	12,01964
{X05.561/X05.562,X05.562/X05.554}	=>	{X05.562/X05.561}	0.02556580	0,953125	9,02443

{X05.561/X05.562,X05.562/X05.554}	=>	{X05.554/X05.561}	0.02556580	0,953125	6,93340
{X05.561/X05.554,X05.561/X05.562}	=>	{X05.562/X05.561}	0.02556580	0,953125	9,02443
{X05.561/X05.554,X05.561/X05.562}	=>	{X05.554/X05.561}	0.02556580	0,953125	6,93340
{X05.554/X05.562,X05.561/X05.554}	=>	{X05.562/X05.561}	0.02556580	0,953125	9,02443
{X05.554/X05.562,X05.561/X05.554}	=>	{X05.554/X05.561}	0.02556580	0,953125	6,93340
{X05.561/X05.562}	=>	{X05.562/X05.561}	0.04987427	0,952000	9,01378
{X05.561/X05.614}	=>	{X05.614/X05.561}	0.04065381	0,941748	20,80565
{X05.561/X05.554}	=>	{X05.554/X05.561}	0.07963118	0,940594	6,84225
{X05.561/X05.559}	=>	{X05.559/X05.561}	0.05280805	0,940299	14,19970
{X05.614/X05.563}	=>	{X05.563/X05.614}	0.03436714	0,931818	22,01305
{X05.561/X05.557}	=>	{X05.557/X05.561}	0.03897737	0,930000	16,19693
{X05.556/X05.563}	=>	{X05.563/X05.556}	0.02221291	0,929825	34,13171
{X05.563/X05.565}	=>	{X05.565/X05.563}	0.02095557	0,909091	34,43001
{X05.559/X05.563}	=>	{X05.563/X05.559}	0.02891869	0,907895	25,48514
{X05.559/X05.562}	=>	{X05.562/X05.559}	0.04568315	0,893443	10,19978
{X05.614/X05.559}	=>	{X05.559/X05.614}	0.04610226	0,880000	16,27659
{X05.557/X05.554}	=>	{X05.554/X05.557}	0.03394803	0,870968	17,76179
{X05.559/X05.554}	=>	{X05.554/X05.559}	0.02472758	0,867647	18,65050
{X05.554/X05.562}	=>	{X05.562/X05.554}	0.04987427	0,850000	13,52067
{X05.559/X05.557}	=>	{X05.557/X05.559}	0.04358759	0,818898	10,56157
{X05.557/X05.562}	=>	{X05.562/X05.557}	0.05238894	0,791139	10,20356
{X05.557/X05.561,X05.562/X05.561}	=>	{X05.554/X05.561}	0.02514669	0,714286	5,19599
{X05.561/X05.557}	=>	{X05.554/X05.557}	0.02724225	0,650000	13,25556
{X05.561/X05.562}	=>	{X05.554/X05.562}	0.03269070	0,624000	10,63474
{X05.557/X05.561}	=>	{X05.562/X05.561}	0.03520536	0,613139	5,80535
{X05.557/X05.561}	=>	{X05.554/X05.561}	0.03520536	0,613139	4,46021
{X05.554/X05.559}	=>	{X05.561/X05.559}	0.02808047	0,603604	10,74775
{X05.561/X05.562}	=>	{X05.554/X05.561}	0.03101425	0,592000	4,30644
{X05.554/X05.557}	=>	{X05.562/X05.557}	0.02891869	0,589744	7,60610
{X05.562/X05.561}	=>	{X05.554/X05.561}	0.06202850	0,587302	4,27226
{X05.561/X05.557}	=>	{X05.554/X05.561}	0.02430847	0,580000	4,21915
{X05.554/X05.559}	=>	{X05.559/X05.561}	0.02640402	0,567568	8,57099
{X05.554/X05.559}	=>	{X05.554/X05.561}	0.02640402	0,567568	4,12871
{X05.561/X05.557}	=>	{X05.562/X05.557}	0.02221291	0,530000	6,83557
{X05.554/X05.562}	=>	{X05.562/X05.561}	0.03101425	0,528571	5,00465
{X05.554/X05.562}	=>	{X05.554/X05.561}	0.03101425	0,528571	3,84504
{X05.562/X05.554}	=>	{X05.561/X05.554}	0.03310981	0,526667	6,22092
{X05.554/X05.614}	=>	{X05.561/X05.614}	0.02137469	0,520408	12,05528
{X05.557/X05.554}	=>	{X05.562/X05.554}	0.02011735	0,516129	8,20989
{X05.557/X05.554}	=>	{X05.561/X05.554}	0.02011735	0,516129	6,09646
{X05.561/X05.562}	=>	{X05.562/X05.554}	0.02682313	0,512000	8,14421
{X05.561/X05.562}	=>	{X05.561/X05.554}	0.02682313	0,512000	6,04768
{X05.559/X05.561}	=>	{X05.554/X05.561}	0.03352892	0,506329	3,68324

Taula 20. Pregunta 3. Regles del anàlisi Matricular-se -> Aprovar (podades)



Il·lustració 28. Pregunta 3. Gràfic regles anàlisi Matricular-se -> Aprovar.

El gràfic mostra que una majoria de regles impliquen que l'alumne es matricula de la assignatura X05.554 i aprova la X05.561 amb una probabilitat molt elevada.

Si canviem les condicions i fem $A_i = \{1,2,3\}, A_j = \{2\}, i \neq j$, estem comparant assignatures matriculades i suspeses.

Annex 11.25. Pregunta 3. Transformació de les dades anàlisi Matricular-se -> Suspendre.

Un cop generat l'arxiu es generen les transaccions:

Annex 11.26. Pregunta 3. Genera transaccions 'apriori' anàlisi Matricular-se-> Suspendre.

El numero de columnes és ara de 898 i 471 files. Denota que òbviament hi ha menys suspensos que aprovats.

La densitat 0.00539297 ens indica que tenim $898 * 471 * 0.00539297 = 2281$ casos certs.

El cas mes freqüent es 1 cas cert per matricula i es repeteix 345 vegades

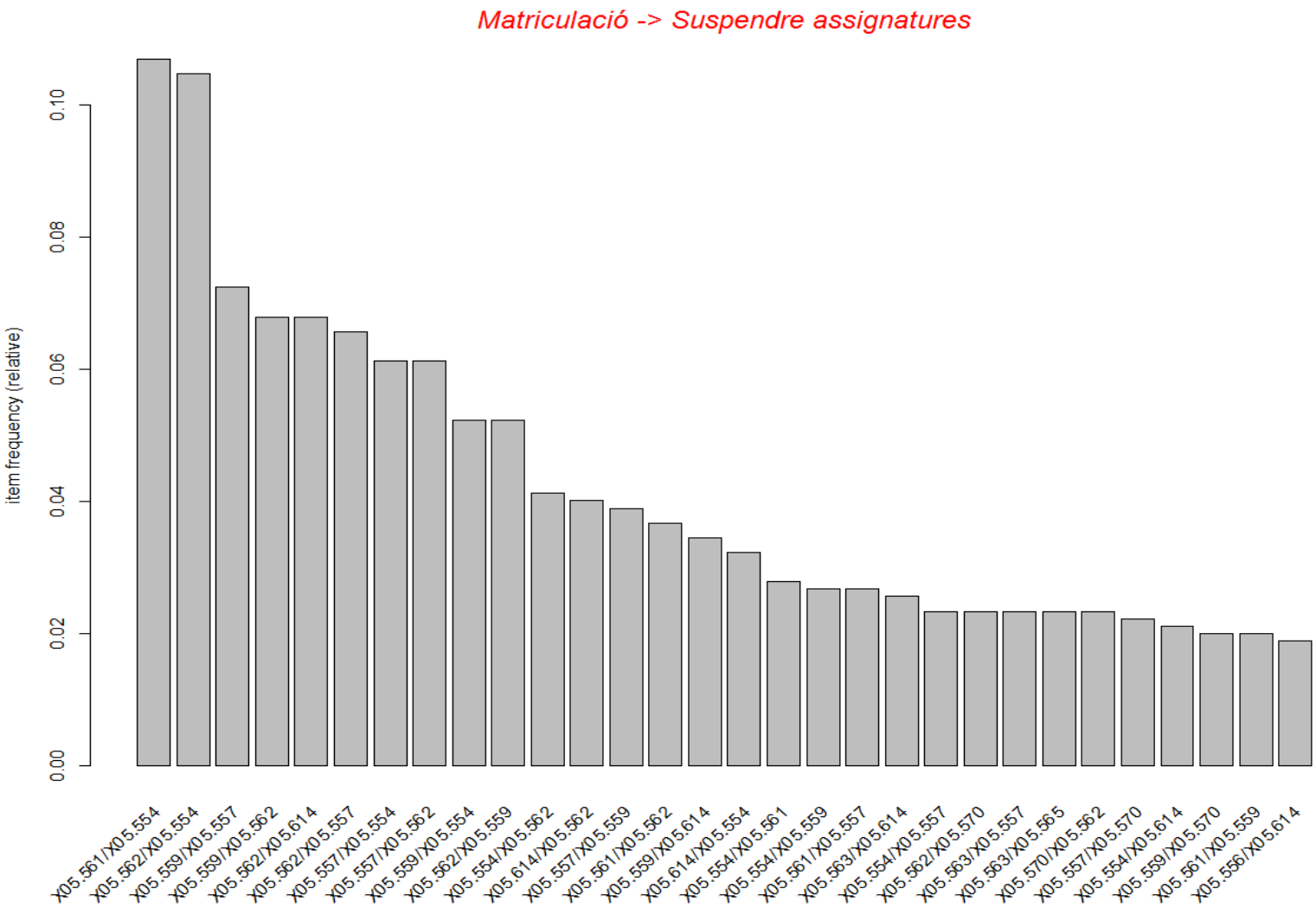
Els 5 items mes freqüents son:

Fet Antecedent	Assignatura	Fet Conseqüent	Assignatura	Freq.	Freq. Rel
Matricular-se	X05.561	Suspèn	X05.554	96	10,69 %
Matricular-se	X05.562	Suspèn	X05.554	94	10.47 %
Matricular-se	X05.559	Suspèn	X05.557	65	7,24 %
Matricular-se	X05.559	Suspèn	X05.562	61	6,79 %
Matricular-se	X05.562	Suspèn	X05.614	61	6,79 %

Taula 21. Pregunta 3. Items mes freqüents anàlisi Matricular-se -> Suspendre

Com a cas mes rellevant podem afirmar que es produeix en un 10,69 % de les matricules quan es matriculen de la assignatura X05.561 suspenen la assignatura X05.554. També en un % similar d'alumnes al matricular-se de l'assignatura X05.562 també suspenen X05.554. És a dir que matricular-se de les assignatures X05.561 i/o X05.562 té com a conseqüència suspendre X05.554 en un 10% dels alumnes.

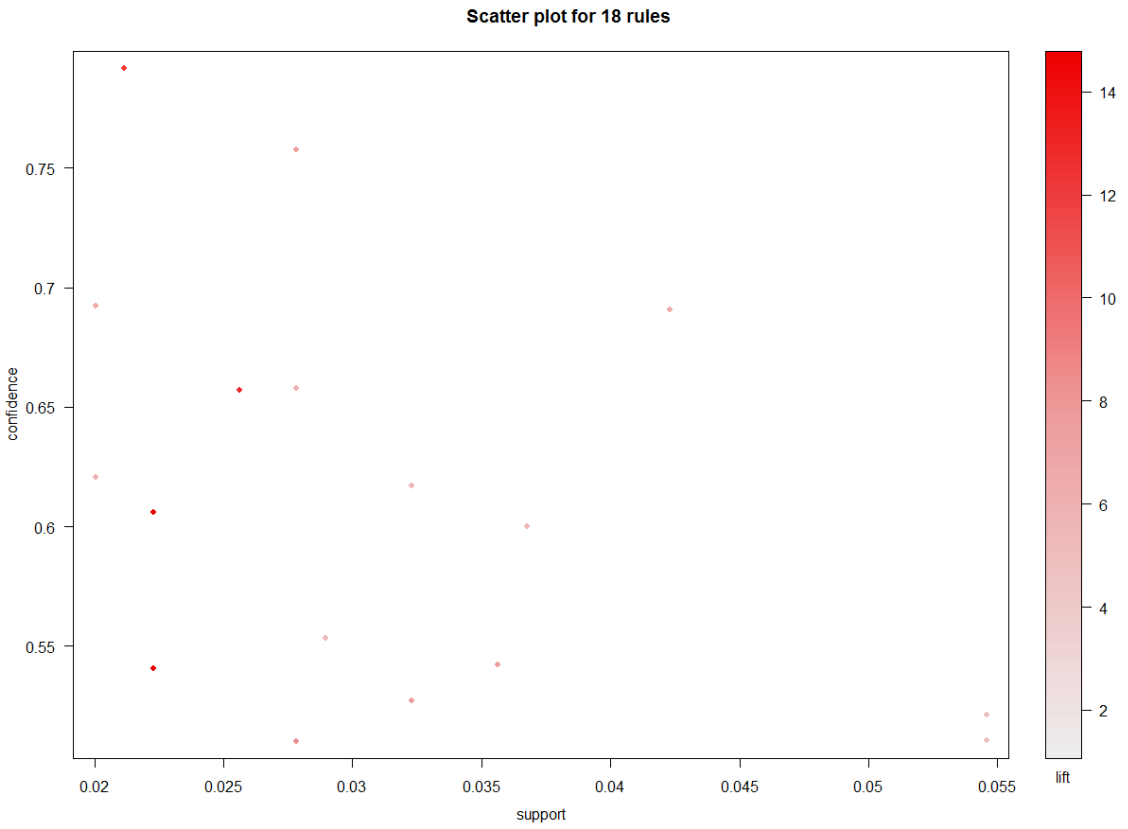
Gràficament els 30 items mes freqüents.



Taula 22. Pregunta 3. 30 transaccions mes freqüents anàlisi Matricular-se -> Suspendre

Con en el cas anterior es generen les regles:

Annex 11. Pregunta 3. Generació regles '*apriori*' anàlisi Matricular-se-> Suspendre.



Il·lustració 29. Pregunta 3. Gràfic de les regles del anàlisi Matricular-se -> Suspendre.

Obtenim un conjunt de 18 regles amb suport > 0.02 i confiança > 0.5

Es mostren les regles ordenades per confiança.

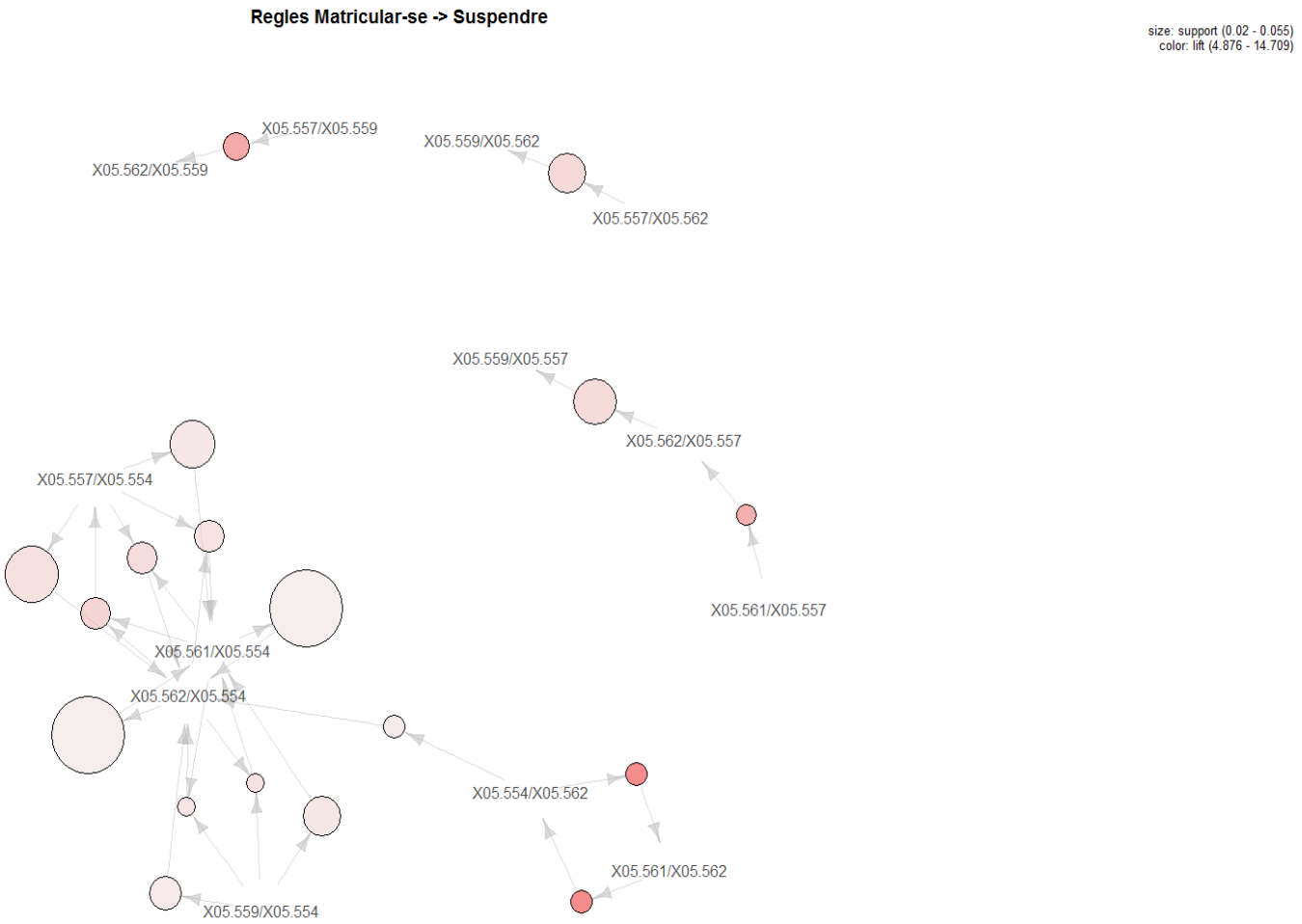
lhs		rhs	support	Conf.	lift
{X05.561/X05.557}	=>	{X05.562/X05.557}	0.02115	0,79166	12,049435
{X05.557/X05.554,X05.561/X05.554}	=>	{X05.562/X05.554}	0.02783	0,75757	7,237266
{X05.559/X05.554,X05.562/X05.554}	=>	{X05.561/X05.554}	0.02004	0,69230	6,475962
{X05.557/X05.554}	=>	{X05.562/X05.554}	0.04231	0,69090	6,600387
{X05.557/X05.554,X05.562/X05.554}	=>	{X05.561/X05.554}	0.02783	0,65789	6,154057
{X05.557/X05.559}	=>	{X05.562/X05.559}	0.02561	0,65714	12,555623
{X05.559/X05.554,X05.561/X05.554}	=>	{X05.562/X05.554}	0.02004	0,62069	5,929567
{X05.559/X05.554}	=>	{X05.561/X05.554}	0.03229	0,61702	5,771720
{X05.561/X05.562}	=>	{X05.554/X05.562}	0.02227	0,60606	14,709255
{X05.557/X05.554}	=>	{X05.561/X05.554}	0.03674	0,60000	5,612500

{X05.559/X05.554}	=>	{X05.562/X05.554}	0.02895	0,55319	5,284744
{X05.562/X05.557}	=>	{X05.559/X05.557}	0.03563	0,54237	7,493090
{X05.554/X05.562}	=>	{X05.561/X05.562}	0.02227	0,54054	14,70925
{X05.554/X05.562}	=>	{X05.562/X05.554}	0.02227	0,54054	5,163887
{X05.557/X05.562}	=>	{X05.559/X05.562}	0.03229	0,52727	7,762146
{X05.562/X05.554}	=>	{X05.561/X05.554}	0.05456	0,52127	4,876108
{X05.561/X05.554}	=>	{X05.562/X05.554}	0.05456	0,51041	4,876108
{X05.561/X05.554,X05.562/X05.554}	=>	{X05.557/X05.554}	0.02783	0,51020	8,330241

Taula 23.Pregunta 3. Taula de les regles del anàlisi Matricular-se -> Suspendre.

