TecnoCampus
Escola Superior
Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona

Grau en Enginyeria Informàtica de Gestió i Sistemes d'Informació

ESTUDI I ANÀLISI DE LES CAUSES D'ABANDONAMENT D'ESTUDIS DELS FUTURS ENGINYERS INFORMÀTICS

Memòria

JON MORALES MARTÍ

SANDRA OBIOL MADRID

2023-24

TecnoCampus
Mataró-Maresme

Abstract

This final degree project analyzes the causes of student dropout at TecnoCampus. Using data analysis techniques, patterns have been identified that help understand this phenomenon. The results obtained allow for proposing strategies to improve student retention and optimize educational resources. This study contributes to improving the quality of education and offers specific recommendations for improving the areas most affected according to former students.

Resum

Aquest treball de fi de grau analitza les causes d'abandonament dels estudiants al TecnoCampus. Mitjançant tècniques d'anàlisi de dades, s'han identificat patrons que ajuden a comprendre aquest fenomen. Els resultats obtinguts permeten proposar estratègies per millorar la retenció d'estudiants i optimitzar els recursos educatius. Aquest estudi contribueix a millorar la qualitat de l'educació i ofereix recomanacions específiques per a la millora de les àrees més afectades per part d'exalumnes.

Resumen

Este trabajo de fin de grado analiza las causas de abandono de los estudiantes en TecnoCampus. Mediante técnicas de análisis de datos, se han identificado patrones que ayudan a comprender este fenómeno. Los resultados obtenidos permiten proponer estrategias para mejorar la retención de estudiantes y optimizar los recursos educativos. Este estudio contribuye a mejorar la calidad de la educación y ofrece recomendaciones específicas para la mejora de las áreas más afectadas según los exalumnos.

1.Introducció

Aquest Treball de Fi de Grau té com a objectiu principal l'anàlisi del rendiment acadèmic i l'abandonament estudiantil al TecnoCampus. Mitjançant l'anàlisi de dades, es busca identificar patrons i raons subjacents d'aquest fenomen amb l'objectiu de millorar la retenció d'estudiants i optimitzar els recursos educatius.

A més, es pretén proporcionar informació rellevant sobre les àrees prioritàries de millora al centre i les assignatures que cal potenciar per enriquir l'experiència educativa.

Aquest treball ofereix una contribució significativa al camp de l'educació superior, proporcionant una comprensió més profunda de les causes de l'abandonament dels estudis universitaris.

2. Marc teòric i anàlisi de referents

Context

[1] El TecnoCampus és un parc tecnològic situat a Mataró (Barcelona) que destaca per les seves tres escoles universitàries vinculades a la Universitat Pompeu Fabra, a més d'incloure un parc empresarial i una incubadora d'empreses.

En els últims anys, s'ha observat una millora tecnològica al centre, per continuar amb aquesta tendència, la universitat ha decidit fer un canvi de direcció i es vol convertir en una de les primeres universitats a nivell mundial en convertir-se en una universitat data-driven.

Amb l'objectiu de convertir el centre en una organització data-driven, el TecnoCampus està implementant una estratègia integral. Aquesta estratègia inclou la realització d'un estudi dedades per identificar patrons de comportament entre els estudiants. Això permetrà millorar l'experiència educativa dels alumnes i desenvolupar un programa que pugui predir amb més precisió si un estudiant completarà o no els seus estudis.

Antecedents

L'escola disposa de tota la informació dels estudiants, incloent aquells que ja s'han graduat, els que han abandonat i els que estan cursant actualment.

Fins ara, aquesta informació no ha estat utilitzada en cap objectiu analític. Amb les eines adequades es pot aprofitar per millorar la situació dins la universitat. En relació al que s'ha esmentat anteriorment, es perd l'oportunitat d'aprofitar al màxim les tecnologies actuals. Aquesta manca d'aprofitament impedeix implementar noves eines que poden proporcionar avantatges significatius. És fonamental explorar com les noves tecnologies poden oferir una idea aproximada dels motius pels quals els estudiants decideixen abandonar les seves carreres.

Necessitats d'informació

Cercant fonts d'informació, es poden trobar exemples de retenció dels millors candidats gràcies a sistemes d'anàlisi de dades. Per tant, seria aconsellable revisar els documents que les empreses ofereixen per comprendre com aquesta eina aporta valor a la retenció de talents. Cal destacar que existeixen articles que aborden la implementació teòrica de l'eina que es proposa [2].

No només es dependrà de fonts d'articles, sinó que també s'haurà de fer recerca de llibreries Python per analitzar les dades i prendre decisions sobre quina llibreria utilitzar per a cada fase del projecte.

No s'ha de deixar de costat la recerca sobre quins són els algoritmes de Machine Learning que s'adapten millor a les necessitats del projecte. Cal analitzar com encaixar això.

Tanmateix, les guies de Power BI per mostrar conjunts de dades segmentades esdevindran fonamentals per a la realització del projecte. S'haurà de realitzar un estudi exhaustiu de com s'haurien de presentar per tenir una visualització de les dades clares i concises per a que el client les pugui interpretar perfectament.
A més, s'utilitzaran llibreries especialitzades en models de Machine Learning per construir el nostre model, que pot ajudar a determinar quins estudiants opten per completar el grau i quins poden decidir abandonar-lo.

Com a font final d'informació, es necessitaran comprendre les pautes d'un estudi de Big Data per realitzar una tasca efectiva. Això inclou adquirir una extensa comprensió sobre el comportament dels estudiants, anàlisi del rendiment dels estudiants per a la millora de programes educatius. Personalització de l'aprenentatge mitjançant l'anàlisi de dades d'estudiants. Optimització de programes acadèmics basada en el rendiment històric.

3.Objectius i abast

1. Millorar la retenció estudiantil en un deu per cent:

   Per avaluar l'èxit d'aquest objectiu, es durà a terme una anàlisi estadística que compararà el nombre d'abandonaments en el curs actual amb els percentatges projectats per al futur. Aquesta mesura proporcionarà una visió clara del progrés i de l'impacte de les iniciatives implementades gràcies a l'eina desenvolupada.

2. Agilitzar el procés d'anàlisi de dades dels estudiants al coordinador de grau corresponent en un trenta per cent:

   Per a avaluar l'èxit d'aquest objectiu, caldrà mesurar el temps que el coordinador dedica a realitzar un estudi i a extreure conclusions de les dades, i comparar aquest temps amb el que requereix l'eina desenvolupada per obtenir informació similar. Això permetrà apreciar de manera clara l'eficàcia de la nova eina en comparació amb els processos existents.

3. Detectar els possibles perfils d'estudiants amb una major tendència a abandonar amb un percentatge d'èxit del setanta per cent.
   Per avaluar l'èxit d'aquest objectiu, es proposa implementar un model de machine learning utilitzant conjunts de dades d'entrenament i prova. Una vegada hagim definit els perfils, el model haurà de ser capaç de proporcionar respostes correctes. Aquest enfocament permetrà una avaluació més precisa i eficaç del rendiment del model.

4. Metodologia

[6] La metodologia àgil és un enfocament de desenvolupament de projectes que es caracteritza per la seva flexibilitat, col·laboració i adaptabilitat a canvis. Aquesta metodologia posa èmfasi en la resposta ràpida als requisits del client, la comunicació contínua i la producció iterativa de lliurables funcionals.

El projecte seguirà la metodologia àgil perquè es necessiten revisions freqüents per garantir que els requisits del client s'estan complint i per obtenir feedback, no obstant es complementarà amb una metodologia de cascada, agafant les millors qualitats de les dues metodologies de treball.

Es farà ús de la Revisió Per Pair per a aquest propòsit. Les reunions es duran a terme cada dues setmanes mitjançant l'eina Microsoft Teams.

També es realitzaran reunions esporàdiques amb el client i altres experts en el tema. Per cada reunió es generarà una acta amb tota la informació treballada en aquesta.

La recopilació principal d'informació serà de font bibliogràfica, amb una estratègia d'anàlisi comparativa. La redacció de resums serà fonamental per sintetitzar la informació recollida

El projecte es divideix en 6 parts:

1. Extracció i modificació de les dades
2. Anàlisi de les dades
3. Explotació de les dades
4. Creació d’un programa implementant machine learning
5. Conclusions basades en els resultats
6. Presentació

Dates d'entrega

Lliurament Memòria 1: Avantprojecte 12 de gener
Lliurament Memòria 2: Memòria intermèdia 15 de març
Lliurament Memòria Final 3 i 4 de juny
Data per al dipòsit digital de la documentació del TFG 3 i 4 de juny
Període per a la defensa del TFG Del 25 al 28 de juny

Fase de desenvolupament

Durant la fase de desenvolupament, s'ha aplicat diligentment aquesta metodologia híbrida, que ha demostrat ser una estratègia efectiva per gestionar el projecte. A mesura que avançava el treball, es van duent a terme els diferents punts definits, fins a arribar a tenir reunions cada dues setmanes aproximadament, en les quals es presenten i discuteixen els avenços realitzats.

El treball realitzat durant aquestes reunions podia prendre diverses formes, ja fos mitjançant la presentació de codi, les modificacions en les interfícies de treball, l'addició d'elements a les taules de dades o la modificació dels propis gràfics que s'utilitzen com a eina visual per a l'anàlisi.

Després de presentar les propostes, es procedeix a una anàlisi conjunt, en la qual s'avaluen tant els aspectes tècnics com els requisits del projecte. En cas de discrepàncies o suggeriments de millora, s'acorden els canvis a implementar per a la següent iteració.

És important destacar que, dins d'aquest marc de treball àgil, l'acabament d'un punt no implica necessàriament que estigui completament finalitzat. Així doncs, és possible tornar enrere al cicle i realitzar modificacions o ajustaments segons les necessitats del projecte.

Durant el transcurs del projecte, aquesta iteració ha estat repetida en diverses ocasions. Això es deu al fet que, en la pràctica, han sorgit diversos problemes en el procés de neteja i tractament dels fitxers CSV. Aquests problemes eren columnes amb dades mal formades, formats de dades incorrectes o fins i tot la necessitat d'incorporar noves taules de dades per abordar nous aspectes de l'anàlisi. Així, l'aplicació d'aquesta metodologia ha permès adaptar-se de manera flexible i eficient als reptes que han anat sorgint al llarg del desenvolupament del projecte.

5. Requisits

Els requisits del projecte son els següents:

1. Preparar el dataset font de l'anàlisi a partir dels arxius generats des de SIGMA
2. Validar i tractar les dades
3. Complir la normativa GDPR
4. Establir les característiques de l'alumnat del TecnoCampus
5. Determinar el mapa de processos que defineix com és el seu pas pels estudis
6. Definir punts crítics causants de l'abandonament
7. Concretar els aspectes clau que determinen l'abandonament
8. Implementar un classificador que detecti els alumnes que potencialment poden arribar a abandonar els estudis
9. Documentar el procés i presentar les conclusions
10. Utilitzar python pel tractament del dataset i implementació de l'analítica
11. Utilitzar Tableau/Power BI per la visualització de dades
12. Utilitzar Microsoft Office per la documentació del procés i presentació de les conclusions
13. Preparar una enquesta per tenir una primera visió sobre els motius de l'abandonament
14. Buscar informació sobre la utilització d'una eina semblant al mercat.

Comparar i analitzar si existeixen diferències entre els abandonaments de carrera masculins vers els femenins.

6. Desenvolupament

Abans de començar a programar i treballar amb les dades proporcionades, és crucial realitzar un estudi preliminar. Mitjançant un marc de dades (data mark), s'ha de determinar quines columnes seran rellevants i quines dades es poden ometre per assegurar resultats més precisos.

Les dades facilitades per la coordinadora de grau consisteixen en taules amb informació general i específica, que inclouen detalls dels estudiants i dades generals sobre cadascun d'ells.

1. Presentació del dataset

Crèdits Matriculats

Conté tots els crèdits matriculats per estudiants del grau d'informàtica i de la carrera de videojocs i informàtica dels cursos que van des del 2014 fins al curs 2023/24.

Info Alumnes

Proporciona informació personal detallada sobre els estudiants, com la procedència geogràfica, estudis previs per a l'accés a la universitat, nivell d'anglès, i estudis dels pares, any de matriculació, grau escollit, ordre de preferència, nota de tall, nota d'accés, convocatòria d'admissió.

Assignatures-Num Alumnes

Conté informació de totes les assignatures del grau per curs acadèmic incloent el nombre de matriculacions per assignatura i convalidacions des del curs 2014 fins al curs 2022-23.

Rendiment Assignatures

Mostra el rendiment dels estudiants matriculats per cada assignatura, incloent el nombre d'estudiants presents, percentatge d'aprovats i taxa de rendiment.

Alumnes Assignatures

Detalla el nombre de matriculacions per cada assignatura i curs, juntament amb el nombre d'estudiants presentats a l'examen i el nombre d'aprovats tant amb percentatges o numèric.

Qualificacions

Conté les qualificacions de cada estudiant per cada una de les seves matrícules durant els cursos 2014 fins al curs 2022/23, així com els crèdits suspesos i superats per assignatura i la qualificació de l'assignatura.

Assignatures

Conté totes les assignatures proposades pel grau d'informàtica juntament amb el tipus de formació que es corresponen, el curs acadèmic que es poden matricular i el nombre de crèdits que representen.

Graduats

Conté tota la informació en relació als alumnes que s'han graduat del grau d'informàtica juntament amb l'any d'inici i l'any de la graduació.

Centre procedència

Conté tota la informació dels estudiants que han intentat matricular-se al grau d'informàtica juntament amb l'acceptació de la pròpia institució i el centre de procedència del mateix.

2. Neteja de taules per la creació dels models

Amb l'objectiu d'assolir els objectius establerts, és important identificar les taules i files que no aporten informació rellevant per a l'estudi.

S'ha de tenir en compte que l'estudi se centrarà en l'anàlisi del conjunt de dades del grup, evitant un enfocament individual per a cada estudiant per assegurar la creació d'un model sòlid i una anàlisi de dades precís.

S'ha decidit realitzar un estudi dels alumnes de grau d'informàtica; per tant, les dades que es corresponen amb les dels alumnes del doble grau seran omeses degut al fet que són perfils d'estudi completament diferents podent arribar a embrutar la mostra que es vol treballar.

1. Crèdits Matriculats

Com s'ha comentat prèviament, les columnes de pla d'estudi on tenim tota la informació de les carreres matriculades, filtrar només pel grau d'informàtica. En un futur, si és possible, es poden recuperar aquelles dades i intentar fer el model juntament amb els estudiants de videojocs.

