Anàlisi i Mineria de Dades UOC 2015

Data mining MOC UPF/C01/2014

Director: Ramon Caihuelas Quiles

Alumne: Josep Maria Bella i Garcia

Juny 2015

"les dades només són dades"

Campus UOC 2014

Índex

Introducció

Aquest treball està orientat a treballar les dades donades referents a diversos cursos fets a finals del 2014.

L'objectiu principal és conèixer les característiques bàsiques dels models de Data mining i el procés d'extracció de coneixement.

El treball està estructurat en tres parts. La primera relacionada amb la comprensió i la neteja de les dades. La segona orientada a l'anàlisi i síntesi de les dades. I la última, l'aplicació dels models de data mining.

Cada etapa és acumulativa respecte de les anteriors. Es farà difícil la comprensió dels models sense una lectura completa del treball, sobretot, d'aquells passos que s'han fet de neteja i discretització de les dades.

Per últim, agrair la paciència, saviesa i comprensió d'en Ramon, estic francament satisfet d'haver escollit aquest TFG.

La naturalesa de les dades

Al llarg del curs s'ha fet arribar dos arxius en format JSON: "tot_arr.json" de 326,7 MB i creat el 5 de març del 2015; i l'arxiu "rest_file.json", de 74,2 MB i creat el 13 de maig.

En ambdós arxius, hi havia els registres de diversos usuaris que havien participat en diversos cursos online. L'objectiu era analitzar les dades i aplicar els models de Data Mining als registres.

Lectura de les dades

Tots dos arxius no complien els requisits del programa que es farà servir (R-PROJECT) així que s'ha tingut de fer un procés de neteja i transformació de dades de caràcter purament tècnic.

Aquest procés de neteja s'ha fet a través d'un programa en JAVA, les principals característiques que té són:

- Transformació dels caràcters de text a UTF-8.
- Transformació de la data que venia en el format "YYYY-MM-DDTHH-MM-SS.MMMMMM+HH:MM" i s'ha transformat en segons des del 1 de gener del 2014 a les 00:00:00.
- Assegurar-se que el format JSON sigui el correcte i en cas que no ho sigui corregir-ho.
- Eliminar registres que no contenen informació destacada i són de difícil correcció.

Després d'aquest procés ens s'han quedat 408173 registres. Encara que queda alguna dada corrompuda que s'anirà netejant posteriorment.

Dades generals

Les dades segueixen l'estructura EDX la qual està destinada a investigadors, experts en data i equips administratius de les universitats i que els serveix per gestionar, administrar i investigar les dades dels cursos.

Aquesta estructura està composta per una base comuna d'atributs que segons l'acció que faci l'usuari se li sumarà un conjunt d'atributs o un altre. Aquests atributs específics estan encapsulats en dos registres: event i context.

Base comuna

Els atributs de la base comuna són:

- agent : ens dóna la informació del navegador que fa servir l'usuari.
- event_source : origen d'on provenen les dades. Pot ser "browser" o "server".
- event_type: especifica el tipus d'acció que es duu a terme en aquell registre, molt important, doncs marcarà quin tipus de registres específics trobarem al event.
- host : la web que visita el usuari
- ip : ip del usuari
- name : té la mateixa funció que event_type, però només en tres accions (matricular-se, desmatricular-se i buscar al fòrum).
- page : és la url de la pàgina que visita l'usuari
- session : cadena de 32 caràcters que identifica les diverses sessions que fan els usuaris.
- time : dóna el temps UTC que es fa l'acció amb el format -> 'YYYY-MM-DDThh:mm:ss.xxxxxx'
- username : cadena de 32 caràcters que identifica els usuaris.

- context : Tot i variar en alguns moments, sempre conté la següent informació.
    - course_id : curs que genera l'event.
    - org_id : organització que genera l'event.
    - path : la url que genera l'event.
    - user_id : usuari que fa l'acció, però no amb la codificació de 32 caràcters, sinó amb un enter.
- event : on es guarden els registres específics.

Tots els registres específics sempre estan dins del registre d'event a no sé que es digui el contrari. Aquests registres variaran segons l'acció que es faci. Aquesta acció queda recollida en l'atribut "event_type".

El nom d'aquests atributs en el nostre programa seguirà la forma "event.nom_atribut".

Per tenir la informació d'una forma esquemàtica, tenim amb un "*" el valor que li correspon a "event_type", i amb un "-" els atributs específics i el seu significat.

Les dades de la matriculació

* edx.couser.enrollment.activated
* edx.course.enrollment.deactivated.

- course_id : curs on ens matriculem o on ens esborrem.
- mode : indica l'estatus dels alumnes, tots són honor.
- user_id : enter que identifica l'usuari.

Les dades dels vídeos

* hide_transcript -> amaga la transcripció
* show_transcript -> mostra la transcripció
* pause_video -> es prem pause o a la fi del vídeo.
* play_video -> prem play.
* stop_video -> final automàtic.

- code : identificador 11 caràcters del vídeo.
- currentTime : punt en segons del vídeo on succeïx l'acció.
- id : identificador que junta el curs i el vídeo.

* load_video -> quan un vídeo s'ha carregat.

- code : identificador 11 caràcters del vídeo.

- id : identificador que junta el curs i el vídeo.

* seek_video -> per anar a un punt concret d'un vídeo.

- code : identificador 11 caràcters del vídeo.

- id : identificador que junta el curs i el vídeo.

- new_time : punt en segons del vídeo on anem.

- old_time : punt en segons del vídeo d'on venim.

- type : tipus de salt, tots són iguals.

* speed_change_video -> canvi de velocitat.

- code : identificador 11 caràcters del vídeo.

- currentTime : punt en segons del vídeo on ocórrer l'acció.

- id : identificador que junta el curs i el vídeo.

- new_speed : nova velocitat que volem.

- old_speed : velocitat anterior.

Les dades de la navegació

* seq_goto -> quan saltem a una unitat concreta.

* seq_next -> quan anem a la unitat següent.

* seq_prev -> quan anem a la unitat anterior.

- id : identifcador EDX

- new : enter de la unitat on anem.

- old : enter de la unitat d'on venim.

Les dades dels problemes d'avaluació

* problem_check -> S'emet quan un problema s'ha comprovat si és correcte.

- answers -> és un parell de valors, "id_problem : resposta". Ens dóna la resposta feta per l'usuari a una pregunta. Són 50 columnes.

- attempts -> un enter, on ens mostra el nombre d'intents que porta l'usuari. Màxim tres intents.

- grade -> nota que ha tret l'alumne en el check.

- max_grade -> màxima puntuació que pot obtenir un alumne, atenció, no sempre és 10.
- problem_id : un total de 7 problemes, cadascun amb el seu identificador.
- success -> ens indica si l'usuari ha conseguit superar la prova, per això s'ha d'obtenir en grade el mateix que en max_grade.

No s'han fet servir la resta de funcions dels problemes d'avaluació ja que amb això ja tenim força informació.

Objectius i línies de treball

En un primer moment aquest treball es va dirigir a relacionar la diversitat de l'alumnat amb la seva feina i la seva avaluació. L'objectiu era clar, buscar quines eren les clau d'èxit de l'aprenentatge en línia.

Però en la mesura que s'entrava en la naturalesa dels cursos estudiats, aquests objectiu s'esvaïa i apareixien els dubtes.

En un primer moment es volia buscar algun tipus de funció, del tipus neuronal per exemple, que relaciones: treball fet a les unitats, treball fet als vídeos, nota obtinguda.

Però el model d'avaluació del curs que més endavant s'analitzarà ho fa impossible, doncs el curs no forma part de l'avaluació d'una assignatura, sinó d'un MOC optatiu per reforçar els conceptes d'àlgebra.

Per tant, ens podem trobar alumnes que veient el temari no creguin necessari seguir-lo. O fins i tot, al ser un punt de reforç, alguns alumnes poden haver fet servir altres suports per reforçar conceptes.

Llavors, quins objectius podem tenir? Per una banda, ens interessarà buscar els perfils d'usuari, el fet que sigui optatiu, no descarta que hi hagi perfils d'usuari, més aviat al contrari, fins i tot podríem pensar en quatre grans grups d'usuaris en relació a la feina i la avaluació:

1) Fan la feina i la avaluació.

2) No fan la feina, però amb la avaluació comproven si dominen el temari.

3) Fan la feina, però no fan l'avaluació.

4) No fan res.

Les dades que ens han fet arribar tenen certes carències que s'haurien pogut solucionar amb una entrevista amb els creadors del curs o obtenint les dades per una altra banda:

a) No tenim constància de l'ordre dels problemes.

b) No coneixem la dificultat dels problemes.

c) Desconeixem l'ordre en que s'han de veure els vídeos.

d) No tenim tampoc les dades demogràfiques dels alumnes, com el sexe, edat, nivell social,…

e)

Tot plegat, afectarà clarament als nostres objectius:

1) Buscar els perfils d'usuari.

2) Buscar si hi ha relació entre la feina feta i l'avaluació.

3) Relacionar el temari amb les proves d'avaluació.

4) Estudiar el seguiment del usuaris en el MOC.

Anàlisi descriptiu

En aquest anàlisi serveix per a poder entendre millor les dades amb les que anirem a treballar. Primer analitzarem els atributs bàsics i després els específics.

Atributs bàsics

---

agent

Tenim multitud de variants. Pel nostre treball no tenen gaire interès.

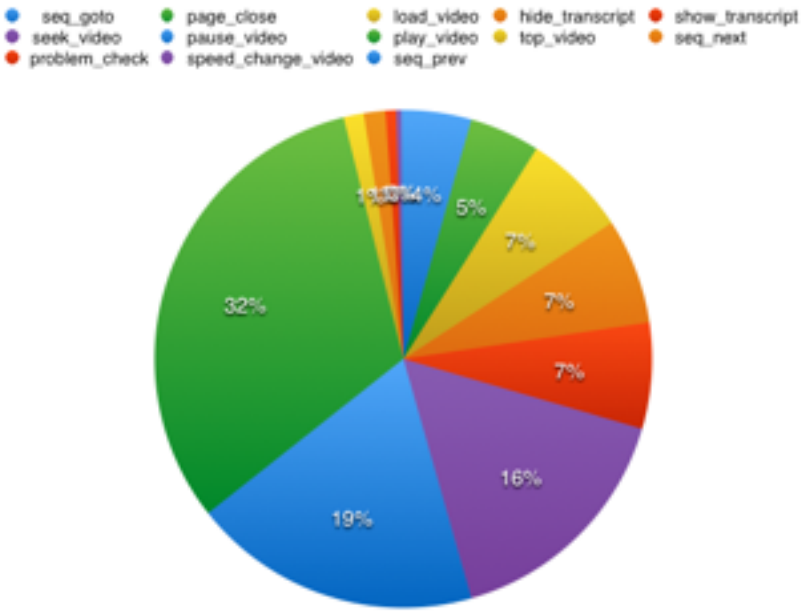
---

event_source

Només tenim dos possibilitats, server amb 140297 ítems i browser amb 181125. Són d'origen intern i no aporten gaire informació.

---

event_type



Només un 55% dels registres entregats inicialment tenen un event_type que nosaltres tractem. Entre tots ells, aquest és el pes relatiu de cadascun d'ells. Sent les accions del vídeo les que més registres acumulen.

---

host

Bàsicament hi ha quatre registres, d'entre ells ressalta www.ucatx.cat el qual té més del 99,5 % del total de registres. Després tenim ucatz-edx.upf.edu amb poc més del 0,3 % i d'altres encara amb menys registres com studio.ucatx.cat i 54.76.219.118.

---

ip

Hi ha una mica més de 2000 ips diferents i totes xifrades amb 32 bits.

---

name

Només hi ha tres registres:

enrollment.deactivated amb 19 registres , enrollment.activated amb 270 registres i edx.forumsearched amb 4 registres. Estan duplicats en el event_type, abans no els he posat.

---

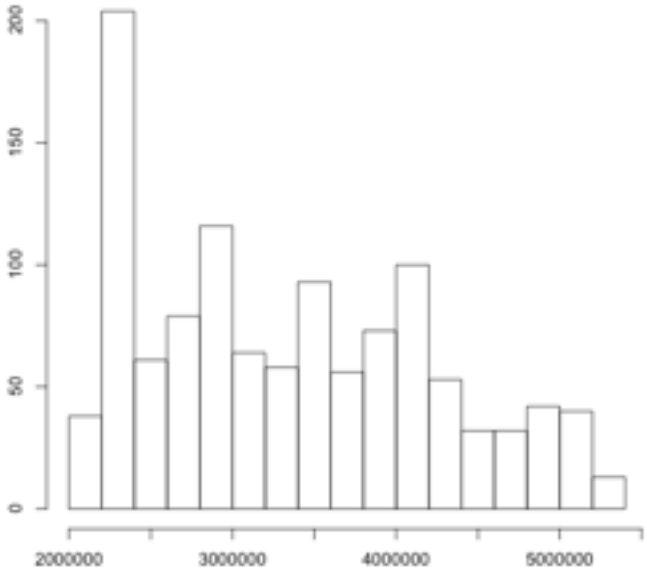
page

Diverses URL sense motius aparents d'extreure informació.

---

session

Hi ha un total de 1436 sessions diferents. Aquestes van de més a menys en la mesura que avança el curs, tot i així hi ha quatre pics importants.



En la mesura que avancem amb el nostre anàlisi descriptiu veurem que molts alumnes es van anar desapuntant. El primer reflex d'això és el nombre de sessions.

A més, les diferents avaluacions no segueixen un ordre, de tal manera, trobarem alumnes que ja han acabat tots els problemes i altres encara segueixen.

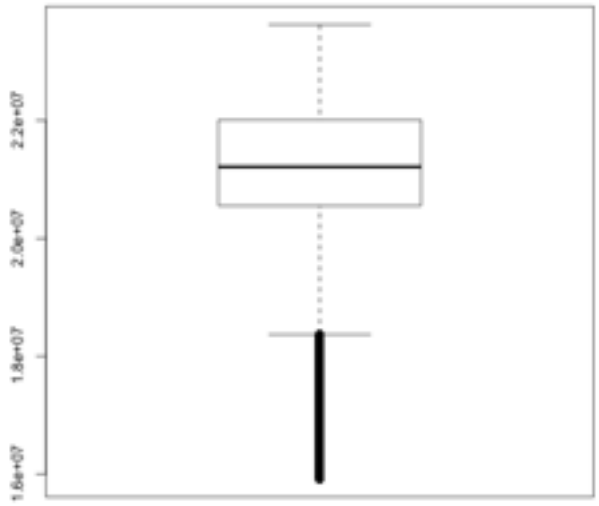
time

És el temps de quan s'efectua la operació, el format és el següent YYYY-MM-DDThh:mm:ss.xxxxxx.

