UNIVERSITAT DE BARCELONA

Treball final de grau

GRAU DE INFORMÀTICA

Facultat de Matemàtiques
Universitat de Barcelona

Data Science aplicat a resultats acadèmics per a la millora del pla d'acció tutorial a la Universitat de Barcelona

Autor: Laura Portell Penadés

Director: Dr. Eloi Puertas Prats
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 30 de juny de 2016

Abstract

Academic tutors at universities find themselves without enough tools or material to help their students due to the lack of knowledge of each student's academic profile. For this reason, we have performed a comprehensive study of student data in order to obtain supplementary material for the tutor. Clusterization algorithms have been used to separate the students into groups to identify general characteristics of the students belonging to a particular cluster. Dropout of each cluster and relationships between groups have been studied as well. A k-NN classifier has been used to create a prediction model which is able to predict in which cluster a student will belong the next academic year. Finally, visualization technics have been used to present the results of the analysis.

Resum

Els tutors d'estudis de les universitats es troben sense eines o material suficient per ajudar als seus alumnes a causa de la manca de coneixement del perfil acadèmic de cada estudiant. Per aquesta raó hem realitzat un estudi exhaustiu de les dades dels estudiants per tal d'obtenir material complementari per al tutor. S'han usat algorismes de clusterització per a agrupar els alumnes i trobar d'aquesta manera les característiques generals dels alumnes que pertanyen a cada grup. També s'ha estudiat l'abandonament i les relacions entre els grups. S'ha usat un classificador k-NN per a crear un model de predicció que és capaç de predir en quin grup es trobarà l'alumne al curs següent. Finalment, s'han usat tècniques de visualització per a mostrar els resultats de l'anàlisi.

Resumen

Los tutores de estudios de las universidades se encuentran sin herramientas o material suficiente para ayudar a sus alumnos a causa de la falta de conocimiento del perfil académico de cada estudiante. Por esta razón hemos realizado un estudio exhaustivo de los datos de los estudiantes para obtener así material complementario para el tutor. Se han usado algoritmos de clustering para agrupar a los alumnos y encontrar de esta manera las características generales de los alumnos que pertenecen a cada grupo. También se ha estudiado el abandono y las relaciones entre los grupos. Se ha usado un classificador k-NN para crear un modelo de predicción que es capaz de predecir en que grupo se encontrará el alumno al curso siguiente. Finalmente, se han usado técnicas de visualización para mostrar los resultados del análisis.

Índex

Índex de figures

Índex de taules

1 Introducció

L'any 2013 es generaven cada segon 30 000 GB de noves dades, i es preveu que pel 2018 se'n generin 50 000 [1]. Trobar patrons i estructures dins d'aquestes dades pot aportar molts beneficis econòmics en les empreses com també pot augmentar la qualitat de vida de les persones. És per això que cada vegada és més important la disciplina de la ciència de les dades o, com és més conegut, Data Science.

L'objectiu principal de la ciència de les dades, o de la persona que treballa en aquest sector, com són els científics de dades, és extreure i examinar les dades, però tenint alhora una visió i coneixement de negoci per a poder interpretar les dades en el context adequat i poder així transmetre recomanacions a l'organització. És important també la visualització de les dades per a poder presentar els resultats d'una manera més atractiva i entenedora. Segons Anjul Bhambhri [2], un científic de les dades ha de ser *"en part analista, en part artista"*. Els defineix com:

> *"A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization."*
>
> Anjul Bhambhri, Vice President of Architecture at Adobe

Alhora de prendre la decisió sobre què realitzar el treball final de grau i al ser el Data Science un sector que em captivava tant, quan se'm va presentar la oportunitat de realitzar-lo sobre aquest àmbit, em va agradar molt la idea de poder aprendre i experimentar sobre la ciència de les dades.

1.1 Motivació

La Universitat de Barcelona disposa d'un *Pla d'Acció Tutorial* que proporciona als seus estudiants un tutor d'estudis, que alhora es tracta també d'un professor. Aquest guia i orienta als alumnes en els seus estudis a la universitat al llarg del grau, i prepara també als alumnes per a la inserció laboral.

El problema que hi ha actualment és que els tutors no només s'encarreguen de tutoritzar a un alumne, sinó que a un grup d'ells, i per tant no coneixen la situació acadèmica de cada alumne. Quan un alumne necessita orientació, el tutor no disposa d'ajuda o de més informació que el seu expedient acadèmic. Per tant, és difícil per al tutor fer el seguiment de l'alumne i poder oferir l'orientació necessària i adaptada per a cada estudiant.

És per resoldre els problemes mencionats anteriorment per el que es va iniciar el projecte d'innovació docent anomenat *"Sistema Intel·ligent de suport al Tutor d'estudis"* [3] del grup consolidat INDOMAIN, format per professors de la Facultat de Matemàtiques, on l'objectiu principal és el de desenvolupar un sistema intel·ligent de suport al Tutor d'estudis que li permeti prendre unes decisions més informades.

Anteriorment al meu projecte s'han realitzat dos treballs finals de grau relacionats amb el PID. Aquests són els del Xavier Moreno Liceras i el Daniel Gabriel Urdas.

Aquest treball forma part del projecte d'innovació docent i es centre en l'anàlisi dels alumnes de diferents graus de la Universitat de Barcelona. D'aquesta manera es vol ajudar al tutor d'estudis a conèixer millor el perfil dels estudiants i així permetre al tutor poder guiar a l'alumne, basant-se en les dades i no només en intuïcions.

1.2 Objectius

L'objectiu principal d'aquest treball és el de poder ajudar als tutors d'estudis dels graus, de manera que puguin saber la situació acadèmica fins al moment d'un alumne i saber si tindrà dificultats o no de cara als següents cursos.

Les tasques globals a realitzar en el projecte per arribar a complir l'objectiu principal són:

1. Obtenció i neteja de dades dels diferents graus de la Universitat de Barcelona
2. Anàlisi i visualització de dades
3. Implementació de models predictius sobre el rendiment futur dels estudiants

4. Mostrar resultats útils per al tutor

S'han enumerat els objectius d'aquest projecte de cara a poder assolir tots els punts per al projecte d'innovació docent, però els objectius personals al realitzar aquest projecte són els d'entendre una mica millor les funcions d'un *data scientist* i com les realitza, aprendre a tractar còmodament les dades i entendre i poder aplicar algorismes d'agrupament i classificació, aprendre a organitzar-me per a projectes de més duració i finalment adquirir coneixements de visualització de dades.

1.3 Estructura de la Memòria

En aquesta secció es pot veure l'estructura de la memòria que resta. Principalment es pot dividir en sis capítols. A continuació es descriurà breument el contingut de cadascun d'ells.

En el capítol 2 es parlarà sobre la planificació inicial i real del treball així com de l'estimació dels costos de la realització del treball. El capítol 3 descriu el plantejament del treball i les preguntes que es realitzen. El capítol 4 està dedicat al disseny del treball, on es veu com hem preparat les dades i els algorismes que hem usat. En el capítol 5 es veu el desenvolupament del treball, on es mostraran les eines que s'han usat i com s'han aplicat els algorismes a les dades. En el capítol 6 es presenten els experiments que hem realitzat i els resultats que hem obtingut. I, finalment, el capítol 7 mostra una anàlisi crític del treball realitzat i es veurà si s'han obtingut els objectius marcats. Es proposa també possibles continuacions del treball.

2 Planificació

Aquesta secció mostra la planificació del projecte abans de començar a fer-lo i com ha acabat sent finalment. Es mostra també una estimació del cost econòmic que hauria tingut el projecte.

2.1 Planificació inicial

Abans de començar el projecte es va fer una planificació del temps que s'estimava que es trigaria a realitzar cada tasca. Per a fer-ho es va haver de fixar les tasques que es realitzarien i estimar quin temps es preveia que es trigaria a realitzar cada una d'elles.

El treball final de grau consta de 18 crèdits ECTS. Com cada crèdit equival a 25 hores de treball, s'estima que el treball final de grau equival a 450 hores de feina. El temps per a realitzar el treball és d'aproximadament 20 setmanes, i per tant s'estima una dedicació de 22,5 hores setmanals. És cert que no cada setmana es dediquen les mateixes hores. Al principi estava plantejat realitzar aquest treball el primer semestre, però finalment no va ser així i s'ha realitzat el segon semestre. És per això que la planificació inicial està plantejada per uns altres mesos als quals han acabat sent.

Les tasques principals amb les que es va separar el projecte van ser:

- Formació i estudis previs
- Plantejament de les preguntes
- Preparació de les dades
- Desenvolupament del projecte
- Visualització i avaluació dels resultats
- Realització de la memòria

En el diagrama de Gantt de la figura 1 es mostra la planificació plantejada en un inici.

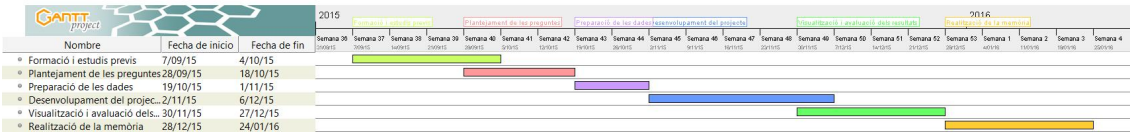


Figura 1: Diagrama de Gantt de la planificació inicial.

2.2 Planificació real

La planificació real, exceptuant el fet d'haver posposat el treball, com es pot veure en el quadre gris de la figura 2, ha estat bastant semblant a l'estimada. Sí que es pot veure que hi ha hagut molta diferència en el temps estimat per a preparar les dades, ja que no s'esperava haver-li de dedicar tant de temps a la neteja de dades. També vam estimar massa poc temps en la tasca de visualització i avaluació dels resultats, ja que finalment se li ha dedicat molt més temps de l'esperat a la visualització de les dades. És per això que aquesta tasca s'ha acabat realitzant alhora amb la del desenvolupament i amb la de l'escriptura de la memòria.

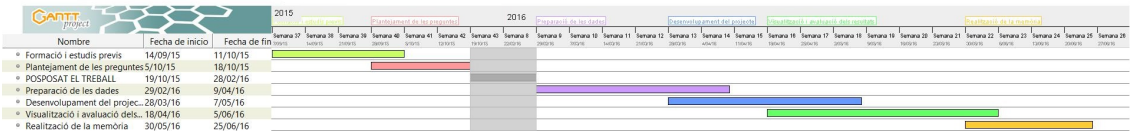


Figura 2: Diagrama de Gantt de la planificació real.

2.3 Avaluació econòmica

En aquesta secció hem fet una estimació econòmica del projecte.

	Hores	**€/hora**	**Preu total (€)**
Formació i estudis previs	30	0	0
Plantejament de les preguntes	35	10	350
Preparació de les dades	90	20	1800
Desenvolupament del projecte	115	35	4025
Visualització i aval. dels resultats	125	35	4375
Realització de la memòria	55	0	0
TOTAL	450		10550

Taula 1: Avaluació econòmica del projecte

L'estimació econòmica del preu del projecte és de 10550 €. Inclou totes les fases del procés que hem seguit per a la realització del projecte.

3 Requeriments del problema

Per a poder començar a fer el projecte cal planejar bé el problema i saber exactament el que es vol. Com s'ha vist abans, el que es vol és poder donar al tutor d'estudis més informació sobre un alumne, per a poder-lo guiar i orientar al llarg dels estudis.

La informació que proporcionarem al tutor estarà dividida en dues seccions diferents: una secció serà sobre la situació acadèmica de l'alumne fins el moment i l'altre sobre la informació predita per un dels models que desenvoluparem en aquest treball.

El primer pas abans de començar a tractar i analitzar les dades és plantejar-se bé les preguntes per a poder saber quina informació es vol, i per tant, poder saber quines són les dades que es necessiten i poder saber què fer amb elles.

3.1 Preguntes plantejades

Com s'ha vist prèviament, abans de començar a tractar les dades i analitzar-les ens hem plantejat les següents preguntes, per a resoldre al llarg del projecte i permetre'ns arribar als objectius proposats.

- Influeix la via d'accés o la nota d'accés a l'hora d'abandonar la carrera? A enginyeria informàtica, i sobretot a matemàtiques, abandona molta gent abans de cursar segon. És per això que ens preguntem quina és la gent que té més tendència a abandonar per a poder informar el tutor de quins són els alumnes que tenen més probabilitats de no seguir.

- Es pot separar en grups els alumnes de manera que cada grup tingui un perfil acadèmic diferent? En aquest cas, en quants grups s'ha de separar cada curs? Es poden trobar les característiques generals dels alumnes que pertanyen a cada grup? Si la resposta fos afirmativa, podríem obtenir molta informació d'un alumne simplement sabent en quin clúster pertany.

- Hi ha diferència d'abandonament entre els grups? És a dir, hi ha perfil d'alumnes que abandonen més que els altres? Si és cert que hi ha més abandonament

en un grup que en un altre es podria anar més amb compte amb els alumnes que es troben en grups amb probabilitat alta d'abandonament.

- Es conserven els grups mencionats prèviament al llarg dels cursos? Si es veu que es conserven els grups al llarg dels cursos significarà que, si un alumne comença malament, tindrà la tendència a seguir així i s'haurà de mirar d'evitar-ho.

- Es pot predir en quin grup es trobarà cada alumne al següent curs? En cas de predir-ho, és fiable el predictor? Si es pot predir i es prediu que un alumne anirà a un grup amb problemes, es podrà aconsellar i tutoritzar a aquest alumne de manera més efectiva.

4 Disseny

Per a realitzar el projecte es seguirà l'esquema que es mostra a la figura 3, on es veuen les principals fases que es realitzaran. Aquestes són "Preparació de les dades", "Anàlisi de les dades" i "Visualització de les dades". Cada fase té les seves parts, que es veuran en les seccions següents.

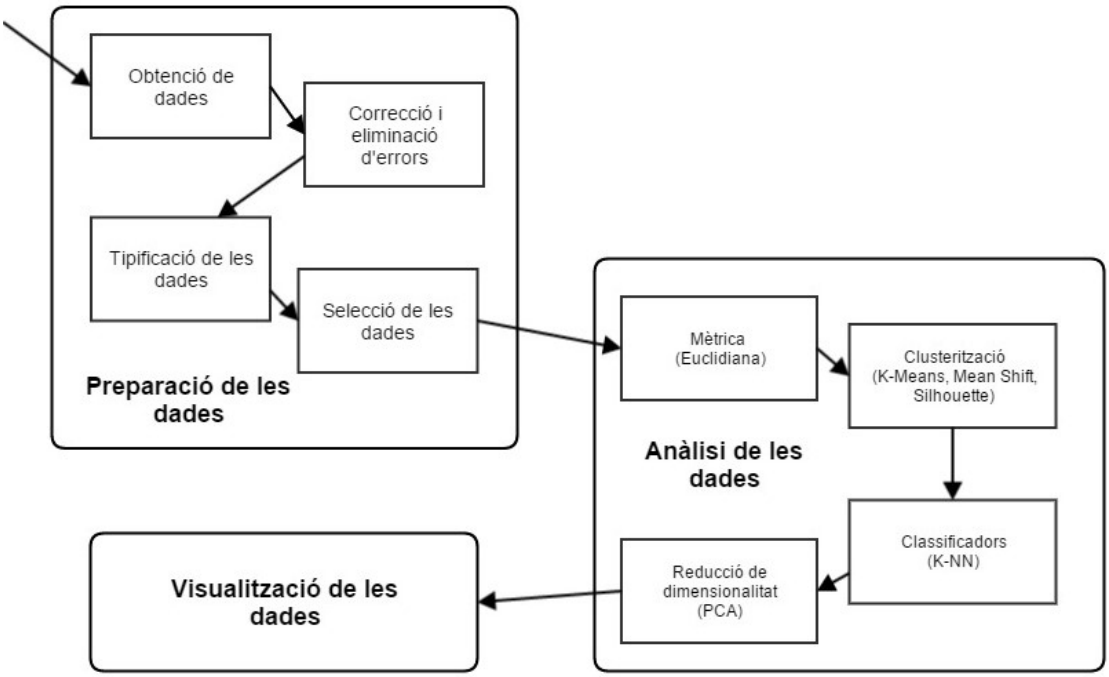


Figura 3: Esquema seguit per a la realització del projecte.

4.1 Preparació de les dades

En aquest capítol es veuen les dades que s'han obtingut i com s'han tractat fins a poder-les començar a analitzar. Per a fer-ho es segueix l'esquema de "Preparació de dades" que s'ha vist abans. Es mostra a la següent figura per tal de poder seguir bé l'ordre que se seguirà.

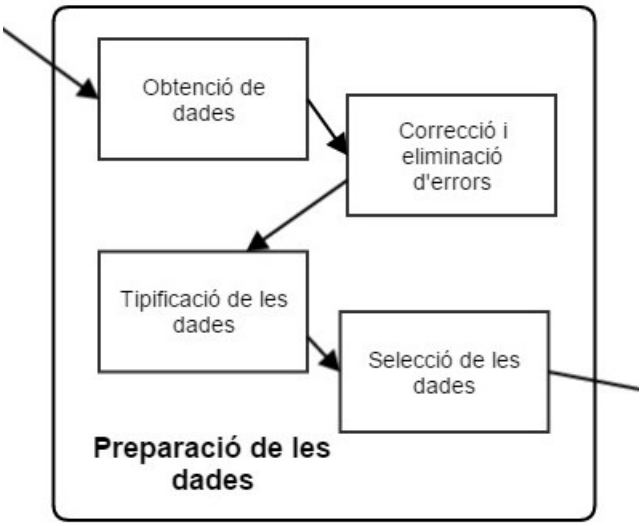


Figura 4: Esquema seguit per a la preparació de les dades.

Es dividirà la secció de preparació de les dades en cinc parts. A la secció 4.1.1 es podrà veure les dades que s'han obtingut i com han sigut tractades fins a poder-les tenir preparades per als següents passos.

Al capítol 4.1.2 es corretgiran els errors que poden tenir les dades i s'eliminaran els duplicats que hi hagi. Més endavant, a la part 4.1.3, es tifipicaran les dades. Finalment, a la secció 4.1.4, s'agafaran les dades que es vol per a realitzar el projecte i es posaran en el format més convenient per a tractar-les.

4.1.1 Obtenció de les dades

Les dades que vam obtenir primer van ser les de la facultat de Matemàtiques, i per tant les del grau d'Enginyeria Informàtica i el grau de Matemàtiques. Més endavant vam rebre les de la facultat de Dret, obtenint així les dades dels graus següents: Ciències Polítiques i de l'Administració, Dret, Gestió i Administració Pública, Criminologia i Relacions Laborals. El projecte s'ha fet analitzant les dades d'aquests 7 diferents graus de la Universitat de Barcelona.

Els primers alumnes matriculats dels quals es disposa dades es van matricular per primera vegada l'any 2009. Els més actuals es van matricular el 2014.

Les dades que se'ns van proporcionar de cada facultat són quatre fulls de càlcul, on cadascun d'ells conté la següent informació:

1. Informació dels alumnes abans d'entrar a la universitat: Conté per a cada alumne la seva informació, com per exemple, la nacionalitat, la nota d'accés, l'any que va fer la PAU, l'any de la primera matrícula, el grau al qual es matricula o el sexe. No totes les caselles estan emplenades.

2. Informació de les assignatures: Cada fila és una assignatura diferent, on es mostra l'identificador del grau en què es cursa l'assignatura, l'identificador de l'assignatura, el nom de l'assignatura i els crèdits d'aquesta.

3. Informació global dels alumnes a la universitat: Cada fila és un alumne, on es mostra el seu identificador, els crèdits superats que porta fins al moment i la seva dedicació, entre d'altres.

4. Qualificacions dels alumnes: Cada fila és d'un alumne en una assignatura en concret. On mostra l'identificador de l'alumne i de l'assignatura, l'any que l'ha matriculat, la nota que ha obtingut en la primera i la segona convocatòria, i més informació.

El primer que s'ha fet ha sigut passar els fulls de càlcul a un fitxer amb format *csv* per a poder manipular les dades amb més facilitat. Després s'han re-anomenat les columnes de manera que les dades de la facultat de Matemàtiques i les de Dret siguin iguals, i de manera que cada fitxer *csv* tingui els mateixos noms a les columnes quan es tracta dels mateixos atributs. D'aquesta manera s'han pogut creuar més endavant les dades, en tenir el mateix nom en les columnes.

Anonimització de les dades

Al llarg de tot el projecte, sempre que es parla de l'identificador de l'alumne o, directament, de l'alumne, no es tracta del NIUB ni de cap informació que permeti conèixer la identitat de l'alumne. Cada alumne s'identifica amb una id generada de forma aleatòria, i les dades de què disposem no mostren en cap moment informació

personal de l'alumne, com seria el nom, telèfon, correu, etc. Tot i això no hem anonimitzat les dades com estableix el protocol de la Universitat de Barcelona, ja que no es faran públiques aquestes dades. Per a anonimitzar correctament les dades s'hauria de fer de tal manera que la seva desanonimització no fos possible. És a dir, que amb la informació anonimitzada de l'alumne fos completament impossible saber de qui es tracta.

4.1.2 Correcció i eliminació d'errors

El primer canvi s'ha realitzat en les dades corresponents al grau d'Enginyeria Informàtica, que al curs 2010 va canviar l'identificador de la carrera. S'ha fet una unió dels dos identificadors del mateix grau per a tenir tot el grau en un sol identificador.

Un altre canvi que s'ha fet és el de canviar les comes per punts, és a dir, a les caselles on ens mostraven les notes dels alumnes ho feien posant una coma abans dels nombres decimals, per tant, quan es volia tractar les notes com a números no es podia.

Per veure la nota final d'un alumne en una assignatura tenim la nota de la primera convocatòria i de la segona. Com molts professors no posen cap nota a la segona convocatòria perquè les posen directament totes a la primera, s'ha decidit unificar la primera i la segona convocatòria, agafant sempre la nota més alta de les dues. D'aquesta manera si un professor fa servir la segona convocatòria, la nota que quedarà de l'assignatura serà aquesta. En cas contrari, la nota serà la de la primera convocatòria.

S'ha modificat també la nota d'accés de l'alumne. Abans la nota d'accés era sobre 10, i des de el 2010 és sobre 14. És per això que si un alumne accedeix a la universitat amb via d'accés de batxillerat o amb nota de la PAU, i aquest alumne ha realitzat la nota de PAU més tard del 2009, passem la nota que estava sobre 14, a sobre 10.

Finalment hem trobat que hi havia assignatures repetides d'un alumne cursades el mateix any, però un cop constaven com a assignatura de "Reconeixement" i un

altre com a "Ordinari". És a dir, que al convalidar l'assignatura l'havien inscrit dues vegades. El que això provocava era que quan miràvem quantes vegades havia cursat cada assignatura un alumne, per a veure quantes vegades la repetia, el fet de tenir-ho dues vegades, contava com a haver realitzat dues vegades l'assignatura. Per aquest motiu s'han eliminat aquests casos repetits.

4.1.3 Tipificació de les dades

Per a poder comparar notes d'una mateixa assignatura però procedents de diferents cursos es normalitzen [4] les notes. D'aquesta manera si es canvia el pla docent, hi ha canvi de professor, o altres factors que puguin influenciar en el canvi de notes, en tipificar les dades això se soluciona i permet tractar totes les dades juntes sense haver de tenir en compte els casos en els quals es poden trobar.