Hi ha regles de 2 i 3 items. En aquest cas els valors de confiança no son tant alts, si ho son els valors de '*lift*' indicant que les associacions apareixen amb una freqüència mes elevada que l'atzar. No hi tantes regles com en les matricules, i tot i que las confiança no es massa alta, indica que estem davant de regles amb prou rellevància com per a ser tingudes en consideració.

Veiem el gràfic:



Il·lustració 30. Gràfic regles anàlisi Matricular-se -> Aprovar.

Es veu com en el cas anterior, que hi ha 2 dues regles que apareixen de forma freqüent com a conseqüents de altres regles i aquestes son:

- X05.562/X05.554
- X05.561/X05.554

És a dir que com a conseqüència, alumnes que es matriculen de X05.562 i X05.561 acaben suspenen X05.554, que com hem vist en la pregunta 2 queda clar que la assignatura X05.554 és la que mes vegades es suspèn. En aquest anàlisi es verifica que efectivament és una de les assignatures que mes apareixen a les regles. Es verifica també que aquest fet te una certa relació amb matricular-se d'assignatures com ara X05.561, X05.557, X05.562 i X05.559.

Com a resum cal dir que aquests anàlisis es poden fer combinant qualsevol parella dels fets comentats anteriorment. Per tant es podria realitzar anàlisis que relacionin *'aprovar'* amb *'suspendre'*, *'presentar-se'* amb *'aprovar'* o *'suspendre'*, en general qualsevol parella d'accions que tingui sentit i òbviament tenint en consideració l'ordre en que es relacionen els fets, donat que les relacions poden ser diferents.

6.6.4 Pregunta 4. Quina relació hi ha entre la matrícula feta i la decisió de tornar-se a matricular el segon semestre?

Per a aquesta qüestió utilitzarem els arbres de decisió com a model de classificació. Els arbres ens permeten explicar i predir a partir de les dades.

Un arbre de decisió és un model de predicció que serveixen per representar i categoritzar una sèrie de condicions que ocorren de forma successiva, per a la resolució d'un problema.
Un arbre de decisió parteix d'un conjunt d'atributs d'entrada i retorna una resposta que és presa a partir de les entrades. Els valors que poden prendre les entrades i les sortides poden ser valors discrets o continus. Quan s'utilitzen valors discrets en les funcions d'una aplicació s'anomena classificació i quan s'utilitzen valors continus es denomina regressió.

Un arbre de decisió és una representació en forma d'arbre a on les branques es bifurquen en funció dels valors presos per les variables i que acaben en una acció concreta. Se sol utilitzar quan el nombre de condicions no és molt gran.

Els arbres de decisió són diagrames de decisions seqüencials que ens mostren els seus possibles resultats, ajudant a determinar quines són les seves opcions al mostrar les diferents decisions i els seus resultats.

En aquet punt intentarem analitzar la influencia que té la matricula feta, en el fet de tornar-se a matricular. Prendrem l'arxiu i farem les conversions pertinents per a generar l'arbre. Essencialment eliminem les variables sociodemogràfiques i partim les dades en un grup d'entrenament i un altre de test amb una proporció d'un 70% i 30 % respectivament. Indicarem en el nostre model de dades quines han estat les assignatures matriculades independentment del resultat obtingut al final del curs, raó per la qual aquest model predictiu pot ser utilitzar en el moment de la matriculació.

Utilitzem la llibreria *'rpart',* que farem servir per a generar el model classificador de l'arbre.

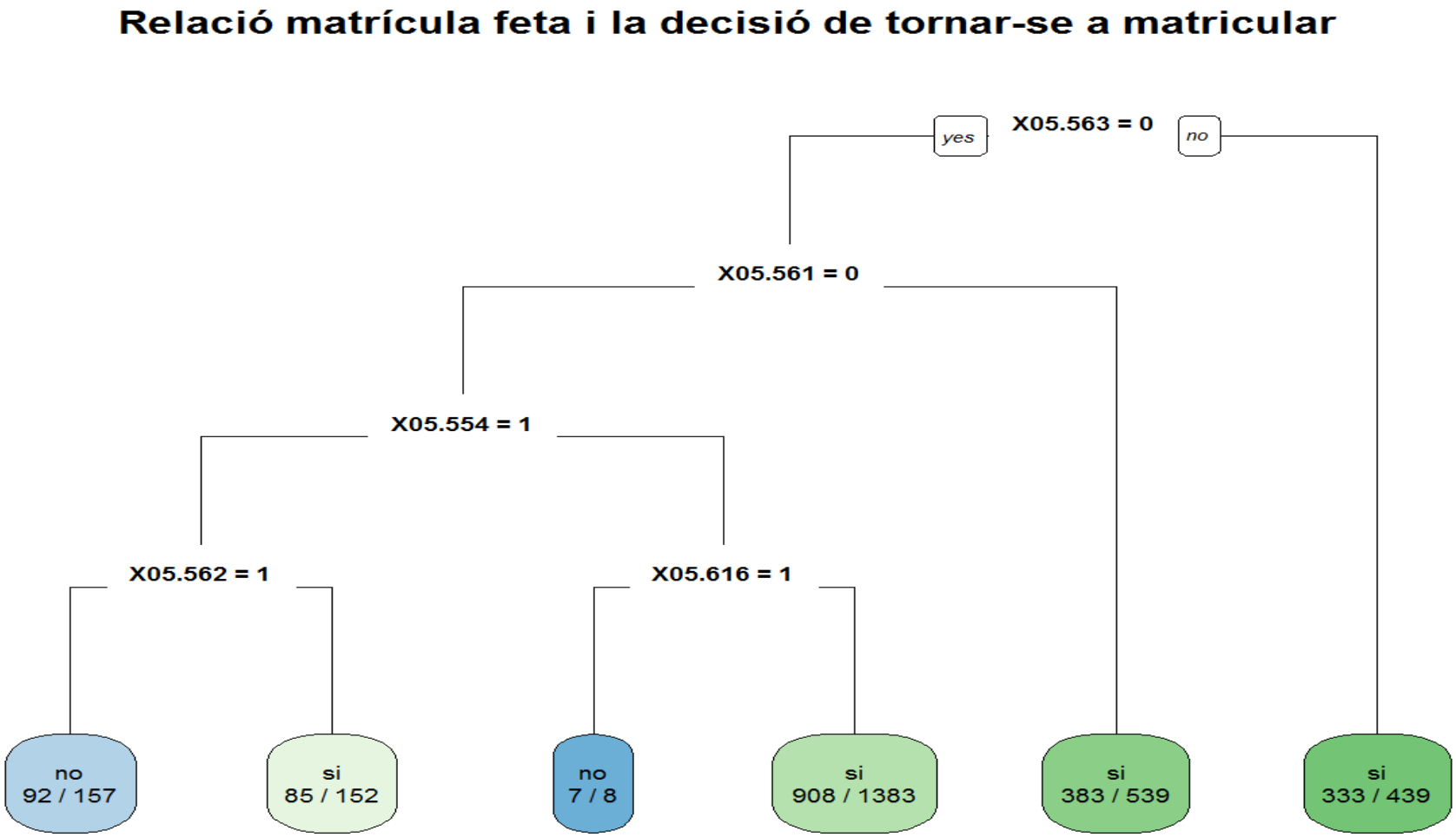
Transformem l'arxiu de dades original i generem un primer arbre separant les dades en un conjunt d'entrenament i un altre de validació.

Annex 11.29. Pregunta 4. Transformació de les dades i generació de l'arbre.

Per les assignatures transformem els valors 0,1,2,3 en nomes 0,1, indicant si l'alumne s'ha matriculat (1) o no (0).

Apliquem el model de classificació a les assignatures i a la variable depenent 'rematricula'.

L'arbre que ens ha generat el procés és el que veiem a continuació. Donat que els conjunts generats d'entrenament i test en prenen de forma aleatòria, l'arbre general cada vegada pot ser diferent.



Il·lustració 31. Pregunta 4. Generació de l'arbre

Com assignatura amb mes pes, l'arbre ha triat la X05.563. Mirem la matriu de confusió fent una predicció.

Annex 11.30. Pregunta 4. Predicció.

	no	si
no	34	352
si	29	731

Taula 24. Pregunta 4. Matriu de confusió

L'error obtingut és 0.3324607, és un error alt i resulta evident que l'arbre no esta funcionant correctament doncs la branca del '*no*' està classificant correctament nomes 34 casos de 386, que implica un error del 91,2% en els casos del '*no*'.

En el '*si*' l'error és nomes del 3,8%. És a dir l'arbre classifica millor els '*si*', però no ho fa amb els '*no*'. El '*si*' és majoritari amb respecte al '*no*', fet pel qual es produeix aquesta descompensació. D'altra banda l'error global de l'arbre és del 33,7%, és a dir si fem la predicció de que tot és '*si*' cometríem un error del 33,7% que és quasi el mateix error que obtenim ara.

Farem una poda de l'arbre per a simplificar-lo si és possible. mirem com abans els valors del error de validació per determinar el coeficient de complexitat a aplicar a la poda.

Annex 11.31. Pregunta 4. Poda de l'arbre.

```
printcp(CART)
```

	CP	nsplit	Rel error	xerror	xstd
1	0.0074751	0	1,000000	1,000000	0,027093
2	0.0066445	4	0.97010	0.99668	0.027070
3	0,0020000	5	0.96346	0.98450	0.026988

Taula 25. Pregunta 4. Taula de complexitat de l'arbre

El valor mínim de l'error està a l'últim nivell de l'arbre per tant no podem fer la poda.

Les regles generades son les següents:

```
Rule number: 3 [rematricula=si cover=439 (16%) prob=0.76]
```

```
   X05.563=1

Rule number: 5 [rematricula=si cover=539 (20%) prob=0.71]
   X05.563=0
   X05.561=1

Rule number: 19 [rematricula=si cover=1383 (52%) prob=0.66]
   X05.563=0
   X05.561=0
   X05.554=0
   X05.616=0

Rule number: 17 [rematricula=si cover=152 (6%) prob=0.56]
   X05.563=0
   X05.561=0
   X05.554=1
   X05.562=0

Rule number: 16 [rematricula=no cover=157 (6%) prob=0.41]
   X05.563=0
   X05.561=0
   X05.554=1
   X05.562=1

Rule number: 18 [rematricula=no cover=8 (0%) prob=0.12]
   X05.563=0
   X05.561=0
   X05.554=0
   X05.616=1
```

Taula 26. Pregunta 4. Regles de l'arbre

Les regles les podem interpretar de la següent forma:

Partim de 2678 mostres de les quals 903 no es rematriculen i 1775 si (33,7% i 66,3 %)

- Si rematriculen els alumnes que s'han matriculat de X05.563.
- Si rematriculen els alumnes que no es matriculen de X05.563 però si de X05.561.
- Si rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 ni de X05.554 ni de X05.554.
- Si rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 ni de X05.562 però si de X05.554.
- NO rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 i si de X05.562 i també de X05.554.
- NO rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 ni de X05.554 però si de X05.616.

Finalment executarem un procés de validació creuada i podrem confirmar que el valor de qualitat és correcte.

Annex 11.31. Pregunta 4. Validació creuada.

Les matrius generades:

1	no	si		2	no	si
no	0	128		no	0	137
si	0	254		si	0	245
3	no	si		4	no	si
no	0	127		no	0	122
si	0	255		si	0	260
5	no	si		6	no	si
no	0	127		no	0	128
si	0	255		si	0	254
7	no	si		8	no	si
no	0	148		no	0	117
si	0	234		si	0	265
9	no	si		10	no	si
no	0	131		no	0	122
si	0	251		si	0	260

Taula 27. Pregunta 4. Matrius de confusió de la validació creuada

La taula amb els error calculats a cada pas:

Pas	Error
1	0,33507850
2	0,35863870
3	0,33246070
4	0,31937170
5	0,33246070
6	0,33507850
7	0,38743460
8	0,30628270
9	0,34293190
10	0,31937170
Mitjana del errors	0,336911

Taula 28. Pregunta 4. Errors generats en la validació creuada

Les matrius de confusió indican que l'error en el '*no*' és del 100% i en el '*si*' és del 0%. Tal como indicàvem, l'arbre no és capaç de predir correctament el '*no*'.

Aquest fet es deu al fet d'utilitzar un gran nombre de variables amb poca variabilitat i la descompensació entre els casos '*si*' i el casos '*no*'.

S'han verificat els resultats amb diferents algorismes de classificació como arbres C5.0, J48, perceptrons multicapa i d'altres, i en tots els casos s'ha obtingut un resultat similar sinó idèntic.

De fet l'interès de la qüestió és l'impacte de les assignatures en la decisió de matricular-se o no, l'arbre generat ens pot dir quines assignatures han estat mes rellevants per al càlcul i amb quina importància ho han fet:

```
CART$variable.importance
```

Assignatura	Importància
X05.554	14,28047719
X05.563	9,62463105
X05.561	6,28577005
X05.616	4,49462080
X05.562	3,25637756
X05.614	0,49274134
X05.559	0,36420012
X05.658	0,25708244
X05.557	0,08569415
X05.558	0,08569415

Taula 29. Pregunta 4. Importància de les variables de l'arbre

Aquestes son les assignatures que mes influeixen.

Aquesta informació és la que es reflecteix a l'arbre. Cal tenir en compte que l'arbre presenta l'inconvenient de la dificultat per a predir els casos '*no*', i de que hi un nombre de variables relativament elevat amb un pes molt similar entre variables.

Hi ha diverses tècniques que intenten millorar la fiabilitat de l'arbre utilitzant diversos algoritmes d'aprenentatge per obtenir un millor rendiment predictiu, per exemple, Random Forest i Gradient Boosting Machine, son ambdós, mètodes combinats d'aprenentatge.

En general destaquen 3 tècniques, com les mes utilitzades:

- Bagging
- Boosting
- Stacking

  - ***Bagging*** (Bootstrap Aggregating) , s'ha concebut per millorar l'estabilitat i la precisió dels algoritmes d'aprenentatge automàtic. Redueix la variància i ajuda a evitar el sobreajust. Bagging és especialment indicat davant de dades sorolloses i valors atípics.
    El mètode utilitza combinacions amb repeticions per produir conjunts múltiples de la mateixa cardinalitat que les seves dades originals
    Un exemple d'aquest mètode és Random Forest, que s'aplica als arbres de decisió.

  - ***Boosting*** és un mètode dissenyat per reduir el biaix i la variància.
    Les dades que estan mal classificades augmenten el seu pes i elles mostres que es classifiquen correctament disminueixen el seu pes. Per tant, el sistema es centre en els elements mal classificats. Això fa que no sigui molt robust davant dades sorolloses i valors atípics.
    Utilitza subconjunts de les dades originals per produir una sèrie de models mitjana i després potencia el seu rendiment mitjançant la combinació de tots ells utilitzant una determinada funció càlcul de cost.
    Bagging i Boosting pretenen crear un model fort a partir dels model febles.

  - ***Stacking*** és similar al '*Boosting'*, també s'apliquen diferents models a les dades originals. Però no tenim nomes una fórmula empírica per a la funció de pes, en el seu lloc s'introdueix un meta-nivell i s'utilitza un altre model per estimar les entrades, juntament amb les sortides de cada model per determinar quins models funcionen bé i quins malament donades les dades d'entrada.

