Gerard Pascual Fontanilles

Eina de recollida i anàlisi automàtic de dades sobre elecció d'estudis a Batxillerat

TREBALL DE FI DE GRAU

Aïda Valls Mateu i Carme Olivé Farré

Grau d'Enginyeria Informàtica

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2025

Resum.

Aquest projecte neix per donar resposta a una necessitat real en l'àmbit educatiu, tecnològic i social. L'objectiu és analitzar informació obtinguda mitjançant enquestes a estudiants de batxillerat per extreure coneixement útil que permeti entendre millor els factors determinants en l'elecció d'estudis, amb especial atenció a la bretxa de gènere en les disciplines STEM. La seva justificació es fonamenta en la importància de disposar de dades estructurades i visualitzacions clares que ajudin els centres educatius en la presa de decisions, així com en la urgència d'incrementar la participació femenina en l'àmbit tecnològic per avançar cap a una societat més igualitària i innovadora.

La metodologia emprada inclou, en primer lloc, l'anàlisi de requeriments i la selecció d'eines específiques (Microsoft Forms per a la recollida de respostes i Python per a l'anàlisi posterior). Les dades han estat tractades mitjançant processos de neteja i codificació, incloent la detecció automatitzada de respostes invàlides i la classificació de textos lliures mitjançant tècniques de similitud semàntica. A partir d'aquestes dades processades, s'han desenvolupat dues eines: una per generar informes gràfics adaptats als centres de batxillerat, i una altra per analitzar dades de manera intel·ligent a través de tècniques d'aprenentatge no supervisat (com k-modes, DBSCAN amb distància de Gower i clústering jeràrquic).

Els resultats obtinguts mostren la viabilitat d'una eina capaç d'oferir informes automatitzats i perfils clars d'estudiants segons sexe, itinerari de batxillerat i interessos, facilitant la identificació de patrons rellevants per dissenyar accions educatives més efectives i inclusives.

Resumen.

Este proyecto nace con la voluntad de dar respuesta a una necesidad real en el ámbito tecnológico y social. El objetivo es analizar información procedente de encuestas de bachillerato para extraer conocimiento útil que permita comprender los factores determinantes en la elección de estudios, con especial énfasis en la brecha de género en áreas STEM. La justificación parte de la importancia de obtener datos estructurados y visualizaciones claras que faciliten la toma de decisiones en los centros educativos, así como de la urgencia de aumentar la participación femenina en estudios tecnológicos para contribuir a la igualdad y la innovación.

La metodología empleada incluye, en primer lugar, el análisis de requerimientos y la selección de herramientas específicas (Microsoft Forms para la recogida de respuestas y Python para el análisis posterior). Los datos han sido tratados mediante procesos de limpieza y codificación, incluyendo la detección automatizada de respuestas no válidas y la clasificación de textos libres mediante técnicas de similitud semántica. A partir de estos datos procesados, se han desarrollado dos herramientas: una para generar informes gráficos adaptados a los centros de bachillerato, y otra para analizar datos de forma inteligente a través de técnicas de aprendizaje no supervisado (como k-modes, DBSCAN con distancia de Gower y clúster jerárquico).

Los resultados obtenidos muestran la viabilidad de una herramienta capaz de ofrecer informes automatizados y perfiles claros de estudiantes según sexo, itinerario de bachillerato e intereses, facilitando la identificación de patrones relevantes para diseñar acciones educativas más efectivas e inclusivas.

Abstract.

This project was born out of the desire to address a real need in both the technological and social spheres. The aim is to analyze information gathered from high school surveys in order to extract useful insights that help understand the key factors influencing students' choice of studies, with a particular focus on the gender gap in STEM fields. The rationale is based on the importance of obtaining structured data and clear visualizations that support decision-making in educational institutions, as well as the urgent need to increase female participation in technological studies to foster equality and innovation.

The methodology used includes, first of all, the analysis of requirements and the selection of specific tools (Microsoft Forms for collecting responses and Python for subsequent analysis). The data was processed through cleaning and coding procedures, including the automated detection of invalid responses and the classification of open-text answers using semantic similarity techniques. Based on this processed data, two tools were developed: one to generate graphical reports tailored to high schools, and another to intelligently analyze data using unsupervised learning techniques (such as k-modes, DBSCAN with Gower distance, and hierarchical clustering).

The results obtained demonstrate the feasibility of a tool capable of providing automated reports and clear student profiles based on gender, high school track, and interests, facilitating the identification of relevant patterns to design more effective and inclusive educational strategies.

Índex

Índex de taules

Índex de figures

1 Introducció

L'elecció dels estudis universitaris és un dels moments més decisius en la trajectòria acadèmica i professional dels joves, amb repercussions significatives tant a nivell personal com social. Aquesta decisió, que sovint es comença a configurar durant el batxillerat, està influïda per múltiples factors: des de les expectatives familiars fins als referents socials, passant per les percepcions sobre les diferents professions i la confiança que tenen els joves en les pròpies capacitats..

En aquest context, l'elecció d'estudis vinculats als àmbits STEM (acrònim en anglès de Science, Technology, Engineering i Mathematics) adquireix una rellevància especial, tant per la seva alta demanda laboral com pel seu paper central en el desenvolupament econòmic i tecnològic. Diversos informes recents adverteixen que, a Espanya, es necessitaran més de 200.000 nous enginyers durant la pròxima dècada [4], un repte que només es podrà afrontar amb una participació més equitativa entre homes i dones.

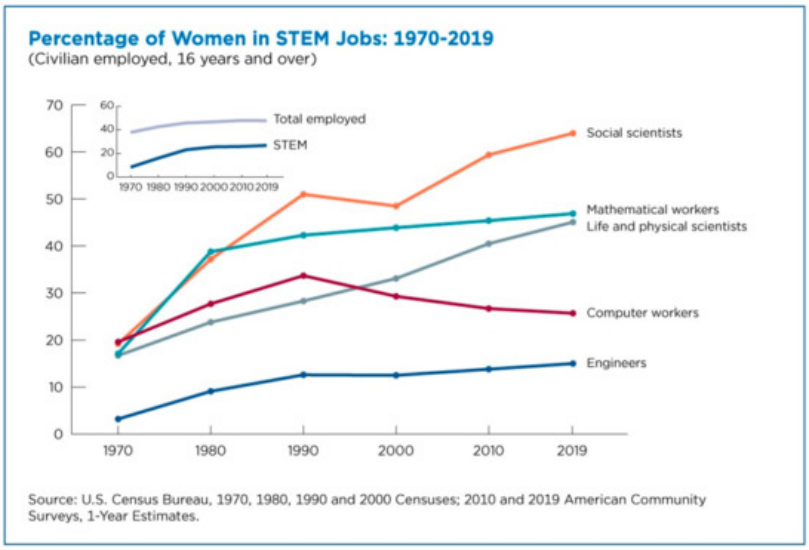


Figura 1. Percentatge dones en treball STEM

Segons dades de la U.S. Census Bureau (veure Figura 1), tot i l'increment de la participació femenina en àmbits com les ciències socials o les matemàtiques, disciplines com l'enginyeria i la informàtica continuen estant fortament masculinitzades. De fet, l'augment de la presència de dones en les àrees STEM no s'ha produït de manera uniforme, i algunes especialitats, particularment la informàtica, han experimentat fins i tot un retrocés en les últimes dècades. Aquest fet posa de manifest la necessitat d'entendre amb més profunditat els factors que influeixen en les decisions vocacionals abans de l'etapa universitària.

Aquest desequilibri de gènere no només té implicacions en termes d'equitat, sinó també profundes conseqüències socials i econòmiques. A Catalunya, per exemple, només entre un 12% i un 14% de l'alumnat dels graus vinculats a la informàtica són dones [1], una dada que evidencia una bretxa persistent en l'accés a l'educació superior tecnològica. Aquesta situació és preocupant perquè pot limitar el potencial de creixement econòmic i d'innovació del país, ja que la manca de talent qualificat en aquests àmbits pot arribar a frenar el desenvolupament tecnològic i la competitivitat.

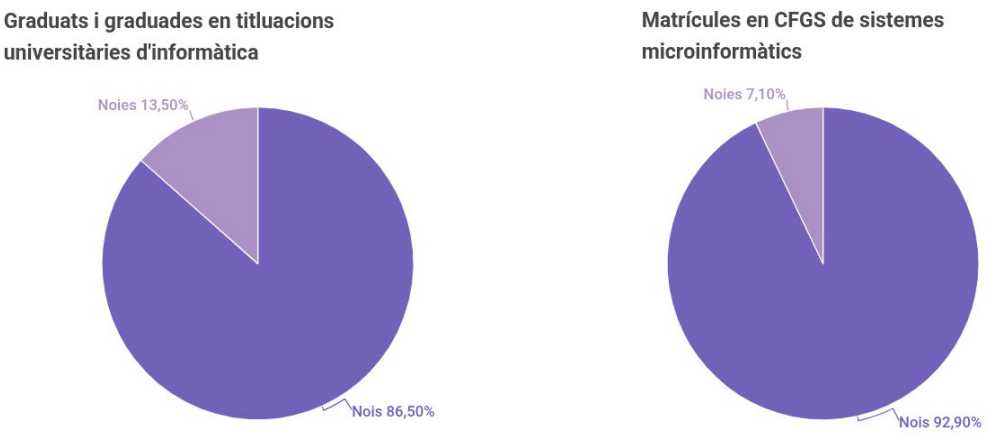


Figura 2. Proporció de nois i noies en estudis d'informàtica (universitaris i CFGS) a Catalunya [3]

Per entendre millor aquesta realitat, resulta fonamental analitzar les percepcions, expectatives i motivacions dels estudiants abans que prenguin decisions acadèmiques decisives, com ara l'elecció de l'itinerari de batxillerat (humanístic, ciències socials, ciències i tecnologia, general o arts). Aquesta etapa educativa és especialment rellevant, ja que els estudiants encara no han escollit definitivament els seus estudis universitaris, la qual cosa permet explorar amb més claredat els factors que influeixen en les seves vocacions i interessos. Sovint, la recerca se centra en decisions ja preses, cosa que dificulta la identificació de les causes subjacents. En canvi, l'anàlisi d'aquesta fase prèvia permet obtenir una visió més precisa de com es construeixen les vocacions i per què nois i noies tendeixen a optar o no per carreres de caire tecnològic.

Una recerca publicada per la UNESCO subratlla que les percepcions sobre la pertinença de les professions STEM per a les dones es construeixen ja en les primeres etapes del sistema educatiu i tendeixen a reforçar-se durant l'adolescència [2]. Aquest fet posa de manifest la importància d'intervencions específiques en l'etapa del batxillerat, un moment clau en què els joves perfilen les seves aspiracions i prenen decisions que condicionaran el seu futur acadèmic i professional.

Altres estudis, com el de la Fundació Bofill, remarquen que la manca de referents femenins visibles en l'àmbit científic i tecnològic contribueix de manera significativa a la persistència d'aquesta bretxa de gènere, ja que limita les expectatives i l'autoeficàcia percebuda de moltes noies a l'hora de considerar aquestes opcions com a viables [3].

Una de les conseqüències més rellevants de la baixa presència femenina en les carreres STEM és la seva incidència directa sobre la bretxa salarial de gènere. Les professions vinculades a la ciència i la tecnologia solen estar associades a millors condicions laborals i salaris més alts; per tant, la infrarepresentació de les dones en aquests sectors contribueix, de manera estructural, a perpetuar desigualtats econòmiques entre homes i dones [1]. Però l'impacte va més enllà de la qüestió econòmica: l'enginyeria, la ciencia i la tecnologia tenen un paper fonamental en el disseny i la implementació de solucions que afecten la vida quotidiana de tota la societat. Quan aquestes decisions es prenen majoritàriament des d'una òptica masculina, es corre el risc que les necessitats i experiències de les dones quedin invisibilitzades o mal representades en els processos d'innovació i desenvolupament tecnològic.

Aquest treball s'emmarca dins un projecte multidisciplinari amb la participació d'investigadors i investigadores dels departaments d'Economia Aplicada, Sociologia i

Informàtica i Matemàtiques. L'objectiu central és aprofundir en aquests factors i identificar els determinants clau que influeixen en l'elecció d'estudis postobligatoris, especialment aquells vinculats als àmbits tecnològics. El projecte pretén entendre per què les vocacions femenines continuen sent clarament minoritàries en aquestes disciplines.

Per això, s'ha dut a terme una anàlisi qualitativa i quantitativa sobre la percepció que tenen els i les estudiants de primer de batxillerat respecte al seu futur acadèmic i professional, amb l'objectiu de detectar possibles diferències de gènere pel que fa a les aspiracions, els interessos i els condicionants interns o externs que influeixen en aquestes decisions.

1.1 Objectius

L'objectiu general d'aquest projecte és dissenyar i implementar una eina informàtica que permeti automatitzar la recollida i anàlisi de dades sobre la percepció i les expectatives dels estudiants respecte als diferents tipus d'estudis, especialment en l'etapa prèvia a la tria definitiva de la seva trajectòria acadèmica, com és el primer curs de batxillerat. Aquesta eina ha de permetre obtenir informació rellevant per comprendre millor les causes de la baixa participació femenina en els àmbits tecnològics i científics.

A partir d'aquest objectiu general, es defineixen els següents objectius específics:

1. **Desenvolupar una eina informàtica intuïtiva i accessible** que permeti administrar enquestes en centres de secundària de manera àgil i eficient, minimitzant la càrrega administrativa per als instituts i garantint una recollida de dades sistemàtica i de qualitat.
2. **Homogeneïtzar i estructurar la informació** recollida per tal de generar informes estandarditzats que facilitin la comparació de resultats entre diferents centres educatius, cursos escolars i classes d'estudiants, promovent una anàlisi transversal rigorosa.
3. **Aplicar tècniques d'Intel·ligència Artificial (IA)** en l'anàlisi de les dades recollides amb la finalitat d'identificar patrons de resposta i definir perfils d'estudiants segons variables com el gènere, el context socioeconòmic, les expectatives familiars o la percepció sobre les professions STEM.
4. **Detectar i caracteritzar els factors clau que influeixen en l'elecció o el rebuig dels estudis tecnològics**, posant èmfasi en les diferències de gènere, per tal de generar coneixement útil per al disseny de polítiques educatives i accions de sensibilització més efectives.

Amb aquests objectius, aquest projecte busca contribuir a una millor comprensió dels factors que distingeixen diferents perfils d'estudiants i que influeixen en la tria dels diferents tipus d'estudis, com els tecnològics i científics versus els humanístics, socials i artístics.

2 Requeriments

Per tal de facilitar una planificació clara i estructurada, els requeriments es divideixen en quatre fases principals, cadascuna centrada en un àmbit específic del procés de treball.

A més, dins de cada fase es distingeixen dos tipus de requeriments:

- **Requeriments funcionals:** descriuen les funcionalitats específiques que el sistema ha de complir, les accions que ha de poder realitzar i els resultats esperats.
- **Requeriments no funcionals:** estableixen criteris de qualitat, restriccions tècniques, usabilitat, eficiència i altres característiques generals que no impliquen accions directes del sistema.

A continuació, es detallen els requeriments associats a cada fase, començant per la fase inicial, que se centra en la construcció del qüestionari i la definició dels criteris per a una recollida de dades estructurada.

2.1 Gestió del qüestionari i recollida de dades

Una de les primeres fases del projecte consisteix a definir com es realitzarà la recollida de dades mitjançant el qüestionari.

2.1.1 Requeriments no funcionals (NF) fase 1

NF1. És fonamental garantir una qualitat de la informació recollida, ja que tota l'anàlisi posterior dependrà de la fiabilitat i precisió d'aquestes dades inicials.

NF2. El sistema utilitzat ha de garantir la seguretat de les dades al llarg de tot el procés, des de la recollida fins a l'emmagatzematge i l'anàlisi.

NF3. L'accés al qüestionari ha de ser universal i accessible des de qualsevol dispositiu, garantint així la participació equitativa de tots els alumnes.

NF4. L'enquesta ha d'estar disponible mitjançant un enllaç públic, sense requerir cap tipus de registre previ.

NF5. Per protegir la identitat dels participants, les respostes han de ser pseudoanònimes, evitant l'ús de noms o codis identificatius directes en tot moment.

2.1.2 Requeriments funcionals (F) fase 1

F1. El projecte parteix d'un qüestionari ja existent (veure Annex 1), dissenyat pel grup de recerca liderat per la Dra. Teresa Corbella i validat per la Comissió d'Ètica en Recerca i Innovació (CERI) de la URV, establint que aquest qüestionari serà la base per garantir una estructura metodològica sòlida.

F2. Es desenvolupa un sistema estructurat de gestió de preguntes que permet la reutilització del qüestionari, facilitant la seva adaptació tant a format digital com en paper, optimitzant així la seva administració.

F3. La versió digital del qüestionari ha de ser implementada mitjançant una eina que ofereixi un entorn segur i controlat per a la recollida de dades, garantint el compliment estricte de la normativa de protecció de dades.

F4. Les preguntes digitals s'han de redactar amb la màxima similitud possible al format en paper, tant en contingut com en estructura, per tal de mantenir la coherència i comparabilitat entre ambdós formats.

F5. Les respostes recollides s'han d'organitzar per centre educatiu i emmagatzemar en fitxers separats per facilitar el seu tractament posterior, assegurant que la propietat d'aquestes dades recaigui exclusivament en la coordinadora del projecte, responsable de la seva gestió i conservació.

F6. En cas d'utilitzar enquestes en paper, cal definir un sistema senzill i eficient per a la seva digitalització manual, permetent introduir les respostes de forma estructurada i fiable.

2.2 Generació i visualització de l'informe de gràfics

Una fase clau del projecte consisteix en la representació visual de les dades recollides a través del qüestionari, amb l'objectiu de facilitar la seva anàlisi i interpretació.

2.2.1 Requeriments no funcionals fase 2

NF6. L'informe gràfic ha de mostrar la informació de manera clara i accessible, evitant dissenys visuals complexos que puguin dificultar la comprensió.

NF7. Cal garantir una coherència visual entre tots els gràfics de l'informe, mitjançant l'ús d'una paleta de colors homogènia i una estructura gràfica consistent, que faciliti la comparació i la lectura.

NF8. La classificació de les respostes ha de respectar estrictament les categories definides prèviament durant la fase de disseny de l'enquesta.

NF9. Per protegir la confidencialitat dels participants, no s'inclouran en l'informe gràfics que mostrin categories amb menys de tres respostes, evitant així la possibilitat d'identificació indirecta d'estudiants en grups reduïts.

NF10. El sistema de recompte ha d'assegurar una coherència interna entre preguntes i categories, aplicant criteris unificats que permetin una interpretació correcta i homogènia dels resultats visuals en tot l'informe.

2.2.2 Requeriments funcionals fase 2

F7. El sistema ha de generar automàticament un informe gràfic personalitzat per a cada centre educatiu, elaborat exclusivament a partir de les seves dades específiques.

F8. A més dels informes personalitzats per a cada institut, s'ha de generar un informe gràfic global que integri les dades agregades de tots els centres participants.

F9. Les dades obtingudes han de ser processades agrupant-les en categories predefinides per a cada pregunta tancada, amb l'objectiu de facilitar un recompte i una representació gràfica estandarditzada.

F10. En el cas de respostes múltiples o obertes, el sistema ha de poder aplicar tècniques de classificació per identificar, agrupar o transformar les respostes en categories vàlides i coherents.

F11. Per a preguntes amb múltiples respostes dins una mateixa categoria, el sistema ha de garantir que cada alumne sigui comptabilitzat només una vegada per categoria, evitant així possibles distorsions estadístiques.

F12. El recompte i la representació gràfica final s'han d'aplicar de manera homogènia en tots els informes, tant institucionals com globals, per assegurar coherència i facilitar la comparació entre resultats.

2.3 Preparació i neteja de dades

Un cop recollides les respostes i generat l'informe preliminar de gràfics, el projecte inclou una fase fonamental de neteja i preparació de les dades per assegurar la qualitat i fiabilitat de l'anàlisi posterior.

2.3.1 Requeriments no funcionals fase 3

NF11. El conjunt de dades resultant ha de ser complet, estructurat i coherent, sense errors que puguin comprometre la validesa dels anàlisis posteriors.

NF12. El procés de classificació de les respostes obertes ha de minimitzar errors d'interpretació i garantir la coherència conceptual amb les categories preestablertes.

NF13. Totes les dades del conjunt final han d'estar codificades de manera estandarditzada i consistent, seguint convencions internacionals reconegudes.

2.3.2 Requeriments funcionals fase 3

F13. El sistema ha de detectar i marcar automàticament les entrades duplicades per evitar que afectin les estadístiques agregades.

F14. Les respostes invàlides, incloses les de persones majors de 20 anys o amb contingut ofensiu, han de ser identificades i etiquetades, però no eliminades; s'han de recollir en un registre separat.

F15. El sistema ha d'analitzar totes les respostes obertes, especialment les provinents de camps de text lliure o "altres", i assignar-les a una categoria coherent quan sigui possible.

F16. Quan una resposta expressi indecisió o manca de coneixement (per exemple, "no ho sé"), aquesta ha de ser etiquetada amb un codi específic de dubte.

F17. Els mètodes de classificació han de garantir un alt percentatge d'encert, validat prèviament.