En aquesta taula de dades apareixen dades que no seran necessàries per l'estudi, com la clau de cada estudiant. No obstant, com no es tenen dades dels anys 2008 fins al 2013, aquestes columnes seran omeses també.

2. Info Alumnes

Pel tractament d'aquesta taula apareixen columnes com les que s'ha mencionat abans que no es necessiten per fer el model, com la clau de l'alumne. Les files on no es corresponen al grau que estem estudiant també no es tindran en compte pel model.

La creació d'una nova columna serà clau, ens determina en base a uns criteris que s'han establert si l'alumne està cursant la carrera, s'ha graduat o ha abandonat.

Finalment, s'ha de mirar si el rang de la nota d'accés i admissió pot aportar valor al model, però la columna que s'ha d'ometre també és la de general degut a que només aporta soroll.

3. Assignatures-Num Alumnes

Es segueix arrossegant el mateix problema de les taules anteriors, el de tenir dades barrejades amb els estudiants de videojocs. Per tant, s'ha decidit eliminar les files on apareixen les dades dels alumnes de videojocs.

4. Rendiment Assignatures

Com s'ha decidit que l'estudi sigui només dels alumnes d'informàtica, en les dades d'assignatures tenim assignatures que es corresponen a les d'altres carreres.Per tant, s'ha decidit no tenir en compte aquelles files on les assignatures no es corresponen a les de la carrera d'informàtica mantenint això sí els crèdits convalidats i les pràctiques externes a empreses.

5. Alumnes Assignatures i Qualificacions

Pel tractament d'aquesta taula, el que s'haurà de fer és filtrar totes les files dels alumnes d'informàtica juntament amb les assignatures, només es vol treballar amb les dades d'aquests estudiants.

6. Qualificacions

Pel tractament d'aquesta taula es seguirà el mateix procediment que la taula alumnes assignatures filtrant només aquells alumnes que coincideixin amb la condició que estiguin cursant la carrera d'informàtica.

7. Assignatures

Aquesta taula no haurà de ser sotmesa a ninguna fase de neteja, serà utilitzada com una taula de consulta per aprofundir la informació que es disposa.

8. Graduats

Aquesta taula no ha estat sotmesa a cap canvi no obstant la informació de la taula s'ha utilitzat per calcular l'estat curricular de la taula info alumnes així com una optimització de la funció i càlcul dels alumnes graduats. S'ha agafat totes les keys de la taula que tenien una referència a la taula info alumnes i s'ha canviat l'estat curricular a Graduat de totes aquelles on coincidia.

9. Centre procedència

La taula centre procedència no ha estat sotmesa a cap canvi, el que s'ha fet amb aquesta taula és modificar novament la taula principal info alumnes degut al fet que la taula en qüestió disposa del voltant de quatre centes files sense referenciar a cap altre taula per tant alumnes que realment no han anat més enllà i mai s'han matriculat a la carrera. Juntament amb la neteja de les files no referenciades a aquelles que si tenen referencia s'ha afegit una nova columna centre procedència per afegir de quin centre ve cada alumne que es matricula a la carrera.

3. Script de neteja de dades per la creació de gràfics

Abans de començar a realitzar gràfics, les dades hauran de ser netejades. Per a la neteja, s'ha decidit utilitzar Python.

La plataforma utilitzada correspon a l'eina Google Colab Notebook, on es pot trobar per a cada conjunt de dades, és a dir, per a cada taula del conjunt de dades, els scripts corresponents al que s'ha treballat per tal de confeccionar les dades de la millor manera per obtenir associacions entre taules i no tenir problemes de malformacions de datasets.

El codi s'ha estructurat d'una manera molt senzilla. Primer s'obté tota la informació de manera local dels arxius de l'Excel que la coordinadora de grau va proporcionar.

A continuació, s'ha dividit el treball en seccions, on per a cada conjunt de dades, s'apliquen les mesures corresponents per disposar d'una matriu de dades ordenada de la millor manera possible.

La última part de l'script descarrega les dades netejades per poder ser manipulades i mostrades en el programari de Microsoft Power BI.

1. Crèdits matriculats

Primer es va començar per la taula de crèdits matriculats. Es va treballar en la duplicació de la clau de cada alumne que tingués files nul·les, degut al fet que quan es van passar les dades d'Excel a CSV i posteriorment a Python, existien files amb valors nul·les degut al propi format de l'Excel.

Quan es va aconseguir solucionar aquest problema, es va procedir al filtratge de files del dataset, seleccionant només aquelles files que es corresponien amb la condició d'estudiants del grau d'informàtica.

2. Info alumnes

Els mètodes aplicats per la neteja d'aquesta taula són semblants als ja esmentats. S'han pres les claus dels alumnes d'informàtica i s'han guardat en una variable auxiliar. A continuació, s'ha filtrat per la columna "plan de estudis" i s'han seleccionat les files corresponents al grau d'informàtica, mentre que les files que no coincideixen amb la condició han estat eliminades.
Per millorar el treball, s'ha decidit etiquetar específicament els estudiants a la taula "info alumnes". S'ha afegit una nova columna anomenada "estat curricular" amb les variables: cursant, graduat i abandonat. Aquestes etiquetes s'utilitzaran com a dades de train i test pel model en un futur.

No obstant això, s'ha detectat un problema amb aquesta categorització i s'ha considerat que establir unes regles per determinar els perfils seria més adequat, ja que els estudiants que estan cursant i els que han abandonat poden generar confusió, i fins i tot els alumnes graduats poden ser mal interpretats.

Primerament, per determinar els alumnes que han completat la carrera, s'han revisat els crèdits aprovats a examen. S'ha observat que no tots els alumnes compleixen la condició dels 240 crèdits a causa d'assignatures convalidades. S'ha establert una condició addicional: si un alumne s'ha matriculat del treball de fi de grau i l'ha aprovat, és considerat candidat a entrar en aquesta categoria.

Per determinar si un estudiant està cursant o ha abandonat la carrera, s'ha revisat la seva última matrícula. Si l'estudiant no compleix amb la condició per ser considerat graduat i no té matrícula per a l'any 2022/23, s'assumeix que ha abandonat la carrera.

Finalment, els alumnes amb matrícula per a l'any 2023/24 s'han categoritzat com a cursants ja que no tenen qualificacions. Aquestes files s'assignaran manualment.

Després de consultar amb la coordinadora de grau i detectar problemes amb les dades de la taula de qualificacions, s'ha decidit mútuament afegir una nova taula anomenada "Assignatures". Aquesta taula contindrà tots els registres de les assignatures impartides des del curs 2014 fins a l'actualitat.

Com a conseqüència d'aquest canvi, la fórmula per calcular els estudiants matriculats i cursants serà diferent. No s'incorporarà el sumatori de crèdits de la columna "Crèdits superats en examen", i les tres columnes referents als crèdits presentats i presentats a examen seran eliminades.

La nova fórmula es basa en la columna "Qualificació" de la taula qualificacions. Per a cada fila, es comprova si el valor corresponent no és "Suspès". S'accedirà a la taula Assignatures per a obtenir la quantitat de crèdits corresponents a aquella assignatura, i es realitza el càlcul. Si el sumatori de crèdits arriba a dos-cents quaranta, s'afegirà l'etiqueta de graduat a la taula "info alumnes".

S'ha afegit una nova condició a causa de la falta d'informació sobre les assignatures convalidades a la taula de qualificacions. Si l'estudiant no té matrícula durant el curs 2022/23 i la suma de crèdits no arriba a dos-cents quaranta, però té el treball de fi de grau aprovat i totes les assignatures de formació obligatòria i bàsica aprovades, es considerarà graduat en cas contrari no graduat.

Com a conseqüència de l'addició de dues noves taules al projecte que proporcionen la informació dels alumnes graduats les fórmules mencionades i processos anteriors no serveixen per determinar si un alumne s'ha graduat el que s'ha fet és treballar amb la taula Graduats i modificar la informació de tots aquelles Key que estan referenciades a la taula.

Per últim s'ha generat una nova columna anomenada Anys al centre que s'encarrega de calcular quants anys ha estat al centre cada un dels alumnes matriculats.

3.Assignatures-Num Alumnes

Durant la fase de programació, el procés realitzat per aquesta taula és similar al de crèdits matriculats. En primer lloc, s'han duplicat totes les files amb valors nuls, els quals provenen de la conversió de valors CSV de l'Excel. En aquest cas, els valors duplicats han estat el curs acadèmic i el tipus de carrera.

Un cop s'ha realitzat la neteja, s'han filtrat les files per aquells alumnes que s'han matriculat únicament a la carrera d'Enginyeria Informàtica.

4. Rendiment assignatures

El treball amb aquesta taula ha seguit un enfocament bastant diferent al mencionat. En primer lloc, s'ha realitzat una inversió de la taula, és a dir, les dades s'han transformat de manera que les capçaleres, en la conversió d'Excel a CSV, no han quedat del tot correctament. Per tant, s'ha optat per transposar les dades de les columnes a files. A conseqüència d'això, s'han perdut algunes columnes i s'han duplicat les files. Aquest procés ha eliminat les columnes repetides d'"alumnes matriculats" per a cada curs acadèmic, i s'ha decidit optar per una única columna que combina "alumnes matriculats" i "curs acadèmic".

Per acabar, s'han filtrat les files per a aquelles assignatures del grau d'Informàtica, així com per a aquelles que s'ha demanat mantenir explícitament, com ara les convalidacions de crèdits o les matriculacions a pràctiques.

5. Alumnes assignatures

Durant la fase de creació del script, la manipulació d'aquest dataset és similar a la taula de crèdits matriculats. Primerament, duplicarem les files del dataset que continguin valors nuls, un problema comú derivat de la conversió d'Excel a CSV.

Una vegada tenim la taula neta, procedim amb la neteja de les files. Utilitzant la llista resultant de l'eliminació d'alumnes de la taula d'informació d'alumnes, examinem les claus per determinar si coincideixen. En cas afirmatiu, conservem únicament els alumnes inscrits en la carrera d'Informàtica.

6. Qualificacions

El conjunt de dades de qualificacions ha passat pel mateix procés de transformació que la majoria dels altres conjunts de dades. En primer lloc, s'han duplicat les files i s'han omplert tots els valors nuls amb la clau corresponent de l'estudiant juntament amb el curs acadèmic.

Un cop s'ha obtingut el conjunt de dades sense cap valor nul, s'ha filtrat per les files que coincideixen amb les claus dels alumnes d'informàtica.

Per acabar, s'han eliminat les columnes "crèdits matriculats" i "crèdits superats en examen" degut als errors detectats en algunes files del conjunt de dades.

7. Assignatures

El dataset assignatures no ha estat modificat ha sigut una taula de referència per obtenir per cada assignatura a quin curs es pot oferir, trimestre, i quants crèdits té.

8. Graduats

El conjunt de dades de graduats ha estat sotmès a un procés de transformació bastant senzill, en primer lloc s'ha estructurat la informació d'una manera per a que sigui llegible per qualsevol persona que revisi el csv, en primer lloc no es disposava de la columna Curs acadèmic i cada columna era un curs per tant s'ha ficat la informació del curs acadèmic en el qual es va graduar l'alumne en una única columna.

Amb aquesta informació modificada el que s'ha decidit fer ha estat calcular quants anys ha estat l'alumne cursant al centre per tant s'ha anat a la taula info alumnes per mirar la primera matrícula i calcular els anys cursats al centre.

Finalment amb tota la informació modificada s'ha comprovat les files que estaven referenciades i s'ha modificat la informació de la taula info alumnes.

9. Centre procedència

La taula centre procedència no ha estat sotmesa a cap canvi la única operació que s'ha fet ha estat afegir el centre de procedència a la taula info alumnes de totes aquelles files que estaven referenciades a la taula destí.

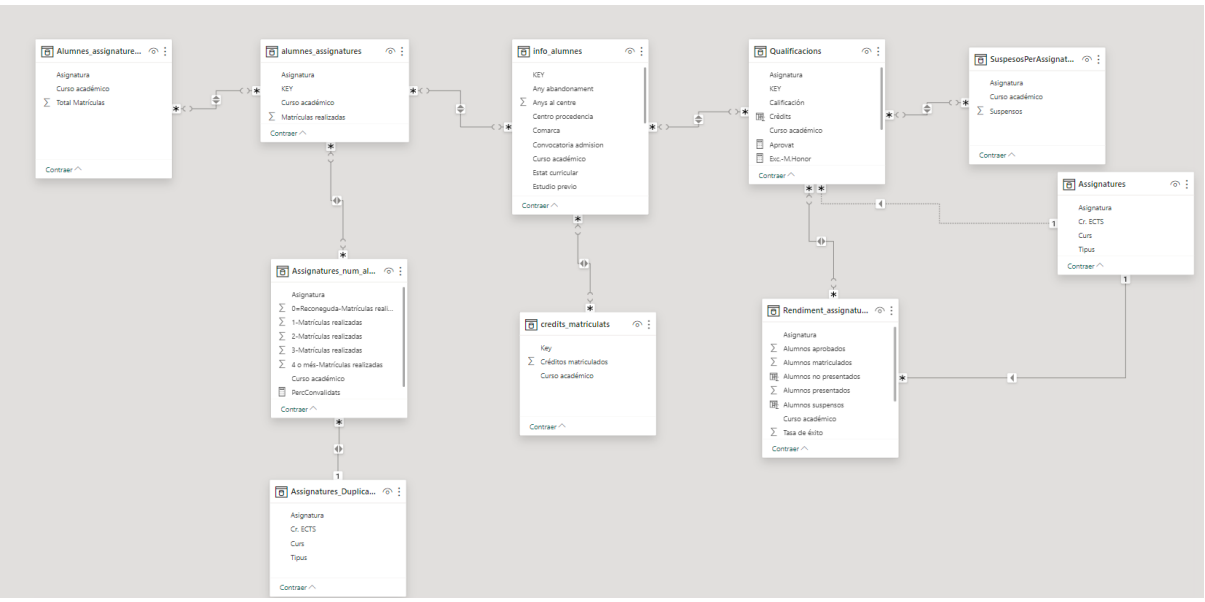
2. Creació d'informes amb Power Bi

Un cop les dades han estat filtrades com se'ns ha demanat s'hauria de mostrar el conjunt d'aquestes amb l'objectiu de poder interpretar relacions entre les dades, tendències i a partir d'aquest punt començar a interpretar el model i veure les relacions entre les variables així com la correcció i normalització dels valors.