Per tal de poder treballar amb el temps tinc una funció que m'ho transforma en segons des de l'1 de gener. Així doncs, si mirem el summary de time següent:

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

2161000 2548000 3249000 3352000 4028000 5264000
```

El podríem traduir per:



Min : 25 d'agost

1st Qu. : 29 d'agost

Median : 6 de setembre

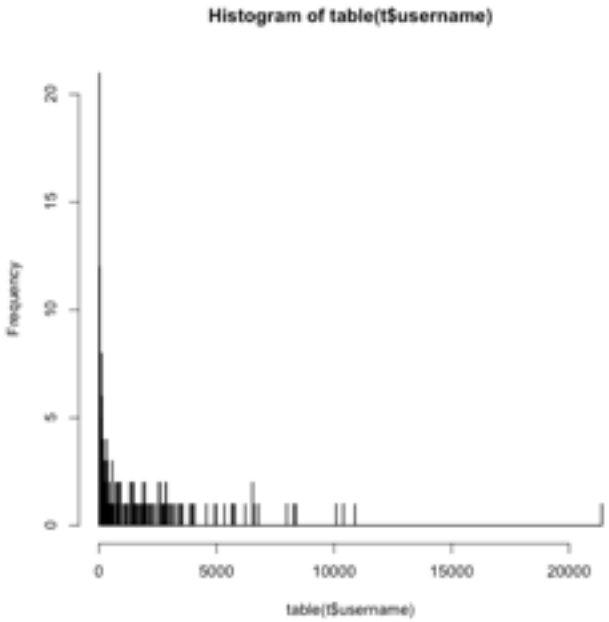
Mean : 7 de setembre

3rd Qu.: 15 de setembre

Max. : 29 de setembre.

Així doncs, podem observar com les observacions es concentren al principi del mes.

username



La major part dels usuaris van efectuar poques accions. Observant el gràfic, ens mostra les accions que han fet els usuaris. Podem observar com la majoria dels usuaris van interaccionar molt poc amb el curs.

En total tenim 241 usuaris. D'aquests, una mica menys del 30% fan menys de 100 interaccions amb la web. Per contra, una mica més d'un 40% dels usuaris ha interactuat més de mil vegades amb la web.

context.course_id

Hi ha una gran diversitat de cursos els quals després posarem a debat, doncs només un serà el que nosaltres treballarem. En total tenim 17 cursos diferents, però només un conté dades suficients com per ser tractat. És el curs UPF/C01/2014 i representa una mica més del 90% dels registres totals.

context.org_id

Aglutina els cursos segons la Universitat. Igual que abans, només tindrem un curs, i per tant, només tindrem una universitat.

path

Url que genera l'event.

user_id

Segueix la mateixa distribució que l'atribut username.

Els atributs dels vídeos

event.code

És un codi que dóna el propi servidor de youtube a tots els vídeos i ens serveix per identificar-los. Hi ha un total de 98 vídeos

En el summary podem observar com hi ha una forta diferència entre les interaccions que tenen alguns vídeos i les interaccions que tenen altres vídeos. El summary s'ha fet en relació a la quantitat de registres que contenen els 98 vídeos. Sent el vídeo amb el codi "b7xgknqkQk8" el que té menys interaccions amb tres registres, i el "b7xgknqkQk8" el que té més registres, amb 6787.

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

3.0   567.2  1098.0  1639.0  1927.0  6787.0
```

event.currentTime

És una marca de temps en segons del punt on s'efectua l'acció. Molt important per poder relacionar les accions que s'efectuen sobre un vídeo per part d'un alumne.

event.id

Fa la mateixa funció que event.code, però aquest cas el codi no el dóna el servidor youtube sinó el servidor del curs. Té exactament els mateixos nombre de registres que abans però canviant el nom.

La resta d'atributs són marques temporals, com new_time, o new_speed,…

Els atributs de la navegació

event.id

Identificació EDX. En total hi ha 30 registres.

event.new - event.old

Són dos atributs que tenen valors entre 1 i 14.

Els atributs dels problemes

answer

En aquest atribut tenim 50 columnes

De cadascun dels 7 problemes que tenim, tenim x exercicis, els quals contenen les respostes dels usuaris. En el primer problema de la UPF tenim 8 exercicis; en els problemes d'ABC tenim un exercici; en la resta de problemes tenim 10 exercicis.

Un exemple és:

"event.answers.i4x_UPF_C01_problem_28218eb1099b436fb84603e95942eb90_3_1"

On després d'"answer" podem observar el nom del problema , el qual també el tenim en problem_id, i després al final el 3 indica que és l'exercici 3.

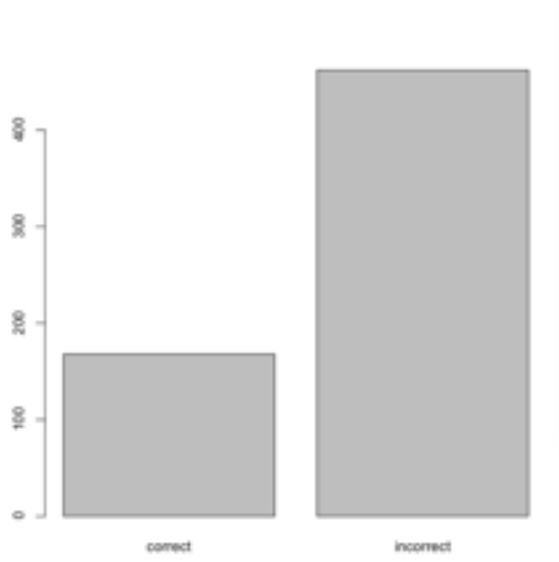
attempts

Els exercicis de la UPF tenen 3 intents. Cada intent es queda guardat en un registre. Els exercicis de l'ABC tenen 5 intents. Tots cinc queden guardats.

puntuació

La puntuació la podem trobar en els atributs grade, max_grade i success.

En grade tenim la puntuació del alumne, de 0 a 10. Però aquesta queda relacionada amb max_grade, que és la nota màxima que poden treure en un exercici. En el primer exercici de la UPF la nota màxima és un 8, en els exercicis de l'ABC és 1 i en la resta 10.



Per últim, a success s'ens mostra si l'exercici s'ha superat.En el gràfic de la dreta podem observar que el nombre d'intents incorrecte és molt superior al nombre d'intents correctes. Tot i que el mateix gràfic però només tenint en compte els registres del tercer intent queda molt més igualat: 88 correctes i 99 incorrectes.

problem_id

En l'anàlisi i síntesi de les dades entrarem més en aquest atribut. Al igual com ens passava amb el nom del curs, aquí també podem observar que els problemes relacionats amb el cursoUPF/C01/2014 són els que contenen més registres.

En total tenim 29 problemes i només els problemes relacionats amb UPF tenen més de 20 registres relacionats.

Anàlisi i síntesi de les dades

A partir d'aquest punt comença l'anàlisi seriós de les dades. L'anàlisi parteix del treball de síntesi que cal fer anteriorment, separant les dades pel seu camp.

Es tracta de fer un anàlisi posant èmfasi en la informació que ens en donen i no en el camp concret. Salvant les distància, ara ens toca fer un anàlisi orientat a objectes.

Començarem amb els cursos i acabarem analitzant els vídeos i els problemes. Tot aquest anàlisi és imprescindible per entendre el conjunt final de dades sobre el qual treballarem.

Els cursos

Tenim un total de 17 cursos. Repartits de forma molt desigual i dels quals en destaca un per sobre de la resta: "UPF/C01/2014", el qual conté 317.923 registres del total de 326.640 registres que tenen un curs associat.

D'aquests 17 n'hi ha 7 que tenen menys de 20 registres, i per tant, els eliminem dins d'aquest procés de neteja de dades, els cursos a eliminar són:

Quantitat de registres	
ABC/C02/2014	273
ABC/C03/2014	886
edX/Open_DemoX/	1539
UAB/E01/2014	1012
ucatx/C02/2014	48
UCATx/C04/2014	121
UOC/CN01/2016	623
UPF/C00/2014	3928

- ExempleX/CS101/2015_T1
- UPF/2014/about
- upf/c01/2014
- UPF/ucatx01upf/2014_T3
- UAB/C02/2014
- UPF/BLABLA/2014
- PF/UPF00/2013_T1
- UPF/ucatx/2014

Així doncs, el que farem és analitzar el paper que tenen tots els cursos restants, i per acabar, acabarem analitzant el curs principal.

Per a tots els cursos analitzarem el paper dels usuaris en el comput total dels registres, la seva interacció amb les unitats i els vídeos, i per últim, la realització de les proves.

ABC/C0/2014

Conté només 273 registres. En total hi ha 6 usuaris diferents, els quals entre tots han fet 7 sessions diferents. És a dir, només un usuari va entrar en dues sessions diferents.

El curs conté una prova "i4x://ABC/C02/problem/a4d50e6c2a3c4c7f983c3e681875b04c", el qual només un dels sis usuaris l'intenta fer.

El usuari que intenta fer la prova no és el que té més registres. El que té més registres en té 169, mentre que l'usuari que fa la prova només en té 58.

El problema té una única pregunta: "What Apple device competed with the portable CD player?". El usuari fa cinc contestacions:"The iPad","The iPod", "The iPod","The vegetable peeler", "The iPod". D'aquestes contestacions la correcta és Ipod.

No hi ha més preguntes, per tant, dóna la sensació que és un curs de prova. Per tot això, l'eliminarem del registre.

ABC/C03/2014

Aquest curs conté 886 registres. En total hi tornen a haver-hi 6 usuaris, els quals tornen a ser els mateixos que el curs ABC/C02/2014. En aquesta ocasió tenim un total de només 8 sessions.

Al igual que abans, el curs només conté una prova: "i4x://ABC/C03/problem/3f0bb393f08849f7b8b52f07330dffc3". Aquest problema, tot i tenir un id diferent, és exactament igual que ABC/C02, té la mateixa pregunta i les mateixes possibles respostes.

Així que entendrem que és part de les proves que es fan i ho eliminarem del registre.

ucatx/C02/2014

Aquest curs conté 48 registres i només conté tres usuaris. Els quals ja sortien en els cursos anteriors. Per tant, entenem que és part dels cursos de proves que hi ha i que caldria eliminar.

edx/Open_DemoX/edx_demo_course

És el segon curs amb més registres, amb un total de 1539 registres. Tenim un total de 16 usuaris, dels quals, al igual que ens està passant tota l'estona, els usuaris que tenen més registres són els mateixos que estem trobant en els cursos anteriors.

Tot i així, hi ha 16 exercicis diferents.Però els valors que tenen en els problemes no tenen sentit, o estan en blanc o tenen preguntes sense gaire sentit. A més, el fet de dir-se el curs com a demo, dóna l'efecte que també és un curs de proves. Així doncs, l'eliminarem amb la resta.

UAB/E01/2014

Aquest curs té 1012 registres, amb un total de 64 usuaris. Tot i així, dels 1012 registres 477 són d'un sol usuari que el teníem en els cursos anteriors i 397 d'un altre usuari que també surt dins del registre.

Els usuaris són: "d41d8cd98f00b204e9800998ecf8427e" i "1253208465b1efa876f982d8a9e73eef". Tots dos tenen el 86% dels registres.

A més, si mirem els registres del event_type ens trobem que són events relacionats amb el control i gestió del curs.

En conclusió, els dos usuaris poden ser consultors o alumnes de proves, i alhora, aquest curs també haurà de ser eliminar.

UCATx/C04/2014

És un altre curs amb pocs registres. En total en té 121 registres i quatre usuaris diferents. Molts d'ells es repeteixen i sense cap problema associat. Sembla doncs que segueix sent un curs de proves el qual eliminarem.

UOC/CN01/2016

Té un total de 623 registres. Té quatre usuaris ja repetits anteriorment i no conté cap problema associat. Així que igual que amb la resta l'eliminem.

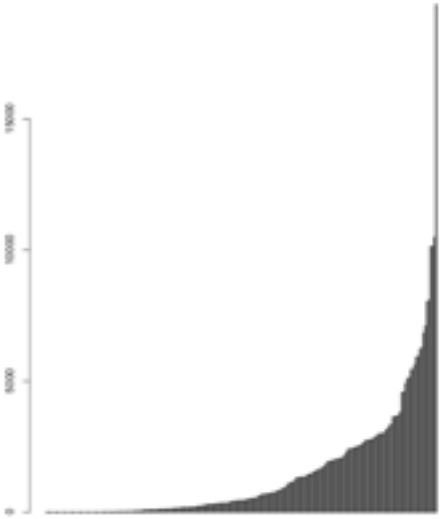
UPF/C00/2014

Conté 13 usuaris, els que tenen més registres ja els teníem anteriorment. Tenen quatre problemes associats, però les preguntes segueixen o sense ser-hi o sense tenir sentit. Així que també ho eliminarem.

Arribats en aquest punt hem descartat tota la resta de cursos. A partir d'ara només ens centrarem amb els registres que estan relacionats amb el curs UPF/C01/2014.

Els usuaris

En total hi ha 317923 registres i tenim 218 usuaris, amb una quantitat de registres molt divers entre usuaris.



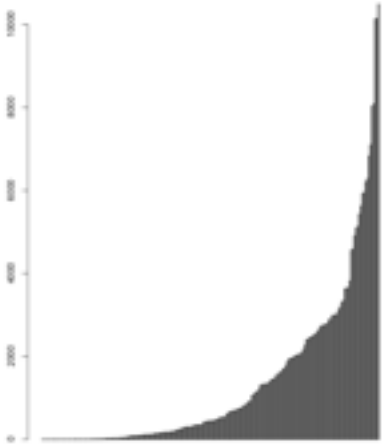
Per una banda destaca que gairebé la meitat dels usuaris tenen un nombre de registres molt baix, mentre que per l'altra banda, podem observar en el gràfic de la dreta com un usuari conté més del doble de registres que el segon.

Aquest usuari és un dels que ja veiem en els cursos anteiors, i per tant, cal suposar que fa la tasca d'instructor. Tal és així, que si mirem els registres que té en el event_type, té registres d'accés on clarament surt el nom de "instructor", com per exemple:

"/courses/UPF/C01/2014/instructor/api/list_instructor_tasks".

Gràcies a trobar aquest tipus d'event_type, ara el podem utilitzar per a trobar a la resta d'usuaris que facin també funcions d'administrador. En total tenim 7 usuaris que han accedit a la llista de tasques del instructor. El nombre de registres d'aquests usuaris és diferent:

```
4394dbe8d222464509f79c7c7800f4b2  ->   70

21232f297a57a5a743894a0e4a801fc3  ->   174

e981764ab5dea3dc11790782165caf60  ->   335

92c3f7069ee15c6ad932bd7ac98d6b4a  ->   337

d41d8cd98f00b204e9800998ecf8427e  ->   1778

0e850994e2b926894579823317240b7c  ->   3363

1253208465b1efa876f982d8a9e73eef  ->   12874
```



Tots aquests usuaris queden eliminar. En el gràfic de la dreta es pot veure el resultat d'haver-los eliminat. Ara ja no només tenim 211 usuaris.

