Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Aplicació de mineria de dades per la detecció i anàlisi de perfils d'estudiants

MEMÒRIA

Autor: Agnès López Soler
Director: Lluis Talavera Mendez

Convocatòria: Gener 2019

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

En aquest projecte s'aplica de forma rigorosa una metodologia pel desenvolupament de projectes de mineria de dades per detectar perfils d'estudiants del Grau en Enginyeria en Tecnologies Industrials (GETI) de forma que es puguin analitzar i extreure caracteritzacions útils per entendre diferents patrons de comportament i de rendiment als estudis d'enginyeria.

Per començar es fa una breu introducció al món de la mineria de dades i a les seves tècniques, fent èmfasi en el clustering i en l'algorisme K-Means que servirà per descriure els grups naturals d'estudiants que puguin existir.

A continuació, es presenta la metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) que comprèn les diferents etapes que cal seguir per poder aplicar un model de mineria de dades.

Finalment, s'aplica aquesta metodologia per agrupar els estudiants del GETI segons la seva semblança. Es documenta el procés de comprensió i preparació de les dades, el procés de construcció del model i l'avaluació d'aquest.

El software utilitzat per dur a terme les etapes del projecte que ho requereixen és Python amb les llibreries Pandas, SciKit-Learn, Matplotlib i Seaborn.

Sumari

1. Introducció

1.1. Objectius del projecte

Els objectius principals d'aquest projecte són:

- Aplicar tècniques de mineria de dades per analitzar dades de rendiment acadèmic dels estudiants del Grau en Enginyeria en Tecnologies Industrials (d'ara endavant GETI)
- Aplicar un algorisme de *clustering*, concretament el K-Means per extreure caracteritzacions útils per entendre diferents patrons de comportament i rendiment als estudis d'enginyeria.
- Descobrir patrons interessants que puguin servir al professorat per entendre el rendiment d'algun grup d'estudiants.

En un segon pla, els objectius són:

- Aplicar de forma rigorosa una metodologia de manera que es puguin identificar clarament les diferents fases, documentar-les i replicar-les en futurs anàlisi.
- Estudiar les limitacions de les dades de les que es disposa.
- Estudiar les limitacions i el rendiment de l'algorisme K-Means.
- Examinar diferents formes de representació gràfica dels resultats.

1.2. Abast del projecte

Hi ha moltes tècniques de mineria de dades i cadascuna requereix estudiar el seu funcionament i portar a terme tot un procés de preparació i validació, és per això que aquest projecte s'ha restringit a un mètode de clustering, el K-Means i a les funcionalitats de la llibreria Scikit-Learn de Python.

Com que és un projecte que servirà per introduir-se al món de la mineria de dades es considera que és millor prendre un volum de dades reduït per realitzar l'estudi i familiaritzar-se amb el tractament de dades i l'aplicació de tècniques de mineria de dades. Per aquest motiu, només es tenen en compte les dades de la fase inicial del GETI.

L'estudi dels perfils de les altres titulacions que puguin cursar-se a l'Escola Tècnica Superior d'Enginyeria Industrial de Barcelona (d'ara endavant ETSEIB) com el Grau en Enginyeria de Materials (GEM) i el Grau en Enginyeria Química (GEQ) també queda fora de l'abast del projecte.

2. Introducció a la Mineria de Dades

La mineria de dades consisteix a analitzar grans quantitats de dades, mitjançant mètodes estadístics o intel·ligència artificial, per extreure'n coneixement rellevant, comprensible i útil per a la presa de decisions en tota mena de contexts.

Va més enllà de l'anàlisi estadístic tradicional. No tan sols en les tècniques que utilitza sinó també en el coneixement que es busca assolir. L'objectiu principal ja no és validar hipòtesi: és descobrir patrons i relacions entre les dades de forma automàtica (o semi-automàtica) sense cap o amb poques hipòtesis prèvies.[1]

Per exemple, la mineria de dades pot servir per determinar perfils de clients (*costumer profiling*) i identificar els clients més rendibles (que interessa conservar) o també pot servir per fer *targetting*, és a dir, determinar les característiques dels clients profitosos que han estat captats per empreses competidores per poder recuperar-los.

Es pot aplicar a moltes àrees, que van des de supermercats fins a la medicina. En els supermercats, els codis de barres proporcionen als establiments grans quantitats de dades que permeten determinar els millors preus, fer inventari i calcular quins productes i quines quantitats cal demanar als proveïdors i també fer balanços de manera eficaç i determinar els beneficis obtinguts. En el camp de la medicina, guardar els historials dels pacients permet identificar el tractament més adequat si es coneix com ha funcionat un tractament en diferents pacients, segons el gènere, l'edat i el seu propi historial mèdic. La mineria de dades aplicada pels bancs permet reconèixer usuaris de targetes de crèdits; aplicada per les companyies asseguradores i de telecomunicacions permet detectar fraus; aplicada per empreses fabricants permet fer controls de qualitat més exhaustius, així, en farmacèutiques implica augmentar de manera notable la seguretat del producte.

La mineria de dades parteix d'una base de dades, que es tracta per poder-hi aplicar algorismes de manera automàtica i extreure'n conclusions. El procés de la mineria de dades recau fortament en la tecnologia de la informació, en la forma de les dades emmagatzemades, i també en el software per analitzar-les. No obstant això, el procés és molt més que simplement aplicar aquest software a les dades. Cal que l'analista comprengui les dades, les seleccioni i transformi de manera adequada i en pugui interpretar els resultats.

2.1. Tècniques de mineria de dades

De manera general, dins la mineria de dades, es poden distingir dos tipus de problemes, els que requereixen tècniques supervisades i els que les requereixen no-supervisades.

Les tècniques supervisades, també anomenades de classificació, fan servir un conjunt de dades etiquetades amb una variable dependent o "classe" d'un conjunt finit i discret de variables per aprendre a deduir-la a partir de la resta de variables per després predir l'etiqueta d'un altre conjunt de dades semblant però pendent de classificar. Per exemple, si s'etiqueten les dades d'un conjunt d'estudiants amb una variable que digui si han aprovat o no una assignatura, es podrien construir models per predir a partir de la resta de notes si aproven o no. També funciona quan les etiquetes no són discretes, i aleshores es fan servir models de regressió.

La classificació és la tècnica que consisteix a predir, per a cada individu d'una població, a quin d'un conjunt (petit) de classes pertany aquest individu. En general, les classes són mútuament excloents. Per exemple, podria aplicar-se per conèixer dins del grup de clients d'una companyia, quins probablement respondran a una oferta determinada. En aquest cas les classes serien dues, els que responen favorablement a l'oferta i els que no. A partir de dades de clients etiquetats (model), es prediu la resposta d'un client nou i s'etiqueta en conseqüència.

Una tasca estretament relacionada amb la classificació és la puntuació o l'estimació de la probabilitat de classe. Un model de puntuació aplicat a un individu produeix, en comptes d'una predicció de classe, una puntuació que representa la probabilitat (o alguna altra quantificació de probabilitat) que aquest individu pertanyi a cada classe. En l'exemple donat per la classificació, un model de puntuació seria capaç d'avaluar cada client individual i produir una puntuació de la probabilitat que cadascun d'ells respongui a l'oferta.

La regressió és la tècnica que consisteix a estimar o predir, per a cada individu, el valor numèric d'alguna variable. La regressió està relacionada amb la classificació, però les dues són diferents. Informalment, la classificació serveix per predir si alguna cosa passarà, mentre que la regressió prediu en quina quantitat (valor numèric) passarà aquesta cosa en qüestió. Per exemple, mentre que la regressió podria predir quina nota obtindrà un estudiant a una assignatura concreta tenint en compte les notes que ha obtingut en les altres assignatures i també un conjunt de dades totes les notes d'altres estudiants; la classificació faria servir etiquetes aprovat/suspès i podria predir el comportament de l'estudiant en aquest sentit.

Les tècniques no-supervisades s'apliquen quan no hi ha dades etiquetades i serveixen per descobrir grups i/o relacions automàticament sense fer suposicions sobre l'estructura de les dades. Aquesta tècnica té l'avantatge de descobrir relacions no previstes.

El *clustering* és la tècnica que consisteix a agrupar els individus d'una població per la seva semblança, però no per cap propòsit específic. L'agrupació busca que els elements d'un grup siguin el més semblant possible i alhora siguin el més diferent possible respecte als elements d'altres grups. Cada clúster pot veure's com una classe d'objectes, del qual se'n poden derivar regles. Una pregunta d'agrupació d'exemple seria: "Els clients formen grups o segments naturals?". L'agrupament és útil en l'exploració preliminar de dominis per veure quins grups naturals existeixen perquè aquests grups poden suggerir altres tasques o enfocaments de mineria de dades. El *clustering* també s'utilitza com a aportació als processos de presa de decisions centrats en qüestions com: Quins productes es poden oferir o desenvolupar? Com s'han d'estructurar els equips d'atenció al client?

L'associació, també coneguda com a mineria d'objectes freqüents o descobriment de regles d'associació, consisteix a trobar associacions entre entitats basades en les transaccions que les impliquen. Una pregunta d'exemple seria: Quins elements es compren conjuntament? Tot i que el *clustering* analitza la similitud entre objectes, l'associació considera similituds d'objectes en funció de la seva aparició junts en les transaccions.

Per exemple, l'anàlisi dels registres de compra d'un supermercat pot descobrir que la carn trinxada es compra juntament amb una salsa concreta amb molta més freqüència del que es podria esperar. Decidir com actuar sobre aquest descobriment pot requerir certa creativitat, però podria suggerir una promoció especial, una visualització de productes o una oferta de combinació. La coincidència de productes en les compres és un tipus comú d'agrupació coneguda com a anàlisi de cistella de mercat. Alguns sistemes de recomanacions també realitzen un tipus d'agrupació d'afinitat trobant, per exemple, parells de llibres que són adquirits freqüentment per les mateixes persones ("les persones que van comprar X també van comprar Y"). El resultat de l'agrupació de coincidència és una descripció d'esdeveniments que es produeixen conjuntament. Aquestes descripcions solen incloure estadístiques sobre la freqüència de la coexistència i una estimació indicadora de si era previsible o no (el factor sorpresa que implica).

2.1.1. El K-Means

La tècnica de mineria de dades en què es basa aquest treball és el *clustering*, ja que es busca conèixer els diferents perfils d'estudiants que hi ha a la fase inicial del GETI.

En aquest apartat es descriu amb més detall aquesta tècnica, en concret, l'algorisme *K-Means* que és el que s'ha decidit utilitzar per dur-lo a terme.

El *K-Means* és un algoritme d'aprenentatge no supervisat que troba un nombre donat k de clústers en un conjunt de dades. Quan s'utilitza un algorisme de *K-Means*, un clúster es defineix mitjançant un centroide, que és un punt (sigui imaginari o real) al centre d'un clúster. Cada punt del conjunt de dades forma part del clúster el centroide del qual li queda més a prop. Per fer-ho, *K-Means* troba un nombre k de centroides i assigna tots els punts de dades al clúster més proper. Per determinar el clúster més proper calcula la distància entre el punt en qüestió i els diferents k centroides, per quedar-se amb el que minimitza la distància. Habitualment s'utilitza la distància euclidiana.

El *K-Means* comença definint aleatòriament k centroides C, tenint $C = c_1, c_2, \dots, c_k$.

A partir d'aquí, treballa en passos iteratius per dur a terme dues tasques:

1. Assignar cada punt de les dades al centroide corresponent més proper.
   Per fer-ho calcula la distància entre cada punt i els diferents centroides i es queda amb la que és mínima. El càlcul de la distància euclidiana és el següent:
   $$d_E(c_i, p) = \sqrt{\left(x_p - x_{c_i}\right)^2 + \left(y_p - y_{c_i}\right)^2}$$

2. Per cada centroide, calcula la mitjana dels valors de tots els punts que hi pertanyen. El valor mitjà es converteix en el nou valor del centroide.
   $$c_i = \frac{1}{|S_i|} \cdot \sum_{p_i \in S_i} p_i$$
   On S_i és el conjunt de tots els punts que pertanyen al clúster i.

Un cop finalitzat el pas 2, tots els centroides tenen nous valors que corresponen a les mitjanes de tots els punts que pertanyen al seu propi clúster. Aquests nous punts es fan passar pels passos 1 i 2 per produir un altre conjunt de valors de centroides. Aquest procés es repeteix una i altra vegada fins que s'aconsegueix que d'una iteració a la següent no hi hagi variació entre els centroides obtinguts, el que significa que les dades s'han agrupat amb precisió.

El K-Means també permet que els centroides inicials no siguin aleatoris sinó definits per l'usuari o també determinar un nombre màxim d'iteracions per aturar l'algorisme.

Per veure millor com funciona s'exposa un cas pràctic.