La fórmula que s'usa és la següent:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

sent x_i la nota de cada assignatura sense normalitzar, μ la mitjana de les notes d'aquella assignatura i d'aquell mateix any, i σ és la desviació típica, també d'aquella assignatura aquell any.

S'obtindran unes dades amb mitjana 0 i desviació típica 1. D'aquesta manera, sempre que es mostri la mitjana de les notes d'una assignatura, el que mostrarà aquell valor és quan per sobre o per sota de la mitjana es trobarà l'alumne. Per això també hi ha números negatius.

4.1.4 Selecció de les dades

Un cop ja es té les dades netejades el que es vol és deixar les dades preparades per a poder-les utilitzar. És a dir, triar la informació que s'usarà i estructurar-la de tal manera que sigui més fàcil realitzar la part d'anàlisi de les dades. Per fer això hem creat una nova variable que anomenarem *rep_i*, on $i \in 1, 2, 3$, que representa si un alumne repeteix moltes assignatures o no. El valor d'aquesta variable es calcula de

la següent manera:

$$rep\_i = \frac{n_i}{x_i} \cdot 10 \text{ , on } rep\_i \in (0, 10]$$

n_i: nombre total d'assignatures del curs i

x_i: nombre total d'assignatures del curs i cursades per l'alumne.

Per tant, s'ha de veure quines són les dades que s'usaran per a la realització del projecte i quines no són necessàries. El que volem obtenir per a realitzar la part de clusterització i predicció són tres matrius diferents les quals passem a explicar a continuació.

- La primera matriu correspon a les dades dels m_1 alumnes que han cursat totes les assignatures del primer curs, juntament amb rep_1.

$$\begin{array}{c} \begin{array}{ccccc} s_1 & s_2 & \dots & s_{n_1} & rep_1 \end{array} \\ \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n_1} & a_{1,n_1+1} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n_1} & a_{2,n_1+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m_1,1} & a_{m_1,2} & \dots & a_{m_1,n_1} & a_{m_1,n_1+1} \end{pmatrix} \end{array} \tag{2}$$

- La segona matriu correspon a les dades dels m_2 alumnes que han cursat totes les assignatures de primer i segon curs, juntament amb rep_1 i rep_2.

$$\begin{array}{c} \begin{array}{cccccc} s_1 & s_2 & \dots & s_{n_1+n_2} & rep_1 & rep_2 \end{array} \\ \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n_1+n_2} & a_{1,n_1+n_2+1} & a_{1,n_1+n_2+2} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n_1+n_2} & a_{2,n_1+n_2+1} & a_{2,n_1+n_2+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{m_2,1} & a_{m_2,2} & \dots & a_{m_2,n_1+n_2} & a_{m_2,n_1+n_2+1} & a_{m_2,n_1+n_2+2} \end{pmatrix} \end{array} \tag{3}$$

- La tercera matriu correspon a les dades dels m_3 alumnes que han cursat totes

les assignatures de primer i segon curs, juntament amb rep_1, rep_2 i rep_3.

$$\begin{array}{ccccccc} s_1 & s_2 & \dots & s_{n_1+n_2+n_3} & rep_1 & rep_2 & rep_3 \\ \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{m_3,1} \end{pmatrix} & \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ \vdots \\ a_{m_3,2} \end{pmatrix} & \begin{pmatrix} \dots \\ \dots \\ \ddots \\ \dots \end{pmatrix} & \begin{pmatrix} a_{1,n_1+n_2+n_3} \\ a_{2,n_1+n_2+n_3} \\ \vdots \\ a_{m_3,n_1+n_2+n_3} \end{pmatrix} & \begin{pmatrix} a_{1,n_1+n_2+n_3+1} \\ a_{2,n_1+n_2+n_3+1} \\ \vdots \\ a_{m_3,n_1+n_2+n_3+1} \end{pmatrix} & \begin{pmatrix} a_{1,n_1+n_2+n_3+2} \\ a_{2,n_1+n_2+n_3+2} \\ \vdots \\ a_{m_3,n_1+n_2+n_3+2} \end{pmatrix} & \begin{pmatrix} a_{1,n_1+n_2+n_3+3} \\ a_{2,n_1+n_2+n_3+3} \\ \vdots \\ a_{m_3,n_1+n_2+n_3+3} \end{pmatrix} \end{array} \tag{4}$$

El quart curs no s'ha tingut en consideració, ja que la majoria d'assignatures, si no totes, són optatives i per tant cada alumne cursa assignatures diferents.

Tot i tenir aquestes tres matrius, per a realitzar la part principal del projecte hem usat altres dades com la via d'accés, la nota d'accés, l'any de matriculació de l'alumne o l'any que s'ha cursat cada assignatura.

4.2 Anàlisi de les dades

En aquest capítol s'explicaran les fórmules o algorismes aplicats a les dades per tal de poder respondre a les preguntes plantejades i poder així assolir els objectius proposats.

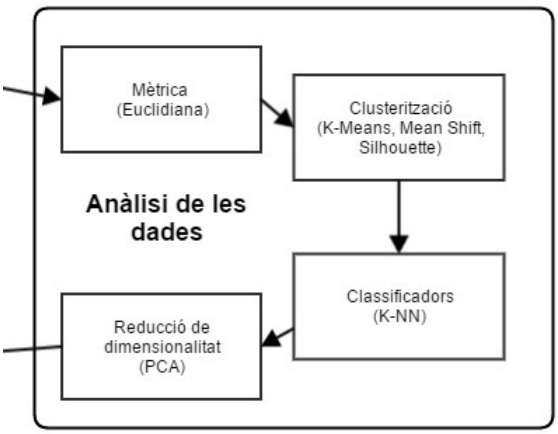


Figura 5: Esquema seguit per a l'anàlisi de les dades.

Es veurà primer quina és la mètrica que s'ha utilitzat. Després, a la secció 4.2.2

s'explicarà en què consisteix la clusterització de les dades i quins algorismes hem usat per a fer-ho. Veurem també (capítol 4.2.3) el classificador que hem usat, per a realitzar el predictor més endavant. I finalment (secció 4.2.4) veurem amb quin mètode reduïm les dimensions de les dades per a poder-les representar en dues dimensions.

4.2.1 Mètrica empleada

Per a poder mesurar la proximitat entre els objectes usem la mètrica euclidiana. Siguin $P = (p_1, p_2, ..., p_n)$ i $Q = (q_1, q_2, ..., q_n)$ dos punts en l'espai de dimensió n $\mathbb{R}^n$ la distància euclidiana entre els dos objectes P i Q és:

$$d_E(P, Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{5}$$

4.2.2 Clusterització

El procés de clusterització o d'agrupament consisteix en la divisió de les dades en grups d'objectes similars, de manera que el que s'intenta aconseguir és fer la millor divisió de manera que cada grup sigui el màxim d'homogeni entre ell i diferent de la resta. Per tant un clúster es defineix com a l'agregació de punts en l'espai d'entrada on la distància entre cada par d'objectes és menor que la distància de qualsevol dels punts del clúster als altres objectes que no pertanyen al clúster. Per a poder mesurar la similitud entre objectes es pot fer amb diferents distàncies, com la Manhattan, Minkowski o Euclidiana. En aquest projecte usarem la distància euclidiana, on la distància es pot veure en la fórmula 5.

Els mètodes d'agrupació els podem classificar en quatre tipus: de particions, jeràrquics, probabilístics i basats en densitats.

1. Els algorismes d'agrupament particionals parteixen l'espai en conjunts de dades. La idea general d'aquests mètodes és la de partir en un inici tots els objectes i a cada iteració intercanviar els objectes a un altre clúster en cas d'obtenir així una partició millor.

2. Els algorismes d'agrupament jeràrquic són aquells que realitzen una descomposició jeràrquica dels objectes. És a dir, els objectes no es divideixen en clústers d'un sol cop, sinó que van fent particions a diferents nivells d'agrupació. Es solen mostrar en arbres binaris o dendrogrames.

3. Els algorismes d'agrupament probabilístic assumeixen que els objectes són generats a partir d'alguna distribució probabilística. Els objectes que es troben en diferents clústers són generats per distribucions de probabilitat diferents.

4. Els algorismes d'agrupament basats en densitats agrupen els objectes que es troben en espais de gran densitat d'objectes, els quals estan separats per regions de menor densitat (soroll).

En el nostre cas hem usat l'algorisme Mean Shift (secció 4.2.2), que és un algorisme basat en densitats i el K-Means (secció 4.2.2), que és particional.

MeanShift

L'algorisme Mean Shift [5] és un algorisme no paramètric iteratiu, és a dir, que no necessita com a paràmetre el nombre de clústers. La idea principal és que cada objecte s'assigna a l'àrea més densa que tingui més a prop, basant-se en una estimació de la densitat del nucli. Finalment els objectes convergeixen a màxims locals de densitat. Els passos que es segueixen són:

1. Es fixa una finestra al voltant de cada punt de dades.

2. Es calcula la mitjana de les dades dins de la finestra.

3. Es desplacen les finestres a les mitjanes

4. Es repeteixen els passos anteriors fins que l'algorisme convergeix a la solució. Finalment les dades estaran agrupades segons indiquin les finestres.

K-Means

L'algorisme K-Means [6] té com a objectiu dividir el conjunt de dades en K grups.

Es basa en la minimització de la distància interna, és a dir, que la suma de les distàncies de cada objecte al centre del clúster sigui mínima (veure equació 6). La distància que s'usa és l'Euclidiana, i es pot usar en els casos en què les variables siguin quantitatives. L'algorisme necessita conèixer el nombre K, és a dir, el nombre de clústers en què es dividirà les dades. Per a trobar el valor de K ho fem usant el coeficient de Silhouette que s'explica posteriorment.

L'algorisme és iteratiu i el procediment és el següent:

- Es comença seleccionant K punts (centroides) sobre l'espai dels objectes de manera aleatòria.
- S'assigna cada objecte al centroide més proper.
- Després es recalcula la posició del centroide, fent que se situï a la mitjana de tots els ítems que estan associats a ell. D'aquesta manera es minimitza el sumatori de la suma dels quadrats dintre de cada grup (WCSS, *within-cluster sum of squares*), com a la fórmula 6.
- Es repeteixen successivament els dos passos anteriors fins que els centroides de tots els grups es quedin fixes, o fins que es compleixi alguna de les condicions de parada, com podria ser assignar un nombre màxim d'iteracions.

$$WCSS(C) = min \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i - \mu_j \right\|^2 \tag{6}$$

on:

-x_i és l'objecte i

-μ_j és el centroide j

-n és el nombre d'objectes que tenim

-k és el nombre de clústers en què volem separar les dades

Coeficient de Silhouette

Els algorismes d'aprenentatge no supervisat necessiten un valor per a identificar

el nombre d'agrupaments a realitzar o com són de separables les agrupacions produïdes. El coeficient de Silhouette [7] ens permetrà obtenir el nombre òptim de clústers. En el nostre cas l'usarem per a saber quina k usar per a l'algorisme K-Means.

Per a calcular el coeficient de Silhouette d'una partició ja donada, es calcula el coeficient de Silhouette de cada element del conjunt de dades i es fa la mitjana. Per a cada element i, el coeficient de Silhouette $s(i)$ es calcula amb la següent fórmula:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{7}$$

on:

-a_i és la distància mitjana entre l'objecte i i tots els altres objectes del mateix clúster.

-b_i és la distància mitjana entre l'objecte i i tots els altres objectes del clúster més proper a l'objecte i (clúster B). Es veu per tant que el clúster B seria la segona millor opció per a l'objecte i, per tant si l'element i estigués en el clúster que no li correspon, $b_i < a_i$ i per tant el numerador $b_i - a_i < 0$, i el coeficient de Silhouette per a l'objecte i seria un nombre negatiu.
Es veu que $-1 \leq s(i) \leq 1$

-Si $s(i)$ és proper a 1 vol dir que l'objecte i està ben classificat, ja que voldrà dir que $a(i) \ll b(i)$, i per tant implicaria que l'objecte i és molt similar als objectes del seu clúster i que a més i no és similar als objectes del clúster més proper.

-Si $s(i)$ és proper a 0 vol dir que l'objecte i es troba entre dos clústers, ja que $a(i) \approx b(i)$.

-Si $s(i)$ és proper a -1 l'objecte i no es troba en el clúster que més li correspon i per tant està mal classificat.

Per tant, donat que el coeficient de Silhouette és la mitjana de tots els $s(i)$ $\forall i$, com més gran sigui el coeficient de Silhouette indicarà una millor partició de clústers.

4.2.3 Classificadors

Els classificadors s'usen per a assignar a un element no etiquetat una categoria concreta coneguda. Hi ha molts algorismes classificadors, i nosaltres usarem el k-NN o veí més proper.

k-NN

El mètode k-NN [8] (*k-Nearest Neighbors*) és un mètode de classificació supervisat, no paramètric. La idea general del k-NN és la de classificar un objecte a la classe més freqüent a la qual pertanyen els seus K veïns més propers.

Per tant, les dades d'entrada que tenim per al mètode k-NN són els n objectes ja classificats: (x_i, c_i), $i = 1...n$, on cada x_i conté les seves N variables predictores, $x_i = (x_{i,1}, ..., x_{i,N})$ i on $c_i \in c^1, ..., c^m$ denota els m possibles valors que podem predir.

El nou cas que volem predir serà de la forma $x = (x_1, ...x_N)$ i voldrem predir la variable C.

Per a fer-ho se segueixen els següents passos:

1. Per a cada objecte dels que ja tenim classificats (x_i, c_i) calculem $d_i = dist(x_i, x)$
2. Ordenem les distàncies en ordre ascendent
3. Ens quedem amb els k casos classificats més propers a x
4. Assignem a x la classe més freqüent dels k objectes seleccionats

4.2.4 Reducció de dimensionalitat

Reduir la dimensionalitat de les dades és molt útil per a representar les dades en gràfics de dues o tres dimensions. Hi ha moltes maneres de fer-ho, sigui amb aprenentatge supervisat (test de múltiples hipòtesis, informació mútua o LDA) o no supervisat (PCA o ICA). En el nostre cas explicarem i usarem el PCA.

PCA

El PCA [9] (*Principal Component Analysis* o anàlisi de components Principals) té com a objectiu trobar un espai de dimensió menor que preservi la major quantitat d'informació continguda en l'espai original.

L'anàlisi de components principals realitza una transformació lineal de les dades en un nou sistema de coordenades ortogonals. Els vectors de projecció de les dades en el nou espai són les direccions de màxima variància de les dades d'entrada. Les noves variables resultants en realitzar la projecció a les dades d'entrada són les components principals (PC). En el nou sistema de coordenades les components principals obtenen de forma seqüencial la màxima variància. És a dir, la primera component principal és la que té una major variància. La segona component principal és la segona variància més gran, i així successivament.

4.3 Visualització de les dades

Per a visualitzar les dades al llarg dels experiments hem usat diferents tipus de gràfiques: diagrames de barres, diagrames de sectors, diagrames de línies, diagrames de dispersió o mapes de calor.

A l'haver separat els alumnes en grups i normalment tenir els gràfics separats en clústers, per a facilitar la visualització de les dades, s'ha seguit un ordre a l'hora d'assignar un color a cada clúster. D'aquesta manera tots els gràfics que mostrin resultats dels clústers tindran aquests colors fixats.

- Alumnes amb bones notes.
- Alumnes que solen aprovar les assignatures a la primera.
- Alumnes que repeteixen assignatures tot i acabar aprovant.
- Alumnes que ho suspenen gairebé tot.

Veurem que a primer curs d'Enginyeria Informàtica i de Matemàtiques s'obtenen

uns grups amb perfils d'alumnes diferents de la resta de graus i cursos. És per això que per aquests casos concrets hem canviat la gamma de colors, per a no associar els colors anteriors amb el perfil d'alumnes que no és el correcte.

 Alumnes que solen aprovar a la primera (amb bones notes o sense).

 Alumnes que repeteixen bastant tot i acabar aprovant.

 Alumnes que suspenen tot.

 Alumnes que suspenen amb pitjors mitjanes que el grup anterior.

5 Desenvolupament

En aquest apartat es veuran les eines que hem usat per a poder realitzar el projecte, i com s'han aplicat els algorismes vistos abans per a poder assolir els objectius del projecte.

5.1 Eines usades

Fem una breu introducció a les eines que hem usat per a la realització del treball. Se separen en eines de gestió, edició i programació.

5.1.1 Eines de gestió

En aquesta secció veurem les eines de gestió que hem usat per a treballar d'una manera més còmoda i organitzada.

Trello

Trello [10] és una plataforma en línia que permet gestionar les tasques, tant a nivell d'equips com personal. Ens permet descompondre per a cada projecte diverses llistes. Cadascuna d'aquestes llistes es pot emplenar amb tasques, documents, imatges o altres coses. Es poden assignar dades límit per a la realització de cada tasca, o assignar a un usuari per a fer-la. Es basa en el mètode Kanban per a la gestió de projectes.

Bitbucket

Bitbucket [11] es tracta d'una plataforma de desenvolupament de software col·laboratiu semblant a GitHub. A més permet disposar d'un nombre il·limitat de repositoris privats, que permeten realitzar projectes de codi tancat sense que el puguis visualitzar persones que no estiguin incloses en el projecte.

5.1.2 Eines d'edició

Les eines d'edició que mostrarem ens han permès la creació i edició de la memòria, així com un entorn de programació molt còmode i clar per a la realització del codi.

ShareLatex

ShareLaTeX [12] és un editor en línia de document LaTeX [13] que serveix per a la creació de tesis o llibres, entre d'altres, sobretot de caire científic. Permet compilar els projectes de forma online en format PDF. A diferència d'altres editors de LaTeX es pot accedir des del web.

Ipython Notebook

Ipython Notebook [14] es tracta d'una aplicació web que permet la combinació de codi, text, equacions matemàtiques, figures i mitjans audiovisuals en un sol document.

5.1.3 Eines de programació

En aquest apartat es pot veure el principal llenguatge de programació que hem usat i també les llibreries que hem empleat per al càlcul, manipulació i visualització de les dades.

Python

Python [15] és un llenguatge de programació d'alt nivell que permet, gràcies a la seva sintaxi, usar menys línies de codi que altres llenguatges. Permet diferents estils de programació, com programació orientada a objectes, o funcional. És molt útil per al desenvolupament de projectes de data science en tenir moltes llibreries que permeten la manipulació d'una gran quantitat de dades de manera eficient i ràpida.

Pandas

Pandas [16] és una llibreria especialitzada en la manipulació i anàlisi de dades. Ofereix una sèrie d'estructures de dades que permeten tractar les taules, tant

numèricament com per a realitzar consultes, agrupar o agregar dades, amb comoditat.

Scikit-learn

Scikit-learn [17] és una llibreria de Python que agrupa algorismes de classificació, regressió i clusterització.

Numpy

Numpy [18] és una llibreria de Python que proporciona funcions específiques pel càlcul numèric de vectors i matrius.

Bokeh

Bokeh [19] és una llibreria de Python que permet realitzar gràfics i figures interactives de grans quantitats de dades.

Seaborn

Seaborn [20] és també una llibreria de visualització de dades de Python. Ofereix sobretot gràfics estadístics.

Matplotlib

Matplotlib [21] és una llibreria que permet la generació de gràfics a partir de dades contingudes en llistes o en arrays. El llenguatge de programació és Python. Permet realitzar gràfics en 2D amb gran facilitat.

5.2 Clusterització de les dades

Per a poder agrupar els alumnes de cada curs s'ha hagut de triar quin algorisme usar per a fer-ho. Es va pensar aplicar l'algorisme DBSCAN o el Mean Shift per a agrupar les dades, però com el DBSCAN elimina alguns alumnes i el MeanShift (capítol 4.2.2) ens separava en un nombre de clústers que no ens és útil, es va decidir fer-ho amb l'algorisme K-Means (veure capítol 4.2.2).

Per a poder-lo usar se li ha de passar un paràmetre k, on k és el nombre de clústers en què volem separar les dades.

Per a triar el paràmetre k per a usar a l'algorisme K-Means s'ha usat el coeficient de Silhouette (capítol 4.2.2). El que s'ha fet ha sigut usar el mètode K-Means per a diferents valors de k, i s'ha calculat el coeficient de Silhouette de cada agrupament. Finalment s'ha escollit la k amb un coeficient de Silhouette major.

Un cop triada la k apliquem l'algorisme K-Means sobre les matrius que tenim de cada curs (vist al capítol 4.1.4 de "Obtenció de les dades") i obtenim així els clústers de cada curs.

5.3 Predictor dels clústers

Per a crear un predictor i així intentar predir a quin clúster anirà l'alumne al curs següent hem usat el *k-nearest neighbor* (secció 4.2.3). D'aquesta manera, donat un alumne, busca els k alumnes més similars a ell i retorna el clúster al qual estaven assignats la majoria d'aquests k alumnes.

5.3.1 Elecció del valor k del k-NN

Per a elegir el valor de k per a usar al mètode k-NN no s'ha pogut usar sempre la mateixa k, ja que el nombre d'alumnes totals varia molt depenent del grau que estiguem analitzant. Per tant s'ha hagut de buscar una manera per a calcular la k.

El que fem és usar la tècnica de "Validació creuada de K iteracions" (o *K-fold cross-validation*).

El que fa la validació creuada és dividir en dos conjunts complementaris els alumnes que disposem. Aquests alumnes tenen la "solució"de la nostra predicció, és a dir, sabem a quin clúster han estat assignats. Un conjunt seran les dades d'entrenament (*training set*) i l'altre les dades de prova (*test set*). D'aquesta manera l'algorisme k-NN fa la fase d'entrenament amb només les dades del training set, i per a veure l'error amb què prediem ho fa amb les dades del test set, ja que no les hem usat per a la fase d'entrenament. Per a què l'error que ens retorni no depengui

tant d'en quins conjunts s'han separat les dades (dades d'entrenament i de prova) el que es fa és realitzar aquest mateix procés diverses vegades, però separant les dades de diferents maneres. Finalment es realitza una mitjana dels errors o encerts obtinguts al llarg de les proves, i obtenim s'obté l'error del k-NN per a aquell valor de la K.

El que es fa és dividir la mostra en k o m subconjunts (direm m per a no confondre la k del k-fold, amb la k del k-NN). D'aquesta manera es realitzaran m iteracions, on cada vegada un dels m subconjunts s'usarà com a dades de prova i els altres $m-1$ subconjunts seran les dades d'entrenament. A cada iteració s'obtindrà l'error de les dades de prova, i finalment es realitzarà la mitjana aritmètica dels resultats de cada iteració per a obtenir un únic resultat (fórmula 8).

En concret, en usar la "Validació creuada de K iteracions", el que es fa és dividir la mostra en k subconjunts, o m (direm m per a no confondre la k del k-fold, amb la k del k-NN). D'aquesta manera es realitzaran m iteracions, on cada vegada un dels m subconjunts s'usarà com a dades de prova i els altres $m-1$ subconjunts seran les dades d'entrenament. A cada iteració s'obtindrà l'error de les dades de prova, i finalment es realitzarà la mitjana aritmètica dels resultats de cada iteració per a obtenir un únic resultat (fórmula 8).

$$E = \frac{1}{m}\sum_{i=1}^{m} E_i \tag{8}$$

Per tant el que fem és provar diferents valors de k (des de k=1 fins a k igual al 20% de la mostra total) i aplicar la "Validació creuada de K iteracions". D'aquesta manera s'obté un error per a cada k. Finalment la k que es tria és la que té un error més petit.