El temps en els registre

Per començar a analitzar el temps de quan s'han fet els registres, podem observar els resultat dels quartils:

Min -> 15950000

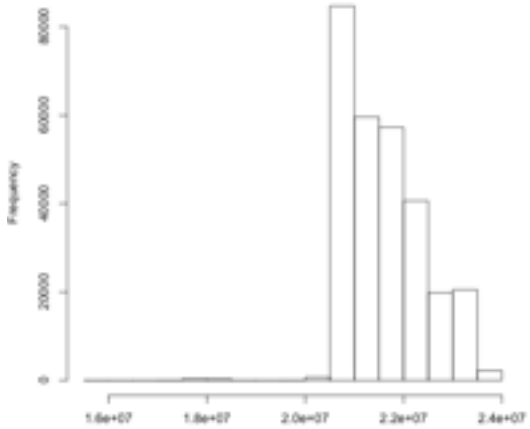
1st Quartil -> 20860000

Median -> 21470000

Mean -> 21590000

3rd Quartil -> 22240000

Max -> 23620000



També podem observar l'histograma per observar-ne la freqüència. El temps està en segons a partir de l'1 de gener del 2014.

En l'histograma podem observar com el curs comença aproximadament a partir del segon 20.000.000. Però ja abans tenim alguns registres. En aquest cas, els registres anteriors no els esborrarem. Després d'observar el event_type no sembla que tinguin molta repercussió, sinó que més aviat són usuaris que es registren i que després miren una mica la web, a l'espera que comenci el curs.

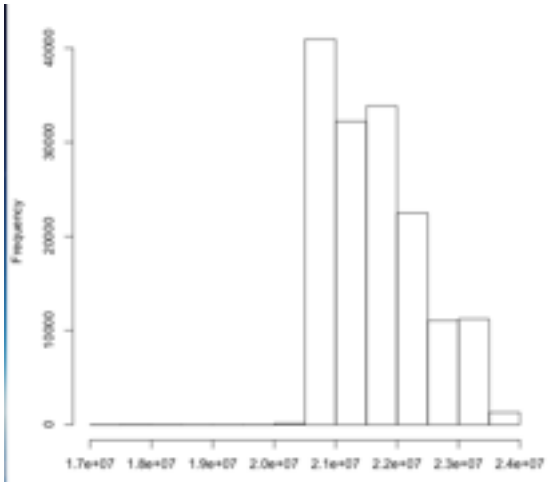
Sí que podem observar en el histograma, que els usuaris es connecten més al principi de curs, marcant una clara davallada durant les setmanes següents. Al final del curs només s'efectuen un 25% dels registres totals que es feien al principi de curs.

Cada columna representa aproximadament entre 5 o 6 dies reals. El curs dura un mes i mig, i si tenim que en total hi ha 286059 registres, podem observar com quasi el 25% dels registres es fan els primers cinc dies. I en les tres primeres setmanes ja s'han fet el 71% dels registres.

Aquest fet distorsionarà molt qualsevol anàlisi, doncs la falta de regularitat en els registres ens obligarà a buscar estratègia per superar-ho, agafant per exemple, només els usuaris que acabin el curs.

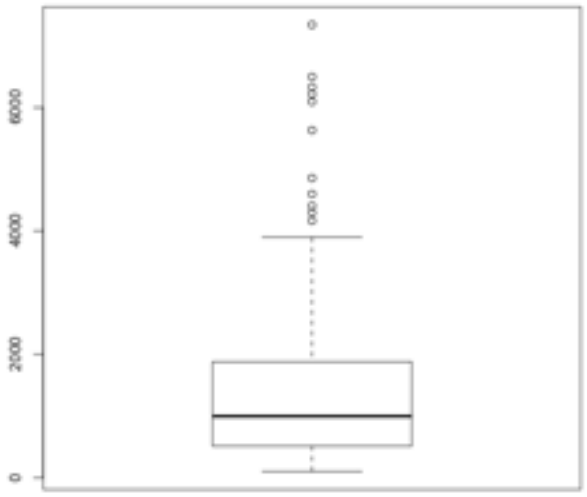
Els vídeos

Si analitzem el temps de quan es fan els registres podem observar com la evolució és força similar a la dels registres en general. Cal tenir en compte que quasi la meitat dels registres que té el curs estan directament relacions amb els vídeos.



En total hi ha 96 vídeos diferents. El nombre de registres que ha provocat cadascun dels vídeos és força dispar.

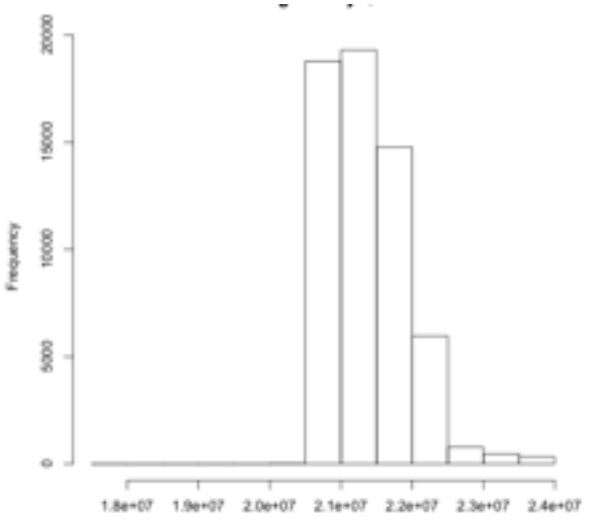
No tots els vídeos han provocat el mateix nombre de registres. Podem observar com hi ha una gran dispersió, i podem observar diversos puntets en el gràfic (BOXPLOT). Aquests punts contrastant amb la major part dels vídeos que tenen relativament molt pocs registres.



Els vídeos més vistos

En aquest apartat analitzarem els vídeos que en el gràfic han sortit per sobre dels 4000 registres.

Si ho comparem amb el histograma anterior, podem observar com l'histograma dels vídeos més vistos correspon amb els temps on va haver-hi més tràfic. Així doncs aquests 11 vídeos poden arribar a representar entre el 50 i el 75 per cent del total dels registres que fan els vídeos.



Però amb la evolució del temps ja podem observar que hi ha una forta relació pel fet de ser vídeos que es visualitzen al principi de curs. No tenim informació sobre quina hauria de ser la posició en que s'hauria de veure cadascun dels vídeos, així que no podrem comparar si aquests eren els primers que s'havien de veure.

El que sí que serà veritat, és que el fet que tothom els vegi al principi fa que tingui més audiència que la resta.

Analitzar la densitat de visió d'un vídeo

Un dels processos que s'ha intentat fer és intentar estudiar quina era la densitat de visions de cada vídeo en cada segon. La densitat la trobàvem calculant la quantitat de vegades que s'havia vist un segon. Però això ha estat impossible ja que el log amb les dades no quadra en molts punts, poso un exemple d'un usuari concret, en un vídeo concret i durant una sessió concreta:

event_type	time	currentTime	old_time	new_time
load_video	21647832	NA	NA	NA
play_video	2167843	0	NA	NA
pause_video	21647850	0	NA	NA
play_video	21647850	0	NA	NA
pause_video	21647883	48,88	NA	NA
play_video	21647899	48,88	NA	NA
seek_video	21647902	NA	53,25	71
play_video	21647902	53,25	NA	NA
pause_video	21647934	118,36	NA	NA

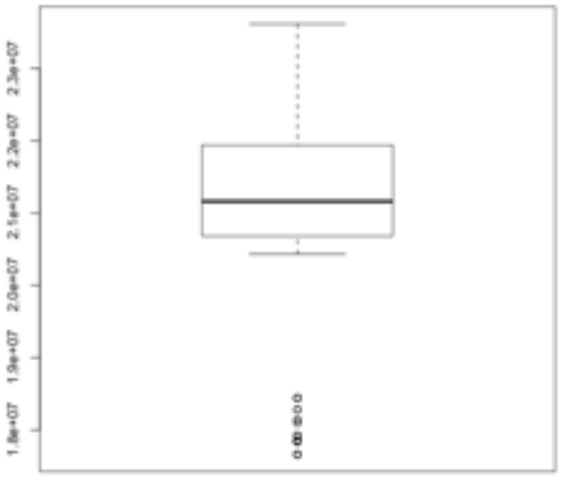
Si s'observa la taula, després del seek no concorda. No queda ben lligat quin és el punt de retrobada, ni 71 ni 53,25 són punts correctes, doncs si mirem el time només passen 32 segons des del seek i el play, fins al pause.

Per tot això, el que farem serà recollir els segons que dedica un usuari a un vídeo en una sessió. Però el problema no queda aquí. Doncs si agafem els temps de les sessions ens podem endur la sorpresa que en una mateixa sessió el temps que hagi passat sigui al voltant de 30000 segons, unes 8 hores.

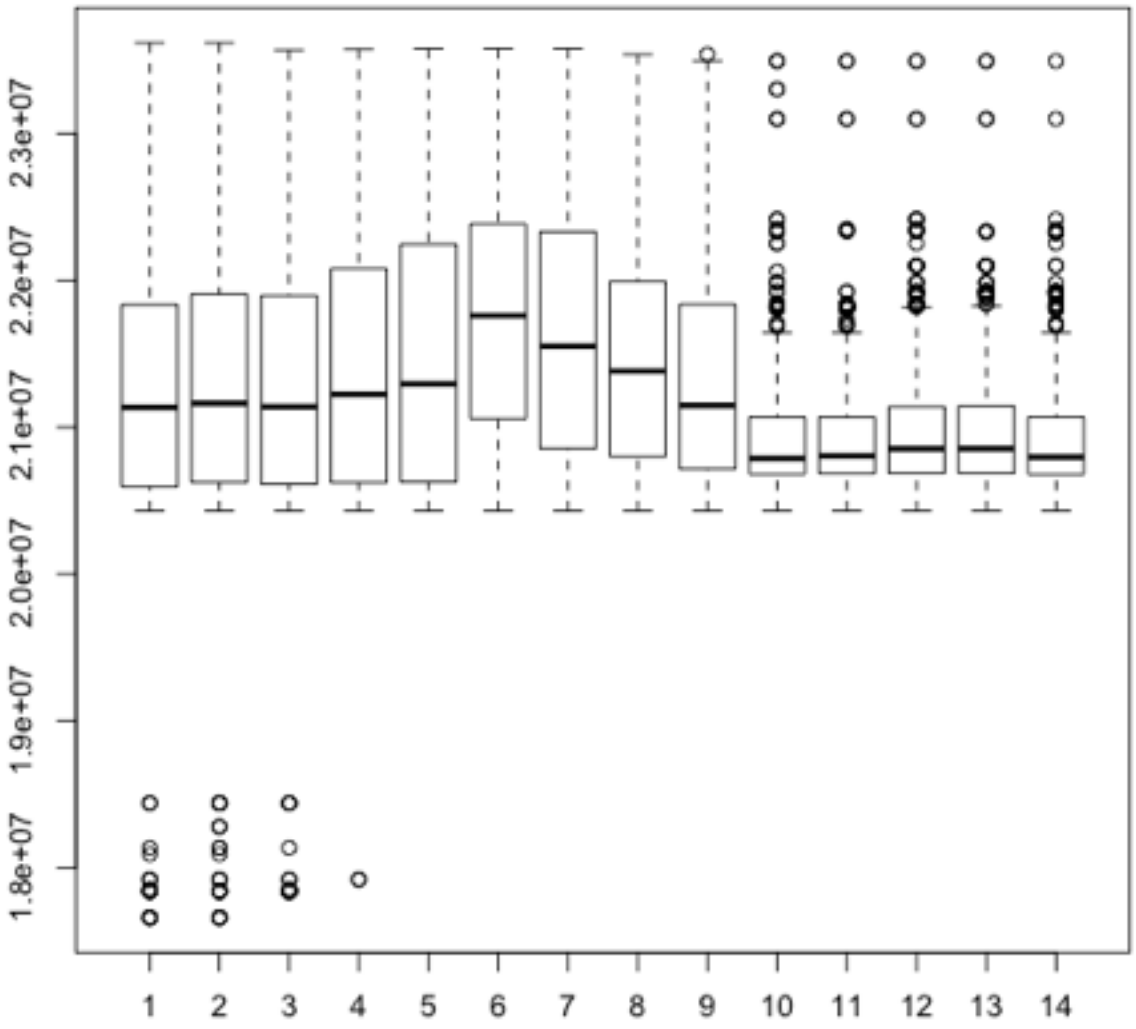
Per evitar aquest fet, més endavant entrarem més a fons en el treball dels vídeos i el que s'ha fet és posar un tope de 10 minuts. En el moment que un usuari no interacciona amb un vídeo durant 10 minuts, s'ha considerat com una nova sessió.

Els temes / lliçons

En total tenim 11858 registres. Si analitzem el temps on es van fer aquests registres, podem observar com tenim registres en un temps inferior a l'inici del curs. Segurament es degui a alumnes que comencessin a mirar alguna part del temari, abans que comenci el curs específicament.



Resulta també interessant observar els boxplots de cadascuna de les unitats en relació al temps que s'han efectuat els seus registres.



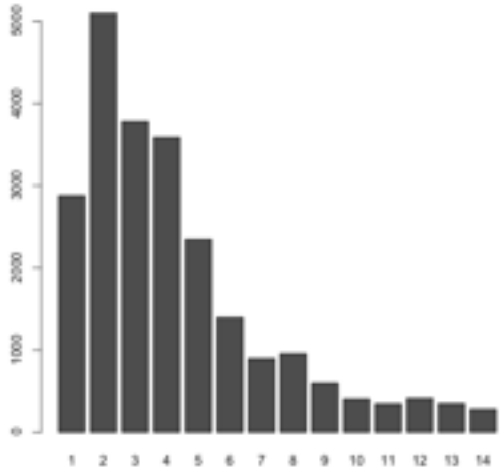
La lògica ens diu que hi hauria d'haver una relació estable entre les unitats i el temps, de tal manera que en la mesura que va passant el temps els estudiants entren més a les unitats

finals. Això succeïx una mica fins a la unitat 6, a partir de llavors sembla que els registres baixen força i no podem seguir aquest procés.

Aquesta interrupció es pot deure a diversos factors, o que els usuaris només mirin el temari al principi del curs, o que la davallada d'usuaris i registres general que hi ha a partir de la segona setmana del curs afecti a l'anàlisi.

Amb la idea d'analitzar aquest segon fet, podem observar l'histograma de registres que genera cada unitat. Que en la unitat 1 hi hagi menys registres té molta lògica, doncs nosaltres agafem aquests valors del "event.old" i "event.new", aquests valors són creats pels usuaris quan es mouen per les unitats. Per tant, és lògic pensar que l'1, al estar a una punta, no tingui tants registres.

Si mirem l'histograma podem veure que fins a les unitat 5 els registres es mantenen una mica, mentre que a partir de la unitat 6 aquests comencen a decaure fins a valors ridículs en comparació amb les primeres unitats.

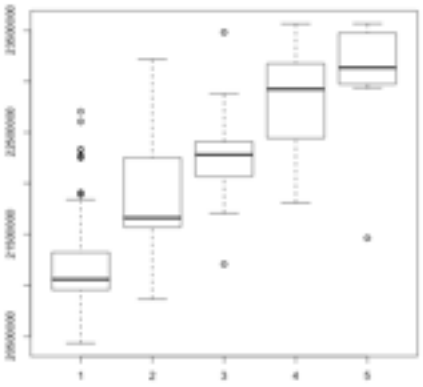


Si ens hi fixem, aquesta davallada és en el mateix punt en que el boxplot de totes les unitats començava a no tenir sentit. Així que tindria força lògica

Els problemes

En total tenim 5 problemes diferents. Cadascun d'aquests problemes està compost per diverses preguntes i només aquells alumnes que aconsegueixen superar totes les preguntes aconsegueixen l'aprovat.