Es disposa d'una base de dades que conté quatre mesures de tres flors d'iris diferents. Les mesures són: longitud del sèpal, amplada del sèpal, longitud del pètal i amplada del pètal. Els tres tipus d'iris són Setosa, Versicolour i Virginica. [2]

Fig. 2.1 D'esquerra a dreta: Iris Versicolor, Setosa i Virginica. A la imatge de l'esquerra pot observar-se la diferència entre pètal i sèpal. Font http://www.lac.inpe.br/~rafael.santos/Docs/R/CAP394/WholeStory-Iris.html

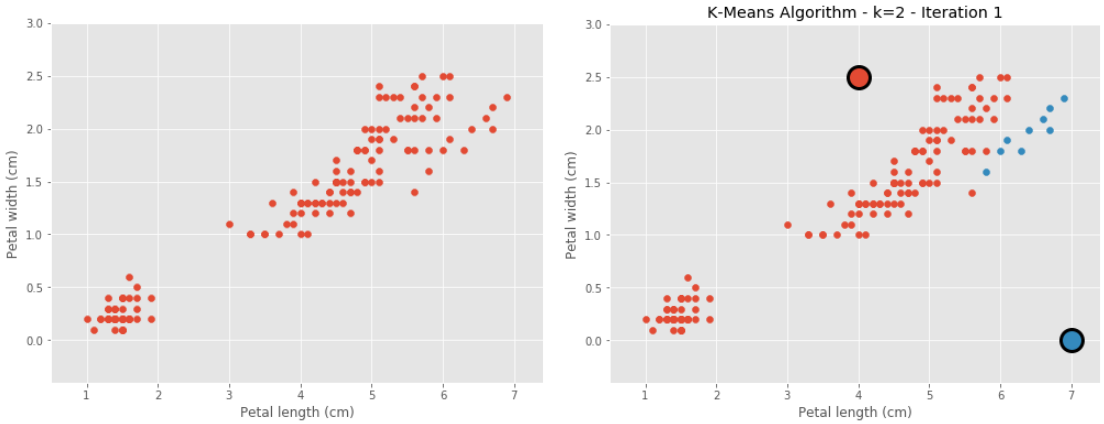

Fig. 2.2 Visualització de les dades (esquerra) i primera iteració de l'algorisme K-Means. Font https://blog.easysol.net/machine-learning-algorithms-3/

Si es visualitza la relació entre l'amplada i l'alçada dels pètals (Fig. 2.2) a simple vista, sembla que es puguin diferenciar dos grups, però no sempre té perquè ser tan evident.

Per l'algorisme K-Means, s'ha d'imposar un nombre k de clústers a trobar, en aquest cas, es tria k=2. A la Fig. 2.2 es veu com es defineixen dos centroides de manera aleatòria i es creen els clústers assignant cada punt al grup del centroide que li queda més a prop. Els centroides són els punts de mida més gran amb el contorn negre i es diferencien els clústers amb dos colors.

El següent pas és tornar a calcular els centroides i torna a ubicar els punts al clúster més proper, de manera iterativa.

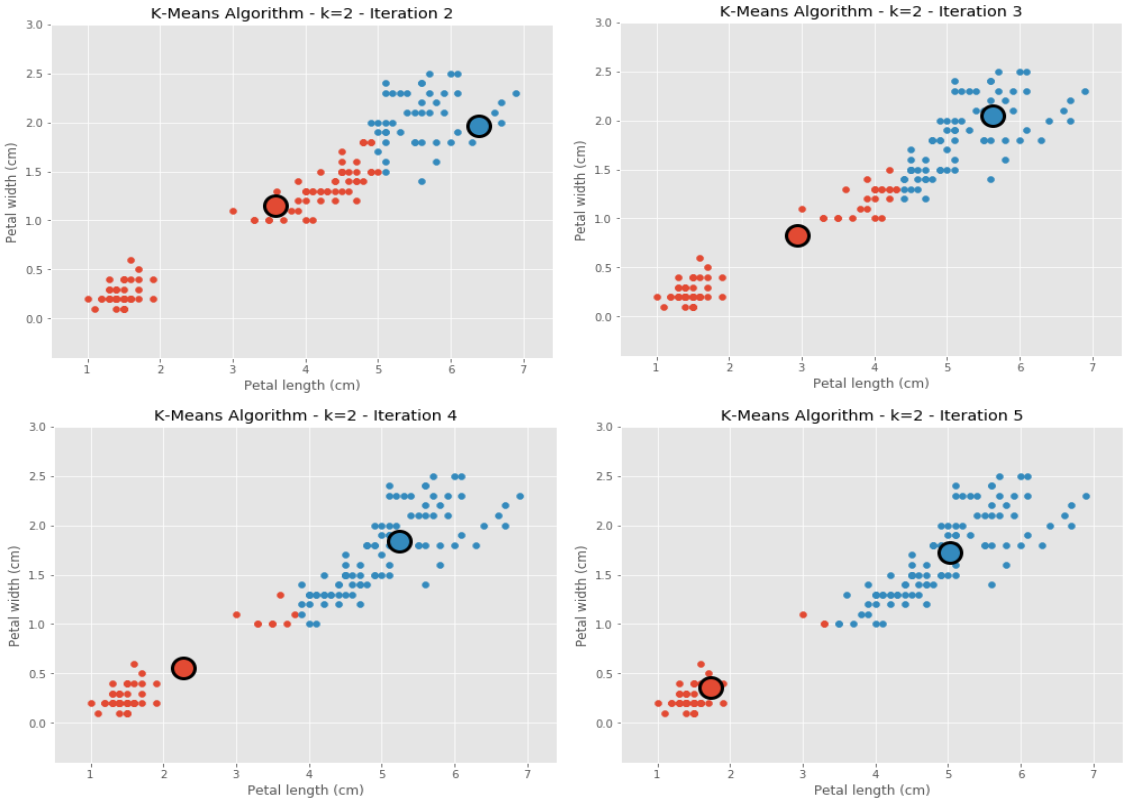

Fig. 2.3 Iteracions 2, 3, 4 i 5 de l'algorisme K-Means. Es pot veure com es desplacen els centroides
Font https://blog.easysol.net/machine-learning-algorithms-3/

La Fig. 2.3 mostra com es realitzen iteracions successives fins que el valor dels centroides no varia. En aquest cas, la iteració 5 és l'última. Com es pot veure a la figura, han quedat definits dos clústers i tots els punts han estat ubicats a un o a l'altre. N'hi ha alguns que en les diferents iteracions han canviat de clúster i alguns que s'han mantingut. També es pot observar el desplaçament que han patit els centroides, primer definits de manera aleatòria.

Tot i haver obtingut una solució, no té per què ser òptima, doncs s'ha fet una suposició a l'hora d'escollir k seguint un criteri completament subjectiu. Per trobar el millor nombre de clústers k de manera objectiva s'ha de trobar una manera de mesurar la qualitat d'aquests.

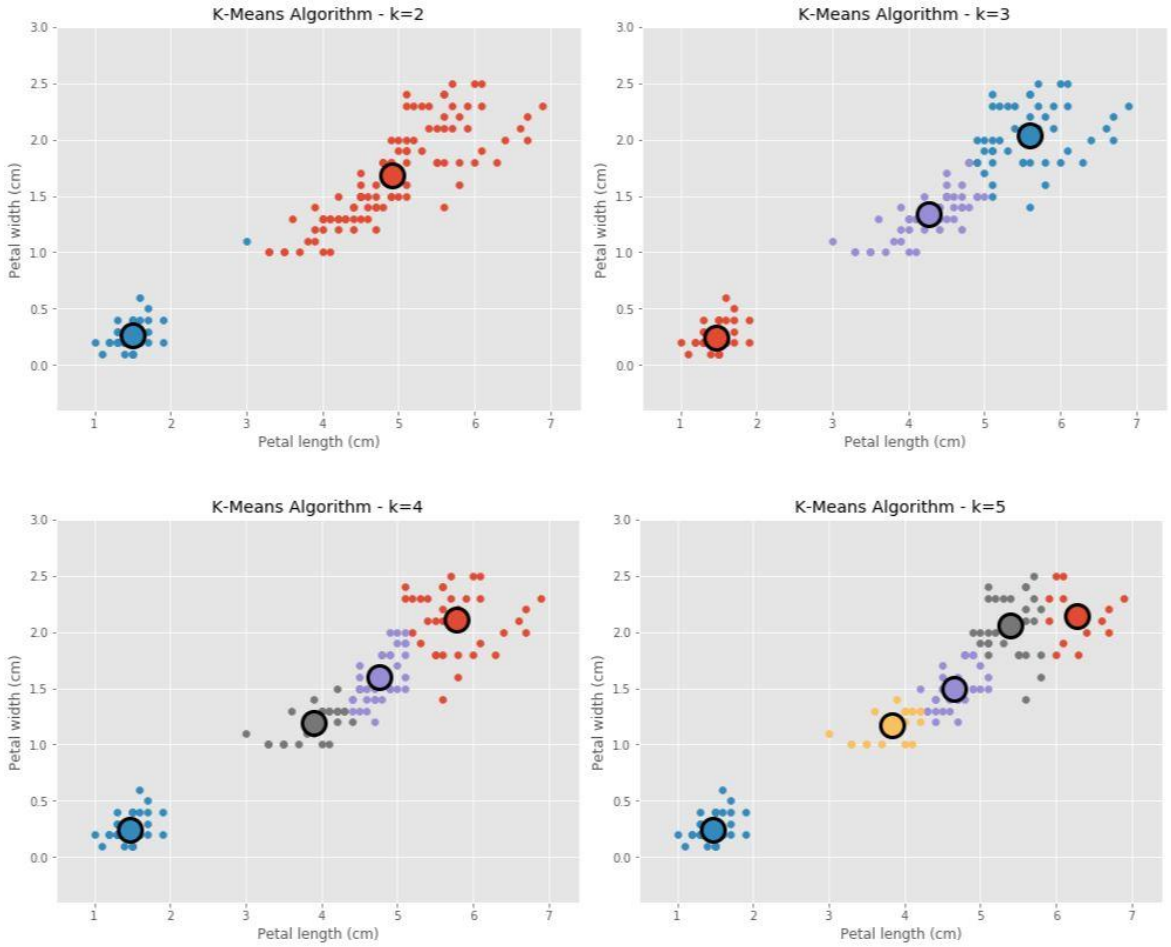

Fig. 2.4 Resultats d'aplicar l'algorisme K-Means amb diferents k. Font https://blog.easysol.net/machine-learning-algorithms-3/

Gràficament, pot veure's quin és el nombre de clústers adequat si s'aplica l'algorisme per diferents valors de k. En aquest cas, el nombre de clústers adequat és k=3. És un resultat coherent, ja que la base de dades conté informació de 3 tipus d'iris.

Per aquest treball en qüestió, a l'apartat de construcció del model s'explica de quina manera s'ha determinat el nombre de clústers k.

2.2. CRISP-DM

Per dur a terme l'anàlisi de les dades de forma rigorosa, és convenient seguir una metodologia com el CRISP-DM (Cross-Industry Standard Process for Data Mining), que s'utilitza de manera habitual a la indústria.[3] L'abast d'aquest treball és molt més reduït que el d'un projecte industrial i no caldrà seguir el model de forma rígida, però dóna una idea de quins aspectes cal tractar, començant per l'exploració de les dades, seguint pel processament

d'aquestes, l'anàlisi, l'extracció d'inferències i la implementació.

La metodologia CRISP-DM engloba la comprensió del negoci, la comprensió de les dades, la construcció del model, l'avaluació dels resultats i la implementació.

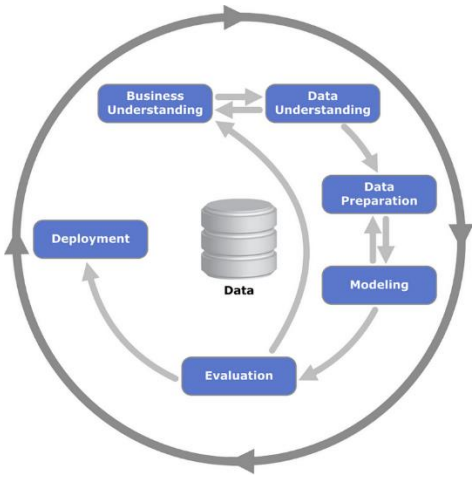

Fig. 2.5 Etapes de la metodologia CRISP-DM. Font: Kenneth Jensen - ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en

-**Comprensió del negoci***:* És la fase on es determinen els objectius, s'estudia la situació actual i s'estableixen les fites a assolir. A la indústria, les fites es referirien per exemple, a descobrir els perfils dels clients o bé quin tipus de clients estan interessats en quins productes. També es desenvolupa el pla del projecte en general. A més, en cas que ho requereixi, s'estableix el pressupost preliminar que es dedicarà a l'estudi.

-**Comprensió de les dades***:* Aquesta fase es centra en les dades i els requeriments d'aquestes. Un cop recollides les dades, caldrà descriure-les, explorar-les i verificar la seva qualitat. En funció de l'objectiu a complir, es necessitaran diferents dades, és per això que és molt important seleccionar-les correctament. Per recollir les dades de manera adequada cal assegurar que amb aquestes es pugui descriure de manera clara i concisa el problema; també cal que es pugui identificar les dades rellevants per la descripció del problema; i que les variables seleccionades per les dades rellevants siguin independents, perquè la informació que aporten no se superposi i pugui crear confusions i complicar l'aplicació d'algorismes, sobretot a l'hora de buscar patrons.

-**Preparació de les dades***:* És la fase a la qual s'arriba un cop es coneixen les dades disponibles. Les dades s'han de seleccionar, netejar i transformar fins que tinguin el format amb què es podrà treballar en les següents fases. Durant la neteja, les dades es filtren i també

s'estudien els valors perduts, per omplir-los si s'escau, ja que hi ha algorismes que no funcionen amb valors buits. Els filtres permeten trobar *outliers* i redundàncies. Un outlier o anomalia és qualsevol dada que sembla no estar en concordança amb la resta de dades. En funció de l'objectiu del procés de mineria de dades, interessa conservar-los i estudiar-los o eliminar-los. Per exemple, conservar aquests valors anòmals pot ajudar als bancs a detectar fraus amb targeta de crèdit. Si s'acostumen a fer transaccions de valor baix i en hores concretes del dia i de cop se'n dona una de valor elevat i a un horari no habitual, per ser diferent, és una transacció que s'estudiarà. Però si el que es vol és detectar el comportament general d'un conjunt de dades, cal obviar els valors anòmals, doncs generen desviacions. Pràcticament la meitat del procés de mineria de dades es dedica a la preparació de les dades.