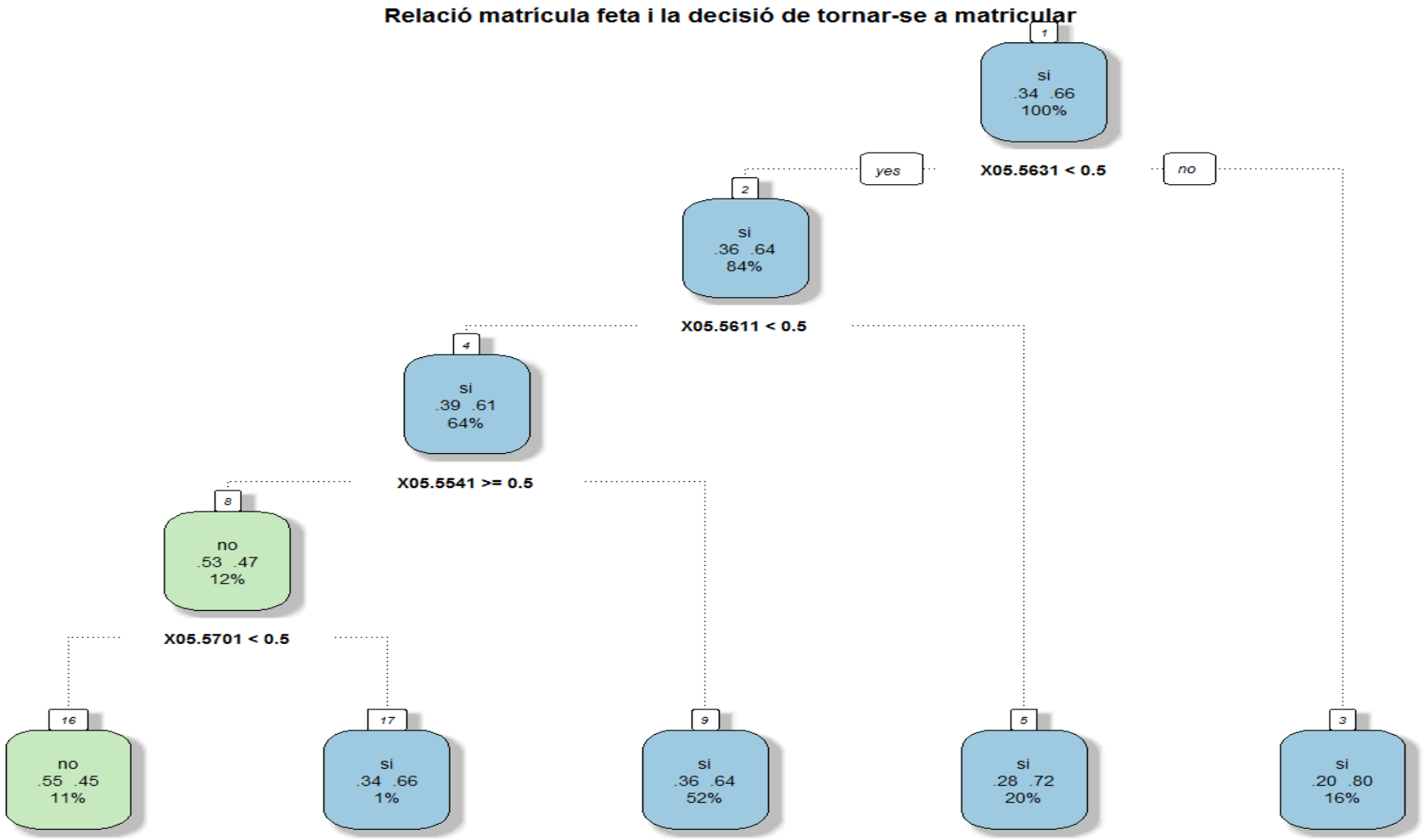
La llibreria 'caret' (Classification and Regression Training), es va crear com una interfície comuna per a treballar amb varis models diferents. Incorpora funcions de bagging, boosting etc. Indiquem amb quin model volem treballar i aplica de forma transparent tècniques combinades prenen com a base el model triat.

Farem un intent de generar el mateix arbre amb les llibreries 'rpart' però en aquest cas utilitzant-les a traves de 'caret'

Utilitzem les mateixes dades, en aquest cas també amb subgrup de dades i entrenament.

Annex 11.33. Pregunta 4. Generar l'arbre amb *'caret'*

L'arbre obtingut és molt semblant a l'anterior, fins i tot la assignatura que determina la primera divisió de l'arbre és la mateixa, X05.563.



Il·lustració 32. Pregunta 4. Arbre generat amb '*caret*'

Fem una predicció per a verificar la matriu de confusió i la precisió de l'arbre. Annex 11.34. Pregunta 4. Predicció amb *'caret'.*

	no	Si
no	53	63
si	333	697

Taula 30: pregunta 4. Matriu de confusió ('*caret*')

Veiem que en aquest cas l'error total comes és lleugerament superior, però el fet important és que la matriu de confusió en els casos del '*no'* ha millorat molt.

Ara tenim que l'error global és 34,6%, l'error en els '*no*' és del 54,3% i d'un 32,3% en els '*si*'.

```
>cart
```

```
2678 samples
  54 predictor
   2 classes: 'no', 'si'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 2678, 2678, 2678, 2678, 2678, 2678, ...
Resampling results across tuning parameters:

  cp           Accuracy   Kappa
  0.002583979  0.6650564  0.09753364
  0.006644518  0.6657135  0.07415808
  0.007751938  0.6641022  0.05828181

Accuracy was used to select the optimal model using
 the largest value.
The final value used for the model was cp = 0.006644518.
```

Taula 31. Pregunta 4. Propietats del arbre generat amb *'caret'*

Veiem que caret ha utilitzat tècniques de '*Bootstrapping*', 25 repositoris. Al final es selecciona l'arbre que ha obtingut els millors valors de precisió. Indicant-lo com '*model final'*

```
asRules(cart$finalModel)
```

```
Rule number: 3 [.outcome=si cover=431 (16%) prob=0.80]
   X05.5631>=0.5

 Rule number: 5 [.outcome=si cover=533 (20%) prob=0.72]
   X05.5631< 0.5
```

```
   X05.5611>=0.5

 Rule number: 17 [.outcome=si cover=32 (1%) prob=0.66]
   X05.5631< 0.5
   X05.5611< 0.5
   X05.5541>=0.5
   X05.5701>=0.5

 Rule number: 9 [.outcome=si cover=1397 (52%) prob=0.64]
   X05.5631< 0.5
   X05.5611< 0.5
   X05.5541< 0.5

 Rule number: 16 [.outcome=no cover=285 (11%) prob=0.45]
   X05.5631< 0.5
   X05.5611< 0.5
   X05.5541>=0.5
   X05.5701< 0.5
```

Taula 32. Pregunta 4. Regles de l'arbre generat amb *'caret'*

Que interpretem com que partim de 2678 mostres de les quals 891 no es rematriculen i 1787 si (33,2% i 66,7 %)

- Si rematriculen els alumnes que s'han matriculat de X05.563.
- Si rematriculen els alumnes que no es matriculen de X05.563 però si de X05.561.
- Si rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 però si de X05.554 i també de X05.570.
- Si rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 ni de 05.554.
- NO rematriculen els alumnes que no es matriculen de X05.563 ni de X05.561 ni de X05.570 i si de X05.554.

Les variables importants son les que es mostren a la taula. Segons la taula la variable que mes influeix és l'assignatura X05.563.

Assignatura	Importància
X05.563	18,816553
X05.554	16,255096
X05.561	9,333812
X05.570	2,552881

Taula 33. Pregunta 4. Variables importants de l'arbre '*caret*'

Podem concloure que amb aquest model que utilitza '*boostrap*' hem millorat la solució pel que fa al problema que presenta la predicció en els models basats en una sol mètode de validació i que presenten problemes deguts

dades sorolloses i valors atípics , amb respecte al algorismes que utilitzen mètodes combinats com '*bootstrap*' que genera n grups de dades de la mateixa grandària que el conjunt original seleccionant el que millor precisió obté.

La precisió del arbre és nomes del 65%. No és una precisió destacable. S'han fet proves amb varis model obtenint resultats si be son similars en precisió, les matrius de confusió mostren el problema de la incapacitat de predir de forma correcta els casos del '*no*'

6.6.5 Pregunta 5. Quines assignatures tenen més impacte en la decisió de tornar-se a matricular el segon semestre?

L'impacte que una assignatura pot tenir en la decisió de matricular-se prové del fet d'aprovar o suspendre la assignatura. Ja hem vist que la variable que indica el numero d'assignatures aprovades és la variable que mes impacte té en la decisió de continuar els estudis.

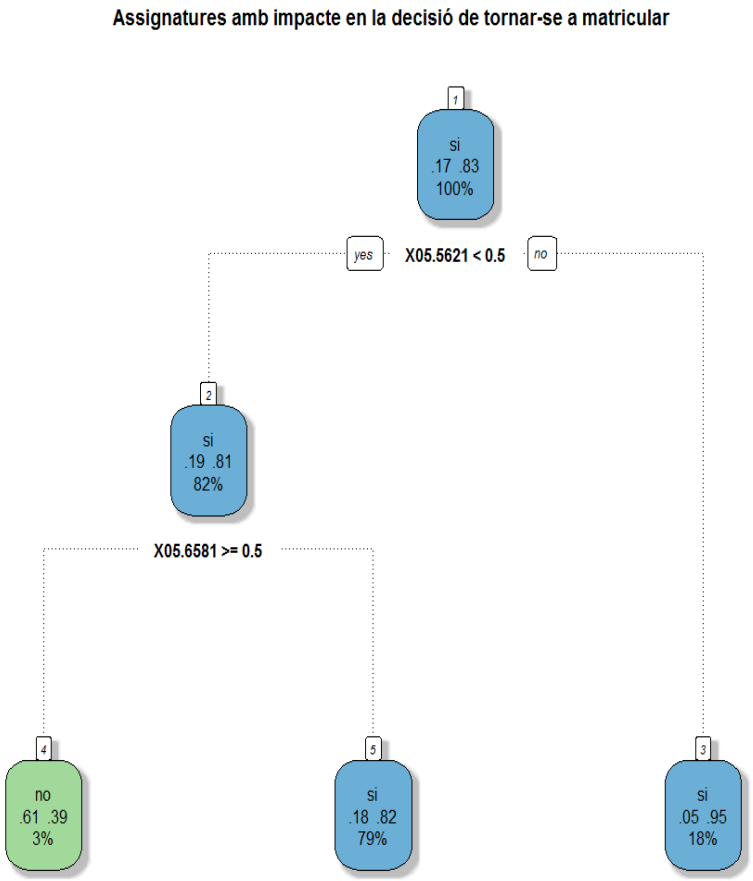
La qüestió en concret tracta de esbrinar si existeixen assignatures que impacten de forma positiva o negativa en l'èxit dels estudiants pel fet de esser aprovades o no.

Farem doncs una doble aproximació al problema, en un primer intent generem una matriu amb les assignatures i l'indicador de si han estat aprovades o no.

Un cop transformat el conjunt de dades generarem un arbre.

Annex 11.35. Pregunta 5. Transformació de les dades i generació de l'arbre. Assignatures aprovades.

Generem l'arbre amb la llibreria '*caret*' directament, doncs l'algorisme s'encarrega de trobar la opció mes optima del mateix després de fer 25 conjunts de dades de la mateixa grandària que el conjunt d'entrenament proveït inicialment



Il·lustració 33. Pregunta 5. Arbre generat – Assignatures aprovades

Un anàlisi previ de l'arbre ens diu que la assignatura que més impacte té en la decisió és X05.562, que curiosament és la assignatura amb mes índex de matriculacions com s'ha demostrat a la pregunta 1.

Aquesta assignatura divideix l'arbre en el seu primer nivell i ens indica que aprovar aquesta assignatura és en la majoria de casos una alta probabilitat d'èxit. Tot i que hi ha altres assignatures que poden variar aquest resultat.

Fem una predicció per a conèixer la precisió de l'arbre.

Annex 11.36. Pregunta 5. Predicció

	no	Si
no	11	122
si	8	651

Taula 34. Pregunta 5. Matriu de confusió de l'arbre d'assignatures suspeses.

Error = 0,164141

Continuen els problemes de predicció en el '*no*' tot i que hem utilitzat '*bagging*', com ja s'ha comentat és un problema provocat per les dades i la seva distribució.

Finalment veiem les variables importants i les regles obtingudes de l'arbre.

```
print(cart$finalModel$variable.importance)
asRules(cart$finalModel)
```

Assignatura	Importància
X05.658	19.49766
X05.562	10.52169

Taula 35. Pregunta 5. Variables importants de l'arbre d'assignatures suspeses.

```
Rule number: 3 [.outcome=si cover=330 (18%) prob=0.95]
   X05.5621>=0.5

 Rule number: 5 [.outcome=si cover=1468 (79%) prob=0.82]
   X05.5621< 0.5
   X05.6581< 0.5

 Rule number: 4 [.outcome=no cover=54 (3%) prob=0.39]
   X05.5621< 0.5
   X05.6581>=0.5
```

Taula 36. Pregunta 5. Regles arbre d'assignatures aprovades.

Si aprovem la assignatura X05.562 tenim èxit (prob. 95%)

Si no aprovem la X05.562 ni la assignatura X05.658 tenim èxit (prob. 82%)

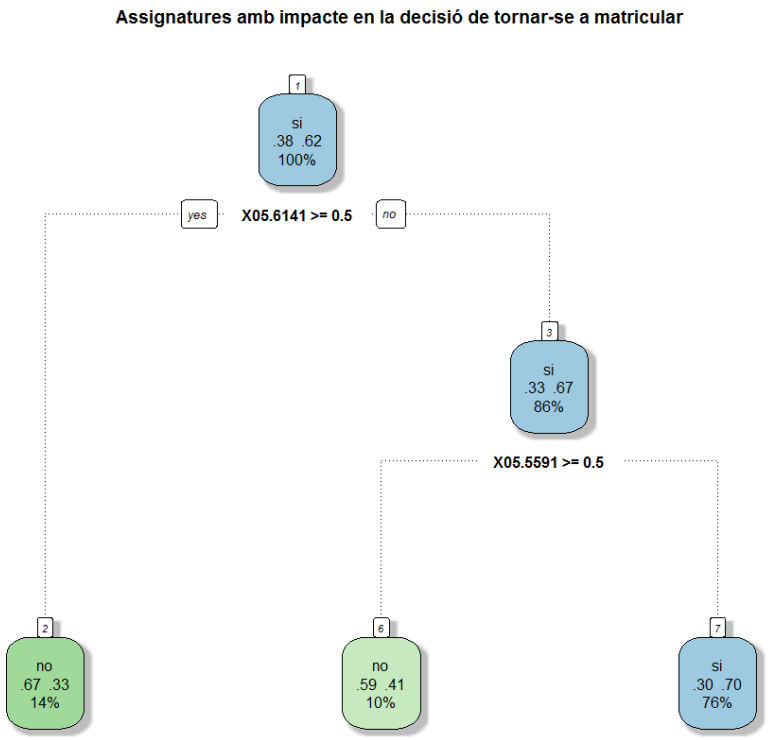
Si no aprovem la X05.562 i si aprovem la X05.658 tindrem fracàs. (prob. 39%)

Veiem que el fet d'aprovar assignatures pot tenir una influencia positiva en l'èxit, tot i que no aprovar també, com diu la regla 5. L'arbre generat pot canviar degut a que els conjunts de dades utilitzats es generen de forma aleatòria, per tant podem obtenir arbres amb mes o menys nivells. Però el significat essencial és molt similar.

Cal fer un aclariment, donat que la matriu de dades indica si la assignatura ha estat aprovada o no, el fet que hagi estat aprovada és evident, en canvi el fet que la assignatura no hagi estat aprovada no significa que hagi estat suspesa, també por indicar que no hagi estat matriculada o be no presentada i cal tenir-ho en compte en el moment d'interpretar l'arbre.

Com a segona fase fem el mateix però amb les assignatures suspeses.

Annex 11.37. Pregunta 5. Transformació de les dades i generació de l'arbre. Assignatures suspeses.



Il·lustració 34. Pregunta 5. Arbre generat – Assignatures suspeses.

Ara constatem que el fet de suspendre assignatures ens aboca al fracàs.

Les assignatures que mes influencien son la X05.614 i X05.559, que estan en el ranking de les assignatures mes suspeses.

Veiem matriu de confusió, variables importants i regles:

	no	si
no	35	74
si	28	150

Taula 37. Pregunta 5. Matriu de confusió d'assignatures suspeses.

Error = 0.3554007

Assignatura	Importància
X05.614	19.08652
X05.559	10.02469

Taula 38. Pregunta 5. Variables importants d'assignatures suspeses.

```
Rule number: 7 [.outcome=si cover=510 (76%) prob=0.70]
   X05.6141< 0.5
   X05.5591< 0.5

 Rule number: 6 [.outcome=no cover=66 (10%) prob=0.41]
   X05.6141< 0.5
   X05.5591>=0.5

 Rule number: 2 [.outcome=no cover=95 (14%) prob=0.33]
   X05.6141>=0.5
```

Taula 39. Pregunta 5. Regles arbre d'assignatures suspeses.

Si suspenem la assignatura X05.614 fracassem (prob. 33%).

Si suspenem la assignatura X05.559 fracassem (prob. 41%).