F18. El sistema ha d'utilitzar codis oficials, com l'ISO 3166-1 per a la representació de països, per assegurar la uniformitat.

F19. Les situacions especials derivades del funcionament del formulari (respostes en blanc, no mostrades per condicions o dubtoses) han d'estar identificades amb codis tècnics únics:

- 991 – Resposta no mostrada
- 992 – Resposta en blanc
- 993 – Resposta dubtosa o incerta
- 999 – Entrada duplicada

F20. Aquest sistema de codificació ha de permetre aplicar mètodes automàtics d'anàlisi i classificació, i ha de ser compatible amb processos com el clústering.

2.4 Anàlisi automàtica i interpretació dels resultats

Un cop completades les fases de neteja, estructuració i codificació de les dades, es defineix una quarta etapa enfocada a l'aplicació d'intel·ligència artificial per dur a terme una anàlisi avançada.

2.4.1 Requeriments no funcionals fase 4

NF14. El sistema ha d'utilitzar mètodes de clústering categòric adequats, prioritzant tècniques validades en l'àmbit acadèmic i reconegudes per la seva robustesa.

NF15. Els resultats obtinguts mitjançant el clústering han de ser interpretables i consistents, evitant la formació de clústers massa amplis o dispersos.

NF16. Cal garantir que el sistema ofereix una avaluació objectiva del rendiment del clústering mitjançant indicadors quantitatius reconeguts.

NF17. L'anàlisi ha de facilitar una interpretació educativa dels resultats, posant en relleu les característiques comunes dins de cada clúster i les diferències entre grups, especialment pel que fa a variables com el sexe i el tipus de batxillerat.

2.4.2 Requeriments funcionals fase 4

F21. El sistema ha de permetre seleccionar entre diversos algorismes de clústering categòric, implementats amb llibreries consolidades i de reconeguda fiabilitat.

F22. Cada clúster ha de ser analitzat internament per identificar la moda (resposta més freqüent) de cada variable, així com els valors secundaris i la seva distribució percentual, per oferir una radiografia completa del perfil típic del grup.

F23. Els resultats han d'incloure una anàlisi exhaustiva de la distribució de respostes dins de cada clúster, permetent avaluar-ne la heterogeneïtat.

F24. A partir dels resultats del clústering, el sistema ha de generar fitxers d'anàlisi que mostrin la distribució dels clústers segons variables clau, facilitant la identificació de patrons, com ara la presència diferencial de perfils segons sexe o itinerari educatiu.

3 Disseny

Un cop definits els requeriments del projecte, es va iniciar la fase de disseny, durant la qual aquests requeriments es van traduir en decisions tècniques concretes. L'objectiu principal d'aquesta etapa és construir una estructura funcional, coherent i ajustada a les limitacions i capacitats de les eines seleccionades.

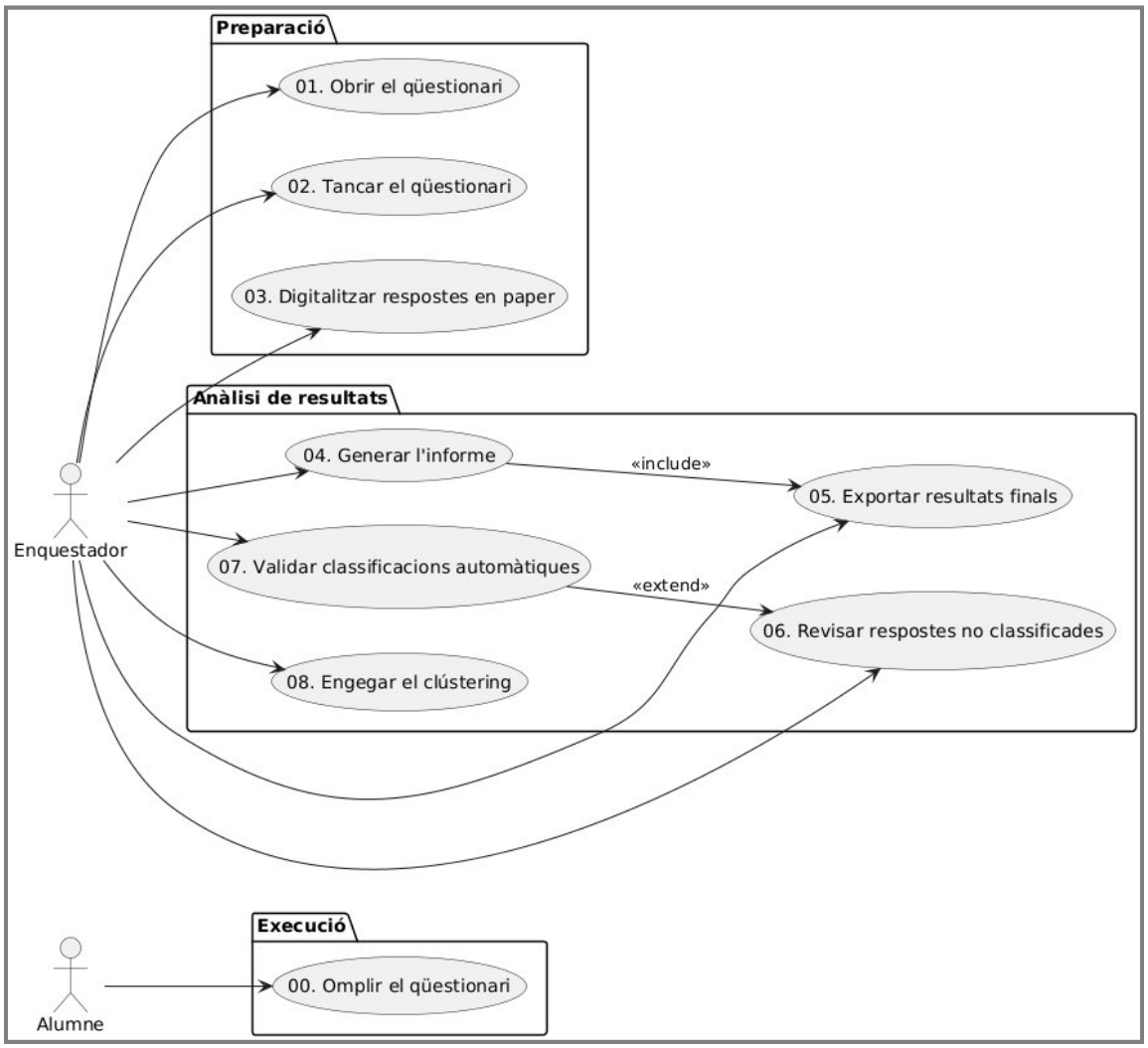


Figura 3. Diagrama de casos d'ús del sistema de gestió del qüestionari i anàlisi de dades

S'ha fet ús de paquets per reflectir la separació funcional de tasques i millorar l'organització visual del diagrama de casos d'ús. (Veure Figura 3)

ID	Resum funcionalitat	Paràmetres entrada	Paràmetres sortida	Actors	Precondició	Postcondició	Procés normal	Excepcions
00. Omplir el qüestionari	L'alumne accedeix al qüestionari, el completa i envia les seves respostes	Accés al qüestionari en línia	Respostes guardades al sistema	Alumne	Qüestionari accessible	Respostes guardades correctament	1. Accedeix a l'enllaç 2. Respon 3. Envia	Qüestionari no disponible / Error de connexió
01. Obrir el qüestionari	L'enquestador habilita el qüestionari per als alumnes	Configuració d'obertura	Qüestionari actiu	Enquestador	Qüestionari creat	Qüestionari disponible públicament	1. Accedeix a l'eina 2. Obre el qüestionari	Error de configuració / L'eina no respon
02. Tancar el qüestionari	L'enquestador desactiva l'accés un cop finalitzat	Identificador del qüestionari	Qüestionari tancat	Enquestador	Qüestionari obert	Ja no es poden enviar respostes	1. Accedeix a l'eina 2. Tanca el qüestionari	Errors de permisos / Bloquejos
03. Digitalitzar respostes	Introdueix manualment respostes en paper	Respostes en paper	Registre digital	Enquestador	Qüestionari en paper completat	Respostes digitals guardades	1. Accedeix al formulari 2. Introdueix les respostes	Errors de transcripció / Entrades incompletes
04. Generar l'informe	Es crea un document amb gràfics i resultats	Fitxer Excel	Informe Word o PDF	Enquestador	Dades carregades i validades	Informe disponible	1. Executa el codi 2. Crea el document	Error de lectura / Errors de gràfics
05. Exportar resultats	Es guarda una còpia dels resultats	Informe generat	Arxiu .docx o .xlsx	Enquestador	Informe generat correctament	Arxiu exportat disponible	1. Tria exportar 2. El sistema desa	Problemes d'escriptura / Format incompatible
06. Revisar respostes no classificades	Revisió manual d'entrades no classificades	Respostes pendents	Classificacions actualitzades	Enquestador	Respostes dubtoses identificades	Respostes classificades	1. Visualitza les entrades 2. Assigna categoria	Classificació no possible o ambigua
07. Validar classificacions	Revisió i aprovació de classificacions automàtiques	Resultats del classificador	Validació o modificació d'etiquetes	Enquestador	Sistema ha classificat respostes	Classificacions validades o corregides	1. Mostra classificacions 2. Valida o edita	Massa errors / Dubtes de context
08. Engegar el clústering	Execució dels algorismes de clústering	Dataset categoritzat	Assignació de clústers	Enquestador	Dataset net i preparat	Entrades assignades a clústers	1. Inicia el procés 2. Executa algorisme	Dades no preparades / Paràmetres incorrectes

Taula 1. Taula de casos d'ús del projecte

El disseny es va organitzar en quatre àmbits diferenciats, corresponents a cada fase de requeriments: el qüestionari digital, la generació de l'informe de gràfics, la preparació de dades i l'anàlisi amb tècniques de clústering.

3.1 Disseny del qüestionari digital

Partint de l'enquesta base en paper, validada pel CERI, es va dissenyar una versió digital mitjançant un formulari web. Durant la selecció de la plataforma per crear l'enquesta, es van considerar diverses opcions populars: Google Forms, Typeform i Microsoft Forms, valorant-ne els punts forts i les limitacions principals.

- Google Forms és àmpliament utilitzat per la seva simplicitat i gratuïtat. Permet crear qüestionaris de forma ràpida, amb preguntes condicionals bàsiques, recopilació automàtica de respostes i exportació directa a Google Sheets. A més, és accessible des de qualsevol dispositiu i no requereix coneixements tècnics avançats. Com a inconvenients, el disseny és força limitat i la personalització visual és molt bàsica.
- Typeform ofereix una experiència d'usuari molt més dinàmica i atractiva, amb preguntes que apareixen una a una i una navegació tipus conversa. Aquesta interacció pot augmentar l'atenció i la motivació per respondre, especialment entre estudiants. Tanmateix, la versió gratuïta limita tant el nombre de respostes com de preguntes, i algunes funcionalitats avançades (com branques complexes, personalització de marca o exportacions completes) només estan disponibles en plans de pagament.
- Microsoft Forms combina una interfície senzilla amb funcionalitats adequades per a enquestes educatives: suport per a preguntes condicionals, obligatòries, diversos tipus de resposta i exportació directa a Excel. Tot i que l'aspecte visual és menys atractiu que el de Typeform, la seva robustesa i facilitat d'ús el fan molt adequat per a contextos escolars. A més, la seva integració amb l'ecosistema Microsoft representa un avantatge en entorns institucionals que ja treballen amb aquesta tecnologia, com és el cas d'aquest projecte. En consonància amb això, el CERI va indicar que Microsoft Forms és l'eina institucional recomanada per a aquest tipus de tasques, ja que compta amb un acord formal amb la URV i garanteix el compliment dels requisits de protecció de dades. Per aquests motius, es va optar finalment per aquesta plataforma com a base per al desenvolupament i distribució del qüestionari digital.

L'anonimització de les respostes s'aconsegueix mitjançant una primera pregunta específica que recull un número identificador de l'estudiant, generat en el moment de signar el consentiment per a la participació. Aquestes dades identificadores no es guarden mai digitalment, però permeten detectar possibles duplicats en les respostes

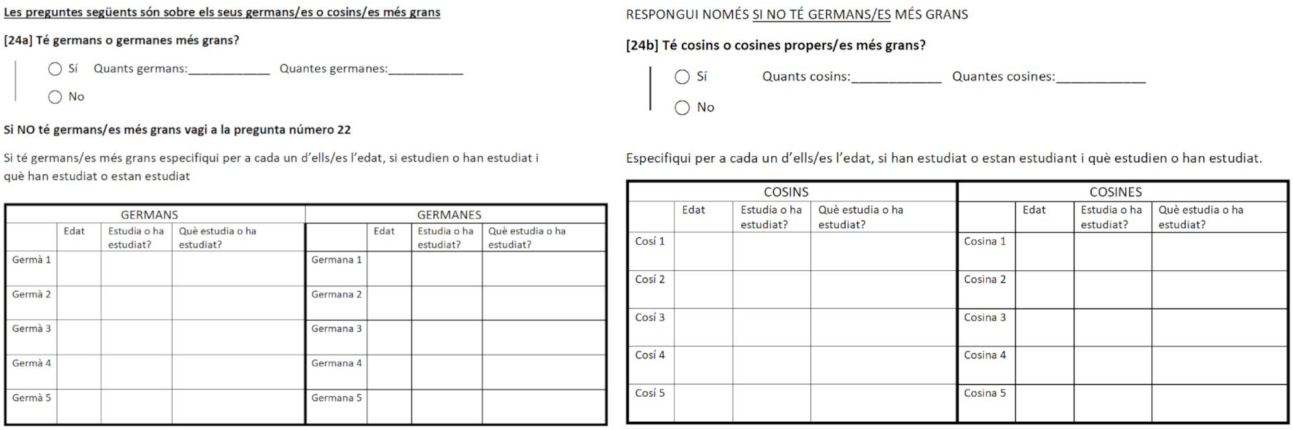
Durant la digitalització del qüestionari original dissenyat en paper (veure Annex 1), l'equip investigador va establir criteris per determinar quines preguntes serien obligatòries i quines opcionals, així com per aplicar restriccions específiques en les respostes. Per exemple, en determinades preguntes es va requerir que les respostes fossin numèriques i dins d'un interval predefinit. No obstant això, Microsoft Forms ofereix capacitats limitades per configurar restriccions avançades, fet que va obligar a simplificar alguns aspectes del disseny respecte a la versió ideal.

També es van identificar i definir quatre tipus de preguntes compatibles amb Microsoft Forms: de selecció única, de selecció múltiple, de resposta lliure i automàtiques

(com l'hora d'inici i finalització de l'enquesta). A cada pregunta, quan s'escau, es van aplicar codis específics com 992 (resposta en blanc), 993 (dubte) o 999 (duplicat), tal com es va preveure durant la fase de preparació de dades.

Pel que fa a la lògica de les preguntes condicionades, es va dissenyar un sistema estructurat basat en quatre configuracions possibles: preguntes obligatòries, opcionals, condicionals obligatòries i condicionals opcionals. Aquest model es va aplicar a preguntes com "Ha escollit vostè el batxillerat que fa?", que activa una subpregunta ("Qui l'ha escollit?") només si l'alumne respon "No". També es van crear preguntes de selecció múltiple condicionada, que despleguen camps de resposta lliure quan es marca una opció concreta.

Un cas especialment complex va ser el disseny de les preguntes relatives a germans i cosins, que en la versió en paper requerien una ramificació específica. Per adaptar aquest flux (veure figura 3) a les possibilitats de Microsoft Forms, es va elaborar un esquema de condicions que assegura que l'alumnat només respongui les preguntes que li pertoquen segons el seu cas particular. Així, per exemple, si declara tenir germans, es mostra la pregunta corresponent; en cas contrari, apareix la referida als cosins. Aquesta estructuració va permetre reproduir amb la màxima fidelitat possible la intenció original de l'enquesta en paper (veure Figura 4).



Les preguntes següents són sobre els seus germans/es o cosins/es més grans

[24a] Té germans o germanes més grans?

○ Sí Quants germans:_____ Quantes germanes:_____

○ No

Si NO té germans/es més grans vagi a la pregunta número 22

Si té germans/es més grans especifiqui per a cada un d'ells/es l'edat, si estudien o han estudiat i què han estudiat o estan estudiat

GERMANS				GERMANES			
	Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?		Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?
Germà 1				Germana 1			
Germà 2				Germana 2			
Germà 3				Germana 3			
Germà 4				Germana 4			
Germà 5				Germana 5			

RESPONGUI NOMÉS SI NO TÉ GERMANS/ES MÉS GRANS

[24b] Té cosins o cosines propers/es més grans?

○ Sí Quants cosins:_____ Quantes cosines:_____

○ No

Especifiqui per a cada un d'ells/es l'edat, si han estudiat o estan estudiant i què estudien o han estudiat.

COSINS				COSINES			
	Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?		Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?
Cosí 1				Cosina 1			
Cosí 2				Cosina 2			
Cosí 3				Cosina 3			
Cosí 4				Cosina 4			
Cosí 5				Cosina 5			

Figura 4. Fragment de l'enquesta en paper sobre germans/es i cosins/es més grans

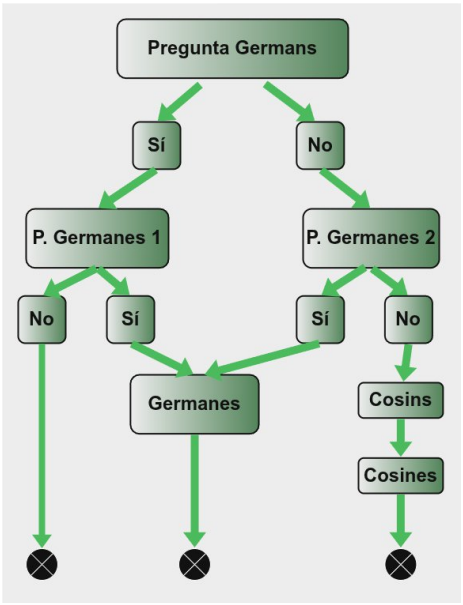


Figura 5. Disseny de la lògica condicional aplicada amb germans/es i cosins/es

Finalment, un cop dissenyat el qüestionari, es va acordar que la Dra. Teresa Corbella crearia una còpia del formulari per assegurar que ella fos la responsable i titular del recull de dades, complint així amb els requisits de responsabilitat i custòdia de la informació.

3.2 Disseny dels l'informe de gràfics

Un cop estructurades les preguntes del qüestionari i definit el seu funcionament dins de Microsoft Forms, es va procedir a dissenyar el conjunt de gràfics que formaran part de l'informe personalitzat per a cada institut. La selecció de les visualitzacions es va basar en les dades recollides a través de l'enquesta, les opcions de resposta disponibles i, especialment, les necessitats i interessos expressats pels centres educatius participants. L'objectiu principal de l'informe gràfic és proporcionar als equips docents una eina clara i visual que faciliti la comprensió de les percepcions, influències i motivacions de l'alumnat en la seva elecció de batxillerat i en les seves decisions futures d'estudi.

3.2.1 Tipologia de gràfics i motius de selecció

L'informe s'inicia amb un gràfic de distribució de freqüències del tipus de batxillerat escollit i del sexe dels estudiants, ja que aquestes dues variables són clau per entendre la composició del grup i serveixen de base per a creuaments amb altres preguntes.

Per a la pregunta sobre els factors que han influït en l'elecció del batxillerat (de resposta múltiple), es va agrupar totes les opcions en quatre categories generals: Positius, Negatius, Tallers i Altres. Aquesta classificació simplifica la interpretació del gràfic i permet als centres identificar ràpidament quins tipus d'influències són més freqüents.

En relació amb la pregunta sobre les persones que han influït en la decisió (també de resposta múltiple), es va aplicar un criteri similar definint les categories següents: Família, Professors o tutors, Altres persones o personatges, i Cap influència. Aquest gràfic ofereix als instituts una visió clara sobre l'origen principal de l'orientació acadèmica dels alumnes.

A més, es va incloure un gràfic que mostra el moment en què l'estudiant va decidir cursar batxillerat, agrupant les respostes en: Abans de l'ESO, entre 1r i 3r d'ESO, a 4t d'ESO abans de veure les notes, i a 4t d'ESO després de veure les notes. Aquesta informació és especialment valuosa per determinar quan els centres haurien de reforçar les tasques d'orientació educativa.

Finalment, una altra pregunta rellevant analitzada és quants companys del mateix grup fan el mateix batxillerat. Les opcions (Cap, Pocs, Meitat, Molts, o Tots) permeten comprendre si l'alumne percep la seva elecció com una opció comuna o minoritària dins del seu entorn proper.

3.2.2 Representació de factors personals i aspiracions acadèmiques

Per a les preguntes de valoració amb escala de Likert (una escala ordinal que mesura el grau d'acord o desacord amb una afirmació, habitualment de 1 a 5), es van identificar tres àmbits principals sobre els quals es generaria un gràfic individual: Interessos i motivació, Autoimatge, i Estrès i pressió. Aquesta classificació facilita que els

centres educatius puguin identificar en quins aspectes emocionals o actitudinals poden sorgir diferències segons el perfil de l'alumnat.