El mostreig de les dades es realitzarà mitjançant power BI i les dades que s'utilitzaran seran les resultants del procés de neteja de l'script realitzat.

1. Model relacional

Abans de començar amb la creació de models s'ha de crear un bon model relacional per tenir la informació necessària de cada classe per a poder ser transmesa a tots els dissenys que es vulguin generar.



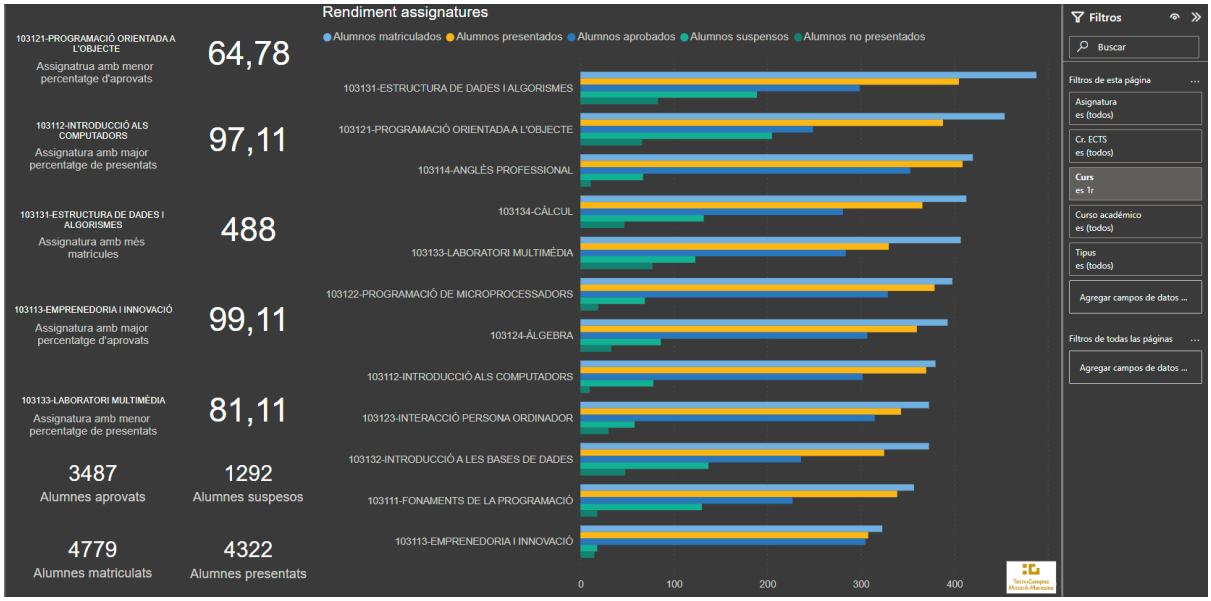
Imatge 1 Model relacional

Com es pot veure al model tenim les taules relacionades per key d'estudiant o per codi d'assignatura i curs acadèmic. Cal comentar que la taula assignatures, taula de suport ha estat duplicada per problemes de relacions entre taules i l'accés a la informació no es feia correctament.

2. Rendiment assignatures

El disseny final d'aquest informe mostra, per a cada assignatura, els alumnes matriculats, els alumnes presentats a examen, els alumnes aprovats, suspesos i no presentats a l'examen, juntament amb els percentatges.

A la part esquerra es mostren les assignatures amb els percentatges més baixos, com el menor percentatge d'aprovats i d'alumnes no presentats, així com el major percentatge d'aprovats. Finalment, es disposa d'un resum general dels alumnes suspesos, aprovats, presentats i matriculats.

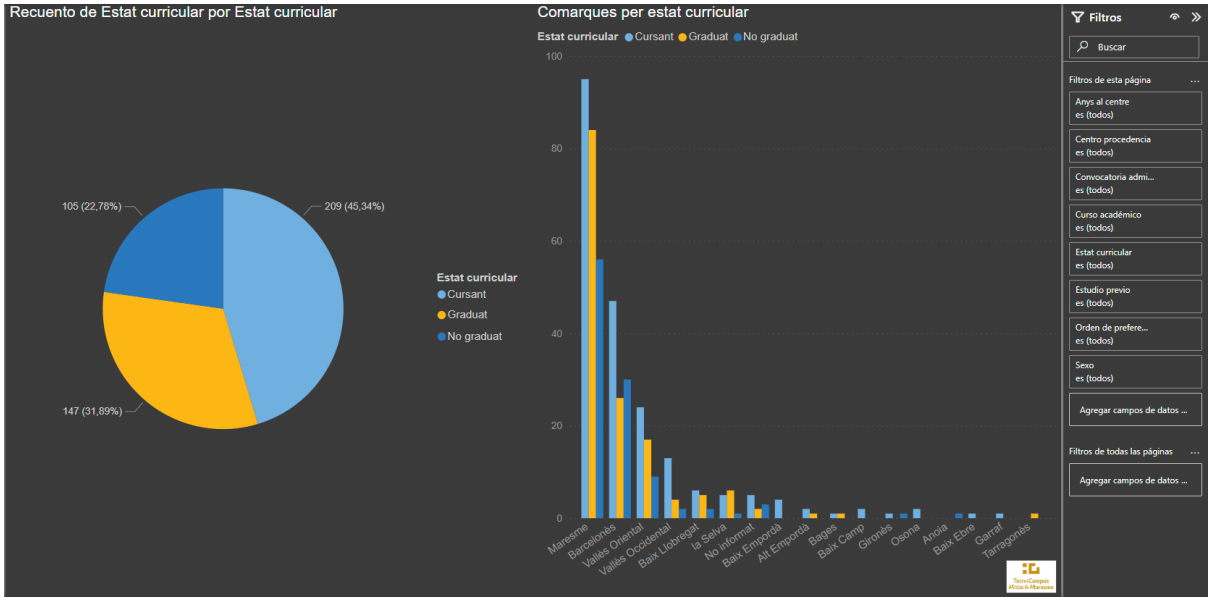


Imatge2 Rendiment assignatures

Pel que fa als filtres aplicats a l'informe, es disposa de filtres com el curs, és a dir, es pot mirar les assignatures de primer, segon, juntament amb el curs acadèmic per mirar anys en concret. Per últim, es pot filtrar per assignatures en específic.

3. Info alumnes

Per la creació i explotació de gràfics d'aquesta taula s'ha decidit ajuntar dos gràfics en un, per una banda tenim un gràfic de formatges amb la quantitat d'alumnes graduats, no graduats i cursant i per altre banda tenim un gràfic de barres amb la procedència de cada alumne.



Imatge 3 Info alumnes

No s'ha vist necessària l'aplicació de cards per marcar els valors més importants del model degut al fet que es poden interpretar de manera senzilla.

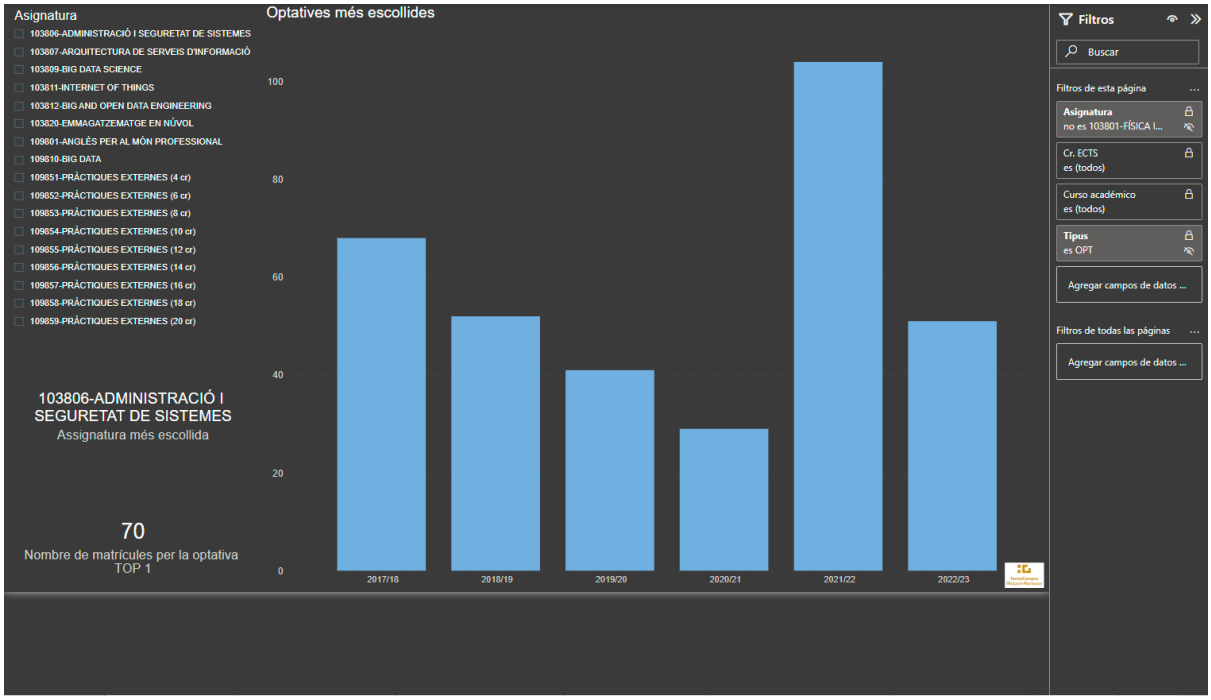
Els filtres aplicats al model en aquest cas són bastant interessants, es pot filtrar per sexe, procedència de l'estudiant és a dir de quina comarca ve, també es pot filtrar per estat curricular, nota de convocatòria entre d'altres.

Amb la utilització de filtres es pot veure com els dos gràfics es transformen de manera dinàmica per tant es disposa d'informació actualitzada amb una simple modificació del filtre de pàgina.

En primer lloc es va decidir canviar el gràfic de barres de territoris per un mapa de calor de Catalunya, degut a problemes de llicència no es pot explotar el següent model. De cara a un futur es tindrà en compte la realització d'aquest gràfic si se'ns proporcionen les eines adequades per la seva creació.

4. Optatives més escollides

Per a la creació del model de la següent taula, es van seguir les peticions de la coordinadora de grau, que no deixa de ser la clienta. Per a les optatives de quart, se'ns va demanar que miréssim quines eren les optatives amb més estudiants per poder veure què deixen de fer els estudiants per escollir pràctiques o si no escollissin pràctiques i feien optatives, quines corresponen.



Imatge 4 Optatives més escollides

Com es pot veure al gràfic, a la part esquerra tenim un filtre per cada assignatura per poder veure precisament la informació d'aquestes juntament amb dues etiquetes on es mostra l'optativa més matriculada i el nombre de matrícules per aquesta assignatura.

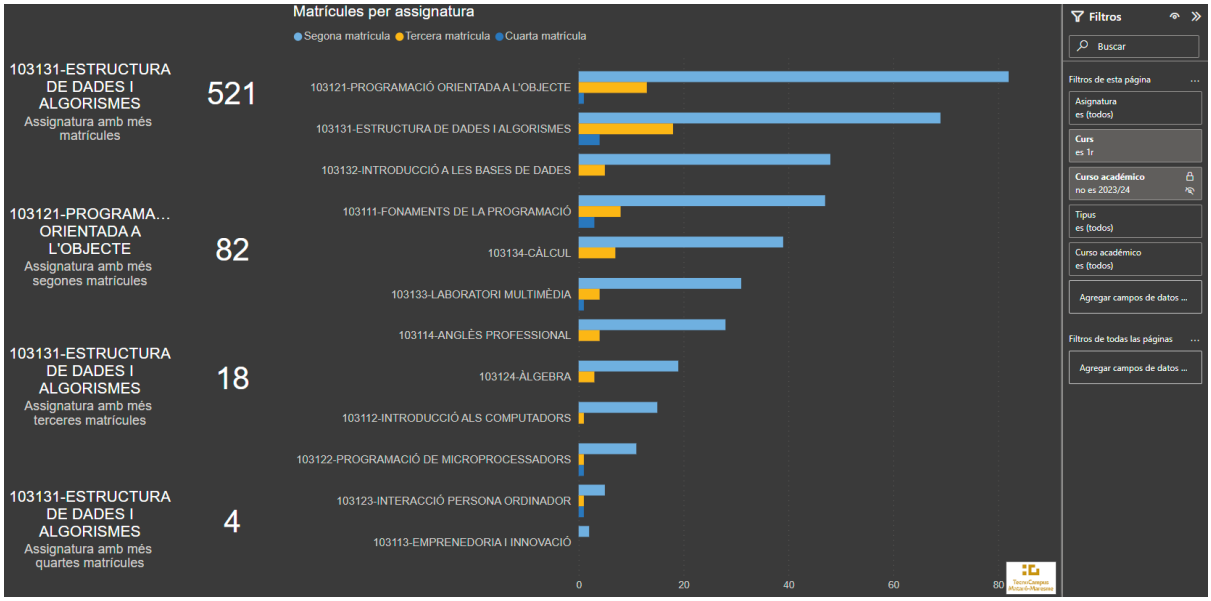
El gràfic consisteix en un gràfic de barres on per a cada optativa es mostra una barra respectiva a la quantitat de matriculacions per alumne, juntament amb l'evolució durant els anys del nombre de matrícules.

Finalment, a la part dreta, tenim els filtres únicament per a curs acadèmic.

5. Matrícules a assignatures

L'informe corresponent consisteix en una visió general del total de matrícules de cada una de les assignatures proposades al grau. Per obtenir el màxim d'informació, s'ha decidit eliminar les convalidacions i primeres matrícules ja que no aporten valor per aquest model. El que es busca és mirar per cada curs quines són les assignatures amb major nombre de segones matrícules, terceres i quartes, per intentar enfocar un canvi i disminuir aquesta tendència.

El gràfic mostra les assignatures juntament amb etiquetes per les assignatures que tenen major representació a cada categoria.