-**Construcció del model***:* És la fase on s'utilitza software de mineria de dades per generar resultats per diverses situacions. Inicialment es visualitzen les dades i se sol aplicar una anàlisi de clúster. Seguidament, depenent de l'objectiu a aconseguir i el tipus de dades s'apliquen diferents models. Si es pretén agrupar dades, i els grups són coneguts, s'aplica una anàlisi discriminant. Si es busca fer estimacions, aplicar regressions és adequat (sempre que les dades siguin continues). En els dos casos, també es podrien aplicar xarxes neuronals. Per classificar dades, una alta opció és fer servir arbres de decisió. En resum, en aquesta fase s'utilitza el software disponible per guanyar coneixement referent a les dades, i es poden fer servir múltiples models, de manera individual o combinats, per fer-ho.

-**Avaluació dels resultats***:* És la fase on s'interpreten les dades i és de les més critiques. Per una banda, és important reconèixer el valor del que s'ha descobert durant la mineria de dades. Per altra banda, també cal decidir quina eina de visualització és apropiada i servirà per exposar els resultats. Cal tenir en compte que l'eina de visualització utilitzada no és l'adequada, pot ser que es perdi informació important. Això no vol dir fer gràfics extremadament complicats, ja que, si no són interpretables, es tindria el mateix problema.

-**Implementació***:* És la fase on l'analista transmet els resultats de la mineria de dades. Es relaciona el coneixement obtingut amb els objectius originals del projecte i s'exposa. Aquest coneixement no té per què ser vàlid per sempre, ja que el comportament de les dades pot variar al llarg del temps, per això cal donar eines per fer-ne un seguiment. En aquesta fase també es pot posar el model en funcionament, per exemple, si el projecte ha servit per detectar clients potencials, es pot iniciar la campanya de marketing per captar-los.

2.3. Eines i entorn de desenvolupament

Exceptuant la comprensió del negoci i la implementació, totes les etapes de la metodologia CRISP-DM necessiten suport computacional per dur-se a terme.

Cal un programa que permeti visualitzar, transformar i manipular les dades, aplicar tècniques i algorismes de mineria de dades i mostrar els resultats obtinguts mitjançant gràfics. En primera instància es valora utilitzar Python com a llenguatge de programació amb el suport de les llibreries Numpy, Pandas, Scikit-Learn, Matplotlib i Seaborn, però cal veure si en aquest cas és la millor opció.

Per a l'anàlisi de dades i la computació exploratòria i la visualització de dades interactiva, Python rivalitza amb els altres llenguatges de programació comercial i de codi obert específics, principalment amb R i SAS però també amb MATLAB, Stata i d'altres.

Segons Wes McKinney al llibre *Python for Data Analysis*, en comparació amb altres programes es distingeix per la seva gran i activa comunitat informàtica científica. També ha millorat de manera notable el suport per a llibreries de Python (principalment Pandas) cosa que ha fet que sigui una bona alternativa per a les tasques de manipulació de dades. [4]

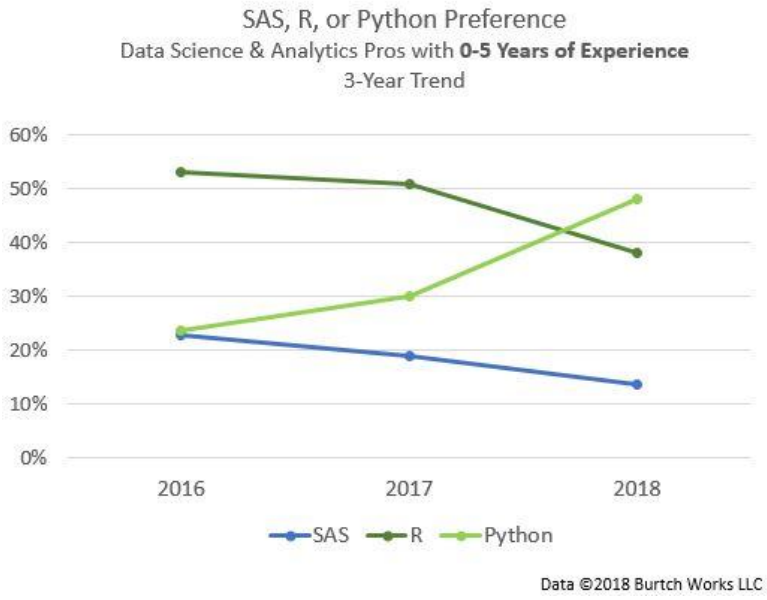

Fig. 2.6 Preferència d'usuaris amb poca experiència entre SAS, R i Python. Font https://www.burtchworks.com/2018/07/16/2018-sas-r-or-python-survey-results-which-do-data-scientists-analytics-pros-prefer/

En comparació amb R i SAS, diferents enquestes i resultats d'estudiar tendències a les xarxes mostren que en els últims anys, la preferència per Python, sobretot en usuaris inexperts, ha augmentat considerablement. El gràfic de la Fig. 2.6 mostra com pels usuaris amb poca experiència SAS queda enrere, i entre R i Python, la tendència convida a fer servir Python.

En conclusió, per fer aquest treball es considera més adequat fer servir Python amb el suport de les llibreries que s'han anomenat anteriorment. A més, un avantatge que es té és que ja es coneix una mica el llenguatge.

Com que es vol treballar amb diferents llibreries s'ha instal·lat Anaconda, que és una distribució de Python, amb eines d'instal·lació i gestió de paquets. Proporciona la possibilitat de crear diferents entorns de Python, cadascun amb la seva pròpia configuració i gestiona la instal·lació de llibreries per evitar problemes de compatibilitat.

Anaconda permet treballar amb Spyder, que és un entorn de desenvolupament integrat (IDE) escrit en i per al llenguatge de programació Python. Comprèn un editor per escriure codi, un terminal per avaluar-lo i veure els resultats en qualsevol moment i un explorador de variables per poder veure les variables definides. També té eines per localitzar errors de codi, permet depurar codis i accedir als manuals i documentació de les diferents funcions de les llibreries que s'utilitzen.

Les llibreries que es fan servir per dur a terme les diferents etapes de la mineria de dades són: NumPy, Pandas, SciKit-Learn, Matplotlib i Seaborn.

Pandas és la llibreria dedicada a les estructures de dades i està desenvolupada sobre la llibreria Numpy.[5][6] Principalment treballa amb DataFrames, que són estructures bidimensionals tabulars de dades orientades a columnes amb etiquetes de files i columnes. Té funcions pròpies dissenyades per treballar amb dades estructurades de forma ràpida i senzilla. Permet importar i exportar dades, visualitzar-les, inspeccionar-les (fer mitjanes, càlculs de desviacions...), seleccionar-les, filtrar-les, ordenar-les, agrupar-les, netejar-les, combinar-les...

SciKit-Learn és la llibreria que proporciona eines per la mineria de dades i l'anàlisi de dades. Proporciona algorismes de classificació, regressió, clustering... [7]

Matplotlib i Seaborn són llibreries per crear gràfics i visualitzar dades.[8][9]

3. Metodologia

L'aplicació de la mineria de dades per la detecció i anàlisi de perfils dels estudiants del GETI de l'ETSEIB es duu a terme adaptant les etapes de la metodologia CRISP-DM. En els següents apartats es descriu el desenvolupament de cada etapa.

3.1. Comprensió del negoci i de les dades

3.1.1. Descripció de les dades

Les dades de què es disposa inicialment són les dades acadèmiques dels estudiants de l'Escola, en dos fitxers separats, un pels resultats de la fase inicial (o fase selectiva) i un altre pels de la fase no inicial.

L'abast d'aquest projecte recau únicament en el tractament de les dades de la fase inicial, tot i que com que les dades de la fase no inicial són semblants, el tractament que es faci a unes dades hauria de ser extrapolable de manera senzilla a les altres.

Les dades de la fase inicial comprenen les matrícules realitzades des de la tardor de 2010 fins a la primavera de 2017 i contenen informació de 3836 estudiants del GETI.

Així doncs, es disposa d'un fitxer excel amb diferents columnes on cada fila correspon a les dades acadèmiques derivades de la matrícula d'un estudiant a una assignatura concreta. Cada vegada que es matricula una assignatura s'ocupa una fila, si es repeteix una assignatura, la informació no se sobreescriu i s'emplenen dues files a la base de dades. La base de dades està feta de manera que cada matrícula a una assignatura s'interpreta com una transacció i cada transacció ocupa una fila de la taula de dades.

COLUMNA	FORMAT	DESCRIPCIÓ
CODI_PROGRAMA	Int64	Indica a quin grau o màster es realitza la inscripció de l'assignatura. Els codi dels estudiants del GETI és el 752.
CODI_EXPEDIENT	Int64	Hi ha un CODI_EXPEDIENT per cada estudiant. Indica a quin estudiant pertanyen les dades que s'exposen, a les columnes de la mateixa fila.

CODI_UPC_UD	Int64	Indica quina assignatura es matricula. Com que la fase inicial consta de 10 assignatures, hi ha 10 CODIS_UPC_UD vàlids. Pren valors 240011, 240012, 240013, 240014, 240015, 240021, 240022, 240023, 240024, 240025 que corresponen respectivament a Àlgebra, Càlcul 1, Mecànica Fonamental, Química 1, Fonaments d'informàtica, Geometria, Càlcul 2, Termodinàmica Fonamental, Química 2 i Expressió Gràfica.
CREDITS	Float64	Indica de quants crèdits és l'assignatura matriculada. Pren valors: 4.5, 6 o 7.5.
CURS	Int64	Indica l'any en què es matricula l'assignatura.
QUAD	Int64	Indica el quadrimestre en què es matricula l'assignatura. Pren valors 1 o 2, corresponents respectivament al quadrimestre de tardor i al de primavera.
SUPERA	String	Indica si se supera o no l'assignatura matriculada. Amb els valors S/N corresponents a si/no.
NOTA_PROF	Float64	Indica la nota obtinguda a l'assignatura matriculada assignada pel professor. Els valors vàlids, van del 0 al 10.
NOTA_NUM_AVAL	Float64	Indica la nota obtinguda a l'assignatura matriculada assignada durant l'avaluació. Els valors vàlids van del 0 al 10.
NOTA_NUM_DEF	Float64	Indica la nota final obtinguda a l'assignatura matriculada, és la definitiva. Els valors vàlids van del 0 al 10.
GRUP_CLASSE	String	Indica el grup classe on s'ha matriculat l'assignatura. Si pren el valor CONV significa que s'ha convalidat l'assignatura.

Taula 3.1 Descripció de l'arxiu qfaseini per columna

3.1.2. Exploració de les dades

Abans de filtrar les dades i transformar-les, s'exploren amb l'objectiu de conèixer les dades i localitzar anomalies (*outliers*) si n'hi ha.

En primera instància es detecta que hi ha valors buits o perduts a la base de dades. A la fase de preparació de les dades caldrà o bé eliminar-los o bé omplir-los, però primer és necessari entendre el seu significat. En concret, hi ha valors buits a les columnes que defineixen les notes obtingudes (NOTA_NUM_PROF, NOTA_NUM_AVAL, NOTA_NUM_DEF) i a les columnes que defineixen el grup classe (GRUP_CLASSE). S'estudien els valors buits de les columnes referents a les notes, ja que aquestes serviran per determinar els clústers, les referents al grup classe no són dades necessàries per a l'estudi que es vol fer i no se seleccionaran quan es faci l'estudi.

Els valors buits en les notes pertanyen a 15 estudiants (CODI_EXPEDIENT) concrets. Del volum total d'estudiants no és un nombre significatiu.

CODI_ EXP.	CODI_ UPC_ UD	CRÈD	CURS	Q.	SUP.	NOTA_ NUM_ PROF	NOTA_ NUM_ AVAL	NOTA_ NUM_ DEF	GRUP_ CLASSE
229928	240011	6.0	2010	1	S	NaN	NaN	NaN	CONV
230843	240023	6.0	2010	2	S	NaN	NaN	NaN	CONV
227073	240025	7.5	2011	2	S	NaN	NaN	NaN	CONV
232924	240022	6.0	2010	2	S	NaN	NaN	NaN	CONV
229036	240025	7.5	2010	2	S	NaN	NaN	NaN	CONV

Taula 3.2 Mostra d'estudiants amb valors buits a les notes.

Com es pot veure a l'exemple de la Taula 3.2, són estudiants que no tenen assignada cap nota i a més el grup classe que els consta és CONV, que indica que s'ha fet una convalidació. S'observa que són matrícules de l'any 2010-2011, que coincideix en el període d'adaptació del canvi de Pla94 al GETI. En tots els casos, la columna SUPERA (a la taula SUP.) indica que s'ha superat l'assignatura.

La particularitat d'aquests casos és que són estudiants que, si s'observen individualment, de les deu assignatures de la fase inicial varies han estat convalidades i no tenen nota. A la taula següent es mostra el cas d'un estudiant concret.