A més, es va decidir representar gràficament les respostes relacionades amb les aspiracions d'estudis futurs, agrupant-les en les categories següents: STEM, Sanitàries, Socials i humanístiques, Serveis, Altres i No ho sap. Aquest gràfic és especialment útil per detectar vocacions primerenques i per comparar-les amb l'itinerari escollit de batxillerat, permetent així avaluar la coherència entre el camí triat i les aspiracions futures.

3.2.3 Format gràfic i visualització

En funció del tipus de dada i de l'objectiu comunicatiu, es va optar per diferents formats gràfics per garantir una visualització clara i efectiva. Concretament, es van seleccionar:

- Gràfics de barres apilats per mostrar proporcions dins d'un mateix grup.
- Gràfics de barres no apilats per facilitar la comparació entre categories independents.
- Gràfics apilats divergents per representar escales com les de Likert (amb valors positius i negatius), facilitant així la lectura de l'equilibri d'opinions.

Durant la fase de desenvolupament, i en funció de les dades obtingudes, es definirà el format més adequat per a cada cas concret.

3.3 Disseny de la preparació i neteja de dades

Un cop dissenyada l'enquesta i establerts els gràfics de sortida, es va planificar el procés de tractament i neteja de les dades amb l'objectiu de garantir que el conjunt fos robust, coherent i adequat per a l'anàlisi posterior. Aquesta fase és fonamental per assegurar la fiabilitat i validesa dels resultats finals. Per aconseguir-ho, es va dissenyar una sèrie de processos automatitzats utilitzant Python i llibreries especialitzades, que permeten una neteja i preparació eficients i reproduïbles.

3.3.1 Tractament del dataset i estructuració bàsica

Per a la manipulació del conjunt de dades, es va optar per fer ús de la llibreria pandas, per la seva potència en el tractament de dades tabulades. En primer lloc, es va decidir normalitzar tant les columnes (preguntes) com les files (respostes). Això implicava transformar tot el text a minúscules, eliminar accents, apòstrofs, caràcters no estàndard i espais innecessaris.

Per aconseguir-ho, es va utilitzar les llibreries unicodedata i re. A les columnes, es va substituir els espais per guions baixos (snake_case), mentre que a les files es va deixar els espais intactes per mantenir la llegibilitat de les respostes, tot i que es tractaven com a text pla.

També es va decidir afegir dues columnes noves al dataset: id_centre, amb valors numèrics del 0 al 12 ordenats alfabèticament segons el nom del centre, i codi_postal, per tenir una referència territorial de les dades. Això permetrà, en fases posteriors, analitzar patrons regionals o comparar respostes entre instituts.

Les columnes buides o innecessàries com correu, nom o llengua, que no aportaven informació rellevant ni eren completes, es van eliminar per mantenir un conjunt net i optimitzat.

3.3.2 Detecció de respostes invàlides

Una part fonamental del disseny de neteja va ser la detecció de respostes que no complien els criteris establerts com a vàlids. Per fer-ho, es van implementar dos mètodes:

- Validació d'edat: qualsevol alumne amb una edat superior a 20 anys es considerava fora del perfil objectiu. Aquestes entrades es van traslladar a un nou dataset específic de respostes invàlides, mantenint-ne la traçabilitat.
- Detecció de llenguatge inadequat: es va crear un fitxer .txt que recull una llista de paraules inapropiades. Mitjançant re, es va analitzar cada fila del dataset com una cadena de text unificada i es van cercar-hi coincidències amb aquest fitxer. En cas de detectar-se alguna paraula inadequada, la resposta també es traslladava al dataset de respostes invàlides

Per cada entrada marcada com a invàlida, es va afegir una columna que especifica el motiu d'eliminació: per edat o per contingut inadequat. Això assegura un registre complet i justificat de totes les exclusions fetes.

3.3.3 Tractament i codificació avançada de respostes especials

Per mantenir una estructura clara i evitar errors durant el processament, es va definir quatre constants que representen valors especials utilitzats durant tot el projecte: CODI_BRANCHING per respostes no mostrades per lògica condicional, CODI_BLANC per respostes deixades buides, CODI_DUBTE per respostes amb indicis d'incertesa i CODI_DUPLICAT per identificadors duplicats.

Aquestes constants estandarditzen la codificació i permeten una automatització més segura i robusta en l'anàlisi.

Una altra fase important va ser identificar aquelles respostes que expressaven dubtes, com "no ho sé", "cap idea", "ho estic pensant", etc. Per fer-ho, es va crear un fitxer .txt amb diferents formes habituals d'expressar incertesa.

Donat que aquestes respostes poden aparèixer amb molta variabilitat, es va decidir utilitzar representació vectorial amb sentence_transformers i la funció util per comparar semànticament cada resposta del dataset amb la llista de frases dubtoses. Quan es detectava una similitud significativa, aquella resposta es marcava amb el CODI_DUBTE.

Per a preguntes obertes com estudis preferits, llengües i països d'origen, es va dissenyar un sistema de recodificació automàtica basat en un fitxer .txt amb format clau:valor. Mitjançant la llibreria thefuzz, concretament el mòdul process, es va comparar cada resposta amb les claus definides. Si la similitud superava un llindar establert, la resposta es recodificava automàticament amb la categoria o codi estandarditzat associat.

TheFuzz és una llibreria de Python que implementa mètodes de fuzzy matching, és a dir, comparació aproximada de cadenes de text. Aquesta tècnica és útil quan les dades poden contenir errors tipogràfics, variants lingüístiques o formes diferents de referir-se a una mateixa entitat.

La llibreria utilitza l'algoritme Levenshtein distance [10], que mesura la distància d'edició entre dues cadenes: quantes insercions, eliminacions o substitucions calen per

transformar una cadena en una altra. TheFuzz retorna un percentatge de similitud, i es poden aplicar funcions com `extractOne()` o `extract()` per trobar les coincidències més properes en una llista predefinida [9].

Això permet transformar entrades com "spain", "tarragona" o "españa" en el codi 724, o associar "mates", "matemáticas" i "matemàtiques" sota la mateixa etiqueta homogènia.

Aquest procés assegura consistència terminològica, facilita la classificació i prepara el dataset per a les fases d'anàlisi i clústering, mantenint l'estructura i semàntica clares.

3.4 Disseny de l'anàlisi amb clústering

Un cop les dades han estat netejades, estructurades i codificades adequadament, el projecte entra en una nova fase centrada en l'aplicació de tècniques de clústering categòric. L'objectiu d'aquesta etapa és identificar agrupacions naturals dins la població estudiada i revelar patrons de comportament o preferències entre els estudiants, tot això sense partir d'una hipòtesi prèvia.

Aquest tipus d'anàlisi s'inscriu dins del camp de la intel·ligència artificial (IA), i més concretament de la mineria de dades. La IA fa referència a tecnologies que permeten a les màquines simular processos humans com l'aprenentatge, la detecció de patrons i la presa de decisions. En aquest context, la mineria de dades és la disciplina encarregada d'extreure coneixement útil a partir de grans volums de dades, sovint no estructurades o no etiquetades, i sense disposar d'una hipòtesi inicial.

Una de les tècniques centrals en aquest àmbit és el clústering, un mètode d'aprenentatge automàtic no supervisat. Tal com defineix IBM, "Clustering is a type of unsupervised learning method where the model is trained using information that is neither classified nor labeled" [11]. Això fa que sigui especialment útil quan es vol evitar introduir biaixos previs i deixar que siguin les pròpies dades les que revelin estructures i agrupacions significatives.

Aplicar tècniques de clústering és especialment pertinent per a aquest estudi, ja que permet detectar grups d'estudiants amb respostes similars. Aquests grups, formats automàticament, poden interpretar-se com a prototipus representatius o perfils d'alumnat, que comparteixen interessos, motivacions, dubtes o condicions socioeducatives. Aquesta informació és especialment valuosa per a millorar els processos d'orientació acadèmica i vocacional, així com per a l'elaboració de polítiques educatives més ajustades a les necessitats reals dels estudiants.

Tanmateix, el clústering amb dades categòriques presenta reptes específics. Tal com s'exposa a l'article "Clustering Categorical Data" [12], no es poden aplicar distàncies euclidianes convencionals i, per tant, cal recórrer a mètodes especialitzats com k-modes, DBSCAN amb distància de Gower o clústering jeràrquic, que estan dissenyats per treballar amb variables qualitatives. En aquest projecte, aquests mètodes han estat seleccionats pel seu bon comportament en contextos on la major part de la informació prové de respostes obertes o tancades no numèriques, com les d'un qüestionari educatiu.

Per tot això, el projecte aplica tècniques de clústering categòric per segmentar la mostra d'alumnes segons criteris com el sexe, l'itinerari acadèmic o les preferències professionals, facilitant així una anàlisi automatitzada, objectiva i significativa de les dades recollides.

3.4.1 Elecció del mètode de clústering

Tenint en compte que totes les dades són categòriques, es va estudiar quines tècniques de clústering són adequades per aquest tipus de variables. Després d'una anàlisi de viabilitat tècnica [12], es va seleccionar tres mètodes:

- KModes (from kmodes.kmodes import KModes): una extensió del conegut KMeans per dades categòriques. Utilitza la moda (valor més freqüent) com a centre del clúster i calcula distàncies per diferència de coincidències. És útil quan es vol determinar un nombre concret de clústers i obtenir un perfil típic per a cada grup.
- DBSCAN (from sklearn.cluster import DBSCAN): un algorisme basat en densitat. Permet detectar clústers de qualsevol forma i també punts que no pertanyen a cap grup (soroll). Amb dades categòriques, s'utilitza conjuntament amb una matriu de distància (com Gower) per identificar densitats.
- Jeràrquic (Hierarchical Clustering) (from scipy.cluster.hierarchy import linkage): aquest mètode construeix un dendograma que mostra les agrupacions successives entre observacions. Permet observar la relació entre els individus sense necessitat de definir el nombre de clústers a priori. Es basa també en una matriu de distàncies entre els objectes (amb distancia de Gower).

A més, per millorar la interpretació dels resultats i validar la qualitat del clústering, es va determinar usar dos indicadors:

- Dendograma (a partir de distàncies de Gower): per representar jeràrquicament les similituds i identificar punts de tall que diferencien els clusters (criteri de Calisnki-Harabaz [25]).
- Silhouette Score: per avaluar la cohesió i separació entre clústers, adaptat a distàncies categòriques.

Mètode	Tipus de dades	Avantatges	Inconvenients	Apte per ús?
KModes	Categòriques	Interpretació clara, ràpid, escalable	Cal especificar el nombre de clústers	Sí
DBSCAN	Categòriques	Detecta soroll, no cal K fix	Paràmetres delicats, sensibilitat a densitat	Sí
Jeràrquic (linkage)	Categòriques	Visualització clara amb dendograma	Cost computacional amb molts individus	Sí
KMeans	Numèriques	Molt conegut, escalable	No apte per categòriques sense codificació	No
Gaussian Mixture model	Numèriques contínues	Flexible, clústering probabilístic amb forma el·líptica	Assumeix distribucions normals; no apte per dades categòriques	No

Taula 2. Anàlisi comparativa de tècniques de clústering aplicables a dades categòriques

Aquest disseny permet aplicar el clústering de manera metodològicament rigorosa, mantenint la compatibilitat amb les dades disponibles i els objectius del projecte, i afavorint la generació de perfils útils per a l'anàlisi de vocacions segons gènere, batxillerat i altres factors clau.

3.4.2 Consideracions metodològiques i criteris de disseny

Tot i que altres mètodes com KMeans són molt coneguts i potents, no són aplicables directament amb dades purament categòriques, ja que assumeixen un espai vectorial continu. Per fer-los servir caldria aplicar un One-Hot Encoding, que converteix les categories en columnes binàries (una per cada valor possible). Aquest procés incrementaria exponencialment el nombre de columnes, fent el model intractable o molt menys eficient.

Per aquesta raó, també es descarten mètriques com Davies-Bouldin Index, que depenen de la distància euclidiana i, per tant, no són aplicables a dades no numèriques.

Perquè l'anàlisi sigui fiable i interpretable, es va establir que el nombre de columnes seleccionades per fer el clústering hauria d'estar entre 5 i 30, i que les categories per columna haurien d'estar entre 2 i 6. D'aquesta manera, es redueix el soroll, es millora la coherència dels resultats i s'evita l'excés de dimensionalitat.

4 Desenvolupament

Un cop definits els requeriments i establert el disseny detallat de les diferents fases del projecte, es va passar a la implementació de totes les funcionalitats previstes. Aquesta etapa de desenvolupament va consistir a convertir les propostes inicials en eines operatives reals, assegurant que cada funcionalitat respongués als objectius definits i complís els criteris establerts: eficiència, fiabilitat, protecció de dades i facilitat d'ús.

Les seccions següents descriuen el desenvolupament corresponent a cadascuna de les quatre fases del projecte: la creació del formulari, la generació de l'informe gràfic, el tractament i neteja de dades, i l'anàlisi mitjançant tècniques de clústering.

4.1 Desenvolupament del qüestionari digital

4.1.1 Creació de la plantilla a Microsoft Forms

El desenvolupament del qüestionari va començar amb la creació d'una plantilla a la plataforma institucional Microsoft Forms, utilitzant un compte universitari. Tot i algunes limitacions pel que fa a la configuració, aquesta eina complia els requisits establerts en la fase de disseny, especialment pel que fa a la seguretat, l'anonimat i l'accessibilitat.

L'estil del formulari es va personalitzar amb una imatge institucional de fons i una paleta de colors en tons marrons clars, escollida per garantir una bona llegibilitat (veure Figura 6). La plantilla es va organitzar en onze seccions temàtiques: introducció, primera part, segona part, tercera part, preguntes sobre el pare, la mare, germans, germanes, cosins, cosines i, finalment, una secció d'agraïments.

Figura 6. Estil del formulàri amb imatge base de la universitat de fons

4.1.2 Disseny i configuració de les preguntes

Un cop establerta l'estructura, es van implementar les preguntes de manera seqüencial. El formulari final contenia 110 preguntes, tot i que el nombre real de preguntes visibles per a cada estudiant oscil·lava entre 32 i 53, depenent dels salts condicionals aplicats.

Cada pregunta es va configurar amb el seu tipus corresponent (opció múltiple, resposta oberta, escala de Likert...), i es van afegir enunciats i subenunciats per clarificar-ne el contingut. En aquelles preguntes on es podien introduir dades numèriques, es van establir restriccions com intervals vàlids o valors mínims/màxims.

També es va tenir en compte la limitació de Microsoft Forms pel que fa a l'ordenació d'opcions. Per exemple, en preguntes d'opció múltiple on s'havia d'incloure l'opció "Cap" o "Altres", aquestes no es podien barrejar lliurement amb la resta. Per tant, es va optar per una ordenació fixa.

4.1.3 Gestió dels salts condicionals

Una part clau del disseny va ser la implementació dels salts condicionals entre preguntes. Aquesta funcionalitat va permetre que cada alumne respongués només a les preguntes que li corresponien segons les seves respostes prèvies, evitant redundàncies i millorant l'experiència d'usuari.

Per exemple, a les preguntes sobre familiars (germans/es i cosins/es), es va aplicar una lògica de salt per evitar que els alumnes haguessin de respondre a preguntes que no els eren aplicables. Aquesta estructuració es va definir prèviament en un esquema que recollia tota la lògica de navegació i que assegurava la fidelitat respecte al formulari original en paper.

4.1.4 Configuració d'accessibilitat i propietat de dades

Un cop finalitzat el formulari, es va configurar perquè pogués ser accessible mitjançant un enllaç directe, sense necessitat de registre. Això permetia la participació de l'alumnat des de qualsevol dispositiu, complint el requisit d'accessibilitat multiplataforma.

A més, es va habilitar l'opció que els estudiants poguessin descarregar les seves respostes en finalitzar, si ho desitjaven. Un cop testejada la plantilla, es va enviar un duplicat del formulari a la Dra. Teresa Corbella, qui havia de ser la responsable final de les dades i la seva custòdia. D'aquesta manera, cada centre educatiu podia rebre una còpia personalitzada del formulari, les respostes del qual es guardarien automàticament en fitxers Excel separats.

4.2 Desenvolupament de l'informe de gràfics

Un cop finalitzada la neteja de les dades (descrita en l'apartat següent), es va procedir a desenvolupar el sistema de generació de l'informe de gràfics. L'objectiu era automatitzar la creació de visualitzacions útils per als instituts, tot assegurant la flexibilitat i reutilització del codi amb diferents conjunts de dades estructurades de forma similar.

Per implementar tècnicament la generació de l'informe i automatitzar el procés d'anàlisi, es va escollir Python per la seva versatilitat i per l'ampli ventall de llibreries disponibles per al treball amb dades i visualització.

Les llibreries triades han sigut:

- matplotlib.pyplot per crear els gràfics [13].
- matplotlib.patches i colormaps per personalitzar l'aspecte visual.

- numpy per fer càlculs i, concretament, per gestionar el recompte de respostes múltiples dins una mateixa categoria. Aquesta funcionalitat era imprescindible per complir el requeriment segons el qual cada alumne només podia comptar una vegada per categoria en preguntes amb selecció múltiple [14].
- rapidfuzz.fuzz per a la comparació de respostes obertes i la seva assignació a la categoria més similar ja existent. Aquest punt responia al requeriment que indicava que les respostes obertes (especialment les de l'opció "altres") s'havien d'intentar classificar dins de categories prèviament definides, i només crear-ne de noves si era estrictament necessari [9].
- python-docx (Document) per generar automàticament un document Word amb tots els gràfics incorporats [15] .

Aquest conjunt d'eines permet crear informes visuals, complets i adaptats a cada centre, garantint una presentació clara i útil dels resultats.

4.2.1 Preparació prèvia i funcions auxiliars

Abans de començar a generar els gràfics finals, es va implementar diverses funcions auxiliars per facilitar i optimitzar el procés:

- Funció `afegir_grafic_al_word(titol)`: Aquesta funció s'encarrega de guardar el gràfic generat amb matplotlib com a imatge temporal, inserir-lo dins un document Word amb el títol especificat i eliminar posteriorment la imatge temporal per evitar conflictes. Això permet construir informes visuals automàtics de manera eficient.
- Funció `etiquetar_barres_percentatge(ax, totals_fila)`: Aquesta funció afegeix, de forma clara i ordenada, els percentatges sobre les barres dels gràfics apilats. Gestiona els solapaments mitjançant un sistema de desplaçament vertical per assegurar la llegibilitat.
- Funció `generar_colors(n)`: Genera una paleta de colors diferenciada i visualment clara, utilitzant esquemes de color com Set3 i tab20, segons el nombre de categories a representar.

Aquest conjunt d'eines auxiliars va permetre crear un marc modular, mantenint el codi net i reutilitzable.

4.2.2 Gràfics apilats i per categories

Per representar preguntes amb múltiples opcions (com les influències o els factors de decisió), es va desenvolupar la funció:

- `grafic_classificacio(df_detall, ordre, noms, variable_grup, titol)`: Genera un gràfic de barres apilat que agrupa les respostes per categoria i segmenta per grups com el sexe o batxillerat. Aplica filtres per eliminar categories amb menys de tres respostes, assegurant l'anonimat, i incorpora les funcions auxiliars per millorar la visualització.

Una versió millorada d'aquesta funció és:

- `grafic_apilat_categoria_batxillerat_total(df_detall, ordre, noms_cat, variable_grup, titol)`: Aquesta funció extén la funcionalitat

anterior permetent creuar categories i tipus de batxillerat. És especialment útil per detectar patrons combinats, com la presència femenina en STEM dins del batxillerat tecnològic. El gràfic resultants mostra dues dimensions d'anàlisi en un únic espai visual.

4.2.3 Gràfics no apilats i divergents

Per a dades que requereixen comparació directa entre valors sense superposició, es va implementar:

- `grafic_classificacio_no_apilat((df_detall, ordre, noms, variable_grup, titol, stacked=False)`: Aquesta funció mostra les barres de manera separada i no apilada. Manté totes les funcionalitats de filtratge, etiquetatge i color, però presenta les dades de forma més directa, útil per a preguntes amb poques opcions excloents.

Per a preguntes de tipus escala Likert (acord/desacord), es va fer ús de:

- `analitza_escala_likert(df, temes_likert, opcions_valides, variables_grup=None)`: Aquesta funció analitza respostes de temes agrupats (com motivació, autoimatge, estrès...), elimina respostes com "no sap" i genera gràfics de barres horitzontals divergents, mostrant les respostes positives a la dreta i les negatives a l'esquerra. És especialment útil per avaluar el clima emocional i actitudinal dels estudiants.

4.2.4 Classificació prèvia de respostes obertes

Una part fonamental per generar gràfics vàlids era assegurar una correcta classificació de respostes obertes. Per això es va crear:

- `classificar_respostes(df, columna_respostes, categories, variables_extres=None.)`: Aquesta funció analitza respostes lliures (com les opcions "altres") i intenta assignar-les a categories definides mitjançant patrons textuals. Si no encaixen, es classifiquen com a "altres". Aquesta funció també gestiona les agrupacions per variables com sexe o centre, filtrant les agrupacions amb menys de tres respostes. Les dades classificades s'integren automàticament en els gràfics corresponents, garantint així que cap resposta rellevant es perdi.