5.4 Avaluació del predictor

Per a veure la precisió del predictor el que fem és usar la tècnica de "Validació creuada deixant-ne un fora" (*Leave-one-out cross-validation, LOOCV*). El que es fa és que en cada iteració només tinguem una mostra per a les dades de prova i tota

la resta de dades formen les dades d'entrenament. D'aquesta manera es realitza aquesta iteració per a totes les n dades que es disposen i s'obté un error (E_i) en cada iteració. Es realitzaran per tant n iteracions. Finalment l'error s'obté usant la mitjana aritmètica dels errors (fórmula 8). Aquest és el resultat que es mostrarà com error del predictor.

6 Experiments i resultats

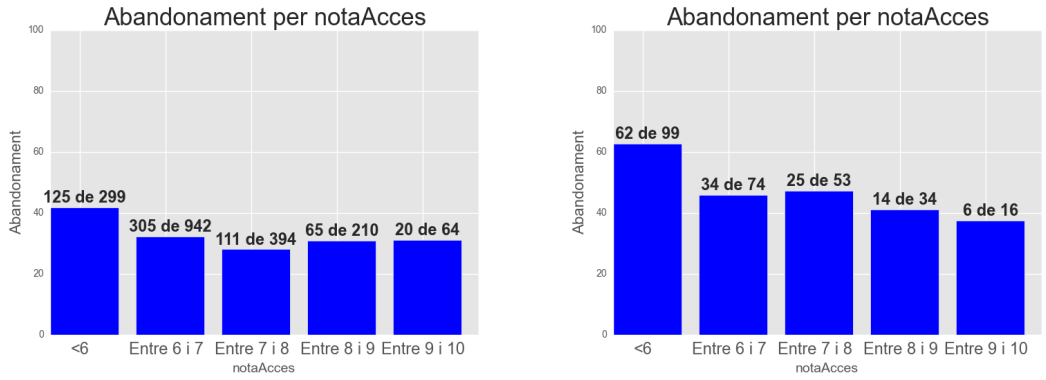
En aquest capítol explicarem les proves que hem realitzat i mostrarem els resultats que hem obtingut. Els resultats gràfics que mostrarem al llarg dels experiments seran del grau de Dret i d'Enginyeria Informàtica, tot i que a les taules que es mostraran hi hauran els resultats dels experiments realitzats als set graus. Els gràfics dels altres graus es poden veure a l'annex B.

6.1 Abandonament dels alumnes

Ens vam plantejar una pregunta abans de començar a fer l'anàlisi de les dades. Ens preguntàvem si era possible veure amb només la informació que teníem de l'alumne abans d'accedir a la carrera, quins alumnes tindrien més probabilitats d'abandonar. Les dades que podem usar per a fer aquesta comprovació són la nota d'accés i la via d'accés. És per això que en els gràfics d'aquesta secció hem mostrat en columnes el percentatge d'abandonament al llarg de la carrera depenent d'aquestes dues variables (per a veure els gràfics de tots els graus, anar a l'annex B.1). Es va mirar d'usar també la variable del sexe, però no hi havia diferències d'abandonament.

Abandonament depenent de la nota d'accés

En netejar les dades vam passar totes les notes d'accés a la mateixa escala, del 0 al 10. D'aquesta manera podem veure quanta gent abandona depenent de la nota d'accés tenint a tots els alumnes en la mateixa escala. Sabem que la nota de PAU sempre serà superior a 5, per tant hem classificat als alumnes en 5 grups diferents, depenent de si la nota és entre el 5 i el 6, el 6 i el 7, i així successivament. Finalment hem mirat quins d'aquests alumnes abandonen al llarg de la carrera i quins no.



(a) Dret

(b) Enginyeria Informàtica

Figura 6: Abandonament dels alumnes depenent de la nota d'accés al grau. L'eix d'ordenades mostra el percentatge d'abandonament i l'eix d'abscisses mostra la nota d'accés al grau.

En les columnes es veu quin percentatge d'alumnes abandona el grau sobre el total que hi ha en aquell grup. El nombre total no es tracta de tots els alumnes que estan en aquell grup, sinó que es tracta de tots els alumnes que podem saber amb certesa si han abandonat o no. És a dir, un alumne que encara està cursant la carrera i no ha acabat ni abandonat, no es trobarà aquí, ja que no sabem si ho deixarà més endavant

Es veu que els alumnes amb nota d'accés entre el 5 i el 6, tenen més tendència a abandonar que la resta de grups.

	Dret (%)	**Info** (%)	**Mates** (%)	**Cienc** (%)	**Crimi** (%)	**Gap** (%)	**Relab** (%)
nota $\in$ **(5,6]**	41.81	62.63	83.93	46.50	43.94	59.35	33.83
nota $\in$ **(6,7]**	32.38	45.95	67.01	41.05	28.16	63.41	57.53
nota $\in$ **(7,8]**	28.17	47.17	64.18	35.21	30.00	48.94	57.71
nota $\in$ **(8,9]**	30.95	41.18	36.67	11.54	38.55	52.63	34.29
nota $\in$ **(9,10]**	31.25	37.5	47.37	40.00	33.33	50.00	14.29

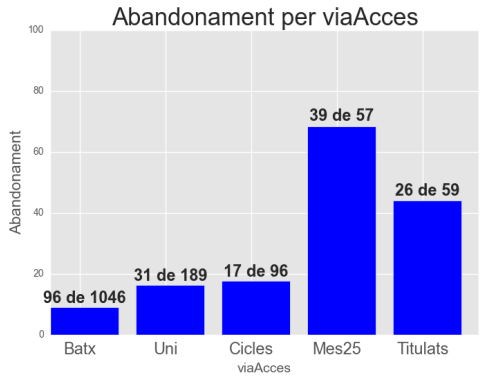
Taula 2: Percentatge d'abandonament al llarg de la carrera depenent de la nota d'accés a la universitat i el grau.

En aquesta taula 2 podem veure l'abandonament depenent de la nota d'accés de tots els graus (per a veure les abreviatures usades per als graus, anar a l'annex A.1). Veiem que en general sempre es manté que els alumnes que més abandonen són els alumnes amb nota d'accés baixa. Cal destacar també que els alumnes de Matemàtiques abandonen bastant més que la resta. Per tant els tutors de matemàtiques han de tenir l'abandonament més present que els altres tutors, tot i que és complicat que puguin prendre alguna acció al respecte.

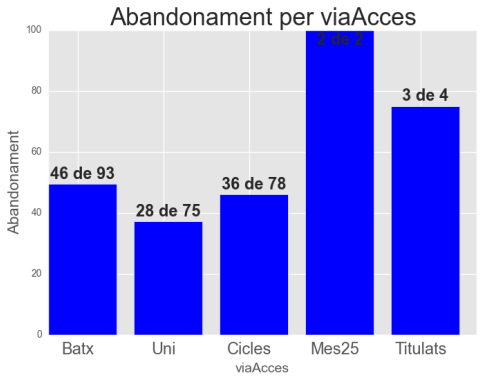
Abandonament depenent de la via d'accés

En aquest cas hem realitzat el mateix que en l'apartat anterior però amb la via d'accés [22]. Les vies d'accés que hi ha són les que es mostren a continuació, tot i que per facilitar la ràpida comprensió del gràfic i la taula hem usat uns diminutius per a facilitar l'enteniment de les vies d'accés.

- **Via 0:** Estudiants provinents de les PAU. Per facilitar la visualització de les taules i gràfics que hi ha més endavant a aquesta via l'anomenarem "Batx" (en ser gairebé tots els alumnes provinents de batxillerat).
- **Via 2:** Titulats superiors, diplomats, tècnics de grau mitjà, etc. A aquesta via l'anomenarem "Titulats".
- **Via 4 i 8:** Estudiants provinents de COU o d'FP de segon grau o mòduls professionals 3 o cicles formatius de grau superior i assimilats. A aquesta via l'anomenarem "Cicles".
- **Via 7:** Estudiants que canvien d'ensenyament o estudiants del mateix ensenyament que vénen a la UB. A aquesta via l'anomenarem "Uni".
- **Via 9:** Estudiants provinents de la prova d'accés per a més grans de 25 anys. A aquesta via l'anomenarem "Mes25".



(a) Dret



(b) Enginyeria Informàtica

Figura 7: Abandonament dels alumnes depenent de la via d'accés al grau.

Veiem a la figura 7 que els alumnes de la columna "Mes25" solen abandonar més que la resta. En el cas d'Enginyeria Informàtica, tot i tractar-se d'un percentatge molt elevat, al no haver-hi gairebé alumnes d'aquella via no es pot treure cap conclusió rellevant. Sí que es pot observar-ho en el cas de dret, que és un percentatge molt més elevat que en la resta de vies. Els "Titulats" tenen també més tendència a abandonar que la resta.

	Dret (%)	**Info** (%)	**Mates** (%)	**Cienc** (%)	**Crimi** (%)	**Gap** (%)	**Relab** (%)
Batx	9.18	49.46	65.49	36.36	24.90	5.21	30.35
Uni	16.40	37.33	55.74	36.11	22.06	3.70	31.39
Cicles	17.71	46.15	90.00	36.67	38.37	6.67	21.53
Mes25	68.42	100	75.00	56.25	50.00	44.44	35.14
Titulats	44.07	75.00	70.00	70.00	64.70	84.78	96.08

Taula 3: Percentatge d'abandonament de cada grau al llarg de la carrera depenent de la via d'accés a la universitat.

A la taula 3 es pot veure com en general els alumnes amb les vies d'accés "Mes25" i "Titulats", abandonen més que la resta. En el cas del grau de Matemàtiques, els alumnes de "Cicle" són els que més abandonen, tot i tractar-se de pocs alumnes.

Veiem com a Enginyeria Informàtica, i sobretot a Matemàtiques, indiferentment de la via d'accés que es tracti, l'abandonament és molt alt.

6.2 Agrupament dels alumnes en perfils similars a la carrera

Volem agrupar els cursos en diferents grups, on cada grup contingui a uns alumnes amb el mateix perfil de notes al llarg de la carrera. D'aquesta manera, sabent en quin curs es troba un alumne podrem tenir una idea del perfil d'aquest alumne.

Per a poder crear aquests grups hem usat l'algorisme K-Means (capítol 4.2.2). Vam voler usar també l'algorisme Mean Shift, havent reduït prèviament les dades a dues dimensions amb el mètode PCA (secció 4.2.4), i tot i haver realitzat bastantes proves amb ell, al obtenir un nombre de particions que no ens era convenient per a l'anàlisi dels perfils (veure taula 4), vam decidir usar el mètode K-Means.

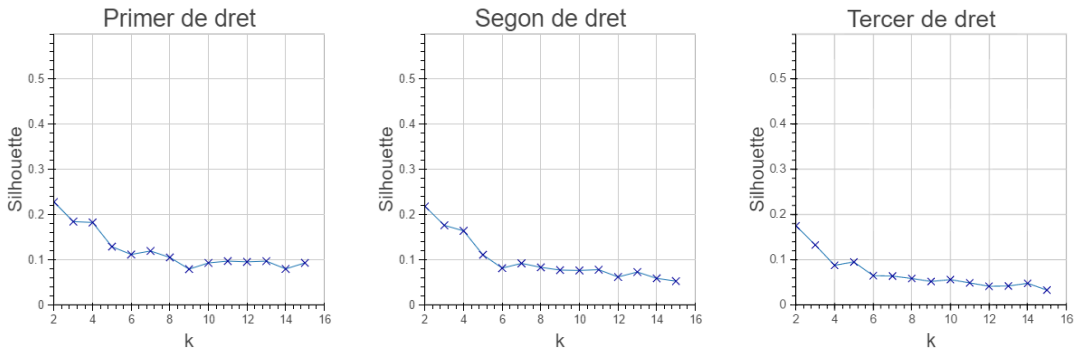
	Dret	**Info**	**Mates**	**Cienc**	**Crimi**	**Gap**	**Relab**
Primer	2	2	2	4	5	2	5
Segon	3	3	2	2	6	4	4
Tercer	3	2	3	3	1	4	4

Taula 4: Mostra el nombre d'agrupacions que realitzava l'algorisme Mean Shift per a cada grau i cada curs.

Veiem que en molts casos, l'algorisme Mean Shift, ens agrupava els alumnes en dos clústers, i que per a poder obtenir suficient informació dels alumnes ens quedarien perfils massa generalitzats.

6.2.1 Nombre d'agrupacions

Al optar per usar el mètode K-Means s'ha de passar la variable k com a paràmetre. Per a fer-ho hem usat el coeficient de Silhouette com s'ha explicat al capítol 5.2.



(a) Primer de Dret (b) Segon de Dret (c) Tercer de Dret

Figura 8: Coeficient de Silhouette aplicat a diferents k per veure quina obté millor resultat per a usar en el mètode k-Means. A l'eix d'ordenades es veu el valor que retorna el coeficient de Silhouette per a les k's de l'eix de les abscisses.

Veiem a les imatges que els valors més alts són sempre per a k igual a 2, però si només ho separem en dos clústers, obtindríem massa poca informació de cada un d'ells. És per això que busquem un altre valor de k. Com més alt sigui el valor de k que triem, en tenir els alumnes separats en més clústers, més informació concreta es podrà donar de cada un d'ells. A primer de Dret veiem que el coeficient de Silhouette per a k igual a 3 o a 4 és pràcticament igual, i que si ho separem en 4 clústers i no en 3 obtindrem més informació del curs. Per això separarem primer curs en 4 clústers. Segon i tercer curs ho separem en 3 clústers.

A l'haver realitzat el mateix procés per als altres graus i haver obtingut gairebé en tots els casos les mateixes k (4 clústers a primer, 3 a segon i 3 a tercer) s'ha decidit usar sempre aquest nombre d'agrupacions. En cas de que haguéssim trobat casos molt diferents dels triats, no haguéssim decidit usar sempre les mateixes k.

6.2.2 Informació dels clústers

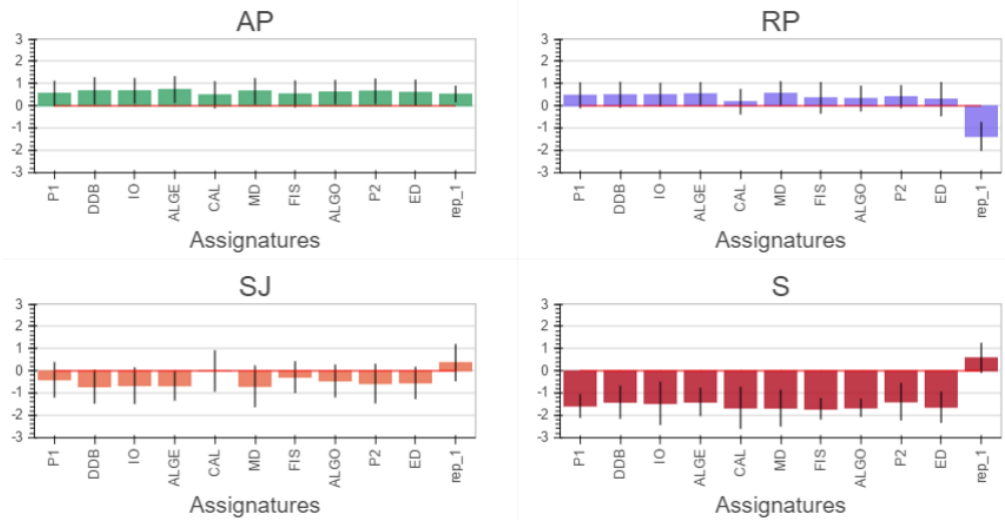
Un cop ja hem usat el mètode k-Means per a agrupar els alumnes es vol veure el perfil d'alumnes que pertanyen a cada un d'ells.

Per a poder veure en detall el perfil d'alumnes de cada clúster hem fet un diagrama de columnes que mostra la mitjana dels alumnes d'aquell clúster en cada

assignatura. Recordem que les notes estan estandarditzades i que per tant un valor significa que estan per sobre de la mitjana dels alumnes en aquella assignatura, un valor neutre significa que han tret la nota mitjana i un valor negatiu és que estan per sota. Es mostra també si un alumne repeteix molt o no en un curs (columnes *rep_1, rep_2 i rep_3*). Un valor alt és que no repeteix, i valors baixos és que sí.

Comentarem els resultats que hem obtingut curs per curs de Dret i d'Enginyeria Informàtica. Per veure els gràfics dels altres graus es pot anar a l'annex B.2.

(a) Primer de Dret

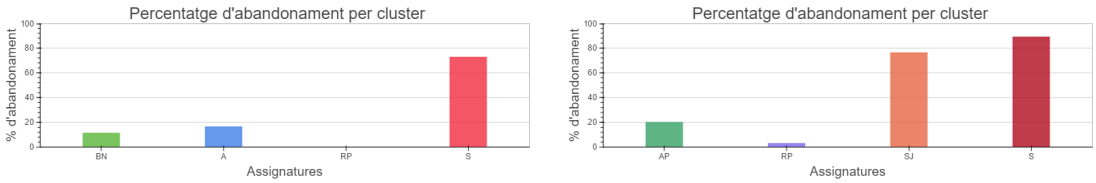


(b) Primer d'Enginyeria Informàtica

Figura 9: Diagrama de columnes de primer curs on es veu per a cada clúster la mitjana de les notes de les assignatures estandarditzades i la variable *rep_1*.

Veiem a la figura 9 que a més de mostrar cada columna la mitjana d'aquell clúster en aquella assignatura mostra també una línia negra vertical en cada columna que mostra la desviació estàndard de cada assignatura. Hi ha també una línia vermella horitzontal en l'eix d'abscisses per a facilitar la visualització de les assignatures que es troben per sobre o per sota de la mitjana. El nom complet de les assignatures es pot veure a l'annex A.2.

Abans de comentar els resultats que observem, mostrem també l'abandonament que es produeix de primer a segon curs de cada clúster.



(a) Primer de Dret

(b) Primer d'Enginyeria Informàtica

Figura 10: Diagrama de columnes que mostra l'abandonament a primer curs de cada clúster.

A primer curs obtenim un perfil d'alumnes per a cada clúster diferent entre Dret i Enginyeria Informàtica. És per això que a primer curs comentarem els dos graus per separat. El grau de Matemàtiques tindria els mateixos clústers que Informàtica, i la resta seguirien els perfils de Dret. Per veure els diagrames de columnes dels altres graus anar a l'annex B.3.
El grau de Dret, que es pot veure al gràfic 9a, se separa en els 4 següents clústers:

- **Clúster "BN"**: És el clúster on es troben les Bones Notes (BN). Veiem que aquests alumnes treuen les millors notes en totes les assignatures. A més, a la columna rep_1 es veu que no repeteixen gairebé mai assignatures.
- **Clúster "A":** És el clúster d'Aprovats ("A"). Els alumnes no treuen les millors notes, sinó unes notes tendint a la mitjana de l'assignatura. No repeteixen gairebé mai assignatures.
- **Clúster "RP":** És el clúster de Repetidors ("RP"). Tot i tenir notes bastant

semblants al clúster anterior es diferencien sobretot per repetir assignatures.

- **Clúster "S"**: És el clúster de suspesos ("S"). Els alumnes d'aquests clústers són els alumnes amb pitjors notes que ho suspenen tot. Veiem que tot i això no repeteixen assignatures. El fet de tenir notes tan baixes i no repetir mai assignatures s'explica en el gràfic d'abandonament (figura 10), ja que es veu com pràcticament tots els alumnes del clúster abandonen, i per tant no repeteixen les assignatures.

En el cas d'Enginyeria Informàtica veiem que el perfil d'alumnes de cada clúster és diferent de l'anterior.

- **Clúster "AP"**: És el clúster on es troben els Aprovats ("AP") a la primera, amb molt bones notes o no. Veiem que aquests alumnes treuen les millors notes en totes les assignatures. A més, a la columna rep_1 es veu que no repeteixen gairebé mai assignatures.

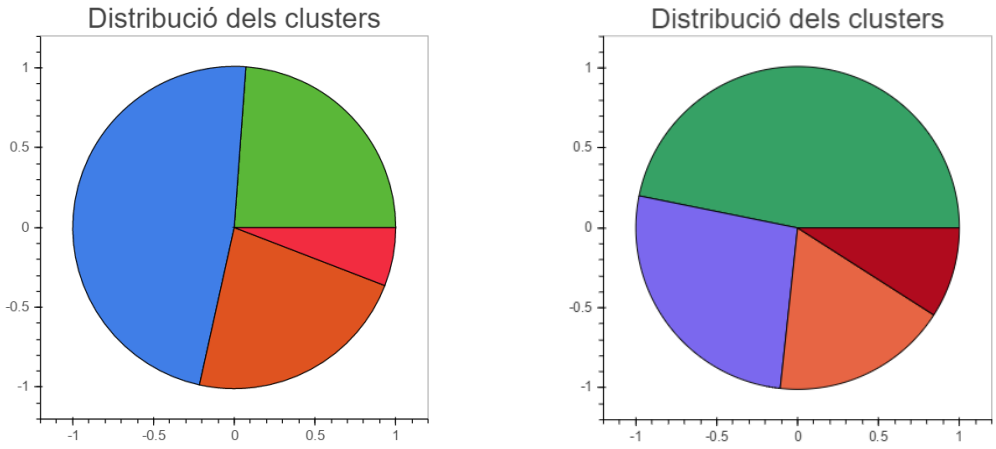
- **Clúster "RP":** És el clúster de Repetidors ("RP"). Veiem que en l'assignatura que treuen notes menys per sobre de la mitjana és en *Càlcul*.

- **Clúster "SJ":** És el clúster on es troben els alumnes que ho suspenen tot però amb notes més justes que el clúster següent. Veiem que el fet de no repetir assignatures es deu al mateix fet d'abans de l'abandonament dels alumnes.

- **Clúster "S"**: És el clúster de suspesos ("S"). Els alumnes d'aquest clúster són els alumnes amb pitjors notes que ho suspenen tot. Tornen a tenir la columna de *rep_1* alta, i per tant no repeteixen i marxen. Veiem que el tutor d'estudis no pot actuar sobre aquests alumnes, ja que en tenir notes tan baixes no hi poden fer res.

S'ha creat un diagrama de sectors per a veure com estan repartits els grups (figura 11). Aquest divideix en sectors els diferents clústers que tenim, on la longitud de l'arc de cada sector és proporcional a la quantitat d'alumnes que pertanyen a cada clúster. Cada sector manté els colors fets servir prèviament i si es passa el cursor

per sobre, apareix el nom del clúster, la nota mitjana estandarditzada que tenen els alumnes d'aquell clúster i el percentatge d'abandonament.