Les cinc proves també tenen un nivell de registres dispers. Nosaltres només analitzarem aquells registres que estiguin relacionats amb el cheking del problema, i que per tant, han comportat un intent per part de l'usuari d'aprovar el problema.



Cada registre que nosaltres tenim representa que un alumne ha fet un intent de resoldre el problema. El màxim d'intents és 3.

Com que no tenim l'ordre de les proves, el que fem és aplicar el mateix gràfic de capses que abans i així li donem un ordre. Podem observar en el boxplot que l'ordre és veu força clar. I per tant, podem pensar en un curs que a cada cert temps s'ha de superar una petita prova.

D'aquesta manera tenim l'ordre següent amb el total de registres:

1) d97b605bcbd04e58a8fd23a7218d9bad -> 236 registres

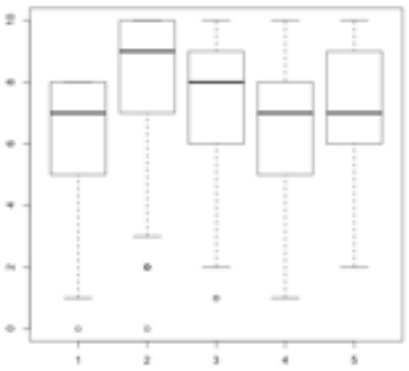
2) 7ec2a5037e7947a49ecfce13700acc1d -> 172 registres

3) 9010e4a261cd4d8ba64bf3c92fb63a09 -> 112 registres

4) 28218eb1099b436fb84603e95942eb90 -> 63 registres

5) 896aedaacc864d84a227b23307865562 -> 41 registres

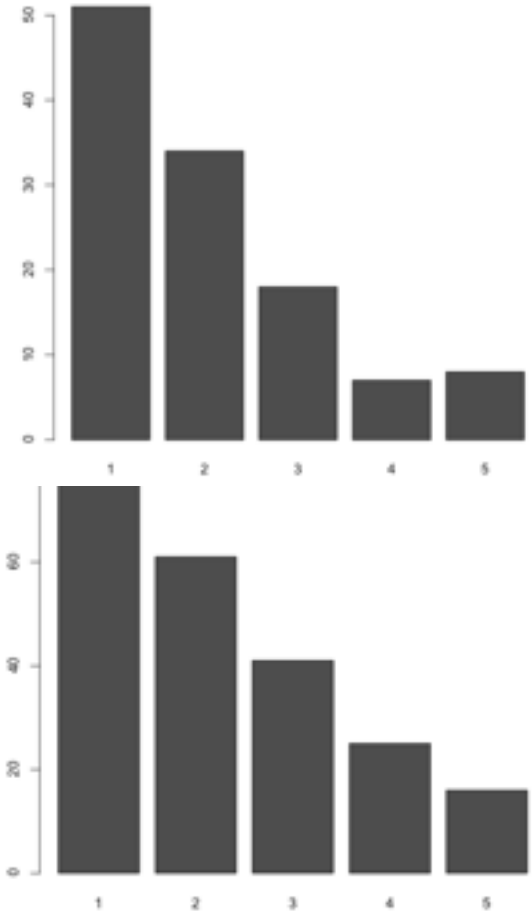
Seguim observant doncs, que hi ha una davallada general que també es plasma en la realització dels exercicis.



Una vegada tenim els problemes, podem mirar a través del histograma com ha variat la nota dels exercicis. En tots, excepte el primer que és 8, la nota màxima és un 10.

Del gràfic podem extreure que els exercicis més fàcils van ser el primer i el segon, mentre que la resta no van obtenir tan bon resultat.

En el gràfic superior de la dreta podem observar un histograma amb el nombre d'alumne que han aconseguit aprovar el problema. Podem observar com on tenim menys aprovats és en el exercici 4, on només l'aproven 7 usuaris.



En el gràfic inferior podem observar el nombre d'usuaris que van intentar superar el problema. El percentatge d'aprovats respecte els usuaris que ho intenten va anar disminuint a cada prova que passava, exceptuant la última(59%, 53%, 43%, 28%, 50%).

Al final, només són 16 usuaris els que intentaran fer la cinquena prova, poc més del 7% del total d'usuaris que inicien el curs. També és veritat que aquest percentatge es manté baix durant tot el curs, doncs només un 40% farà la primera prova(Evolució: 40%, 28%, 19%, 11% i finalment 7%).

Els 16 magnífics

Dels 16 estudiants que acaben el curs, nomes quatre superen amb èxit totes les proves.

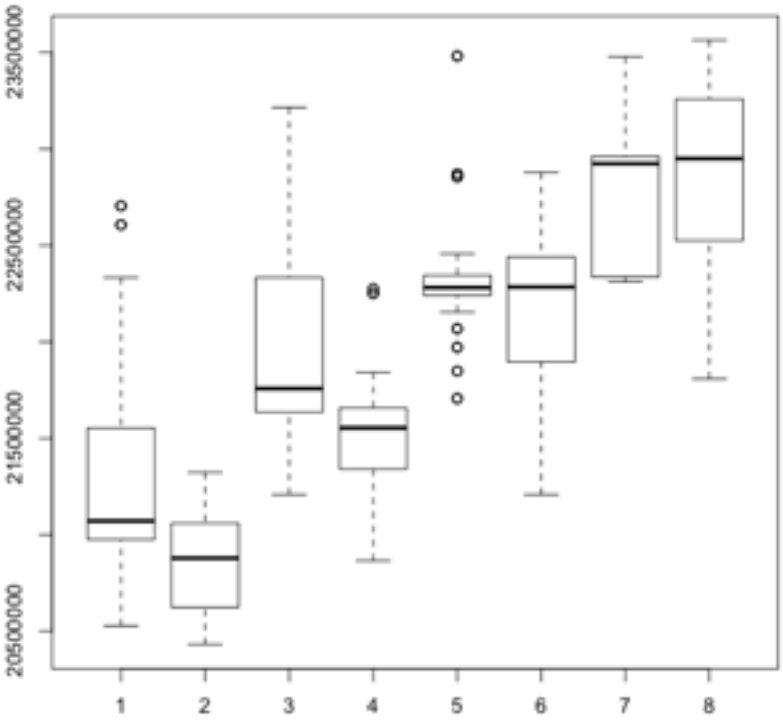
La evolució dels que ho tenen tot superat és la següent: 51, 24, 11, 6 i finalment 4.

En conclusió, no ens serviria gaire agafar els quatre usuaris que van acabar, doncs són massa pocs per aplicar qualsevol tipus d'estadística. El que sí que podem agafar són el grup d'usuaris que va fer les cinc proves.

Aquest és un total de 16 alumnes, i el que es tractarà ara és de buscar informació referent aquest grup, el qual se suposa que hauria de ser un grup més involucrat que la resta d'usuaris.

Els 16 i els problemes

Ja en un primer moment podem veure com aquests 16 usuaris treballaran una mica diferent a la resta d'usuaris. Si fem un gràfic de capses i comparem en cada exercici els 16 magnífics amb els altres 70 usuaris que van començar a fer algun problema ens trobarem amb el següent gràfic:



El gràfic funciona de la següent manera, cada problema ocupa dos capses, les capses 1 i 2, són del problema 1, les capses 3 i 4 capses són del problema 2, i així successivament fins al problema 4. En la capsa imparella tenim tots la capsa de tots els usuaris, en la capa parella tenim la capsa creada pels 16.

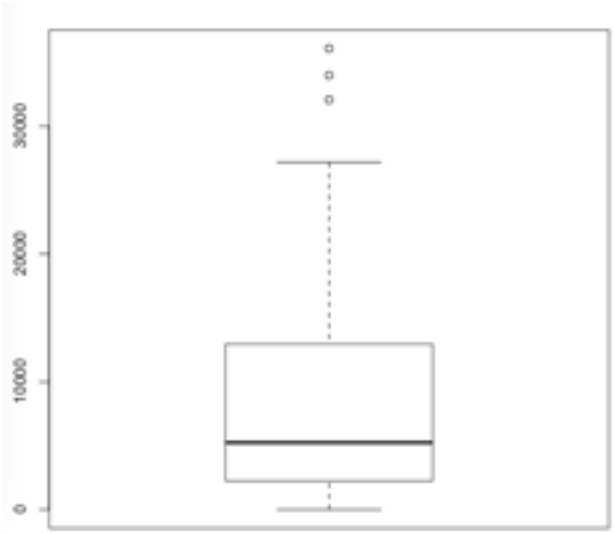
La gràcia és que podem observar com els 16 usuaris que acaben el curs es diferencien de la resta ja que acostumen a fer els exercicis una mica abans que la resta de la classe.

Aquesta diferència es veu molt clara en els problemes 1 i 2, mentre que en els 3 i 4 no tant.

Una altra característica important a destacar és que el grup dels 16 sembla tenir una repartició més harmònica del temps, sembla que les caixes puguin encaixar molt millor que la resta.

Els 16 i els vídeos

Un dels aspectes que crida l'atenció és la poca importància que semblen tenir els vídeos. Un esperaria que el temps que es dedica als vídeos sigui, en part, força homogeni, però no sembla que aquesta sigui la tònica.



En el gràfic de la dreta podem observar un diagrama de capses amb tots els vídeos i el total de temps que li han dedicat els 16 usuaris. Degut al problema que comentàvem, en aquest exercici s'ha trencat les sessions que entre una acció i just l'acció següent de l'usuari ha passat més de deu minuts. En aquest punt, s'han creat dues sessions independent alhora de comptabilitzar el temps.

El gràfic recorda molt al gràfic que ja teníem quan fèiem l'anàlisi dels vídeos que veuen tots els usuaris.

Hi ha tres vídeos que sobresurten per dalt, que són:

"M_1hwM7aZKY" amb 36089 segons.

"EczAxW0QoNI" amb 33992 segons.

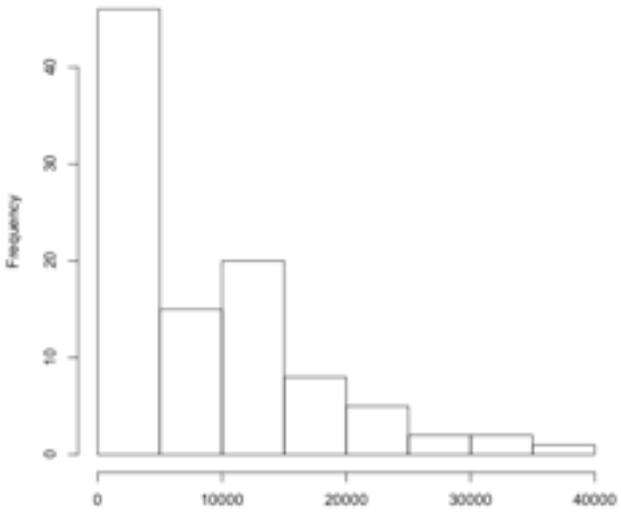
"mp_PKrHR6B8" amb 32078 segons.

Per contra, en tenim dos que sobresurten per baix:

"Dd1iE_NtBho" amb 1 segon.

"hoZ8lZpzvwk" amb 3 segons.

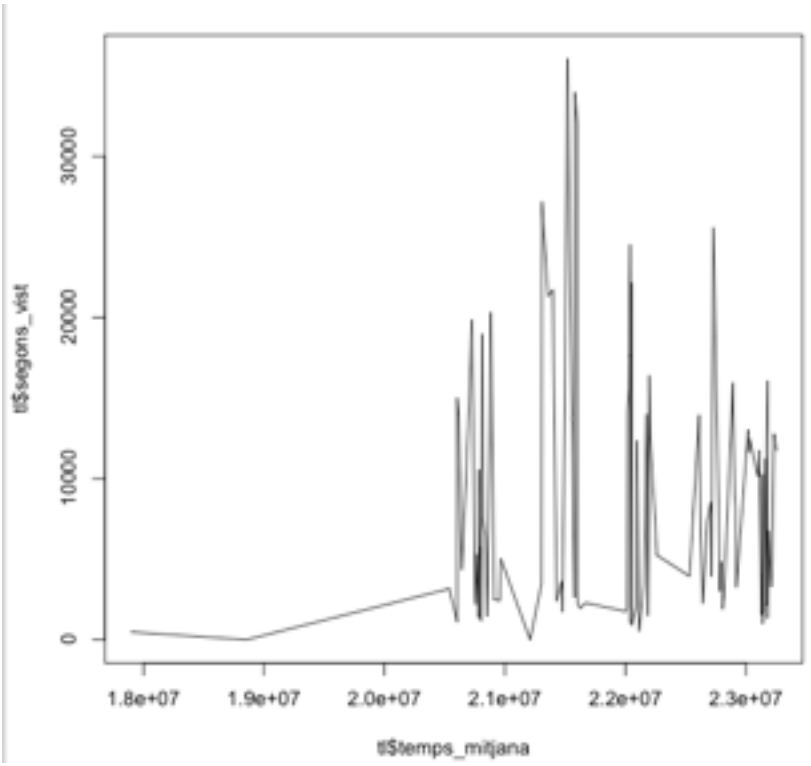
Així doncs, aquells vídeos més vistos han comportat uns 30 minuts de mitjana als 16 usuaris. Per contra podem veure un gran nombre de vídeos que no han acumulat gaire temps.



Si fem l'anàlisi de la mediana la tenim a 5248 segons, uns cinc minuts. Així que al final el que podem veure és que tenim una part de vídeos que no se li dedica ni 5 minuts de mitjana, i per contra, una altra meitat de vídeos que es dedica més de cinc minuts de mitjana.

Entendre aquesta disparitat ha estat un dels reptes més importants d'aquest treball i la clau ha estat agafar aquests vídeos i fer algú similar al que fèiem amb les unitats, buscar en quin moment es van veure i buscar a veure si hi ha algun tipus de relació.