Un cop implementades totes aquestes funcions, el procediment per generar cada gràfic s'ha simplificat notablement. Només cal definir la columna de respostes que es vol analitzar, establir les categories corresponents i cridar la funció classificar_respostes seguida de la funció de visualització adequada, passant els paràmetres concrets. D'aquesta manera, es poden analitzar automàticament múltiples preguntes amb una estructura comuna, mantenint el control sobre les respostes no classificades, que es recullen a part per garantir el seguiment complet de la informació, tal com s'havia definit als requisits inicials del disseny.

4.2.5 Generació d'una interfície gràfica per a l'ús dels informes de gràfics

Amb l'objectiu de facilitar l'ús de les funcions de generació d'informes gràfics per part de l'equip investigador, especialment en el cas de persones sense coneixements avançats de programació, es va desenvolupar una interfície gràfica d'usuari mitjançant la llibreria Streamlit (Veure Figura 7). Aquesta interfície permet carregar un arxiu Excel amb les respostes, introduir el nom i l'identificador del centre, i executar automàticament tot el procés de generació de gràfics i creació del document Word amb els resultats.

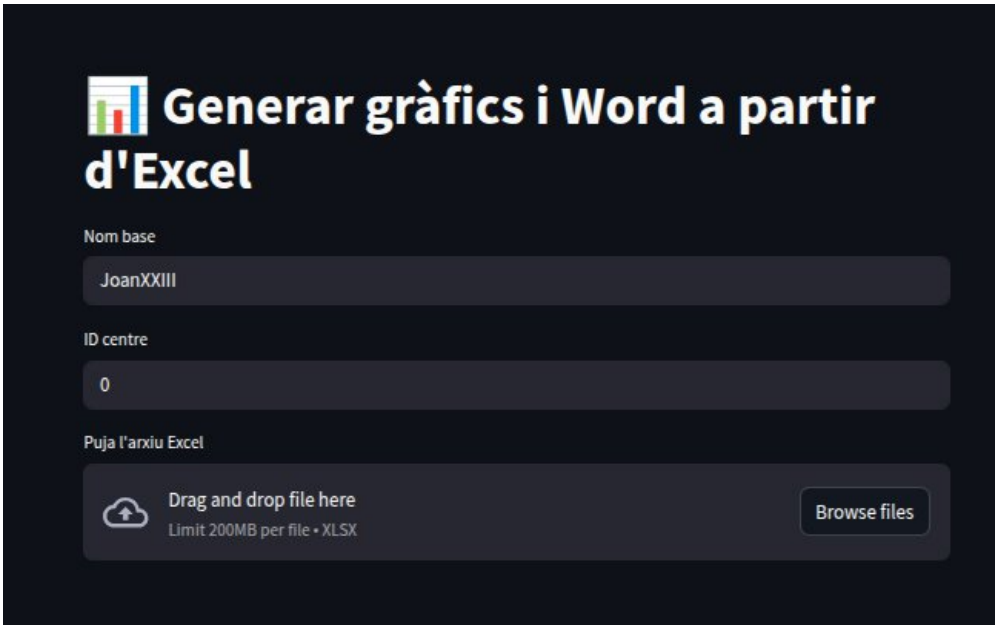


Figura 7. Interfície gràfica per generar informes a partir d'un fitxer Excel

Per aconseguir-ho, es va combinat diversos elements:

- Streamlit: per generar una interfície web senzilla i interactiva [16].
- nbconvert [17] i nbformat [18]: per executar de manera programada un notebook Jupyter (Analyze_ui.ipynb) que conté tot el codi per analitzar les dades i generar l'informe.
- webview i subprocessos: per gestionar la visualització del projecte com una finestra d'aplicació d'escriptori [19].

La funcionalitat es distribueix en dues parts. D'una banda, el codi principal escriu un fitxer de configuració (config.json) amb els paràmetres d'execució (com el nom de l'informe, el fitxer Excel o l'identificador del centre). D'altra banda, el notebook és executat automàticament i llegeix aquest fitxer per generar un document Word amb els gràfics personalitzats.

El procés es gestiona mitjançant fils paral·lels per garantir que es poden iniciar tant el servidor Streamlit com l'execució del notebook en temps real. Això permet mantenir una experiència d'usuari fluïda i intuïtiva, complint així el requeriment d'oferir una eina funcional, reutilitzable i fàcil d'usar per generar els informes visuals a partir de les dades recollides.

Amb aquesta eina, només cal introduir les dades d'un centre i pujar el fitxer Excel corresponent per obtenir automàticament un document de resultats, fet que agilitza enormement el procés d'anàlisi i el posa a l'abast de qualsevol membre de l'equip.

4.3 Preparació i neteja de dades

La fase de preparació de dades va ser clau per assegurar que el conjunt fos robust, coherent i adequat per a l'anàlisi posterior. Aquest procés es va estructurar en dues grans etapes: d'una banda, la neteja i normalització inicial de les dades recollides a través del qüestionari; de l'altra, la creació d'un nou conjunt categòric pensat específicament per a l'aplicació de tècniques de clústering.

4.3.1 Neteja inicial i normalització del conjunt de dades

Per començar, es van definir unes constants bàsiques com original_filename, codi_postal, id_centre (de 0 a 12 segons el nom del centre) i folder_name, que permetien identificar correctament els arxius i organitzar els resultats.

Es va fer ús de Python i la llibreria pandas [20] per carregar el fitxer Excel, així com de diverses funcions per dur a terme la normalització del text. La funció `normalize_ascii(text)` eliminava accents i caràcters diacrítics mitjançant la llibreria unicodedata, aplicant una descomposició en forma NFKD. Aquest procés assegurava la coherència entre respostes semblants.

Les funcions `normalize_column_name(col_name)` i `normalize_text_cell(val)` permetien normalitzar tant els noms de les columnes com les entrades de text. Es van utilitzar mètodes com `.lower()`, `.sub()` i `.strip()` per transformar el text a minúscules, eliminar caràcters no desitjats i ajustar els espais. Les columnes es convertien a snake_case i els valors es mantenien amb espais i puntuació rellevant.

També es van aplicar instruccions com `df.insert(...)` per afegir l'ID del centre i `df.drop(...)` per eliminar columnes innecessàries com correu, nom i llengua.

La funció `load_fitxer(path)` va servir per carregar llistes auxiliars en format .txt com la de paraulotes, països o estudis. Cada línia es llegia i normalitzava per facilitar les cerques posteriors.

Per detectar respostes invàlides, es va implementar la funció `detect_invalid_response(...)`, que analitzava cada fila per comprovar si contenia una edat superior a 20 anys o paraulotes detectades. En aquest cas, es movien les respostes a un nou DataFrame amb una columna de motius d'eliminació.

4.3.2 Classificació de respostes i detecció específica

Per les preguntes obertes com assignatures, estudis, llengües i països, es va aplicar un procediment comú. Primer, es carregava un fitxer .txt amb parelles clau-valor que definien cada categoria. A continuació, es feia ús de la llibreria thefuzz, concretament del mètode process.extract, per comparar cada entrada amb les claus del fitxer. Si la similitud entre l'entrada i una clau superava un llindar predefinit, l'entrada es classificava automàticament amb el valor associat. En cas que no es trobés cap coincidència acceptable, s'aplicava la constant CODI_BLANC o CODI_DUBTE segons el context i la naturalesa de la pregunta.

La classificació de respostes amb dubtes va requerir un tractament més específic. Tot i que també es partia d'un fitxer .txt amb frases típiques de dubte, les entrades a

analitzar sovint consistien en frases més llargues i complexes. Per tant, es va optar per un enfocament basat en representacions semàntiques amb la llibreria sentence-transformers.

Primer, es carregaven les frases de referència i es generaven els seus embeddings amb un model transformador preentrenat:

```
    embeddings_dubtes = model.encode(frases_dubtoses,
convert_to_tensor=True)
```

Després, per a cada resposta a analitzar, es calculava la seva similitud cosinus amb tots els embeddings de referència:

```
    if text.strip() and util.cos_sim(model.encode(text,
convert_to_tensor=True), embeddings_dubtes).max().item() >= 0.60:
        # Es considera una resposta dubtosa
```

Si la màxima similitud era igual o superior a 0.60, es classificava la resposta com a dubtosa i se li assignava el CODI_DUBTE. En cas contrari, el valor original es mantenia.

Aquest enfocament va permetre identificar respostes amb intencions dubtoses encara que no coincidissin literalment amb cap de les frases del fitxer, fent que el sistema de detecció fos molt més robust i fiable.

4.3.3 Construcció del DataFrame per al clústering

Amb el conjunt de dades normalitzat, es va iniciar la construcció del DataFrame definitiu per al clústering. Per a cada columna, es va revisar si ja estava categòricament preparada mitjançant `df[col].unique()`.

Per a la columna de factors de decisió, es va dissenyar la funció `classificar_respostes_en_columna_nova(...)`, que, segons el nombre de categories detectades en una entrada múltiple, afegia una nova entrada al DataFrame amb la combinació de categories trobades. Això evitava la pèrdua d'informació important, assignant "X i Y" si es trobaven dues categories, o "general" si hi havia més de dues.

En el cas de la pregunta sobre persones influents en la decisió, es va aplicar la funció `classificar_respostes_per_categories(...)`, que creava una columna per a cada categoria (família, amics, professors, etc.). Es diferenciava entre respostes amb una sola entrada i combinacions múltiples. Així, s'assignava "altres" si no s'identificava cap categoria coneguda i es manté una categoria general quan hi ha respostes amb moltes opcions.

Per la resta de columnes, es va aplicar la funció `classificar_resposta_unica(...)`, que afegeix una nova columna al DataFrame amb la categoria corresponent, o "altres" si no encaixa amb cap patró.

Aquest procés es va repetir per totes les columnes escollides, fins a obtenir un conjunt final de dades completament categoritzat i llest per ser analitzat mitjançant clústering categòric.

4.4 Desenvolupament del clústering categòric

Un cop construït el conjunt de dades categòric a partir de les respostes a l'enquesta, es va procedir a aplicar diferents tècniques de clústering per identificar perfils d'estudiants amb característiques similars. Seguint el disseny establert, es van implementar tres mètodes diferents: K-Modes [21], Jeràrquic [22] i DBSCAN [23] amb distància de Gower [24].

4.4.1 Preparació prèvia del conjunt de dades

Per garantir una aplicació correcta dels mètodes de clústering, es va eliminar la variable sexe del conjunt abans de l'anàlisi, ja que es volia emprar únicament com a variable d'anàlisi posterior. Aquesta es va afegir de nou un cop finalitzat el procés de classificació.

4.4.2 Mètode K-Modes

El mètode K-Modes és una extensió del K-Means pensada exclusivament per a dades categoriques. És l'únic dels tres mètodes aplicats que permet treballar directament sobre variables categòriques. En comptes d'utilitzar la mitjana (com fa K-Means), utilitza la moda (el valor més freqüent) per definir el centre dels clústers. A més, mesura la distància entre punts mitjançant el nombre de discrepàncies (Hamming distance) entre les categories [27].

L'algoritme segueix aquests passos bàsics:

1. Inicialització dels modes:

   Es seleccionen aleatòriament o amb alguna estratègia heurística els valors inicials dels centres dels clústers, anomenats modes. A diferència del K-Means, no són valors mitjans sinó les modes (valors més freqüents) per a cada atribut categòric.

2. Assignació de punts als clústers:

   Cada punt de dades es compara amb els modes existents utilitzant una distància de coincidència (comptar quants atributs no coincideixen). El punt s'assigna al clúster amb el mode més proper (és a dir, amb menys discrepàncies).

3. Actualització dels modes:

   Un cop tots els punts han estat assignats, es recalculen els modes de cada clúster. Per a cada atribut dins d'un clúster, es tria el valor més freqüent (la moda) per definir el nou centre.

4. Repetició fins a convergència:

   Els passos 2 i 3 es repeteixen fins que els modes ja no canvien significativament o les assignacions de punts es mantenen estables. Això vol dir que s'ha arribat a una solució estable.

Amb les dades carregades es va aplicar el model K-Modes de la llibreria kmodes.kmodes, amb la següent configuració:

```
km = KModes(n_clusters=n_clusters, init="Huang", n_init=5,
verbose=0)
clusters = km.fit_predict(df_cat)
```

Aquest model calcula els centroides categòrics i assigna cada mostra al clúster més proper.

4.4.3 Mètodes amb distància de Gower: Jeràrquic i DBSCAN

Per aplicar tècniques de clústering sobre dades categòriques o mixtes (combinació de variables categòriques i numèriques), cal utilitzar una mesura de dissimilitud que permeti comparar correctament la diferència entre observacions. En aquest projecte s'ha emprat la distància de Gower, una eina especialment útil per treballar amb aquest tipus de dades.

La distància de Gower és una mesura de dissimilitud que permet calcular com diferents són dues observacions entre si quan aquestes contenen variables de diferent naturalesa (numèriques, categòriques, binàries, etc.). A diferència d'altres distàncies com l'euclidiana, que només són aplicables a variables contínues, Gower permet combinar diferents tipus de variables en un mateix càlcul.

Concretament, el mètode calcula la dissimilitud entre dues observacions comparant cada variable de forma individual, normalitzant el resultat, i fent una mitjana ponderada. Per a les variables numèriques, utilitza la diferència absoluta dividida pel rang. Per a les categòriques, assigna 0 si els valors coincideixen i 1 si són diferents.

El resultat final és una mitjana ponderada de totes les dissimilituds individuals, cosa que fa aquesta mesura especialment adequada per al nostre cas, en què moltes preguntes tenen respostes tancades o codificades. [26].

Clústering jeràrquic amb distància de Gower

El clustering jeràrquic és una tècnica d'agrupament no supervisat que construeix una estructura jeràrquica de clústers, organitzant les dades en diferents nivells d'agrupació. A diferència de mètodes com el K-Modes, no requereix especificar prèviament el nombre de clústers, i permet explorar les relacions entre punts a diferents escales. Aquest mètode pot aplicar-se tant a dades numèriques com categòriques, sempre que es defineixi una matriu de distàncies adequada [28].

L'algoritme segueix aquestes etapes:

1. Inicialització dels clústers:

   Cada punt de dades es considera inicialment com un clúster independent. Si es tenen n observacions, es comença amb n clústers individuals.

2. Càlcul de distàncies entre clústers:

   Es calcula la distància entre totes les parelles de clústers segons un criteri d'enllaç. El mètode d'enllaç pot ser:

   a. Enllaç senzill: es considera la distància mínima entre dos punts dels clústers.

b. Enllaç complet: es considera la distància màxima.
c. Enllaç mitjà: es fa la mitjana de totes les distàncies entre parells de punts (mètode utilitzat en aquest treball).
d. Ward: minimitza la variància interna dels clústers (aplicable principalment a dades numèriques).

3. Fusió iterativa:

   Es fusionen els dos clústers més propers segons el mètode d'enllaç seleccionat. Un cop fusionats, la matriu de distàncies s'actualitza per reflectir aquest canvi:

   - S' elimina la fila i columna corresponents als dos clústers originals.
   - Es calcula la distància del nou clúster fusionat respecte a la resta de clústers encara actius, segons el criteri d'enllaç triat.

   Per exemple, si s'utilitza l'enllaç mitjà, la nova distància es calcula com la mitjana de les distàncies entre tots els punts del nou clúster i els punts dels altres clústers.

   Aquest procés de fusió i actualització es repeteix de manera iterativa fins que totes les observacions estan agrupades en un únic clúster final o fins que s'atura en un nombre de clústers determinat.

4. Construcció del dendrograma:

   El procés de fusió es representa visualment mitjançant un dendrograma, que mostra com es van agrupant les observacions a mesura que augmenta la distància entre clústers.

5. Tall del dendrograma:

   Per obtenir un nombre concret de grups, es talla el dendrograma a una certa altura. Aquest tall permet seleccionar el nombre de clústers que millor s'adapta a l'anàlisi.

DBSCAN amb distància de Gower

El mètode DBSCAN (Density-Based Spatial Clustering of Applications with Noise) és una altra tècnica d'agrupament no supervisat que identifica clústers basats en la densitat de punts dins de l'espai de distàncies. Tal com passa amb el clústering jeràrquic, no cal especificar el nombre de clústers per avançat, i és especialment útil per detectar clústers de forma arbitrària i per identificar punts aïllats (soroll). Pot aplicar-se a dades categòriques si es fa servir una matriu de distàncies adequada, com la distància de Gower [29].

L'algoritme segueix aquests passos generals:

1. Càlcul de la distància entre punts:

   Primer, es construeix una matriu de distàncies entre tots els punts del conjunt de dades. En aquest cas, s'utilitza la distància de Gower, que permet calcular distàncies mixtes entre variables categòriques i numèriques.

2. Identificació de punts densament connectats:

Per a cada punt, es compta quants veïns té dins d'un radi determinat (eps). Si un punt té almenys min_samples veïns dins d'aquest radi, es considera un nucli de clúster.

3. Expansió del clúster:

   Els punts que es troben dins del radi eps d'un nucli s'assignen al mateix clúster. El procés es repeteix recursivament per als nous punts que també compleixin la condició de densitat, fins a completar el clúster.

4. Identificació de soroll:

   Els punts que no tenen suficients veïns per formar part d'un clúster es consideren soroll i reben l'etiqueta -1. Aquests punts no s'assignen a cap grup i es poden analitzar com a anomalies o casos especials.

5. Resultat i avaluació:

   El resultat són clústers de forma flexible, adaptats a la densitat de les dades. Per avaluar la qualitat de l'agrupament, es pot calcular el Silhouette Score, tenint en compte només els punts que no han estat classificats com a soroll (cluster ≠ -1).

Aquest mètode és especialment útil quan s'espera que els clústers tinguin formes irregulars, o quan es vol detectar puntes atípiques (outliers) de manera natural. No obstant això, la seva sensibilitat als valors de eps i min_samples requereix una certa calibració.

Per a l'aplicació del clústering jeràrquic i DBSCAN, es va generar una matriu de distància mitjançant la funció `gower_matrix(df_gower)` de la llibreria gower, adaptada específicament per a dades categòriques.

- Clústering jeràrquic:
  - Es va utilitzar el mètode d'enllaç mitjà `(method="average")` mitjançant la funció `linkage()` de scipy.cluster.hierarchy.
  - Els clústers es van formar amb la funció `fcluster(...)` a partir de la linkage matrix.
- DBSCAN:
  - El model es va definir amb `DBSCAN(eps=0.4, min_samples=5, metric="precomputed")` de la llibreria sklearn.cluster.
  - L'agrupament es va aplicar directament sobre la matriu de distàncies de Gower calculada prèviament.

4.4.4 Avaluació dels resultats i generació d'informes

Per valorar la qualitat dels agrupaments obtinguts amb els models basats en distància, es va utilitzar la mètrica Silhouette Score, adaptada a distàncies precomputades:

```
silhouette = silhouette_score(gower_dist, clusters,
metric="precomputed")
```

Aquest índex permet avaluar la cohesió i separació entre els diferents clústers.

Un cop assignat un clúster a cada entrada del DataFrame, es va executar un conjunt de funcions d'anàlisi per entendre millor la composició de cada grup:

- Per a cada clúster es va calcular la moda i el segon valor més habitual per a cada variable, així com el seu percentatge.

- També es va generar una distribució completa de respostes per clúster, amb l'objectiu d'analitzar la seva variabilitat interna.
- Es va calcular la distribució del gènere (homes, dones, altres) dins de cada clúster, tant en nombre absolut com en percentatge.

Tota aquesta informació es va recollir automàticament en un fitxer Excel amb múltiples fulls, que inclou:

- Les dades classificades amb el número de clúster.
- El perfil típic de cada clúster.
- La distribució detallada de les respostes.
- La distribució del sexe dins de cada clúster.

Aquest procés ha permès obtenir agrupacions interpretables i útils per a la comprensió dels patrons de decisió acadèmica dels estudiants, complint amb els objectius d'identificació de perfils plantejats inicialment.

5 Experimentació

En aquesta secció es detallen les proves realitzades per validar el correcte funcionament de les eines desenvolupades, així com l'aplicació real del qüestionari als centres educatius participants. L'objectiu principal d'aquesta fase era verificar la robustesa de la plataforma, la fiabilitat del sistema de recollida de dades i la idoneïtat de les funcionalitats en el context real d'administració de l'enquesta en entorns escolars.

5.1 Validació i execució del qüestionari

5.1.1 Proves inicials de funcionalitat

Abans de distribuir el qüestionari als instituts, es va dur a terme una fase de proves simulades per assegurar el correcte funcionament de la plataforma Microsoft Forms. Es van generar respostes fictícies, com si fossin alumnes reals, per comprovar diversos aspectes clau del procés:

ID	Cas de prova	Descripció	Resultat esperat	Validació
1	Respondre preguntes obligatòries	Comprovar si les preguntes obligatòries bloquegen l'avanç si no es responen	No es pot avançar si no es repon	OK
2	Restricció per edat	Verificar que no es pot continuar si s'introdueix una edat fora del rang	Es bloqueja si l'edat és incorrecta	OK
3	Salts condicionals	Comprovar que les preguntes es mostren o amaguen segons la resposta anterior	Les preguntes s'adapten a la resposta prèvia	OK
4	Exportació a Excel	Revisar si el fitxer Excel mostra correctament columnes i respostes múltiples	Les dades s'exporten correctament	OK
5	Identificadors i numeració	Assegurar que no es repeteixen identificadors i que la numeració és coherent	Identificadors únics i numeració correcta	OK

Taula 3. Validació del comportament del formulari i les seves funcionalitats

Figura 8. Pantalla inicial del qüestionari – Correcte ús restringit

Aquestes proves van permetre validar la configuració del qüestionari i garantir que la implementació complia els requisits tècnics i metodològics establerts en la fase de disseny.