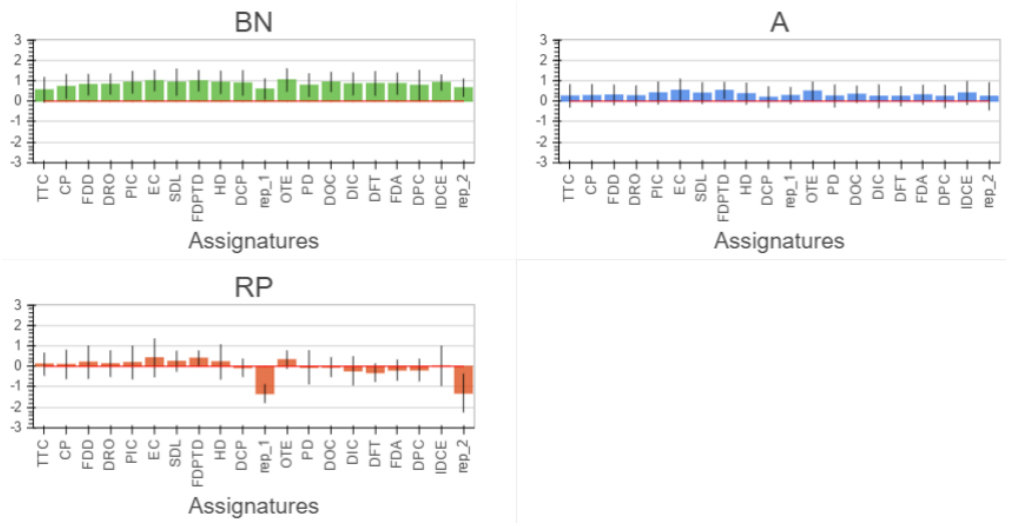
(a) Primer de Dret (b) Primer d'Enginyeria Informàtica

Figura 11: Diagrama de sectors que mostra com estan repartits els clústers de primer.

En el cas de Dret, hem vist que l'abandonament es dóna pràcticament només en el clúster "S" (vermell). En el diagrama de sectors, aquest clúster es tracta d'un percentatge molt petit de gent, i per tant l'abandonament a Dret no hauria de ser un problema. En el cas d'Enginyeria Informàtica l'abandonament es trobava en el clúster "SJ" i "S". (taronja i vermell), i per tant veiem que hi ha molt més abandonament que a Dret i s'hauria d'anar amb compte.

A segon curs hem realitzat els mateixos gràfics, però agrupant als alumnes en 3 clústers.

A la figura 12 es veu com a segon curs ja no ens trobem amb el perfil de gent que ho suspèn tot, com era l'últim clúster que teníem a primer. Això es deu a què aquests estudiants ja han deixat la carrera i no cursen segon.

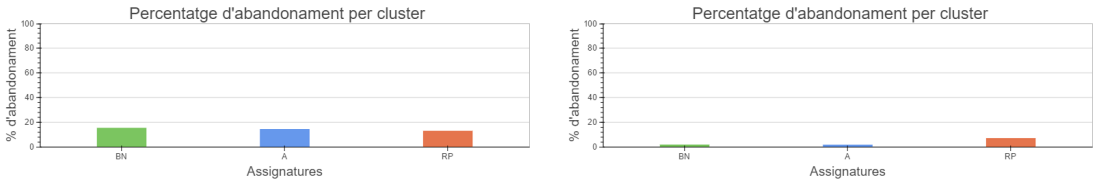


(a) Segon de Dret

(b) Segon d'Enginyeria Informàtica

Figura 12: Diagrama de columnes de segon curs on es veu per a cada clúster la mitjana estandarditzada de les notes de les assignatures i les variables *rep_1* i *rep_2*.

Es pot veure a la figura 13 que l'abandonament a segon és molt poc, i que per tant, el tutor ja no ha de vigilar tant l'abandonament en un alumne.



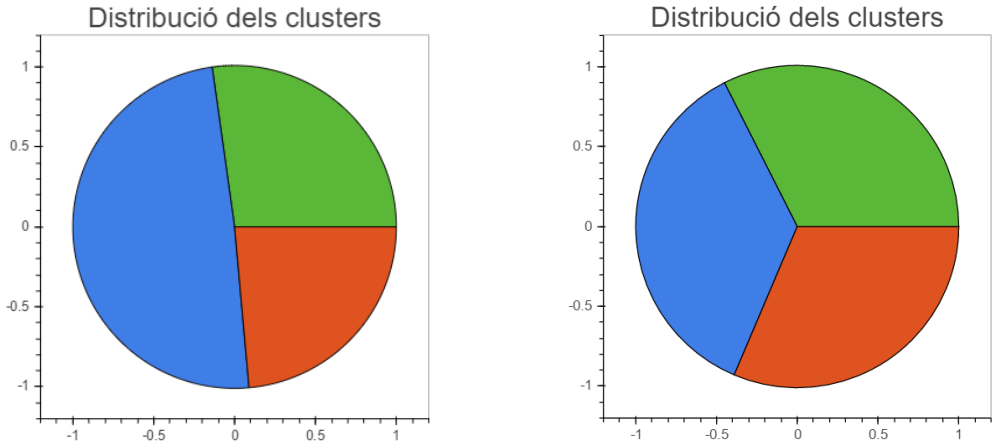
(a) Segon de Dret (b) Segon d'Enginyeria Informàtica

Figura 13: Diagrama de columnes que mostra l'abandonament a segon curs de cada clúster.

Veiem també que en aquest cas els clústers de Dret i Enginyeria Informàtica tenen els mateixos perfils d'alumnes per clúster. Per tant reben els mateixos noms els clústers i els comentarem conjuntament.

Els clústers que obtenim es podrien explicar com:

- **Clúster "BN"**: Alumnes amb millors notes que mai repeteixen. Veiem que aquest clúster és molt semblant al qual tenim a primer. Més endavant veurem si es tracta dels mateixos alumnes que es trobaven en el clúster de millors alumnes de primer curs.
- **Clúster "A"** (Aprovats): Alumnes amb notes una mica per sobre de la mitjana, que repeteixen poques assignatures, i que per tant suspenen poc.
- **Clúster "RP"** (Repetidors): Alumnes amb les pitjors notes, que destaquen sobretot per ser els que més assignatures repeteixen.



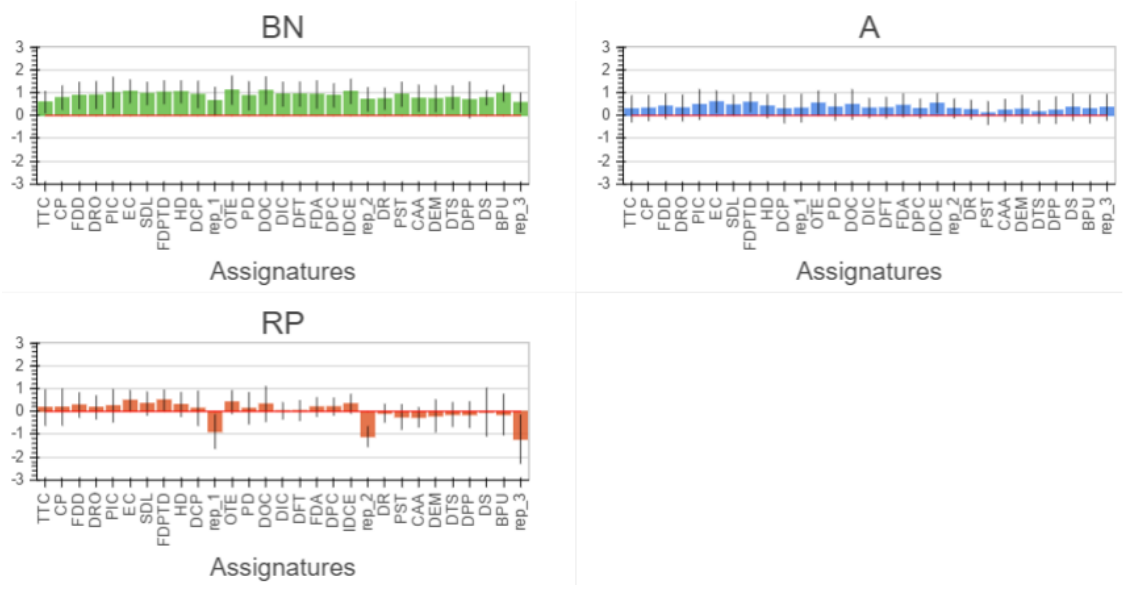
(a) Segon de Dret (b) Segon d'Enginyeria Informàtica

Figura 14: Diagrama de sectors que mostra com estan repartits els clústers de segon curs.

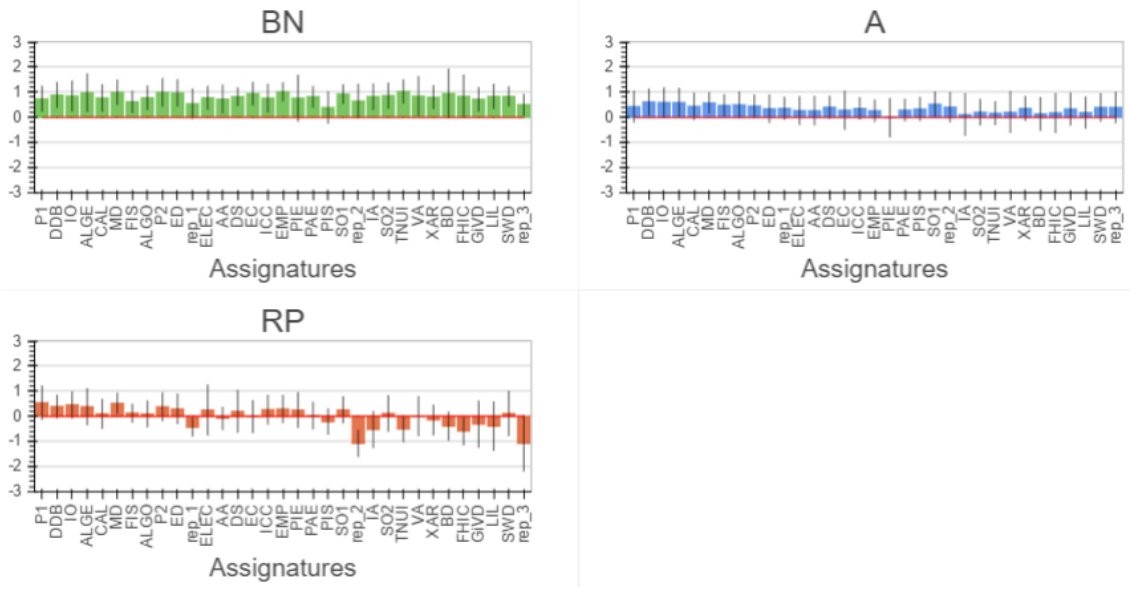
Veiem que els clústers estan ben repartits. En el cas de Dret hi ha més alumnes del clúster "A", que del clúster "BN" o el clúster "RP". En el cas d'Enginyeria Informàtica, en canvi, estan més repartits. Veiem que el tutor d'estudis hauria d'anar més amb compte amb els alumnes que es troben en el clúster "RP", al ser els que més repeteixen.

Tercer curs és l'últim curs que hem tingut en compte. Quart curs no l'hem analitzat, ja que hi ha moltes assignatures optatives i per tant els alumnes fan assignatures molt diferents. Hem vist prèviament que el nombre de clústers en què separarem tercer curs és 3.

Es veu que a tercer curs els clústers que obtenim segueixen sent els mateixos que havíem obtingut a segon curs. Resumint-ho serien els tres clústers: "BN" (bones notes), "A" (aprovats) i "RP" (repetidors).

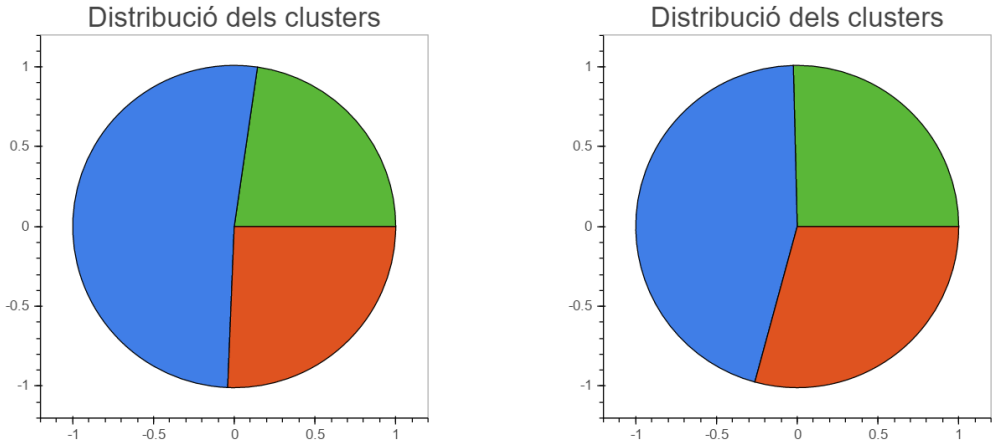


(a) Tercer de Dret



(b) Tercer d'Enginyeria Informàtica

Figura 15: Diagrama de columnes de tercer curs on es veu per a cada clúster la mitjana estandarditzada de les notes de les assignatures i les variables *rep_1*, *rep_2* i *rep_3*.



(a) Tercer de Dret (b) Tercer d'Enginyeria Informàtica

Figura 16: Diagrama de sectors que mostra com estan repartits els clústers de tercer curs.

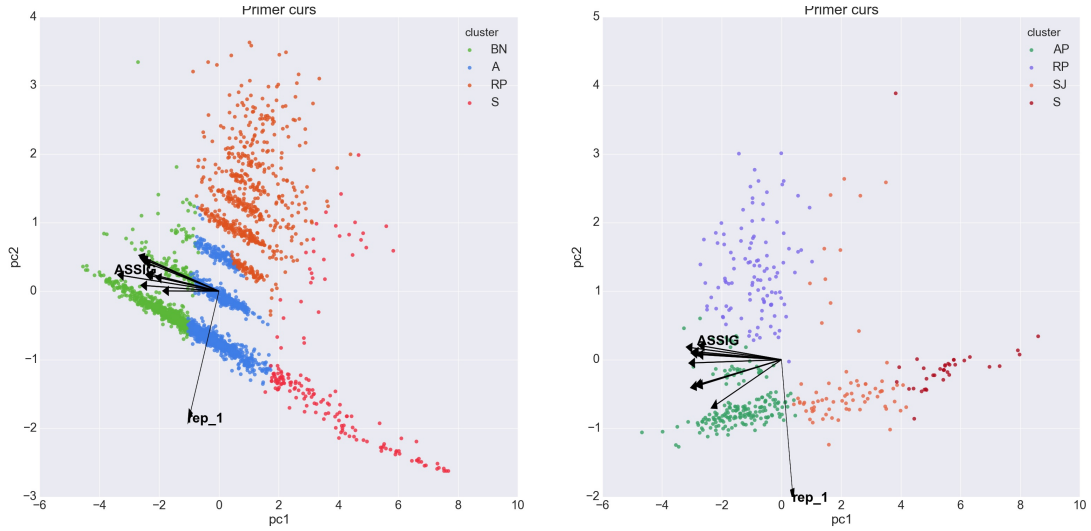
Veiem que el diagrama de sectors és pràcticament igual que a segon curs. Per això voldrem veure si es tracta sempre dels mateixos alumnes els que es troben en cada clúster. Ho veurem al capítol 6.4

Per tant, al llarg d'aquesta secció hem pogut crear uns clústers, veient el perfil d'alumnes que pertanyen a cada un d'ells.

6.3 Visualització general dels cursos

Per a fer una visualització general de com s'han separat els alumnes en els diferents clústers hem volgut mostrar en un mapa de punts a cada alumne amb el color del seu respectiu grup. En tenir cada estudiant un vector amb diverses variables, però voler mostrar els alumnes en un mapa de punts en 2D, hem redimensionat la taula d'alumnes de cada curs a només dues dimensions per alumne. Això ho hem fet amb el mètode d'Anàlisi de Components Principals (vist al capítol 4.2.4).

Ho mostrem només per a primer curs. Per a veure els altres cursos, i els altres graus, es pot veure a l'annex B.4.



(a) Primer de Dret (b) Primer d'Enginyeria Informàtica

Figura 17: Visualització en 2D dels alumnes de primer curs separat en clústers amb el mètode PCA i K-Means.

Veiem en els gràfics, que al mostrar els alumnes en dues dimensions, els punts formen unes "línies" gairebé horitzontals (es pot apreciar bé en el cas de Dret, en haver-hi més alumnes). Per a veure quina era la causa es va decidir mostrar els "eixos" que havia creat el mètode PCA. Veiem que les línies es creen a causa del vector *rep_1*, al ser amb diferència el que té variància més gran. La resta d'assignatures es mostren totes agrupades cap a la mateixa direcció.

Per tant, mirant el gràfic de Dret es pot veure com els alumnes que més repeteixen són els del clúster "RP" (color taronja) en trobar-se en el sentit contrari al qual apunta el vector *rep_1*, mentre que del clúster "S" (vermell) no repeteix pràcticament ningú. Havíem vist que això es devia a què abandonaven directament. Veiem també que els del clúster "BN" (verd) es troben més a l'esquerra de la imatge, és a dir, cap on apunten els vectors de les assignatures, i que per tant es veu que són els que tenen més bones notes. Veiem també que hi ha pocs alumnes d'aquest clúster que repeteixin. La interpretació del significat de les diagonals i la seva amplada ve donada per la diferència entre variàncies de la primera i la segona component creades per el PCA. Veiem que *rep_1* té molta més variància en la segona component que en la primera, fet que genera les diagonals que es poden veure en el gràfic. En

canvi els vectors de les assignatures varien molt més en la primera component que en la segona, determinant així l'amplada d'aquestes diagonals.

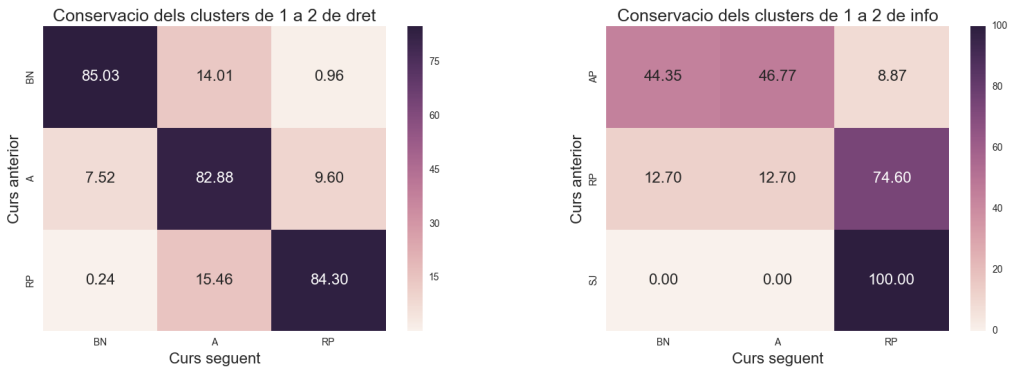
En el cas d'Enginyeria Informàtica passa el mateix, però es veu que els alumnes que repeteixen són els del clúster "RP" (blau), ja que havíem vist que en aquest cas els alumnes del clúster "SJ" i "S" abandonaven gairebé tots. Per això es veu que no repeteixen assignatures, però que es troben a la dreta de la imatge (en sentit contrari als vectors de les assignatures).

Per tant, amb una bona interpretació d'aquest gràfic es pot veure ràpidament el perfil d'alumnes de cada clúster.

6.4 Conservació dels clústers

Al parlar de conservació dels clústers ens referim a veure si els alumnes es mantenen al mateix clúster al llarg dels cursos. És a dir, si un alumne en passar els cursos se segueix mantenint en el grup d'alumnes amb el mateix perfil d'estudiants o va a parar a grups on es troba gent amb diferent perfil que al curs anterior. Per a poder-ho veure hem usat un mapa de calor (*heatmap* [23]. Les files mostren a quin curs es troba l'alumne a un curs donat, i les columnes mostren on es troba al curs següent. Si un quadrat es troba a la posició i, j (on i és la fila i j és la columna), vol dir que l'alumne que es trobava al clúster i, al curs següent ha anat a parar al clúster j. Si la diagonal de la matriu està molt marcada (colors molt foscos) vol dir que els alumnes tenen la tendència a mantenir-se al mateix clúster que es trobaven.

Hem creat dos *heatmaps* diferents per a cada grau. Un per a veure la conservació dels clústers de primer a segon curs, i l'altre de segon a tercer. Per a veure les imatges dels altres graus, anar a l'annex B.5.



(a) Primer a segon de Dret (b) Primer a segon d'Enginyeria Informàtica

Figura 18: Mapa de calor on es veu la conservació dels clústers de primer a segon curs. Els valors de les files són percentatges. Sumant cada fila el 100%. Les files mostren el clúster en què es trobaven a primer curs i les columnes el clúster en què es troben a segon curs.

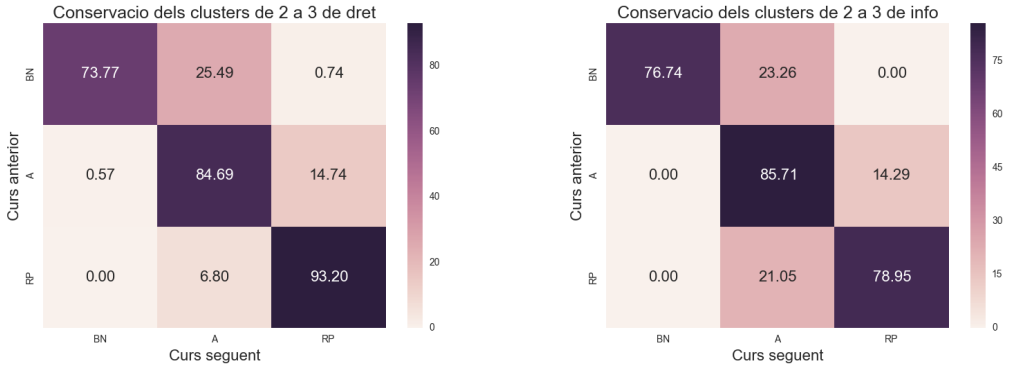
Veiem que tot i haver-hi quatre clústers a primer curs, només en mostrem els tres primers, ja que l'últim clúster, en abandonar gairebé tots els alumnes, no l'hem posat al gràfic.

El grau de Dret es veu ràpidament que té les diagonals molt marcades, i que per tant es pot veure que els alumnes es mantenen de primer curs a segon en els mateixos clústers, havent-hi poques excepcions.

Per altra banda, en el cas d'Enginyeria Informàtica, es veu que els alumnes es mantenen en el clúster en què es trobaven o baixen un clúster, és a dir, van a un perfil d'estudiants amb pitjors notes. Això es deu al fet que a primer curs pràcticament tots els alumnes dels últims dos clústers abandonaven i per tant els alumnes que no eren tan bons que es trobaven en el primer clúster passen al segon clúster, on es trobaran els alumnes "A" (aprovats, però sense destacar amb bones qualificacions). Per altra banda, els alumnes dels últims dos clústers que no van abandonar (són molt pocs), es mantenen tots en el clúster "RP" (repetidors).