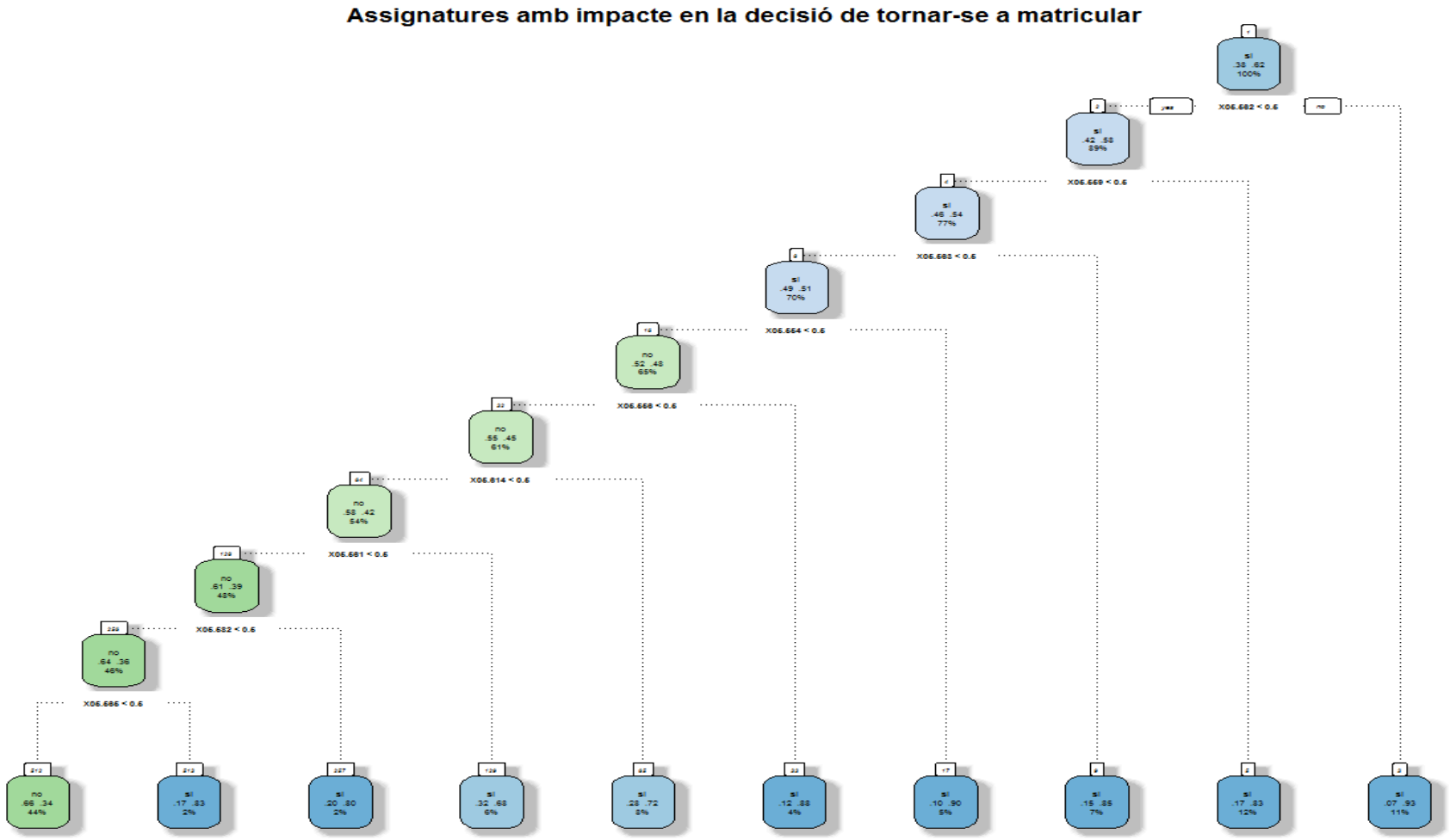
En qualsevol altre cas tindrem èxit. (prob. 70%).

En aquest cas també és vàlid l'aclariment del fet que una assignatura no ha estat suspesa no significa que hagi estat aprovada.

Aquest arbres que hem obtingut son molt simples i degut al problema que comentem de la distribució de les dades, és possible que els resultats no siguin molt acurats. El fet evident és que un arbre senzill quan la dispersió de les dades en tan gran no pot predir amb cura.

Donat que el problema ve provocat per la distribució de les classes '*si*', '*no*' de la variable dependent, podem provar de generar un conjunt d'entrenament que contingui menys casos '*si*' i mes mostres de la classe '*no*', sense repetir-ne, ni generar-les de forma automàtica, nomes fent una repartició diferent a partir del conjunt original, i generem un arbre amb els aprovats.

Annex 11.38. Pregunta 5. Generar arbre amb menys desequilibri.



Il·lustració 35. Pregunta 5. Arbre equilibrat. Assignatures aprovades

La informació mostrada per l'arbre és interesant. Tots el nodes

La matriu de confusió ens diu:

	no	si
no	425	155
si	355	1039

Taula 40. Pregunta 5. Matriu de confusió del arbre equilibrat

Error = 0,2583587

Les variables importants:

Assignatura	Importància
X05.562	43,6611544
X05.559	31,6425443
X05.563	28,8412080
X05.554	28,7243163
X05.556	25,6922103
X05.614	22,5690212
X05.582	17,0513019
X05.561	15,5308282
X05.565	13,8081197
X05.601	1,1120414
X05.586	0,3706805
X05.589	0,3706805

Les regles:

```
Rule number: 3 [.outcome=si cover=199 (11%) prob=0.93]
   X05.562>=0.5

 Rule number: 17 [.outcome=si cover=88 (5%) prob=0.90]
   X05.562< 0.5
   X05.559< 0.5
   X05.563< 0.5
   X05.554>=0.5

 Rule number: 33 [.outcome=si cover=75 (4%) prob=0.88]
   X05.562< 0.5
   X05.559< 0.5
   X05.563< 0.5
   X05.554< 0.5
   X05.556>=0.5

 Rule number: 9 [.outcome=si cover=138 (7%) prob=0.85]
   X05.562< 0.5
   X05.559< 0.5
   X05.563>=0.5
```

```
Rule number: 513 [.outcome=si cover=30 (2%) prob=0.83]
  X05.562< 0.5
  X05.559< 0.5
  X05.563< 0.5
  X05.554< 0.5
  X05.556< 0.5
  X05.614< 0.5
  X05.561< 0.5
  X05.582< 0.5
  X05.565>=0.5

Rule number: 5 [.outcome=si cover=221 (12%) prob=0.83]
  X05.562< 0.5
  X05.559>=0.5

Rule number: 257 [.outcome=si cover=46 (2%) prob=0.80]
  X05.562< 0.5
  X05.559< 0.5
  X05.563< 0.5
  X05.554< 0.5
  X05.556< 0.5
  X05.614< 0.5
  X05.561< 0.5
  X05.582>=0.5

Rule number: 65 [.outcome=si cover=139 (8%) prob=0.72]
  X05.562< 0.5
  X05.559< 0.5
  X05.563< 0.5
  X05.554< 0.5
  X05.556< 0.5
  X05.614>=0.5

Rule number: 129 [.outcome=si cover=102 (6%) prob=0.68]
  X05.562< 0.5
  X05.559< 0.5
  X05.563< 0.5
  X05.554< 0.5
  X05.556< 0.5
  X05.614< 0.5
  X05.561>=0.5

Rule number: 512 [.outcome=no cover=812 (44%) prob=0.34]
  X05.562< 0.5
  X05.559< 0.5
  X05.563< 0.5
  X05.554< 0.5
  X05.556< 0.5
  X05.614< 0.5
  X05.561< 0.5
  X05.582< 0.5
  X05.565< 0.5
```

Si interpretem les regles tenint en consideració l'aclariment fet en els punts anterior d'aquest apartat podem dir de forma simple:

Si aprovem la assignatura X05.614 tenim èxit (prob. 93%).

Si aprovem la assignatura X05.554 tenim èxit (prob. 90%).

Si aprovem la assignatura X05.556 tenim èxit (prob. 88%).

Si aprovem la assignatura X05.563 tenim èxit (prob. 85%).

Si aprovem la assignatura X05.565 tenim èxit (prob. 83%).

Si aprovem la assignatura X05.559 tenim èxit (prob. 83%).

Si aprovem la assignatura X05.582 tenim èxit (prob.80%).

Si aprovem la assignatura X05.614 tenim èxit (prob. 68%).

Si aprovem la assignatura X05.561 tenim èxit (prob. 34%).

En qualsevol altre cas fracàs.

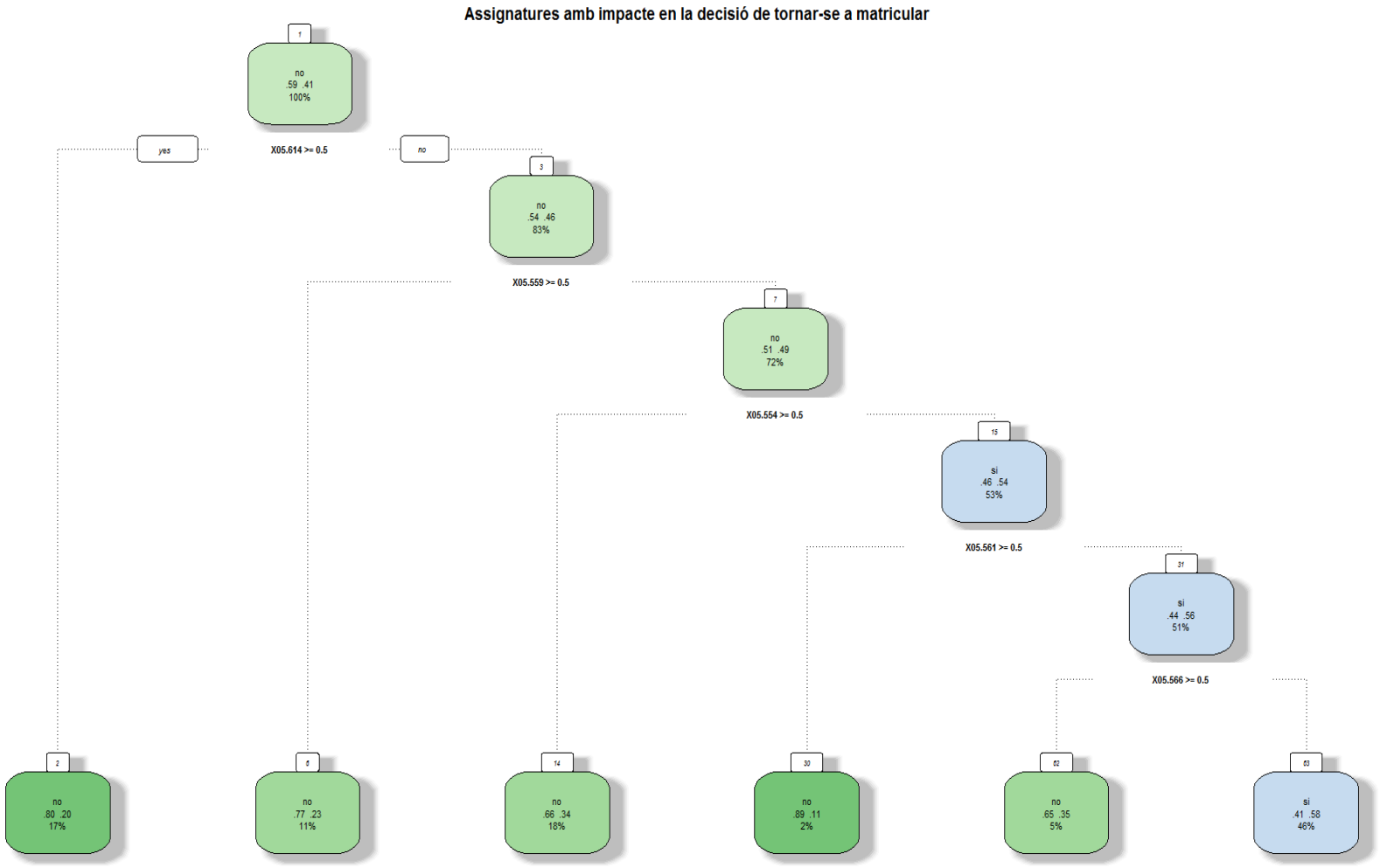
L'arbre és molt significatiu, totes les branques acaben en '*si*' i ho fan en funció d'una assignatura, la ultima branca és un '*no',* fet que es produeix quan no has aprovat cap assignatura.

La matriu de confusió és mes equilibrada i l'error produït és del 25,8%.

Farem el mateix amb les suspeses.

Generem un arbre augmentant la proporció dels '*no*' . Cal trobar l'equilibri doncs en el cas dels suspensos s'eliminen força files de l'arbre cosa que encara desequilibra mes la proporció entre els elements de la classe '*rematricula'.*

Annex 11.39. Pregunta 5. Generar arbre amb menys desequilibri.



Taula 41. Pregunta 5. Arbre d'assignatures suspeses.

La informació mostrada per l'arbre és interesant. Tots el nodes

La matriu de confusió ens diu:

	no	si
no	72	37
si	161	254

Taula 42. Pregunta 5. Matriu de confusió del arbre equilibrat

Error = 0,3778626

Les variables importants:

Assignatura	Importància
X05.614	8,184062
X05.559	5.743419
X05.554	4.777572
X05.561	3.494682
X05.566	2.320688

Les regles:

```
Rule number: 63 [.outcome=si cover=200 (46%) prob=0.58]
   X05.614< 0.5
   X05.559< 0.5
   X05.554< 0.5
   X05.561< 0.5
   X05.566< 0.5

 Rule number: 62 [.outcome=no cover=23 (5%) prob=0.35]
   X05.614< 0.5
   X05.559< 0.5
   X05.554< 0.5
   X05.561< 0.5
   X05.566>=0.5

 Rule number: 14 [.outcome=no cover=79 (18%) prob=0.34]
   X05.614< 0.5
   X05.559< 0.5
   X05.554>=0.5

 Rule number: 6 [.outcome=no cover=48 (11%) prob=0.23]
   X05.614< 0.5
   X05.559>=0.5

 Rule number: 2 [.outcome=no cover=75 (17%) prob=0.20]
   X05.614>=0.5

 Rule number: 30 [.outcome=no cover=9 (2%) prob=0.11]
   X05.614< 0.5
   X05.559< 0.5
```

```
    X05.554< 0.5
    X05.561>=0.5
```

Si interpretem les regles simplificant:

Si suspenem la assignatura X05.566 fracassem (prob. 35%).

Si suspenem la assignatura X05.554 fracassem (prob. 24%).

Si suspenem la assignatura X05.559 fracassem (prob. 23%).

Si suspenem la assignatura X05.614 fracassem (prob. 20%).

Si suspenem la assignatura X05.561 fracassem (prob. 11%).

En qualsevol altre cas èxit (prob. 58%).

En aquest arbre totes les branques acaben en '*no*' i ho fan en funció d'una assignatura, la última branca és un '*si',* fet que es produeix quan no has suspès cap assignatura.

La matriu de confusió també és mes equilibrada però en aquest cas l'error produït és del 37,8%.

Veiem dos arbres prou significatius amb significats oposats però un missatge clar, si aproves continues el estudis, si no aproves fracasses.

7 Conclusions

L'anàlisi realitzat amb el model associatiu del Market Basket Analisys ha demostrat que la gran dispersió de les dades ha impedit trobar relacions fortes entre les assignatures triades pels alumnes en la seva primera matriculació. Hi ha una gran varietat de combinacions de matricules.

Si que es veuen clarament les assignatures mes matriculades o les que es suspenen amb mes freqüència. No és tan clar que existeixin relacions fortes entre assignatures, tant a l'hora de matricular-se com als suspensos.

Les combinacions de suspensos encara tenen uns nivells de relacions mes febles donat que el numero de registres és inferior al de matriculacions.

Si que existeixen relacions en les influencies a les assignatures entre el fets de la matriculació i els fets d'aprovar o suspendre. Aquestes relacions s'han generat a partir d'algorismes escrits en R, generant una estructura de transaccions per a ser tractada amb l'algorisme '*apriori'*. Degut a la naturalesa de l'algorisme l'anàlisi '*apriori*' es generen regles que tenen sentit de forma individual com a relacions freqüents, i també es veu que hi ha fortes relacions especialment entre les matricules i aprovar certes assignatures, amb confiances del 100% de probabilitat.

En l'anàlisi amb el model de classificació de les assignatures triades per l'alumne en el moment de la matriculació veiem que els arbres de decisió no han funcionat correctament degut a varis factors, entre d'altres el numero de mostres no sembla ser suficient per a l'anàlisi, la dispersió de dades junt amb un numero de variables elevat amb un comportament molt homogeni provoca l'efecte de *overfitting* a l'arbre, (*sobreajust*) de forma que l'algorisme no funciona amb els registres de la classe *'no'* al ser minoritària davant el '*si*'. Les tècniques de models combinats com el '*bagging'* ens han permès finalment generar un model, amb un nivell d'error elevat, però mes equilibrat a l'hora d'avaluar les dues classes de la variable dependent '*rematricula'*. Fet que ens ha indicat que fracassen en els seus estudis els alumnes que no es matriculen de X05.563 ni de X05.561 ni de X05.570 i si de X05.554 a la seva primera matricula.

Finalment al analitzar les assignatures que mes impacte tenen en la decisió de continuar o no els estudis hem vist que hi ha una relació directa entre aprovar i èxit o be suspendre i fracàs. També hi ha una relació directa entre les assignatures que son les mes matriculades donat que tenen mes influencia positiva en l'èxit dels estudis i en el cas contrari, les assignatures mes suspeses o 'letals' son les que mes impacte negatiu tenen en la decisió de continuar els estudis.

8 Línies d'evolució a futur

Accions a tenir en consideració per a propers estudis, seria poder disposar d'un numero de mostres mes ampli que permeti establir relacions associatives mes fortes.

D'altra banda caldria reconsiderar el fet d'incloure dades sociodemogràfiques per a anàlisis futurs donat que la seva influencia en l'objectiu cercat és nul·la. Com ja s'ha vist, la variable que és decisiva és el numero d'assignatures aprovades.

No es pot descartar, si mes no, que algun factor sociodemogràfic pugui haver tingut una influencia directa en el fet d'aprovar assignatures, com ara la classe social, si treballa o no, temps disponible per als estudis etc. Però en aquest treball no s'han considerat altres que el sexe, edat i semestre de matriculació.

També seria interessant poder estudiar altres mètodes o algorismes d'anàlisi mes 'forts' davant de estructures de dades com les que s'han utilitzat en aquest estudi, per a aconseguir model predictius mes acurats.

9 Glossari

apriori. és un algoritme utilitzat en mineria de dades que permet trobar 'conjunts d'ítems freqüents', que serveixen de base per generar regles d'associació.

Bagging. Bootstrap Aggregation, és un meta algorisme d'aprenentatge automàtic dissenyat per millorar l'estabilitat i precisió d'algoritmes d'aprenentatge usats en classificació estadística i regressió. A més redueix la variància i ajuda a evitar el sobreajust.

BBDD. Una base de dades o banc de dades (en ocasions abreujada amb la sigles BD o amb l'abreviatura bd) és un conjunt de dades pertanyents a un mateix context i emmagatzemades sistemàticament per al seu posterior ús.

BIG DATA. És un concepte que fa referència a l'emmagatzematge de grans quantitats de dades i als procediments usats per trobar patrons repetitius dins d'aquestes dades.