Imatge 5 Matrícules a assignatures

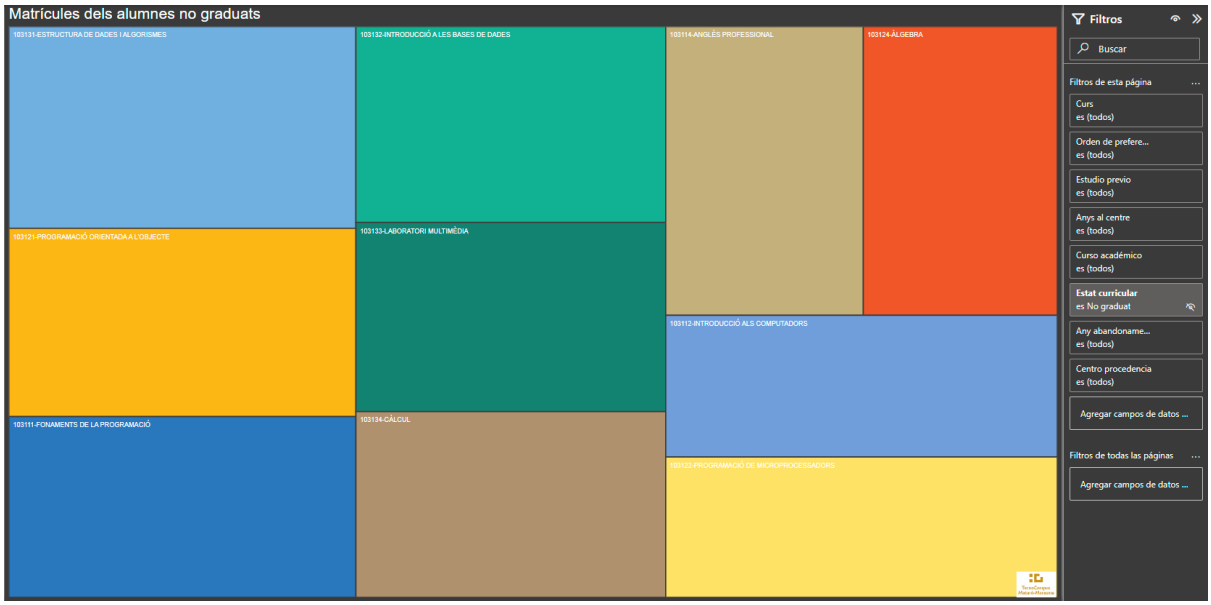
La proposta realitzada representa un gràfic de barres on per cada segona tercera i quarta matrícula tenim registre juntament amb l'addició de cards a la part esquerra per tenir una visió general i completa de les dades amb els rànkings sobre quines son les assignatures que tenen més matrícules.

Pel que fa als filtres, es pot filtrar per curs acadèmic, assignatura, tipus d'assignatura i finalment curs.

6. Matrícules dels alumnes no graduats

L'informe elaborat representa les matrícules realitzades pels alumnes no graduats, aplicant un filtre per curs acadèmic i mostrant un TOP 10 de les assignatures amb major nombre de matrícules.

El gràfic que es mostra consisteix en un treemap de colors on per cada assignatura es representa amb colors la quantitat d'alumnes i el grossor de cada figura representa la quantitat d'alumnes matriculats a cada una de les assignatures

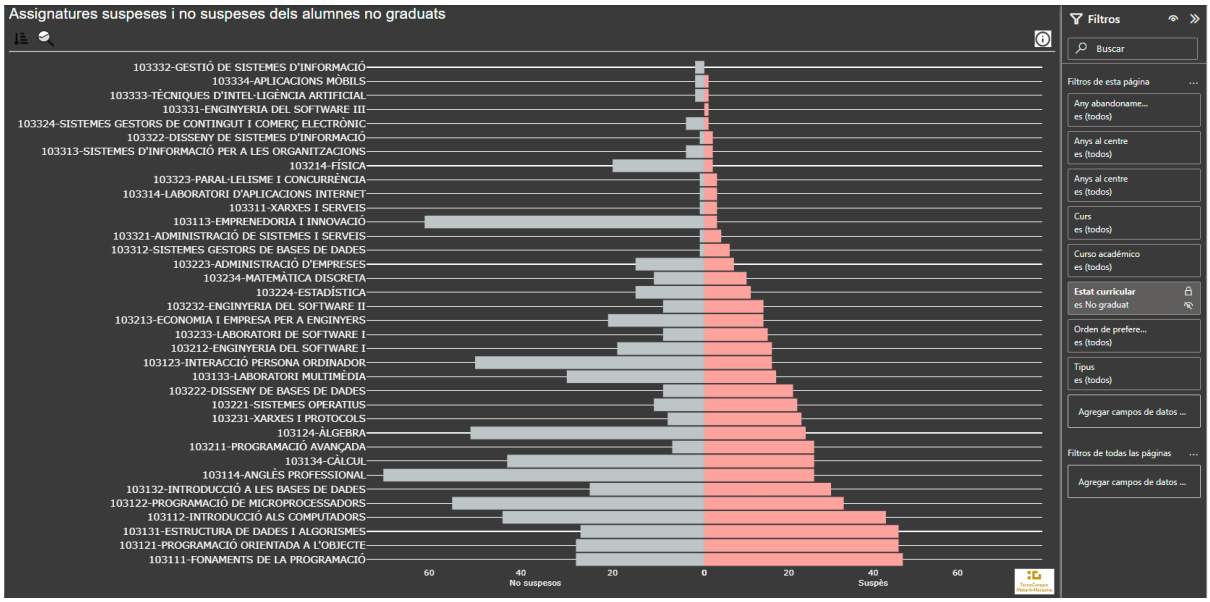


Imatge 6 Matrícules dels alumnes no graduats

Els filtres que es poden aplicar inclouen el curs, el curs acadèmic, l'any d'abandonament i el centre de procedència.

7. No suspesos i suspesos per assignatures i estat curricular

L'informe correspon a un gràfic d'arbre on la part esquerra representa el nombre d'estudiants aprovats mentre que la part dreta representa el nombre de suspesos per assignatura.



Imatge 7 No suspesos i suspesos per assignatures

El gràfic mostra la informació de totes les assignatures on per cada una d'aquestes es fa un recompte de les matrícules aprovades i suspeses i es mostra als eixos, el color gris representa la població on les assignatures han estat aprovades mentre que el color vermell consisteix en els alumnes que han suspès.

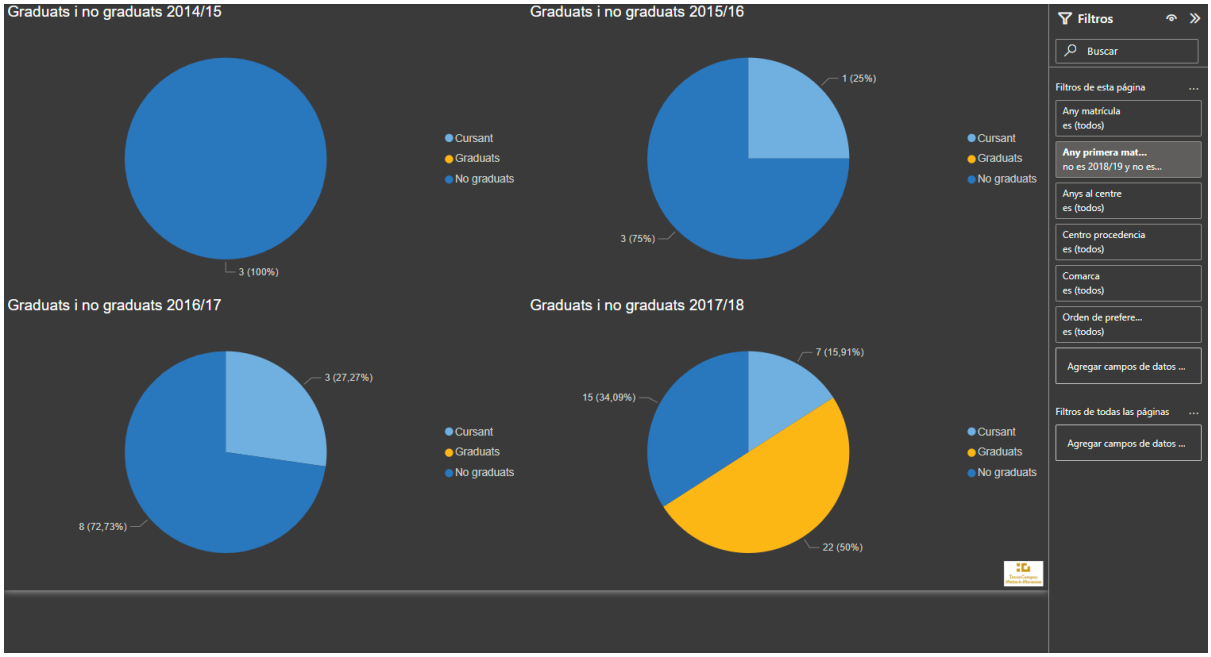
Per la visualització d'aquest informe s'ha aplicat un filtre d'alumnes no graduats no obstant es pot canviar per veure les tendències entre els diferents perfils d'estudiants.

Per filtrar el gràfic es pot fer per curs, curs acadèmic, estat curricular, anys al centre, any d'abandonament, tipus d'assignatura i ordre de preferència.

8. Graduats i no graduats període 2014-2023

El gràfic corresponent disposa de quatre gràfics de formatge on per cada gràfic es mostra en primer lloc els alumnes que van abandonar a cada un dels anys juntament amb els alumnes que es van graduar en un any en concret mentre que els alumnes cursant es mostren aquells en el curs que van començar i no acabar.

El gràfic adjunt representa els cursos 2014-2018

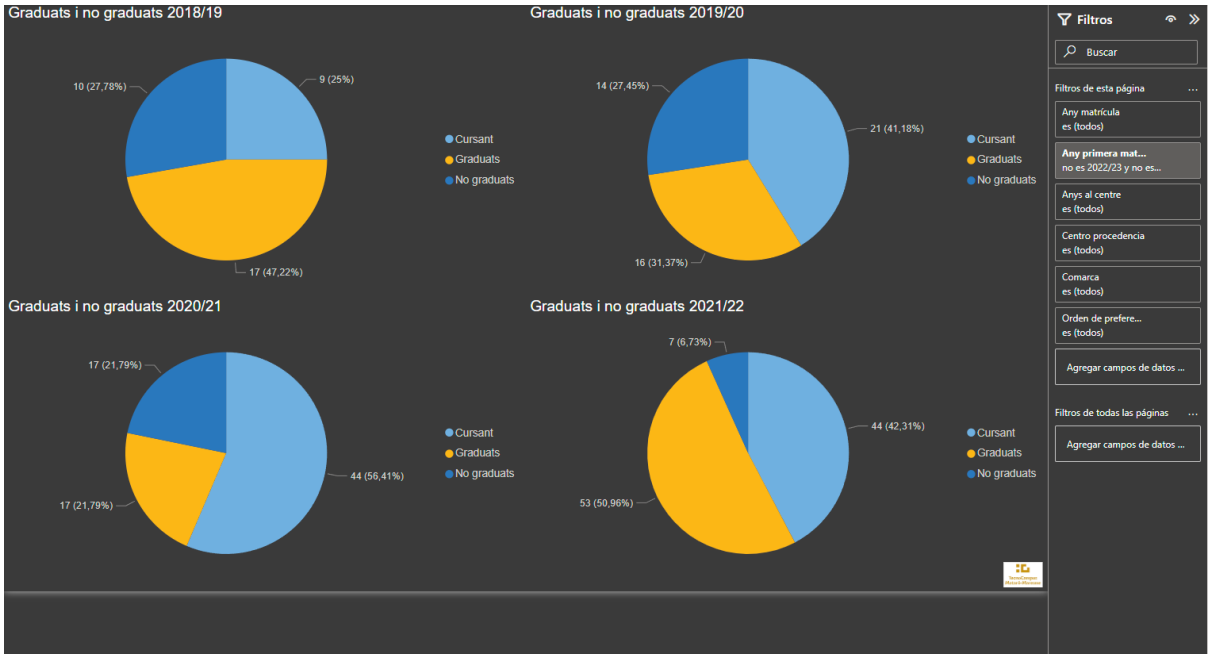


Imatge 8 Graduats i no graduats i cursant 2014-18

Cada formatge representa un curs acadèmic, i per cada gràfic es fica la informació dels alumnes que s'han graduat aquell any, han abandonat i si encara estan cursant s'indica l'any el qual van començar com es pot veure al gràfic del curs 2015/16 existeix un alumne que encara està cursant mentre que el curs 2017/18 ja es comença a tenir alumnes graduats degut al fet que es comencen a tenir registres a partir del curs 2014/15.

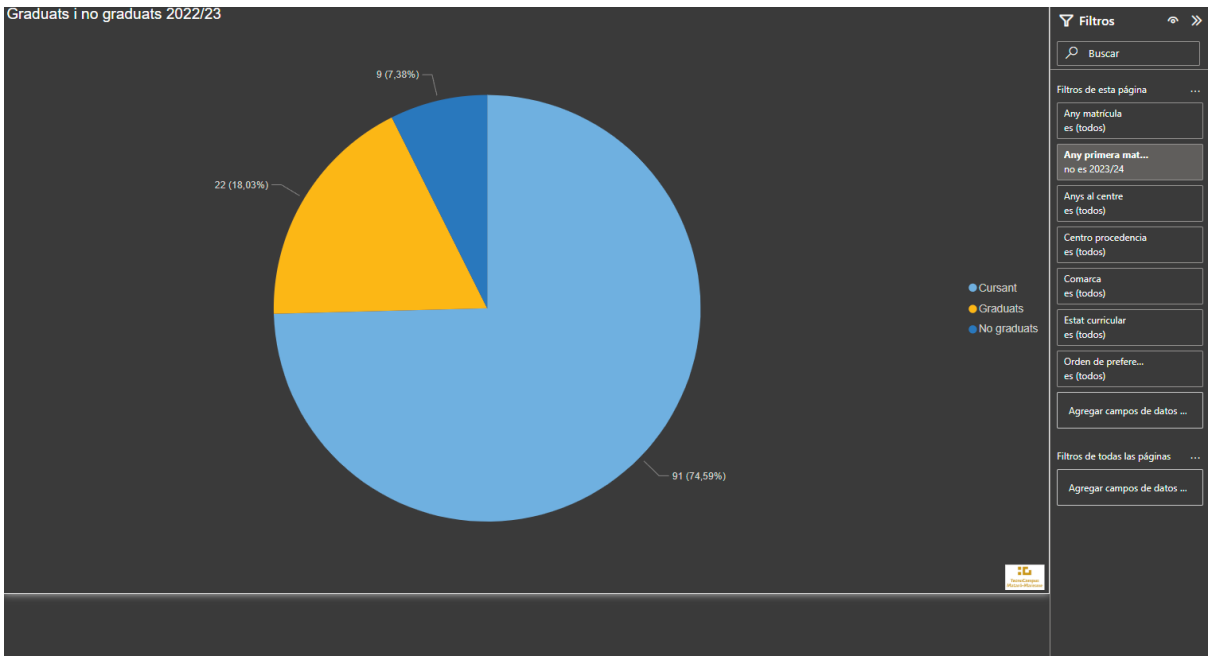
Els filtres que es poden aplicar al gràfic són els següents: Ordre de preferència, any matrícula i anys al centre.