A la imatge següent veiem la conservació dels clústers de segon a tercer curs. En aquest cas ja es veu com els dos graus tenen la diagonal molt marcada, i que, per tant, els alumnes d'Enginyeria Informàtica en general ja es mantenen en el mateix

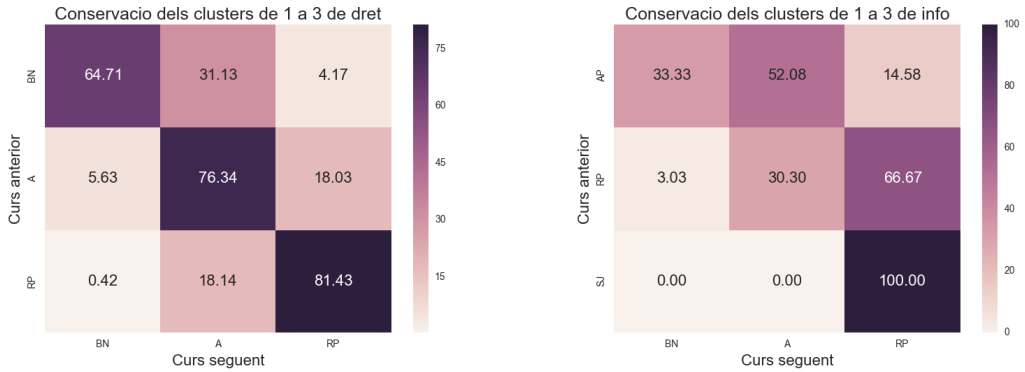
clúster que a segon, i no van al clúster amb un perfil de notes pitjor.



(a) Segon a tercer de Dret (b) Segon a tercer d'Enginyeria Informàtica

Figura 19: Mapa de calor on es veu la conservació dels clústers de segon a tercer curs. Les files mostren el clúster en què es trobaven a segon curs i les columnes el clúster en què es troben a tercer curs.

Per acabar el capítol de conservació dels clústers volem veure directament de primer a tercer curs. D'aquesta manera podem veure si al llarg dels cursos un alumne pot canviar molt acadèmicament, o no. Això voldria dir que un alumne que comença malament acabarà de la mateixa manera (en cas de continuar) i un alumne bo seguirà sent-ho al llarg de tota la carrera.



(a) Primer a tercer de Dret (b) Primer a tercer d'Enginyeria Informàtica

Figura 20: Mapa de calor on es veu la conservació dels clústers de primer a tercer curs. Els valors de les files són percentatges. Les files mostren el clúster en què es trobaven a primer curs i les columnes el clúster en què es troben a tercer curs.

Al veure la figura 20, podem veure que en general Dret, tot i poder haver-hi variacions, hi ha la tendència a romandre en el mateix clúster. En canvi, en el cas d'Enginyeria Informàtica els alumnes no canvien en cap cas cap a un clúster millor i sí que ho fan en molts casos cap a un clúster amb un perfil d'alumnes amb pitjors notes.

6.5 Predictor

Per a poder fer el seguiment d'un alumne, no només del que ha fet fins al moment, sinó el que farà, hem fet un predictor de clústers. D'aquesta manera si volem saber si un alumne concret tindrà moltes dificultats o no al curs següent, farem servir el predictor per veure a quin grup ens preveu que anirà a parar l'alumne. D'aquesta manera si l'alumne va a parar als dos primers clústers, l'alumne no hauria de tenir massa dificultat de cara al següent curs. En cas d'anar a parar al tercer clúster, s'hauria de mirar si reduir el nombre d'assignatures de l'alumne o ajudar amb material suplementari a l'alumne.

	Dret	**Info**	**Mates**	**Cienc**	**Crimi**	**Gap**	**Relab**
Primer a segon	38	9	30	7	19	29	108
Segon a tercer	31	10	11	8	38	10	44

Taula 5: k escollida per a usar en el mètode k-NN per a cada curs i cada grau.

Per a fer el predictor hem usat el classificador k-NN (vist al capítol 4.2.3). Com hem explicat, hem usat una k diferent per a cada grau i cada curs, ja que aquesta k depèn molt del nombre de dades que tinguem i altres factors de cada grau. En la taula 5 es veu la k que hem usat per a cada grau per a usar al mètode k-NN per a la predicció dels alumnes de primer a segon curs i de segon a tercer.

D'aquesta manera hem creat un mètode que si passem com a paràmetre l'identificador d'un alumne, ens retornarà el clúster en què prediem que anirà a parar al seu curs següent, juntament amb les mitjanes de les assignatures del clúster al qual predim que anirà. A continuació mostrem un exemple del que retornaria el predictor

en passar-li com a paràmetre l'identificador d'un alumne d'Enginyeria Informàtica que es troba a primer curs i cursarà segon.

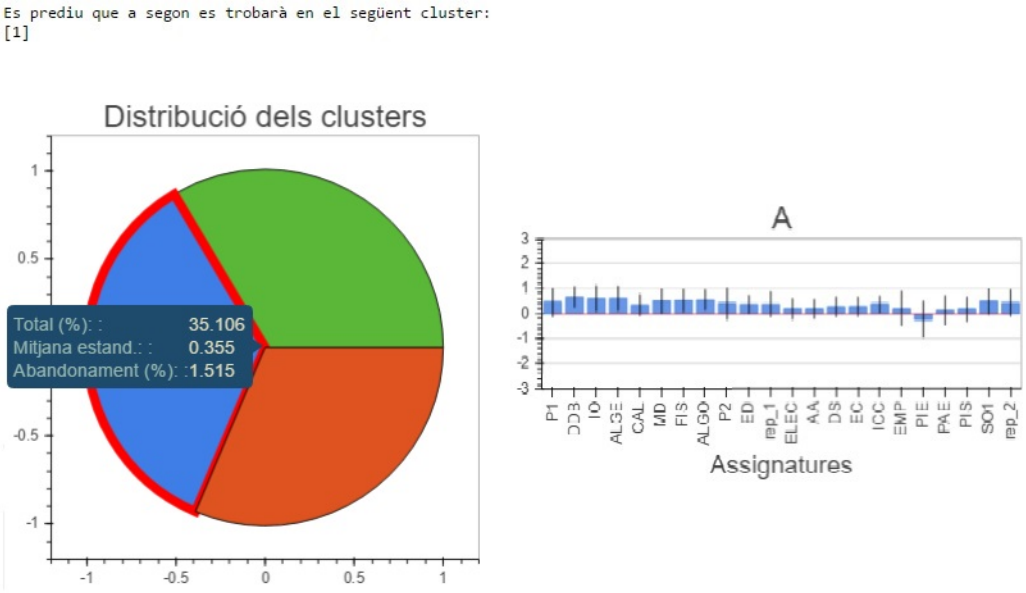


Figura 21: Exemple de predicció d'un alumne.

6.6 Precisió del Predictor

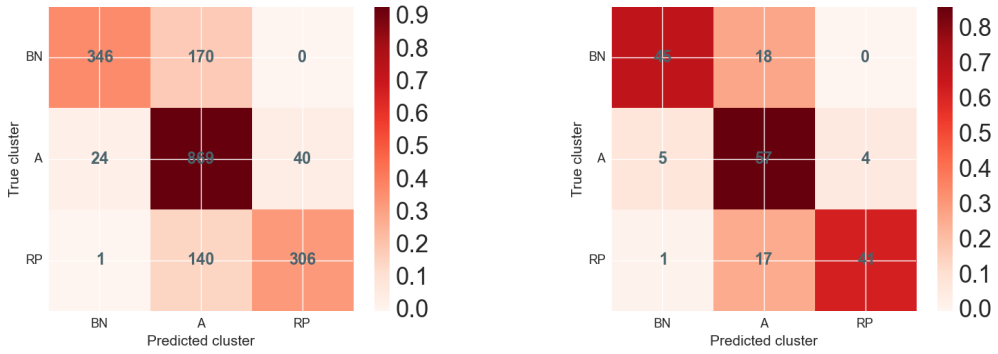
Per a obtenir la precisió de l'algorisme ho hem fet amb el *Leave-one-out cross-validation*, com hem explicat al capítol 5.4.

	Dret (%)	**Info** (%)	**Mates** (%)	**Cienc** (%)	**Crimi** (%)	**Gap** (%)	**Relab** (%)
Primer a segon	81.49	77.13	76.92	84.19	80.97	84.31	76.92
Segon a tercer	83.79	87.69	85.48	83.01	80.49	89.29	81.1

Taula 6: Precisió del predictor. Es pot veure per a cada grau, el percentatge d'encert del predictor d'un curs a l'altre.

La precisió que hem obtingut per a cada curs es pot veure a la taula 6. Es veu que el predictor té en tots els casos un encert superior al 75%, i no hi ha variacions molt grans entre els graus.

Per a veure on falla més el predictor, és a dir, quin clúster té més problemes per a predir, s'ha creat una matriu de confusió [24]. Això ens permet veure a quin clúster hem predit que aniran els alumnes i on han anat realment.



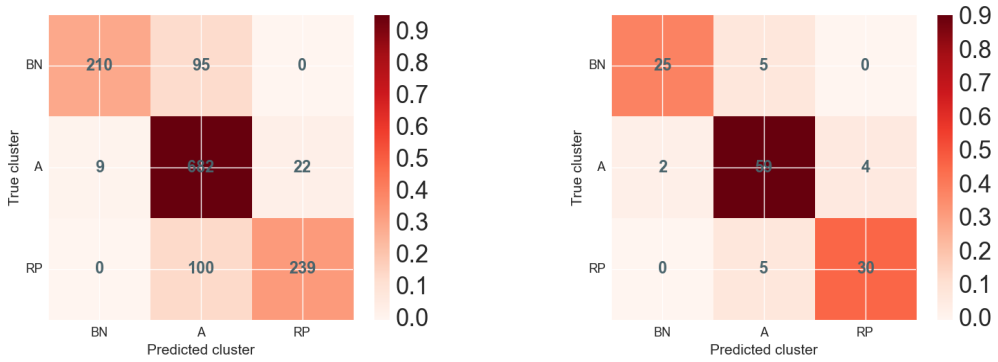
(a) Primer a segon de Dret (b) Primer a segon d'Enginyeria Informàtica

Figura 22: Matriu de confusió amb la precisió del predictor de primer a segon curs. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.

La diagonal principal de les matrius anteriors indica el nombre d'alumnes que han sigut predits correctament per a cada clúster. En canvi, els valors ubicats fora de la diagonal principal indiquen l'error d'assignació.

Aquesta matriu de confusió l'hem mostrada en un *heatmap*, com es pot veure a la figura 22 (la resta de graus es poden veure a l'annex B.6).

Veiem que l'error que comet el predictor amb més freqüència, tant de primer curs a segon com de segon curs a tercer, és el de tenir tendència a predir que els alumnes es trobaran al clúster d'"A" (segona columna), mentre que realment es troben al clúster "MB" (primera fila) o de "RP" (tercera fila). Això és a causa del fet que on més alumnes tenim és al clúster d'"A" (veure figura 14), per tant en cas que es tracti d'un alumne que no destaca massa cap a cap clúster (no és ni molt dolent ni molt bo) l'algorisme predirà que l'estudiant es trobarà al clúster d'"A" al ser on es troba la majoria de gent.



(a) Segon a tercer de Dret

(b) Segon a tercer d'Enginyeria Informàtica

Figura 23: Matriu de confusió amb la precisió del predictor de segon a tercer curs. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.

En el cas de segon a tercer curs (figura 23) veiem que passa el mateix. Les diagonals estan clarament marcades, i per tant el predictor encerta en la gran majoria de casos. En el cas de Dret els errors que es cometen tornen a ser els d'en alguns casos tenir la tendència a predir als alumnes en el clúster "A", en trobar-se la gran majoria d'alumnes. En el cas d'Enginyeria Informàtica, en trobar-se els alumnes més repartits entre els tres clústers, aquesta tendència desapareix i fa que el predictor tingui més precisió.

6.7 Aplicació

Per a orientar al tutor sobre un alumne en concret hem creat un mètode que l'ajudi. Aquest mètode ens permetrà obtenir informació de l'alumne fins al curs actual, i ens predirà informació futura. El mètode funciona passant-li com a variable d'entrada el nombre d'identificació de l'alumne i ens mostrarà dues parts diferents.

Una d'elles és què ha fet fins al curs actual l'alumne. Això ho farà mostrant-nos la seva nota mitjana estandarditzada de cada curs, és a dir, quan per sobre o per sota de la resta d'alumnes s'ha trobat l'alumne fins al moment. També ens mostrarà en quins clústers ha estat classificat l'alumne, i on es trobaria aquell alumne en un mapa de punts.

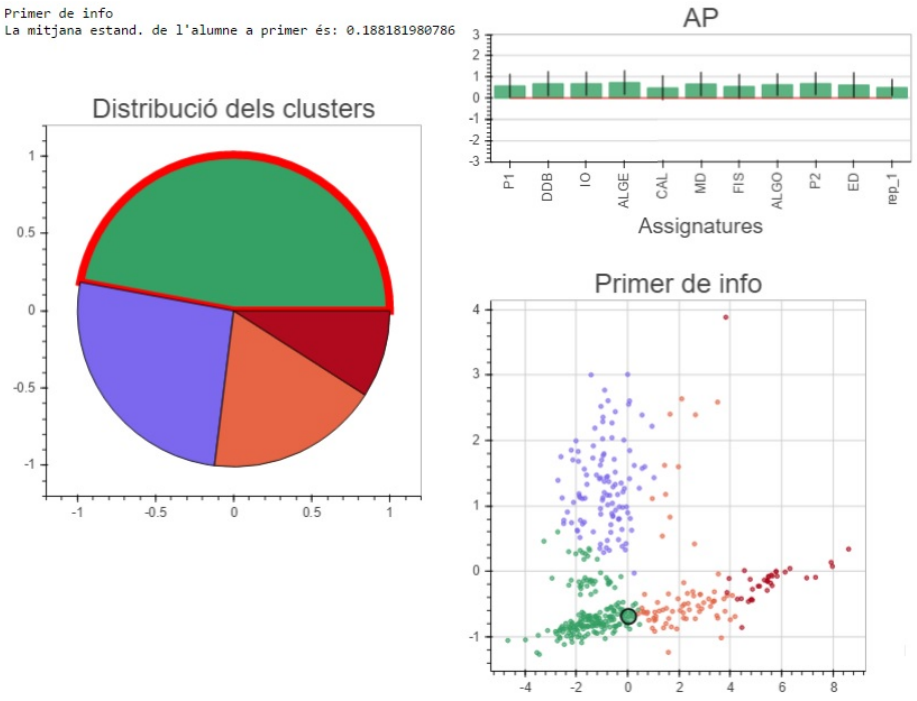


Figura 24: Exemple de la informació que mostrem d'un alumne fins al curs actual. En aquest cas es tacta d'un alumne que ha cursat primer curs.

Mostrem el mateix alumne que havíem mostrat a la figura 21, en mostrar el que es veia en usar el predictor. Veiem que aquest alumne en el mapa de punts es troba proper als alumnes d'un altre clúster, i hem vist que el predictor, ens predeia que aquest alumne aniria a un clúster amb perfil de notes pitjor.

L'altra part que ens mostra aquest mètode és la mateixa que mostra el predictor vist a la figura 21.

D'aquesta manera, el tutor d'estudis, mirant ràpidament aquests resultats pot saber si es tracta d'un alumne que ha tingut molts problemes o no al llarg de la carrera, i veu, amb un error possible, si tindrà dificultats al curs següent.

7 Conclusió i treballs futurs

En iniciar el projecte s'havien marcat quatre objectius a completar, amb la finalitat d'assolir així l'objectiu principal. Aquest era el de poder ajudar als tutors d'estudis a poder saber les dificultats que ha tingut cada alumne fins al moment i les que es preveu que tindrà en un futur.

1. *Obtenció i neteja de dades dels diferents graus de la Universitat de Barcelona:* En un inici es creia que s'hauria de realitzar aquesta preparació de les dades de només dos graus. En realitzar-ho finalment per a set graus diferents, li hem dedicat a aquest objectiu més temps de l'esperat. Considero que hem aconseguit tenir unes dades fiables, tractables i representatives per a realitzar una anàlisi còmode i correcte d'aquestes dades.

2. *Anàlisi i visualització de dades:* S'han realitzat diferents anàlisis de les dades obtenint resultats d'abandonament, on hem pogut veure les diferències d'abandonament que hi ha entre els graus i també com depenent de la via d'accés o la nota d'accés pot variar bastant la probabilitat d'abandonar. Hem realitzat també la clusterització dels alumnes, i ens hem adonat que és possible agrupar-los en clústers representatius de perfils acadèmics ben diferenciats. Hem vist també que aquests grups tendeixen a conservar-se al llarg de tota la carrera. A més, tots aquests resultats els hem pogut mostrar gràficament, i per tant considero que s'han assolit els objectius d'anàlisi i visualització de dades.

3. *Implementació de models predictius sobre el rendiment futur dels estudiants:* Hem realitzat un predictor per a què ens indiqui en quin clúster es trobarà l'alumne el curs següent. Avaluant la precisió d'aquest predictor hem obtingut uns bons resultats, sempre tenint en compte que no hem considerat factors externs que poden afectar l'alumne.

4. *Mostrar resultats útils per al tutor:* En l'apartat 6.7 s'ha desenvolupat una eina que mostra informació acadèmica d'un alumne que pot ser útil per a un tutor. Tot i això no s'ha creat un producte final que pugui ser utilitzat com a eina de

tutorització.

Veient els resultats del treball i els objectius que s'havien marcat, es pot concloure que s'ha assolit amb èxit l'objectiu principal del projecte.

En l'àmbit personal m'ha resultat molt interessant poder adquirir coneixements sobre l'aprenentatge automàtic i la manera de visualitzar les dades. He pogut aprendre molt sobre temes que no coneixia de l'anàlisi de dades i també sobre tecnologies que desconeixia i han resultat molt útils, com llibreries noves per a la realització de gran part del projecte, o trello i bitbucket per a mantenir una bona organització del projecte.

Tot i haver mostrat només les millors tècniques trobades al llarg del projecte, m'ha resultat també molt enriquidor haver pogut tractar amb comoditat les dades i així haver pogut aplicar molts algorismes i mètodes que desconeixia. D'aquesta manera he pogut experimentar amb diferents tècniques i trobar així resultats útils per al projecte o completament sense sentit, però també útils a nivell personal.

7.1 Treballs futurs

En formar part aquest treball d'un projecte d'innovació docent (PID), les propostes de treballs futurs són per a finalitzar el PID.

Una proposta seria realitzar una pàgina web per a què el tutor pugui accedir de manera còmoda a les funcionalitats i recursos desenvolupats en aquest treball. Alguna de les parts que podria contenir serien la informació completa de l'alumne, recomanacions personalitzades de material addicional i recomanacions a l'hora de fer la matrícula i l'itinerari curricular.

També es podria realitzar una avaluació d'aquesta eina. Per a això els tutors l'haurien d'usar i avisar dels problemes que puguin trobar i de les millores que els agradaria que es realitzessin. També s'hauria de comprovar si amb aquesta ajuda els alumnes tenen un millor rendiment acadèmic.

Referències

[1] Every day big data statistics. http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/.

[2] Ibm - what is a data scientist? – bringing big data to the enterprise. https://www-01.ibm.com/software/data/infosphere/data-scientist/.

[3] Sistema intel·ligent de suport al tutor d'estudis — programa de millora i innovació docent. http://mid.ub.edu/webpmid/content/sistema-intel%E2%80%A2ligent-de-suport-al-tutor-d%E2%80%99estudis.

[4] Standard score - wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Standard_score.

[5] Mean shift de la biblioteca scikit-learn. http://scikit-learn.org/stable/modules/clustering.html#mean-shift.

[6] K-means de la biblioteca scikit-learn. http://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html.

[7] Silhouette-coefficient scikit-learn. http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient.

[8] Nearest neighbors scikit-learn. http://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification.

[9] Pca de la biblioteca scikit-learn. http://scikit-learn.org/stable/modules/decomposition.html#pca.

[10] Tableros — trello. https://trello.com/.

[11] Bitbucket — the git solution for professional teams. https://bitbucket.org/.

[12] Your projects - sharelatex, online latex editor. https://www.sharelatex.com/project.

[13] Latex – a document preparation system. https://www.latex-project.org/.

[14] The jupyter notebook — ipython. https://ipython.org/notebook.html.

[15] Welcome to python. https://www.python.org/.

[16] Python data analysis library — pandas. http://pandas.pydata.org/.

[17] Scikit-learn: machine learning in python. http://scikit-learn.org/stable/.

[18] Numpy — numpy. http://www.numpy.org/.

[19] Bokeh docs. http://bokeh.pydata.org/en/latest/.

[20] Seaborn: statistical data visualization — seaborn 0.7.1 documentation. https://web.stanford.edu/~mwaskom/software/seaborn/.

[21] Matplotlib: python plotting — matplotlib 1.5.1 documentation. http://matplotlib.org/.

[22] Documentació acadèmica que cal adjuntar (via d'accés). http://www.ub.edu/dret/secretaria/matricula/docs/matricula_documentacio_presentar.pdf.

[23] Heat map - wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Heat_map.

[24] Confusion matrix scikit-learn. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html.