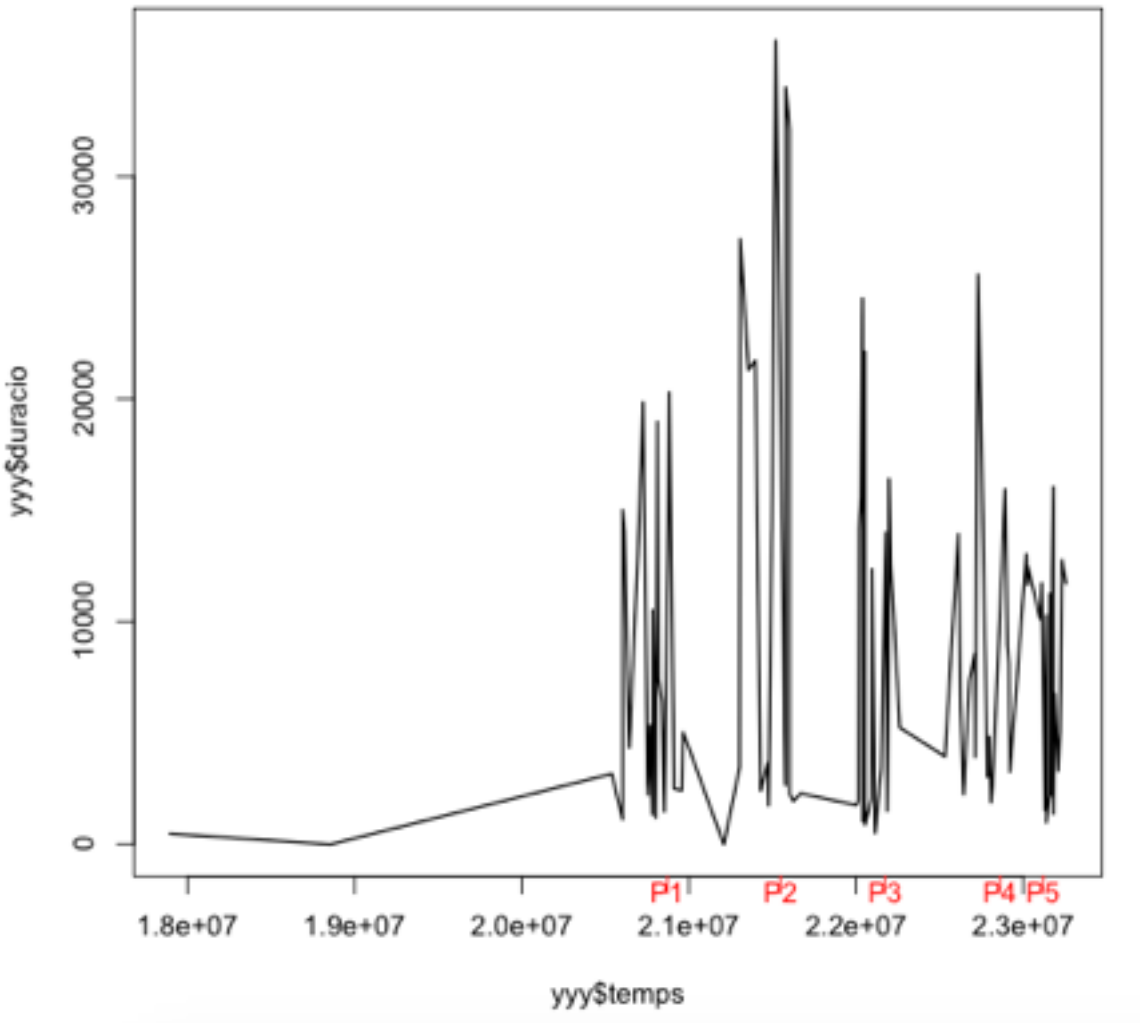
La solució és el gràfic lineal següent. En el eix de la abscissa tenim la mitjana de temps de quan es van veure els vídeos. Per exemple, podríem dir, aquest vídeo es va veure de mitjana el 5 de setembre. A l'eix de la ordenada, hi situem el nombre de segons que s'ha vist en total un vídeo. En l'exemple anterior, ara diríem, es va veure de mitjana el 5 de setembre i en total el van veure 15.000 segons.



Fet això el resultat és esplèndid. Podem veure que hi ha quatre etapes de visualització de vídeo, i tres etapes intermitges on no es visualitzen gaires vídeos.

La clau està en unir aquests vídeos amb la data mitjana de quan els 16 van fer els problemes, i observarem que les tres primeres etapes estan directament relacionades amb tres dels problemes:

```
 [,1]     [,2]

[1,]    1 20872054

[2,]    2 21548713

[3,]    3 22174034

[4,]    4 22861901

[5,]    5 23119089
```



Per tant, els 16 usuaris van veient els vídeos en la mesura que ho necessiten per fer els exercicis. Aquest fet ens permet fer una diferenciació dels vídeos força interessant, doncs ara podem diferenciar els vídeos que els usuaris van veure perquè el crien útil per resoldre el problema, dels vídeos que no van creure útils.

Només en els problemes 4 i 5 aquesta diferència no és prou clara, per això, alhora de fer la selecció dels vídeos s'ha decidit adjuntar-ho.

Per fer la divisió entre els vídeos que els usuaris troben útils, nosaltres hem fet servir la barrera dels 8.000 segons (una mica més de 8 minuts de mitjana per usuari).

Problema 1

code	segons vist	code	segons vist
BU_GRenoMMY	20300	hoZ8lZpzvwk	3
Lj66BSDhrs0	19849	DZuQCcSM3o0	482
9O6uIPmccEM	18979	GTtdpWjasxU	1116
Ck4OsRIoBWg	15020	V7PGMERwoZ8	1190
Vc_GC7ggAEo	14413	cXMjnGHpKQc	1341
6zyhrg_3OgU	14064	Qd2A3_TjWn0	1468
bHHSPzO5b4s	12865	HPHxvuts6jc	1547
zLCqll1EXPY	10538	joeYFnGt6OE	2218
		FhbGpydLg4M	2426
		jEU_ypg890I	2502
		Ek8_onLK2nI	2700
		6_zn4WCeX0o	4347
		6Yx1d61UbDw	5044
		Abkh3XF7Alo	5086
		YTti3PmBBwU	5332
		B6TigUmu7X4	5761
		fgjuc6g18N4	6433
		igOYd1a_iTc	7460

Problema 2

code	segons vist	code	segons vist
M_1hwM7aZKY	36089	Dd1iE_NtBho	1
EczAxW0QoNI	33992	q1EcsSLjoy0	1747
mp_PKrHR6B8	32078	JzM_ThSkFJQ	1950
dxvHaoj76ic	27182	GCzf03XhiYE	2259
yx70uoP4KtQ	21732	gCxi5ZgfAiE	2293
jMjkq96q4g0	21308	if0wJ8_ZNcM	2406
GBMp_Pifm0w	14064	svI3Kq5JSKQ	2647
		vsfQ6RM17bk	2782
		DZKeDwb7vIk	3456
		8_0klo0ge44	3658

Problema 3

code	segons vist	code	segons vist
Qv0DZFSIaHc	24516	QG_A54VzPpw	510
dxc3yS4bwXc	22126	bDHLYrfpYVY	879
4wJfHVt0zU4	17597	a9gu6GttvTg	1039
qHgLG4CnWcU	16400	n7Duojao3z4	1497
sX7bTwVrgFU	15571	1N1sFnMFSH0	1762
grgITlPHYvo	14167	Pq7zs9VYS_I	1970
QS3mSoDavIE	14004	QGcMvbMpOUk	2077
uhYfep5RBYE	12565	cqqFwj9F2mM	3399
rrpSQSR_TIY	12361	lhDHe0p6P1Q	5248
		YjZwIhpLkGg	7825

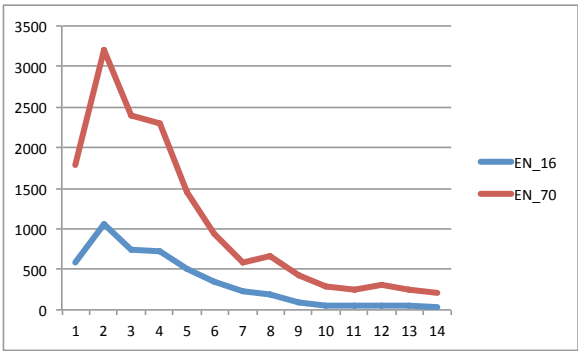
Problema 4

code	segons vist	code	segons vist
yS6h8d65KsA	25594	89dy86zYUIY	979
15NoeJZv638	16064	f_bCNKdbYL8	1368
vkWM_rMSv_c	15945	Ee_qeG_AAso	1540
m2ItrCT6bWM	13940	Nbt6Bf2S8Bs	1892
i0EepE_lDe8	13039	1XmFyGxzRk8	2206
pw1NOdpfTQw	12774	CzdkZSIXXso	2218
cdjSjXaijy8	12455	Eysn011TDjg	2249
NrYk45gGSJA	12290	VWd2pSJTICY	2579
xoQfS_jvEwU	11744	x6iDdVOWxBM	3023
Lol_Aci1pyo	11707	1iBEE30nO38	3256
YbCIuadfArY	11647	gMAYzGSSZXQ	3298
un3WAEN9G7o	11246	7UnxGrZWUbU	3917
zEiUSQAfepk	10292	WRxYrNwTzmk	3947
6NU4xj_eVy4	10112	vY0E46XQXHw	3979
_rWVuJ0mVAE	8952	ep6VxUo5NAM	4834
_mtVl0CxMDY	8562	NY2h7TIXjko	5035
KZboOJgFACk	8117	ylAJicpLSsQ	6487
		jNNr_K7BAJU	6772
		kbX1vkrjX1g	7374

En conclusió, si nosaltres volem agafar la totalitat dels vídeos i comparar-los, no funcionarà, doncs la major part dels vídeos pels usuaris no van ser importants. En canvi, si observem els vídeos més vistos en cadascun dels problemes sí que ens podem fer una idea de la importància d'aquests alhora de fer els problemes. En total tenim uns 40 vídeos que els usuaris veuen importants per a fer els seus exercicis.

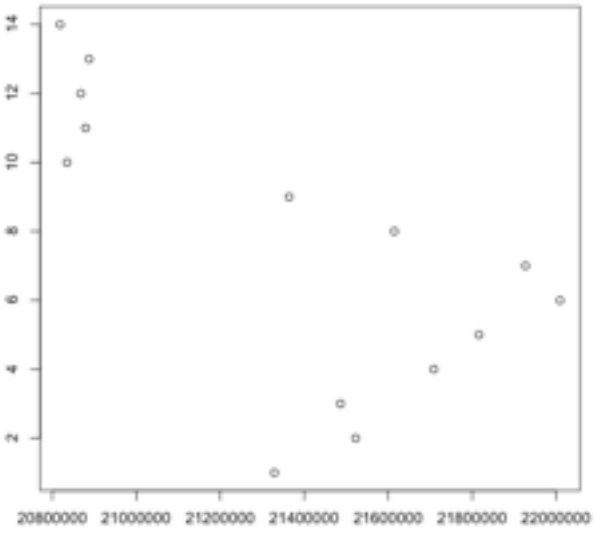
Els 16 i les unitats

Un dels aspectes que caldria esperar algun tipus de diferència és amb el tractament de les unitats. Esperaríem que el grup dels 16 hagués treballat més les unitats, o les hagués treballat diferent. Però sembla que les dades no ens ho acaben de mostrar.



Si ens fixem amb el gràfic de la dreta, podem observar el nombre de registres que provoca cadascuna de les unitats. Observem com la evolució és bàsicament idèntica. Hi ha un boom inicial que en la mesura que passen les unitats es va apagant. Això hauria estat d'esperar del grup que va abandonant, però no del grup que es queda fins al final.

Fins i tot, si fem un anàlisi com el d'abans, hauríem de trobar alguna lògica entre el temps que es visita una unitat i el temps. Si fem el mateix gràfic observem el resultat.



Fixem-nos, que per les unitats més altes, de la 14 a la 10, els valors mitjans de visita es situen al principi del curs, quan els estudiants comencen a visitar el curs. Només veiem una mica de lògica entre les unitats 1 i la unitat 7.

A partir de la unitat 7 aquest procés ascendent s'atura. Si recordem els valors del temps mitjà dels problemes -> 1: 20872054, 2: 21548713, 3: 22174034, 4: 22861901, 5: 23119089.

Veiem que les unitats no segueixen cap ordre en relació als problemes. Si volguéssim fer alguna cosa similar al que hem fet abans amb els vídeos ens sortiria una taula així.

problema	mitjana problema	unitats
1	20872054	10, 11, 12, 13, 14
2	21548713	1, 2, 3, 9
3	22174034	4, 5, 8
4	22861901	6, 7
5	23119089	-

Ja veiem que no s'estableix cap tipus de relació lògica amb la qual en puguem treure conclusions.

Conclusions anàlisi i síntesi de les dades

Les principals conclusions que podem extreure són:

1) El nombre de registres disminueix força en la mesura que avança el curs.

2) Els vídeos, les unitats i els problemes amb més audiència són els que estan al principi del curs.

3) Els usuaris més tardant alhora de fer el primer i el segon problema són els que tenen més possibilitats d'abandonar el curs.

4) Hi ha un conjunt d'uns 40-50 vídeos amb una relació clara entre aquests vídeos i la realització dels problemes d'avaluació.

5) Aquesta relació, no existeix entre la visita de les unitats i els problemes.

Discretització i preparació de les dades pels models

A partir d'aquest punt es començarà a treballar amb un conjunt de dades diferents. Aquest nou conjunt de dades s'ha creat a través d'un programa JAVA que aglutina les dades segons l'usuari.

L'objectiu era establir un marc d'atributs que fossin identificadors de cada usuari. Entenc que el nom dels atributs ja és prou clarificador com per haver d'explicar cadascun dels atributs. En cas que no sigui així, faig una petita explicació entre parèntesis. Quan surt "DESV" es refereix a la desviació típica. En total han estat 24 atributs:

Per una banda tenim els atributs que hem extret directament dels registres que teníem en les dades inicials: USER, NOTA_1, NOTA_2, NOTA_3, NOTA_4, NOTA_5.

Un segon grup serien els atributs que provenen de la combinació aritmètica de registres inicials: NUM_PROBLEMES_FETS, NOTA_MITJANA, NUM_SUCCESS (nombre de problemes superats), DESV_TIP_NOTA, NUM_SESSIONS_PROBLEMES, UNITATS_VISITADES, NUM_SESIONS_UNITATS, NUM_VIDEOS_VIST_DIFERENTS, NUM_SESSIONS_VIDEOS, NUM_TOTAL_REGISTRES.

El tercer grup té en compte el temps que un usuari dedica en les seves sessions:

TEMPS_SESSIONS_PROBLEMES, MITJANA_TEMPS_SESSIONS_PROBLEMES, DESV_TIP_TEMPS_SESSIONS_PROBLEMES, TEMPS_SESSIONS_UNITATS, MITJANA_TEMPS_SESSIONS_UNITATS, DESV_TIP_SESSIONS_UNITATS, TEMPS_SESSIONS_VIDEOS, MITJANA_TEMPS_SESSIONS_VDIEOS, DESV_TIP_SESSIONS_VIDEOS, TEMPS_SESSIONS_TOTALS, TEMPS_SESSIONS_PRE_CURS, TEMPS_SESSIONS_1ERA_SET, TEMPS_SESSIONS_2ONA_SET, TEMPS_SESSIONS_3ERA_SET, TEMPS_SESSIONS_4RTA_SET, TEMPS_SESSIONS_PLUS_SET, DESV_TIPICA_TEMPS_SESSIONS_1ERA_A_4RTA_SET, TEMPS_SESSIONA_1ERA_QUINZENA.

Una vegada aconseguides aquestes noves dades, caldria començar a discretitzar. Però només tenim un total de 211 registres, amb 34 atributs cadascun, per tant no es veu la necessitat de discretitzar per a reduir el conjunt total de valors.

Quan ho fem, que en algun cas ho farem, ja s'exposarà en el propi model.

Models d'agregació

Amb aquests models tenim per objectiu amb quins grups podem dividir els nostres usuaris. Es tracta d'anar creant grups que a posteriori ens serveixi per fer un gran model que els aglutini a tots.