El gràfic corresponent representa els cursos 2018-2022



Imatge 9 Graduats i no graduats i cursant 2018-22

El gràfic corresponent representa l'últim curs el qual es tenen dades que és el 2022-23

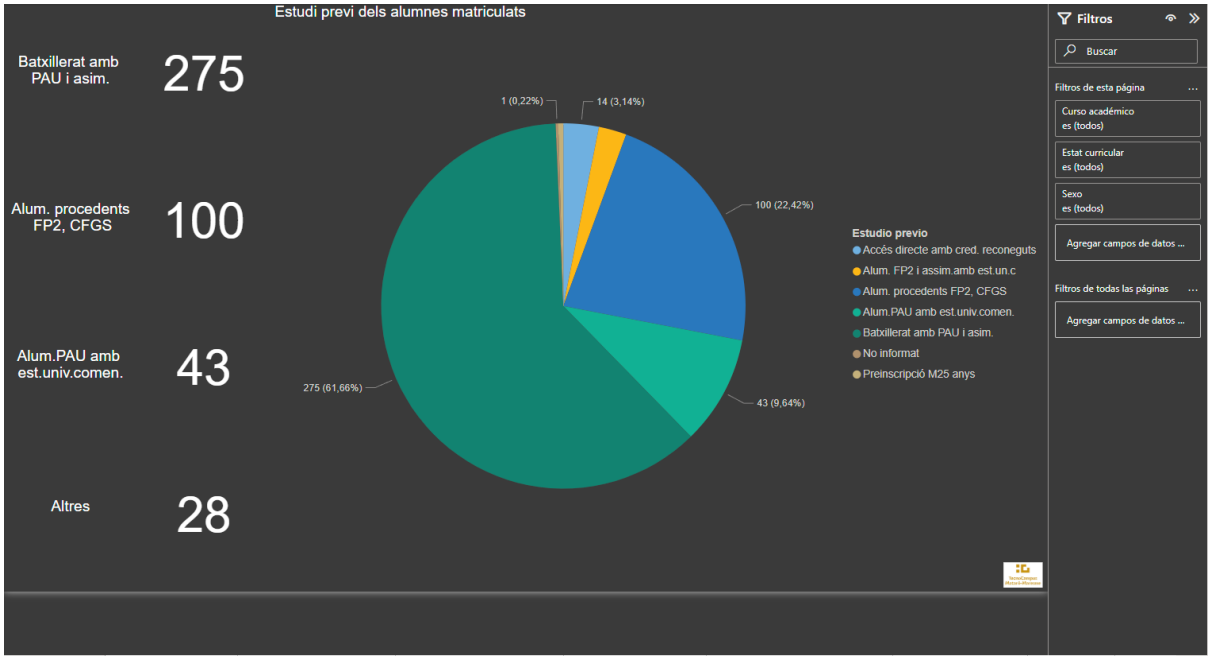


Imatge 10 Graduats i no graduats i cursant 2022-23

9. Estudis previs dels alumnes matriculats

L'informe corresponent mostra la procedència dels estudiants matriculats al grau, és a dir un gràfic de formatge on hi ha un recompte de la quantitat dels alumnes procedents de batxillerat, cicles formatius entre d'altres.

A la part esquerra del gràfic es mostren unes cards amb el recompte d'alumnes procedents, com poden ser accés mitjançant proves PAU, cicles superiors, mitjançant convalidacions i altres que entraria no informat, preinscripció M 25 anys i crèdits reconeguts.

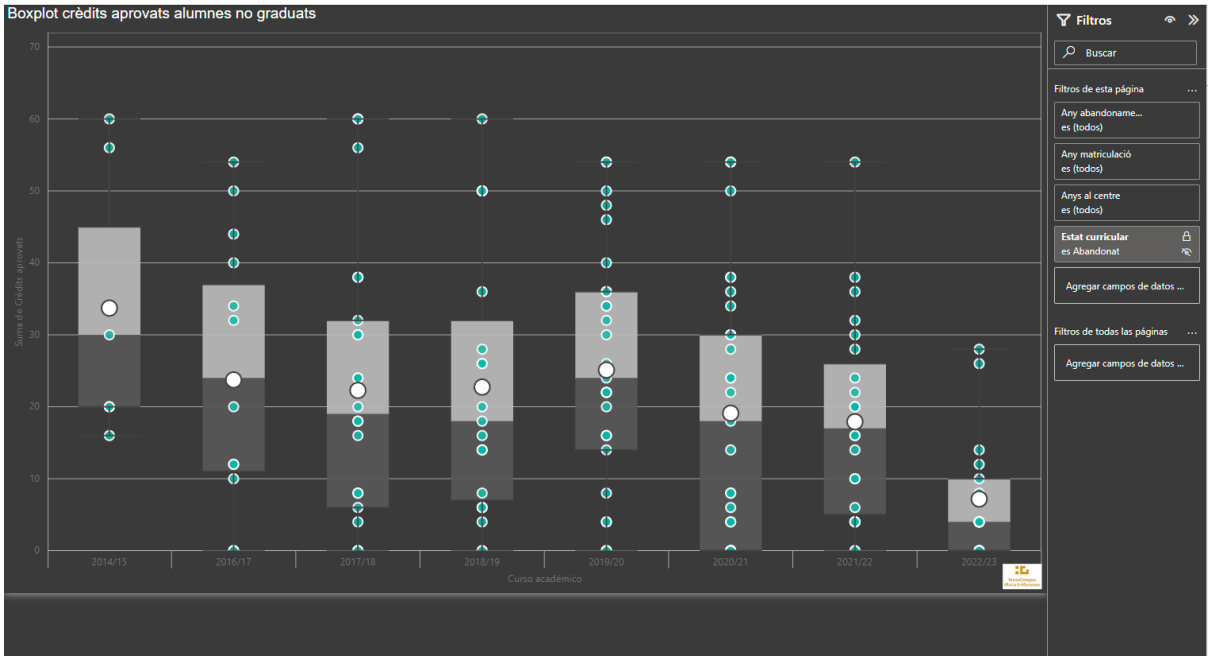


Imatge 11 Estudis previs dels alumnes matriculats

Per la part dels filtres es pot filtrar per curs acadèmic per veure les tendències, també es pot filtrar per sexe i estat curricular.

10. Boxplot de crèdits aprovats alumnes no graduats

L'informe corresponent consisteix en una proposta per part de la clienta. Va sol·licitar generar un boxplot amb els crèdits aprovats dels alumnes no graduats; no obstant això, es poden modificar els filtres per alterar la visualització de les dades.



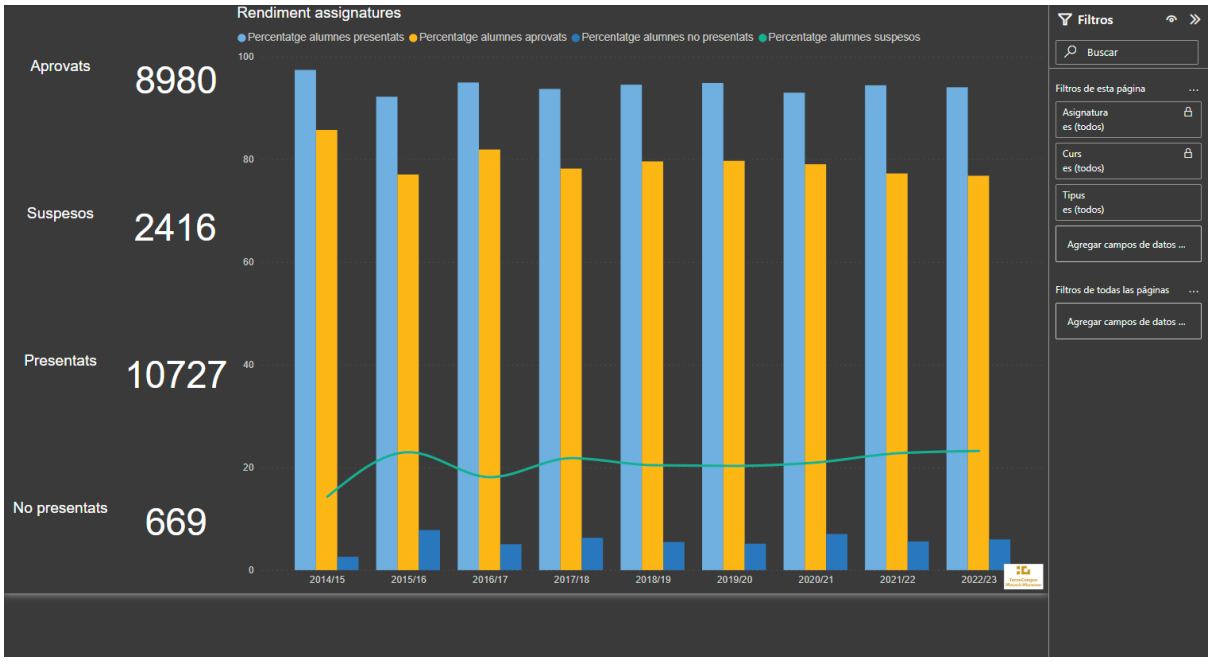
Imatge 12 Boxplot crèdits aprovats d'alumnes no graduats

L'informe consisteix en un boxplot on per cada curs acadèmic que es tenen dades es mostren la mitja de crèdits aprovats per part de tots els alumnes que han abandonat la carrera.

Els filtres treballats per a aquest informe són els següents: anys al centre, estat curricular que per defecte està en alumnes que han abandonat la carrera, any matriculació i any abandonament.

11. Rendiment assignatures

L'informe adjunt mostra l'evolució del rendiment de les assignatures amb un gràfic de barres durants els cursos 2014-2023. Com es pot veure al gràfic mostra el total d'alumnes no presentats, aprovats presentats i matriculats en barres i per ultim una línia per tot l'eix de les x que mostra l'evolució dels suspesos.



Imatge 13 Rendiment assignatures

El gràfic mostra l'evolució de les assignatures durant els períodes que es tenen dades on cada color representa una variable les barres de color blau clar representen el percentatge d'alumnes que es presenten a examen és a dir que no donen per perduda l'assignatura, mentre que el color groc representa el percentatge d'alumnes aprovats.

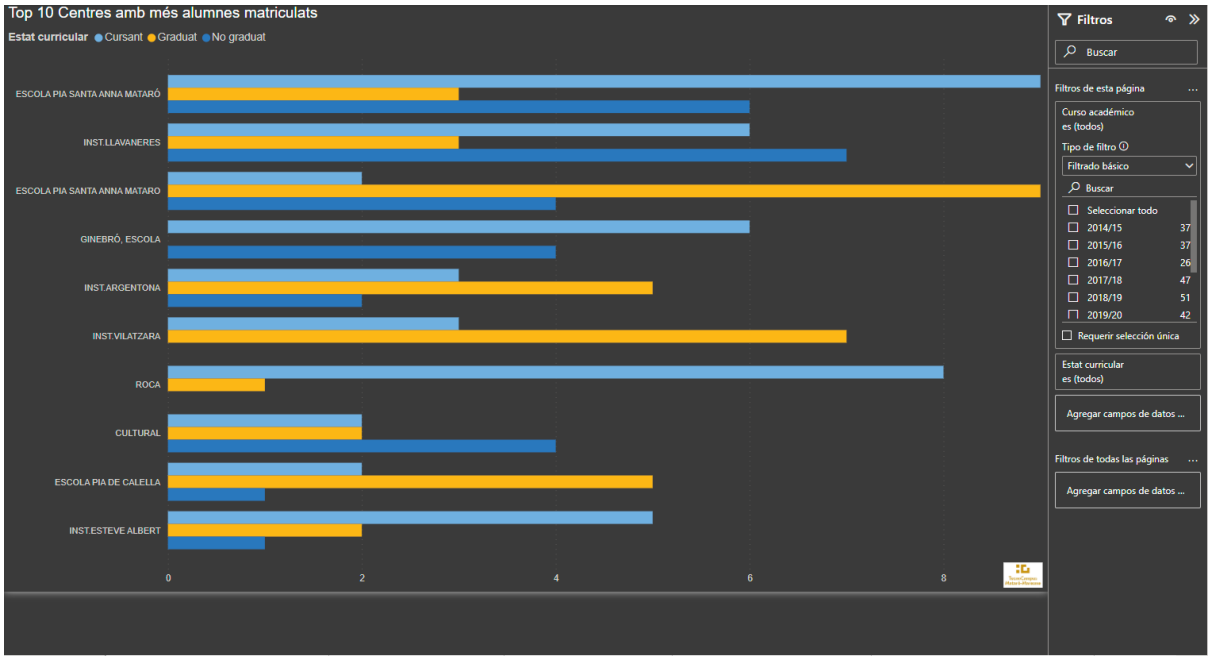
En relació al color blau fosc representa el percentatge d'alumnes que no es presenten a l'examen.

Per acabar per cada curs es disposa d'una línia sobre l'eix de cada curs la qual representa el percentatge de suspesos per curs.

Per la part dels filtres es pot filtrar per assignatures, curs i tipus d'assignatura.

12. Top 10 centres de procedència amb major nombre d'estudiants matriculats

Amb l'addició al model de noves taules i dades se'ns va demanar realitzar un nou gràfic el qual mostra la procedència dels alumnes que han passat pel grau d'informàtica per tant es va decidir realitzar un gràfic de barres mostrant quins son els centres de procedència amb major nombre d'estudiants matriculats a la carrera d'Enginyeria informàtica.