CODI_ EXP.	CODI_ UPC_ UD	CRÈD	CURS	Q.	SUP.	NOTA_ NUM_ PROF	NOTA_ NUM_ AVAL	NOTA_ NUM_ DEF	GRUP_ CLASSE
230149	240011	6.0	2010	1	S	NaN	NaN	NaN	CONV
230149	240012	6.0	2010	1	S	NaN	NaN	NaN	CONV
230149	240014	6.0	2010	1	S	7.1	7.1	7.1	62
230149	240015	6.0	2010	1	S	7.8	7.8	7.8	11
230149	240025	7.5	2010	1	S	NaN	NaN	NaN	CONV
230149	240013	6.0	2010	2	S	NaN	NaN	NaN	CONV
230149	240021	6.0	2010	2	S	NaN	NaN	NaN	CONV
230149	240022	6.0	2010	2	S	5.8	5.8	5.8	31
230149	240023	6.0	2010	2	S	5.6	5.6	5.6	31
230149	240024	4.5	2010	2	S	8.4	8.4	8.4	31
230149	240011	6.0	2010	1	S	NaN	NaN	NaN	CONV

Taula 3.3 Mostra d'estudiant concret amb valors buits a les notes

La Taula 3.3 mostra el cas d'un estudiant que l'any 2010 va matricular quatre assignatures al primer quadrimestre i sis al segon quadrimestre. Al primer quadrimestre (Tardor), va cursar Àlgebra, Càlcul 1, Química 1 i Fonaments d'informàtica. Les dues primeres consten com a convalidades i els valors referents a les notes estan buits; les dues últimes sí que tenen notes i són notables. Al segon quadrimestre (Primavera), va cursar Mecànica Fonamental, Geometria, Expressió Gràfica, Càlcul 2, Termodinàmica Fonamental i Química 2. Les tres primeres consten com a convalidades i els valors referents a les notes estan buits, les altres tres sí que tenen notes. La columna SUPERA (a la taula SUP.) en els deu casos és una S, que indica que s'ha superat l'assignatura. Respecte al total de dades disponibles referents a les notes que ha obtingut l'estudiant en qüestió, n'hi ha més de buides.

Es poden fer suposicions sobre per què es tenen aquestes dades d'aquest estudiant, però no es té prou informació per conèixer la validesa de les hipòtesis que es facin. Per exemple, podria ser que fos un estudiant que iniciés els seus estudis l'any 2009 i superés cinc assignatures de les deu, després l'any 2010 amb el canvi de pla acadèmic es convalidessin les assignatures aprovades (perdent-se la nota amb el canvi de pla) i l'estudiant cursés les cinc assignatures pendents de superar. Quedaria pendent entendre perquè les convalidacions d'assignatures es van fer en quadrimestres diferents i no totes de cop. En ser dades de l'any 2010, no es disposa de dades anteriors per ajudar a confirmar o desmentir aquesta hipòtesi.

Les opcions per tractar els valors nuls podrien ser eliminar-los, però no només els valors nuls sinó tots els que representen aquests estudiants; omplir-los, o bé a partir de mitjanes de les seves pròpies notes, que en aquest cas pot ser que no es tinguin notes per fer mitjanes, o bé a partir d'aplicar tècniques supervisades i predir els seus valors, però són casos especials i no té sentit predir-los en funció de casos de circumstàncies que podrien ser completament diferents.

Per tant, després d'aquesta exploració s'arriba a la conclusió que a l'etapa de preparació de les dades caldrà eliminar aquests estudiants.

A continuació, donat que es treballa amb dades de la fase inicial, també anomenada selectiva, es creu convenient veure el volum d'estudiants que supera la fase selectiva i el volum d'estudiants que no. Malgrat la redundància, la fase inicial queda superada quan se superen les deu assignatures que la conformen. Superar la fase inicial no implica aprovar totes les assignatures a la primera convocatòria.

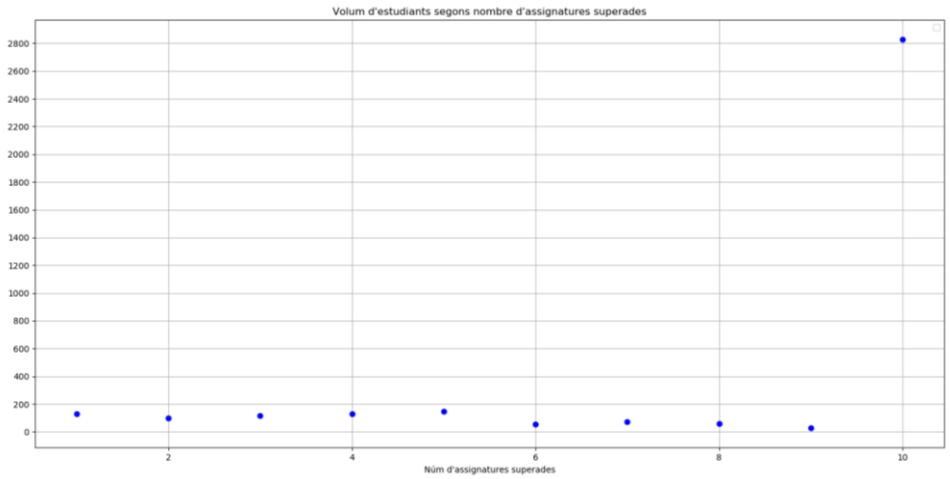

Fig. 3.1 Nombre d'estudiants en funció de les assignatures superades

El gràfic de la Fig. 3.1 mostra que del total de 3836 estudiants dels qual es tenen dades, 2819 superen la fase inicial (un 73.48%). La resta d'estudiants són aquells que el juny de 2017 tenen pendent superar alguna de les assignatures, o bé perquè seguiran cursant la fase inicial el setembre de 2017 i pretenen superar-les o bé perquè han abandonat els estudis.

3.2. Preparació de les dades

Les dades de les quals es disposa contenen informació no necessària i/o redundant per l'anàlisi que a realitzar. A més la disposició de les dades no és l'adequada.

Com s'ha comentat durant la descripció de les dades, la base de dades conté informació de les diferents matrícules que els estudiants han realitzat, enteses com a transaccions. El que interessa en aquest punt és unir la informació de totes les transaccions que ha realitzat cada estudiant, de manera que ja no ocupi vàries files de la taula de dades sinó una de sola.

L'etapa de preparació de les dades es divideix en dues fases, la primera de neteja i selecció i la segona de transformació.

3.2.1. Neteja i selecció de les dades

L'objectiu d'aquesta fase és seleccionar els estudiants del GETI i eliminar els estudiants outliers, és a dir els que tenen valors anòmals, detectats durant la comprensió de les dades.

La columna CODI_PROGRAMA de les dades indica la titulació a la qual es realitza la matrícula de l'assignatura. Com que es volen estudiar únicament els estudiants del GETI, es seleccionen les dades dels estudiants amb el codi 752.

A continuació, es construeix un filtre per eliminar les dades dels estudiants que tenen algun valor buit a les seves notes. Concretament s'identifiquen els estudiants que tenen un valor buit (NaN) a la columna NOTA_NUM_DEF i s'esborren de la taula de dades totes les files amb el seu CODI_EXPEDIENT.

Seguidament s'aplica un altre filtre per seleccionar únicament els estudiants que superen la fase inicial. El motiu d'aquesta decisió s'explica al punt 3.2.2.

Per conèixer els estudiants que superen la fase inicial n'hi ha prou en observar la columna SUPERA. La fase inicial consta de deu assignatures, es considera la fase inicial superada quan s'han aprovat totes. La columna supera, com s'ha explicat a la Taula 3.1, pren valor S quan s'aprova una assignatura i N quan no. Els estudiants que superen la fase inicial tindran deu vegades el valor S a la columna SUPERA.

El filtre detecta els estudiants que tenen menys de deu assignatures superades i esborra de la taula de dades totes les files amb el seu CODI_EXPEDIENT.

3.2.2. Transformació de les dades

L'objectiu d'aquesta fase és tenir una matriu de dades, com que es treballa amb pandas també s'anomena dataframe, que permeti analitzar les notes obtingudes per cada estudiant. La informació necessària serà la nota obtinguda a les assignatures i s'obviarà la informació referent als crèdits, al curs, al quadrimestre i al grup classe. També interessa conèixer en quina convocatòria s'ha obtingut cada qualificació, és a dir, si es repeteix o no. Aquesta informació no apareix de manera explícita a la taula i caldrà trobar-la a partir de les dades existents.

Per fer-ho, es valoren dues opcions:

- La construcció d'una matriu de dades on cada estudiant (CODI_EXPEDIENT) ocupi una fila i les columnes siguin les notes obtingudes en funció de la convocatòria. Amb

tantes columnes com convocatòries realitzades.

CODI_EXP.	N-Àlg_1	N-Àlg_2	N_Càl1_1	N-Càl1_2	...	N-ExpGr_1	N-ExpGr_2
A	6.1	NaN	3.9	7.6	...	4.5	5.3
B	0	6.7	5.0	NaN	...	5.5	NaN
C	3.2	5	4.0	5.7	...	5.3	NaN
D	4.0	NaN	3.2	NaN	…	NaN	NaN
E	8.5	NaN	8.1	NaN	…	7.1	NaN

Taula 3.4 Exemple de la primera proposta de construcció de *dataframe* prèvia a l'anàlisi

- La construcció d'una taula on cada estudiant (CODI_EXPEDIENT) ocupi una fila i les columnes continguin l'última nota que consta de les deu assignatures que conformen la fase inicial i la convocatòria on s'ha obtingut aquesta nota.

CODI_EXP.	N-Àlg	Conv-Àlg	N_Càl1	Conv-Càl1	...	N-ExpGr	Conv-ExpGr
A	6.1	1	7.6	2	...	5.3	2
B	6.7	2	5.0	1	...	5.5	1
C	4.9	2	5.7	2	...	5.3	1
D	4.0	1	3.2	1	…	NaN	0
E	8.5	1	8.1	1	…	7.1	1

Taula 3.5 Exemple de la segona proposta de construcció de *dataframe* prèvia a l'anàlisi

Per construir aquestes matrius de dades, es fa servir la llibreria Pandas per importar el fitxer excel i transformar la taula de dades inicial. El procediment és similar per les dues propostes de taules i segueix el següent esquema:

1. Crear una columna indicadora de la convocatòria a la qual s'ha assolit cada nota. Com s'ha explicat a l'apartat de descripció de les dades, inicialment cada fila de la taula conté informació de la matrícula d'un estudiant a una assignatura concreta. Així que per cada fila, l'element d'aquesta columna contindrà informació de la convocatòria en què s'ha matriculat una assignatura concreta.

2. Crear una columna indicadora de la convocatòria i de l'assignatura. (Aquest és un pas intermedi necessari pel pas 3.)

3. Pivotar la matriu de dades amb l'eina *pivot* de Pandas. Aquesta eina permet pivotar la taula i transformar-la de manera que els valors d'una columna concreta, que

ocupen files, passen a ser les columnes d'una nova taula, que és la taula pivotada. Caldrà pivotar de manera que els valors de la columna indicadora de la convocatòria i l'assignatura siguin les columnes de la taula i com a valors, la taula tingui les notes obtingudes a l'assignatura (NOTA_NUM_DEF). La informació innecessària, referent als crèdits, al curs, al quadrimestre i al grup classe, no apareix a la taula pivotada.

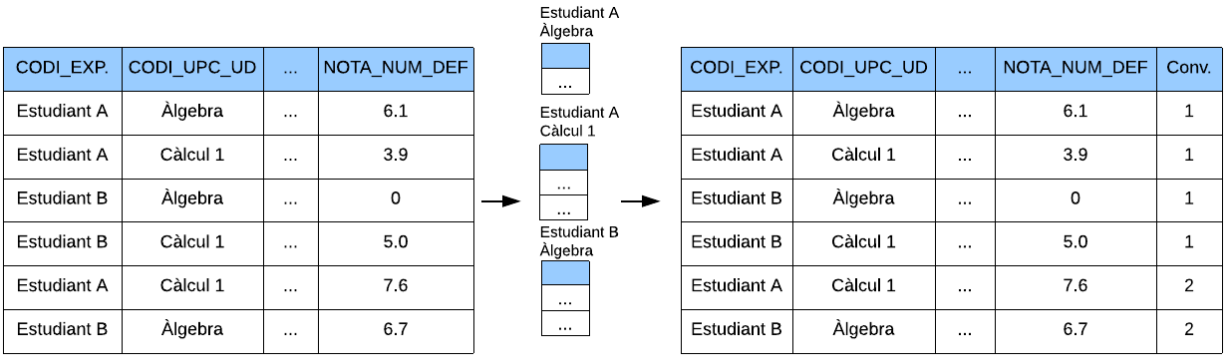

Fig. 3.2 Esquema del primer pas a seguir per obtenir la taula desitjada

Per crear la columna indicadora de la convocatòria s'utilitzen les eines de Pandas *groupby* i c*umcount*. Primer s'agrupen les dades de tots els estudiants en funció del CODI_EXPEDIENT i del CODI_UPC_UD amb *groupby* de manera que es fan grups de les dades disponibles de cada assignatura de cada estudiant. Això permet calcular el nombre de convocatòries realitzat amb *cumcount*, que numera els diferents elements que hi ha a cada grup. Com que numera els elements per ordre, és important que les agrupacions s'ordenin en funció de l'any i el quadrimestre tenint en compte les columnes CURS i QUAD. La nova columna creada conté els resultats d'aplicar *cumcount*.