5.1.2 Implementació i supervisió de la recollida de dades

Un cop verificat el qüestionari, es va procedir a preparar la seva distribució definitiva. Es va crear una còpia independent del formulari per a cada institut, de manera que la persona responsable en pogués gestionar i recollir les respostes per separat. Aquesta estratègia garantí una organització clara de les dades i permeté etiquetar-les correctament des del moment de la seva recollida.

Per iniciar la col·laboració amb els centres educatius, es va enviar una carta de presentació, un document informatiu sobre l'estudi i un full de consentiment informat que havia de ser signat per l'alumnat participant. Aquest pas assegurava que la participació fos voluntària i d'acord amb la normativa vigent en matèria de protecció de dades.

La Dra. Teresa Corbella va visitar personalment cadascun dels 13 instituts participants, repartits entre diverses localitats del Camp de Tarragona. Durant aquestes visites, va presentar els objectius de l'estudi i va oferir suport directe a l'alumnat durant la realització del qüestionari.

En 11 dels centres, l'enquesta es va dur a terme en format digital a través de Microsoft Forms. En els dos restants, es va administrar en paper, i posteriorment les respostes es van transcriure manualment a la versió digital per tal d'unificar el conjunt de dades. En total es van recollir 695 qüestionaris, aconseguint una base de dades sòlida i representativa. En aquells centres amb un nombre elevat d'estudiants, es va optar per aplicar l'enquesta en petits grups successius, facilitant-ne així la gestió i garantint una millor atenció individual. Cal destacar que no es van registrar incidències tècniques durant les sessions digitals.

Indicador	Valor
Total de centres participants	13
Centres amb enquesta digital	11
Centres amb enquesta en paper	2
Total de qüestionaris recollits	**695**
Àrea geogràfica de recollida de dades	Camp de Tarragona

Taula 4. Resum de la participació dels centres i dades recollides

Les enquestes completades en paper es van digitalitzar manualment seguint l'estructura del formulari original. Tot i que aquest procés es va realitzar amb cura, es van detectar algunes respostes incorrectes o incompletes, especialment en preguntes condicionals o amb restriccions, ja que el format paper no aplica validacions automàtiques.

Les respostes recollides es van revisar i corregir manualment, aplicant els criteris definits durant la fase de disseny per garantir que només s'incloguessin les respostes completes i consistents. Aquest procés va permetre unificar i depurar el conjunt de dades, deixant-lo preparat per a les etapes posteriors d'anàlisi i visualització.

5.2 Generació i validació de l'informe de gràfics

5.2.1 Aplicació del sistema de visualització

Amb el sistema de neteja i classificació completat i el conjunt de dades preparat, es va procedir a validar la funcionalitat de les eines de visualització gràfica desenvolupades durant la fase anterior. Per fer-ho, es van analitzar les dades obtingudes dels 13 instituts participants aplicant les funcions de generació de gràfics dissenyades en Python.

L'anàlisi es va dur a terme institut per institut, generant tots els gràfics definits per a cadascun i comprovant el seu correcte funcionament tant a nivell visual com a nivell de càlculs.

5.2.2 Validació de l'aspecte visual

Per a cada gràfic, es va revisar detalladament que la representació visual complís els següents criteris:

ID	Cas de prova	Descripció	Resultat esperat	Validació
1	Representació de barres	Comprovar l'alçada i alineació correcta de les barres	Les barres es mostren proporcionalment i alineades correctament	OK
2	Colors consistents	Verificar que els colors són diferenciats i coherents entre gràfics	Els colors són clars, no es repeteixen i es mantenen en gràfics similars	OK
3	Percentatges i recomptes visibles	Assegurar que les xifres són correctes i ben posicionades sobre les barres	Els valors numèrics apareixen centrats i amb el percentatge correcte	OK
4	Llegenda coherent	Comprovar que la llegenda coincideix amb els colors i categories mostrades	La llegenda reflecteix correctament la codificació visual del gràfic	OK
5	Títols dels eixos	Revisar que els eixos tinguin títols clars i adequats	Els eixos apareixen amb títols descriptius de les variables representades	OK

Taula 5. Casos de prova i validació de la visualització de gràfics

Aquest procés es va repetir per cada tipus de gràfic: barres apilades, no apilades, divergents per escales de Likert i gràfics dobles amb segmentació per batxillerat i sexe. Tots els gràfics es van generar correctament i van ser inserits automàticament en documents Word, tal com s'havia dissenyat.

5.2.3 Comprovació dels càlculs interns

A banda de la validació visual, es va dur a terme una comprovació manual dels càlculs corresponents a cada gràfic. Aquest procés consistia a contrastar els resultats representats gràficament amb les dades originals contingudes als fitxers .xlsx generats per Microsoft Forms.

Es va verificar que el criteri de recompte per categoria s'aplicava correctament, assegurant que cada alumne es comptés una sola vegada per categoria en preguntes amb múltiples respostes, d'acord amb els requisits definits. Igualment, es va comprovar que les classificacions automàtiques de les respostes obertes assignaven amb precisió cada entrada a la categoria corresponent, o bé la deixaven degudament marcada com a no classificada per a la seva revisió posterior.

5.2.4 Resultats i conformitat amb els requisits

Els resultats obtinguts van confirmar que el sistema de visualització gràfica desenvolupat funcionava de manera òptima, complint tots els requisits establerts en la fase de disseny, tant pel que fa a l'estructura de la informació com a la seva claredat visual i facilitat d'interpretació.

A continuació, es presenten diverses figures representatives que il·lustren els diferents tipus de gràfics utilitzats en l'anàlisi de dades dels estudiants:

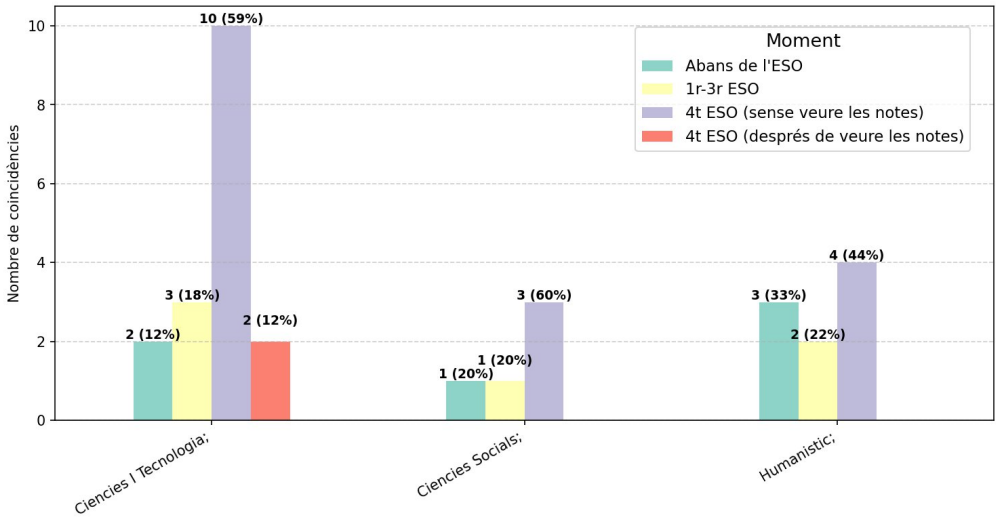


Figura 9. Gràfic de barres no apilat

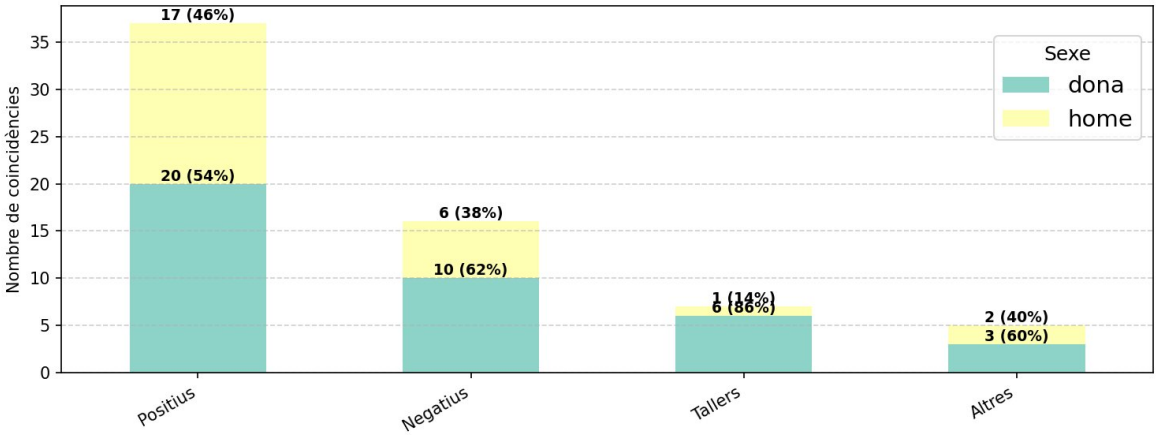


Figura 10. Gràfic de barres apilat

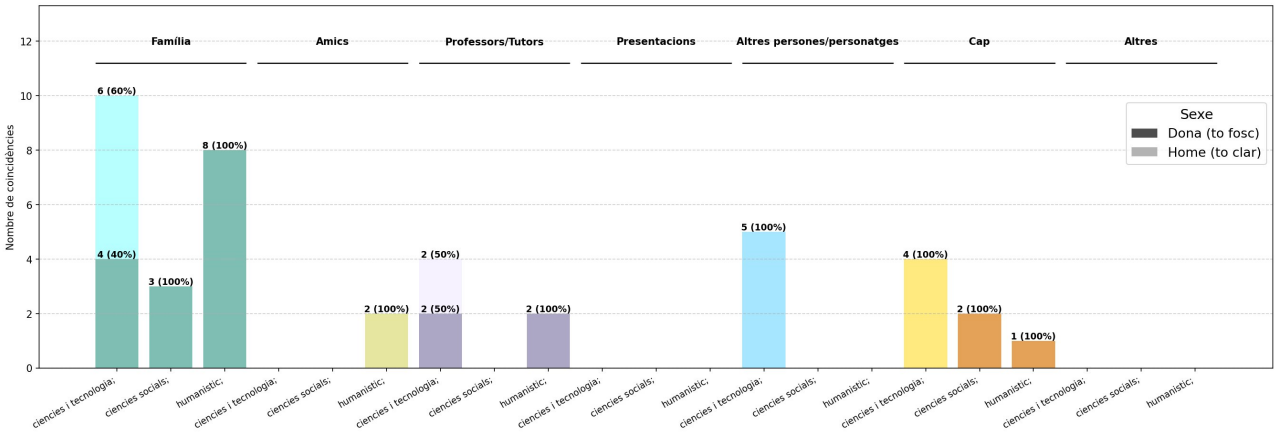


Figura 11. Gràfic de barres apilat amb repetició de batxillerats

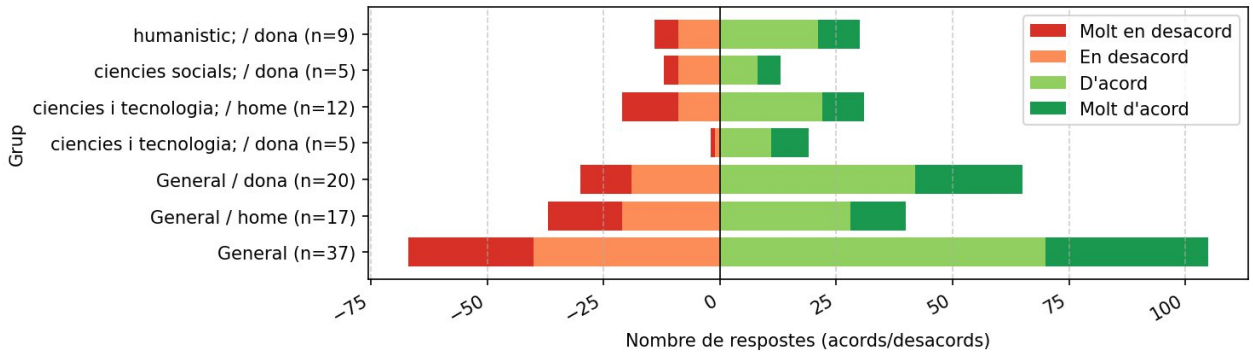


Figura 12. Gràfic de barres apilat i divergent

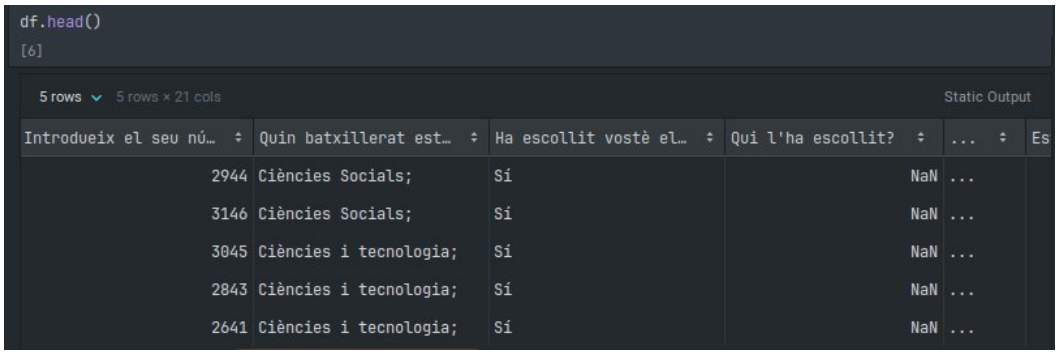
5.3 Neteja i preparació de dades

Un cop recopilades totes les respostes, es va iniciar la fase de neteja i preparació de dades, amb l'objectiu d'assegurar que el conjunt final fos adequat per a l'anàlisi gràfica i el clústering posterior. Aquesta etapa va incloure una revisió exhaustiva de cada pas implementat durant el desenvolupament, validant-ne tant l'eficàcia com la precisió en la classificació.

5.3.1 Validació de la normalització i l'estructura del DataFrame

Per comprovar que la normalització dels valors s'aplicava correctament, es va fer ús de la funció .head() de la llibreria pandas [20], que permet obtenir una previsualització de les primeres files del DataFrame. Aquesta eina resulta molt útil per confirmar que els canvis s'han aplicat tal com s'esperava, tant als noms de les columnes (preguntes) com als valors de les files (respostes dels estudiants).

Es va revisar visualment si els caràcters no desitjats com accents, caràcters especials o espais innecessaris havien estat eliminats o substituïts correctament. També es va validar que les columnes identificatives com id_centre i codi_postal s'haguessin introduït correctament, i que els codis especials com el CODI_BLANC o CODI_BRANCHING es col·loquessin on tocava, especialment en els camps amb resposta buida o condicional.



```
df.head()
[6]
```

5 rows · 5 rows × 21 cols — Static Output

Introdueix el seu nú…	Quin batxillerat est…	Ha escollit vostè el…	Qui l'ha escollit?	...	Es
2944	Ciències Socials;	Sí	NaN	...	
3146	Ciències Socials;	Sí	NaN	...	
3045	Ciències i tecnologia;	Sí	NaN	...	
2843	Ciències i tecnologia;	Sí	NaN	...	
2641	Ciències i tecnologia;	Sí	NaN	...	

Figura 13. Utilització de la funcio .head()

5.3.2 Proves específiques per la classificació automàtica

Per a les variables amb respostes obertes que havien de ser classificades automàticament (com assignatures, estudis, llengües, països i dubtes), es va desenvolupar un conjunt de datasets de prova. Aquests contenien mostres representatives de respostes reals, amb l'objectiu de comprovar l'efectivitat de cada sistema de classificació.

Validació del reconeixement d'assignatures

En el cas de les assignatures, es va fer ús de la llibreria TheFuzz per comparar cada entrada amb una llista de claus i detectar la més similar. Es va comprovar manualment si la categoria assignada era correcte. Amb un cutoff del 80%, es va obtenir un alt nivell d'encert, tal com es mostra a la taula següent:

Assignatura (resposta estudiant)	**Assignatura detectada**	**Percentatge de similitud (%)**	**Categoria assignada**
llati	llati	89%	lletres
Física	fisica	100%	ciencies
Si literatura	literatura	95%	lletres
Física i química	fisica i quimica	93%	ciencies

Taula 6. Validació del procés de reconeixement d'assignatures

Validació del reconeixement d'estudis

Per a la classificació d'estudis futurs, on les entrades podien ser molt variades, es va aplicar el mateix mètode amb un cutoff més permissiu (60%) per evitar pèrdues d'informació. El sistema va aconseguir classificar correctament la majoria d'entrades

Estudis (resposta estudiant)	**Estudi detectat**	**Percentatge de similitud (%)**	**Categoria assignada**
magisteri especialitzat en educacio especial	educacio	90%	socials_i_humanistiques
acabar el bachillerato		0%	altres
magisterio	magisteri	95%	socials_i_humanistiques
993		0%	no_ho_saben
doble grau de matematiques i fisica	matematiques	90%	stem
psicologia o ingenieria	psicologia	90%	sanitaries
despres de batxillerat full fer un cicle mitja de disseny de moda	disseny de moda	90%	artistic

Taula 7. Validació del procés de reconeixement d'estudis

Validació del reconeixement de llengües i països

Les llengües i països també van ser validats seguint el mateix patró. El percentatge d'encert va ser molt elevat, especialment en els casos de llengües habituals i països amb noms fàcilment identificables.

Llengua text (resposta estudiant)	**Llengua detectada**	**Percentatge de similitud (%)**	**Llengua assignada**	**Pais text** (resposta estudiant)	**Pais detectat**	**Percentatge de similitud (%)**	**Pais codificat assignat**
arab i amazigh	arab	90%	arab	espana	espanya	92%	724
espanyol	espanol	93%	castella	marrocc	marroc	92%	504
frances pero en se poquissim	frances	90%	frances	italia	italia	100%	380
angles	angles	100%	angles	brazil	brasil	83%	076
castella catala	catala	90%	catala	paisos baixos	paisos baixos	100%	528
espanyol	espanol	93%	castella	regne despanya	espanya	90%	724

Taula 8. Validació del procés de reconeixement de llengües i països

Validació del reconeixement de dubtes

El cas dels dubtes va requerir una aproximació més complexa amb embeddings semàntics. Es van provar tres models diferents de representació vectorial per identificar la frase més similar dins d'un conjunt de dubtes predefinits. El model que va oferir un millor rendiment va ser all-MiniLM-L6-v2, amb un percentatge d'encert superior al 83%.

Model	Encerts	Errors	Total	Percentatge d'encert (%)
MiniLM-v2	50	10	60	83.33%
Paraphrase-MiniLM	49	11	60	81.67%
DistilUSE	46	14	60	76.67%

Taula 9. Validació del procés de reconeixement de dubtes

5.3.3 Validació de la categorització final

Amb el conjunt de dades net i normalitzat, es va procedir a construir un nou DataFrame per al clústering, aplicant el sistema de classificació desenvolupat a cadascuna de les columnes seleccionades.

Per assegurar el bon funcionament del procés, es va revisar manualment que les columnes categoritzades contenien únicament les categories desitjades, i no apareguessin valors no previstos. Es van utilitzar les funcions .unique() i .value_counts() per detectar possibles errors i assegurar que les etiquetes estiguessin ben definides.

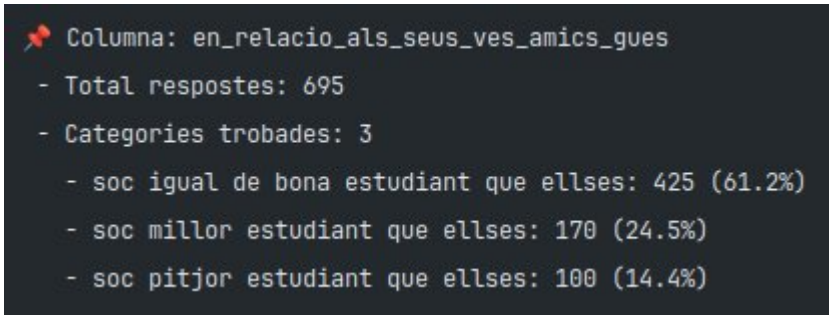


Figura 14. Ús de la funció .unique()

Especial atenció es va posar en columnes amb classificacions compostes, com les de factors d'influència o les persones influents, on diverses respostes s'agrupaven sota una mateixa etiqueta. Es va confirmar que els valors assignats seguien els criteris establerts al disseny i reflectien fidelment les respostes originals dels estudiants.

5.3.4 Conclusions de la validació

Totes les proves han confirmat que el sistema de neteja, classificació i preparació de dades funciona correctament, amb alts percentatges d'encert en la detecció de categories i sense introducció d'errors greus en cap columna.

5.4 Validació dels models de clústering

Un cop obtingut el dataframe finalment categoritzat, es procedeix a la validació i aplicació dels diferents models de clústering. Aquest dataframe inclou les columnes seleccionades per a l'anàlisi, així com una columna addicional amb el sexe dels participants, que s'utilitza exclusivament per fer un seguiment i interpretació posterior, però que es retira abans d'aplicar qualsevol algorisme per tal que no interfereixi en la classificació automàtica.