Model 1: Usuaris segons la nota

Objectiu -> Es tracta d'esbrinar quins són els grups d'usuaris que podem formar a través de les notes que han aconseguit.

Dades -> NOTA_1, NOTA_2, NOTA_3, NOTA_4, NOTA_5, NOTA_MITJANA, DESV_TIP_NOTA

A tractar -> no cal fer cap operació prèvia.

Model -> k-means

Comentaris -> agafem la mitjana i la desviació típica per donar valor a la regularitat.

R Project -> funció kmeans(dades, num_clusters). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Resultat:

	square by cluster
k=2	62,4%
k=3	77,8%
k=4	82,6%
k=5	94,0%
k=6	82,5%

DISC_MOD1	SIZE	NOTA_1	NOTA_2	NOTA_3	NOTA_4	NOTA_5	NOTA_MITJ	DESV_TIP
1	23	7,3	0	0	0	0	1,4	6,5
2	127	0,06	0	0	0	0	0,0	0,0
3	20	7,5	9,2	0	0	0	3,3	9,2
4	25	6,8	8,8	8,3	2,7	0	5,3	8,6
5	16	7,1	9,6	9,4	8,4	8,75	8,7	2,6

Conclusions:

Per una banda veiem que la nota mitjana i la desviació típica no tenen un gran efecte. Els grups aconseguits van en relació a les proves d'avaluació fetes.

Model 7:

1->NOMES_COMEN; 2->NOTA_RES; 3->DOS_EXERCICIS; 4->FINAL_CANSA; 5-> TOT_FET

Model 2: Usuaris segons el treball amb els problemes

Objectiu -> Trobar com dividir els grups d'usuaris segons el temps que li han dedicat als problemes.

Dades -> TEMPS_SESSIONS_PROBLEMES, DESV_TIP_TEMPS_SESSIONS_PROBLEMES, MTIJANA_TEMPS_SESSIONS_PROBLEMES, NUM_SESSIONS_PROBLEMES.

A tractar -> no cal fer res.

Model -> k-means.

Comentaris -> li donem el mateix pes el temps total, que a la regularitat, que a les sessions, ho fem així per evitar que algun usuari estigui molt de temps amb una sessió i no entri mai més.

R Project -> funció kmeans(dades, num_clusters). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Resultat:

	square by cluster
k=2	70,2%
k=3	86,9%
k=4	89,1%
k=5	90,1%

DISC_MOD2	SIZE	TEMPS_SESSIO_PROB	NUM_SESS_PROBL	MITJ_TT_SS_PROB	DESV_TIP_SESS_PROB
1	16	634489,37	2,25	388088,79	273967,78
2	189	3492,35	0,78	1794,85	1806,03
3	6	1584551,83	1,16	1430622,58	2178518,18

Conclusions:

Tenim tres grups, un que ha treballat més temps i ha fet servir menys sessions, i en canvi, l'altre ha treballat menys però ha fet més sessions. Per últim, un tercer grup que quasi no ha fet res.

Model 7:

1 -> MES_SESS_PROB; 2->PROB_RES; 3 ->MES_TEMPS_PROB

Model 3: Usuaris segons el treball de les unitats

Objectiu -> Trobar com dividir els grups d'usuaris segons el temps que li han dedicat a visitar les unitats

Dades -> UNITATS_VISITADES, TEMPS_SESSIONS_UNITATS, DESV_TIP_TEMPS_SESSIONS_UNITATS, NUM_SESSIONS_UNITATS, MTIJANA_TEMPS_SESSIONS_UNITATS.

A tractar ->no cal fer res.

Model -> k-means

Comentaris -> Igual que abans, en el treball intentem que tot el pes no el tingui el temps en les sessions, sinó que la regularitat i entrar moltes vegades a veure les unitats se li dóna valor.

R Project -> funció kmeans(dades, num_clusters). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Resultats:

	square by cluster
k=2	62,8%
k=3	78,7%
k=4	89,2%
k=5	90,6%

DICS_MOD3	SIZE	UNI_VIS	TEMPS_SS_UN	DESV_TIP_TE_SS	NUM_SS_UNI	MITJ_TT_SS_UN
1	157	5,15	38954,22	17290,62	2,94	13554,05
2	4	14	2383078,00	0	1	2383078,00
3	37	12,24	839548,03	452693,92	4,37	324842,82
4	13	13,61	1729719,38	1044509,71	4,38	463862,47

Conclusió: Tenim un grup 2 divertit, doncs amb una sola sessió van mirar les 14 sessions. Després tenim dos grups inicials que es diferencien perquè el primer ha treballat més i mirant més o menys els mateixos temes i durant les mateixes sessions, li ha acabat dedicant més temps. Per últim, el grup nombrós ha treballat poc i no ha visitat la totalitat de les unitats.

Model 7:

1 ->POQUES_UNITATS; 2-> TOT_UNITAT_1_SS; 3->FORCA_TT_UNITATS; 4 -> MOLT_TT_UNITATS

Model 4: Usuaris segons el treball dels vídeos

Objectiu -> Trobar com dividir els grups d'usuaris segons el temps que li han dedicat a visitar els vídeos

Dades -> NUM_VIDEOS_VIST_DIFERENTS, TEMPS_SESSIONS_VIDEOS, DESV_TIP_TEMPS_SESSIONS_VIDEOS, NUM_SESSIONS_VIDEOS, MITJANA_TEMPS_SESSIONS_VIDEOS.

A tractar -> cal normalitzar les dades 0-1.

Model -> k-means fins aconseguir un 95%

Comentaris -> Igual que abans, en el treball intentem que tot el pes no el tingui el temps en les sessions, sinó que la regularitat i entrar moltes vegades a veure les unitats se li dóna valor.

R Project -> funció kmeans(dades, num_clusters). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

	square cluster
k=2	62,0%
k=3	78,3%
k=4	89,0%
k=5	91,6%

DISC_MOD4	SIZE	NUM_VID	TT_SS_V	DESV_TT_SS_V	NUM_SS_VID	MITJ_TT_SS_V
1	41	37,09	733720,20	416575,78	5,09	238833,72
2	7	57,57	2098650,29	0	1	2098650,29
3	22	65,68	1604015,09	942568,11	5,86	366529,30
4	141	15,70	41876,54	15781,68	3,51	15522,72

Conclusions:

Tornem a tenir quatre grups, el més gran de tots no ha vist ni un 20% del total dels vídeos. Després, el segon més gros, igualment també ha vist pocs vídeos. Després entre els altres dos, que han vist més vídeos, podem veure que un ho ha fet amb més sessions, mentre que tenim l'altre que ha dedicat molt de temps, però que ho ha fet tot en una sessió.

Model 7:

1-> POCS_VIDEOS; 2-> TOT_VID_1_SS; 3->MOLTS_VIDEOS; 4->VID_RES;

Model 5: Usuaris segons la dedicació total

Objectiu -> Trobar els usuaris segons la seva dedicació total a l'assignatura sense tenir en compte la resta de valors.

Dades -> NUM_PROBLEMES_FETS, NUM_TOTAL_REGISTRES, TEMPS_SESSIONS_TOTAL.

A tractar -> no cal fer res.

Model -> k-means

Comentaris -> Pel que portem vist fins ara amb les dades, haurem de vigilar que aquest grup no acapari la resta. He fet entrar el nombre total de problemes fets, d'aquesta manera relacionem la feina amb la voluntat que aquesta sigui avaluada.

R Project -> funció kmeans(dades, num_clusters). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Resultat:

	square cluster
k=2	76,7%
k=3	90,1%
k=4	94,6%
k=5	95,6

DISC_MOD5	SIZE	NUM_PROB	NUM_TOTAL_REGISTRES	TT_SS_TOTAL
1	21	3,14	3927,28	2072535,1
2	50	1,6	2237,4	912876,8
3	140	0,59	655,11	54401,81

Conclusió:

Tenim tres grups diferenciats força per l'esforç que li han dedicat al curs. Podem veure que hi ha relació entre el nombre de problemes, el nombre de registres i el temps final dedicat.

Model 7:

1 -> TREB_MOLT; 2-> TREB_REGULAR; 3-> TREB_RES

Model 6: Usuaris segons la dedicació durant el curs

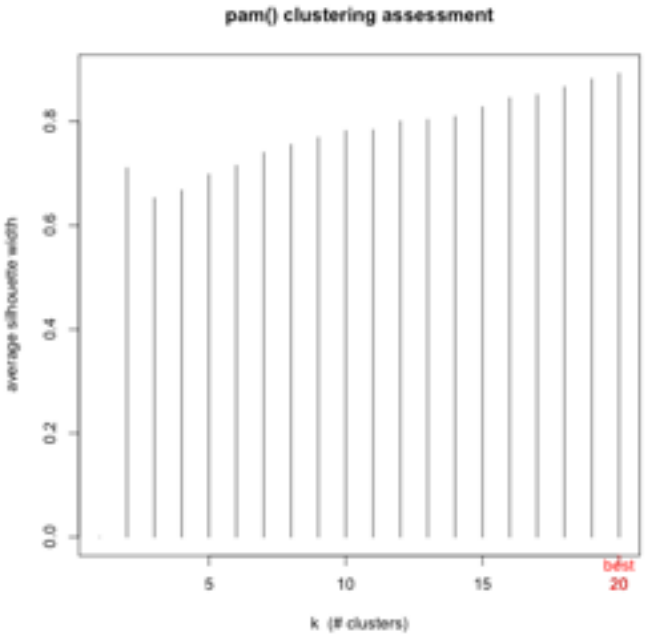
Objectiu -> Trobar els grups d'usuaris segons la feina que han anat desenvolupant al llarg de les setmanes. Es tracta de buscar, no només el temps total, sinó també com l'han repartit durant les poc més de quatre setmanes.

Dades -> TEMPS_SESSIONS_PRE_CURS, TEMPS_SESSIONS_1ERA_SET, TEMPS_SESSIONS_2ONA_SET, TEMPS_SESSIONS_3ERA_SET, TEMPS_SESSIONS_4RTA_SET.

A tractar -> Dades discretitzades en cinc trams iguals. Molt Alt, Alt, Regular, Baix, Molt Baix (4, 3, 2, 1, 0)

Model -> PAM

Comentaris -> Li donem una mica de valor a la regularitat fent entrar la desviació típica.



R Project -> funció pam(dades, nodes) http://cran.r-project.org/web/packages/cluster/cluster.pdf

En el assessorament ens recomanen fer servir més de 20 clústers. Però com que això no és pràctic, en farem servir 6, doncs serien les divisions que tenim i així observar si hi ha relació.

id clúster	id clúst-model	size	PRE CURS	1ERA SETM	2ONA SETM	3ERA SETM	4RTA SETM	PLUS SETM
1	211	149	MOLT BAIX	MOLT BAIX	MOLT BAIX	MOLT BAIX	MOLT BAIX	MOLT BAIX
2	103	15	MOLT BAIX	BAIX	MOLT BAIX	MOLT BAIX	MOLT BAIX	MOLT BAIX
3	153	29	MOLT BAIX	MOLT BAIX	ALT	MOLT BAIX	MOLT BAIX	MOLT BAIX
4	43	10	MOLT BAIX	ALT	MOLT BAIX	MOLT BAIX	MOLT BAIX	MOLT BAIX
5	101	5	MOLT BAIX	MOLT BAIX	MOLT BAIX	ALT	MOLT BAIX	MOLT BAIX
6	78	3	MOLT BAIX	BAIX	MOLT BAIX	MOLT BAIX	ALT	MOLT BAIX

Model 7:

1 i 2 -> SEMPRE_BAIX; 3-> SEGONA; 4-> PRIMERA; 5-> TERCERA; 6-> QUARTA

Conclusions -> En un primer moment aquest model s'havia fet amb els valors sense discretitzar, després de fer-ho sembla que s'observa un resultat més harmònic. En total tenim que cada grup té prioritat en una setmana.

Model 7: Trobar els grups d'usuaris

Objectiu -> En aquest últim model d'agregació es tractarà de posar totes les dades en comú. Aprofitant els grups d'usuaris anteriors, ho intentarem posar tot en comú en un model únic.

Dades -> MODEL_1, MODEL_2, MODEL_3, MODEL_4, MODEL_5, MODEL_6.

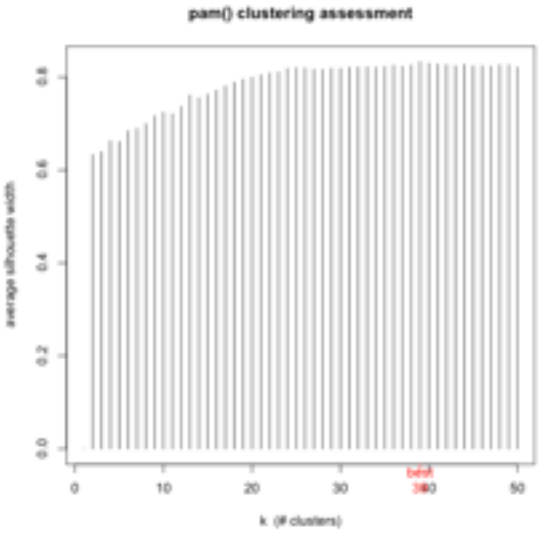
A tractar -> Agafem les dades discretitzades dels models anteriors.

Model -> PAM.

Comentaris -> Aquest models ens deixa treballar bé amb valors categòrics

R Project -> funció pam(total_dai, nodes) http://cran.r-project.org/web/packages/cluster/cluster.pdf

Una de les funcions del PAM és la possibilitat que ens aconselli sobre quin seria el nombre més òptim de clústers. En el nostre cas ens ha donat 36.



Però són tants clústers que no ens interessa. Per això ens decantem per fer-ne 10, ja que entre el 8 i el 12 és el més alt.