Annexos

A Abreviatures

A.1 Abreviatures dels graus

cienc: Ciències Polítiques i de l'Administració

crimi: Criminologia

dret: Dret

gap: Gestió i Administració Pública

info: Enginyeria Informàtica

mates: Matemàtiques

relab: Relacions Laborals

A.2 Abreviatures de les assignatures

Dret

Primer curs

TTC: Tècniques de Treball i Comunicació

CP: Ciència Política

FDD: Fonaments del Dret

DR: Dret Romà

PIC: Principis i Institucions Constitucionals

EC: Economia

SDL: Sistema de Drets i Llibertats

FDPTD: Fonaments del Dret Penal i Teoria del Delicte

HD: Història del Dret

DCP: Dret civil de la persona

Segon curs

OTE: Organització territorial de l'estat

PD: Penes i delictes

DOC: Dret d'Obligacions i Contractes

DIC: Dret internacional públic

DFT: Dret financer i tributari

FDA: Fonaments del dret administratiu

DPC: Dret processal civil

IDCE: Institucions de dret comunitari europeu

Tercer curs

DR: Drets Reals

PST: Procediments i sistema tributari

CAA: Contractació i activitat de l'Administració

DEM: Dret de l'empresa i del mercat

DTS: Dret del treball i la seguretat social

DPP: Dret processal penal

DS: Dret de societats

BPU: Béns públics i urbanisme

Enginyeria Informàtica

Primer curs

P1: Programació I

DDB: Disseny digital bàsic

IO: Introducció als ordinadors

ALGE: Àlgebra

CAL: Càlcul

MD: Matemàtica discreta

FIS: Física

ALGO: Algorísmica

P2: Programació II

ED: Estructura de dades

Segon curs

ELEC: Electrònica

AA: Algorísmica avançada

DS: Disseny de software

EC: Estructura de computadors

ICC: Introducció a la computació científica

EMP: Empresa

PIE: Probabilitat i estadística

PAE: Programació d'arquitectures encastades

PIS: Projecte integrat de software

SO1: Sistemes operatius I

Tercer curs

IA: Intel·ligència Artificial

SO2: Sistemes Operatius II

TNUI: Taller de nous usos de la informàtica

VA: Visió Artificial

XAR: Xarxes

BD: Bases de dades

FHIC: Factors humans i de la computació

GiVD: Gràfics i visualització de dades

LIL: Lògica i llenguatge

SWD: Software distribuït

Ciències Polítiques i de l'Administració

Primer curs

IDP: Introducció al dret públic

CP1: Ciència Política

SG: Sociologia General

HPSC: Història Política i Social Contemporànea

IS: Instrumentàrium

TEIS: Tècniques Estadístiques d'Investigació Social

CA: Ciència de l'Administració

SPE: Sistema Politic Espanyol

ES: Estructura Social

EP: Economia Política

Segon curs

TP1: Teoria Politica

TC: Teoria Constitucional

EM: Economia Mundial

SPC: Sistemes Polítics Comparats

TIIS: Tècniques Informàtiques d'Investigació Social

CP2: Ciència Política II

EE: Economia Espanyola

DA: Dret Administratiu

PP1: Polítiques Públiques

DIPRI: Dret Internacional Públic i Relacions Internacional

Tercer curs

TP2: Teoria Política II

UEIP: Unió Europea: Institucions Polítiques

CP3: Ciència Política II

DA2: Dret Administratiu II

PP2: Polítiques Públiques II

IMP: Ideologies i Moviments Polítics

CoP: Comportament Politic

Criminologia

Primer curs

IDP: Introducció al dret públic

IS: Introducció a la Sociologia

TTC: Tècniques de Treball i Comunicació

HPSC: Història Política i Social Contemporànea

IC: Introducció a la Criminologia

DCDF: Drets Constitucionals i Drets Fonamentals

AC: Anglès Criminològic

EAD: Estadística Anàlisi de Dades

IP: Introducció a la Psicologia

ISP: Introducció al Sistema Polític

Segon curs

TC1: Teories Criminològiques

MC: Metodologia Científica

MLCF: Medicina Legal i Ciències Forenses

FDPTD: Fonaments de Dret Penal i Teoria del Delicte

TC2: Teories Criminològiques II

TP: Tipologies Penals

SD: Sociologia del Dret

PC: Psicologia Criminal

Tercer curs

FC: Formes de Criminalitat

TIC1: Tècniques d'Investigació en Criminologia I

DJ: Delinqüència Juvenil

DCS: Delinqüència i Control Social

POC: Política Criminal

TIC2: Tècniques d'Investigació en Criminologia

EC: Enjudiciament Criminal

Gestió i Administració Pública

Primer curs

ID: Introducció al Dret

TIAP: Tractament de la Informació a l'Admin. Pública

So: Sociologia

HPSC: Història Política i Social Contemporànea

TTC: Tecniques de Treball i Comunicació

DC: Dret Constitucional

AEP: Economia Política

TMGP: Tècniques i Mètodes de la Gestió Pública

CA: Ciència de l'Administració

EA: Estructures Administratives

Segon curs

EAAP1: Estadística Aplicada a l'Admin. Pública I

SPE: Sistema Polític Espanyol

RJAP1: Règim Jurídic de les Administracions Públiques I

HP: Hisenda Pública

IDA: Informació i Documentació Administrativa

EAAP2: Estadística Aplicada a l'Admin. Pública 2

IGF: Introducció a la Gestió Financera

DAUE: Dret i Administració de la Unió Europea

RJAP2: Règim Jurídic de les Administracions Públiques II

ROP: Règim d'Ocupació Pública

Tercer curs

GP1: Gestió Pressupostària I

GT1: Gestió Tributària I

PP: Polítiques Públiques

AA1: Activitat Administrativa I

DTSS: Dret del Treball i de la Seguretat Social

GP2: Gestió Pressupostària II

GT2: Gestió Tributària II

CP: Comptabilitat Pública

AA2: Activitat Administrativa II

DGP: Direcció i Gestió de Persones

Matemàtiques

Primer curs

ADIP: Anàlisi de dades i introducció a la probabilitat

ELPR: Elements de programació

IACD: Introduccio al càlcul diferencial

LIRM: Llenguatge i raonament matemàtic

MAVE: Matrius i vectors

ALLI: Àlgebra lineal

ARIT: Aritmètica

FISI: Física

IACI: Introducció al càlcul integral

PRCI: Programació científica

Segon curs

CDDV: Càlcul diferencial en diverses variables

ESAL: Estructures algebraiques

GELI: Geometria lineal

GRAF: Grafs

MNU1: Mètodes numèrics I

CIDV: Càlcul integral en diverses variables

GEPR: Geometria projectiva

HIMA: Història de les matemàtiques

MMSD: Models matemàtics i sistemes dinàmics

TOPO: Topologia

Tercer curs

ANMA: Anàlisi matemàtica

EQAL: Equacions algebraiques

GDCS: Geometria diferencial de corbes i superfícies

MNU2: Mètodes numèrics II

PROB: Probabilitats

ANCO: Anàlisi complexa

EQDI: Equacions diferencials

ESTA: Estadística

MODE: Modelització

TGGS: Topologia i geometria global de superfícies

Relacions Laborals

Primer curs

IE: Introducció a l'Economia

SPDC: Sistema Polític i Dret Constitucional

PT: Psicologia del Treball

EARL: Estadística Aplicada a les Relacions Laborals

TTC: Tècniques de Treball i Comunicació

ID: Introducció al Dret

OE: Organització d'Empreses

ET: Economia del Treball

HES: Història Econòmica i Social

ST: Sociologia del Treball

Segon curs

DT1: Dret del Treball I

DE: Dret Empresarial

OMT: Organització i Mètodes de Treball

PO: Polítiques d'Ocupació

DT2: Dret del Treball II

GP: Gestió de Persones

C1: Compatibilitat 1

DSS1: Dret de Seguretat Social I

Tercer curs

DPRL: Dret de la Prevenció de Riscos Laborals

RJPAC: Règim Jurídic i Procediment Administratiu Comú

RJEP: Règim Jurídic dels Empleats Públics

C2: Compatibilitat II

DSS2: Dret de la Seguretat Social II

DS1: Dret Sindical I

SSL: Seguretat i Salut Laboral

RJIT: Règim Jurídic Internacional del Treball

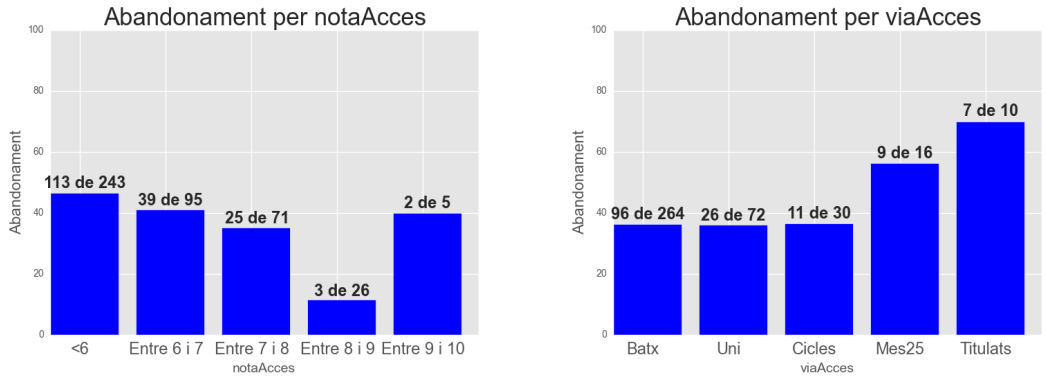
GOP: Gestió de l'Ocupació Pública

FE: Fiscalitat en l'Empresa

B Gràfics de tots els graus

B.1 Abandonament dels alumnes per nota i via d'accés

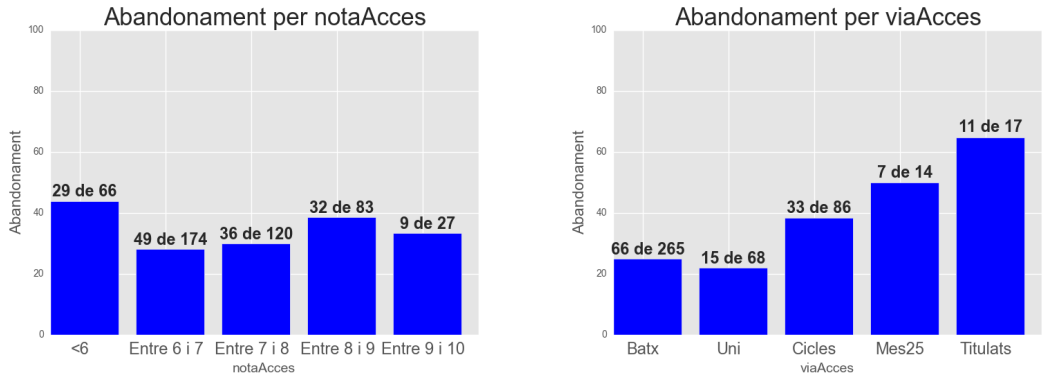
Ciències Polítiques i de l'Administració



(a) Abandonament per nota d'accés (b) Abandonament per via d'accés

Figura 25: Abandonament a Ciències Polítiques i de l'Administració dels alumnes depenent de la nota i la via d'accés al grau.

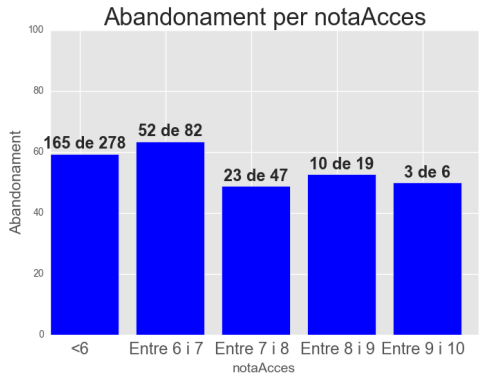
Criminologia



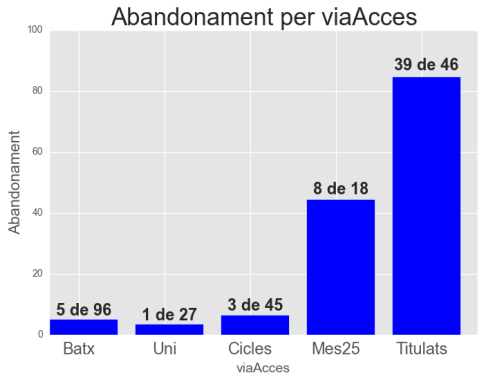
(a) Abandonament per nota d'accés (b) Abandonament per via d'accés

Figura 26: Abandonament a Criminologia dels alumnes depenent de la nota i la via d'accés al grau.

Gestió i Administració Pública



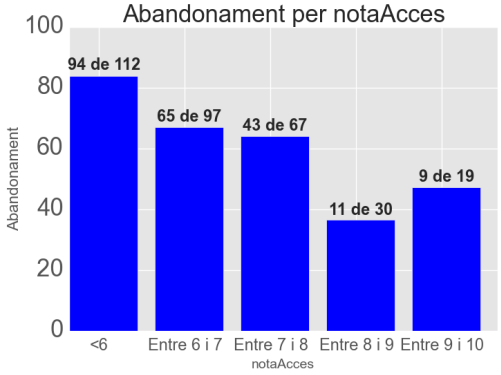
(a) Abandonament per nota d'accés



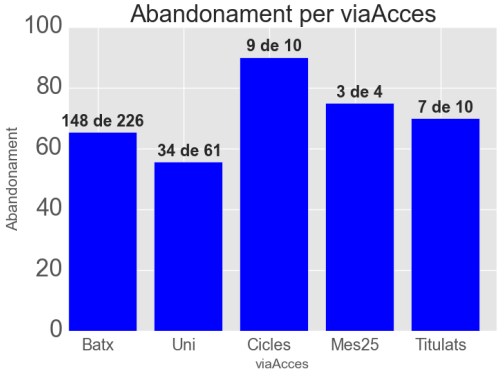
(b) Abandonament per via d'accés

Figura 27: Abandonament a Gestió i Administració Pública dels alumnes depenent de la nota i la via d'accés al grau.

Matemàtiques



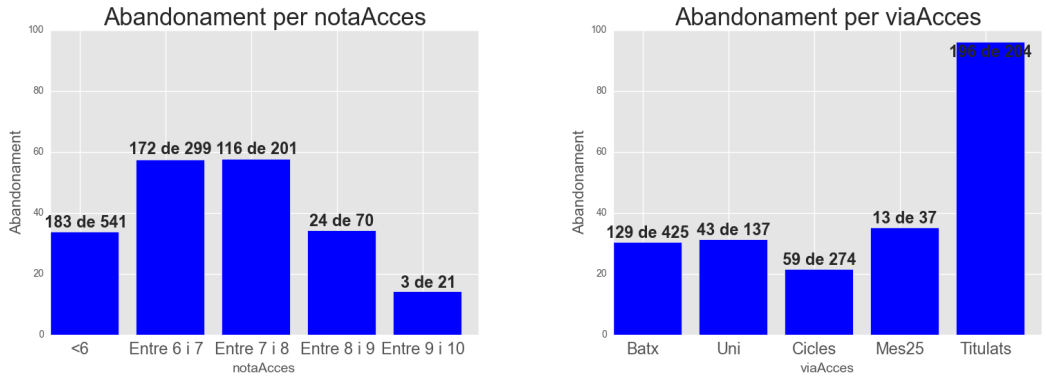
(a) Abandonament per nota d'accés



(b) Abandonament per via d'accés

Figura 28: Abandonament a Matemàtiques dels alumnes depenent de la nota i la via d'accés al grau.

Relacions Laborals



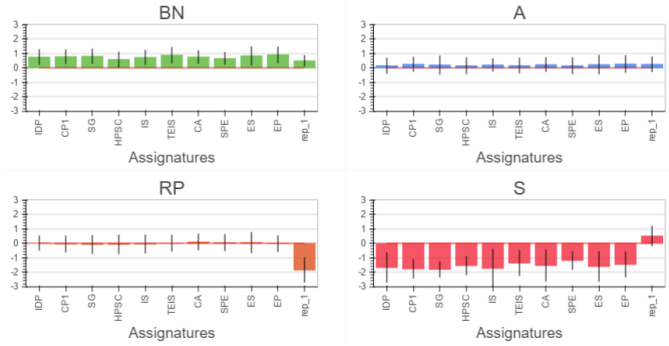
(a) Abandonament per nota d'accés

(b) Abandonament per via d'accés

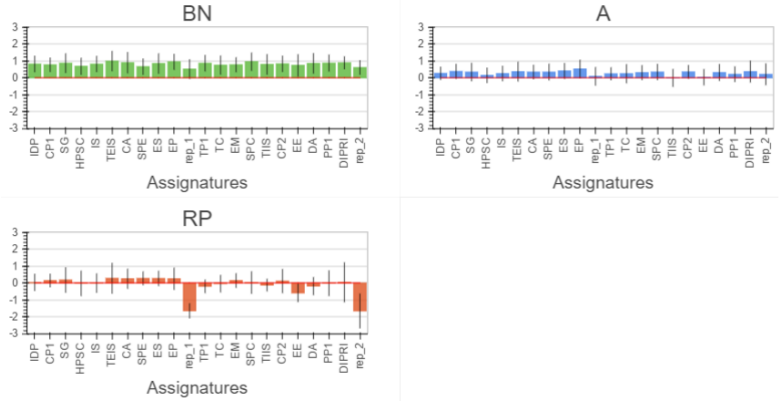
Figura 29: Abandonament a Relacions Laborals dels alumnes depenent de la nota i la via d'accés al grau.

B.2 Perfil dels clústers

Ciències Polítiques i de l'Administració



(a) Primer curs

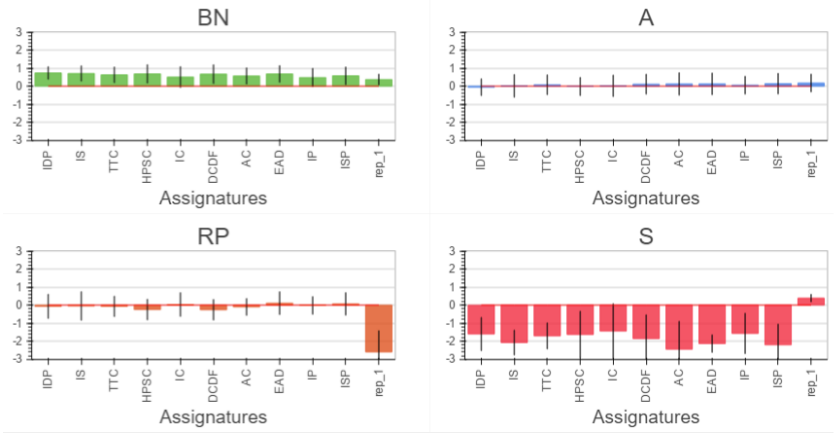


(b) Segon curs

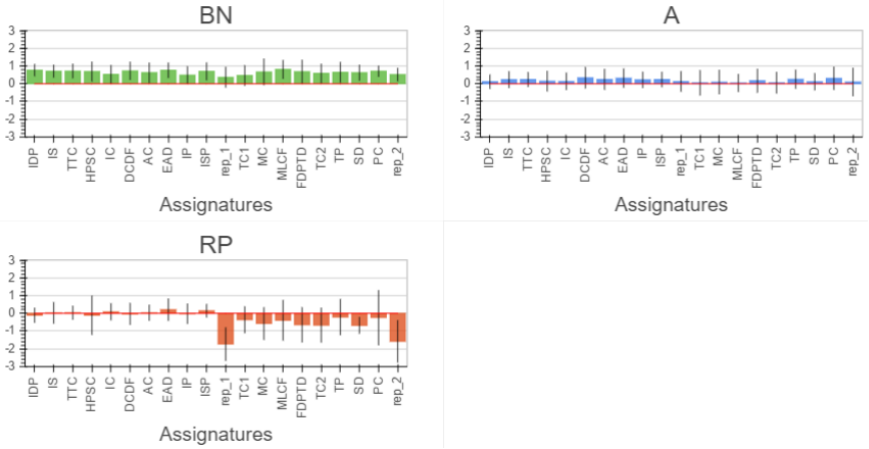
(c) Tercer curs

Figura 30: Diagrama de columnes on es veu per a cada clúster i cada curs la mitjana de les notes de les assignatures estandarditzades.

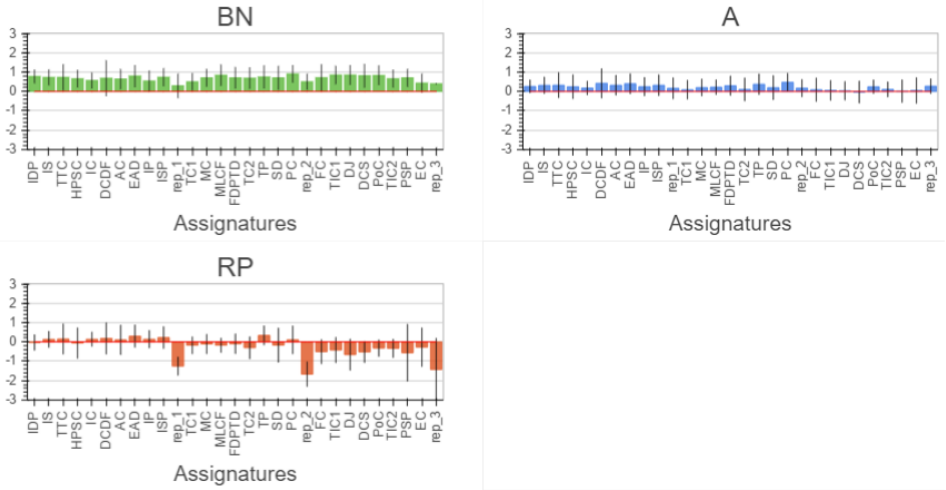
Criminologia



(a) Primer curs



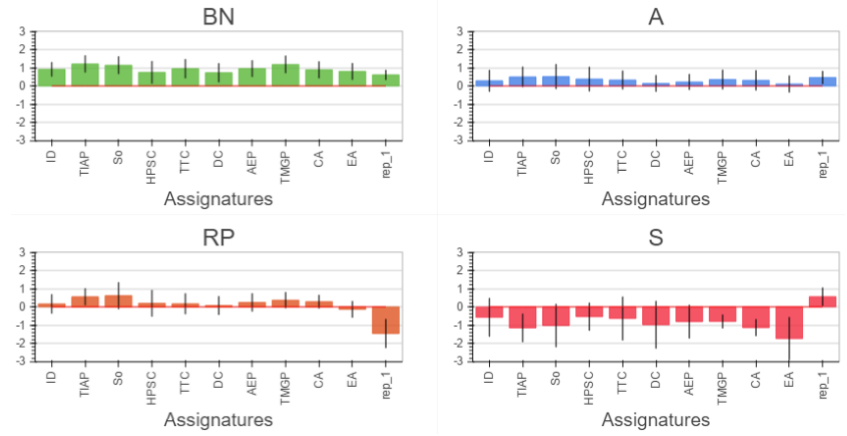
(b) Segon curs



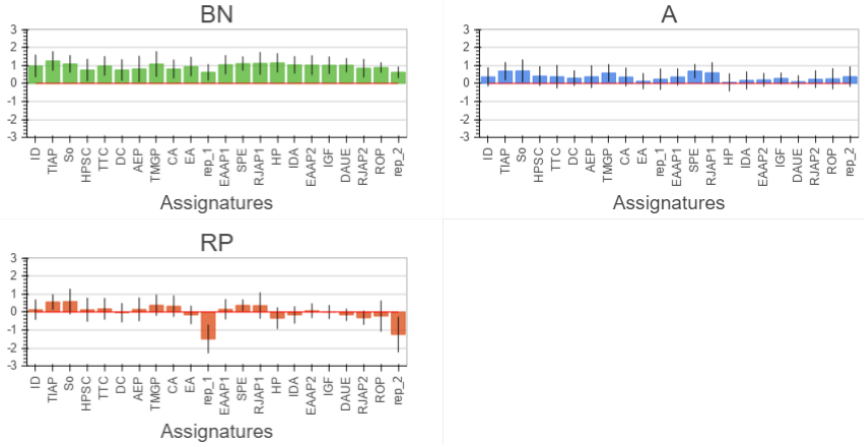
(c) Tercer curs

Figura 31: Diagrama de columnes on es veu per a cada clúster i cada curs la mitjana de les notes de les assignatures estandarditzades.

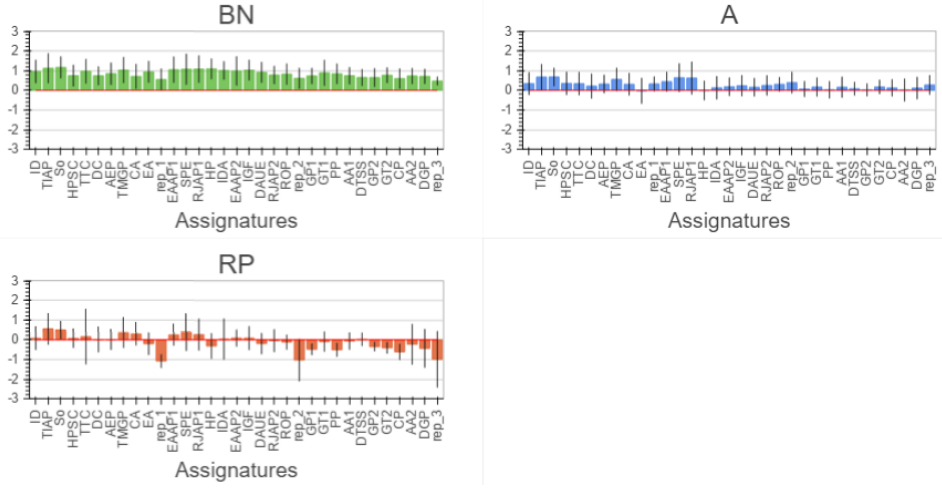
Gestió i Administració Pública



(a) Primer curs



(b) Segon curs



(c) Tercer curs

Figura 32: Diagrama de columnes on es veu per a cada clúster i cada curs la mitjana de les notes de les assignatures estandarditzades.