Abans d'iniciar el procés de clústering, es verifica una vegada més la correcta estructuració del conjunt de dades mitjançant la funció .head() per fer una inspecció visual ràpida, així com la revisió del nombre de columnes i categories per assegurar que el format és l'esperat. Un cop confirmat que les dades són consistents, es passa a la validació i execució dels tres models previstos: jeràrquic, DBSCAN i KModes.

5.4.1 Clústering jeràrquic i validació de la distància Gower

Per aplicar el clústering jeràrquic i DBSCAN, és imprescindible calcular prèviament la matriu de distàncies de Gower, ja que aquests mètodes treballen amb mesures de similitud en espais categòrics. Per validar aquesta matriu, es genera la seva representació gràfica, que permet detectar visualment possibles errors o anomalies.

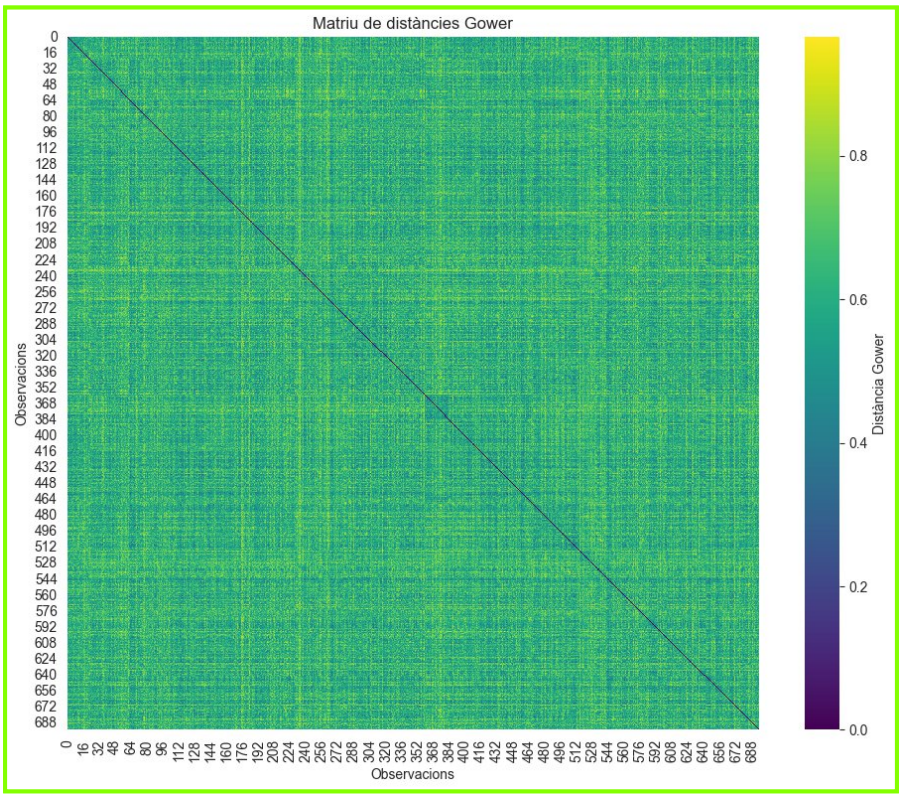


Figura 15. Matriu de distàncies de Gower

Posteriorment, es calcula el dendograma per avaluar visualment l'estructura jeràrquica de les dades i suggerir possibles nombres òptims de clústers. En aquest cas, s'identifiquen dues opcions consistents: una amb 5 clústers (línia verda) i una altra amb 8 (línia vermella), segons els punts de tall més nets del dendograma.

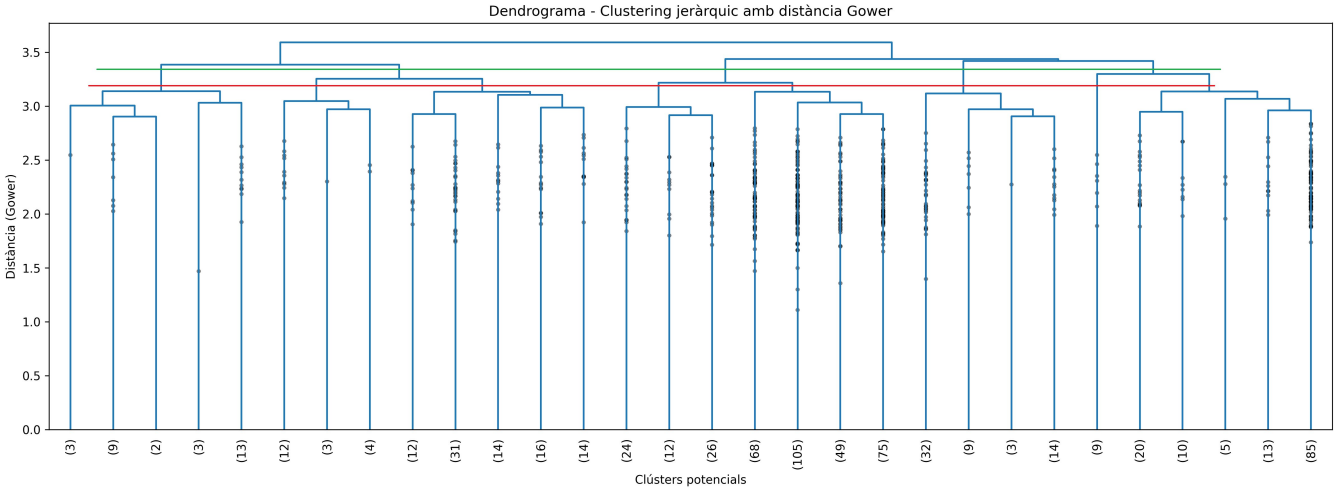


Figura 16. Dendograma a partir de la distància de Gower

Aquests resultats són positius, ja que mostren una certa estabilitat i diversitat en les agrupacions. S'apliquen, per tant, les dues configuracions (5 i 8 clústers) amb el mètode jeràrquic, obtenint fitxers Excel amb els resultats ja classificats. Tot i això, el càlcul del silhouette score, que mesura la qualitat de separació entre grups, dona un valor baix (0.038), indicant que els clústers tenen molt de solapament i no es defineixen amb claredat. Això pot afectar el seu valor predictiu, però no impedeix una possible lectura descriptiva si es fa una bona interpretació.

5.4.2 Resultats del model DBSCAN

Amb el model DBSCAN, s'experimenta amb diferents combinacions dels paràmetres `eps` i `min_samples` per intentar obtenir agrupacions útils. Malgrat això, en cap cas s'assoleix una classificació viable: en la millor configuració, només 141 de les 695 observacions (aproximadament un 20.3%) són classificades, mentre que la resta són marcades com a soroll. Aquest percentatge de classificació és massa baix per considerar el model com a útil en aquest context, i per tant no s'aprofundeix en la seva anàlisi.

5.4.3 Aplicació i resultats del model KModes

Per aplicar el model KModes primer de tot es comprova el conjunt de dades amb les funcions .head(), .unique() i .value_counts() per garantir que no s'han introduït errors en la transformació.

	quin_batxiller…	ha_escollit…	quins_factors…	volia_ev…	familia	amics	professors/tutors
0	ciencies socials	si	positius	ciencies	els pares	altres	la meva tutora de rt deso
1	humanistic	si	positius	mates	la meva mare	altres	altres
2	humanistic	si	positius	irrellevants	pares + germans	altres	professorat

Figura 17. Validació de les categories

Un cop confirmada la qualitat del dataset, s'executa el model K-Modes amb diferents configuracions de nombre de clústers. De manera interessant, s'observa que les millors separacions s'aconsegueixen també amb 5 i 8 clústers, igual que en el cas del dendograma. Aquesta coincidència reforça la idea que aquests nombres de clústers poden reflectir estructures reals dins el conjunt d'alumnes.

Els resultats obtinguts amb K-Modes són satisfactoris: es generen agrupacions amb diferències clares i útils per identificar perfils d'alumnat. Aquesta classificació es considera la més útil del projecte, ja que permet extreure patrons interpretables i comparables, i es fa servir com a base per a les anàlisis finals de perfils.

5.5 Resultats i interpretació del clústering

Un cop aplicats els diferents mètodes de clústering sobre el conjunt de dades netejat i codificat, es procedeix a comparar els resultats obtinguts per determinar quina tècnica proporciona una segmentació més informativa i útil de l'alumnat segons les seves característiques.

S'han aplicat els mètodes jeràrquic i K-Modes amb dues configuracions de nombre de clústers: 5 i 8. El mètode DBSCAN ha estat descartat a causa de la gran quantitat de punts classificats com a soroll, fet que l'invalida per a aquesta anàlisi.

En aquesta fase, la variable sexe s'ha utilitzat com a criteri principal de validació visual i inicial, ja que és una variable rellevant per a l'estudi de vocacions i diferències d'interessos en contextos educatius. La diferència de distribució entre homes i dones en cada clúster s'ha emprat com a indicador per identificar agrupacions amb característiques diferenciades.

5.5.1 Resultats del clústering jeràrquic

Resultats amb 5 clústers:

Clúster	Sexe	Comptador	Percentatge (%)	Diferència de sexe (%)	Total clúster
0	dona	16	53,3%	6,6	30
	home	14	46,7%		
1	dona	54	50,9%	1,8	106
	home	52	49,1%		
2	altre	2	0,6%		360
	dona	219	61%	22,6	
	home	138	38,4%		
3	dona	20	34,5%	31	58
	home	38	65,5%		
4	altre	1	0,7%		143
	dona	71	50%	0,7	
	home	70	49,3%		

Taula 10. Resultats clústering jeràrquic amb 5 clústers

L'anàlisi mostra que només dos dels cinc clústers presenten una diferència clara entre sexes, concretament els clústers 2 i 3, on la proporció entre dones i homes és significativament desigual (amb una diferència superior al 20%), indicant agrupacions amb perfils potencialment diferenciats segons el gènere. Aquests clústers s'han ressaltat en verd per la seva rellevància interpretativa.

En canvi, els altres tres clústers (0, 1 i 4), tot i mostrar una lleugera inclinació cap a un sexe o un altre, presenten diferències mínimes (inferiors al 7%) i es poden considerar equilibrats pel que fa a la composició de gènere. Aquestes agrupacions, marcades en groc, no evidencien un biaix de sexe prou significatiu per permetre extreure conclusions específiques sobre preferències diferenciades.

Resultats amb 8 clústers:

Clúster	Sexe	Comptador	Percentatge (%)	Diferència de sexe (%)	Total clúster
0	dona	16	53,3%	6,6	30
	home	14	46,7%		
1	dona	14	73,7%	47,4	19
	home	5	26,3%		
2	dona	40	46%	8	87
	home	47	54%		
3	dona	49	79%	58	62
	home	13	21%		
4	altre	2	0,7%		
	dona	170	57,2%	15,1	297
	home	125	42,1%		
5	dona	20	34,5%	31	58
	home	38	65,5%		
6	dona	3	33,3%	33,4	9
	home	6	66,7%		
7	altre	1	0,8%		
	dona	68	51,1%	3	133
	home	64	48,1%		

Taula 11. Resultats clústering jeràrquic amb 8 clústers

Amb la configuració de vuit clústers, l'anàlisi mostra una segmentació més detallada pel que fa a la variable sexe. Concretament, quatre dels vuit clústers (1, 3, 5 i 6) presenten diferències significatives entre dones i homes, amb una discrepància superior al 30%, indicant perfils clarament diferenciats segons el gènere. Dues d'aquestes agrupacions tenen una majoria destacada de dones, mentre que les altres dues són clarament masculines; per aquest motiu s'han assenyalat en verd com a clústers rellevants des del punt de vista de la variable sexe.

Els altres quatre clústers (0, 2, 4 i 7), tot i mostrar lleugeres diferències en la distribució per sexe, presenten proporcions més equilibrades, amb discrepàncies inferiors al 15%. Aquestes agrupacions s'han marcat en groc, ja que no evidencien una inclinació suficient per considerar-les fortament diferenciades per gènere. Aquesta major granularitat permet identificar subgrups amb patrons més específics, tot i que també posa de manifest que no totes les agrupacions es defineixen principalment per la variable sexe.

5.5.2 Resultats del clústering amb K-Modes

Resultats amb 5 clústers:

Clúster	Sexe	Comptador	Percentatge (%)	Diferència de sexe (%)	Total clúster
0	dona	156	70,9%	41,8	220
	home	64	29,1%		
1	altre	1	0,8%		132
	dona	34	25,8%	47,7	
	home	97	73,5%		
2	altre	1	0,8%		120
	dona	64	53,3%	7,5	
	home	55	45,8%		
3	altre	1	0,8%		123
	dona	52	42,3%	14,6	
	home	70	56,9%		
4	dona	74	74%	48	100
	home	26	26%		

Taula 12. Resultats clústering k-modes amb 5 clústers

Aquest model ha generat cinc clústers amb una distribució força equilibrada pel que fa a la mida de cada grup, facilitant així la comparació entre perfils. Dels cinc clústers, tres (0, 1 i 4) presenten diferències de sexe clarament marcades, superiors al 40%, cosa que indica agrupacions amb característiques molt diferenciades segons el gènere. Aquests clústers, destacats en verd, ofereixen una base sòlida per a l'anàlisi de patrons específics vinculats al sexe.

En canvi, els clústers 2 i 3 mostren una distribució més equitativa entre homes i dones, amb diferències inferiors al 15%. Aquesta baixa diferenciació limita el seu valor interpretatiu pel que fa a la variable sexe, tot i que poden aportar informació rellevant respecte a altres variables del conjunt de dades.

En conjunt, el model K-Modes demostra ser més robust que el jeràrquic en termes de coherència i claredat en la segmentació, permetent una millor identificació de subgrups amb perfils ben definits.

Resultats amb 8 clústers:

Clúster	Sexe	Comptador	Percentatge (%)	Diferència de sexe (%)	Total clúster
0	altre	1	1%		98
	dona	52	53,1%	7,2%	
	home	45	45,9%		
1	altre	1	0,8%		126
	dona	34	27%	45,2%	
	home	91	72,2%		
2	dona	70	66,7%	33,4%	105
	home	35	33,3%		
3	dona	60	67,4%	34,8%	89
	home	29	32,6%		
4	dona	38	49,4%	1,2%	77
	home	39	50,6%		
5	dona	42	58,3%	16,6%	72
	home	30	41,7%		
6	dona	60	77,9%	55,8%	77
	home	17	22,1%		
7	altre	1	2%		51
	dona	24	47,1%	3,9%	
	home	26	51%		

Taula 13. Resultats clústering k-modes amb 8 clústers

Amb la configuració de vuit clústers, el model K-Modes continua identificant agrupacions amb una diferenciació clara pel que fa a la variable sexe. Concretament, quatre clústers (1, 2, 3 i 6) presenten diferències superiors al 30% entre homes i dones, evidenciant perfils amb una marcada predominança de gènere. Aquests clústers, destacats en verd, són especialment valuosos per analitzar comportaments o interessos diferenciats segons el sexe.

En canvi, la resta de clústers (0, 4, 5 i 7) mostren distribucions més equilibrades, amb diferències de sexe inferiors al 20%, fet que en redueix el potencial discriminant per a l'anàlisi basada en aquesta variable. Això indica una segmentació més difusa pel que fa al gènere dins d'aquest subconjunt.

Tot i que la segmentació amb vuit clústers aporta més detall i revela patrons específics, la configuració amb cinc clústers es confirma com la més òptima en aquest cas, ja que manté una bona diferenciació entre grups sense generar clústers massa petits o amb un valor analític limitat.

5.5.3 Perfils identificats

Aprofundint en l'anàlisi del model K-Modes amb cinc clústers, que ha demostrat oferir la segmentació més clara i significativa entre perfils, s'observa la possibilitat d'identificar grups d'alumnat ben diferenciats. Aquesta classificació facilita l'estudi de les seves característiques tenint en compte variables com el sexe, les preferències acadèmiques i les motivacions personals. En els apartats següents es presenten detalladament aquests perfils, posant especial èmfasi en els trets més rellevants de cada agrupació.

Clúster	% dones	Itineràri principal	Itineràri secundàri	Què volen estudiar?	Mitjana Actual	Mitjana 4rt ESO
0	**70,9%**	ciencies socials (44,5%)	humanistic (21,8%)	Socials i Humanístics **(55%)**	**Notable (55,9%)**	**Notable (68,2%)**
					Aprovat (34,5%)	Excel·lent (16,8%)
1	25,8% **(73,5% homes)**	ciencies i tecnologia **(77,3%)**	ciencies socials (11,4%)	STEM **(57.6%)**	Notable (50%)	Notable (58,3%)
				Socials i Humanístics (17,5%)	Aprovat (24,2%)	Excel·lent (31,8%)
2	~50%	ciencies i tecnologia **(71,7%)**	ciencies socials (12,5%)	STEM (42.5%)	**Notable (55%)**	**Notable (60,8%)**
				Sanitàries (16,7%)	Aprovat (26,7%)	Excel·lent (33,3%)
3	~40%	ciencies i tecnologia **(64,2%)**	ciencies socials (12,2%)	STEM (36.6%)	**Aprovat (65%)**	**Notable (60,2%)**
				Socials i Humanístics (22,8%)	Notable (17,1%)	Aprovat (22%)
4	**74%**	ciencies i tecnologia (42%)	humanistic (18%)	Socials (41%)	Notable (50%)	**Excel·lent (56%)**
				Sanitàries (20%)	Aprovat (23%)	Aprovat (22%)

Taula 14. Taula resum global de clústers (Perfil general)

Clúster	Quan vas triar el batxillerat?	Quants dels teus amics fan el mateix batxillerat que tu?	En relació als teus amics, et consideres...
0	4t ESO (sense notes) (54,1%)	Molts (24,5%)	Igual de bon/a estudiant (70%)
	1r/2n/3r ESO (20,9%)	Meitat (20,5%)	Millor estudiant (16,4%)
1	1r/2n/3r ESO (41,7%)	Pocs (35,6%)	Igual de bon/a estudiant (49,2%)
	4t ESO (sense notes) (35,6%)	Meitat (22,7%)	Millor estudiant (38,6%)
2	4t ESO (sense notes) (46,7%)	Tots o gairebé tots (34,2%)	Igual de bon/a estudiant (66,7%)
	1r/2n/3r ESO (27,5%)	Molts (22,5%)	Millor estudiant (25%)
3	4t ESO (sense notes) (63,4%)	Pocs (32,5%)	Igual de bon/a estudiant (61%)
	1r/2n/3r ESO (18,7%)	Meitat (20,3%)	Pitjor estudiant (22%)
4	4t ESO (sense notes) (42%)	Cap o gairebé cap (32%)	Igual de bon/a estudiant (51%)
	1r/2n/3r ESO (26%)	Tots o gairebé tots (20%)	Millor estudiant (32%)

Taula 15. Taula resum variables acadèmiques i socials

Clúster	Nervis examen general	Nervis exàmens matemàtiques	Sóc bo/na en matemàtiques	Nervis exàmens llengües	Sóc bo/na en llengües
0	Poc nerviós/a (35,9%)	**Molt nerviós/a (43,6%)**	**Molt en desacord (42,7%)**	**Nerviós/a (45,5%)**	**D'acord (50%)**
	Molt nerviós/a (24,1%)	Nerviós/a (20,5%)	En desacord (20,5%)	Poc nerviós/a (23.6%)	
1	**Molt poc nerviós/a (56,8%)**	**Poc nerviós/a (45,5%)**	**D'acord (50,8%)**	**Poc nerviós/a (48,5%)**	**D'acord (48,5%)**
	Poc nerviós/a (16,7%)		Molt d'acord (25%)	**Molt poc nerviós/a (22%)**	Molt en desacord (22%)
2	Poc nerviós/a (39,2%)	**Nerviós/a (50%)**	**D'acord (53,3%)**	**Nerviós/a (56,7%)**	**En desacord (55%)**
	Nerviós/a (30%)	Molt nerviós/a (20,8%)			
3	Nerviós/a (39,8%)	**Molt nerviós/a (39,8%)**	En desacord (36,6%)	Poc nerviós/a (45.5%)	D'acord (36,6%)
		Molt poc nerviós/a (19,5%)	Molt d'acord (18,7%)		En desacord (24,4%)
4	**Molt nerviós/a (53%)**	**Molt nerviós/a (47%)**	D'acord (32%)	Molt nerviós/a (33%)	**D'acord (46%)**
	Molt poc nerviós/a (13%)	Nerviós/a (24%)	En desacord (24%)	Molt poc nerviós/a (25%)	**Molt d'acord (29%)**

Taula 16. Taula resum nervis exàmens i relació matemàtiques/llengües

Clúster	Preocupació per les notes	Em considero un/a bon/a estudiant	M'agrada aprendre
0	**Molt d'acord (61,8%)**	D'acord (61,8%)	**D'acord (65,9%)**
	D'acord (33,2%)		Molt d'acord (19,5%)
1	**D'acord (64,4%)**	D'acord (62,1%)	D'acord (59,1%)
	Molt d'acord (28%)		Molt d'acord (31,1%)
2	**Molt d'acord (65,8%)**	D'acord (59,2%)	**D'acord (60,8%)**
	D'acord (30%)	Molt d'acord (16,7%)	Molt d'acord (32,5%)
3	**D'acord (65,9%)**	D'acord (47,2%)	D'acord (61,8%)
	Molt d'acord (24,4%)		Molt d'acord (18,7%)
4	**Molt d'acord (81%)**	**Molt d'acord (44%)**	**Molt d'acord (70%)**
		D'acord (27%)	D'acord (23%)