El PAM és un algoritme que situa a un node com a centre del clúster, per tant, identifica usuaris tipus. El resultat és el següent:

num clus	ID USUARI MODEL	SIZE
1	211	111
2	92	13
3	7	18
4	9	6
5	173	22
6	15	11
7	182	8
8	130	7
9	83	6
10	157	9

Ara ens queda visualitzar les característiques de cadascun dels clústers:

cluster	SIZE	NOTA	PROBLEMES	UNITATS	VIDEOS	TEMPS	PERIODICITAT
1	111	NOTA_RES	PROB_RES	POQUES_UNITATS	VID_RES	TREB_RES	SEMPRE_BAIX
2	13	NOTA_RES	PROB_RES	POQUES_UNITATS	POCS_VIDEOS	TREB_REGULAR	SEGONA
3	18	NOTA_RES	PROB_RES	FORCA_TT_UNITATS	POCS_VIDEOS	TREB_REGULAR	SEGONA
4	6	TOT_FET	MES_SESS_PROB	MOLT_TT_UNITATS	MOLTS_VIDEOS	TREB_MOLT	SEGONA
5	22	FINAL_CANSA	PROB_RES	POQUES_UNITATS	VID_RES	TREB_RES	SEMPRE_BAIX
6	11	FINAL_CANSA	PROB_RES	FORCA_TT_UNITATS	POCS_VIDEOS	TREB_REGULAR	SEMPRE_BAIX
7	8	NOMES_COMEN	PROB_RES	POQUES_UNITATS	VID_RES	TREB_RES	SEMPRE_BAIX
8	7	NOTA_RES	PROB_RES	FORCA_TT_UNITATS	MOLTS_VIDEOS	TREB_REGULAR	SEGONA
9	6	FINAL_CANSA	MES_TEMPS_PROB	TOT_UNITAT_1_SS	TOT_VID_1_SS	TREB_MOLT	PRIMERA
10	9	FINAL_CANSA	MES_SESS_PROB	MOLT_TT_UNITATS	MOLTS_VIDEOS	TREB_MOLT	TERCERA

Abans d'analitzar els perfils d'usuari obtinguts, hem de tenir molt clar que significa la periodicitat. La periodicitat només remarca on més ha treballat un usuari, en cap cas vol dir que la resta de setmanes no li dediqui temps, però sí que ens dóna una idea de quan li ha dedicat més temps.

clúster 1 : Visitant fugaç

Podem veure que tenim 111 usuaris que estan considerats com a visitants fugaços, van mirar que hi havia i van decidir no seguir.

clúster 2 i 3: Es miren el temari

Aquest conjunt d'uns 30 usuaris es miren una mica el temari, una mica de vídeos, però no fan les proves d'avaluació.

clúster 4: Perfecte

Són els 6 usuaris que ja teníem detectats i que acaben tot el curs amb tot aprovat. Aquest ho fan tot.

clúster 5 i 6: Nomes avaluació

Aquest conjunt d'usuaris només es dedica a fer les proves d'avaluació de forma ràpida i no es miren gaire els vídeos ni les unitats. Potser un clúster treballa una mica més que un altre, però no es tant com per crear un nou grup, sinó podríem caure en una sobreespecialització.

clúster 7: Nomes comencen

Aquest petit grupet només comença el curs i el deixa just al principi de tot sense haver vist res. Són uns usuaris que no interactuen i l'únic que fan és el primer problema.

clúster 8: Miren molt el temari, no fan l'avaluació

Aquest grupet de 7 alumnes destaca per ser un grup que treballa molt, li dedica molt de temps a tot el curs i a mirar els vídeos, però no farà les avaluacions.

clúster 9: En una sessió ho fan tot

Aquest grup destaca per haver fet tot el curs en una sola sessió. Han dedicat moltes hores i han fet bona part de tot el temari, encara que no l'acaben com els que ho fan tot perfecte.

clúster 10: Treballen molt, però al final es cansen

Aquest grup fa bona part del temari i li dedica moltes hores, l'únic que al final no fan la última avaluació.

Els 16 i els perfils d'usuari

Per acabar de comprovar els nostres perfils, ho farem servir amb els 16 usuaris que ja teníem identificats anteriorment.

clúster	quantitat
clúster 4: Perfecte	4
clúster 5 i 6: Nomes avaluació	9
clúster 9 : En una sessió ho fan tot	1
clúster 10 : Treballen molt, però al final es cansen	2

Crida l'atenció que bona part dels usuaris que han acabat el curs no estan dins del d'usuaris que treballen molt, sinó més aviat, usuaris que s'han dedicat a fer les proves d'avaluació.

Fins i tot, podríem agafar els quatre usuaris que han aprovat tot i els dividiríem de la següent manera:

clúster	quantitat
clúster 4: Perfecte	1
clúster 5 i 6: Nomes avaluació	3

Conclusions models de classificació

1) S'observen diferències clares entre tipus d'usuari.
2) Més enllà que molts no facin els problemes d'avaluació, no els hem de descartar com usuaris que no treballin el temari.
3) Els usuaris que més han treballat, no són els que han aconseguit aprovar-ho tot.
4) Un 75% dels usuaris que ho ha aprovat tot, estaria classificat com que ha treballat poc o gens.
5) Dins els 16 usuaris que fan totes les proves, hi ha una diversitat clara entre els que han treballat al llarg del curs, i els que no.
6) Tenim un gran conjunt d'usuaris que són visitants fugaços.

Models Arbres de decisió

Amb l'objectiu de poder comparar diversos aspectes dels nostres usuaris, farem servir arbres de decisió per a comparar i extreure'n la informació.

Tenim dos tipus d'arbres de decisió, aquells que tenen atributs numèrics i els que tenen atributs categòrics. En els primers farem servir el model CART, mentre que en els segons farem servir el model CHAID.

Tot i així, els models de classificació ens han donat un problema. Hem vist la dificultat de classificar els nostres 16 usuaris, i encara més els quatre que ho han aprovat tot.

Per tot això, als nostres arbres els hi donarem dues dimension, una per estudiar el resultat segons el nombre de problemes que han fet, i una altra per estudiar segons la nota mitjana final.

S'ha pensat en fer l'arbre de decisió amb els 16 usuaris que han acabat el curs, però no hi ha prou dades com per obtenir resultats òptims.

Model 10: Comparar total vs regularitat

Objectiu -> Es tracta d'observar si per obtenir una bona nota o fer molts exercicis, si ha tingut més importància el volum total del temps dedicat o la regularitat de les sessions.

Dades -> TEMPS_SESSIONS_PROBLEMES, DESV_TIP_TEMPS_SESSIONS_PROBLEMES, TEMPS_SESSIONS_UNITATS, DESV_TIP_TEMPS_SESSIONS_UNITATS, TEMPS_SESSIONS_VIDEOS, DESV_TIP_TEMPS_SESSIONS_VIDEOS.

(a) NOTA_MITJANA, (b) NUM_POROBLEMES_FETS.

A tractar -> Les dades han estat discretitzades en hores ja que el programa no suportava la continuïtat dels segons. La desviació típica també. Així doncs, un usuari amb 120 segons, tindrà valor 1, i un usuari amb 3601, tindrà valor 2.

Model -> CART

Comentaris -> La idea no és negar un o l'altre, sinó observar què ha tingut més pes.

R Project -> funció rpart(formula, dades, opcions…). http://cran.r-project.org/web/packages/rpart/

Resultat -> Taula.

	Nota Mitjana	Problemes fets
PROBLEMES	TOTAL	TOTAL
UNITATS	TOTAL	TOTAL
VIDEOS	REGULARITAT	REGULARITAT

Conclusió->

Crida l'atenció que només amb els vídeos, la regularitat sigui més important que el temps total. És veritat que durant tot el treball ens estem trobant que els vídeos calia treballar-los diferents que les unitats. Però aquest model no és correcte degut a la manera que tenen de treballar els arbres.

Això es deu a que si busquem els usuaris i les seves notes, segur que ens trobarem amb una bona part que no ha mirat ni un vídeo i deu tenir molt males notes, això s'acaba traduint en una desviació típica pròxima a 0 i molt normalitzada en aquest grup.

Així doncs, aquest model cal descartar-lo.

Model 11: Comparar els temps v/p/u

Objectiu -> En aquest arbre buscarem trobar en quin grup (vídeos, problemes, unitats) han dedicat més temps aquells alumnes que han tret més bones notes i aquells que han fet més exercicis.

Dades -> TEMPS_SESSIONS_VIDEOS, TEMPS_SESSIONS_UNITATS, TEMPS_SESSIONS_PROBLEMES, NOTA_MITJANA, NUM_PROBLEMES_FETS.

A tractar -> Agafem les dades convertides en el model 10.

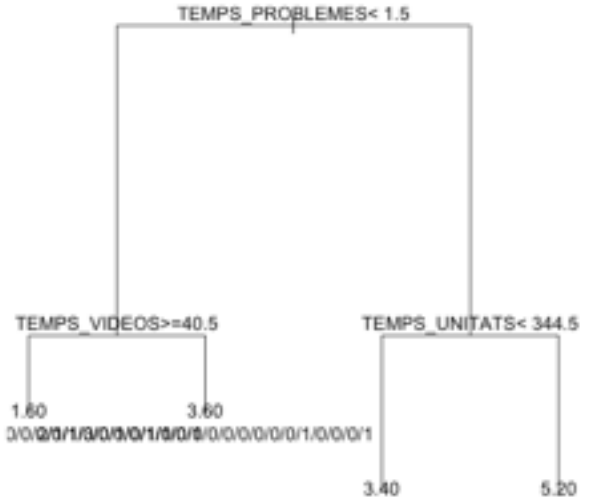
Model -> CART

Comentaris -> Cal vigilar molt com treballem amb aquest model per no caure amb errors com abans.

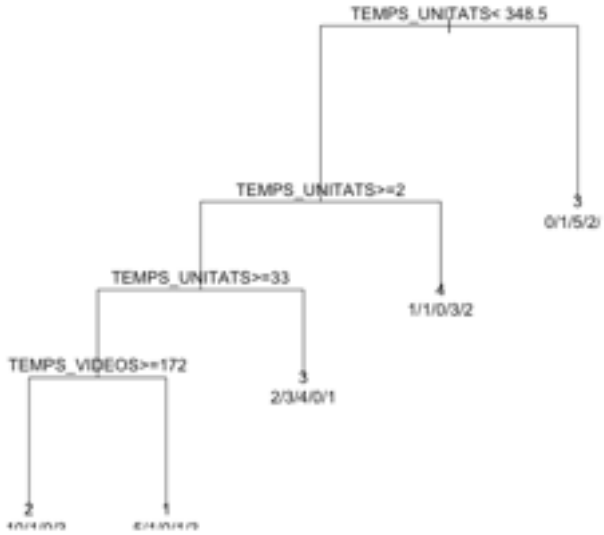
R Project -> funció rpart(formula, dades, opcions…). http://cran.r-project.org/web/packages/rpart/

Resultat ->

NOTA_MITJANA



PROBLEMES_FETS



En el gràfic de la esquerra (NOTA_MITJANA) podem observar que l'element principal és el temps dedicat als problemes, mentre que el temps dedicat als videos i a les unitats estan en paral·lel.

En el gràfic de la dreta(NOMBRE_PROBLEMES) no hem posat el temps dedicat als problemes, doncs semblava obvi que estaria molt relacionat amb el nombre de problemes fets, així que només tenim el temps dedicat a les unitat i als vídeos, i podem veure, que el temps dedicat a les unitats ha estat més decisiu.

Conclusió ->

Les conclusions d'aquest model són sorprenents, doncs deixen el temps dedicat a veure els vídeos en un tercer pla, sense gaire pes alhora de decantar un alumne amb els seus resultats.

Ja hem vist abans que és difícil marcar una línia d'alumnes que ho han fet tot bé, doncs molts que han acabat el curs ho van fer de forma ràpida i sense una gran feina.

Però aquest model desperta preguntes. Podria ser que al igual que abans, no sigui el temps sinó el nombre de vídeos vistos la clau per entendre el paper dels vídeos?

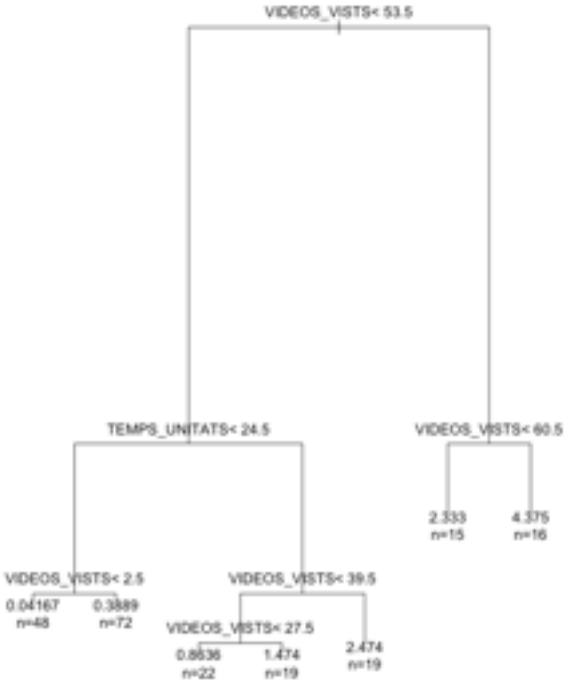
Amb la voluntat de respondre aquesta pregunta repetim el model, però ara no amb el temps dels vídeos sinó amb el nombre de vídeos vists.

Resultat ->

Conclusió->

Les conclusions són clares, una vegada fem servir el nombre de vídeos vists, aquest atribut puja de consideració i esdevé un ítem clau alhora d'adjudicar una nota a un alumne, tal com podem veure en el gràfic de l'esquerra. Si observem bé el gràfic, tenim que els usuaris que obtinguin més d'un 7 de mitja al final, haurien de veure un mínim d'entre 50 i 60 vídeos. A més, les unitats queden en un tercer pla.

Però si mirem el gràfic de la dreta, les dades són igual d'aplastants, la quantitat de vídeos vista està directament relacionada amb el nombre total d'exercicis que farà un usuari. I al igual que abans, sembla que el nombre màgic és entre 50 i 60 vídeos.



Model 12: Comparar les setmanes

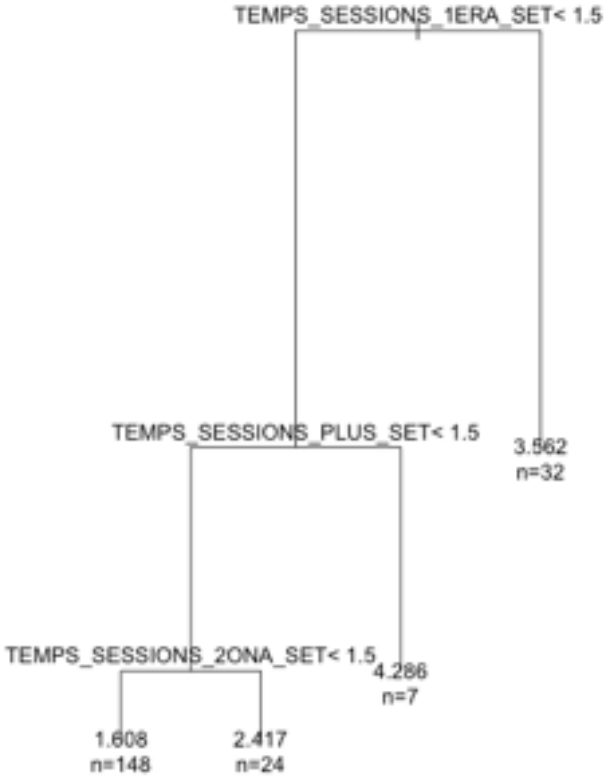
Objectiu -> En aquest model haurem de mirar quina és la setmana més rellevant alhora de relacionar els usuaris amb la nota.

Dades -> TEMPS_SESSIONS_PRE_CURS, TEMPS_SESSIONS_1ERA_SET, TEMPS_SESSIONS_2ONA_SET, TEMPS_SESSIONS_3ERA_SET, TEMPS_SESSIONS_4RTA_SET, TEMPS_SESSIONS_PLUS_SET, NOTA_MITJANA, NUM_PROBLEMES_FETS.

A tractar -> En aquest cas, agafem les dades discretitzades del model 6.

Model -> CART

Comentaris -> Tenir en compte que també buscarem si hi ha diferències entre la nota dels alumnes i el nombre de problemes fets.



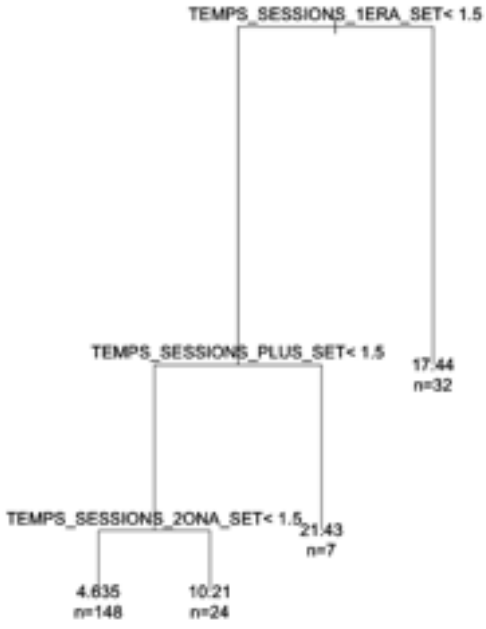
R Project -> funció rpart(formula, dades, opcions…). http://cran.r-project.org/web/packages/rpart/

Resultat num problemes fets-> SUP

Resultat nota mitjana -> INF

Conclusió-> En ambdós resultats podem observar que el temps dedicat durant la primera setmana és el que més classifica els alumnes.

En tots registres podem veure que hi ha un grup d'alumnes que es dedica molt la primera setmana, però són els usuaris que es dediquen més temps durant la última setmana els que aconsegueixen millors resultats. Segurament degut a que el temps dedicat al final del curs els hi serveix per acabar el curs, mentre que la resta d'usuaris no fan res durant aquella última setmana ja que han abandonat el curs.



És a dir, tenim un grup d'uns 30 alumnes que al principi li dediquen molt de temps, però que no acaben el curs. D'altra banda, i dit així acaba tenint molta lògica: aquells alumnes que destaquen per treballar a la última setmana del curs, són aquells alumnes que acaben el curs.

Model 13: Comparar model 7

Objectiu -> Es tracta d'agafar el model 7 i aplicar un arbre de decisió sobre la nota i sobre els exercicis fets.

Dades -> MODEL_7(del model 2 al model 5), NOTA_MITJANA, NUM_PROBLEMES_FETS.

A tractar -> Passar les dades a data.frame. Agafem la discretització feta en els models anteriors.

Model -> CART

Comentaris -> Model molt interessant i que posarà de manifest la correcta execució dels models anteriors. En aquest cas no agafarem la periodicitat en el temps, ja que ja es veu clarament que els alumnes que es dediquen la última setmana de curs són els que tenen més bones notes i els que fan més exercicis, ja que fan la part final del curs.

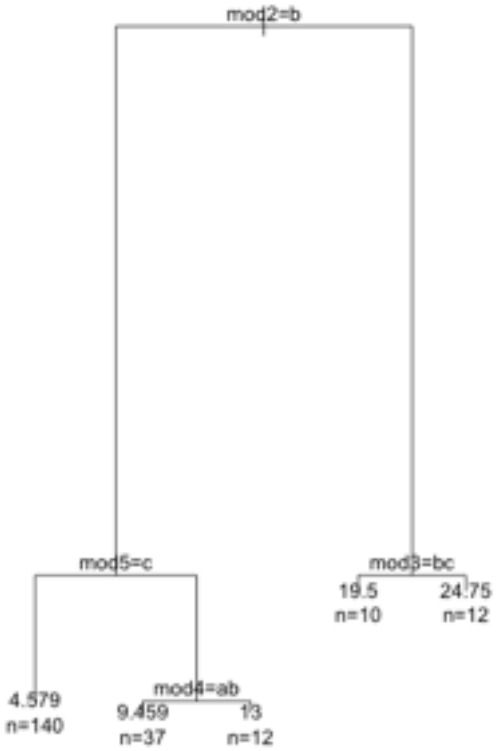
Nosaltres teníem la següent selecció la qual podem mirar fàcilment quina ha estat la nota mitjana del grup i quants problemes han fet de mitjana:

id clust	size	discr_model_7	nota_mitjana	prob_fets_mitj
1	111	FUGAC	0	0
2-3	31	MIREN_UNA_MICA	1	0,7
4	6	PERFECTE	7,4	4,3
5-6	33	NOMES_AVAL	5,8	3,4
7	8	NOMES_COMN	1,4	1
8	7	FEINA_NO_AVL	1,8	1,1
9	6	UNA_SESSIO	5,7	3,6
10	9	FINAL_CANSA	4,9	2,8

On podem observar que hi ha una clara relació entre el tipus d'usuari i la nota aconseguida. Però la idea del model és observar si hi ha relació amb les categories treballades en el model 7 a través dels models de mineria de dades.

R Project -> funció rpart(formula, dades, opcions…). http://cran.r-project.org/web/packages/rpart/

Resultat



	mod2 PROBLEMES	**mod3** UNITATS	**mod4** VIDEOS	**mod5** DEDICACIO
a	MES_SESS_PROB	POQUES_UNITATS	POCS_VIDEOS	TREB_MOLT
b	PROB_RES	TOT_UNITAT_1_SS	TOT_VID_1_SS	TREB_REGULAR
c	MES_TEMPS_PROB	FORCA_TT_UNITATS	MOLTS_VIDEOS	TREB_RES
d		MOLT_TT_UNITATS	VID_RES	

Resultat nota mitjana:

En l'arbre podem observar com la lògica al final és aplastant. En la mesura que els alumnes hagin treballat més els problemes, mod2, més bona nota mitjana han aconseguit.

Si mirem la branca dels que han treballat, llavors veiem que la manera de treballar les unitats acaba decantant més la nota final. Sent els alumnes que han treballat molt les unitats els que millor nota treuen.

Aquestes dades contrasten amb les que teníem amb el model 11. El fet que els vídeos no siguin tan importants en aquest model es deu haver degut a que no comptabilitza el nombre total de vídeos, sinó la manera com s'han visualitzat aquests vídeos.

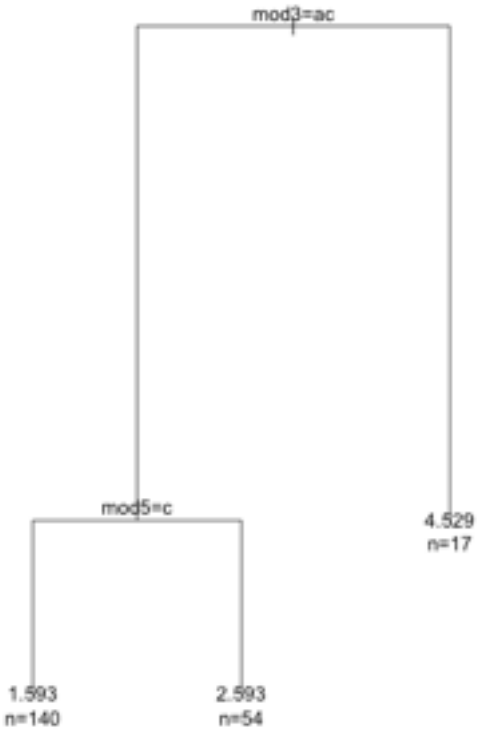
Resultat nombre problemes fets:

Aquest gràfic encara és més interessant, doncs ens ha trobat un grup format per 17 alumnes (nosaltres teníem els 16 magnífics) els qual fan la major part dels problemes.

Aquests usuaris el model els detecta a través del treball de les unitats, fet que seguiria contrastant amb el model 11.

Així doncs, aquells alumnes que treballen molt les unitats són els que acaben fent la major dels problemes.

Alhora, també podem observar que la manera com hem treballat els vídeos no surten per enlloc.



Conclusions arbres de decisió

1) No podem afirmar que la regularitat o el temps total siguin classificadors alhora d'avaluar els resultats obtinguts.

2) El temps més rendible és aquell que es dedica als problemes.

3) El temps dedicat als vídeos no és més rendible que el temps dedicat a les unitats.

4) Per contra, el nombre de vídeos vistos sí que és més determinant que el temps dedicat a les unitats alhora d'avaluar un alumne.

5) Seguim observant que hi ha un conjunt de 50-60 vídeos clau.

6) Els qui acaben completament el curs són aquells que treballen la última setmana.

7) Hi ha una relació clara entre el perfil del alumne i la nota mitjana i nombre de problemes fets.

8) Els models 2, 3 i 5 són més determinants que el model 4.

9) Els alumnes que treballen la major part de les unitats, tenen més possibilitats d'acabar tot el curs.

Conclusions finals

Aquestes conclusions finals les he dividit en dues parts, una primera de forma esquemàtica fent un recull de les principals conclusions que hem anat extraient del treball, i una segona part on s'intenta respondre als objectius plantejats a l'inici del treball

Conclusions esquemàtiques

1) De totes les dades només un curs es pot estudiar i cal prestar atenció amb els usuaris que fan d'administrador.

2) L'activitat del curs dràsticament disminueix al llarg del curs.

3) Hi ha un conjunt de vídeos que no s'haurien d'analitzar, doncs els propis usuaris, en quasi la seva totalitat, els ha descartat.

4) No hem pogut trobar un la línia del temps lògica en la visita de les unitats.

5) Sí hem trobat una cronologia lògica alhora de veure els vídeos i fer els problemes.

6) No tenim prou usuaris que ho hagin aprovat tot com per a treballar amb ells.

7) Els usuaris que acaben el curs, acostumen a fer la feina inicial abans que la resta de companys.

8) Hi ha un conjunt de 50-60 vídeos clau.

9) No hi ha relació entre la realització dels problemes i el fet de visitar les unitats.

10) Els models 1, 2, 3, 4 , 5 i 6 aporten informació interessant alhora de discretitzar els grups d'usuaris.

11) El model 7 dibuixa un escenari de 8 perfils d'usuari.

12) Hi ha un grup d'usuaris amb el qual, l'avaluació no és reflex de la feina que han fet durant el curs.

13) Els 16 alumnes que han fet els cinc problemes tenen una complexitat interior gran, amb perfils que han treballat molt poc i perfils que han treballat molt.

14) El nombre de vídeos vistos és clau per aconseguir una bona nota.

15) Durant la primera i la última setmana els alumnes que acabaren el curs, o aconseguiran millor nota, destaquen més.

16) Hi ha una relació evident entre els perfils dibuixat en el model 7 i la nota i exercicis que han fet.

Conclusions basades en objectius

Alhora d'exposar les conclusions finals ho farem en relació als objectius inicials que ens havíem marcat.

Buscar perfils d'usuari

Els diversos models efectuats ens han mostrat clarament un conjunt de 8 perfils d'usuaris, els quals no només destacant per característiques comunes, sinó que també observem que aquestes conjuguem amb la nota mitjana que obtenen i el nombre de problemes fets.

En un primer moment pensàvem en dividir els grups segons la feina i la avaluació, si ho féssim podríem dividir els nostres perfils de la següent manera:

	TREBALLEN **13%**	**NO TREBALLEN** **86%**
FAN AVALUACIÓ **25%**	PERFECTE UNA SESSIÓ FINAL CANSA 10%	NOMES AVALUACIÓ 15%
NO FAN AVALUACIÓ **74%**	FEINA NO AVALUACIÓ 3%	FUGAÇ MIREN UNA MICA NOMES COMENCEN 71 %

Al final, el conjunt de 8 perfils no deixa de ser una graduació d'aquests quatre grans grups. Però la gràcia de fer aquesta agrupació e´s que podem distribuir el percentatge en cadascuna de les dimensions.

Podem veure doncs, que els usuaris han fet més feina d'avaluació que treball amb els vídeos o les unitats.

Tot i així, cal tenir present que al final només hem tingut 211 usuaris i que parlem d'un curs no avaluable, per tant, faltaria obtenir més dades.

Per últim, si no volem perfils tan amplis d'usuaris, la resta de models d'agregació també ens han mostrat que és possible classificar els grups segons com han vist les unitats, els vídeos, els problemes o com han treballat.

Buscar la relació entre la feina i l'avaluació

És una relació difusa i difícil de descriure. Sí que és veritat que tenim un grup molt identificat que treballa molt i que fa les feines d'avaluació. Però crida molt l'atenció quan observem el perfil dels 16 que fan tota les proves i dels 4 que ho aproven tot.

Al final només podem afirmar que el fet de fer la feina és positiu alhora d'aconseguir una avaluació, però el fet de no fer la feina no és un impediment per aconseguir la mateixa avaluació.

Això segurament es degui a la naturalesa del curs, optatiu i repàs de conceptes. Es tractava d'un curs preparatori per una assignatura universitària. Seria molt interessant aconseguir les notes d'aquests usuaris en l'assignatura i observar si hi ha relacions.

Crec que aquest és el punt que més falta de totes aquestes dades, doncs si al final tot plegat serveix per preparar-se per a un avaluació posterior, necessitaríem les dades d'aquesta avaluació posterior per a poder avaluar correctament aquest curs.

Fins i tot, no aniria malament una enquesta feta pels usuaris marcant on es pogués identificar la relació entre fer aquest curs i la facilitat després per seguir l'assignatura.

Relacionar temari i proves d'avaluació

Aquest era un altre aspecte clau i només en el cas dels vídeos hem pogut aconseguir-ho. Podem afirmar que hi ha un conjunt de vídeos, entre 40 i 60, que són claus per aconseguir obtenir una nota mitjana alta, i alhora, aquells usuaris que veuen aquests vídeos són els que acabaran el curs.

Aquesta relació ha estat impossible amb les unitats. Només entre les unitats 1 i 7 hi ha una relació estable en el temps que es podria relacionar amb els problemes, però a partir de la 7 no segueix cap ordre.

Va ser un dels objectius que més fàcilment es va assolir. Ja vam veure que dels 211 usuaris que van interaccionar en algun moment amb el curs, només 70 van fer la primera prova, i d'aquests, només 16 acabaran fent les avaluacions.

Un altre exemple el tenim per exemple en la visió dels vídeos. Si agafem el vídeo més vist entre els 16 que acaben el curs, poc més de 60 alumnes interaccionaran amb aquest vídeo. Però encara més dràstic és si agafem el quinzè vídeo més vist entre els grups de 16, només 21 usuaris interaccionen amb ell.

Tot plegat ens dóna una idea del poc seguiment que hi hagut per la major part dels usuaris. Sí que és veritat que els models ens han reflectit que hi ha un grup d'usuaris que mira uns 50 vídeos i obté bons resultats, però aquest grup és molt petit i no és reflex de la major part dels usuaris.

Bibliografia

Llibre, apunts i TFG anteriors

Mor i Pera, E.; et Altri, Mineria de Dades, UOC, 2010, Barcelona

Han, Jiawei; et Altri, Data Mining. Concepts and Techniques, ELSEVIER, Waltham , 2012

Blanco Carpintero, Antonio, TFG Educational data minin learning analytics, UOC, 2014

Adroher Salvia, A, TFG Patrons de connexió al cv de la UOC, UOC, 2014

Recursos electrònics

Tota la informació referent a R-Project es troba al CRAN, allà es on es troben bona part de les llibreries que he fet servir:

http://cran.r-project.org , CRAN, 2015.

Per la comprensió de les dades en el seu format original:

http://edx.readthedocs.org/en/latest/index.html , EDX, 2015

API JAVA:

http://docs.oracle.com/javase/7/docs/api/ , JAVA, 2015