Boosting. És una conjunt de meta-algoritme d'aprenentatge computacional per reduir el biaix principalment, i també la variància en l'aprenentatge supervisat, que converteixen els sistemes d'aprenentatge 'febles' en 'forts'

Caret. Paquet d'utilitats diverses per a l'entrenament i impressió de models de classificació i regressió en llenguatge R

CART. Classification and Regression tres. És un tipus d'algorisme d'aprenentatge computacional per al modelatge predictiu basat en arbres de decisió..

CRISP-DM. Cross Industry Standard Process for Data Mining. És un model Mineria de Dades que descriu l'enfoc que utilitzen els experts en mineria de dades. L'altre estàndard és el SEMMA però CRISP-DM és el mes utilitzat.

CSV. Els arxius CSV (comma separated values) son un tipus de document en format obert senzill per a representar dades en forma de taula, en les que les columnes es separen per comes (o punt i coma a on la coma és el separador decimal) y les files per salts de línia.

Diagrama Gantt. És una eina útil que te com a objectiu mostrar gràficament el temps de dedicació previst per diferents tasques o activitats al llarg d'un temps determinat total.

DM. Data Mining o Mineria de dades És Un camp de l'estadística i les ciències de la computació referit al procés que intenta descobrir patrons en grans volums de conjunts de dades. Utilitza mètodes d'intel·ligència artificial i aprenentatge computacional.

EDM. Descriu un camp d'investigació que s'ocupa de l'aplicació de la mineria de dades, aprenentatge computacional i l'estadística a la informació generada a partir dels entorns educatius (per exemple, les universitats).

KDD. Knowledge Discovery in Databases. Descobriment de coneixement en Bases de Dades, es refereix al Procés no trivial de descobrir coneixement i informació útil dins de les dades continguts en qualsevol repositori d'informació.

LA. Learning Analytics. (anàlisi de l'aprenentatge) és la mesura, recopilació, anàlisi i presentació de dades sobre els estudiants, els seus contextos i les interaccions que allí es generen, amb la finalitat de comprendre el procés d'aprenentatge que s'està desenvolupant i optimitzar els entorns en els que es produeix.

Matriu de confusió. és una eina que permet la visualització de l'acompliment d'un algoritme que s'empra en aprenentatge supervisat. Cada columna de la matriu representa el nombre de prediccions de cada classe, mentre que cada fila representa les instàncies en la classe real.

MBA. Market Basket Analisys. És un dels tipus més comuns i útils d'anàlisi de dades per a la distribució i venda al detall. El propòsit del MBA és determinar quins productes compren els clients conjuntament.

PMML. El Predictive Model Markup Language (PMML) és un llenguatge de marcat de text XML desenvolupat pel Data Mining Group (DMG) per proveir a les aplicacions una manera de definir models relacionats amb els anàlisis predictius i la mineria de dades per compartir aquests models entre les aplicacions PMML.

Poda. És l'acció de canviar el model mitjançant la supressió dels nodes fills d'un altre node de la branca. El node tallat es tracta com un node final. Els nodes finals no es poden tallar. L'objectiu és obtenir un arbre amb menys branques però amb una precisió similar.

R. R és un entorn i llenguatge de programació amb un enfocament a l'anàlisi estadística.

Random Forest. És una combinació d'arbres predictors tal que cada arbre depèn dels valors d'un vector aleatori provat independentment i amb la mateixa distribució per a cada un d'aquests. És una modificació substancial del '*bagging*' que construeix una llarga col·lecció d'arbres no correlacionats i després els fa la mitjana.

Rpart. Recursive partitioning for classification, regression and survival trees. És una llibreria per al llenguatge R que implementa aquesta mena d'arbres de classificació.

Sobreajust. Overfitting. És una situació en la què el model és com si aprengués les dades d'entrenament de memòria, en lloc d'aprendre les generalitats de les dades de prova, això succeeix quan el model és massa complex pel que fa a la mida de les dades d'entrenament, és a dir, quan la mida de les dades d'entrenament és petita en comparació amb la complexitat del model.

Stacking. És un algoritme d'aprenentatge per combinar les prediccions de diversos altres algoritmes d'aprenentatge. En primer lloc, tots els altres algoritmes estan capacitats utilitzant les dades disponibles, llavors un algoritme combinador és capaç de fer una predicció final utilitzant totes les prediccions dels altres algoritmes com entrades addicionals

Validació creuada. és una tècnica de validació de models per avaluar com els resultats d'una anàlisi estadística depenen d'un conjunt de dades. S'utilitza quan es vol estimar la precisió d'un model predictiu, i consisteix a dividir el conjunt en k parts provant el model amb k-1 parts com a dades d'entrenament i la resta com a test. Es prova k vegades de forma aleatòria i el resultat és la mitjana dels errors de les k proves.

10 Bibliografia i referències

Bali, Raghav, Sarkar,Dipanjan. *What you need to know about R*. Packt Publishing, 2016

Cichosz,Pawet. *Data Mining Algorithms Explained using R*. John Wiley & Sons, Ltd, 2015

Guisande González, Cástor, Vaamonde Liste, Antonio. *Graficos Estadisticos Y Mapas Con R*. Ediciones Díaz de Santos, 2013

Ledolter, Johannes. *Data Minig and Business Analytics with R*. John Wiley & Sons, Inc, 2013

Yu-Wei, Chiu. *Machine Learning with R*. Packt Publishing, 2015

R and Data Mining.
[*http://www.rdatamining.com/*]

The R Project for Statistical Computing
[*https://www.r-project.org/*]

The Comprehensive R Archive Network (CRAN)
[https://cran.r-project.org/]

11 Annexos

11.1 Script de conversió del arxiu inicial.

Carrega Inicial.R

```
#neteja la memòria
rm(list=ls())
#Carrega les dades
EDMLA <- read.csv("EDMLA.csv", sep=",")

# Elimina les columnes dummy
for (i in 1:57){
 j = 7 + i
 EDMLA[,j:(j+5)] = NULL
}

# canvia Any per edat
EDMLA["any"] = EDMLA["semestre"] %/% 10 - EDMLA["any"]
colnames(EDMLA)[ colnames(EDMLA)=="any"] = "edat"

# discretitza la edat
EDMLA$edat[as.numeric(EDMLA$edat) <= 25] = "Grup 1"
EDMLA$edat[as.numeric(EDMLA$edat) <= 40] = "Grup 2"
EDMLA$edat[as.numeric(EDMLA$edat) <= 55] = "Grup 3"
EDMLA$edat[as.numeric(EDMLA$edat) < 200] = "Grup 4"

# discretitza el semestre
EDMLA$semestre[as.numeric(EDMLA$semestre) %%2 == 1] = "estiu"
EDMLA$semestre[as.numeric(EDMLA$semestre) %%2 == 0] = "hivern"

# discretitza la matricula
EDMLA$rematricula[EDMLA$rematricula == 1] = "si"
EDMLA$rematricula[EDMLA$rematricula == 0] = "no"

#Elimina columnes sense valor
EDMLA$X05.581 = NULL
EDMLA$X05.610 = NULL
EDMLA$X05.615 = NULL

#convertim columnes en factors (variables qualitatives)
for (i in 1:3){
 EDMLA[,i] = as.factor(EDMLA[,i])
}

for (i in 7:ncol(EDMLA)){
 EDMLA[,i] = as.factor(EDMLA[,i])
}
```

11.2 Valor únics per cada variable

Variable	Valors
sexe	F,M
semestre	estiu ,hivern
edat |	grup 1, grup 2, grup 3, grup 4
X05.554 X05.555 X05.556 X05.557 X05.558 X05.559 X05.560 X05.561 X05.562 X05.563 X05.564 X05.565 X05.566 X05.567 X05.568 X05.569 X05.570 X05.571 X05.572 X05.573 X05.575 X05.578 X05.582 X05.585 X05.586 X05.587 X05.589 X05.590 X05.591 X05.593 X05.601 X05.604 X05.607 X05.611 X05.613 X05.614 X05.616	0, 1, 2 ,3
X05.574 X05.576 X05.577 X05.579 X05.580 X05.584 X05.592 X05.594 X05.595 X05.596	0, 1, 3

X05.597 X05.599	
X05.588 X05.598 X05.600 X05.658	0, 3
rematricula	no, si

Taula 43. Valors únics per cada variable

11.3 Freqüències dels valors per cada variable.

Variable	Freqüències
Sexe	F: 384 M:3440
Semestre	estiu : 2370 hivern: 1454
edat	Grup 1: 936 Grup 2:2337 Grup 3: 540 Grup 4: 11
matr	Min. :1.000 1st Qu.:2.000 Median :2.000 Mean :2.519 3rd Qu.:3.000 Max. :7.000
pres	Min. :0.000 1st Qu.:1.000 Median :2.000 Mean :1.672 3rd Qu.:2.000 Max. :7.000
aprv	Min. :0.000 1st Qu.:0.000 Median :1.000 Mean :1.368 3rd Qu.:2.000

	Max. :7.000
rematricula	N:1289 S:2535

Taula 44. 1.20 Freqüències dels valors per cada variable

11.4 Freqüències dels valors de les assignatures.

X05.554	X05.555	X05.556	X05.557	X05.558	X05.559	X05.560
0: 2906 1: 364 2: 191 3: 363	0: 3709 1: 58 2: 7 3: 50	0: 3433 1: 122 2: 16 3: 253	0: 2887 1: 411 2: 137 3: 389	0: 3684 1: 73 2: 19 3: 48	0: 2857 1: 289 2: 97 3: 581	0: 3771 1: 13 2: 4 3: 36
X05.561	X05.562	X05_563	X05_564	X05_565	X05_566	X05_567
0: 2988 1: 168 2: 35 3: 633	0: 2536 1: 678 2: 139 3: 471	0: 3183 1: 180 2: 43 3: 418	0: 3716 1: 38 2: 16 3: 54	0: 3503 1: 76 2: 72 3: 173	0: 3724 1: 24 2: 47 3: 29	0: 3759 1: 20 2: 16 3: 29
X05_568	X05_569	X05_570	X05_571	X05_572	X05_573	X05_574
0: 3670 1: 67 2: 11 3: 76	0: 3778 1: 27 2: 5 3: 14	0: 3460 1: 141 2: 65 3: 158	0: 3712 1: 35 2: 6 3: 71	0: 3775 1: 17 2: 6 3: 26	0: 3668 1: 81 2: 23 3: 52	0: 3814 1: 4 3: 6
X05_575	X05_576	X05_577	X05_578	X05_579	X05_580	X05_582
0: 3779 1: 19 2: 1 3: 25	0: 3816 1: 4 3: 4	0: 3816 1: 1 2: 1 3: 7	0: 3814 1: 2 3: 7	0: 3822 1: 1 3: 1	0: 3812 1: 5 3: 7	0: 3622 1: 41 2: 14 3: 147
X05_583	X05_584	X05_585	X05_586	X05_587	X05_588	X05_589
0: 3822 1: 2	0: 3812 1: 3 3: 9	0: 3758 1: 24 3: 25	0: 3796 1: 8 2: 17 3: 18	0: 3816 1: 3 2: 2 3: 4	0:3823 3: 1	0: 3756 1: 18 2: 9 3: 41
X05_590	X05_591	X05_592	X05_593	X05_594	X05_595	X05_596
0: 3716 1: 41 2: 1 3: 66	0: 3813 1: 3 2: 2 3: 6	0: 3815 1: 2 3: 7	0: 3803 1: 3 2: 2 3: 16	0: 3808 1: 8 3: 8	0: 3793 1: 7 3: 24	0: 3797 1: 6 3: 21
X05_597	X05_598	X05_599	X05_600	X05_601	X05_604	X05_607
0: 3806 1: 3 3: 15	0: 3819 3: 5	0: 3813 1: 3 3: 8	0: 3819 3: 5	0: 3809 1: 3 2: 4 3: 8	0: 3777 1: 13 2: 3 3: 31	0: 3817 1: 1 2: 2 3: 4
X05_611	X05_613	X05_614	X05_616	X05_658	rematricula	

0: 3697 1: 59 2: 15 3: 53	0: 3807 1: 4 2: 1 3: 12	0: 3046 1: 65 2: 130 3: 583	0: 3815 1: 1 3: 7	3: 128 0: 3696 2:	no:1289 si:2535	

Taula 45. Freqüències dels valors de les assignatures

11.5 Taula de valors faltants i únics per cada variable.

Atribut	Observacions
sexe	n missing unique 3824 0 2 F (384, 10%), M (3440, 90%)
edat	n missing unique 3824 0 4 Grup 1 (936, 24,5%), Grup 2 (2337, 61,1%), Grup 3 (540, 14,1%), Grup 4 (11, 0%)
matr	n missing unique Info Mean 3824 0 7 0.88 2.519 1 2 3 4 5 6 7 Frequency 452 1782 1002 392 142 45 9 % 12 47 26 10 4 1 0
pres	n missing unique Info Mean 3824 0 8 0.94 1.672 0 1 2 3 4 5 6 7 Frequency 767 915 1280 622 169 51 18 2 % 20 24 33 16 4 1 0 0
aprv	n missing unique Info Mean 3824 0 8 0.93 1.368 0 1 2 3 4 5 6 7 Frequency 1180 923 1078 474 125 33 10 1 % 31 24 28 12 3 1 0 0
X05_554	n missing unique 3824 0 4 2 (363, 9%), 0 (2906, 76%), 1 (364, 10%), 3 (191, 5%)
X05_555	n missing unique 3824 0 4 2 (50, 1%), 0 (3709, 97%), 1 (58, 2%), 3 (7, 0%)
X05_556	n missing unique 3824 0 4 2 (253, 7%), 0 (3433, 90%), 1 (122, 3%), 3 (16, 0%)
X05_557	n missing unique 3824 0 4 2 (389, 10%), 0 (2887, 75%), 1 (411, 11%), 3 (137, 4%)
X05_558	n missing unique 3824 0 4 2 (48, 1%), 0 (3684, 96%), 1 (73, 2%), 3 (19, 0%)
X05_559	n missing unique

	3824 0 4 2 (581, 15%), 0 (2857, 75%), 1 (289, 8%), 3 (97, 3%)
X05_560	n missing unique 3824 0 4 2 (36, 1%), 0 (3771, 99%), 1 (13, 0%), 3 (4, 0%)
X05_561	n missing unique 3824 0 4 2 (633, 17%), 0 (2988, 78%), 1 (168, 4%), 3 (35, 1%)
X05_562	n missing unique 3824 0 4 2 (471, 12%), 0 (2536, 66%), 1 (678, 18%), 3 (139, 4%)
X05_563	n missing unique 3824 0 4 2 (418, 11%), 0 (3183, 83%), 1 (180, 5%), 3 (43, 1%)
X05_564	n missing unique 3824 0 4 2 (54, 1%), 0 (3716, 97%), 1 (38, 1%), 3 (16, 0%)
X05_565	n missing unique 3824 0 4 2 (173, 5%), 0 (3503, 92%), 1 (76, 2%), 3 (72, 2%)
X05_566	n missing unique 3824 0 4 2 (29, 1%), 0 (3724, 97%), 1 (24, 1%), 3 (47, 1%)
X05_567	n missing unique 3824 0 4 2 (29, 1%), 0 (3759, 98%), 1 (20, 1%), 3 (16, 0%)
X05_568	n missing unique 3824 0 4 2 (76, 2%), 0 (3670, 96%), 1 (67, 2%), 3 (11, 0%)
X05_569	n missing unique 3824 0 4 2 (14, 0%), 0 (3778, 99%), 1 (27, 1%), 3 (5, 0%)
X05_570	n missing unique 3824 0 4 2 (158, 4%), 0 (3460, 90%), 1 (141, 4%), 3 (65, 2%)
X05_571	n missing unique 3824 0 4

	2 (71, 2%), 0 (3712, 97%), 1 (35, 1%), 3 (6, 0%)
X05_572	n missing unique 3824 0 4 2 (26, 1%), 0 (3775, 99%), 1 (17, 0%), 3 (6, 0%)
X05_573	n missing unique 3824 0 4 2 (52, 1%), 0 (3668, 96%), 1 (81, 2%), 3 (23, 1%)
X05_574	n missing unique 3824 0 3 2 (6, 0%), 0 (3814, 100%), 1 (4, 0%)
X05_575	n missing unique 3824 0 4 2 (25, 1%), 0 (3779, 99%), 1 (19, 0%), 3 (1, 0%)
X05_576	n missing unique 3824 0 3 2 (4, 0%), 0 (3816, 100%), 1 (4, 0%)
X05_577	n missing unique 3824 0 3 2 (7, 0%), 0 (3816, 100%), 1 (1, 0%)
X05_578	n missing unique 3824 0 4 2 (7, 0%), 0 (3814, 100%), 1 (2, 0%), 3 (1, 0%)
X05_579	n missing unique 3824 0 3 2 (1, 0%), 0 (3822, 100%), 1 (1, 0%)
X05_580	n missing unique 3824 0 3 2 (7, 0%), 0 (3812, 100%), 1 (5, 0%)
X05_582	n missing unique 3824 0 4 2 (147, 4%), 0 (3622, 95%), 1 (41, 1%), 3 (14, 0%)
X05_583	n missing unique 3824 0 2 0 (3822, 100%), 1 (2, 0%)
X05_584	n missing unique 3824 0 3 2 (9, 0%), 0 (3812, 100%), 1 (3, 0%)
X05_585	n missing unique 3824 0 4

	2 (25, 1%), 0 (3758, 98%), 1 (24, 1%), 3 (17, 0%)
X05_586	n missing unique 3824 0 4 2 (18, 0%), 0 (3796, 99%), 1 (8, 0%), 3 (2, 0%)
X05_587	n missing unique 3824 0 4 2 (4, 0%), 0 (3816, 100%), 1 (3, 0%), 3 (1, 0%)
X05_588	n missing unique 3824 0 2 2 (1, 0%), 0 (3823, 100%)
X05_589	n missing unique 3824 0 4 2 (41, 1%), 0 (3756, 98%), 1 (18, 0%), 3 (9, 0%)
X05_590	n missing unique 3824 0 4 2 (66, 2%), 0 (3716, 97%), 1 (41, 1%), 3 (1, 0%)
X05_591	n missing unique 3824 0 4 2 (6, 0%), 0 (3813, 100%), 1 (3, 0%), 3 (2, 0%)
X05_592	n missing unique 3824 0 3 2 (7, 0%), 0 (3815, 100%), 1 (2, 0%)
X05_593	n missing unique 3824 0 4 2 (16, 0%), 0 (3803, 99%), 1 (3, 0%), 3 (2, 0%)
X05_594	n missing unique 3824 0 3 2 (8, 0%), 0 (3808, 100%), 1 (8, 0%)
X05_595	n missing unique 3824 0 3 2 (24, 1%), 0 (3793, 99%), 1 (7, 0%)
X05_596	n missing unique 3824 0 3 2 (21, 1%), 0 (3797, 99%), 1 (6, 0%)
X05_597	n missing unique 3824 0 3 2 (15, 0%), 0 (3806, 100%), 1 (3, 0%)
X05_598	n missing unique