Taula 17. Taula resum variables emocionals i motivacionals

Clúster	M'agrada dibuixar / pintar / fer manualitats	M'agrada cantar / tocar instruments / compondre música	M'agraden els jocs de taula / scape rooms / trencaclosques /màgia	M'agrada provar coses noves o fer experiments	M'agrada llegir
0	**D'acord (39,5%)**	D'acord (35%)	**D'acord (57,7%)**	**D'acord (62,7%)**	En desacord (36,4%)
	Molt d'acord (20,5%)	En desacord (23,6%)	Molt d'acord (18,2%)	Molt d'acord (15%)	Molt d'acord (25%)
1	En desacord (37,1%)	**Molt en desacord (48,5%)**	**D'acord (50,8%)**	**Molt d'acord (47,7%)**	En desacord (36,4%)
	Molt en desacord (26,5%)	En desacord (18,9%)	Molt d'acord (25%)	D'acord (30,3%)	D'acord (28,8%)
2	D'acord (36,7%)	En desacord (41,7%)	**D'acord (55,8%)**	**D'acord (61,7%)**	En desacord (36,4%)
	Molt en desacord (20,8%)	Molt en desacord (16,7%)	Molt d'acord (24,2%)	Molt d'acord (25%)	D'acord (28,8%)
3	**Molt en desacord (43,1%)**	No sap (26%)	**D'acord (49,6%)**	**D'acord (61%)**	**Molt en desacord (38,2%)**
	D'acord (17,9%)	Molt en desacord (25,2%)	Molt d'acord (20,3%)	Molt d'acord (18,7%)	D'acord (24,4%)
4	**Molt d'acord (45%)**	Molt d'acord (31%)	**Molt d'acord (50%)**	**Molt d'acord (64%)**	**Molt d'acord (57%)**
	D'acord (21%)	En desacord (19%)	D'acord (26%)	D'acord (17%)	D'acord (21%)

Taula 18. Taula resum d'interessos i aficions personals

Clúster	Factors	Família	Presentacions	Amics	Professors /tutors	Altres persones	Cap factor
0	Positius (46,8%)	Els pares (42,3%)	Presentacions a la uni/saló enseny. (3,6%)	(22,7%)	(12,3%)	(9,5%)	(13,6%)
	Negatius i positius (22,7%)						
1	Positius (57,6%)	Els pares (38,6%)	Presentacions a la uni/saló enseny. (5,3%)	(18,9%)	(19,7%)	(12,1%)	(18,9%)
	Negatius i positius (11,4%)						
2	Positius (49,2%)	**Pares + germans (49,2%)**	Altres coneguts (13,3%)	**(56,7%)**	(24,2%)	(13,3%)	(10%)
	Negatius i positius (11,7%)	Els pares (20%)					
3	Positius (55,3%)	Pares + germans (15,4%)	Presentacions a la uni/saló enseny. (4,1%)	(11,4%)	(8,9%)	(13,8%)	(35%)
	Negatius i positius (13%)						
4	Positius (48%)	Els pares (25%)	Presentacions a la uni/saló enseny. (3%)	(15%)	(16%)	(14%)	(29%)
	Negatius i positius (14%)						

Taula 19. Taula resum factors i persones/personatges que han influït en la decisió

Clúster 0

Aquest clúster destaca per tenir una clara majoria de noies (70,9%) i una preferència per itineraris de ciències socials (44,5%) i humanístics (21,8%). Les preferències d'estudi es centren principalment en àmbits socials i humanístics (55%) i, en menor mesura, en les ciències de la salut (Veure taula 14). Acadèmicament, la majoria tenen mitjanes de Notable tant actuals com a 4t d'ESO (55,9% i 68,2%, respectivament), tot i que també hi ha una presència significativa d'alumnes amb Aprovat o Excel·lent (Veure taula 14).

Pel que fa a l'elecció del batxillerat, més de la meitat ho decideixen a 4t d'ESO (54,1%) i acostumen a tenir molts amics que fan el mateix itinerari (24,5%) (Veure taula 15). Emocionalment, tendeixen a mostrar nervis en exàmens, sobretot de matemàtiques, i una percepció negativa de les seves capacitats en aquesta àrea, tot i que se senten més còmodes amb les llengües (Veure taula 16). Són alumnes preocupats per les notes (61,8%) i amb una actitud positiva cap a l'aprenentatge (65,9%) (Veure taula 17). Mostren interès en activitats creatives i experimentals (dibuixar, jocs de taula, fer experiments), però no tant en la lectura (Veure taula 18). L'entorn familiar, especialment els pares, és el principal factor d'influència (Veure taula 19).

Clúster 1:

Amb un 73,5% d'alumnes homes, aquest clúster es focalitza en itineraris de ciències i tecnologia (77,3%) i mostra una clara orientació cap als estudis STEM (57,6%) (Veure taula 14). Les seves mitjanes acadèmiques es concentren al voltant del Notable, i destaca el percentatge d'alumnes amb Excel·lent a 4t d'ESO (31,8%) (Veure taula 14). Sovint decideixen l'itinerari abans de 4t d'ESO (41,7%) i tenen pocs amics amb el mateix itinerari (35,6%) (Veure taula 15).

Aquests estudiants mostren una baixa resposta emocional davant els exàmens, especialment en matemàtiques, on també tenen una percepció molt positiva de les seves capacitats (Veure taula 16). Es consideren bons estudiants (62,1%) i els agrada aprendre (59,1%) (Veure taula 17). Tenen menys interès en activitats artístiques, però sí en activitats d'experimentació i resolució de problemes (Veure taula 18). L'entorn familiar és rellevant però també valoren altres referents (Veure taula 19).

Clúster 2:

Aquest grup presenta una distribució paritària entre sexes (\~50%) i una orientació predominant cap a itineraris de ciències i tecnologia (71,7%), amb interès en estudis STEM (42,5%) i sanitaris (16,7%) (Veure taula 14). Acadèmicament, també dominen les mitjanes de Notable, amb una notable presència d'Excel·lents (33,3%) a 4t d'ESO (Veure taula 14). Trien batxillerat principalment a 4t d'ESO (46,7%) i acostumen a tenir molts o tots els amics al mateix itinerari (56,7%) (Veure taula 15).

Tot i una certa tensió davant els exàmens, tenen una percepció positiva de les seves habilitats tant en matemàtiques com en llengües (Veure taula 16). La majoria es preocupen per les notes i es consideren bons estudiants, tot i que també destaca el gust per aprendre (60,8%) (Veure taula 17). Tenen interessos diversos: els agraden els jocs de taula, fer experiments i, en menor mesura, la lectura (Veure taula 18). Les influències familiars (pares i germans) i les amistats tenen un pes rellevant en la seva decisió (Veure taula 19).

Clúster 3:

Aquest clúster presenta un perfil masculinitzat (\~60% homes) i una orientació cap a ciències i tecnologia (64,2%), amb preferència relativa per estudis STEM (36,6%) i socials-humanístics (22,8%) (Veure taula 14). El seu rendiment acadèmic és més baix, amb una majoria amb mitjana d'Aprovat actualment (65%) (Veure taula 14). L'elecció del batxillerat sovint es fa tard (63,4% a 4t d'ESO) i tenen pocs amics al mateix itinerari (Veure taula 15).

Mostren un nivell de nervis elevat als exàmens, especialment en matemàtiques, i perceben menys competència en aquesta àrea (Veure taula 16). Els seus interessos personals són més limitats, amb poc gust per la lectura o activitats artístiques (Veure taula 18). El suport familiar és més dèbil i molts indiquen no haver tingut cap influència clara (35%) (Veure taula 19).

Clúster 4:

Aquest és el perfil amb més presència femenina (74%) i una orientació variada: ciències i tecnologia (42%), humanístic (18%) i un interès destacat pels estudis socials i sanitaris (Veure taula 14). Acadèmicament, tenen bones qualificacions, amb un 56% amb Excel·lent a 4t d'ESO i un 50% de Notables actuals (Veure taula 14). També és rellevant l'elevada motivació per les notes (81%) i per aprendre (70%) (Veure taula 17).

Emocionalment, mostren molts nervis davant els exàmens, però tenen una bona percepció d'habilitats, especialment en llengües (Veure taula 16). Els agraden activitats creatives i culturals: dibuixar, llegir, fer experiments... (Veure taula 18). Les influències familiars hi són presents, tot i que molts també reconeixen altres factors com amics, professors o presentacions institucionals (Veure taula 19).

Diferències i similituds

Els cinc clústers comparteixen alguns trets comuns com ara una tendència general cap a mitjanes de Notable i una elecció majoritària dels itineraris de ciències i tecnologia. La majoria dels alumnes decideixen l'itinerari a 4t d'ESO, valoren les notes i tenen influència familiar en la seva decisió. També és comú una certa inseguretat davant els exàmens, especialment en matemàtiques, tot i que aquesta varia en intensitat segons el perfil.

Les diferències principals es troben en el gènere, la motivació, el rendiment acadèmic i els interessos personals. Els clústers 1 i 2 destaquen per una orientació STEM amb seguretat i bon rendiment. El clúster 0 reflecteix un perfil femení amb preferència per l'àmbit social-humanístic i el 4, també femení, es caracteritza per un alt rendiment i interès divers. Així, els perfils varien tant en actitud com en orientació vocacional i emocional davant els estudis.

6 Conclusions

En conclusió, aquest treball ha assolit amb èxit l'objectiu principal: dissenyar, distribuir i analitzar un qüestionari per comprendre els factors que influeixen en la tria dels itineraris de batxillerat entre l'alumnat. A més, s'ha aconseguit automatitzar la generació d'informes visuals personalitzats per a cada centre educatiu i implementar tècniques de clústering que han permès identificar perfils diferenciats dins del conjunt d'estudiants.

Un dels focus principals ha estat analitzar les diferències de gènere, especialment per entendre per què les noies opten menys per itineraris vinculats amb l'àmbit STEM, amb l'objectiu de generar eines que ajudin a reduir aquesta bretxa en el futur.

Durant el desenvolupament del projecte, s'han superat diversos reptes tècnics i metodològics. Un dels més rellevants ha estat el tractament de les respostes obertes, que s'ha resolt mitjançant tècniques de fuzzy matching per codificar-les de manera coherent. També s'ha abordat la detecció i gestió de respostes amb dubtes, definint codis específics que asseguren la coherència en l'anàlisi. Pel que fa al clústering, la naturalesa categòrica de les dades ha suposat un repte addicional, ja que moltes tècniques convencionals no són aplicables. Per això, s'ha treballat amb mètodes com K-Modes i l'ús de distàncies de Gower, validant els resultats tot i les limitacions inherents a aquestes mètriques.

Un altre repte important ha estat interpretar quins factors diferenciaven realment els perfils identificats. Aquesta dificultat no ha estat només tècnica sinó també analítica, per tal d'entendre què tenen en comú o què els distingeix més enllà de la variable de gènere.

Pel que fa als perfils identificats, s'han pogut detectar agrupacions diferenciades d'alumnes segons variables com el gènere, el rendiment acadèmic, el nivell de preocupació per les notes i els interessos d'estudi. Aquesta informació obre la porta a una millor comprensió dels factors que intervenen en la tria educativa i pot tenir aplicacions útils per part dels equips orientadors dels centres. En concret, s'han identificat cinc perfils ben definits que es diferencien tant en les seves aspiracions acadèmiques com en els seus trets emocionals i motivacionals.

Tot i que s'han assolit els objectius principals plantejats, una possible millora hauria estat aprofundir més en l'anàlisi dels resultats del clústering. Hauria estat enriquidor realitzar més proves comparatives entre diferents models i explorar amb més detall les diferències entre clústers, per extreure conclusions més precises i fonamentades sobre els patrons que distingeixen cada perfil d'alumnat. Aquesta anàlisi més exhaustiva es planteja com una línia de treball futura per ampliar l'impacte i la utilitat del projecte.

6.1 Competència CT7 – Ètica i responsabilitat social

El projecte desenvolupat posa en el centre la igualtat d'oportunitats en la tria dels itineraris educatius, amb especial atenció a les possibles diferències de gènere. Els formularis van ser dissenyats per ser neutres i accessibles per a tot l'alumnat, independentment de la seva identitat, mentre que els algorismes de clústering aplicats no van imposar patrons previs, sinó que van permetre identificar desigualtats existents a partir de les respostes reals. Això ha facilitat l'observació de perfils diferenciats que poden ajudar a comprendre millor les desigualtats estructurals i socials que influeixen en les decisions educatives. A més, la mostra recollida és representativa de diversos tipus de centres (públics i privats) i contextos socials, afavorint una mirada inclusiva i transversal.

Pel que fa a la responsabilitat social, aquest treball pretén generar un impacte real en la comunitat educativa. L'anàlisi i evidència de possibles desigualtats en la tria del batxillerat poden contribuir a que futurs projectes institucionals o docents dissenyin intervencions més justes i orientadores. La voluntat final és reduir bretxes i fomentar decisions acadèmiques informades i lliures de condicionants externs, com la pressió social o els estereotips de gènere.

Quant a l'impacte ambiental, el projecte no implica processos amb una petjada ecològica significativa. L'anàlisi de dades es realitza amb eines de baix cost computacional, sense necessitat d'entrenar models massius ni d'un ús intensiu de servidors externs. Totes les execucions i proves s'han dut a terme des d'un ordinador personal, i el processament de dades s'ha limitat a un conjunt reduït, cosa que minimitza l'impacte sobre els recursos naturals i el consum energètic.

Finalment, el projecte ha complert en tot moment amb els criteris ètics establerts per la universitat. El qüestionari va ser validat pel comitè ètic, i tots els centres van rebre la documentació prèvia, inclòs el full de consentiment informat per a l'alumnat. Durant la recollida de dades, es va garantir la privacitat dels participants i es van aplicar mecanismes d'anonimització perquè els resultats finals fossin sempre agregats i no identificables. Aquesta rigorosa protecció del tractament de dades reforça el compromís del projecte amb la responsabilitat i el respecte als drets individuals.

6.2 Valoració personal

Aquest projecte ha estat una experiència profundament enriquidora que m'ha permès aplicar de manera pràctica els coneixements adquirits durant el grau, com ara programació, estadística i tècniques d'intel·ligència artificial, a un problema real amb fortes implicacions socials i educatives. Desenvolupar un projecte des del disseny conceptual del qüestionari fins a la generació automàtica d'informes i l'anàlisi mitjançant clústering m'ha ajudat a comprendre en profunditat tot el cicle d'un projecte de dades.

A més, he consolidat habilitats tècniques com el tractament i la neteja de dades, la codificació de respostes obertes, la creació de visualitzacions clares i interpretables, i la presentació de resultats de manera accessible per a públics no tècnics. També he après a gestionar millor el temps i a adaptar el treball davant dels obstacles trobats, ja siguin problemes tècnics o reptes metodològics.

Aquest projecte ha reforçat el meu interès per l'anàlisi de dades, i per això, durant l'any vinent continuaré la meva formació amb un màster en ciència de dades. Sens dubte, aquest treball final de grau ha estat un pas fonamental per confirmar la meva trajectòria professional i continuar creixent en aquest àmbit.

6.3 Futurs desenvolupament i millores

Aquest treball representa l'inici d'un estudi més ampli que ja ha començat a prendre forma i que, a mesura que avanci, es preveu que incorporarà més instituts, enriquint així la base de dades i fent-la més representativa. L'objectiu final és aprofundir en la comprensió dels factors que influeixen en la tria dels itineraris de batxillerat per part de l'alumnat i aportar evidència empírica útil per a la presa de decisions educatives.

Entre les possibles línies de millora per a futures fases, es proposa incloure noves variables d'interès que no s'han recollit en aquesta primera etapa, com ara la influència

familiar, el context socioeconòmic o l'accés a serveis d'orientació acadèmica. Aquestes dimensions podrien oferir una visió més completa dels condicionants que afecten les decisions dels estudiants.

També seria interessant perfeccionar els models de classificació utilitzats, aplicant tècniques més avançades d'anàlisi descriptiva i explorant capacitats predictives. A més, es podrien comparar els resultats entre diferents promocions o realitzar un seguiment longitudinal de les trajectòries dels estudiants, per tal de verificar si les expectatives manifestades en el qüestionari es corresponen amb els camins educatius realment seguits.

7 Referències

[1] Gibert, K., & Valls, A. (2022). Building a Territorial Working Group to Reduce Gender Gap in the Field of Artificial Intelligence. Applied Sciences, 12(6), 3129. https://doi.org/10.3390/app12063129

[2] UNESCO. (2023). Girls' and Women's Education in STEM. Recuperat de https://www.unesco.org/en/gender-equality/education/stem [Consultat el: 18/02/2025]

[3] Fundació Bofill. (2023). 8 dades per parlar d'educació i gènere el 8M. Recuperat de https://fundaciobofill.cat/blog/8-dades-per-parlar-d-educaci-i-gnere-el-8m [Consultat el: 21/02/2025]

[4] Fulls d'Enginyeria. (s.d.). Evolució, referents i biaixos inconscients de les dones a les STEM. Recuperat de https://www.fullsdenginyeria.cat/evolucio-referents-i-biaixos-inconscients-dones-les-stem [Consultat el: 21/02/2025]

[5] United Nations Statistics Division. (s.d.). Standard country or area codes for statistical use (M49). Recuperat de https://unstats.un.org/unsd/methodology/m49/ [Consultat el: 07/03/2025]

[6] MindOnMap. (s.d.). What is Bar Graph: Definition, Types, Uses, and Examples. Recuperat de https://www.mindonmap.com/ca/blog/knowledge/what-is-bar-graph/ [Consultat el: 08/03/2025]

[7] FusionCharts. (2021). Choosing the Right Chart Type: Column Charts vs. Stacked Column Charts. Recuperat de https://www.fusioncharts.com/blog/choosing-the-right-chart-type-column-charts-vs-stacked-column-charts/ [Consultat el: 28/03/2025]

[8] Naouali, S., Ben Salem, S., & Chtourou, Z. (2019). Clustering Categorical Data: A Survey. International Journal of Information Technology & Decision Making, 19(1), 7–41. https://doi.org/10.1142/S0219622019300064

[9] Cochet, S. (2021). TheFuzz: Fuzzy string matching in Python. GitHub. https://github.com/seatgeek/thefuzz [Consultat el: 28/03/2025]

[10] Nam, E. (2019). Understanding the Levenshtein Distance Equation for Beginners. Recuperat de https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0 [Consultat el: 30/03/2025]

[11] IBM. (s.d.). What is clustering in machine learning? IBM. https://www.ibm.com/topics/clustering [Consultat el: 12/04/2025]

[12] Matplotlib Developers. (s.d.). Matplotlib: Visualization with Python. Recuperat de https://matplotlib.org/ [Consultat el: 13/04/2025]

[13] NumPy Developers. (s.d.). NumPy. Recuperat de https://numpy.org/ [Consultat el: 19/04/2025]

[14] python-docx Developers. (s.d.). python-docx Documentation. Recuperat de https://python-docx.readthedocs.io/en/latest/# [Consultat el: 19/04/2025]

[15] Streamlit Inc. (s.d.). Streamlit. Recuperat de https://streamlit.io/ [Consultat el: 19/04/2025]

[16] Jupyter Project. (s.d.). Nbconvert Documentation. Recuperat de https://nbconvert.readthedocs.io/en/latest/ [Consultat el: 11/03/2025]

[17] Python Package Index. (s.d.). nbformat. Recuperat de https://pypi.org/project/nbformat/ [Consultat el: 11/03/2025]

[18] PyWebView Developers. (s.d.). pywebview. Recuperat de https://pywebview.flowrl.com/ [Consultat el: 11/03/2025]

[19] pandas Development Team. (s.d.). User Guide — pandas Documentation. Recuperat de https://pandas.pydata.org/docs/user_guide/index.html [Consultat el: 21/04/2025]

[20] Python Package Index. (s.d.). kmodes. Recuperat de https://pypi.org/project/kmodes/ [Consultat el: 21/04/2025]

[21] SciPy Developers. (s.d.). scipy.cluster.hierarchy.linkage — SciPy Documentation. Recuperat de https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html [Consultat el: 21/04/2025]

[22] scikit-learn Developers. (s.d.). sklearn.cluster.DBSCAN — scikit-learn Documentation. Recuperat de https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html [Consultat el: 21/04/2025]

[23] Python Package Index. (s.d.). gower. Recuperat de https://pypi.org/project/gower/ [Consultat el: 21/04/2025]

[24] Ranjan, A. (2020). Calinski-Harabasz Index for K-Means Clustering Evaluation Using Python. Towards Data Science. Recuperat de https://towardsdatascience.com/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python-4fefeeb2988e/ [Consultat el: 08/05/2025]

[25] Yadav, N. (2020). Gower's Distance. Medium – Analytics Vidhya. Recuperat de https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553 [Consultat el: 12/05/2025]

[26] GeeksforGeeks. (s.d.). K-Modes Clustering in Python. Recuperat de https://www.geeksforgeeks.org/k-mode-clustering-in-python/ [Consultat el: 19/05/2025]

[27] GeeksforGeeks. (s.d.). Hierarchical Clustering. Recuperat de https://www.geeksforgeeks.org/hierarchical-clustering/ [Consultat el: 19/05/2025]

[28] GeeksforGeeks. (s.d.). DBSCAN Clustering in ML | Density Based Clustering. Recuperat de https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ [Consultat el: 19/05/2025]

Annex 1

(Omplir per l'enquestador)

Número:

ID: _____

UNIVERSITAT ROVIRA i VIRGILI

PRIMERA PART

[1] Quin batxillerat està fent?