Matemàtiques

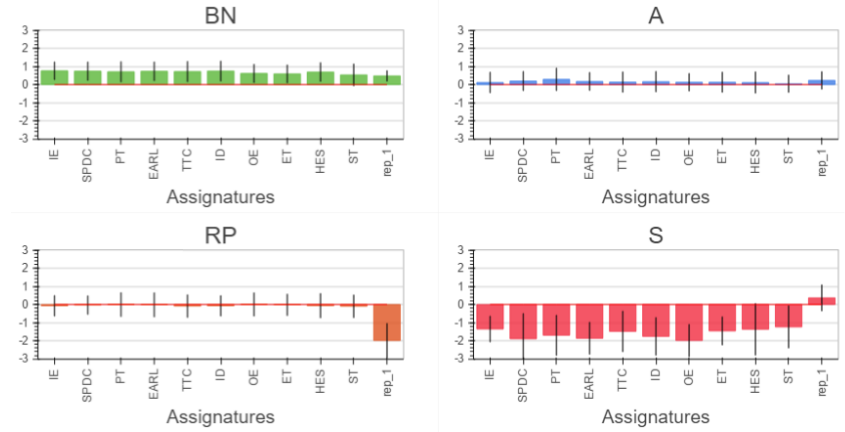
(a) Primer curs

(b) Segon curs

(c) Tercer curs

Figura 33: Diagrama de columnes on es veu per a cada clúster i cada curs la mitjana de les notes de les assignatures estandarditzades.

Relacions Laborals



(a) Primer curs

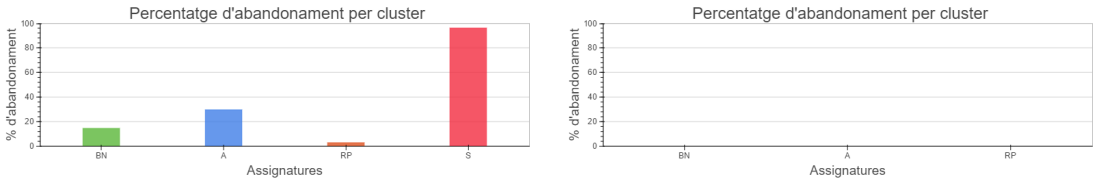
(b) Segon curs

(c) Tercer curs

Figura 34: Diagrama de columnes on es veu per a cada clúster i cada curs la mitjana de les notes de les assignatures estandarditzades.

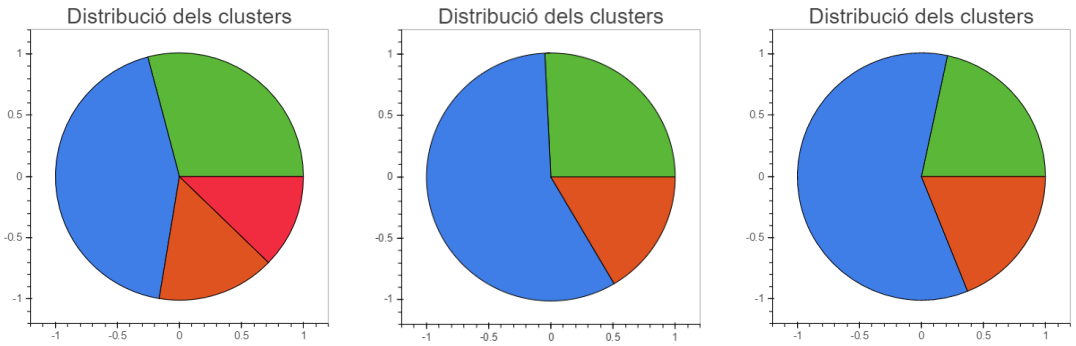
B.3 Abandonament i distribució dels clústers

Ciències Polítiques i de l'Administració



(a) Primer curs (b) Segon curs

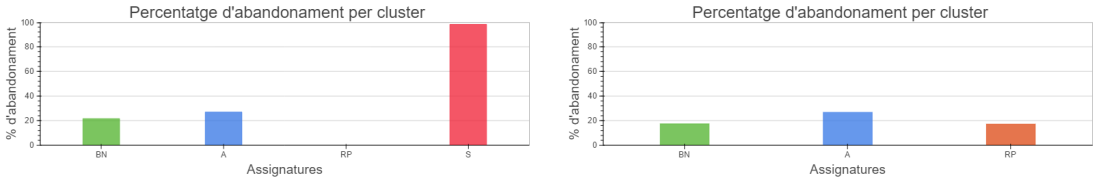
Figura 35: Diagrama de columnes que mostra l'abandonament de cada clúster de Ciències Polítiques i de l'Administració.



(a) Primer curs (b) Segon curs (c) Tercer curs

Figura 36: Diagrama de sectors que mostra com estan repartits els clústers de cada curs de Ciències Polítiques i de l'Administració

Criminologia



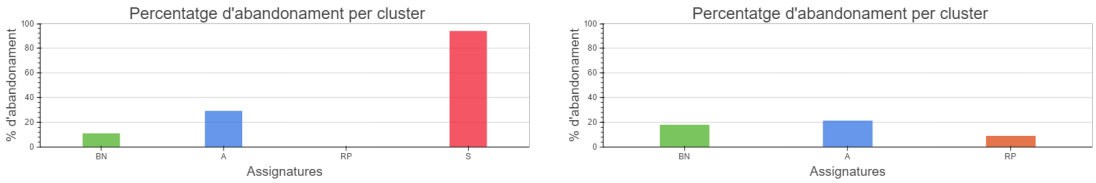
(a) Primer curs (b) Segon curs

Figura 37: Diagrama de columnes que mostra l'abandonament de cada clúster de Criminologia.

(a) Primer curs (b) Segon curs (c) Tercer curs

Figura 38: Diagrama de sectors que mostra com estan repartits els clústers de cada curs de Criminologia.

Gestió i Administració Pública



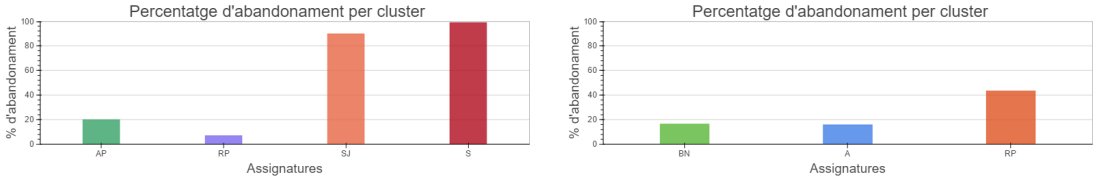
(a) Primer curs (b) Segon curs

Figura 39: Diagrama de columnes que mostra l'abandonament de cada clúster de Gestió i Administració Pública.

(a) Primer curs (b) Segon curs (c) Tercer curs

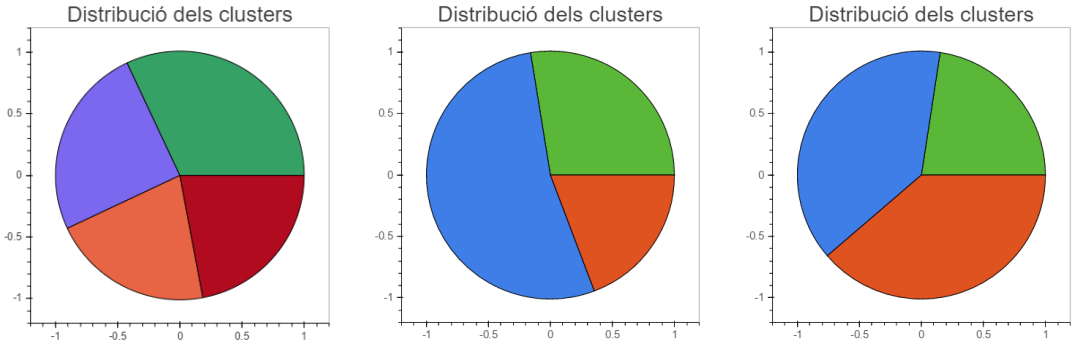
Figura 40: Diagrama de sectors que mostra com estan repartits els clústers de cada curs de Gestió i Administració Pública.

Matemàtiques



(a) Primer curs (b) Segon curs

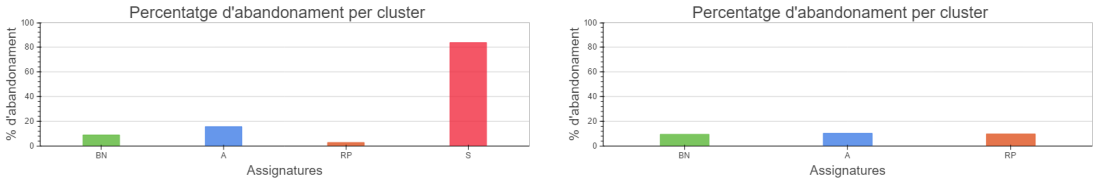
Figura 41: Diagrama de columnes que mostra l'abandonament de cada clúster de Matemàtiques.



(a) Primer curs (b) Segon curs (c) Tercer curs

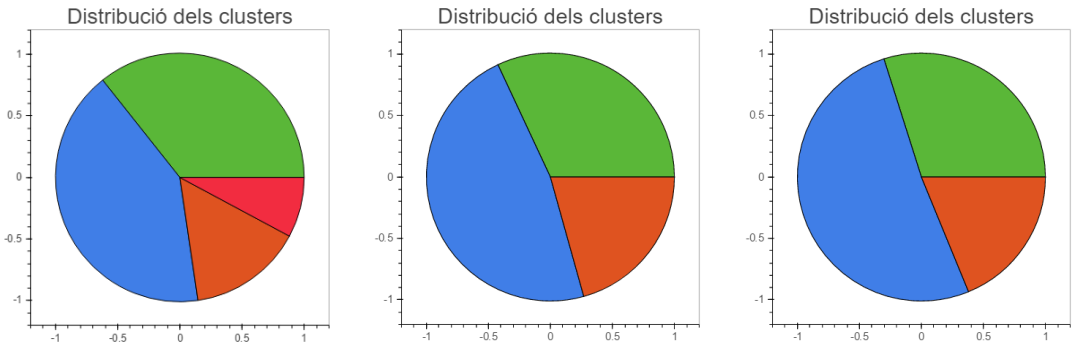
Figura 42: Diagrama de sectors que mostra com estan repartits els clústers de cada curs de Matemàtiques.

Relacions Laborals



(a) Primer curs (b) Segon curs

Figura 43: Diagrama de columnes que mostra l'abandonament de cada clúster de Relacions Laborals.

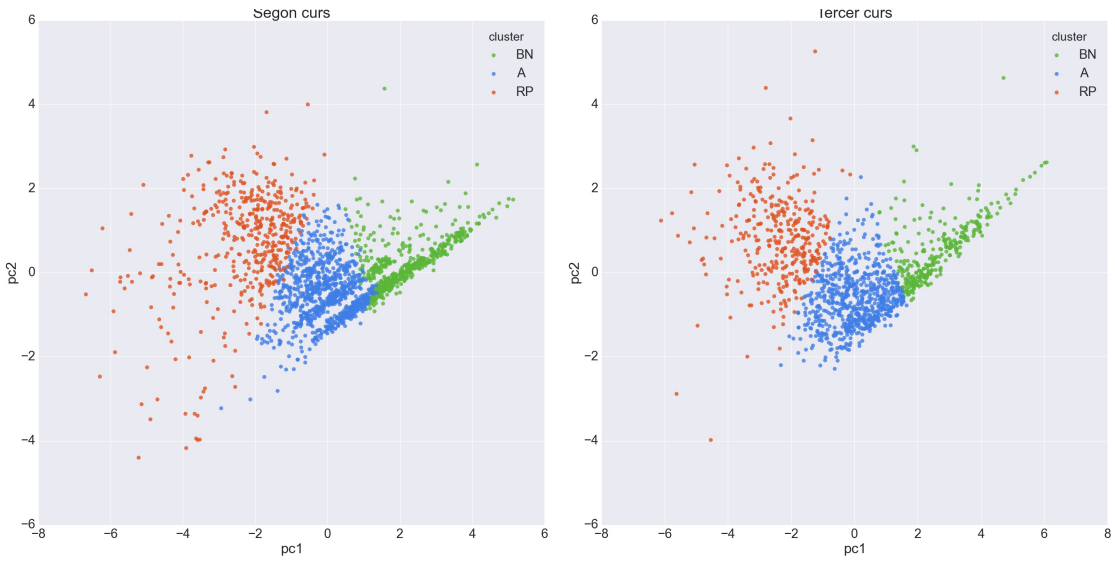


(a) Primer curs (b) Segon curs (c) Tercer curs

Figura 44: Diagrama de sectors que mostra com estan repartits els clústers de cada curs de Relacions Laborals.

B.4 Visualització en 2D dels clústers

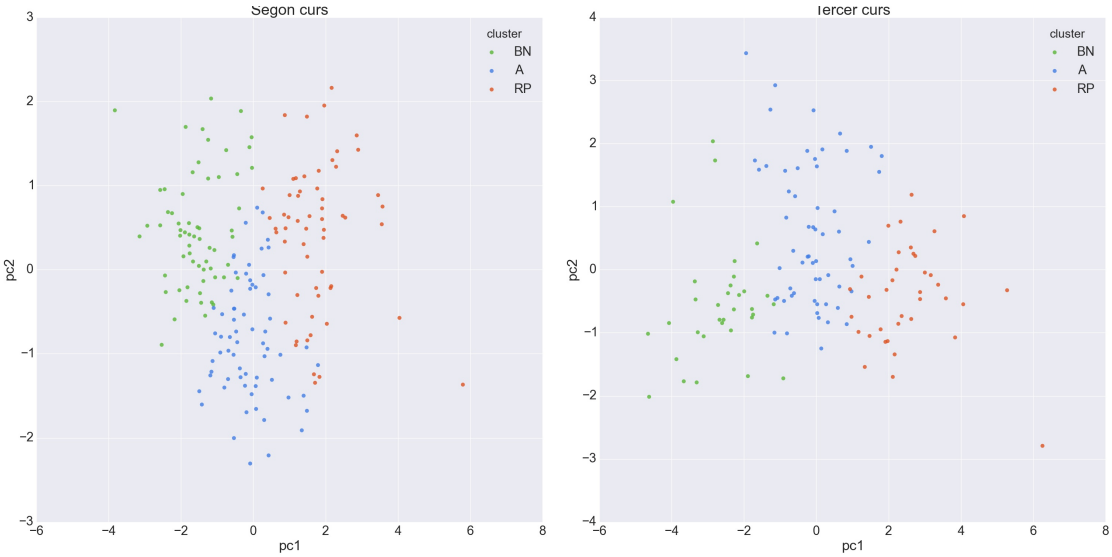
Dret



(a) Segon curs (b) Tercer curs

Figura 45: Visualització en 2D dels alumnes i els clústers de segon i tercer curs de Dret amb el mètode PCA i K-Means.

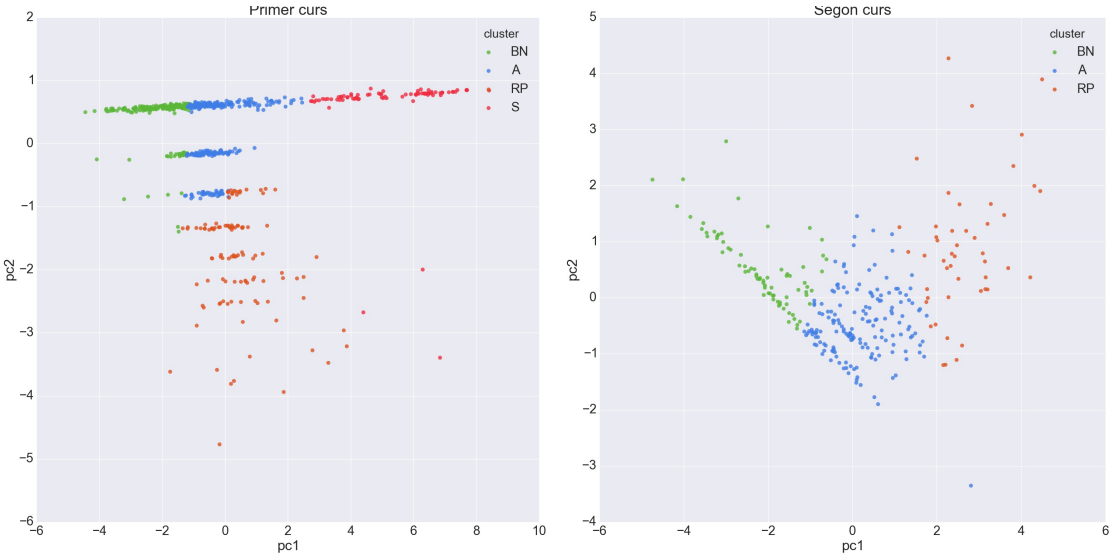
Enginyeria Informàtica



(a) Segon curs

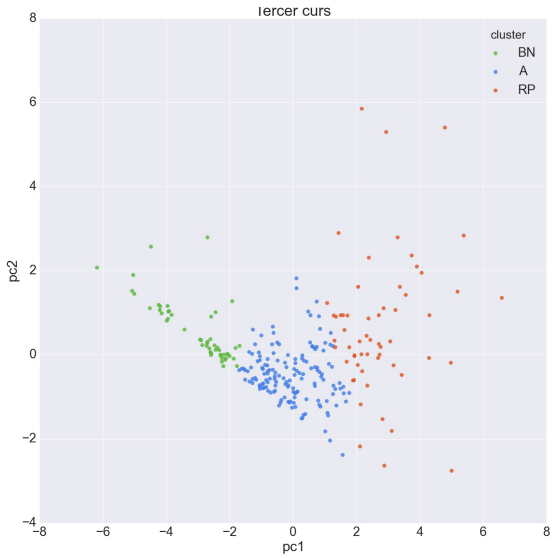
(b) Tercer curs

Figura 46: Visualització en 2D dels alumnes i els clústers de segon i tercer curs d'Enginyeria Informàtica amb el mètode PCA i K-Means.



(a) Primer curs

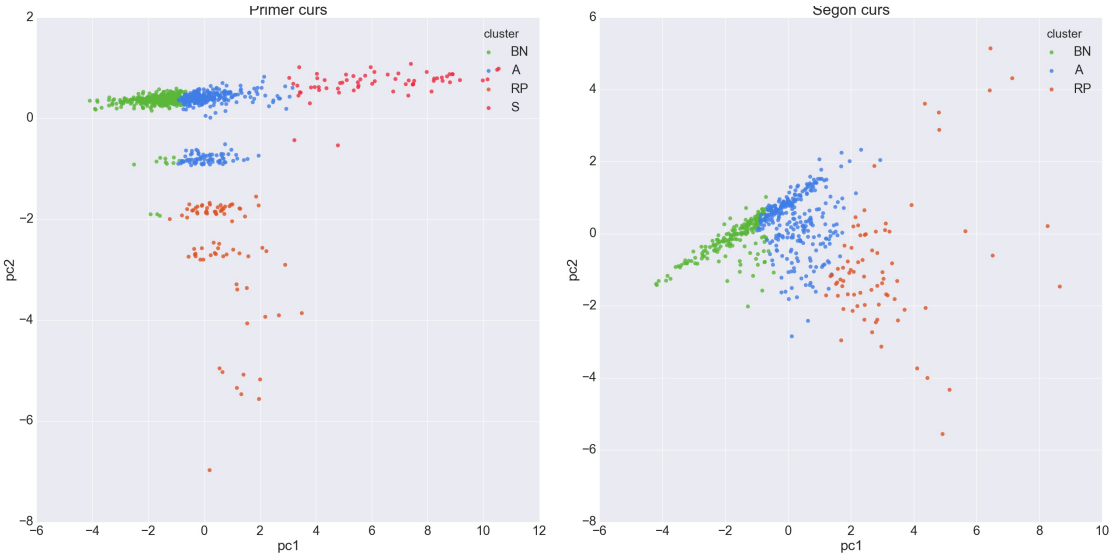
(b) Segon curs



(c) Tercer curs

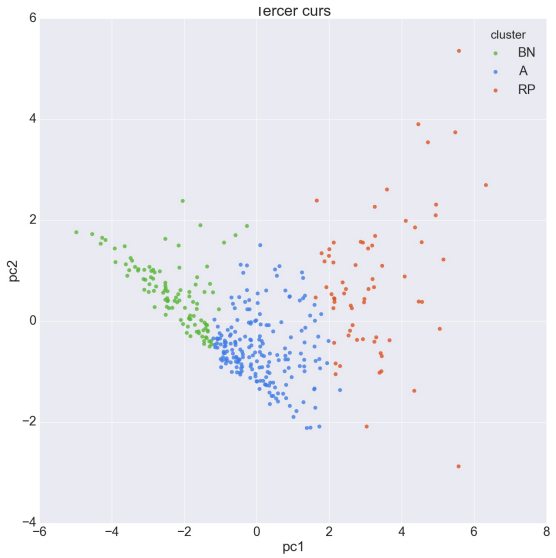
Figura 47: Visualització en 2D dels alumnes i els clústers de Ciències Polítiques i de l'Administració amb el mètode PCA i K-Means.

Criminologia



(a) Primer curs

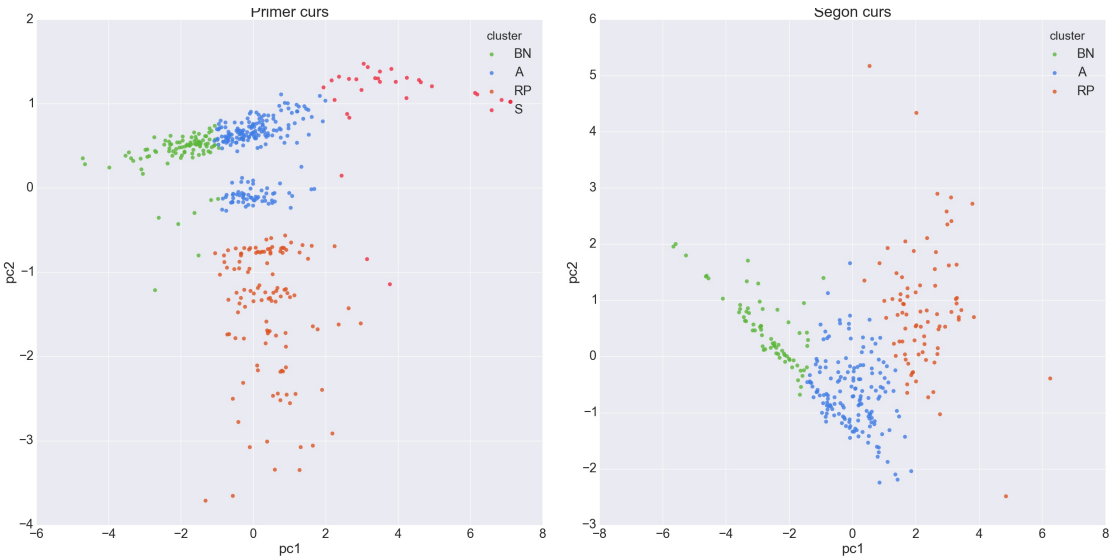
(b) Segon curs



(c) Tercer curs

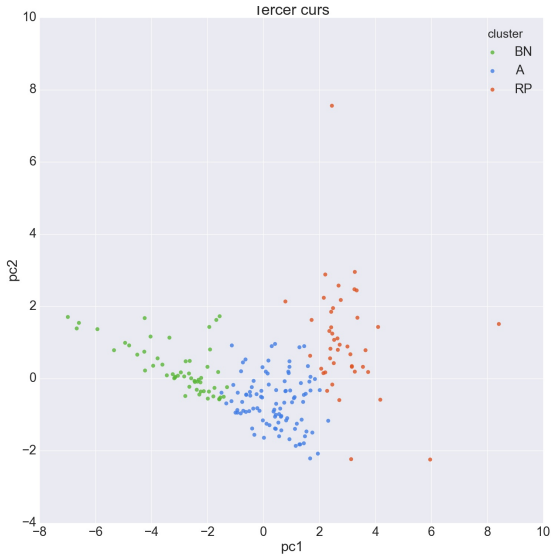
Figura 48: Visualització en 2D dels alumnes i els clústers de Criminologia amb el mètode PCA i K-Means.