Utilitzant la informació de la nova columna creada, la indicadora de la convocatòria, anomenada CONV a partir d'ara, i de la columna CODI_UPC_UD es crea la columna per pivotar.

CODI_EXP.	CODI_UPC_UD	...	NOTA_NUM_DEF	Conv.	C-A
Estudiant A	Àlgebra	...	6.1	1	N-Àlg-1
Estudiant A	Càlcul 1	...	3.9	1	N-Càl1-1
Estudiant B	Àlgebra	...	0	1	N-Àlg-1
Estudiant B	Càlcul 1	...	5.0	1	N-Càl1-1
Estudiant A	Càlcul 1	...	7.6	2	N-Càl1-2
Estudiant B	Àlgebra	...	6.7	2	N-Àlg-2

→

CODI_EXP.	N-Àlg-1	N-Àlg-2	N-Càl1-1	N-Càl1-2
Estudiant A	6.1	NaN	3.9	7.6
Estudiant B	0	6.7	5.0	NaN

Fig. 3.3 Esquema de la construcció de la taula de la proposta 1

Per la primera proposta, cal crear una columna que contingui informació de l'assignatura i la convocatòria, és a dir, que combini la informació de la columna CODI_UPC_UD amb la de CONV. Després s'aplica la funció pivot per pivotar la taula de manera que s'obtingui una taula on cada CODI_EXPEDIENT ocupi una sola fila, les columnes indiquin assignatura i convocatòria i els valors de la taula siguin les notes emmagatzemades inicialment a la columna NOTA_NUM_DEF.

CODI_EXP.	CODI_UPC_UD	...	NOTA_NUM_DEF	Conv.	N-A	C-A
Estudiant A	Àlgebra	...	6.1	1	N-Àlg	C-Àlg
Estudiant A	Càlcul 1	...	3.9	1	-	-
Estudiant B	Àlgebra	...	0	1	-	-
Estudiant B	Càlcul 1	...	5.0	1	N-Càl1	C-Càl1
Estudiant A	Càlcul 1	...	7.6	2	N-Càl1	C-Càl1
Estudiant B	Àlgebra	...	6.7	2	N-Àlg	C-Àlg

1 →

CODI_EXP.	N-Àlg	N-Càl1
Estudiant A	6.1	7.6
Estudiant B	6.7	5.0

+

CODI_EXP.	C-Àlg	C-Càl1
Estudiant A	1	2
Estudiant B	2	1

2 →

CODI_EXP.	N-Àlg	C-Àlg	N-Càl1	C-Càl1
Estudiant A	6.1	1	7.6	2
Estudiant B	6.7	2	5.0	1

Fig. 3.4 Esquema de la construcció de la taula de la proposta 2

Per la segona proposta, se seleccionen només les files de la taula inicial que contenen informació de l'última matrícula realitzada a cada assignatura. Es creen dues columnes, una servirà per crear una taula amb els valors de les notes en funció de l'assignatura, i l'altra per crear una taula amb els valors de la convocatòria en funció de l'assignatura. A partir d'aquestes columnes, es construeixen dues taules diferents amb *pivot* i a continuació s'uneixen amb la funció *merge* de pandas per construir la taula final.

En ambdues propostes anteriors, al construir les taules apareixen valors buits. Els valors buits han de desaparèixer perquè l'algorisme K-Means funcioni. Les opcions, igual que pel cas de les anomalies durant l'exploració de les dades, són eliminar-los o donar-los un valor.

Per la primera proposta els valors buits són deguts a que no s'ha matriculat una assignatura en una convocatòria concreta. És evident que un cop s'ha aprovat una assignatura no tornarà a matricular-se, per tant, per l'estructura de la taula, les columnes referents a una mateixa assignatura successives a la columna on s'aprova l'assignatura quedaran buides.

Per exemple, si s'aprova Àlgebra la primera vegada que es cursa, la columna nota-Àlgebra-1 contindrà nota però les columnes de la forma nota-Àlgebra-i amb i>1 quedaran buides. Per aquest cas, resulta "positiu" que quedin valors buits.

Es valora substituir els valors buits per un valor impossible i extrem però també considerat positiu, com per exemple el 20, així tot i afectar el càlcul de les distàncies de l'algorisme K-Means, el que es pretén és augmentar la similitud entre estudiants que tenen valors buits valorats positivament i augmentar la diferència envers els estudiants que han repetit l'assignatura i, per tant, tenen valor assignat.

El problema és que si es té en compte el volum total d'estudiants, n'hi ha que no arriben a matricular mai una assignatura, i aquests també tenen valors buits, considerats "negatius".

Per exemple un estudiant que després de suspendre-ho tot el primer quadrimestre decideix abandonar els estudis i per tant, no matricula ni Geometria ni cap de les assignatures del segon quadrimestre. Quan s'apliqui l'algorisme, aquest estudiant no pot ser similar a un altre estudiant que sí que les hagi matriculat. També es descarta predir el valor a assignar donat que podria manipular els futurs clústers.

Veient que és inviable assignar un nombre als valors buits que impedeixi que hi hagi similituds entre contraris, es descarta seguir utilitzant la taula de la proposta 1.

Seleccionar només els estudiants que superen la fase inicial solucionaria el problema dels valors buits "negatius" però obligaria a tractar els valors buits "positius" i tampoc es considera adequat per la desviació que comportaria.

Per la segona proposta, només es creen valors buits quan un estudiant no ha matriculat mai una assignatura (per exemple, l'estudiant D de la Taula 3.5), aleshores el valor de la columna indicadora de la nota de l'assignatura queda buit mentre que el valor de columna indicadora de la convocatòria d'aquesta assignatura és el 0. En aquest cas, tots els valors buits tenen el mateix significat.

Una opció seria assignar-los el 0, o un valor extrem negatiu com el -10 però, igual que amb la proposta 1 assignar qualsevol valor comportaria afegir una desviació a les dades.

Si se seleccionen només els estudiants que superen la fase inicial, aquesta taula no té valors buits a tractar i es considera que és la millor opció per treballar.

3.3. Construcció del model

L'etapa de construcció del model, és l'etapa on s'aplica el K-Means, l'algorisme de *clustering*, i comença de seguida que es disposa d'una matriu de dades adequada.

L'algorisme K-Means, tal com s'ha descrit al punt 2.1.1, necessita que es defineixi un nombre de clústers k per implementar-se.

Per determinar el nombre de clústers k de manera objectiva es fa servir el mètode del punt de colze o *elbow,* que és un mètode heurístic que valora la qualitat dels clústers aconseguits a partir d'aplicar el K-Means amb diferents valors de nombre de clúster k.

La llibreria Scikit-Learn proporciona una classe anomenada KMeans que en si mateixa és l'algorisme K-Means. Només treballa amb valors numèrics, no accepta ni valors buits ni categòrics. Tenint això present, a l'etapa de preparació de les dades s'han evitat els valors nuls i la matriu de dades només conté valors numèrics.

El random State és un paràmetre propi de la classe. L'algorisme K-Means comença definint k centroides aleatoris, donar un valor al random State permet determinar aquesta aleatorietat. Conèixer el random State serveix perquè cada vegada que s'apliqui l'algorisme s'obtingui el mateix resultat. El random State generat es farà servir a l'apartat de validació del model per descriure els clústers.

La inertia és un atribut propi de la classe, i és el resultat de calcular la suma de distàncies quadrades de mostres respecte al seu centroide. És una mesura de la qualitat del clúster. Conèixer-la serà necessari per decidir el nombre de clústers k.

3.3.1. Mètode del punt de colze

El mètode de buscar el punt de colze és el més tradicional i senzill de tots els mètodes que poden ajudar a trobar el nombre de clústers k.

Consisteix a executar l'algorisme K-Means amb un nombre k aleatori. Aleshores calcular una suma en funció de les distàncies entre cada punt i el seu centroide més proper, entenent aquesta suma com a mesura de l'error del clustering.

Com que l'augment dels clústers es correlaciona amb agrupacions i distàncies menors, aquesta suma sempre disminuirà quan k augmenta; com a exemple extrem, si escollim un valor k que sigui igual al nombre de punts de dades que tenim, la suma serà zero perquè cada

punt serà un clúster, i el seu centroide serà ell mateix.

L'objectiu d'aquest procés és trobar el punt en què augmentar k provocarà una disminució molt petita de la suma d'errors, mentre que la disminució de k augmentarà de forma pronunciada la suma d'error. Aquest punt s'anomena "punt del colze".

Per fer els càlculs de la qualitat dels clústers per diferents nombres de clúster k, es defineix una funció que donada una matriu de dades amb les dades dels estudiants i un nombre màxim de clústers max_clusters, calcula la qualitat del resultat d'aplicar l'algorisme K-Means per cada k ϵ (1, max_clusters). La mesura de la qualitat és l'atribut inertia.

La funció retorna un gràfic que mostra el "punt de colze" i una llista ordenada de millor a pitjor del nombre de clúster, la inèrcia i el RandomState calculat per cada k ϵ (1, max_clusters).

Per implementar la funció es tria un nombre màxim de clústers max_clusters=10. Com s'ha comentat abans cada vegada que augmenta k millora la inèrcia dels clústers i per tant la qualitat, però no interessa trobar el punt que fa que la inèrcia sigui millor sinó el que proporciona un canvi significatiu en la qualitat. Per això s'ha valorat que no és necessari augmentar el nombre màxim de clústers.

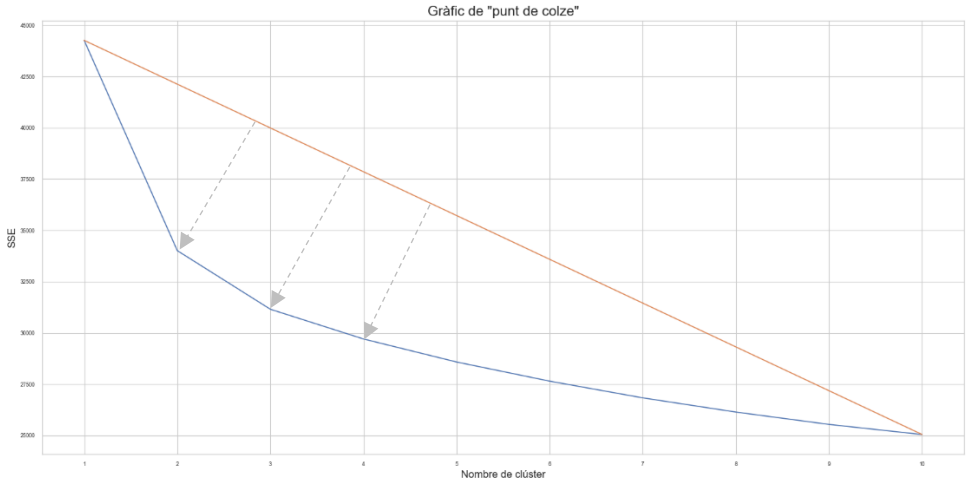

Fig. 3.5 Gràfic de punt de colze

Com pot veure's a la Fig. 3.5 no hi ha un punt de colze pronunciat però per k=2,3 i 4 sí que es produeix una inflexió destacada. S'estudiaran els resultats d'agafar aquestes mides de clúster.

3.4. Validació dels resultats

En aquesta etapa es descriuen els clústers trobats després d'aplicar l'algorisme K-Means

tenint en compte els resultats d'aplicar el mètode del punt de colze.

Per valorar els resultats obtinguts, es fa servir la classe SelectKBest de la llibreria Scikit-Learn. Aquesta classe, a partir dels valors de cada columna i dels clústers assignats, calcula una mesura de dependència entre cada columna i clúster, retornant les k columnes més correlacionades amb el clúster.

Per aquest projecte, es té una matriu de dades de 20 columnes (10 per notes de les assignatures de la fase inicial i 10 indicadores de la convocatòria a la qual s'ha aprovat cada assignatura), s'ha decidit buscar les 4 més rellevants. Per tant, les columnes seleccionades seran o bé corresponents a notes d'assignatures o bé corresponents a la convocatòria en què s'han aprovat.

La funció creada a partir de SelectKBest retorna el nom de les columnes més rellevants i fa una descripció estadística de cada clúster indicant-ne la mitjana, la desviació, els percentils, i els valors mínim i màxim.

Després d'aplicar SelectKBest, s'estudien els clústers obtinguts mitjançant gràfics de les llibreries matplotlib i seaborn. I s'interpreten tant les descripcions de les columnes més rellevants com els gràfics per caracteritzar els grups obtinguts.

3.4.1. K-Means amb nombre de clústers k=2

CLUSTER 0

	nota-Càl1	nota-MecFon	nota-Quí1	nota-Geo
count	687.00	687.00	687.00	687.00
mean	7.49	7.11	7.82	7.09
std	1.16	1.06	1.18	1.08
min	5.00	5.00	5.00	5.00
25%	6.70	6.40	7.00	6.30
50%	7.40	7.10	7.90	7.00
75%	8.20	7.80	8.70	7.80
max	10.00	10.00	10.00	10.00

CLUSTER 1

	nota-Càl1	nota-MecFon	nota-Quí1	nota-Geo
count	2132.00	2132.00	2132.00	2132.00
mean	5.93	5.75	6.24	5.83
std	0.89	0.75	1.01	0.77
min	5.00	5.00	5.00	5.00
25%	5.10	5.00	5.38	5.10
50%	5.70	5.60	6.10	5.70
75%	6.50	6.20	7.00	6.30
max	9.80	9.40	9.80	9.00

Fig. 3.6 Descripció dels clústers per k=2

A Fig. 3.6 pot observar-se que les columnes més rellevants són les referides a les notes de Càlcul 1, Mecànica Fonamental, Química 1 i Geometria.