- ◯ Humanístic
- ◯ Ciències Socials
- ◯ Ciències i Tecnologia
- ◯ General
- ◯ Arts

[2] Ha escollit vostè el batxillerat que fa?

- ◯ Sí
- ◯ No Qui l'ha escollit? _____

<u>Si ha escollit vostè el batxillerat respongui, si us plau, a les tres preguntes següents:</u>

[3] Quins factors han influït en la seva decisió? (pot marcar-ne tants com vulgui)

- ◯ Se'm donen bé aquestes matèries
- ◯ No se'm donen bé algunes matèries dels altres batxillerats
- ◯ He participat en tallers de ciència o de tecnologia
- ◯ Volia evitar determinades assignatures. Especificar: _____
- ◯ Crec que són matèries que m'obriran portes
- ◯ He participat en tallers d'escriptura o en concursos literaris
- ◯ És el batxillerat que em permetrà estudiar el que jo vull estudiar
- ◯ No sabia què escollir i ho he fet per eliminació
- ◯ Crec que és l'únic batxillerat on puc sortir-me'n i aprovar
- ◯ Jo vull guanyar molts diners i crec que aquest batxillerat és bo per això
- ◯ És el que fan els meus amics/gues
- ◯ He vist una presentació sobre aquestes matèries i m'han agradat
- ◯ Aquest batxillerat em permetrà seguir amb la tradició o el negoci familiar
- ◯ No soc prou llest/a per a fer determinats batxillerats
- ◯ Altres: _____
- ◯ Cap

[4] Quines persones o personatges han influït positivament en la seva decisió? (pot marcar tots els que vulgui)

- ◯ El meu pare
- ◯ La meva mare
- ◯ Els meus germans/es i/o cosins/es
- ◯ Els meus amics/gues
- ◯ Els meus professors (homes)
- ◯ Les meves professores (dones)
- ◯ El meu tutor de 4rt d'ESO
- ◯ La meva tutora de 4rt d'ESO
- ◯ Altres persones que conec (amics de la família, veïns...). Especificar:_____
- ◯ Persones que han vingut a fer presentacions a l'escola o a l'institut
- ◯ Persones que van fer les presentacions quan vaig visitar la universitat, el saló d'ensenyament...
- ◯ Persones que no conec (personalitats de les xarxes socials, escriptors...)
- ◯ Personatges de sèries, pel·lícules, llibres...
- ◯ Altres. Especificar: _____
- ◯ Cap

[5] Quan va decidir fer aquest batxillerat?

- ◯ Abans de la ESO
- ◯ A 1er, 2on o 3er d'ESO
- ◯ Durant 4rt d'ESO (al llarg del curs)
- ◯ Al final de 4rt d'ESO sense preocupar-me per les notes
- ◯ Al final de 4rt d'ESO després de veure les notes
- ◯ Altres _____

[6] Del seu grup d'amics/gues quants/es fan el mateix batxillerat que vostè?

- ◯ Cap
- ◯ Gairebé cap
- ◯ Pocs
- ◯ Aproximadament la meitat
- ◯ Molts
- ◯ Tots o gairebé tots

[7] És el primer cop que fa primer de batxillerat?

- ◯ SÍ
- ◯ No → Ha canviat de batxillerat?
  - ◯ Sí → Quin feia abans? _____
  - ◯ No

SEGONA PART

[8] Si us plau completi la taula següent amb l'opció més adient:

	Molt d' acord	*D' acord*	*En desacord*	*Molt en desacord*	*No sap*
Em preocupa treure bones notes					
M'agraden els jocs de taula, els *scape room*, els trencaclosques o la màgia					
Em poso molt nerviós/a quan faig/feia exàmens de matemàtiques					
Em poso molt nerviós/a quan faig/feia exàmens de llengua					
Em considero un/a bon/a estudiant					
M'agrada dibuixar, pintar, fer manualitats ...					
M'agrada aprendre					
Soc bo/na en llengües					
M'agrada provar coses noves o fer experiments					
Dormo malament el dia abans de fer un examen					
M'agrada llegir					
Soc bo/na en matemàtiques					
M'agrada cantar, tocar instruments, compondre música...					

[9] En relació als seus/ves amics/gues:

- ◯ Soc millor estudiant que ells/es
- ◯ Soc igual de bon/a estudiant que ells/es
- ◯ Soc pitjor estudiant que ells/es

[10] Quina va ser la seva nota mitjana de 4rt d'ESO?

- ◯ Aprovat
- ◯ Notable
- ◯ Excel·lent

[11] En mitjana quina és la seva nota actual?

- ◯ Suspès
- ◯ Aprovat
- ◯ Notable
- ◯ Excel·lent

[12] Vol seguir estudiant en el futur?

◯ Sí Què vol estudiar? _____

◯ No

TERCERA PART

[13] Si us plau indiqui el seu sexe:

◯ Home

◯ Dona

◯ Altres

[14] Edat: _____

[15] País de naixement: _____

[16] Llengua materna: _____

[17] Llengua pròpia (sentida): _____

[18] Si us plau, indiqui amb qui conviu:

◯ Visc amb el meu pare i la meva mare conjuntament (estan junts)

◯ Visc amb el meu pare o la meva mare depenent del dia (estan separats)

◯ Visc amb el meu pare

◯ Visc amb la meva mare

◯ Altres. Especificar: _____

[19] Com és la connexió a internet de la seva llar?

◯ Bona

◯ Dolenta

◯ No n'hi ha

[20] Quants ordenadors hi ha a la seva llar?

◯ Cap o 1

◯ 2 o més

[21] Es considera membre de grups o minories socials o culturals?

◯ Sí Quines? _____

_____

_____

◯ No

Les preguntes següents són sobre el seu PARE

[22a] País de naixement del pare: _____

[22b] Quin nivell de formació té el seu pare?

- ◯ Cap
- ◯ Formació bàsica/obligatòria
- ◯ Formació professional Quina?_____
- ◯ Formació universitària Quina?_____
- ◯ No ho sé

[22c] En general, el seu pare té una feina remunerada (guanya diners per la feina que fa)?

- ◯ Sí De què treballa? _____
- ◯ No

Les preguntes següents són sobre la seva MARE

[23a] País de naixement de la mare: _____

[23b] Quin nivell de formació té la seva mare?

- ◯ Cap
- ◯ Formació bàsica/obligatòria
- ◯ Formació professional Quina?_____
- ◯ Formació universitària Quina?_____
- ◯ No ho sé

[23c] En general, el seva mare té una feina remunerada (guanya diners per la feina que fa)?

- ◯ Sí De què treballa? _____
- ◯ No

Les preguntes següents són sobre els seus germans/es o cosins/es més grans

[24a] Té germans o germanes més grans?

◯ Sí Quants germans:_____ Quantes germanes:_____

◯ No

Si NO té germans/es més grans vagi a la pregunta número 22

Si té germans/es més grans especifiqui per a cada un d'ells/es l'edat, si estudien o han estudiat i què han estudiat o estan estudiat

GERMANS				GERMANES			
	Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?		Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?
Germà 1				Germana 1			
Germà 2				Germana 2			
Germà 3				Germana 3			
Germà 4				Germana 4			
Germà 5				Germana 5			

RESPONGUI NOMÉS SI NO TÉ GERMANS/ES MÉS GRANS

[24b] Té cosins o cosines propers/es més grans?

◯ Sí Quants cosins:_____ Quantes cosines:_____

◯ No

Especifiqui per a cada un d'ells/es l'edat, si han estudiat o estan estudiant i què estudien o han estudiat.

COSINS				COSINES			
	Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?		Edat	Estudia o ha estudiat?	Què estudia o ha estudiat?
Cosí 1				Cosina 1			
Cosí 2				Cosina 2			
Cosí 3				Cosina 3			
Cosí 4				Cosina 4			
Cosí 5				Cosina 5			

MOLTES GRÀCIES PER LA SEVA COL·LABORACIÓ

Annex 2

(Plantilla) Eleccions d'itinerari durant el batxillerat

Universitat Rovira i Virgili

25 d'abril de 2025

* Required

1

Introdueix el seu número *

Number must be between 100000 ~ 999999

Primera Part

2

Quin batxillerat està fent? *

- [ ] Humanístic
- [ ] Ciències Socials
- [ ] Humanístic i socials
- [ ] Ciències i tecnologia
- [ ] General
- [ ] Arts

3

Ha escollit vostè el batxillerat que fa? *

- ◯ Sí
- ◯ No

4

Qui l'ha escollit? *

Enter your answer

5

Quins factors han influït en la seva decisió? *

(Pot marcar-ne tants com vulgui)

- [ ] Se'm donen bé aquestes matèries
- [ ] No se'm donen bé algunes matèries dels altres batxillerats
- [ ] He participat en tallers de ciència o de tecnologia
- [ ] Crec que són matèries que m'obriran portes
- [ ] He participat en tallers d'escriptura o en concursos literaris
- [ ] És el batxillerat que em permetrà estudiar el que jo vull estudiar
- [ ] No sabia què escollir i ho he fet per eliminació
- [ ] Crec que és l'únic batxillerat on puc sortir-me'n i aprovar
- [ ] Jo vull guanyar molts diners i crec que aquest batxillerat és bo per això
- [ ] És el que fan els meus amics/gues
- [ ] He vist una presentació sobre aquestes matèries i m'han agradat
- [ ] Aquest batxillerat em permetrà seguir amb la tradició o el negoci familiar
- [ ] No soc prou llest/a per a fer determinats batxillerats
- [ ] Cap
- [ ] Other

6

Volia evitar alguna/es assignatures?

(Especifiqui quina/es)

Enter your answer

7

Quines persones o personatges han influït positivament en la seva decisió? *

(Pot marcar-ne tants com vulgui)

- [ ] El meu pare
- [ ] La meva mare
- [ ] Els meus germans/es i/o cosins/es
- [ ] Els meus amics/gues
- [ ] Els meus professors (homes)
- [ ] Les meves professores (dones)
- [ ] El meu tutor de 4rt d'ESO
- [ ] La meva tutora de 4rt d'ESO
- [ ] Altres persones que conec (amics de la família, veïns...).
- [ ] Persones que han vingut a fer presentacions a l'escola o a l'institut
- [ ] Persones que van fer les presentacions quan vaig visitar la universitat, el saló d'ensenyament...
- [ ] Persones que no conec (personalitats de les xarxes socials, escriptors...)
- [ ] Personatges de sèries, pel·lícules, llibres...
- [ ] Cap
- [ ] Other

8

Quan va decidir fer aquest batxillerat? *

- ◯ Abans de la ESO
- ◯ A 1er, 2on o 3er d'ESO
- ◯ A 4rt d'ESO sense preocupar-me per les notes
- ◯ A 4rt d'ESO després de veure les notes
- ◯ Soc repetidor
- ◯ Other

9

Quin batxillerat feia abans? *

- ◯ Humanístic
- ◯ Ciències Socials
- ◯ Ciències i tecnologia
- ◯ General
- ◯ Arts

10

Del seu grup d'amics/gues quants/es fan el mateix batxillerat que vosté? *

- ◯ Cap
- ◯ Gairebé cap
- ◯ Pocs
- ◯ Aproximadament la meitat
- ◯ Molts
- ◯ Tots o gairebé tots

Segona Part

11

Si us plau completi la taula següent amb l'opció més adient *

	Molt d'acord	D'acord	En desacord	Molt en desacord	No sap
Em preocupa treure bones notes	◯	◯	◯	◯	◯
M'agraden els jocs de taula, els scape room, els trencaclosques o la màgia	◯	◯	◯	◯	◯
Em poso molt nerviós/a quan faig/feia exàmens de matemàtiques	◯	◯	◯	◯	◯
Em poso molt nerviós/a quan faig/feia exàmens de llengua	◯	◯	◯	◯	◯
Em considero un/a bon/a estudiant	◯	◯	◯	◯	◯
M'agrada dibuixar, pintar, fer manualitats ...	◯	◯	◯	◯	◯
M'agrada aprendre	◯	◯	◯	◯	◯
Soc bo/na en llengües	◯	◯	◯	◯	◯
M'agrada provar coses noves o fer experiments	◯	◯	◯	◯	◯
Dormo malament el dia abans de fer un examen	◯	◯	◯	◯	◯
M'agrada llegir	◯	◯	◯	◯	◯
Soc bo/na en matemàtiques	◯	◯	◯	◯	◯
M'agrada cantar, tocar instruments, compondre música...	◯	◯	◯	◯	◯

12

En relació als seus/ves amics/gues *

- ◯ Soc millor estudiant que ells/es
- ◯ Soc igual de bon/a estudiant que ells/es
- ◯ Soc pitjor estudiant que ells/es

13

Quina va ser la seva nota mitjana de 4rt d'ESO? *

- ◯ Aprovat
- ◯ Notable
- ◯ Excel·lent

14

En mitjana quina és la seva nota actual? *

- ◯ Suspès
- ◯ Aprovat
- ◯ Notable
- ◯ Excel·lent

15

Vol seguir estudiant en el futur? *

- ◯ Sí
- ◯ No

16

Què vol estudiar? *

Enter your answer

Tercera Part

17

Si us plau indiqui el seu sexe *

- ◯ Home
- ◯ Dona
- ◯ Altre

18

Edat *

Please enter a number greater than or equal to 14

19

País de naixement *

Enter your answer

20

Llengua materna *

Enter your answer

21

Llengua que sent com a pròpia *

Enter your answer

22

Indiqui amb qui conviu: *

- ◯ Visc amb el meu pare i la meva mare conjuntament (estan junts)
- ◯ Visc amb el meu pare o la meva mare depenent del dia (estan separats)
- ◯ Visc amb el meu pare
- ◯ Visc amb la meva mare
- ◯ Other

23

Com és la connexió a internet de la seva llar? *

- ◯ Bona
- ◯ Dolenta
- ◯ No n'hi ha

24

Quants ordinadors hi ha a la seva llar? *

- ◯ Cap o 1
- ◯ 2 o més

25

Es considera membre de grups o minories socials o culturals? *

- ◯ Sí
- ◯ No

26

Quines? *

Enter your answer

Les preguntes són sobre el teu PARE

27

País de naixement del pare

Enter your answer

28

Quin nivell de formació té el seu pare?

- ◯ Cap
- ◯ Formació bàsica/obligatòria
- ◯ Formació professional
- ◯ Formació universitària
- ◯ No ho sé

29

Quina formació professional/universitària té?

Enter your answer

30

En general, el seu pare té una feina remunerada?

(Guanya diners per la feina que fa)?

- ◯ Sí
- ◯ No

31

De què treballa?

Les preguntes són sobre la teva MARE

32

País de naixement de la mare

Enter your answer

33

Quin nivell de formació té la seva mare?

- ◯ Cap
- ◯ Formació bàsica/obligatòria
- ◯ Formació professional
- ◯ Formació universitària
- ◯ No ho sé

34

Quina formació professional/universitària té?

Enter your answer

35

En general, la seva mare té una feina remunerada?

(Guanya diners per la feina que fa)?

- ◯ Sí
- ◯ No

36

De què treballa?

Enter your answer

Preguntes sobre GERMANS

37

Té germans **(masculins) més grans**? *

- 1 germà
- 2 germans
- 3 germans
- 4 o més germans
- No

38

Quina edat té el seu germà? *

Please enter a number greater than or equal to 14

39

Quina edat tenen els seus dos germans? *

(Indica-ho separat per una coma: "Edat 1, Edat 2")

Please enter text that contains ,

40

Quina edat tenen els seus germans? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

41

Quina edat tenen els seus germans? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

42

El seu germà estudia/ha estudiat? *

- ◯ Sí
- ◯ No

43

Estudis del seu germà *

Enter your answer

44

El seu primer germà estudia/ha estudiat? *

- ◯ Sí
- ◯ No

45

Estudis del seu primer germà *

Enter your answer

46

El seu segon germà estudia/ha estudiat? *

- ◯ Sí
- ◯ No

47

Estudis del seu segon germà *

Enter your answer

48

El seu primer germà estudia/ha estudiat? *

- Sí
- No

49

Estudis del seu primer germà *

Enter your answer

50

El seu segon germà estudia/ha estudiat? *

- Sí
- No

51

Estudis del seu segon germà *

Enter your answer

52

El seu tercer germà estudia/ha estudiat? *

- Sí
- No

53

Estudis del seu tercer germà *

Enter your answer

54

Què estudien/han estudiat els seus germans? *

Indica-ho separat per comes

55

Té alguna germana més gran? *

- ◯ Sí
- ◯ No

Preguntes sobre GERMANES

56

Té germanes **més grans**? *

- 1 germana
- 2 germanes
- 3 germanes
- 4 o més germanes
- No

57

Quina edat té la seva germana? *

Please enter a number greater than or equal to 14

58

Quina edat tenen les seves dos germanes? *

(Indica-ho separat per una coma: "Edat 1, Edat 2")

Please enter text that contains ,

59

Quina edat tenen les seves germanes? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

60

Quina edat tenen les seves germanes? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

61

La seva germana estudia/ha estudiat? *

- Sí
- No

62

Estudis de la seva germana *

Enter your answer

63

La seva primera germana estudia/ha estudiat? *

- Sí
- No

64

Estudis de la seva primera germana *

Enter your answer

65

La seva segona germana estudia/ha estudiat? *

- Sí
- No

66

Estudis de la seva segona germana *

Enter your answer

67

La seva primera germana estudia/ha estudiat? *

- Sí
- No

68

Estudis de la seva primera germana *

Enter your answer

69

La seva segona germana estudia/ha estudiat? *

- Sí
- No

70

Estudis de la seva segona germana *

Enter your answer

71

La seva tercera germana estudia/ha estudiat? *

- Sí
- No

72

Estudis de la seva tercera germana *

Enter your answer

73

Què estudien/han estudiat les seves germanes? *

(Indica-ho separat per comes)

Preguntes sobre COSINS

74

Té cosins **(masculins) més grans**? *

- 1 cosí
- 2 cosins
- 3 cosins
- 4 o més cosins
- No

75

Quina edat té el seu cosí? *

Please enter a number greater than or equal to 14

76

Quina edat tenen els seus dos cosins? *

(Indica-ho separat per una coma: "Edat 1, Edat 2")

Please enter text that contains ,

77

Quina edat tenen els seus cosins? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

78

Quina edat tenen els seus cosins? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

79

El seu cosí estudia/ha estudiat? *

- Sí
- No

80

Estudis del seu cosí *

Enter your answer

81

El seu primer cosí estudia/ha estudiat? *

- Sí
- No

82

Estudis del seu primer cosí *

Enter your answer

83

El seu segon cosí estudia/ha estudiat? *

- Sí
- No

84

Estudis del seu segon cosí *

Enter your answer

85

El seu primer cosí estudia/ha estudiat? *

- Sí
- No

86

Estudis del seu primer cosí *

Enter your answer

87

El seu segon cosí estudia/ha estudiat? *

- Sí
- No

88

Estudis del seu segon cosí *

Enter your answer

89

El seu tercer cosí estudia/ha estudiat? *

- Sí
- No

90

Estudis del seu tercer cosí *

Enter your answer

91

Què estudien/han estudiat els seus cosins? *

(Indica-ho separat per comes)

Preguntes sobre COSINES

92

Té cosines **més grans**? *

- 1 cosina
- 2 cosines
- 3 cosines
- 4 o més cosines
- No

93

Quina edat té la seva cosina? *

Please enter a number greater than or equal to 14

94

Quina edat tenen les seves dos cosines? *

(Indica-ho separat per una coma: "Edat 1, Edat 2")

Please enter text that contains ,

95

Quina edat tenen les seves cosines? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

96

Quina edat tenen les seves cosines? *

(Indica-ho separat per una coma: "Edat 1, Edat 2"...)

Please enter text that contains ,

97

La seva cosina estudia/ha estudiat? *

- Sí
- No

98

Estudis de la seva cosina *

Enter your answer

99

La seva primera cosina estudia/ha estudiat? *

- Sí
- No

100

Estudis de la seva primera cosina *

Enter your answer

101

La seva segona cosina estudia/ha estudiat? *

- Sí
- No

102

Estudis de la seva segona cosina *

Enter your answer

103

La seva primera cosina estudia/ha estudiat? *

- Sí
- No

104

Estudis de la seva primera cosina *

Enter your answer

105

La seva segona cosina estudia/ha estudiat? *

- Sí
- No

106

Estudis de la seva segona osina *

Enter your answer

107

La seva tercera cosina estudia/ha estudiat? *

- Sí
- No

108

Estudis de la seva tercera cosina *

Enter your answer

109

Què estudien/han estudiat les seves cosines? *

(Indica-ho separat per comes)

Moltes gràcies per la seva col·laboració

URV