Gestió i Administració Pública



(a) Primer curs

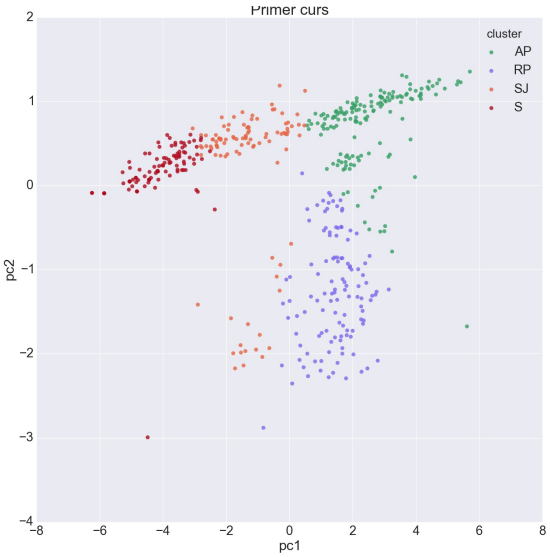
(b) Segon curs



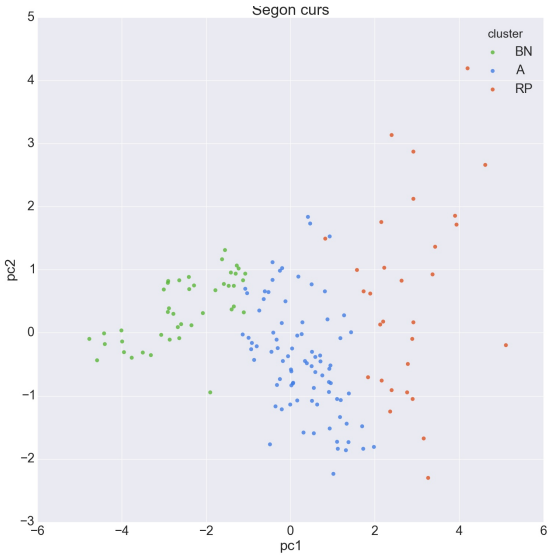
(c) Tercer curs

Figura 49: Visualització en 2D dels alumnes i els clústers de Gestió i Administració Pública amb el mètode PCA i K-Means.

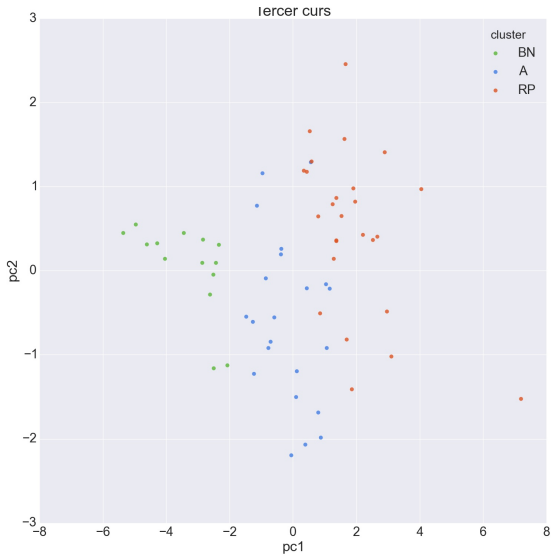
Matemàtiques



(a) Primer curs



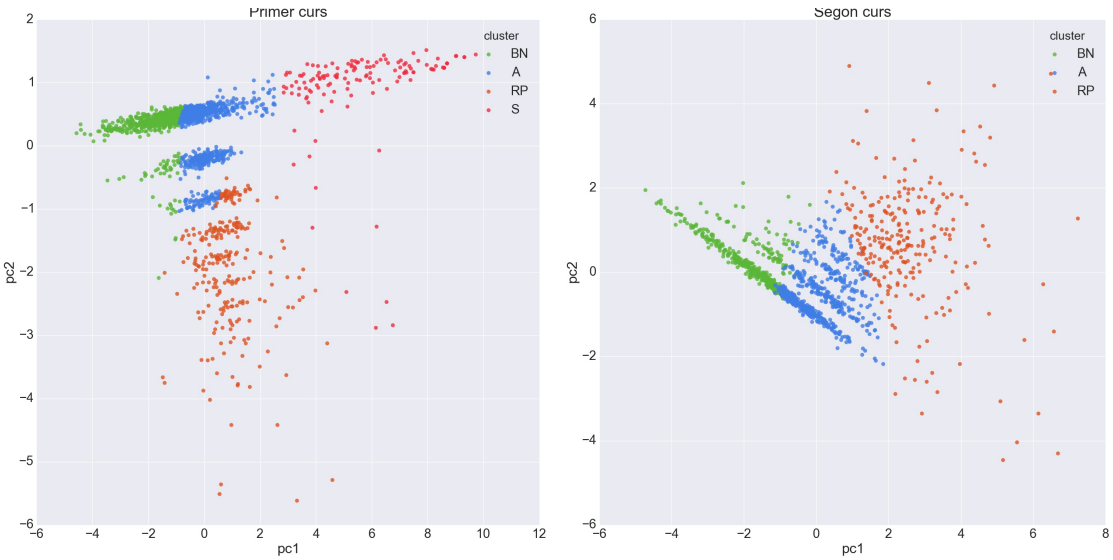
(b) Segon curs



(c) Tercer curs

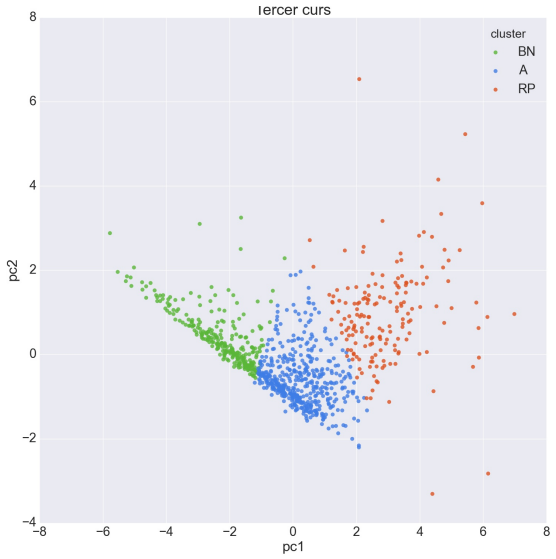
Figura 50: Visualització en 2D dels alumnes i els clústers de Matemàtiques amb el mètode PCA i K-Means.

Relacions Laborals



(a) Primer curs

(b) Segon curs

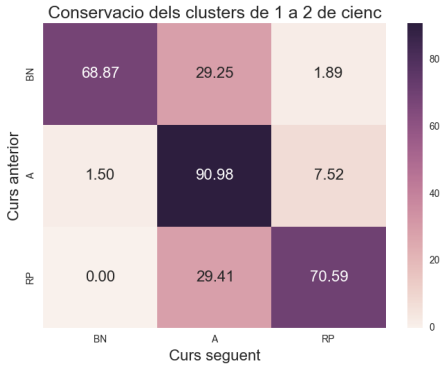


(c) Tercer curs

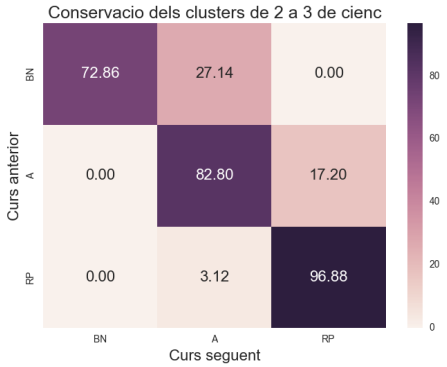
Figura 51: Visualització en 2D dels alumnes i els clústers de Relacions Laborals amb el mètode PCA i K-Means.

B.5 Conservació dels clústers

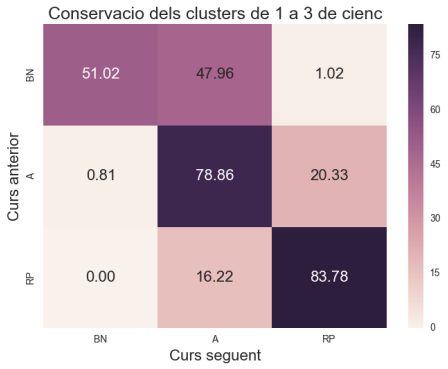
Ciències Polítiques i de l'Administració



(a) Primer a segon curs



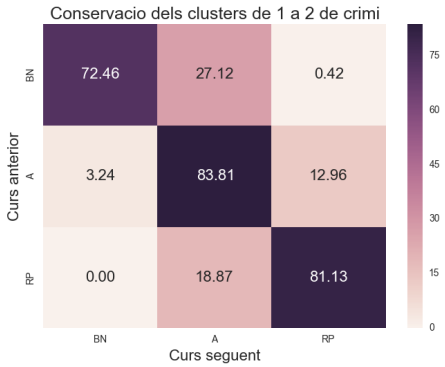
(b) Segon a tercer curs



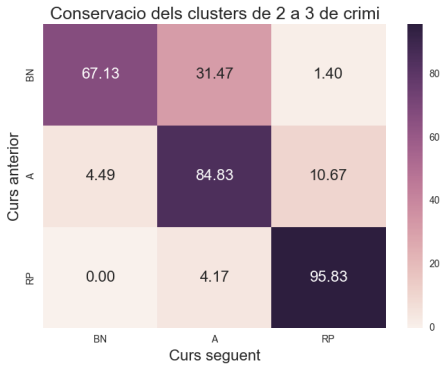
(c) Primer a tercer curs

Figura 52: Mapa de calor on es veu la conservació dels clústers de Ciències Polítiques i de l'Administració. Els valors de les files són percentatges. Sumant cada fila obtenim el 100%. Les files mostren el clúster en què es trobaven al curs anterior i les columnes el clúster en què es troben al curs següent.

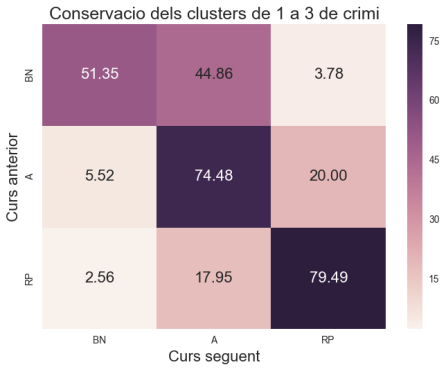
Criminologia



(a) Primer a segon curs



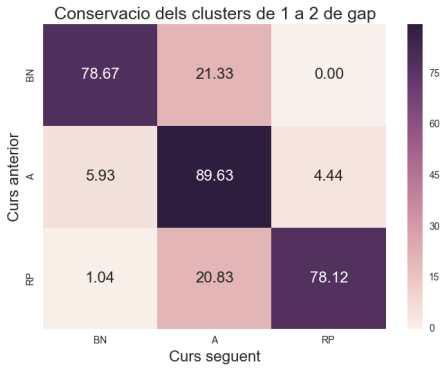
(b) Segon a tercer curs



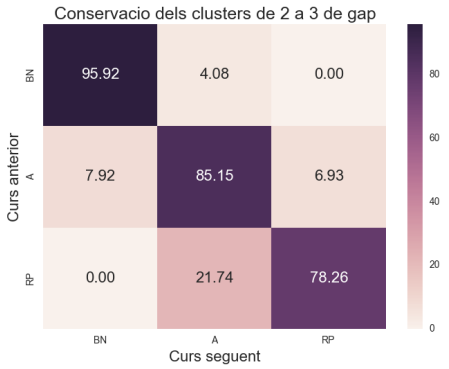
(c) Primer a tercer curs

Figura 53: Mapa de calor on es veu la conservació dels clústers de Criminologia. Els valors de les files són percentatges. Sumant cada fila obtenim el 100%. Les files mostren el clúster en què es trobaven al curs anterior i les columnes el clúster en què es troben al curs següent.

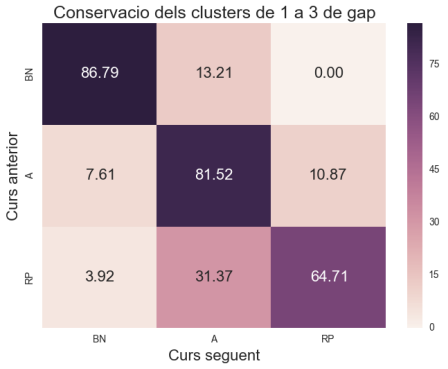
Gestió i Administració Pública



(a) Primer a segon curs



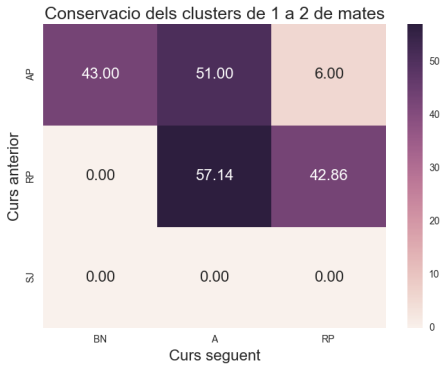
(b) Segon a tercer curs



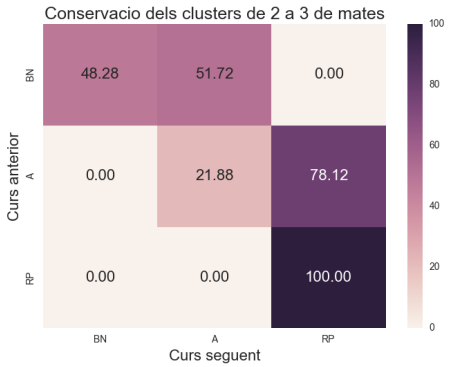
(c) Primer a tercer curs

Figura 54: Mapa de calor on es veu la conservació dels clústers de Gestió i Administració Pública. Els valors de les files són percentatges. Sumant cada fila obtenim el 100%. Les files mostren el clúster en què es trobaven al curs anterior i les columnes el clúster en què es troben al curs següent.

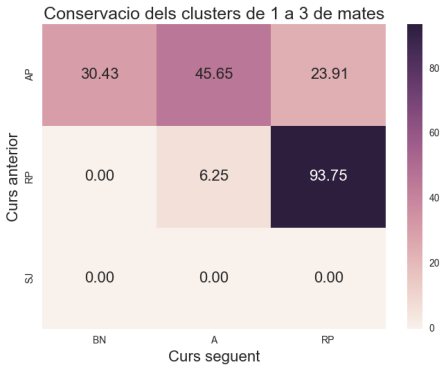
Matemàtiques



(a) Primer a segon curs



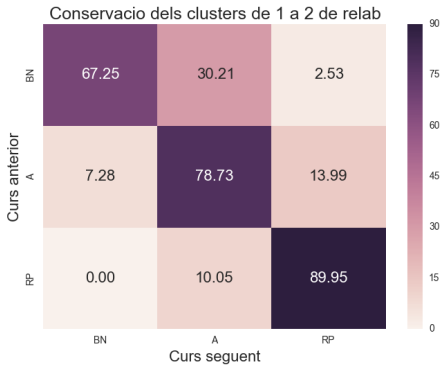
(b) Segon a tercer curs



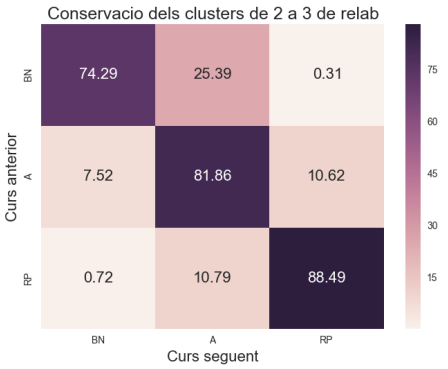
(c) Primer a tercer curs

Figura 55: Mapa de calor on es veu la conservació dels clústers de Matemàtiques. Els valors de les files són percentatges. Sumant cada fila obtenim el 100%. Les files mostren el clúster en què es trobaven al curs anterior i les columnes el clúster en què es troben al curs següent.

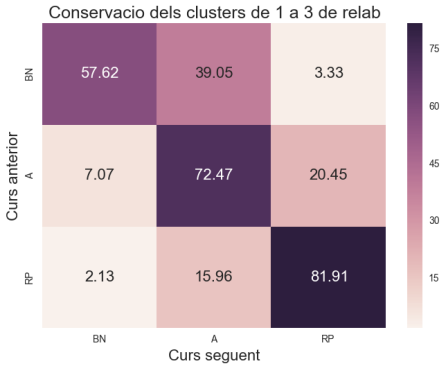
Relacions Laborals



(a) Primer a segon curs



(b) Segon a tercer curs

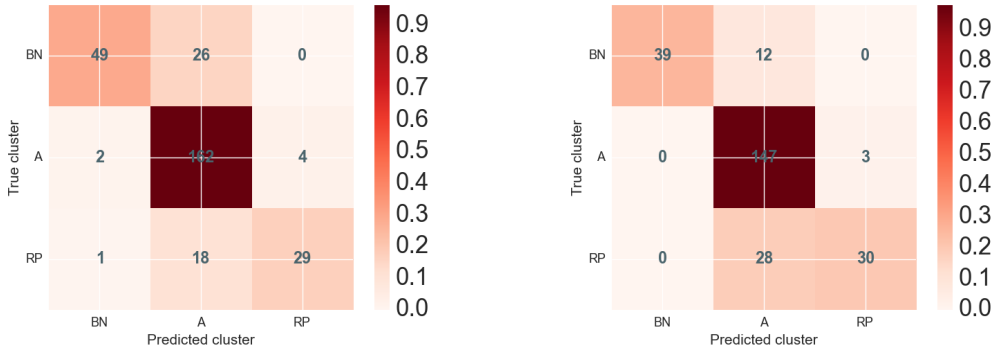


(c) Primer a tercer curs

Figura 56: Mapa de calor on es veu la conservació dels clústers de Relacions Laborals. Els valors de les files són percentatges. Sumant cada fila obtenim el 100%. Les files mostren el clúster en què es trobaven al curs anterior i les columnes el clúster en què es troben al curs següent.

B.6 Avaluació del predictor

Ciències Polítiques i de l'Administració



(a) Primer a segon curs

(b) Segon a tercer curs

Figura 57: Matriu de confusió amb la precisió del predictor de Ciències Polítiques i de l'Administració. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.

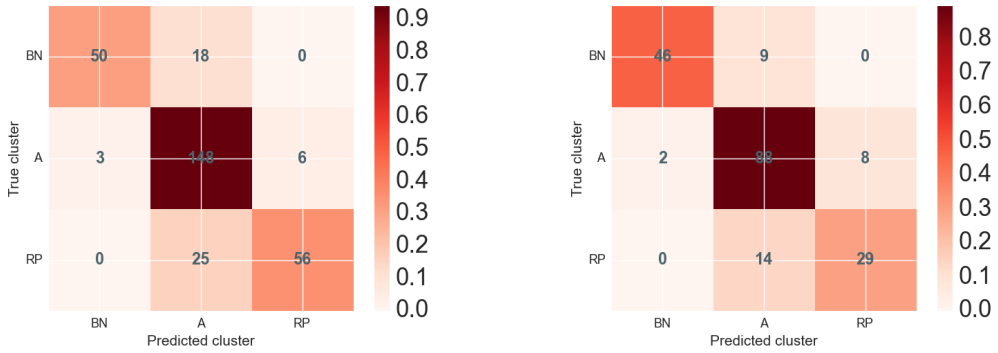
Criminologia

(a) Primer a segon curs

(b) Segon a tercer curs

Figura 58: Matriu de confusió amb la precisió del predictor de Criminologia. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.

Gestió i Administració Pública

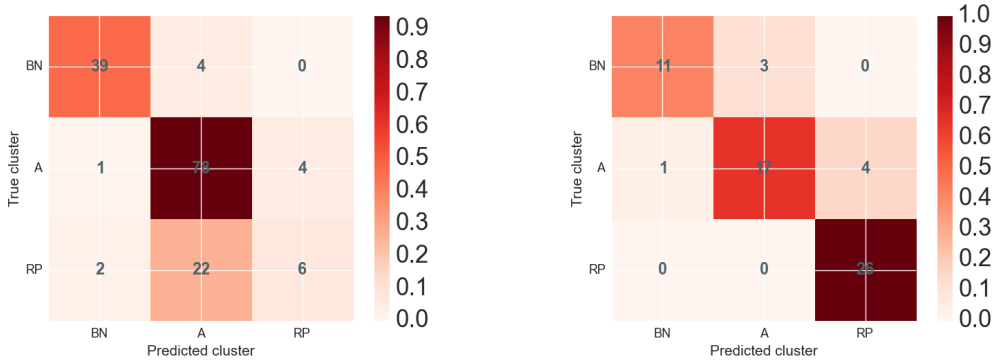


(a) Primer a segon curs

(b) Segon a tercer curs

Figura 59: Matriu de confusió amb la precisió del predictor de Gestió i Administració Pública. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.

Matemàtiques

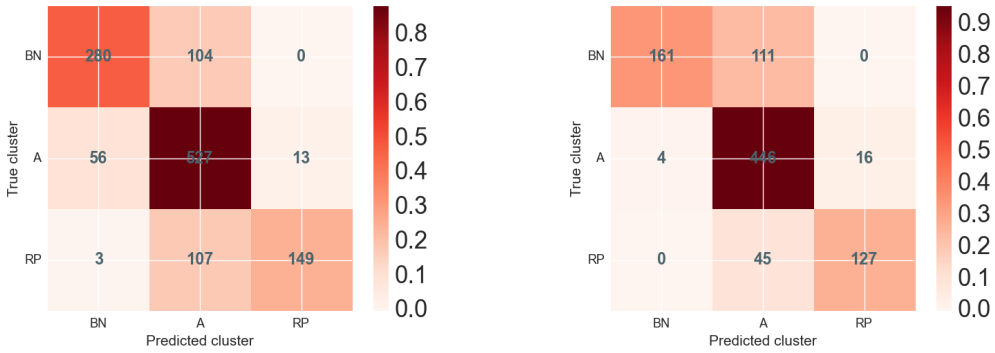


(a) Primer a segon curs

(b) Segon a tercer curs

Figura 60: Matriu de confusió amb la precisió del predictor de Matemàtiques. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.

Relacions Laborals



(a) Primer a segon curs

(b) Segon a tercer curs

Figura 61: Matriu de confusió amb la precisió del predictor de Relacions Laborals. Les files mostren la freqüència d'alumnes que pertanyen a aquell clúster i les columnes mostren la freqüència d'alumnes en aquell clúster segons el predictor.