Per visualitzar i poder entendre els clústers creats, s'ha decidit projectar en un scatterplot les columnes rellevants cadascuna envers les altres i en un altre scatterplot la distribució de les notes i convocatòries de cada clúster per les diferents assignatures de la fase inicial.

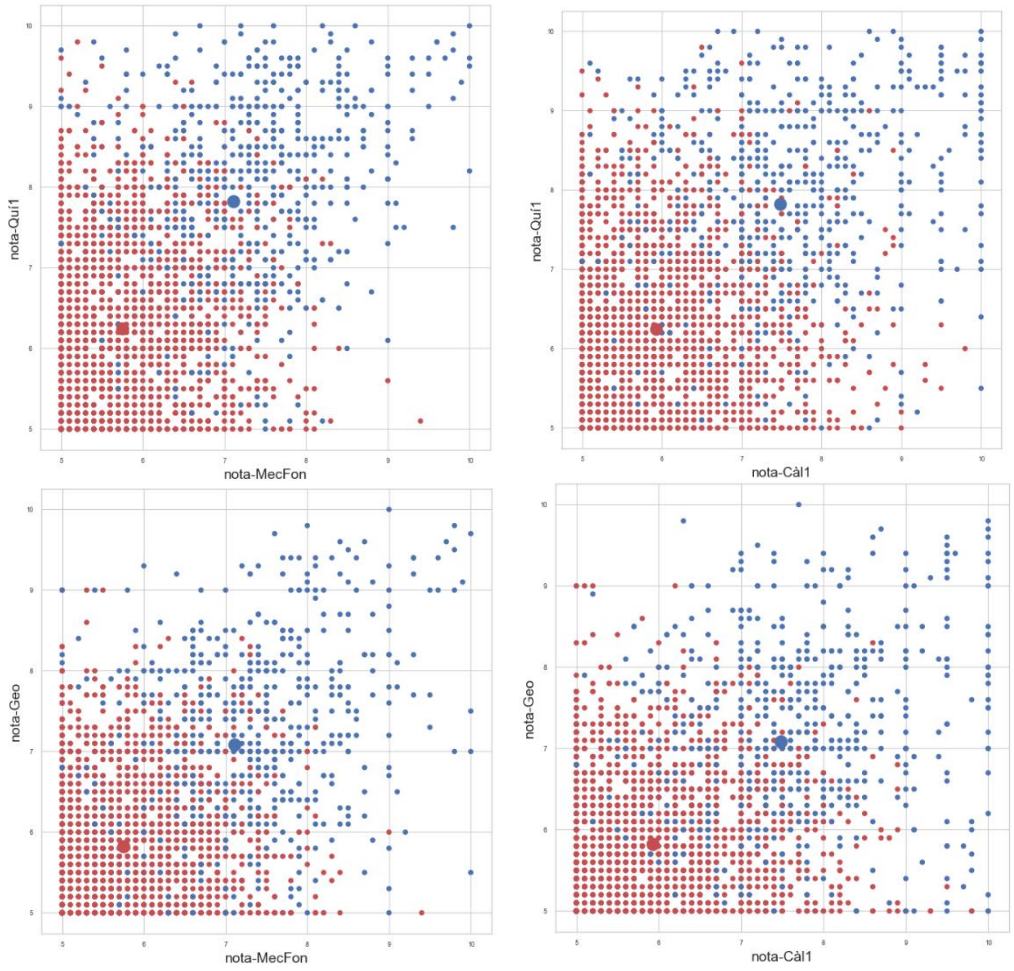

Fig. 3.7 Selecció d'scatterplots de les columnes més discriminants per k=2

La Fig. 3.7 mostra els dos clústers creats i els seus centroides. Els punts de color blau i de color vermell corresponen respectivament als estudiants del CLUSTER 0 i del CLUSTER 1 de la Fig. 3.6.

S'observa que els estudiants del clúster 0 tenen tendència a treure notes que van del notable a l'excel·lent i els estudiants del clúster 1 aprovats. Els centroides dels dos clústers estan alineats en una diagonal, que per Química 1 es percep desplaçada cap a notes més altes. La tendència dels estudiants a treure millor nota a Química 1 pot veure's en els dos clústers i també es veuria sense indicar els clústers. A més, es pot intuir una línia diagonal perpendicular a la que dibuixen els centroides que separa els clústers. El clúster 0 conté menys estudiants i és dispers, el clúster 1 agrupa un volum d'estudiants més gran i és més compacte, amb una

aparença similar a la d'un triangle. El triangle que es forma no és un equilàter perfecte perquè dins del clúster 1 hi ha estudiants que tot i treure notes més justes a una assignatura milloren el seu rendiment en d'altres. Per exemple, el triangle que es forma quan es fa una comparativa amb Química 1 tendeix a créixer cap a notes més altes d'aquesta assignatura. En canvi, el triangle que es forma al comparar Mecànica Fonamental i Geometria sí que té una aparença més simètrica.

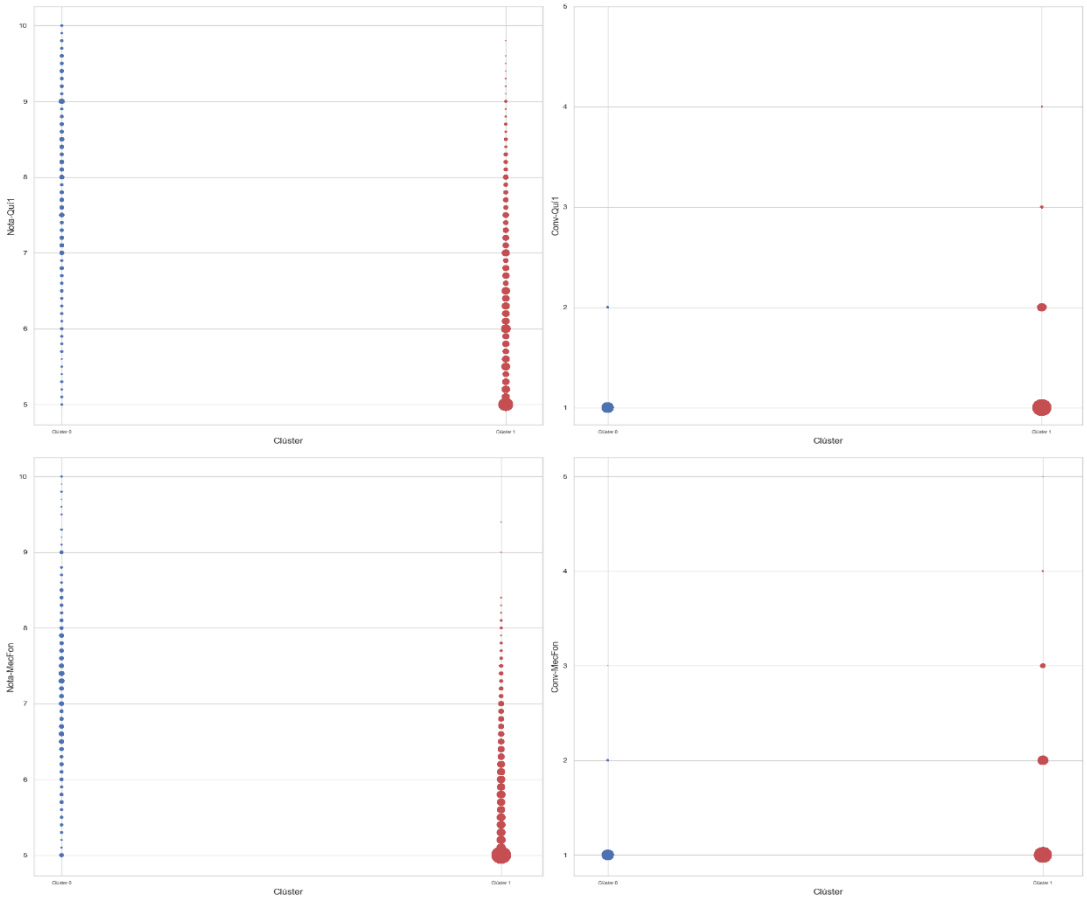

Fig. 3.8 Selecció d'scatterplots que mostren el clúster envers la nota i envers la convocatòria per k=2

La Fig. 3.8 conté dos scatterplots que mostren els clústers creats envers la nota que obtenen (esquerra) i la convocatòria en què l'obtenen (dreta). La mida dels punts que conformen l'scatterplot es dibuixa en funció del volum d'estudiants que representa, serveix per veure que hi ha més estudiants al clúster 1 i també per veure que gran part dels estudiants d'aquest clúster treuen un 5 a les assignatures que es mostren. La convocatòria en què s'aproven les assignatures no és un factor gaire rellevant, tot i això al visualitzar la distribució de convocatòries en funció dels clústers (dreta de la Fig. 3.8), s'observa que els estudiants del

primer clúster participen a menys convocatòries que els altres. Pel que fa a la tendència dels clústers envers les notes i el volum d'estudiants es considera més entenedora la visualització de la Fig. 3.7.

En conclusió, amb dos clústers podria dir-se que s'obté una divisió entre estudiants que treuen notes molt bones (a partir de notable) i estudiants que no destaquen i "només aproven". També pot veure's, encara que no caldria tenir en compte els clústers, que hi ha assignatures on les notes acostumen a ser més altes (Química 1) i assignatures on el rendiment dels estudiants és similar, com Geometria i Mecànica Fonamental.

3.4.2. K-Means amb nombre de clústers k=3

S'estudien els resultats d'aplicar l'algorisme K-Means amb k=3.

```
CLUSTER 0
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count     421.00       421.00     421.00       421.00
mean        7.85         7.52       8.13         8.25
std         1.10         0.95       1.13         1.17
min         5.00         5.00       5.00         5.00
25%         7.10         6.90       7.40         7.50
50%         7.80         7.40       8.20         8.40
75%         8.50         8.10       9.00         9.20
max        10.00        10.00      10.00        10.00
CLUSTER 1
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count    1397.00      1397.00    1397.00      1397.00
mean        5.79         5.63       5.95         6.03
std         0.83         0.69       0.85         0.91
min         5.00         5.00       5.00         5.00
25%         5.00         5.00       5.20         5.20
50%         5.50         5.40       5.80         5.90
75%         6.20         6.00       6.50         6.60
max         9.50         8.40       9.00         9.50
CLUSTER 2
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count    1001.00      1001.00    1001.00      1001.00
mean        6.38         6.11       6.95         7.62
std         1.01         0.85       1.09         1.09
min         5.00         5.00       5.00         5.00
25%         5.60         5.40       6.10         7.00
50%         6.30         6.00       7.00         7.60
75%         7.00         6.70       7.70         8.30
max        10.00         9.40       9.80        10.00
```

Fig. 3.9 Descripció dels clústers per k=3

Les columnes més rellevants són d'assignatures del primer quadrimestre, concretament Càlcul 1, Mecànica Fonamental, Química 1 i Fonaments d'Informàtica.

Pot veure's que la mitjana dels estudiants del CLUSTER 0 és un notable alt per les quatre assignatures i la del CLUSTER 1 és propera al 6 per les quatre assignatures mentre que, pel CLUSTER 3 la mitjana de Càlcul 1, Mecànica Fonamental i Química és superior a 6 i la de Fonaments d'Informàtica és superior a 7. Així, mentre que la diferència entre el primer clúster i la resta es manté en totes les assignatures, la diferència entre el segon i el tercer clúster s'accentua quan l'assignatura és Fonaments d'Informàtica.