	3824 0 2 2 (5, 0%), 0 (3819, 100%)
X05_599	n missing unique 3824 0 3 2 (8, 0%), 0 (3813, 100%), 1 (3, 0%)
X05_600	n missing unique 3824 0 2 2 (5, 0%), 0 (3819, 100%)
X05_601	n missing unique 3824 0 4 2 (8, 0%), 0 (3809, 100%), 1 (3, 0%), 3 (4, 0%)
X05_604	n missing unique 3824 0 4 2 (31, 1%), 0 (3777, 99%), 1 (13, 0%), 3 (3, 0%)
X05_607	n missing unique 3824 0 4 2 (4, 0%), 0 (3817, 100%), 1 (1, 0%), 3 (2, 0%)
X05_611	n missing unique 3824 0 4 2 (53, 1%), 0 (3697, 97%), 1 (59, 2%), 3 (15, 0%)
X05_613	n missing unique 3824 0 4 2 (12, 0%), 0 (3807, 100%), 1 (4, 0%), 3 (1, 0%) ----------
X05_614	n missing unique 3824 0 4 2 (583, 15%), 0 (3046, 80%), 1 (65, 2%), 3 (130, 3%)
X05_616	n missing unique 3824 0 4 2 (7, 0%), 0 (3815, 100%), 1 (1, 0%), 3 (1, 0%)
X05_658	n missing unique 3824 0 2 2 (128, 3%), 0 (3696, 97%)
rematricula	n missing unique 3824 0 2 no (1289, 34%), si (2535, 66%)

Taula 46. Taula de valors faltants i únics per cada variable

11.6 Variables Importants.

```
# Carrega de les dades i les llibreries
```

```
source('Carrega Inicial.R')
library (mlbench)
library (caret)

datos = EDMLA

control <- trainControl (method ="repeatedcv", number =30 ,repeats =10)
model <- train(rematricula~., data =R, method ="lvq", preProcess ="scale", trControl = control)

# variables mes importants
 rimp <- varImp (model , scale= FALSE)
 print(rimp )
 plot ( rimp )
```

11.7 Variables importants (assignatures)

```
# Carrega de les llibreries
source('Carrega Inicial.R')
library (mlbench)
library (caret)

datos = EDMLA[,7:61]
control <- trainControl (method ="repeatedcv", number =30 ,repeats =10)
model <- train( rematricula~., data =datos, method ="lvq", preProcess ="scale", trControl =
control )

# variables més importants
 rimp <- varImp (model , scale= FALSE)
 print(rimp )
 plot ( rimp )
```

11.8 Variables importants (variables sociodemogràfiques)

```
# Carrega de les llibreries
source('Carrega Inicial.R')
library (mlbench)
library (caret)

datos = EDMLA[,-(7:60)]
control <- trainControl (method ="repeatedcv", number =30 ,repeats =10)
model <- train( rematricula~., data =datos, method ="lvq", preProcess ="scale", trControl =
control )

# variables mas importantes
 rimp <- varImp (model , scale= FALSE)
 print(rimp )
 plot ( rimp )
```

11.9 Dispersió de les dades. Assignatures matriculades

```
# Combinació d'assignatures matriculades (Dispersió de les dades)
source('Carrega Inicial.R')
library(plyr)
#Esborrem columnes innecessàries
datos = (EDMLA[, 7:60])
#Cerquem les matricules (1,2,o 3)
for (i in 1:ncol(datos)){
  datos[,i] = as.character(datos[,i])
  datos[datos[,i] == 1 | datos[,i] == 2 | datos[,i] == 3, i] = "1"
  #datos[datos[,i] == 3 , i] = "1"
}
#Fem el càlcul i ordenem per la columna 'freq'
res1 = count(datos)
res1 = res1[ order(-res1[,55]), ]
plot(res1$freq)
title(main = list("Distribució de combinacions (count) d'assignatures matriculades", cex = 1.5,
                  col = "red", font = 3))
```

11.10 Dispersió de les dades. Assignatures suspeses

```
# Combinació d'assignatures suspeses
source('Carrega Inicial.R')
library(plyr)
#Esborrem columnes innecessàries
datos = (EDMLA[, 7:60])
#Cerquem les suspeses (2)
for (i in 1:ncol(datos)){
  datos[,i] = as.character(datos[,i])
  datos[datos[,i] == 0 | datos[,i] == 1 | datos[,i] == 3, i] = "0"
  datos[datos[,i] == 2 , i] = "1"
}
#Fem el càlcul i ordenem per la columna 'freq'
res1 = count(datos)
res1 = res1[ order(-res1[,55]), ]
#Esborrem la primera fila (cap suspens)
res1 = res1[-1,]
plot(res1$freq)
title(main = list("Distribució de combinacions (count) d'assignatures suspeses", cex = 1.5,
                  col = "red", font = 3))
```

11.11 Pregunta 1. Transformació de dades per a aplicar l'algorisme '*apriori*'.

```
# Pregunta 1
#Càrrega de les dades
source('Carrega Inicial.R')
#Carrega de les llibreries
library(plyr)
```

```
library(arules)
library(arulesViz)


#Nomes deixa les assignatures
EDMLA = (EDMLA[, 7:60])
# Canvia valors 0,1,2,3 per 0,1
for (i in 1:ncol(EDMLA)){
  EDMLA[,i] = as.character(EDMLA[,i])
  EDMLA[EDMLA[,i] > 0 ,i ] = 1
  EDMLA[,i] = as.numeric(EDMLA[,i])
}

datos = EDMLA

trx = as(as.matrix(datos),"transactions")
```

11.12 Pregunta 1. Taula de assignatures mes matriculades ordenades per freqüència.

```
sort(itemFrequency(trx), decreasing=TRUE)
```

X05.562	X05.559	X05.557	X05.554	X05.561	X05.614	X05.563	X05.556	X05.570
0,33682	0,25287	0,24503	0,24006	0,21861	0,20345	0,16762	0,10224	0,09518
X05.565	X05.582	X05.573	X05.568	X05.558	X05.658	X05.611	X05.555	X05.571
0,08394	0,05282	0,04079	0,04027	0,03661	0,03347	0,03321	0,03007	0,02928
X05.564	X05.590	X05.566	X05.589	X05.585	X05.567	X05.560	X05.572	X05.604
0,02824	0,02824	0,02615	0,01778	0,01725	0,01699	0,01386	0,01281	0,01229
X05.569	X05.575	X05.595	X05.586	X05.596	X05.593	X05.597	X05.613	X05.594
0,01202	0,01176	0,00810	0,00732	0,00706	0,00549	0,00470	0,00444	0,00418
X05.601	X05.580	X05.584	X05.591	X05.599	X05.574	X05.578	X05.592	X05.616
0,00392	0,00313	0,00313	0,00287	0,00287	0,00261	0,00261	0,00235	0,00235
X05.576	X05.577	X05.587	X05.607	X05.598	X05.600	X05.579	X05.583	X05.588
0,00209	0,00209	0,00209	0,00183	0,00130	0,00130	0,000523	0,000523	0,000262

Taula 47. Pregunta 1. Assignatures mes matriculades ordenades per freqüència.

11.13 Pregunta 1. Obtenció de regles d'associació '*apriori*'

```
reglas = apriori(trx, parameter = list(support = .01, confidence =.01,minlen=2))
reglas <-sort(reglas, by="supp", decreasing=TRUE) # ordena regles
summary(reglas)
```

```
set of 159 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
88 51 20
```

```
  Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
 2.000   2.000   2.000   2.572   3.000   4.000

summary of quality measures:
 support            confidence          lift
 Min.   :0.01020   Min.   :0.04162   Min.   :0.3444
 1st Qu.:0.01308   1st Qu.:0.14030   1st Qu.:0.9048
 Median :0.02118   Median :0.27005   Median :1.2298
 Mean   :0.03053   Mean   :0.30372   Mean   :1.3912
 3rd Qu.:0.03452   3rd Qu.:0.39699   3rd Qu.:1.6160
 Max.   :0.12082   Max.   :1.00000   Max.   :3.5946

mining info:
 data ntransactions support confidence
  trx         3824            0.01       0.01
```

11.14 Pregunta 1. Poda de les regles.

```
# troba les regles redundant
subset <- is.subset(reglas, reglas)
subset[lower.tri(subset, diag=T)] <- NA
redundants <- colSums(subset, na.rm=T) >= 1
which(redundants)
# esborra les regles redundant
poda <- reglas[!redundants]
inspect(poda)
```

11.15 Pregunta 1. K-means

```
#Pregunta 5
#Carrega dades i llibreries
source('Carrega Inicial.R')
n_cluster = 8
#Nomes deixa les assignatures matriculades
EDMLA = EDMLA[,-(1:6)]
EDMLA = EDMLA[,-(55)]
for (i in 1:(ncol(EDMLA))){
  EDMLA[,i] = as.character(EDMLA[,i])
  EDMLA[EDMLA[,i] =="1" |EDMLA[,i] =="2" |EDMLA[,i] =="3",i ] = 1
  EDMLA[,i] = as.numeric(EDMLA[,i])
}

quecluster = function (cluster, pos) {
  max = 0
  j = 0
  for (i in 1:n_cluster) {
    if (cluster[i,pos]>max) {
      max = cluster[i,pos]
      j = i
```

```
   }
  }
  print(j)
  return(j)
}

#executa k-menas
km = kmeans(EDMLA, centers = n_cluster)

noms = names(EDMLA)
cluster = c(length(noms))


for (i in 1:length(noms)) {
  cluster[i]= quecluster(km$centers, i)
}
#crea vector assignatura - cluster
noms = data.frame(noms,cluster)

#Crea un data frame cluster - num. matricules - assignatures
matr = c(n_cluster)
for (i in 1:n_cluster) {
  matr[i]= sum(km$cluster==i)
}

ass= c(n_cluster)
for (i in 1:n_cluster) {
  ass[i] = paste(noms[noms$cluster==i,1],collapse=",")
}

mkm =data.frame(c(1:n_cluster),matr,ass )
colnames(mkm) = c("Cluster","matricules","assignatures")
```

11.16 Pregunta 2. Conversió de dades.

```
# Pregunta 2
#Carrega de les dades
source('Carrega Inicial.R')
#Carrega de les llibreries
library(plyr)
library(arules)
library(arulesViz)


#Nomes deixa les assignatures
EDMLA = (EDMLA[, 7:60])
# Canvia valors 2 per 1
#     i  valors 0,1,3 per 0
for (i in 1:ncol(EDMLA)){
  EDMLA[,i] = as.character(EDMLA[,i])
```

```
 EDMLA[EDMLA[,i] == "0"| EDMLA[,i] == "1"|EDMLA[,i] == "3"  ,i ] = 0
 EDMLA[EDMLA[,i] == "2" ,i ] = 1
 EDMLA[,i] = as.numeric(EDMLA[,i])
}
#Redueix columnes amb tot '0'
c = 1
eliminar = c(ncol(EDMLA)*ncol(EDMLA)-ncol(EDMLA))
for (i in 1:ncol(EDMLA)) {
 if (sum(EDMLA[,i])==0) {
  eliminar[c]=i
  c = c + 1
 }
}
EDMLA[,eliminar] = NULL

#redueix files amb tot '0'
c = 1
eliminar = c(nrow(EDMLA))
for (i in 1:nrow(EDMLA)) {
 if (sum(EDMLA[i,])==0) {
  eliminar [c]=i
  c = c + 1
 }
}
EDMLA = EDMLA[-eliminar,]
datos = EDMLA
trx = as(as.matrix(datos),"transactions")
summary(trx)
```

```
transactions as itemMatrix in sparse format with
 958 rows (elements/itemsets/transactions) and
 37 columns (items) and a density of 0.03275405

most frequent items:
X05.554 X05.562 X05.557 X05.614 X05.559 (Other)
      191       139       137       130        97     467

element (itemset/transaction) length distribution:
sizes
  1   2   3   4
781 152  24   1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.212   1.000   4.000

   Min.    1st Qu.  Median    Mean   3rd Qu.    Max.
 0.0000  0.0000   0.0000    0.3036  1.0000    4.0000
```

11.17 Pregunta 2. Taula de assignatures mes suspeses ordenades per freqüència.

```
itemFrequency (dades)
```

X05,554	X05,562	X05,557	X05,614	X05,559	X05,565	X05,570	X05,566	X05,563
0,19937	0,14509	0,14300	0,13569	0,10125	0,07515	0,06784	0,04906	0,04489
X05,561	X05,573	X05,558	X05,585	X05,556	X05,564	X05,567	X05,611	X05,582
0,03653	0,02400	0,01983	0,01774	0,01670	0,01670	0,01670	0,01565	0,01461
X05,568	X05,589	X05,555	X05,571	X05,572	X05,569	X05,560	X05,601	X05,604
0,01148	0,00939	0,00730	0,00626	0,00626	0,00521	0,00417	0,00417	0,00313
X05,586	X05,591	X05,593	X05,607	X05,575	X05,578	X05,587	X05,590	X05,613
0,00209	0,00208	0,00208	0,00208	0,00104	0,00104	0,00104	0,00104	0,00104
X05,616								
0,00104								

Taula 48. Pregunta 2. Assignatures mes suspeses ordenades per freqüència

11.18 Pregunta 2. Obtenció de regles d'associació '*apriori*'.

```
reglas = apriori (dades, parameter = list (support = .01, confidence = .01,
minlen = 2))
reglas <-sort(reglas, by="supp", decreasing=TRUE) # ordena reglas
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
maxlen target   ext
          0.01      0.1       1 none FALSE            TRUE       5    0.01
2         10  rules FALSE

Algorithmic control:
 Filtre  tree  heap memopt load sort verbose
    0.1 TRUE      FALSE TRUE    2      TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[37 item(s), 958 transaction(s)] done [0.00s].
sorting and recoding items ... [19 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [8 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

```
summary(reglas)
```

```
set of 8 rules

rule length distribution (lhs + rhs):sizes
2
8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support         confidence          lift
 Min.   :0.01253   Min.   :0.06806   Min.   :0.4759
 1st Qu.:0.01331   1st Qu.:0.09275   1st Qu.:0.6602
 Median :0.01566   Median :0.11351   Median :0.7885
 Mean   :0.01618   Mean   :0.11611   Mean   :0.8313
 3rd Qu.:0.01853   3rd Qu.:0.12904   3rd Qu.:0.9595
 Max.   :0.02088   Max.   :0.18462   Max.   :1.2724

mining info:
 data ntransactions support confidence
  trx           958    0.01       0.01
```

11.19 Pregunta 2. K-Means.

```
#Pregunta 2
#Carrega dades i llibreries
source('Carrega Inicial.R')
n_cluster = 8
#Nomes deixa les assignatures suspeses
EDMLA = (EDMLA[, 7:60])
# Canvia valors 2 per 1
#       valors 0,1,3 per 0
for (i in 1:ncol(EDMLA)){
 EDMLA[,i] = as.character(EDMLA[,i])
 EDMLA[EDMLA[,i] == "0"| EDMLA[,i] == "1"|EDMLA[,i] == "3"  ,i ] = 0
 EDMLA[EDMLA[,i] == "2" ,i ] = 1
 EDMLA[,i] = as.numeric(EDMLA[,i])
}

#Redueix columnes amb tot '0'
c = 1
eliminar = c(ncol(EDMLA)*ncol(EDMLA)-ncol(EDMLA))
for (i in 1:ncol(EDMLA)) {
 if (sum(EDMLA[,i])==0) {
  eliminar[c]=i
  c = c + 1
 }
}
EDMLA[,eliminar] = NULL

#redueix files amb tot '0'
c = 1
eliminar = c(nrow(EDMLA))
```

```
for (i in 1:nrow(EDMLA)) {
 if (sum(EDMLA[i,])==0) {
  eliminar [c]=i
  c = c + 1
 }
}
EDMLA = EDMLA[-eliminar,]


quecluster = function (cluster, pos) {
 max = 0
 j = 0
 for (i in 1:n_cluster) {
  if (cluster[i,pos]>max) {
   max = cluster[i,pos]
   j = i
  }
 }
 print(j)
 return(j)
}

#executa k-menas
km = kmeans(EDMLA, centers = n_cluster)

noms = names(EDMLA)
cluster = c(length(noms))


for (i in 1:length(noms)) {
 cluster[i]= quecluster(km$centers, i)
}
#crea vector assignatura - cluster
noms = data.frame(noms,cluster)

#Crea un data frame cluster - num. matricules - assignatures
matr = c(n_cluster)
for (i in 1:n_cluster) {
 matr[i]= sum(km$cluster==i)
}

ass= c(n_cluster)
for (i in 1:n_cluster) {
 ass[i] = paste(noms[noms$cluster==i,1],collapse=",")
}

mkm =data.frame(c(1:n_cluster),matr,ass )
colnames(mkm) = c("Cluster","matricules","assignatures")
```

11.20 Pregunta 3. Transformació de les dades anàlisi Matricular-se -> Aprovar.