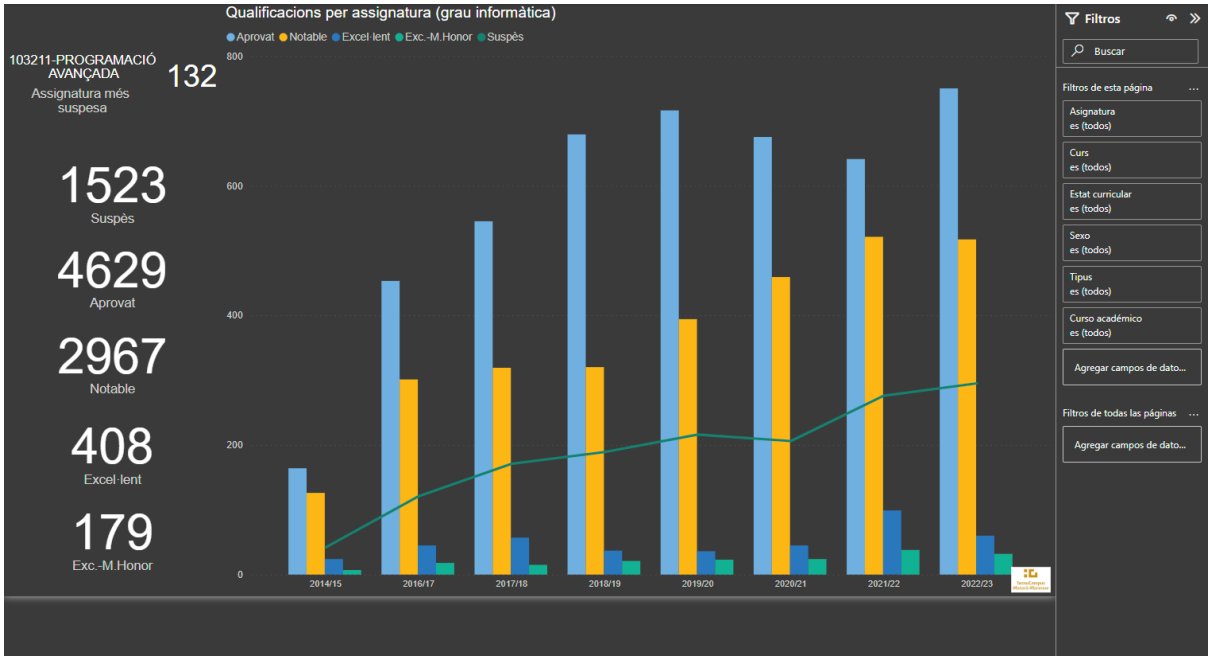
Imatge 14 TOP 10 centres de procedència amb major nombre d'estudiants

El gràfic mostra per cada centre acadèmic la quantitat d'alumnes que hi ha cursant, que han abandonat i s'han graduat de la carrera corresponent. Per la part dels colors de les barres es respecte la llegenda tenint els alumnes cursant amb un color blau clar mentres que els graduats representen el groc i finalment es que no s'han graduat es mostren amb un color blau fosc.

Per la part dels filtres es pot filtrar per estat curricular és a dir es pot veure quins són els centres amb major nombre de graduats, alumnes cursant i alumnes no graduats.
Per últim es pot aplicar un filtre per curs acadèmic on es pot mirar per cada un dels anys o un conjunt d'anys quins han estat els centres de procedència dels alumnes.

13. Qualificacions per assignatura

El darrer informe treballat es molt semblant al gràfic generat per rendiment assignatures no obstant aquest informe mostra les dades amb els valors originals sense tenir en compte percentatges i al treballar amb taules diferents els filtres que es poden aplicar canvien.



Imatge 15 Qualificacions per assignatura

Com es pot veure al gràfic es tenen barres amb el recompte total de cada una de les possibles qualificacions, les quals són aprovat representat amb blau, notable amb groc, excel·lent amb blau fosc i finalment matrícula amb un verd més clar. Pel que fa a l'eix de les x s'ha generat una línia que representa la quantitat de suspesos que han hagut durant aquells cursos.

Al treballar amb taules diferents els filtres que es poden aplicar al gràfic canvien, filtres com estat curricular que anteriorment a l'altre gràfic no es podia aplicar. En aquest informe es pot aplicar per tenir una visió general sobre les tendències dels diferents perfils d'estudiants juntament amb filtres d'assignatures, tipus d'assignatura, curs acadèmic entre d'altres.

3. Generació de models mitjançant algorismes de machine learning

Finalitzada la implementació del dashboard i un cop assegurat que aquest cobreix totes les necessitats d'informació de la clienta, comença la fase final del projecte: l'entrenament d'n model de machine learning que ajudi a identificar quins estudiants es graduaran i quins abandonaran la carrera.

Es treballa amb el mateix dataset en el que s'han realitzat les tasques d'analítica descriptiva, tot i que totes les taules que no tinguin informació de l'alumne seran omeses pel simple fet que no aportaran significat a l'estudi que es vol realitzar. Si la clienta vol que es realitzi algun estudi a part amb les taules que no es faran servir, es pot mirar d'estudiar la possibilitat sempre i quan els terminis s'adeqüin al que s'està demanant.

S'han utilitzat els següents algoritme
- SVM
- XGBoosting
- K-Means
- Random forest

1. Modificació del dataset

Pel que fa a les necessitats per dur a terme l'estudi, es requereixen les següents dades sobre els alumnes:

- Comarca
- Sexe
- Estudis previs
- Nivell B2 d'anglès
- Estudis de la mare
- Estudis del pare
- Universitat
- Convocatòria d'admissió
- Ordre de preferència
- Rang de la nota d'accés
- Rang de notes d'admissió
- Estat curricular
- Centre
- Anys al centre
- Any abandonament

Pel que fa al que no es necessita, s'han de descartar les següents categories: "Key" no aporta res juntament amb el pla d'estudi.

Les taules assignatures-num alumnes , rendiment assignatures i assignatures tampoc son necessàries.

Es requereix la taula alumnes-assignatures, on per cada alumne cal calcular els crèdits totals matriculats juntament amb el nombre de primeres, segones, terceres i quartes matrícules.

Per últim, la taula qualificacions també és necessària. Cal crear una columna extra on per cada estudiant s'ha de fer el recompte d'assignatures aprovades, suspeses, notables, excel·lents i matrícules.

2. Transformació de dades categòriques a numèriques

Un cop les dades han estat netejades per a poder ser explotades és necessari transformar aquelles columnes amb valors categòrics a numèrics i per això que s'ha utilitzat la llibreria LabelEncoder [9] i s'ha creat un diccionari per poder recuperar les dades amb els valors originals.

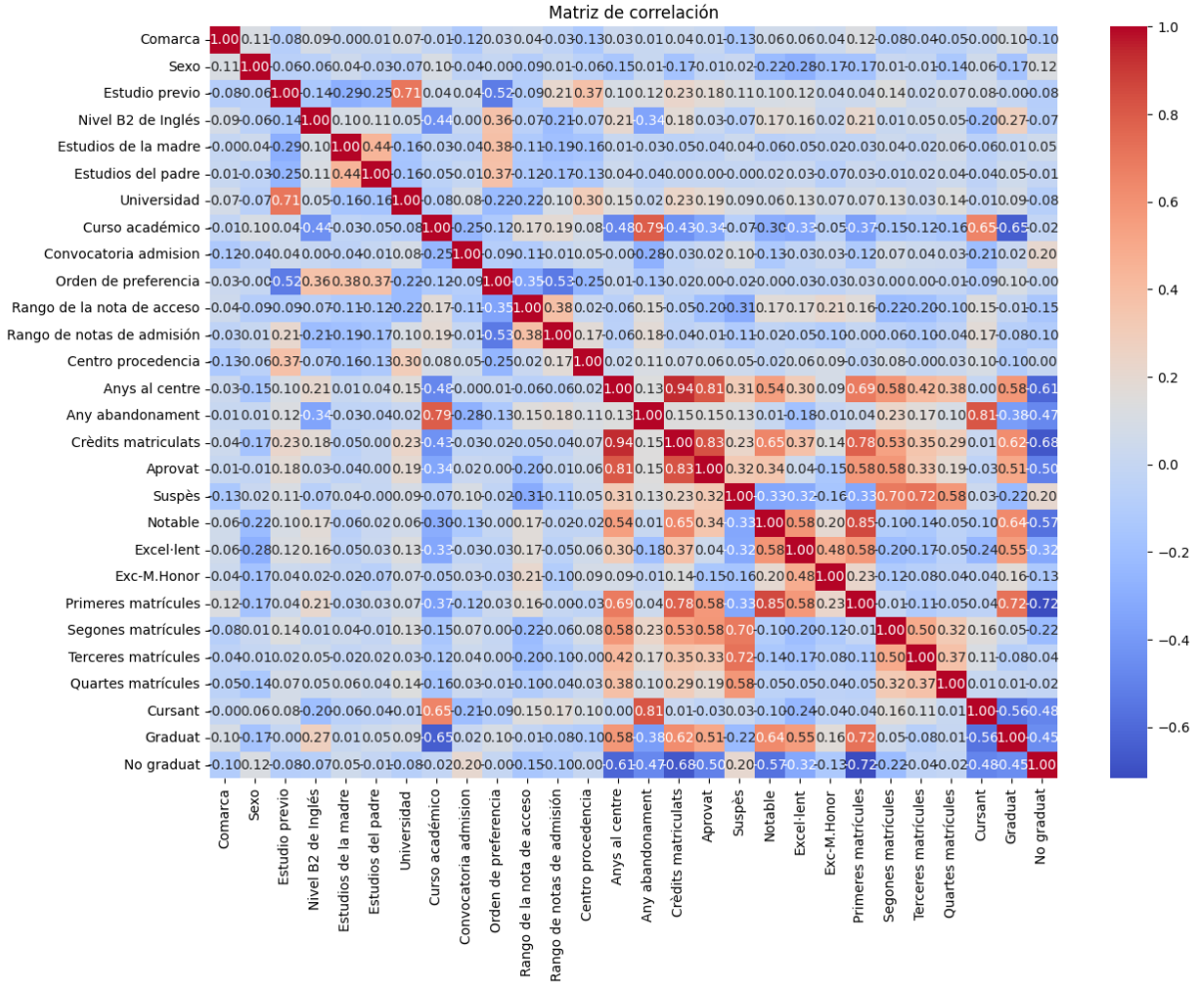
3. Separació dades train, test i de classificació

Per poder tenir una millor visió dels models s'ha decidit dividir les dades en quatre, en primer lloc s'ha separat totes les dades de train i test, Filtrarem les dades per considerar només els alumnes que s'han graduat i els que han abandonat. Per evitar overfitting aplicarem remostreig dividint el set de dades en dades per entrenat el model o dades de train 85% i dades per avaluar el model o dades de test 15%.

D'altra banda, es disposarà de dos dataframes, on s'utilitzarà el primer conjunt de dades en la seva totalitat juntament amb la informació dels alumnes que estan cursant actualment per predir si es graduaran o abandonaran. S'uitilitzarà a més el primer conjunt de dades en la seva totalitat per aplicar tècniques de clustering i fer per separat el perfil dels alumnes que s'han graduat i els que han abandonat.

4. Matrius de correlació

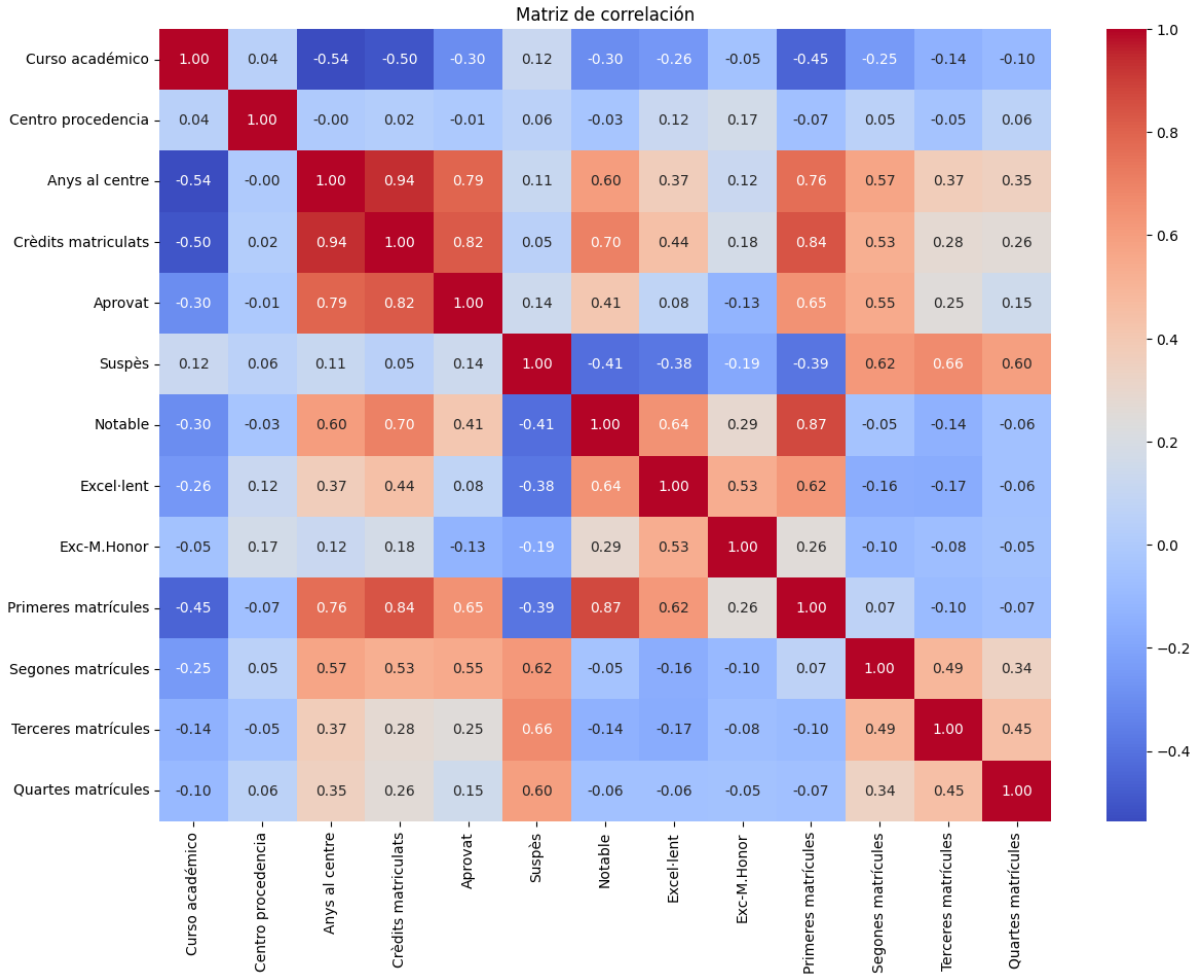
Amb l'objectiu de generar un model el màxim precís possible s'ha de generar una matriu de correlació per així poder veure quines són les columnes amb menys relació entre elles i eliminar-les del model.