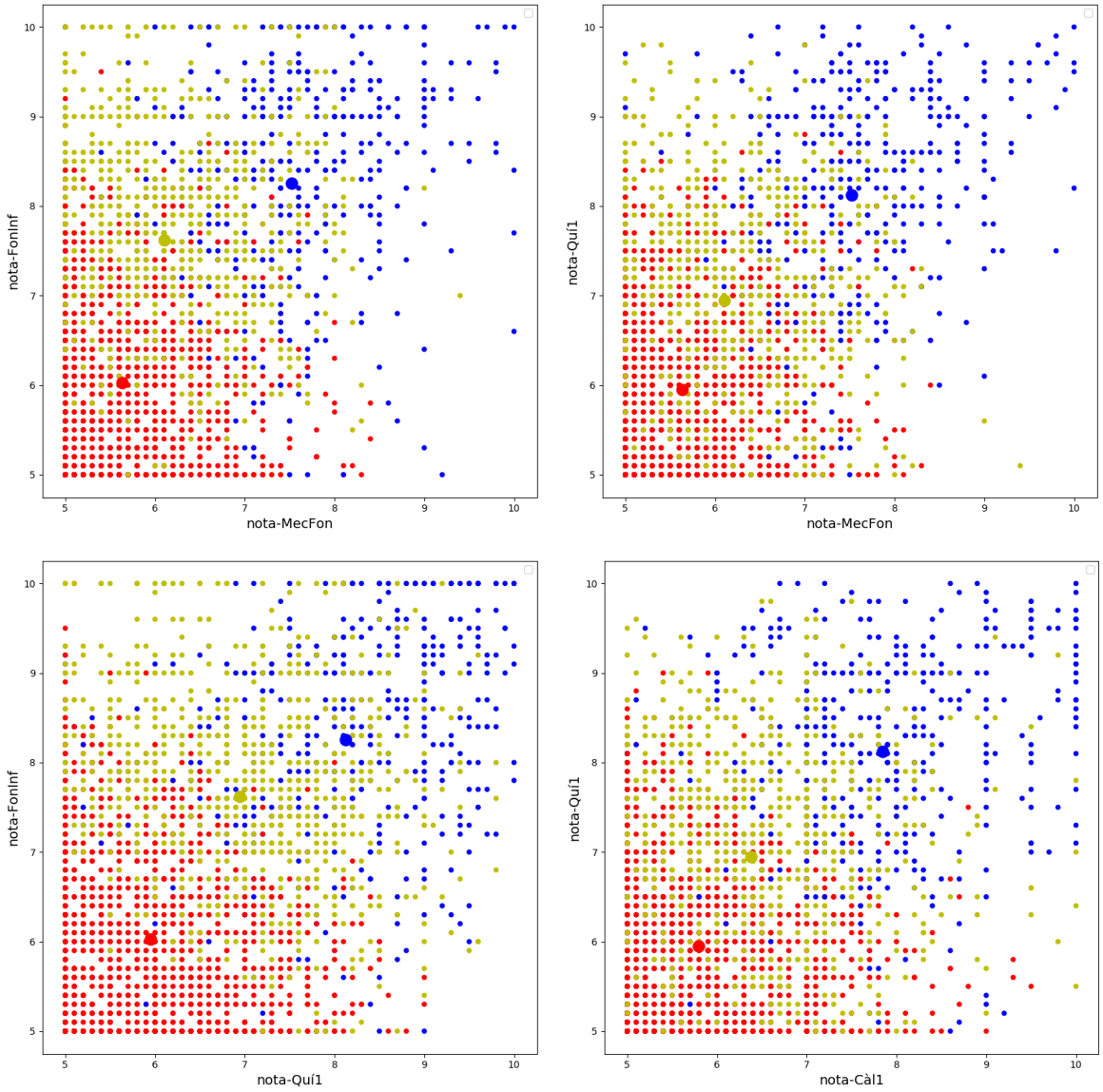

Fig. 3.10 Selecció d'scatterplots de les columnes més discriminants per k=3

A la Fig. 3.10 poden veure's els clústers creats i els seus centroides amb els colors blau, vermell i groc corresponent al CLUSTER 0, al CLUSTER 1 i al CLUSTER 2 de la Fig. 3.9 respectivament.

Si s'uneixen els centroides dels clústers, pot veure's que el clúster 2 es desplaça cap a les notes més altes de Química 1 i encara més altes de Fonaments d'Informàtica. En general, també es veu la tendència dels estudiants a treure millors notes a Química i a Fonaments d'Informàtica que a Càlcul 1 i a Mecànica Fonamental. El clúster 1 és compacte en comparació

als altres.

Fig. 3.11 Selecció d'scatterplots de les columnes més discriminants separats per clúster per k=3

Per millorar la visualització, especialment havent afegit més clústers, s'introdueix un nou gràfic que permet veure per separat la distribució dels estudiants a cada clúster. Es fa servir la classe FacetGrid de la llibreria Seaborn. Respecte al comportament de cada clúster, a la Fig. 3.11

s'observa que el clúster 0 (blau) és bastant dispers i poc compacte però en general queda desplaçat cap a les notes altes. A mot intuir-se una correlació o relació més lineal entre les notes de Mecànica Fonamental i Química 1 que entre les notes de Mecànica Fonamental i Càlcul 1. El clúster 1 (vermell) és bastant compacte i dibuixa un triangle, que és més equilàter quan es compara Mecànica Fonamental i Càlcul 1 i que perd la simetria quan es compara qualsevol assignatura amb Química 1 i amb Fonaments d'Informàtica, anant cap a notes més altes d'aquestes assignatures. El clúster 2 (groc), és menys estable i canvia el seu comportament en funció de l'assignatura cosa que fa que el seu estudi resulti interessant.

Si s'estudia amb més detall el clúster 2, al gràfic superior el comportament envers Càlcul 1 i Mecànica Fonamental és similar al del clúster 1 encara que el dibuix del clúster 2 tendeix més a un quadrat que a un triangle indicant que els estudiants d'aquest grup es troben aproximadament dins dels mateixos límits de nota en ambdues assignatures. Al gràfic central pot intuir-se que el grup 2 dibuixa un rectangle delimitat per notes altes de Química 1 i no tan altes de Mecànica Fonamental, el comportament pel que fa a Mecànica Fonamental és similar al del clúster 1. Al gràfic inferior, a la comparativa entre Química 1 i Fonaments d'Informàtica hi ha força dispersió, cosa que suggereix que dins aquest grup hi ha un ventall molt variat de resultats. Això en particular fa pensar que pot ser interessant veure el comportament en aquestes assignatures provant k = 4.

Fig. 3.12 Scatterplot tipus swarmplot de la convocatòria en que s'aprova envers la nota que s'obté.

Per veure com es distribueixen els clústers segons la convocatòria en què s'aprova l'assignatura i la nota que s'obté s'ha decidit canviar de gràfic respecte al que s'utilitza a la Fig. 3.8 per un gràfic de la llibreria seaborn anomenat swarmplot. (Fig. 3.12). Aquest gràfic és un scatterplot on no se superposen els punts. L'eix horitzontal indica la convocatòria en què s'ha aprovat l'assignatura i el vertical la nota que s'ha tret. En termes generals s'observa que la convocatòria no és rellevant envers la nota que treuen els estudiants dels diferents clústers. Els estudiants del clúster 0 (en blau) treuen millors notes i la majoria només realitzen una convocatòria, en canvi, els estudiants del clúster 1 realitzen més convocatòries. Resulta curiós veure que, per moltes convocatòries que realitzin no obtenen notes més altes.

En conclusió, amb tres clústers podria dir-se que s'obté una divisió entre estudiants que treuen notes molt bones, els estudiants que treuen notes "mitjanes" (notables en algunes assignatures i aprovats en altres) i estudiants que no destaquen i "només aproven".

3.4.3. K-Means amb nombre de clústers k=4

```
CLUSTER 0
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count     364.00       364.00     364.00       364.00
mean        7.87         7.61       8.26         8.46
std         1.12         0.94       1.06         1.06
min         5.00         5.00       5.00         5.00
25%         7.10         7.00       7.57         7.80
50%         7.80         7.50       8.40         8.50
75%         8.52         8.20       9.10         9.30
max        10.00        10.00      10.00        10.00
CLUSTER 1
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count    1155.00      1155.00    1155.00      1155.00
mean        5.76         5.60       5.94         5.84
std         0.80         0.68       0.85         0.73
min         5.00         5.00       5.00         5.00
25%         5.00         5.00       5.20         5.10
50%         5.50         5.40       5.80         5.70
75%         6.20         6.00       6.50         6.40
max         9.50         8.40       9.00         8.00
CLUSTER 2
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count     607.00       607.00     607.00       607.00
mean        6.87         6.23       7.02         6.85
std         1.04         0.88       1.14         1.05
min         5.00         5.00       5.00         5.00
25%         6.10         5.50       6.20         6.00
50%         7.00         6.10       7.00         6.90
75%         7.50         6.90       7.80         7.60
max        10.00         9.40       9.80         9.40
CLUSTER 3
       nota-Càl1  nota-MecFon  nota-Quí1  nota-FonInf
count     693.00       693.00     693.00       693.00
mean        5.91         5.95       6.58         8.01
std         0.83         0.81       1.06         0.88
min         5.00         5.00       5.00         5.60
25%         5.20         5.20       5.80         7.40
50%         5.70         5.80       6.50         8.00
75%         6.50         6.50       7.30         8.50
max         9.50         9.00       9.60        10.00
```

Fig. 3.13 Descripció dels clústers per k=4

A la Fig. 3.13 es mostra la descripció estadística de les columnes més rellevants que són Càlcul 1, Mecànica Fonamental, Química 1 i Fonaments d'Informàtica. El resultat d'aplicar SelectKBest diferents vegades també ha considerat Àlgebra en comptes de Càlcul 1 perquè quan es fa el càlcul i es produeix un empat es decideix aleatòriament proporcionar una columna o l'altra. Això vol dir que el comportament dels clústers envers ambdues assignatures és similar.

Els estudiants del clúster 0 tenen tendència a treure notes properes a l'excel·lent, els del clúster 1 a treure notes properes al 6, els del clúster 2 properes al 7 i els del clúster 3 mentre que tenen tendència a treure notes properes al 6 a la majoria d'assignatures i tenir un comportament semblant al dels estudiants del clúster 1, a Fonaments d'Informàtica són estudiants que milloren el seu rendiment.

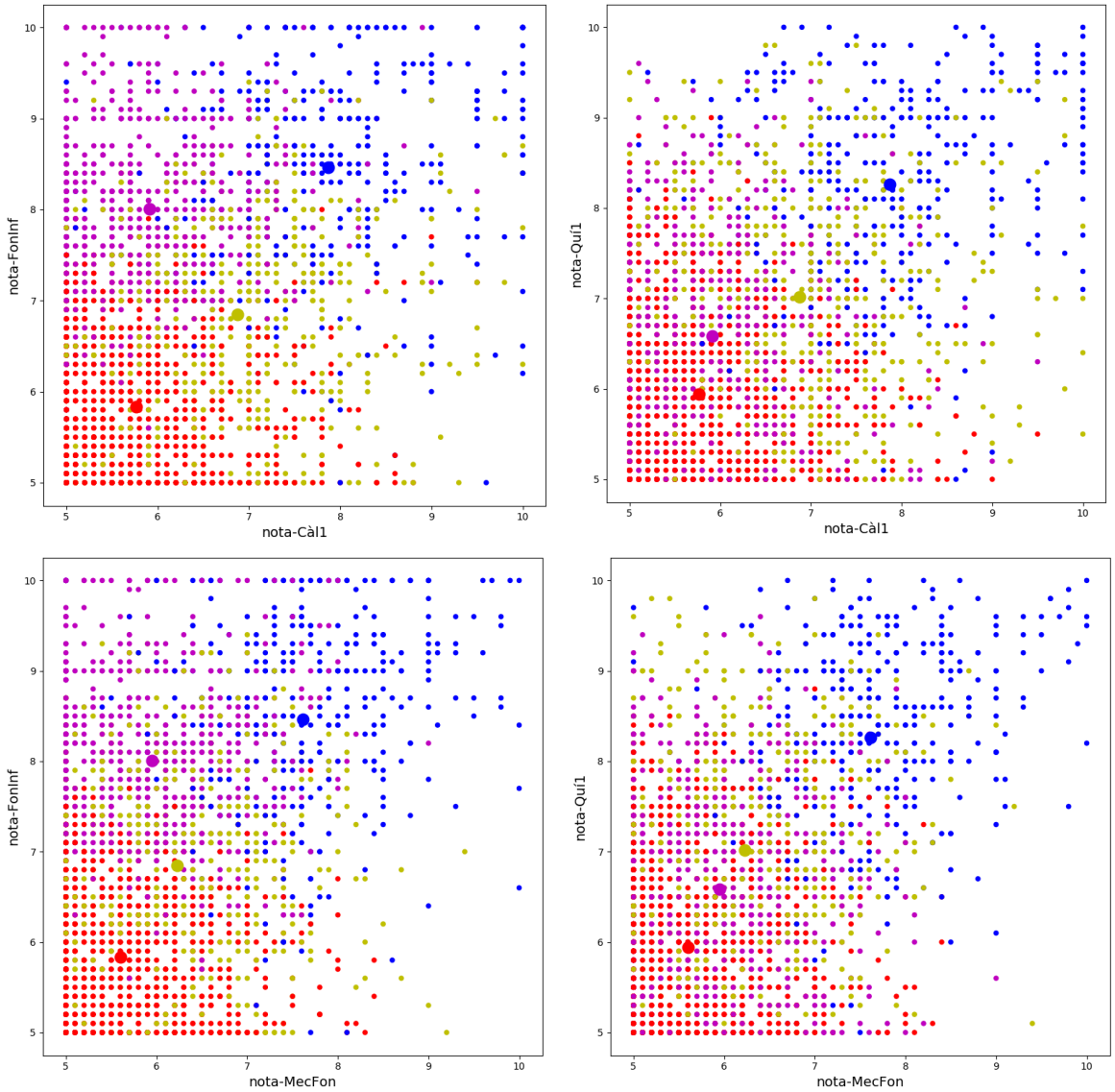

Fig. 3.14 Selecció d'scatterplots de les columnes més discriminants per k=4

A la Fig. 3.14 poden veure's els clústers creats i els seus centroides amb els colors blau, vermell groc i lila corresponents al CLUSTER 0, al CLUSTER 1, al CLUSTER 2 i al CLUSTER 3 de la Fig. 3.13 respectivament.

Si s'uneixen els centroides dels clústers, és habitual trobar els dels clústers 0,1 i 2 força alineats.

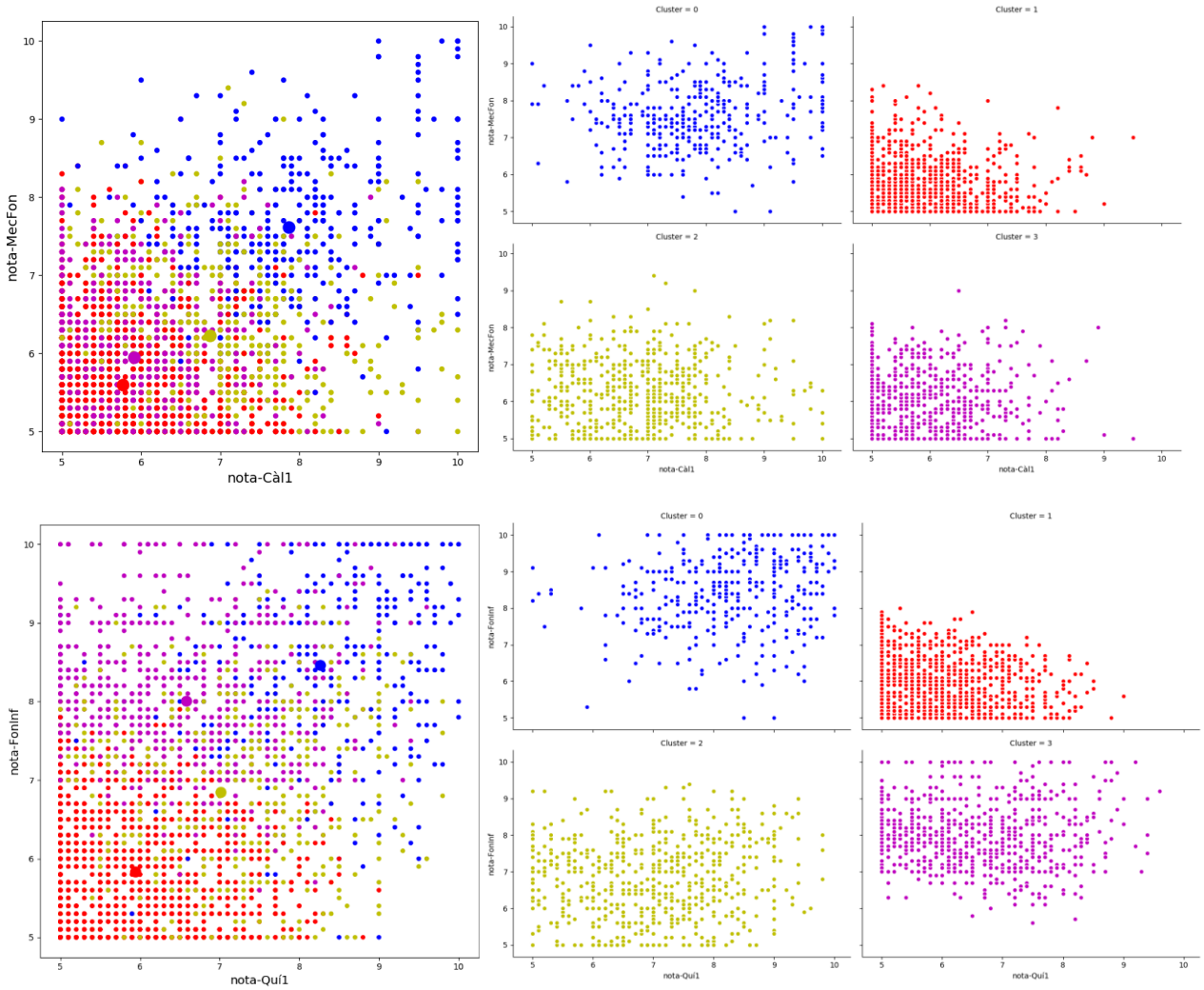

Fig. 3.15 Selecció d'scatterplot conjunt (esquerra) i separat (dreta) pels diferents clústers amb k=4

A la Fig. 3.15 s'observen els diferents clústers i la seva distribució, s'han seleccionat dues comparatives interessants. Als gràfics superiors la comparativa entre Càlcul 1 i Mecànica Fonamental mostra com el comportament del clúster 1 i del clúster 3 és molt similar. Es pot intuir una línia vertical, corresponent al 8.5 a Càlcul 1 que delimita tots dos clústers i també una línia horitzontal, corresponent al 8.5 a Mecànica fonamental que també els delimita. El clúster 2 també queda delimitat per aquesta línia horitzontal. Als gràfics inferiors, la comparativa entre Química 1 i Informàtica Fonamental mostra que el clúster 3 està format per estudiants que treuen notes més altes a Fonaments d'Informàtica, en canvi, es comporten de manera semblant al clúster 2 pel que fa a les notes que obtenen a Química 1.

4. Planificació del projecte

Abans de començar el projecte es va fer aquesta planificació per marcar el ritme de treball durant el quadrimestre.

Tasca	Durada (set.)	SET.	OCT.	NOV.	DES.	GEN.
Introducció a la mineria de dades	3					
Formació en les eines de mineria de dades	12					
Comprensió del negoci	2					
Comprensió de les dades	4					
Preparació de les dades	7					
Construcció del model	8					
Validació del model	4					
Redacció de la memòria	3					

Taula 4.1 Planificació del projecte

En termes generals s'ha seguit la planificació, però les etapes de construcció i validació del model s'han allargat una mica més del que s'esperava.

5. Pressupost

Els costos a considerar per realització d'aquest projecte corresponen principalment a costos de personal. També es consideren costos derivats de llicències i l'amortització de l'ordinador amb la que s'ha realitzat tot el projecte.

Pel que fa al personal, es calculen les hores dedicades a la realització del projecte i també les hores dedicades a consultes amb el tutor. S'han realitzat entre dues i tres reunions al mes.

	Dedicació (h)	Preu (€/h)	Cost (€)
Introducció a la mineria de dades	40	20	480
Comprensió del negoci	12	20	240
Comprensió de les dades	32	30	960
Preparació de les dades	126	30	3780
Construcció del model	120	30	3600
Validació del model	40	30	1200
Redacció de la memòria	26	20	520
Consultes amb expert	18	60	1.080
		TOTAL	12.180

Taula 5.1. Costos de personal

Pel que fa a les llicències, el software utilitzat per realitzar les etapes del procés de mineria de dades és de programari lliure i de codi obert i gratuït. Es calcula el cost proporcional de l'ús realitzat envers el preu de la llicència de Microsoft Office, encara que es podria haver fet servir una alternativa de programari lliure. Es considera que es fa servir unes 1500 h a l'any.

	Preu llicència (€/any)	Dedicació (h)	Preu (€/h)	Cost (€)
Microsoft Office 2013	75,99	480	0,05	24,3
			TOTAL	24,3

Taula 5.2. Costos de llicències

Pel que fa a l'amortització de l'ordinador, va costar 800 € i es va comprar fa un any. Es considera un cost de manteniment anual del 10% del preu de compra.

	Amortització (€/any)	Dedicació (h)	Preu (€/h)	Cost (€)
Amortització ordinador	80	480	0,06	25,6
			TOTAL	25,6

Taula 5.3. Costos d'amortització de l'ordinador

6. Impacte Ambiental

Aquest projecte de mineria de dades no produeix un impacte ambiental rellevant. No existeix ni en deriva cap fase experimental en el medi.

Donat que s'ha treballat amb ordinador, es calculen les emissions de CO2 derivades al consum elèctric de l'ús de l'ordinador per veure de quin ordre són a partir del mix elèctric.

El mix elèctric és el valor que expressa les emissions de CO_2 associades a la generació de l'electricitat que es consumeix. Segons l'Oficina Catalana del Canvi Climàtic el mix de la xarxa elèctrica peninsular de 2017 s'estima en 392 g CO_2/kWh.

Tenint en compte la planificació del projecte, es considera un mínim de 450 hores d'utilització de l'ordinador.

Si el consum de l'ordinador és d'uns 200 W, el consum d'energia és de 90kWh i l'emissió és de 35 kg CO_2.

No es té en compte el possible consum energètic a causa de la il·luminació o climatització del lloc de treball, per una banda perquè en gran part s'ha treballat de dia amb llum natural i per l'altre perquè hi hauria hagut consum encara que no es treballés en el projecte.

Conclusions

Es valora molt positivament el fet que s'han assolit els objectius proposats a l'inici del projecte.

S'ha adaptat la metodologia CRISP-DM a les característiques del treball i durant tot el procés s'han justificat la presa de decisions. A l'etapa de comprensió de les dades, s'han detectat estudiants anòmals i s'ha raonat com tractar-los. Després s'han seleccionat, filtrat i transformat les dades fins que han tingut una forma adequada per aplicar l'algorisme K-Means. A l'etapa de construcció del model s'ha decidit quants clústers estudiar a partir del mètode del punt de colze i s'ha aplicat l'algorisme K-Means per grups de 2, 3 i 4. Finalment, a l'etapa de validació, s'ha vist la utilitat del clustering i s'han pogut caracteritzar els clústers creats. S'han estudiat diversos formats de visualització per validar els resultats i s'han trobat els més adients per poder entendre les tendències dels grups. S'ha vist que hi ha assignatures, com Química 1, on el rendiment dels estudiants tendeix a ser millor independentment del grup a què corresponen. Quan es fan tres grups, hi ha un grup interessant, el del nivell mitjà, que seria interessant estudiar per separat perquè mostra diferents comportaments segons les assignatures. S'ha vist que la relació en aquest grup entre les notes de Fonaments d'Informàtica i Química 1 no acaba de mostrar cap patró i la forma que pren el clúster és molt dispersa. En fer quatre grups s'ha vist que hi ha estudiants que tenen un millor rendiment a l'assignatura de Fonaments d'Informàtica. Per tres i quatre grups, també s'ha vist que el grup de nivell alt, sí que podria haver-hi una relació més directa (o lineal) entre les notes de les diferents assignatures.

Per caracteritzar millor els grups, caldria enriquir les dades de què es disposa. Es podria fer servir la informació referent al codi postal i/o a la nota de la selectivitat. Per fer clústers seria interessant veure si té algun efecte l'any de naixement de l'estudiant, per poder comparar el comportament de les diferents generacions. També seria interessant que les dades tinguessin més informació sobre el rendiment dels estudiants, per exemple informació sobre l'avaluació continuada, els parcials, la seva assistència a classe...

L'algorisme K-Means de la llibreria SciKit-Learn presenta algunes limitacions, no es poden tractar els valors buits ni valors no numèrics, a més s'han realitzat els càlculs amb la distància euclidiana. Caldria estudiar aplicar algun altre algorisme de clustering que detectés altres tipus de grups o que fes servir una altra mesura per la distància per trobar grups amb formes diferents. També seria interessant poder incloure els estudiants amb valors buits a l'estudi. Les limitacions han servit per veure que queda molt per fer i que es poden portar a terme treballs futurs molt diversos relacionats amb aquest tema.

Personalment, aquest treball m'ha servit per ser conscient de com és d'important l'organització i tenir clars no només els objectius del projecte sinó les diferents etapes a seguir. Ha estat interessant aplicar el clustering perquè, al ser una tècnica no supervisada, no hi ha una mesura ferma per saber si s'està fent bé (amb tècniques supervisades sí que es pot mesurar el % d'encert) i durant totes les etapes del projecte he hagut de pensar i prendre decisions sense tenir una seguretat darrere ni una única resposta correcta.

Des del meu punt de vista en aquest tipus de treballs no hi ha males decisions si són justificades i el conjunt de decisions que es prenen és coherent, però és difícil actuar amb seguretat quan no s'està acostumat a treballar així. En aquest aspecte, potser trobo a faltar al Grau alguna assignatura més on es treballi d'aquesta manera.

M'ha agradat molt tornar a treballar amb Python, sobretot valoro positivament l'entorn d'Spyder que penso que és molt útil. També m'ha agradat buscar diferents maneres de visualitzar els clústers per entendre'ls.

En conclusió valoro molt positivament aquest projecte, no tan sols per tot el que he après de mineria de dades sinó també pel que m'ha ensenyat de com desenvolupar un projecte i prendre decisions.

Agraïments

Al meu tutor Lluis Talavera. Per la seva disponibilitat i ajuda.

A la família i amics. Pel seu suport, sempre.

Bibliografia

Referències bibliogràfiques

[1] DAVID L. OLSON, DURSUN DELEN. *Advanced Data Mining Techniques*, 2008

[2] Machine Learning Algorithms Explained - K-Means Clustering. [https://blog.easysol.net/machine-learning-algorithms-3/]

[3] KD-Nuggets.*CRISP-DM, still the top methodology for analytics, data mining, or data science projects.* [https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html]

[4] Wes McKinney. *Python for Data Analysis*. O'Reilly, first edition octubre 2012.

[5] NumPy [http://www.numpy.org/]

[6] Pandas.Python Data Analysis Library [https://pandas.pydata.org/]

[7] SciKit-Learn. *Machine Learning in Python* [https://scikit-learn.org]

[8] Matplotlib.[https://matplotlib.org/]

[9] Seaborn:statistical data visualization [https://seaborn.pydata.org/]

[10] Oficina Catalana del Canvi Climàtic. *Factor d'emissió associat a l'energia elèctrica: el mix elèctric* [http://canviclimatic.gencat.cat/ca/redueix_emissions/com-calcular-emissions-de-geh/factors_demissio_associats_a_lenergia/]

Bibliografia complementària

[1] JIAWEI HAN, MICHELINE KAMBER, JIAN PEI. *Data Mining Concepts and Techniques*, 3ra edició 2012

[2] HERNÁNDEZ, RAMÍREZ, FERRI *Introducción a la Minería de Datos*. Edició 2008

[3] K-Means Clustering in Python [https://mubaris.com/posts/kmeans-clustering/]

[4] https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

[5] https://dzone.com/articles/r-or-python-data-scientists-delight

[6] https://www.stoodnt.com/blog/r-vs-python-metareview-usability-popularity-pros-cons-jobs-salaries/