```
# Pregunta 3 - Matricular-se -> Aprovar
source('Carrega Inicial.R')
library(plyr)
library(arules)
library(arulesViz)

#elimina les columnes que no son assignatures
EDMLA[,1:6] = NULL
EDMLA[,ncol(EDMLA)]=NULL

#Genera un nou data.frame per matriculat / versus aprovats
EDMLA5 = data.frame()

muestra = c(1:(ncol(EDMLA)*ncol(EDMLA)-ncol(EDMLA)))

for (r in 1:nrow(EDMLA)) {
 c = 1
 for (i in 1:ncol(EDMLA)) {
  for (j in 1:ncol(EDMLA)) {
   if (i!=j) {
    if ((EDMLA[r,i] == 1 | EDMLA[r,i] == 2 | EDMLA[r,i] == 3) & EDMLA[r,j] == 3 )
    { val = 1 }
    else
    {val = 0}
    muestra[c] = val
    c = c + 1
   }
  }
 }
 EDMLA5= rbind (EDMLA5 , muestra )
}

#genera la capçalera
k = 1
for (i in 1:ncol(EDMLA)) {
 for (j in 1:ncol(EDMLA)) {
  if (i != j) {
   names(EDMLA5)[k] = paste(names(EDMLA)[i] ,names(EDMLA)[j], sep ="/")
   k = k+1
  }
 }
}

#Redueix columnes amb tot '0'
c = 1
muestra = c(ncol(EDMLA)*ncol(EDMLA)-ncol(EDMLA))
for (i in 1:ncol(EDMLA5)) {
 if (sum(EDMLA5[,i])==0) {
```

```
   muestra[c]=i
   c = c + 1
  }
}
EDMLA5[,muestra] = NULL

#redueix files amb tot '0'
c = 1
muestra = c(nrow(EDMLA))
for (i in 1:nrow(EDMLA5)) {
  if (sum(EDMLA5[i,])==0) {
   muestra[c]=i
   c = c + 1
  }
}
EDMLA5 = EDMLA5[-muestra,]
#força columnes numèriques
datos = EDMLA5
datos = as.numeric(datos)
```

11.21 Pregunta 3. Genera transaccions 'apriori' anàlisi Matricular-se-> Aprovar.

```
#calcula les transaccions a partir de les dades
trx = as(as.matrix(datos),"transactions")

inspect(trx)
summary(trx)
itemFrequency(trx)
itemFrequencyPlot(trx,topN=30,type="relative")
title(main = list("Matriculació -> Superar assignatures", cex = 1.5, col = "red", font = 3))
```

```
transactions as itemMatrix in sparse format with
 2386 rows (elements/itemsets/transactions) and
 1029 columns (items) and a density of 0.004079515


most frequent items:
X05.554/X05.561 X05.562/X05.561 X05.562/X05.559 X05.561/X05.554 X05.562/X05.614         (Other)
            328             252             209             202             192            8833


element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    8    9   10   12   15   16   18   20   24   25   30   36   42
405  953   63  223    6  433   26   72    3  123    8   20    1   30    2    6   10    1    1
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   4.198   6.000  42.000
```

11.22 Pregunta 3. Generació regles '*apriori*' anàlisi Matricular-se-> Aprovar.

```
reglas = apriori(trx, parameter = list(support = .02, confidence =.5 ,minlen=2))
reglas <-sort(reglas, by="supp", decreasing=TRUE) # ordena reglas
summary(reglas)
```

```
set of 305 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6
 61 124  84  30   6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.331   4.000   6.000

summary of quality measures:
      support         confidence          lift
   Min.   :0.02012   Min.   :0.5000   Min.   : 3.637
   1st Qu.:0.02431   1st Qu.:0.8243   1st Qu.: 9.024
 Median :0.02557   Median :0.9531   Median :12.897
   Mean   :0.02720   Mean   :0.8756   Mean   :13.253
   3rd Qu.:0.02682   3rd Qu.:1.0000   3rd Qu.:17.043
    Max.   :0.07963   Max.   :1.0000   Max.   :34.430

mining info:
 data ntransactions support confidence
  trx          2386    0.02        0.5
```

11.23 Pregunta 3. 100 regles amb mes confiança' anàlisi Matricular-se-> Aprovar.

lhs	rhs	support	Conf.	lift
{X05.557/X05.658}	{X05.562/X05.658}	0,0210	1	18,6406
{X05.658/X05.562}	{X05.562/X05.658}	0,0231	1	18,6406
{X05.554/X05.614,X05.614/X05.561}	{X05.561/X05.614}	0,0201	1	23,1650
{X05.554/X05.561,X05.554/X05.614}	{X05.561/X05.614}	0,0201	1	23,1650
{X05.554/X05.561,X05.561/X05.614}	{X05.554/X05.614}	0,0201	1	24,3469
{X05.554/X05.614,X05.614/X05.561}	{X05.554/X05.561}	0,0201	1	7,2744
{X05.554/X05.561,X05.554/X05.614}	{X05.614/X05.561}	0,0201	1	22,0926
{X05.554/X05.561,X05.561/X05.614}	{X05.614/X05.561}	0,0201	1	22,0926

{X05.554/X05.557,X05.557/X05.561}	{X05.561/X05.557}	0,0243	1	23,8600
{X05.554/X05.561,X05.561/X05.557}	{X05.554/X05.557}	0,0243	1	20,3932
{X05.554/X05.557,X05.554/X05.561}	{X05.561/X05.557}	0,0243	1	23,8600
{X05.557/X05.561,X05.562/X05.557}	{X05.561/X05.557}	0,0205	1	23,8600
{X05.561/X05.557,X05.562/X05.561}	{X05.557/X05.561}	0,0205	1	17,4161
{X05.554/X05.561,X05.561/X05.557}	{X05.557/X05.561}	0,0243	1	17,4161
{X05.561/X05.557,X05.562/X05.561}	{X05.562/X05.557}	0,0205	1	12,8973
{X05.562/X05.557,X05.562/X05.561}	{X05.561/X05.557}	0,0205	1	23,8600
{X05.554/X05.559,X05.559/X05.561}	{X05.561/X05.559}	0,0264	1	17,8060
{X05.554/X05.559,X05.554/X05.561}	{X05.561/X05.559}	0,0264	1	17,8060
{X05.554/X05.561,X05.561/X05.559}	{X05.554/X05.559}	0,0264	1	21,4955
{X05.554/X05.559,X05.559/X05.561}	{X05.554/X05.561}	0,0264	1	7,2744
{X05.554/X05.559,X05.554/X05.561}	{X05.559/X05.561}	0,0264	1	15,1013
{X05.557/X05.559,X05.559/X05.562}	{X05.557/X05.562}	0,0201	1	15,1013
{X05.557/X05.559,X05.557/X05.562}	{X05.559/X05.562}	0,0201	1	19,5574
{X05.557/X05.559,X05.559/X05.562}	{X05.562/X05.559}	0,0201	1	11,4163
{X05.557/X05.562,X05.562/X05.559}	{X05.559/X05.562}	0,0201	1	19,5574
{X05.554/X05.557,X05.557/X05.561}	{X05.554/X05.561}	0,0243	1	7,2744
{X05.554/X05.557,X05.554/X05.561}	{X05.557/X05.561}	0,0243	1	17,4161
{X05.559/X05.561,X05.562/X05.559}	{X05.561/X05.559}	0,0210	1	17,8060
{X05.561/X05.559,X05.562/X05.561}	{X05.559/X05.561}	0,0210	1	15,1013
{X05.554/X05.561,X05.561/X05.559}	{X05.559/X05.561}	0,0264	1	15,1013
{X05.561/X05.559,X05.562/X05.561}	{X05.562/X05.559}	0,0210	1	11,4163
{X05.562/X05.559,X05.562/X05.561}	{X05.561/X05.559}	0,0210	1	17,8060
{X05.557/X05.561,X05.562/X05.557}	{X05.562/X05.561}	0,0205	1	9,4683
{X05.562/X05.557,X05.562/X05.561}	{X05.557/X05.561}	0,0205	1	17,4161
{X05.561/X05.562,X05.562/X05.554}	{X05.554/X05.562}	0,0268	1	17,0429
{X05.561/X05.554,X05.561/X05.562}	{X05.554/X05.562}	0,0268	1	17,0429
{X05.554/X05.562,X05.561/X05.554}	{X05.561/X05.562}	0,0268	1	19,0880
{X05.554/X05.562,X05.562/X05.561}	{X05.561/X05.562}	0,0310	1	19,0880
{X05.554/X05.561,X05.561/X05.562}	{X05.554/X05.562}	0,0310	1	17,0429
{X05.554/X05.561,X05.554/X05.562}	{X05.561/X05.562}	0,0310	1	19,0880
{X05.561/X05.562,X05.562/X05.554}	{X05.561/X05.554}	0,0268	1	11,8119
{X05.561/X05.554,X05.561/X05.562}	{X05.562/X05.554}	0,0268	1	15,9067
{X05.554/X05.561,X05.561/X05.562}	{X05.562/X05.561}	0,0310	1	9,4683
{X05.557/X05.559,X05.557/X05.562}	{X05.562/X05.559}	0,0201	1	11,4163
{X05.557/X05.562,X05.562/X05.559}	{X05.557/X05.559}	0,0201	1	12,8973
{X05.559/X05.561,X05.562/X05.559}	{X05.562/X05.561}	0,0210	1	9,4683
{X05.562/X05.559,X05.562/X05.561}	{X05.559/X05.561}	0,0210	1	15,1013
{X05.554/X05.562,X05.561/X05.554}	{X05.562/X05.554}	0,0268	1	15,9067
{X05.554/X05.562,X05.562/X05.561}	{X05.554/X05.561}	0,0310	1	7,2744
{X05.554/X05.561,X05.554/X05.562}	{X05.562/X05.561}	0,0310	1	9,4683
{X05.562/X05.554,X05.562/X05.561}	{X05.561/X05.554}	0,0310	1	11,8119
{X05.561/X05.554,X05.562/X05.561}	{X05.562/X05.554}	0,0310	1	15,9067

{X05.554/X05.561,X05.562/X05.554}	{X05.561/X05.554}	0,0310	1	11,8119
{X05.562/X05.554,X05.562/X05.561}	{X05.554/X05.561}	0,0310	1	7,2744
{X05.554/X05.561,X05.562/X05.554}	{X05.562/X05.561}	0,0310	1	9,4683
{X05.561/X05.554,X05.562/X05.561}	{X05.554/X05.561}	0,0310	1	7,2744
{X05.554/X05.614,X05.561/X05.614,X05.614/X05.561}	{X05.554/X05.561}	0,0201	1	7,2744
{X05.554/X05.561,X05.554/X05.614,X05.561/X05.614}	{X05.614/X05.561}	0,0201	1	22,0926
{X05.554/X05.561,X05.554/X05.614,X05.614/X05.561}	{X05.561/X05.614}	0,0201	1	23,1650
{X05.554/X05.561,X05.561/X05.614,X05.614/X05.561}	{X05.554/X05.614}	0,0201	1	24,3469
{X05.554/X05.557,X05.557/X05.561,X05.561/X05.557}	{X05.554/X05.561}	0,0243	1	7,2744
{X05.554/X05.557,X05.554/X05.561,X05.561/X05.557}	{X05.557/X05.561}	0,0243	1	17,4161
{X05.554/X05.561,X05.557/X05.561,X05.561/X05.557}	{X05.554/X05.557}	0,0243	1	20,3932
{X05.554/X05.557,X05.554/X05.561,X05.557/X05.561}	{X05.561/X05.557}	0,0243	1	23,8600
{X05.557/X05.561,X05.561/X05.557,X05.562/X05.557}	{X05.562/X05.561}	0,0205	1	9,4683
{X05.557/X05.561,X05.561/X05.557,X05.562/X05.561}	{X05.562/X05.557}	0,0205	1	12,8973
{X05.561/X05.557,X05.562/X05.557,X05.562/X05.561}	{X05.557/X05.561}	0,0205	1	17,4161
{X05.557/X05.561,X05.562/X05.557,X05.562/X05.561}	{X05.561/X05.557}	0,0205	1	23,8600
{X05.554/X05.559,X05.559/X05.561,X05.561/X05.559}	{X05.554/X05.561}	0,0264	1	7,2744
{X05.554/X05.559,X05.554/X05.561,X05.561/X05.559}	{X05.559/X05.561}	0,0264	1	15,1013
{X05.554/X05.559,X05.554/X05.561,X05.559/X05.561}	{X05.561/X05.559}	0,0264	1	17,8060
{X05.554/X05.561,X05.559/X05.561,X05.561/X05.559}	{X05.554/X05.559}	0,0264	1	21,4955
{X05.557/X05.559,X05.557/X05.562,X05.559/X05.562}	{X05.562/X05.559}	0,0201	1	11,4163
{X05.557/X05.559,X05.559/X05.562,X05.562/X05.559}	{X05.557/X05.562}	0,0201	1	15,1013
{X05.557/X05.562,X05.559/X05.562,X05.562/X05.559}	{X05.557/X05.559}	0,0201	1	12,8973
{X05.557/X05.559,X05.557/X05.562,X05.562/X05.559}	{X05.559/X05.562}	0,0201	1	19,5574
{X05.559/X05.561,X05.561/X05.559,X05.562/X05.559}	{X05.562/X05.561}	0,0210	1	9,4683
{X05.559/X05.561,X05.561/X05.559,X05.562/X05.561}	{X05.562/X05.559}	0,0210	1	11,4163
{X05.561/X05.559,X05.562/X05.559,X05.562/X05.561}	{X05.559/X05.561}	0,0210	1	15,1013
{X05.559/X05.561,X05.562/X05.559,X05.562/X05.561}	{X05.561/X05.559}	0,0210	1	17,8060
{X05.554/X05.562,X05.561/X05.562,X05.562/X05.554}	{X05.561/X05.554}	0,0268	1	11,8119
{X05.554/X05.562,X05.561/X05.554,X05.561/X05.562}	{X05.562/X05.554}	0,0268	1	15,9067
{X05.561/X05.554,X05.561/X05.562,X05.562/X05.554}	{X05.554/X05.562}	0,0268	1	17,0429
{X05.554/X05.562,X05.561/X05.554,X05.562/X05.554}	{X05.561/X05.562}	0,0268	1	19,0880
{X05.561/X05.562,X05.562/X05.554,X05.562/X05.561}	{X05.554/X05.562}	0,0256	1	17,0429
{X05.554/X05.562,X05.562/X05.554,X05.562/X05.561}	{X05.561/X05.562}	0,0256	1	19,0880
{X05.554/X05.561,X05.561/X05.562,X05.562/X05.554}	{X05.554/X05.562}	0,0256	1	17,0429
{X05.554/X05.561,X05.554/X05.562,X05.562/X05.554}	{X05.561/X05.562}	0,0256	1	19,0880
{X05.561/X05.554,X05.561/X05.562,X05.562/X05.561}	{X05.554/X05.562}	0,0256	1	17,0429
{X05.554/X05.562,X05.561/X05.554,X05.562/X05.561}	{X05.561/X05.562}	0,0256	1	19,0880
{X05.554/X05.561,X05.561/X05.554,X05.561/X05.562}	{X05.554/X05.562}	0,0256	1	17,0429
{X05.554/X05.561,X05.554/X05.562,X05.561/X05.554}	{X05.561/X05.562}	0,0256	1	19,0880
{X05.554/X05.562,X05.561/X05.562,X05.562/X05.561}	{X05.554/X05.561}	0,0310	1	7,2744
{X05.554/X05.561,X05.554/X05.562,X05.561/X05.562}	{X05.562/X05.561}	0,0310	1	9,4683
{X05.554/X05.561,X05.561/X05.562,X05.562/X05.561}	{X05.554/X05.562}	0,0310	1	17,0429
{X05.554/X05.561,X05.554/X05.562,X05.562/X05.561}	{X05.561/X05.562}	0,0310	1	19,0880

{X05.561/X05.562,X05.562/X05.554,X05.562/X05.561}	{X05.561/X05.554}	0,0256	1	11,8119
{X05.561/X05.554,X05.561/X05.562,X05.562/X05.561}	{X05.562/X05.554}	0,0256	1	15,9067
{X05.554/X05.561,X05.561/X05.562,X05.562/X05.554}	{X05.561/X05.554}	0,0256	1	11,8119
{X05.554/X05.561,X05.561/X05.554,X05.561/X05.562}	{X05.562/X05.554}	0,0256	1	15,9067

Taula 49. Pregunta 3. 100 regles del anàlisi Matricular-se -> Aprovar ordenades per 'confiança'.

11.24 Pregunta 3. Poda regles de l'anàlisi Matricular-se-> Aprovar.

```
# Trobar les regles redundants
subset <- is.subset(reglas, reglas)
subset[lower.tri(subset, diag=T)] <- NA
redundants <- colSums(subset, na.rm=T) >= 1
which(redundants)

# esborrar regles redundants
poda <- reglas[!redundants]
poda <-sort(poda, by="conf", decreasing=TRUE) # ordena regles
summary(poda)
```

```
set of 46 rules

rule length distribution (lhs + rhs):sizes
 2  3
41  5

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.109   2.000   3.000

summary of quality measures:
    support          confidence          lift
 Min.   :0.02012   Min.   :0.5063   Min.   : 3.683
 1st Qu.:0.02232   1st Qu.:0.5707   1st Qu.: 6.375
 Median :0.02892   Median :0.8344   Median :10.202
 Mean   :0.03275   Mean   :0.7608   Mean   :11.946
 3rd Qu.:0.03803   3rd Qu.:0.9382   3rd Qu.:15.923
 Max.   :0.07963   Max.   :1.0000   Max.   :34.430

mining info:
 data ntransactions support confidence
  trx          2386    0.02        0.5
```

11.25 Pregunta 3. Transformació de les dades anàlisi Matricular-se -> Suspendre.

```
# Pregunta 3 - Matricular-se -> Suspendre
source('Carrega Inicial.R')
library(plyr)
```

```
library(arules)
library(arulesViz)

#elimina les columnes que no son assignatures
EDMLA[,1:6] = NULL
EDMLA[,ncol(EDMLA)]=NULL

#Genera un nou data.frame per matriculat / versus suspens
EDMLA5 = data.frame()

muestra = c(1:(ncol(EDMLA)*ncol(EDMLA)-ncol(EDMLA)))

for (r in 1:nrow(EDMLA)) {
 c = 1
 for (i in 1:ncol(EDMLA)) {
  for (j in 1:ncol(EDMLA)) {
   if (i!=j) {
    if ((EDMLA[r,i] == 1 | EDMLA[r,i] == 2 | EDMLA[r,i] == 3) & EDMLA[r,j] == 2 )
    { val = 1 }
    else
    {val = 0}
    muestra[c] = val
    c = c + 1
   }
  }
 }
 EDMLA5= rbind (EDMLA5 , muestra )
}

#genera la capçalera
k = 1
for (i in 1:ncol(EDMLA)) {
 for (j in 1:ncol(EDMLA)) {
  if (i != j) {
   names(EDMLA5)[k] = paste(names(EDMLA)[i] ,names(EDMLA)[j], sep ="/")
   k = k+1
  }
 }
}

#Redueix columnes amb tot '0'
c = 1
muestra = c(ncol(EDMLA)*ncol(EDMLA)-ncol(EDMLA))
for (i in 1:ncol(EDMLA5)) {
 if (sum(EDMLA5[,i])==0) {
  muestra[c]=i
  c = c + 1
 }
}
EDMLA5[,muestra] = NULL
```

```
#redueix files amb tot '0'
c = 1
muestra = c(nrow(EDMLA))
for (i in 1:nrow(EDMLA5)) {
 if (sum(EDMLA5[i,])==0) {
  muestra[c]=i
  c = c + 1
 }
}
EDMLA5 = EDMLA5[-muestra,]

write.csv(EDMLA5,file="EDMLA5b.csv")


#força columnes numèriques
datos = EDMLA5
datos = as.numeric(datos)
```

11.26 Pregunta 3. Genera transaccions 'apriori' anàlisi Matricular-se-> Suspendre.

```
#calcula les transaccions a partir de les dades

trx = as(as.matrix(datos),"transactions")


inspect(trx)

summary(trx)

itemFrequency(trx)

itemFrequencyPlot(trx,topN=30,type="relative")

title(main = list("Matriculació -> Suspendre assignatures", cex = 1.5, col = "red", font = 3))
```

```
transactions as itemMatrix in sparse format with
 898 rows (elements/itemsets/transactions) and
 471 columns (items) and a density of 0.00539297

most frequent items:
X05.561/X05.554 X05.562/X05.554 X05.559/X05.557 X05.559/X05.562 X05.562/X05.614          (Other)
             96              94              65              61              61            1904

element (itemset/transaction) length distribution:
sizes
  1   2   3   4   5   6   8   9  10  12  15
345 268  88  94  12  49  20   6   6   6   4

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    2.00    2.54    3.00   15.00
```

11.27 Pregunta 3. Generació regles '*apriori*' anàlisi Matricular-se-> Suspendre

```
reglas = apriori(trx, parameter = list(support = .02, confidence =.5 ,minlen=2))
reglas <-sort(reglas, by="supp", decreasing=TRUE) # ordena reglas
summary(reglas)
```

```
set of 18 rules

rule length distribution (lhs + rhs):sizes
 2  3
13  5

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.278   2.750   3.000

summary of quality measures:
    support          confidence          lift
 Min.   :0.02004   Min.   :0.5102   Min.   : 4.876
 1st Qu.:0.02227   1st Qu.:0.5405   1st Qu.: 5.652
 Median :0.02784   Median :0.6030   Median : 6.538
 Mean   :0.03081   Mean   :0.6076   Mean   : 7.866
 3rd Qu.:0.03480   3rd Qu.:0.6577   3rd Qu.: 8.188
 Max.   :0.05457   Max.   :0.7917   Max.   :14.709

mining info:
 data ntransactions support confidence
  trx           898    0.02        0.5
```

11.28 Pregunta 3. Poda regles de l'anàlisi Matricular-se-> Suspendre.

```
# Trobar les regles redundants
subset <- is.subset(reglas, reglas)
subset[lower.tri(subset, diag=T)] <- NA
redundants <- colSums(subset, na.rm=T) >= 1
which(redundants)

# esborrar regles redundants
poda <- reglas[!redundants]
poda <-sort(poda, by="supp", decreasing=TRUE) # ordena regles
summary(poda)
```

```
set of 11 rules

rule length distribution (lhs + rhs):sizes
 2
11

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support          confidence          lift
```

```
 Min.   :0.02116   Min.   :0.5104   Min.   : 4.876
 1st Qu.:0.02394   1st Qu.:0.5415   1st Qu.: 5.449
 Median :0.03229   Median :0.6000   Median : 6.600
 Mean   :0.03219   Mean   :0.6033   Mean   : 7.989
 3rd Qu.:0.03619   3rd Qu.:0.6371   3rd Qu.: 9.906
 Max.   :0.05457   Max.   :0.7917   Max.   :14.709

mining info:
 data ntransactions support confidence
  trx           898    0.02        0.5
```

11.29 Pregunta 4. Transformació de les dades i generació de l'arbre.

```
#Pregunta 4
#Carrega dades i llibreries
source('Carrega Inicial.R')
library(rpart)
library(rpart.plot)

#Eliminem les variables sociodemogràfiques
EDMLA = EDMLA[,-(1:6)]
#Canviem 1,2,3 -> 1
for (i in 1:(ncol(EDMLA)-1)){
 EDMLA[,i] = as.character(EDMLA[,i])
 EDMLA[EDMLA[,i] =="1" |EDMLA[,i] =="2"|EDMLA[,i] =="3",i ] = 1
 EDMLA[,i] = as.factor(EDMLA[,i])
}

#Separem conjunt de test i entrenament a partir de les dades
inTrain <- createDataPartition(y=EDMLA$rematricula, p=0.7, list=FALSE)
train <- EDMLA[inTrain,]
test <- EDMLA[-inTrain,]

#genera l'arbre
CART<- rpart(rematricula ~ . ,  data= train,
        control=rpart.control(minsplit=5, minbucket = 5, cp = 0.002 , surrogatestyle =
1),method="class")

rpart.plot(CART, type=0, extra=2, varlen=0, ycompress = TRUE, main="Relació matrícula feta i
la decisió de tornar-se a matricular", cex.main=2, cex=1.2)

printcp(CART)
print(CART)
summary(CART)
```

11.30 Pregunta 4. Predicció

Farem una predicció per determinar la qualitat de l'arbre:

```
#predicció
pred = predict(CART, newdata = test, type = "class")
```

```
mc = table(test$rematricula,pred)
err <- 1.0 - (mc[1,1]+mc[2,2])/sum(mc)
err
mc
```

```
> err
[1] 0.3376963
> mc
     pred
       no  si
  no   31 355
  si   32 728
```

11.31 Pregunta 4. Poda de l'arbre

```
#poda
CART = prune.rpart(CART, cp=0,0033223)
rpart.plot(CART, type=0, extra=2, varlen=0, ycompress = TRUE, main=" Relació matrícula feta i la decisió de tornar-se a matricular ", cex.main=2, cex=1.2)
```

11.32 Pregunta 4. Validació Creuada

```
#cross validation
n <- nrow(EDMLA) #nombre d'observacions
K <- 10 # nombre de folds
size <- n%/%K # calculem la grandària de cada bloc
set.seed(5) #per a obtenir sempre la mateixa seqüencia
alea <- runif(n) #genera números aleatoris (n)
rang <- rank(alea) #associa un rang a cada individu
bloc <- (rang-1)%/%size + 1 # associa un numero de bloc a cada individu
bloc <- as.factor(bloc) #transformar en factor
print(summary(bloc)) #imprimeix el summary del bloc

#executa la validació creuada
errors <- numeric(0)
for (k in 1:K){
 #genera un arbre amb tots els blocs menys k
 arbre <- rpart(rematricula ~., data = EDMLA[bloc!=k,], method = "class")
 #fa una predicció amb el bloc = k
 pred <- predict(arbre,newdata=EDMLA[bloc==k,], type = "class")
 #genera la matriu de confusió
 mc <- table(EDMLA$rematricula[bloc==k],pred)
 #calcula l'error
 err <- 1.0 - (mc[1,1]+mc[2,2])/sum(mc)
 #afegeix el registre
 errors <- rbind(errors,err)
 print(mc)
}
```

```
#mostra els error generats a cada validació
print(errors)

#calcula la mitjana dels errors
print(mean(errors))
```

11.33 Pregunta 4. Generar l'arbre amb '*caret*'

```
# Pregunta 5 - caret
# Càrrega de les dades i les llibreries
source('Carrega Inicial.R')
library(rattle)
library(rpart.plot)
library(rpart)
library(caret)
library(knitr)

#Eliminem les variables sociodemogràfiques
EDMLA = EDMLA[,-(1:6)]
#Canviem 1,2,3 -> 1
for (i in 1:(ncol(EDMLA)-1)){
  EDMLA[,i] = as.character(EDMLA[,i])
  EDMLA[EDMLA[,i] =="1" |EDMLA[,i] =="2"|EDMLA[,i] =="3",i ] = 1
  EDMLA[,i] = as.factor(EDMLA[,i])
}

# Reordenar al atzar les dades
EDMLA <- EDMLA [sample(nrow(EDMLA)),]

train <- EDMLA[1:2678,]
test <- EDMLA[2679:3824,]

## generem l'arbre
cart <- train(rematricula ~ .,data=EDMLA, method="rpart")

fancyRpartPlot(cart$finalModel,sub="", main=" Relació matrícula feta i la decisió de tornar-se
a matricular " )
```

11.34 Pregunta 4. Predicció amb '*caret*'.

```
# Fem la predicció
pred <- predict(cart, test)
mc <- confusionMatrix(pred, test$rematricula)
kable(mc$table)
err <- 1.0 - (mc$table[1,1]+mc$table[2,2])/sum(mc$table)
err
```

```
|    |  no|  si|
|:---|---:|---:|
```

```
|no |  53|  63|
|si | 333| 697|

 [1] 0.3455497
```

11.35 Pregunta 5. Transformació de les dades i generació de l'arbre. Assignatures aprovades

```
#Pregunta 5
#Carrega dades i llibreries
source('Carrega Inicial.R')
library(rpart)
library(rpart.plot)
library(caret)
library(rattle)

#Eliminem les variables sociodemogràfiques
EDMLA = EDMLA[,-(1:6)]
#Canviem  3 -> 1"
#       0,1,2 -> 0
# Assignatures aprovades
for (i in 1:(ncol(EDMLA)-1)){
  EDMLA[,i] = as.character(EDMLA[,i])
  EDMLA[EDMLA[,i] =="0" |EDMLA[,i] =="1"|EDMLA[,i] =="2" ,i ] = 0
  EDMLA[EDMLA[,i] =="3",i ] = 1
  EDMLA[,i] = as.numeric(EDMLA[,i])
}

datos  = EDMLA[,1:(ncol(EDMLA)-1)]
for (i in 1:ncol(datos)){
  datos[,i] = as.numeric(datos[,i],na.rm=TRUE)
}
#Redueix columnes amb tot '0'
c = 1
borrar = c(ncol(datos))
for (i in 1:ncol(datos)) {
  if (sum(datos[,i])==0) {
    borrar[c]=i
    c = c + 1
  }
}
EDMLA[,borrar] = NULL

#redueix files amb tot '0'
c = 1
borrar = c(nrow(datos))
for (i in 1:nrow(datos)) {
  if (sum(datos[i,],na.rm = TRUE)==0) {
    borrar[c]=i
    c = c + 1
  }
```

```
}
EDMLA = EDMLA[-borrar,]
# Transformem tot a factor
for (i in 1:(ncol(EDMLA)-1)){
  EDMLA[,i] = as.factor(EDMLA[,i])
}

# Reordenar al atzar les dades
EDMLA <- EDMLA [sample(nrow(EDMLA)),]
#Separem conjunt de test i entrenament a partir de les dades
inTrain <- createDataPartition(y=EDMLA$rematricula, p=0.7, list=FALSE)
train <- EDMLA[inTrain,]
test <- EDMLA[-inTrain,]

# generem l'arbre
cart <- train(rematricula ~ .,data=train, method="rpart", na.action = na.exclude)
fancyRpartPlot(cart$finalModel,sub="", main="Assignatures amb impacte en la decisió de
tornar-se a matricular" )
```

11.36 Pregunta 5. Predicció

```
pred = predict(cart, newdata = test)
mc = table(test$rematricula,pred)
err <- 1.0 - (mc[1,1]+mc[2,2])/sum(mc)
err
mc
```

11.37 Pregunta 5. Transformació de les dades i generació de l'arbre. Assignatures suspeses

```
#Pregunta 5 – menys desequilibri
#Carrega dades i llibreries
source('Carrega Inicial.R')
library(rpart)
library(rpart.plot)
library(caret)
library(rattle)

#Eliminem les variables sociodemogràfiques
EDMLA = EDMLA[,-(1:6)]
#Canviem  2 -> 1"
#        0,1,3 -> 0
# Assignatures suspeses
for (i in 1:(ncol(EDMLA)-1)){
  EDMLA[,i] = as.character(EDMLA[,i])
  EDMLA[EDMLA[,i] =="0" |EDMLA[,i] =="1"|EDMLA[,i] =="3" ,i ] = 0
  EDMLA[EDMLA[,i] =="2",i ] = 1
  EDMLA[,i] = as.numeric(EDMLA[,i])
}
```

```
datos  = EDMLA[,1:(ncol(EDMLA)-1)]
for (i in 1:ncol(datos)){
  datos[,i] = as.numeric(datos[,i],na.rm=TRUE)
}
#Redueix columnes amb tot '0'
c = 1
borrar = c(ncol(datos))
for (i in 1:ncol(datos)) {
  if (sum(datos[,i])==0) {
    borrar[c]=i
    c = c + 1
  }
}
EDMLA[,borrar] = NULL

#redueix files amb tot '0'
c = 1
borrar = c(nrow(datos))
for (i in 1:nrow(datos)) {
  if (sum(datos[i,],na.rm = TRUE)==0) {
    borrar[c]=i
    c = c + 1
  }
}
EDMLA = EDMLA[-borrar,]
# Transformem tot a factor
for (i in 1:(ncol(EDMLA)-1)){
  EDMLA[,i] = as.factor(EDMLA[,i])
}

# Reordenar al atzar les dades
EDMLA <- EDMLA [sample(nrow(EDMLA)),]
#Separem conjunt de test i entrenament a partir de les dades
inTrain <- createDataPartition(y=EDMLA$rematricula, p=0.7, list=FALSE)
train <- EDMLA[inTrain,]
test <- EDMLA[-inTrain,]

# generem l'arbre
cart <- train(rematricula ~ .,data=train, method="rpart", na.action = na.exclude)

fancyRpartPlot(cart$finalModel,sub="", main="Assignatures amb impacte en la decisió de
tornar-se a matricular" )
```

11.38 Pregunta 5. Generar arbre amb menys desequilibri.

```
#Pregunta 5
#Carrega dades i llibreries
source('Carrega Inicial.R')
library(rpart)
```

```
library(rpart.plot)
library(caret)
library(rattle)

#Eliminem les variables sociodemogràfiques
EDMLA = EDMLA[,-(1:6)]
#Canviem  3 -> 1"
#        0,1,2 -> 0
# Assignatures aprovades
for (i in 1:(ncol(EDMLA)-1)){
  EDMLA[,i] = as.character(EDMLA[,i])
  EDMLA[EDMLA[,i] =="0" |EDMLA[,i] =="1"|EDMLA[,i] =="2" ,i ] = 0
  EDMLA[EDMLA[,i] =="3",i ] = 1
  EDMLA[,i] = as.numeric(EDMLA[,i])
}

# Reordenar al atzar les dades
EDMLA <- EDMLA [sample(nrow(EDMLA)),]
data_si = EDMLA[EDMLA$rematricula =="si",]
data_no = EDMLA[EDMLA$rematricula =="no",]

#Separem conjunt de test i entrenament a partir de les dades
inTrain_si <- createDataPartition(y=data_si$rematricula, p=0.45, list=FALSE)
inTrain_no <- createDataPartition(y=data_no$rematricula, p=0.55, list=FALSE)

#Combinem si i no en la proporció establerta
train = data_si[inTrain_si,]
test = data_si[-inTrain_si,]
train_2 = rbind(train, data_no[inTrain_no,])
test_2 = rbind(test,data_no[-inTrain_no,] )

# generem l'arbre
cart <- train(rematricula ~ .,data=train_2, method="rpart", na.action = na.exclude)
fancyRpartPlot(cart$finalModel,sub="", main="Assignatures amb impacte en la decisió de
tornar-se a matricular" )
```

11.39 Pregunta 5. Generar arbre amb menys desequilibri.

```
#Pregunta 5
#Carrega dades i llibreries
source('Carrega Inicial.R')
library(rpart)
library(rpart.plot)
library(caret)
library(rattle)

#Eliminem les variables sociodemogràfiques
EDMLA = EDMLA[,-(1:6)]
#Canviem  2 -> 1"
#        0,1,3 -> 0
# Assignatures aprovades
for (i in 1:(ncol(EDMLA)-1)){
```

```
 EDMLA[,i] = as.character(EDMLA[,i])
 EDMLA[EDMLA[,i] =="0" |EDMLA[,i] =="1"|EDMLA[,i] =="3" ,i ] = 0
 EDMLA[EDMLA[,i] =="2",i ] = 1
 EDMLA[,i] = as.numeric(EDMLA[,i])
}

# Reordenar al atzar les dades
EDMLA <- EDMLA [sample(nrow(EDMLA)),]
data_si = EDMLA[EDMLA$rematricula =="si",]
data_no = EDMLA[EDMLA$rematricula =="no",]

#Separem conjunt de test i entrenament a partir de les dades
inTrain_si <- createDataPartition(y=data_si$rematricula, p=0.45, list=FALSE)
inTrain_no <- createDataPartition(y=data_no$rematricula, p=0.55, list=FALSE)

#Combinem si i no en la proporció establerta
train = data_si[inTrain_si,]
test = data_si[-inTrain_si,]
train_2 = rbind(train, data_no[inTrain_no,])
test_2 = rbind(test,data_no[-inTrain_no,] )

# generem l'arbre
cart <- train(rematricula ~ .,data=train_2, method="rpart", na.action = na.exclude)

fancyRpartPlot(cart$finalModel,sub="", main="Assignatures amb impacte en la decisió de

tornar-se a matricular" )
```