Imatge 16 Matriu de correlació dades info alumnes

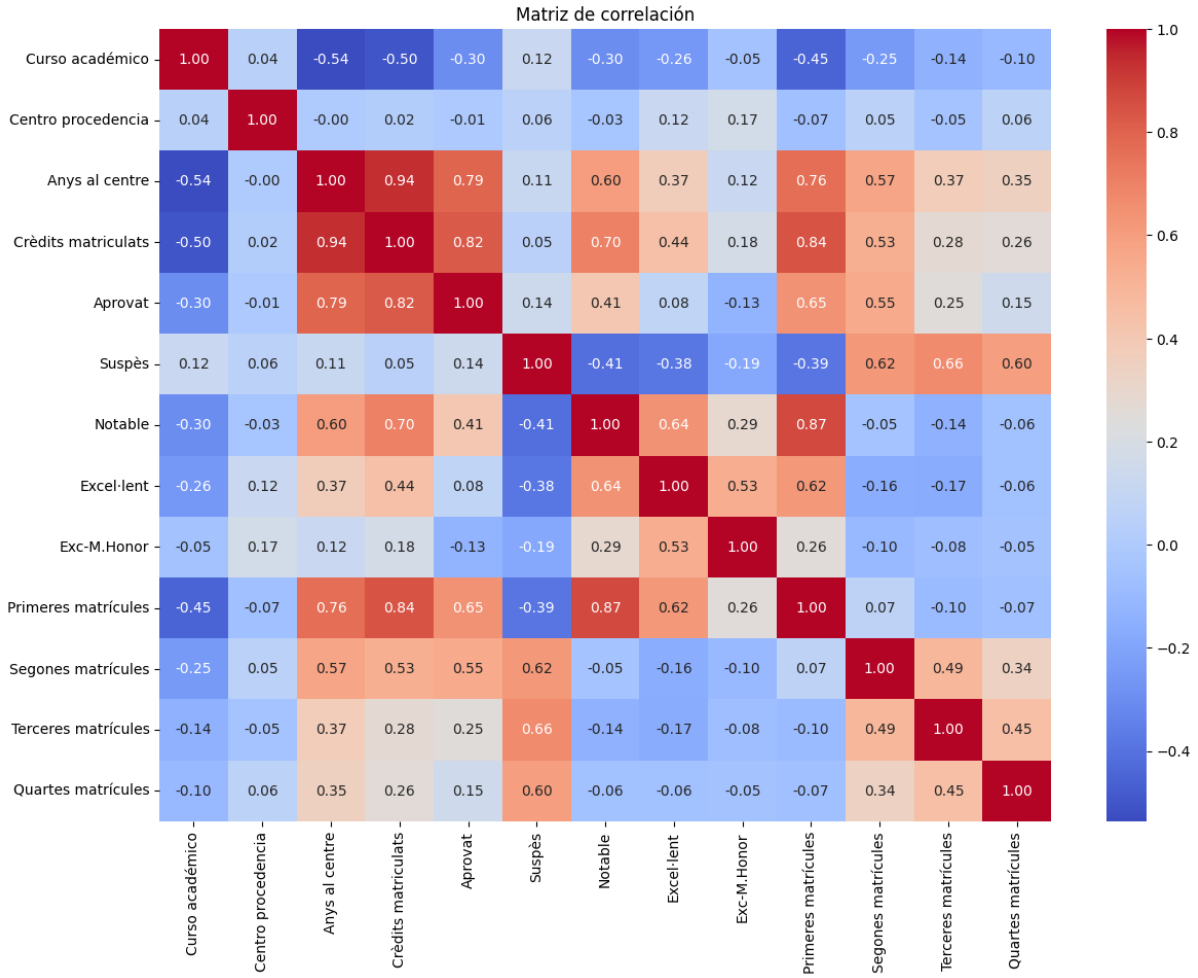
Com es pot veure a la imatge existeixen bastantes columnes on la correlació és molt baixa per tant aquestes seran eliminadas i al final el model es queda amb les següents columnes.

El gràfic adjunt correspon a les dades d'entrenament.



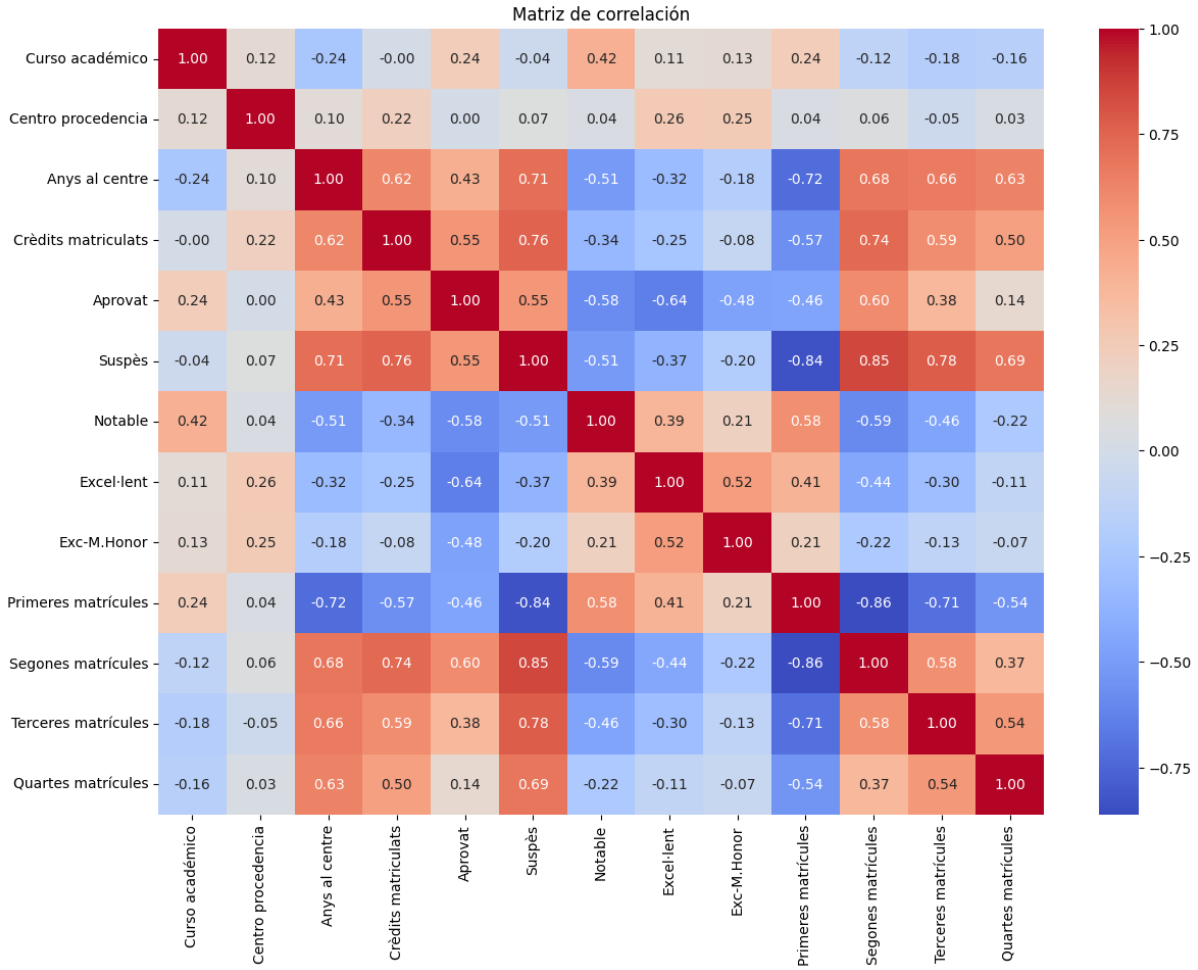
Imatge 17 Matriu de correlació dades train

El gràfic adjunt correspon a les dades de test



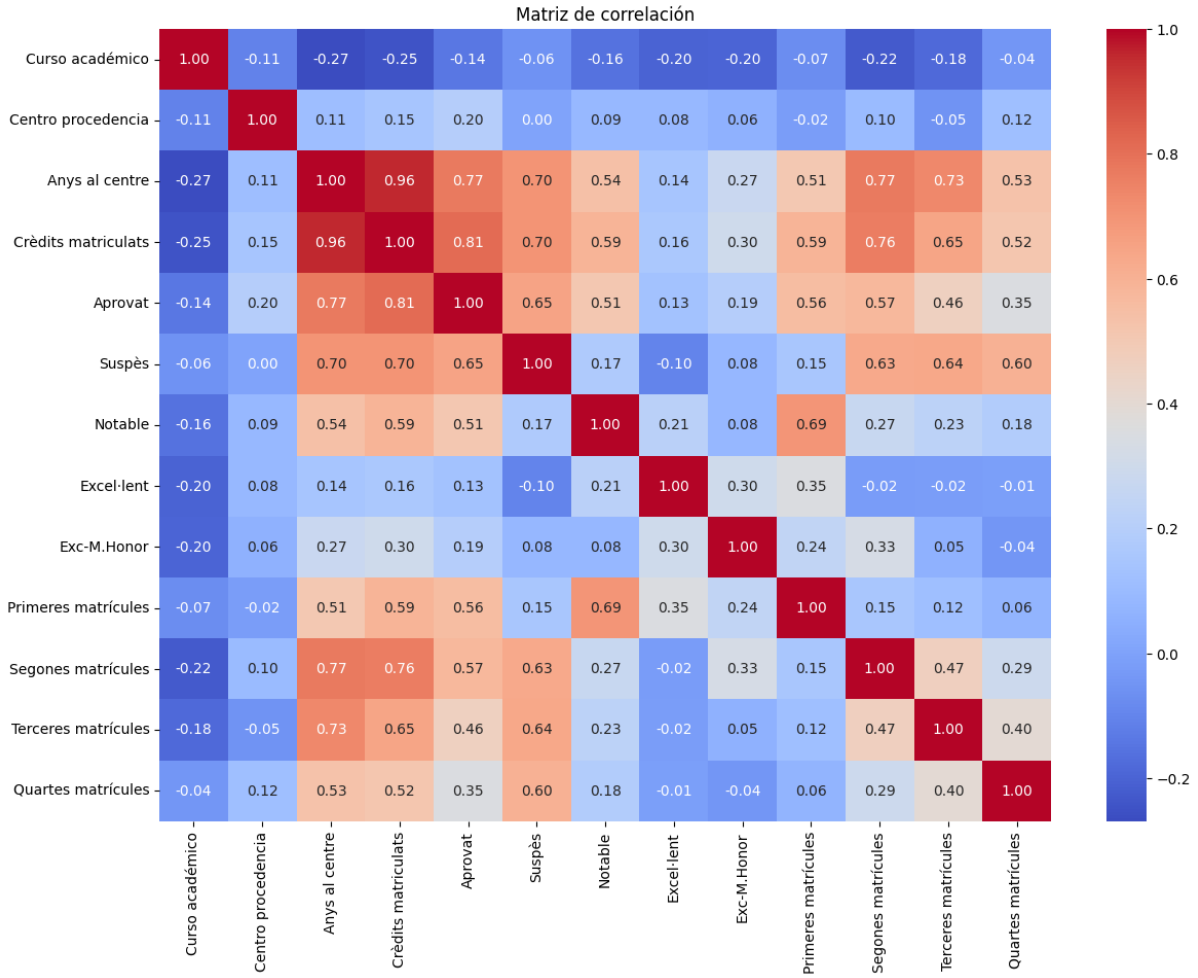
Imatge 18 Matriu de correlació dades test

El gràfic adjunt correspon a les dades d'alumnes graduats



Imatge 19 Matriu de correlació alumnes graduats

El gràfic adjunt correspon a les dades d'alumnes no graduats



Imatge 20 Matriu de correlació alumnes no graduats

5. Aplicació d'algorismes per classificar

Els algorismes plantejats per la perfilació dels alumnes consisteixen en l'aplicació de K-Means, on les dades d'alumnes graduats i no graduats seran dividides en dos dataframes per així poder realitzar la classificació .

Per a poder treballar les dades, s'ha modificat el dataframe per treballar amb la distància cosinus i generar un nou dataframe de similaritats al qual es pot aplicar l'algoritme per fer la classificació.

1. Dataset alumnes no graduats

1. K-Means

Per a la perfilació s'ha optat per aquest algoritme, és senzill i precís no obstant per la obtenció dels perfils s'ha aplicat diferents tècniques.

1. Centroide

En primer lloc, s'han identificat els punts més propers als centroides dels clústers generats. S'ha seleccionat un punt aleatori entre ells. Donat que aquests valors estaven força separats, no resultaven òptims, i la manca d'observacions tendia a produir dos perfils diferents d'estudiants:

Estudiants que es graduen en primer lloc o provenen de cicles formatius i batxillerat, però no tenen gaire èxit amb la majoria d'assignatures, el que els porta a abandonar en el primer any.

Estudiants que suspenen algunes assignatures però no moltes, i cada any van acumulant assignatures d'altres cursos fins arribar a un punt on es veuen superats per la quantitat d'assignatures pendents i acaben abandonant..

2. Mitjana dels clústers

La proposta dels centroides no acabava de donar els resultats desitjats, per tant, s'han agafat tots els punts dels clústers generats i s'ha fet una mitjana de totes les mostres obtingudes per tenir una visió extra de la generada.

Els resultats són semblants als anteriors, on es poden veure dos clústers diferenciats. Tot i que el model en genera quatre, la majoria de les observacions es concentren en dos clústers principals: vuitanta observacions corresponen a alumnes que abandonen al primer any, mentre que vint-i-cinc observacions corresponen a alumnes que aguanten uns quants anys al centre però finalment acaben abandonant.

El gràfic adjunt mostra una taula amb els 4 clústers generats.

Anys al centre	Crèdits matriculats	Aprovat	Suspès	Notable	Excel·lent	Exc-M.Honor	Primeres matricules	Segones matricules	Terceres matricules	Quartes matricules
1	69	3	4	0	0	0	14	0	0	0
3	136	9	7	3	0	0	18	2	1	1
1	60	5	7	0	0	0	12	0	0	0
3	144	10	9	2	0	0	20	3	1	0

Imatge 21 Mitjana dels clústers

2. Dataset alumnes graduats

1. K-Means

Per a la perfilació s'ha optat per aquest algoritme, és senzill i precís no obstant per la obtenció de les dades s'ha aplicat diferents tècniques.

1. Centroide

S'han identificat els punts més propers als centroides dels clústers generats. Un cop els punts més propers als centroides han estat escollits, es transformen les dades per a poder ser interpretades, i es pot veure com realment la gent que es matricula té un perfil bastant semblant. Els estudiants es divideixen en aquells que completen la carrera sense haver de repetir cap assignatura i aquells que han hagut de repetir alguna assignatura, però les assignatures suspeses representen una minoria.

2. Mitjana dels clústers

L'aplicació d'aquesta metodologia no ha proporcionat els resultats desitjats, ja que per calcular la quantitat de crèdits matriculats s'ha de revisar la informació d'altres taules, i hi ha alumnes que no tenen aquesta informació. Per tant, columnes com crèdits matriculats o anys al centre no es calculen de la manera més òptima, i en fer la mitjana de tots els valors dins de cada clúster, els resultats no tenen sentit tot i que aquells valors amb zero han estat omesos per fer la mitjana.

Com es pot observar al gràfic adjunt, la mitjana dels crèdits matriculats de cada un dels clústers mai acaba donant un resultat correcte.

Anys al centre	Crèdits matriculats	Aprovat	Suspès	Notable	Excel·lent	Exc-M.Honor	Primeres matrícules	Segones matrícules	Terceres matrícules	Quartes matrícules
3	193	5	0	18	8	5	43	0	0	0
4	192	11	1	16	4	1	41	2	0	0
4	180	7	1	17	6	1	42	0	0	0
4	212	5	0	19	8	7	44	0	0	0

Imatge 22 Mitjana dels clústers

6. Model de predicció

Les institucions educatives s'enfronten al repte de l'abandonament escolar, un fenomen que afecta negativament tant als estudiants com a les pròpies institucions. Per tal d'abordar aquesta problemàtica, es proposa el desenvolupament d'un model predictiu capaç d'identificar els alumnes amb risc d'abandonament i, per tant, implementar mesures preventives que afavoreixin la seva continuïtat en els estudis.

El model es construirà a partir de dades d'alumnes actuals, graduats i no graduats, incloent informació demogràfica, acadèmica i altres variables rellevants. Per tal d'assegurar la robustesa del model, es realitzarà una modificació de les dades, utilitzant informació fins al curs 2022-23 per a la selecció de l'algorisme més adequat i dades posteriors per a la seva validació.

S'espera que el model desenvolupat sigui capaç de predir amb precisió quins alumnes acaben la carrera i quins abandonen. No obstant això, cal tenir en compte que la precisió pot ser menor per a aquests últims, ja que la mostra de dades està desequilibrada a favor dels alumnes graduats.

La mostra de les dades no acaba d'estar del tot compensada, si parlem de percentatges de representació de les dues mostres es pot veure com les dades d'alumnes no graduats és a dir aquells que abandonen representen un quaranta per cent de les mostres totals mentres que el seixanta per cent de les dades restants són d'alumnes graduats. Alhora de fer prediccions els models faran millors prediccions d'aquells alumnes que acabaran graduant-se mentres que els que abandonaran no seran tant correctes.

En la configuració dels algorismes no s'ha especificat cap paràmetre concret. S'ha diferenciat entre les dades d'entrenament i de prova, i s'ha aplicat l'algoritme segons les indicacions de la biblioteca corresponent. L'execució del codi no ha presentat cap problema.

1. XGBoost

Els resultats proporcionats per l'ús d'aquest algoritme han estat un èxit del cent per cent. La mostra de dades representa unes dues-centes cinquanta observacions i les dades que s'està intentant predir unes quaranta.

```
Percentage d'encerts en el conjunt de proba: 100.0 %
Matriu de confusió en el conjunt de proba:
[[20  0]
 [ 0 18]]
Prediccions correctes en el conjunt de proba: 38
Prediccions incorrectes en el conjunt de proba: 0
```

Imatge 23 Accuracy XGBoost

Si ens fixem en els resultats, de la predicció amb les dades actuals es pot observar que una gran part dels alumnes que estan cursant actualment acabaran graduant-se, mentre que una part notable acaba abandonant.

```
Total mostres: 141
Graduats: 99
No graduats: 42
```

Imatge 24 Predicció XGBoost

2. Random forest

Els resultats proporcionats per aquest algoritme són idèntics als d'XGBoost, és a dir que l'accuracy és del cent per cent però a l'hora de realitzar les prediccions amb les dades actuals es pot veure com els alumnes que es graduaran o abandonaran varien en unes poques unitats.

```
Total mostres: 141
Graduats: 102
No graduats: 39
```

Imatge 25 Predicció Random forest

3. SVM

Els resultats de la utilització d'aquest algoritme són els pitjors, ja que la taxa d'encert en aquest cas no és del cent per cent, sinó del vuitanta per cent. Per tant, es pot descartar la utilització d'aquest algoritme per realitzar la predicció d'aquest estudi.

```
Percentage d'encerts en el conjunt de proba: 81.57894736842105 %
Matriu de confusió en el conjunt de proba:
[[18  2]
 [ 5 13]]
Prediccions correctes en el conjunt de proba: 31
Prediccions incorrectes en el conjunt de proba: 7
```

Imatge 26 Predicció SVM.

7. Anàlisi de resultats, conclusions i possibles ampliacions

1. Perfilat d'alumnes

Per finalitzar el treball i un cop s'han analitzat tots els algorismes i les dades disponibles, es poden identificar les possibles causes d'abandonament del grau d'informàtica.

En primer lloc, s'ha observat que hi ha dos perfils principals d'estudiants que decideixen abandonar la carrera. El primer perfil correspon a aquells estudiants que cursen un any i suspenen la majoria d'assignatures relacionades amb la informàtica, especialment les de programació i computació. Això suggereix que aquests alumnes no estan satisfets amb l'elecció del grau.

El segon perfil preocupa més. Inclou estudiants que han estat cursant la carrera durant anys, amb una gran part dels crèdits matriculats aprovats i algunes assignatures suspeses. Tot i el seu progrés, aquests alumnes decideixen abandonar.

Després de contactar amb alguns individus en aquesta situació, es poden identificar problemes com la insatisfacció amb el contingut de la carrera, ja que consideren que la formació rebuda no és òptima per als seus interessos. A més, molts tenen dificultats amb els horaris de matí, ja que compaginen feina i estudis, el que porta a pensar que són alumnes que han trobat feina o ja disposaven de feina durant la carrera i fer el sacrifici de compaginar ambdues coses amb el benefici que acabaran tenint no els acaba de convèncer.

És important destacar que aquestes observacions es basen en les experiències d'uns pocs alumnes que han abandonat, i que seria necessari un estudi més avançat per extreure conclusions definitives. Amb les dades disponibles de tots els cursos, resulta difícil obtenir conclusions precises.

No obstant això, els informes i l'anàlisi dels perfils mostren que les causes d'abandonament poden variar per a cada individu. Tot i així, es pot observar que la majoria d'estudiants que abandonen ho fan després de suspendre assignatures de la branca de programació, com ara

Fonaments de Programació, Programació Avançada, Programació Orientada a Objectes, Estructures de Dades i Algorismes.

D'altra banda, s'ha fet una maquetació dels alumnes que acaben graduant-se, on es pot veure amb els perfilats realitzats dels alumnes graduats es passen entre tres i quatre anys per acabar la carrera. Els que tarden tres anys és degut al fet que disposen de crèdits convalidats.

Si es fa una visualització més a fons de les qualificacions, es pot veure com les notes que s'obtenen són bastant bones, on notable representa la qualificació que té la majoria mentre que aprovat representa una part petita. Per concloure aquesta secció, les assignatures suspeses no representen un problema per aquests alumnes, tot i que existeixen alumnes amb assignatures suspeses, però la mitjana d'assignatures suspeses per tots els alumnes dóna un valor d'un.

2. Biaix de gènere

Com s'havia previst, un cop realitzat l'estudi de dades, s'ha constatat que les mostres del sexe femení són tan poques que s'ha decidit eliminar la diferenciació entre gèneres. De les gairebé cinc-centes observacions, només cinquanta representaven persones de sexe femení. Amb tan poques mostres, s'ha arribat a la conclusió que separar les mostres per gènere no seria significatiu per a la realització de l'estudi i la generació de diferents models.

No obstant això, en la part de mostreig de dades s'ha tingut en compte el gènere. És a dir, pels informes on es pot accedir a dades d'homes i dones, existeixen filtres per diferenciar les dades, però no s'ha fet cap comparació, no hauria estat una observació justa.

3. Possibles ampliacions

Per concloure la part de conclusions, s'hauria de tractar el tema de les ampliacions del projecte. Tot i haver complert tots els requeriments inicials, es poden fer millores significatives. Per exemple, es podria modificar Power BI per accedir a les dades en temps real, eliminant la necessitat de descarregar-les localment i tractar-les manualment. També es podria dissenyar nous informes dins del producte i facilitar la modificació i addició de noves dades.

En relació amb el que s'ha esmentat en l'apartat anterior, es podria unificar els dos scripts generats, un per al tractament i processament de les dades i l'altre per a la generació dels models, perquè treballin dins del propi producte de Power BI, facilitant així la usabilitat del producte generat.

La darrera ampliació que es podria realitzar és l'addició dels alumnes de cadascuna de les carreres del TecnoCampus, per tenir informes similars als realitzats per a cada carrera que es presenta al centre.

4. Conclusions

Les tècniques de classificació i clusterització utilitzades han permès identificar perfils diferenciats d'alumnes que abandonen i que es graduen. Tot i així, les limitacions de les dades i la necessitat d'estudis més avançats impedeixen treure conclusions definitives. Les prediccions realitzades amb XGBoost i Random Forest han demostrat ser les més efectives per avaluar els alumnes actuals, mentre que SVM s'ha descartat degut a la seva menor precisió. És recomanable seguir treballant en l'obtenció de més dades per millorar els models i les conclusions.

8. Testing

En el desenvolupament del projecte, no s'han realitzat proves específiques detallades en un moment concret del cicle de desenvolupament. En lloc d'això, s'ha adoptat una estratègia de testing basada en iteracions periòdiques amb la clienta. Aquest enfocament ha permès una revisió contínua del producte, assegurant que les millores i ajustaments necessaris es poguessin implementar de manera eficient i oportuna.

El procés d'iteracions amb la clienta s'ha estructurat en cicles de 3-4 setmanes. Cada iteració s'ha centrat en els següents passos:

Durant cada sessió, es presenta a la clienta l'estat actual del producte, incloent-hi les funcionalitats desenvolupades i els canvis realitzats des de l'última iteració.
La clienta proporciona comentaris i suggeriments sobre el producte, destacant aspectes que funcionen correctament i àrees que requereixen millores.
Es discuteixen les possibles millores a implementar en les següents iteracions, tenint en compte el feedback rebut. Aquestes millores poden incloure nous informes o ajustos a informes ja existents.
Es defineixen els objectius per a la següent iteració, assegurant-se que les prioritats de la clienta estan clarament establertes i alineades amb els objectius del projecte.

Per concloure, el testing basat en iteracions amb la clienta ha estat una part fonamental del procés de desenvolupament del projecte. Aquesta metodologia ha permès una col·laboració estreta amb la clienta, assegurant que el producte s'ajusta a les seves necessitats i preferències.

9. Bibliografia

[1] Col·laboradors dels projectes Wikimedia, "Tecnocampus - Viquipèdia, l'enciclopèdia lliure", Viquipèdia, l'enciclopèdia lliure. [En línia]. Consultat el 14 de novembre de 2023. Disponible: https://es.wikipedia.org/wiki/Tecnocampus.

[2] C. E. Aguirre and J. C. Pérez, "Predictive data analysis techniques applied to dropping out of university studies," 2020 XLVI Latin American Computing Conference (CLEI), Loja, Ecuador, 2020, pp. 512-521, doi: 10.1109/CLEI52000.2020.00066.

[3] "Calculadora sueldo neto 2023 y 2024: Pasa tu salario bruto a neto | Bankinter," Bankinter. Consultat el 8 de desembre de 2023 [En línia]. Disponible: https://www.bankinter.com/blog/economia/como-varia-sueldo-neto-funcion-salario-bruto-graficos.

[4] Generalitat de Catalunya, "Portal Jurídic de la Generalitat de Catalunya", Portal Jurídic de la Generalitat de Catalunya. [Consultat el 15 de desembre de 2023]. Disponible: https://portaljuridic.gencat.cat/ca/inici/.

[5] Blogthinkbig, "Ahorra en tu factura de luz con soluciones en la nube," BlogThinkBigEmpresas, Gener 15, 2022. [En Línea]. Consultat el 18 de desembre de 2023. Disponible: https://empresas.blogthinkbig.com/ahorra-factura-de-luz-con-soluciones-cloud/.

[6] G. Navarro, "Títol del document," Treball de Fi de Grau, Universitat Oberta de Catalunya, juliol 2016. [En Línia]. [Data D'accés: 01/01/2024]. Disponible: https://openaccess.uoc.edu/bitstream/10609/53504/7/gnavarroTFG0716memòria.pdf.

[7] Universidad Pablo de Olavide, "Títol del document o pàgina," Universitat Pablo de Olavide, Consultoria Data Mining, Machine Learning y Big Data. [En línia]. [Data d'accés: 5 de gener de 2024]. Disponible: https://www.upo.es/upotec/catalogo/consultoria-gestion-y-servicios-empresariales/consultoria-data-mining-machine-learning-y-big-dat/.

[8] Parlament Europeu i Consell, "Proposta de Reglament del Parlament Europeu i del Consell per a l'establiment de normes harmonitzades en matèria d'intel·ligència artificial (Llei d'intel·ligència artificial) i modificació de certes actes legislatives de la Unió," COM/2021/206 final, 2021.

[9] Autor(s), "Scikit-learn: Descobreix la biblioteca Python," DataSciencetests. Disponible: https://datascientest.com/es/scikit-learn-decubre-la-biblioteca-python . Consultat: 5 d'abril de 2024.

Índex de figures

Jon Morales Martí

Grau en Enginyeria Informàtica de Gestió i Sistemes d'Informació

Fundació TecnoCampus
Mataró-Maresme
Avinguda d'Ernest Lluch, 32
08302 Mataró(Barcelona)
Tel 931696501
www.tecnocampus.cat

TecnoCampus
Escola Superior Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona