UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

DISSENY I ANÀLISI DE TÈCNIQUES DE CLUSTERING APLICADES A SISTEMES DE RECOMANACIÓ

POL FONOYET GONZÁLEZ

Treball Final de Grau

19 de març de 2025

Director/a
CONRADO MARTÍNEZ PARRA (Departament de Ciències de la Computació)

Titulació
Grau en Enginyeria Informàtica (Computació)

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Índex

Índex de figures

Índex de taules

Capítol 1

Contextualització i Abast del projecte

1.1 Introducció

Aquest treball de final de grau s'ha dut a terme en el Grau d'Enginyeria Informàtica impartit en el marc de la Facultat d'Informàtica de Barcelona (FIB), pertanyent a la Universitat Politècnica de Catalunya (UPC). Dins d'aquest programa acadèmic, s'ofereixen diverses àrees d'especialització, entre les quals destaca la de Computació. El present projecte se situa en aquesta especialització, centrant el seu enfocament en temàtiques clau com la Intel·ligència Artificial (IA) i l'Aprenentatge Automàtic (AA).

1.1.1 Context

A mesura que la quantitat de dades generades en l'entorn digital segueix augmentant, la necessitat de tècniques avançades per analitzar i processar aquesta informació es torna cada vegada més crítica. En aquest escenari, els sistemes de recomanació han emergit com a eines essencials per personalitzar l'experiència d'usuari, filtrant i suggerint continguts rellevants en funció de les seves preferències i comportaments [1]. Aquests sistemes exerceixen un paper fonamental en àmbits tan diversos com el comerç electrònic, l'entreteniment i l'educació, permetent millorar la interacció i satisfacció dels usuaris [2]. Segons un informe de McKinsey, el 35% del que els consumidors compren a Amazon i el 75% del que veuen a Netflix provenen de recomanacions [3].

La personalització efectiva requereix comprendre les preferències i comportaments dels usuaris. El filtratge col·laboratiu, una tècnica àmpliament utilitzada, enfronta dos grans desafiaments: l'escalabilitat i la dispersió. A mesura que augmenta la quantitat d'usuaris i ítems, les matrius d'utilitat es tornen enormes, cosa que incrementa el cost computacional [4]. A més, la majoria dels usuaris interactuen amb només una fracció dels ítems, resultant en matrius molt disperses, cosa que dificulta el càlcul precís de similituds [5].

Les tècniques de clustering, com k-means, DBSCAN i k-NN, permeten agrupar

usuaris amb interessos similars, facilitant la identificació de patrons i la generació de recomanacions més precises [6]. En aplicar clustering com a pas previ al filtratge col·laboratiu, es poden crear grups més petits i homogènics, reduint la complexitat de les matrius i el temps de còmput. A més, en treballar amb grups d'usuaris similars, es redueix el problema de la dispersió, ja que els usuaris dins d'un mateix clúster tendeixen a tenir més interaccions en comú [2].

Aquest projecte se centra en el disseny i l'anàlisi de tècniques de clustering aplicades a sistemes de recomanació, amb l'objectiu d'optimitzar la qualitat i el rendiment de les recomanacions.

1.1.2 Conceptes

A continuació, es defineixen els termes i conceptes clau relacionats amb el tema d'estudi.

1.1.2.1 Sistema de Recomanació

Un sistema de recomanació és un sistema que calcula i proporciona contingut rellevant a l'usuari basant-se en el coneixement de l'usuari, el contingut i les interaccions entre l'usuari i l'element. [7]

El propòsit d'una plataforma de recomanació de contingut és facilitar als usuaris l'accés a informació rellevant de manera ràpida i eficient, millorant la seva experiència i fomentant la seva permanència en el servei. Al mateix temps, busca optimitzar recursos i maximitzar beneficis per al proveïdor, assegurant que el contingut ofert sigui atractiu i sostenible econòmicament. Això es fa mitjançant sistemes intel·ligents que analitzen el comportament dels usuaris per oferir opcions personalitzades, tot mantenint un equilibri entre satisfacció i rendibilitat.

1.1.2.2 Filtratge Col·laboratiu

La idea principal dels enfocaments de recomanació col·laborativa és aprofitar la informació sobre el comportament passat o les opinions d'una comunitat d'usuaris existent per predir quins elements li agradaran o interessaran més a l'usuari actual del sistema [8].

La figura 1.1 il·lustra com funciona el filtrat col·laboratiu. El cercle exterior representa el catàleg complet. El cercle intermedi representa un grup d'usuaris que han consumit elements similars. El sistema de recomanació suggereix elements del cercle més proper (davant), partint de la idea que si aquests usuaris coincideixen en preferències amb l'usuari actual, aquest també preferirà els elements que han consumit.

El grup es determina mitjançant la superposició entre els elements que han agradat als usuaris i els que ha agradat a l'usuari actual. Les recomanacions corresponen a la part del cercle intermedi no coberta pel cercle de l'usuari actual (és a dir, els elements que ell/ella no ha consumit però sí el grup semblant).

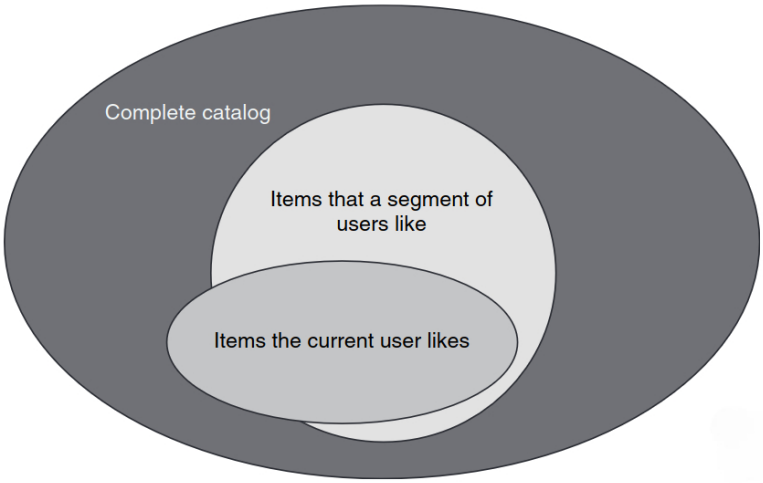

Figura 1.1: Diagrama de filtratge col·laboratiu [7]

1.1.2.3 Clustering

El clustering, segons la perspectiva de [6], es concep fonamentalment com l'organització de dades dins de grups cohesionats, on la semblança entre els elements d'un mateix grup supera la que mantenen amb els d'altres grups. Aquesta tècnica, que es fonamenta en mesures de similitud o distància, és clau en l'anàlisi de dades sense supervisió, ja que permet identificar estructures o patrons latents en els conjunts de dades sense necessitat d'etiquetes prèvies. Així, el clustering facilita la comprensió i la interpretació de grans volums d'informació, sent una eina fonamental tant en l'exploració de dades com en l'aprenentatge automàtic.

1.1.3 Identificació del problema

En aquest treball es desenvoluparà una comparativa entre diferents tècniques de clustering aplicades als sistemes de recomanació. El problema es pot dividir en tres blocs de treball:

1.1.3.1 Selecció i tractament del conjunt de dades

Aquest primer bloc consisteix en la recerca, selecció i/o creació de datasets, tant reals com sintètics, que representin diferents volums de dades i nivells de dispersió, assegurant que els escenaris simulats reflecteixin les diverses situacions que es poden presentar en entorns reals.

1.1.3.2 Desenvolupament del sistema de recomanació

En aquest segon bloc es construirà el sistema de recomanació, on el clustering actuarà com a pas previ al filtratge col·laboratiu. Aquest també capturarà diferents mètriques

de precisió i eficiència computacional. A més, estarà dissenyat de manera que es pugui escollir amb quina tècnica de clusterització ha de funcionar.

1.1.3.3 Comparació de resultats i anàlisi del rendiment

Finalment, es compararan els resultats de les diferents tècniques emprades amb els diferents datasets amb l'objectiu de determinar quina tècnica ofereix el millor equilibri entre eficàcia i eficiència en diferents escenaris.

1.1.4 Actors implicats

En aquest projecte hi ha diversos actors implicats:

- **Estudiant/investigador**: Responsable de desenvolupar, implementar i analitzar les tècniques de clustering aplicades als sistemes de recomanació.
- **Director/tutor acadèmic**: Responsable de supervisar el progrés del projecte i oferir orientació acadèmica.
- **Comunitat acadèmica**: Inclou altres estudiants, investigadors i professionals interessats en els resultats del projecte. Els principals beneficiaris dels resultats serien els investigadors en temes de sistemes de recomanació i aprenentatge automàtic.
- **Proveïdors de datasets**: Organitzacions o comunitats que proporcionen conjunts de dades reals i sintètics per a l'entrenament i proves dels algorismes de clustering.

1.2 Justificació

L'augment exponencial de la informació digital ha generat un repte important per als sistemes de recomanació, ja que els mètodes tradicionals, com el filtratge col·laboratiu, es veuen limitats per problemes d'escala i dispersió de dades. Aquesta situació posa de manifest la necessitat d'explorar noves estratègies per millorar tant la precisió com l'eficiència en la generació de recomanacions.

En aquest context, l'aplicació de tècniques de clustering com a pas previ ofereix un potencial significatiu: agrupar usuaris i ítems en clústers més cohesius redueix la dimensió de la matriu usuari-ítem, minimitza el cost computacional i afavoreix una major densitat en les dades per a càlculs més precisos. Tot i que existeixen diverses solucions informàtiques que integren el clustering en sistemes de recomanació, en la revisió de la literatura no s'han trobat estudis que comparin explícitament els diferents algoritmes de clustering en aquest context de manera sistemàtica.

1.2.1 Estat de l'art

En els darrers anys, l'interès per optimitzar els sistemes de recomanació ha condüït a l'aplicació de diverses tècniques d'aprenentatge automàtic, amb especial èmfasi en el filtratge col·laboratiu. Paral·lelament, s'ha observat una creixent utilització de tècniques de clustering per reduir la dimensionalitat de la matriu usuari-ítem i millorar la densitat de les dades, aspecte clau per a l'eficiència computacional i la precisió de les recomanacions.

Diversos estudis han abordat l'ús d'algoritmes com k-means i DBSCAN, entre d'altres, per agrupar usuaris amb interessos similars, facilitant així la generació de recomanacions personalitzades. Aquestes investigacions han demostrat que el preprocessament de les dades mitjançant clustering pot reduir significativament el temps de càlcul i millorar els resultats en entorns amb grans volums de dades.

No obstant això, en la revisió de la literatura existent s'ha constatat la manca d'estudis que realitzin una comparativa explícita i sistemàtica dels diferents algoritmes de clustering aplicats específicament en el context dels sistemes de recomanació. La major part dels treballs se centra en la implementació d'un sol mètode o en l'aplicació general del clustering sense aprofundir en la comparació directa dels resultats obtinguts per cadascun d'ells.

Aquest buit en la recerca representa una oportunitat per aportar una anàlisi més detallada i una valoració crítica dels avantatges i inconvenients de cada tècnica. El present projecte pretén omplir aquesta llacuna aportant una comparativa exhaustiva dels principals algoritmes de clustering, la qual cosa servirà com a referència tant per a la comunitat acadèmica com per a la indústria, i contribuirà a la millora dels sistemes de recomanació en entorns reals.

1.2.2 Selecció d'eines

En aquest projecte s'ha realitzat una acurada elecció de les eines i entorns de treball, tenint en compte la necessitat d'executar càlculs numèrics, processar grans volums de dades, implementar tècniques d'aprenentatge automàtic i visualitzar els resultats d'una manera clara i professional.

Python és el llenguatge de programació seleccionat per la seva sintaxi senzilla, la seva versatilitat i la seva àmplia comunitat, que garanteix un suport continu i una gran quantitat de recursos per al desenvolupament d'aplicacions científiques i d'aprenentatge automàtic.

Per al processament de dades i càlculs numèrics, s'empren les llibreries **NumPy** i **Pandas**. NumPy facilita la manipulació d'arrays i l'execució d'operacions matemàtiques d'alt rendiment, mentre que Pandas proporciona estructures de dades eficients (com els DataFrames) per a la manipulació i anàlisi de dades, essent essencials per a la preparació i neteja dels datasets.

La implementació i comparativa dels diferents algoritmes de clustering es realitzarà amb la llibreria **Scikit-learn**, que inclou una àmplia varietat d'eines per a l'aprenen-

tatge automàtic, incloent mètodes de classificació, regressió, clustering i reducció de dimensionalitat. Aquesta eina és fonamental per a dur a terme experiments rigorosos i comparatius.

Per a la visualització dels resultats i l'anàlisi exploratòria, s'utilitzarà **Matplotlib**, que permet generar gràfics i representacions visuals personalitzables i de gran qualitat, facilitant així la interpretació de les dades.

A més, s'incorpora el paquet **Surprise**, especialitzat en sistemes de recomanació, que facilita la implementació de tècniques de filtratge col·laboratiu i la comparativa amb els mètodes basats en clustering.

Pel que fa als entorns de desenvolupament, s'ha optat per **Jupyter Notebook**, que permet combinar codi, visualitzacions i documentació en un entorn interactiu, afavorint la iteració i validació ràpida de hipòtesis. El control de versions es gestionarà amb **Git**, assegurant un seguiment acurat dels canvis. Finalment, **LaTeX** s'utilitzarà per a la redacció i formatatge del document final, garantint una presentació clara, estructurada i professional dels resultats.

1.3 Abast

1.3.1 Objectius

L'objectiu principal d'aquest projecte és desenvolupar i comparar diferents tècniques de clustering en el context dels sistemes de recomanació. Això implica:

- Implementar un mòdul de clustering basat en tècniques particionals clàssiques.
- Desenvolupar versions alternatives que utilitzin tècniques com clustering difús i jeràrquic.
- Analitzar i comprendre el funcionament i les particularitats de cada mètode de clustering.
- Avaluar, mitjançant mètriques específiques, la precisió, l'eficiència computacional i la robustesa de cada enfocament.
- Contribuir al coneixement en l'àmbit dels sistemes de recomanació, aportant una comparativa rigorosa i documentada que pugui servir de referència tant per a la comunitat acadèmica com per a futurs desenvolupaments en l'àrea.

1.3.2 Subobjectius

Per assolir l'objectiu principal es plantegen els següents subobjectius:

- Dur a terme una revisió bibliogràfica exhaustiva sobre tècniques de clustering i la seva aplicació en sistemes de recomanació.

- Dissenyar una arquitectura modular que permeti integrar fàcilment diferents algoritmes de clustering en el sistema de recomanació.
- Implementar els algorismes seleccionats de manera que es puguin comparar de forma sistemàtica.
- Capturar i analitzar mètriques clau, com ara el temps d'execució, la densitat de dades i la precisió de les recomanacions.
- Elaborar una documentació detallada que reculli els resultats experimentals i les conclusions obtingudes.

1.3.3 Requeriments

1.3.3.1 Funcionals

- **Recollida i tractament de dades:** Seleccionar conjunts de dades (reals o sintètics) que representin diferents escenaris en sistemes de recomanació. Aplicar tècniques de preprocessament i neteja per garantir la qualitat de les dades per a l'anàlisi.
- **Implementació dels algorismes de clustering:** Desenvolupar implementacions de diferents tècniques de clustering (particional clàssic, difús i jeràrquic). Adaptar els algorismes per estudiar el seu comportament i comparativa en diferents contextos.
- **Anàlisi dels resultats:** Definir i aplicar mètriques d'avaluació que permetin comparar de manera rigorosa els diferents mètodes. Interpretar els resultats per identificar avantatges, limitacions i possibles àrees de millora de cada tècnica.

1.3.3.2 No funcionals

- **Claredat i documentació:** Redactar una documentació completa i exhaustiva que reculli la metodologia, els procediments experimentals i l'anàlisi dels resultats.
- **Reproductibilitat:** Assegurar que els experiments es puguin reproduir seguint la metodologia descrita, facilitant la verificació i validació dels resultats.
- **Validesa dels resultats:** Garantir que les tècniques aplicades permetin arribar a conclusions rigoroses i ben fonamentades, centrant-se en l'anàlisi comparativa més que en la implementació de solucions optimitzades.

1.3.4 Obstacles i riscos

1.3.4.1 Obstacles

Els principals obstacles que es poden trobar en el desenvolupament del projecte són:

- **Falta de coneixement profund en la matèria:** L'autor del treball no té un gran coneixement previ de les diferents tècniques de clustering ni de sistemes de recomanació. Menys de la seva aplicació conjunta.
- **Potència computacional limitada:** El processament i anàlisi de grans volums de dades poden requerir recursos computacionals elevats.
- **Documentació extensa:** La normativa actual dels TFG exigeix la generació d'una àmplia documentació, incloent-hi entregables de GEP i la memòria. Tot i que la seva elaboració demanda una inversió de temps considerable, aquests materials permeten demostrar que s'han aplicat correctament les competències adquirides durant la titulació.

1.3.4.2 Riscos

Els riscos associats al projecte inclouen:

- **Elements imprevistos en el desenvolupament:** Dificultats tècniques o situacions no anticipades poden afectar el progrés del projecte.
- **Gestió del temps:** La coordinació entre la implementació, els experiments i la redacció de la documentació pot resultar més complexa del previst. L'autor també està cursant assignatures que podrien demandar pics d'activitat inesperats.
- **Accidents o imprevistos personals:** Eventualitats que afectin la disponibilitat de temps per al desenvolupament del projecte.

1.4 Metodologia i rigor

L'èxit de qualsevol projecte depèn en gran mesura de la implementació de mètodes estructurats i rigorosos, que estableixen un marc organitzat per afavorir l'eficiència, la qualitat i una comunicació efectiva.

1.4.1 Metodologia de treball

Atès que és un projecte individual, una metodologia àgil i flexible com Kanban és ideal. El mètode Kanban és una metodologia àgil que es fonamenta en la visualització del projecte per millorar la transparència i la col·laboració entre els membres de l'equip. Va ser creat per Taiichi Ohno als anys 40 per a la gestió de la producció, però es va adaptar al món de la gestió de projectes. Es distingeix per la seva simplicitat i per la seva capacitat d'adaptar-se a organitzacions amb estructures jeràrquiques tradicionals. Kanban funciona amb un sistema visual que permet fer un seguiment clar del progrés del projecte [9].

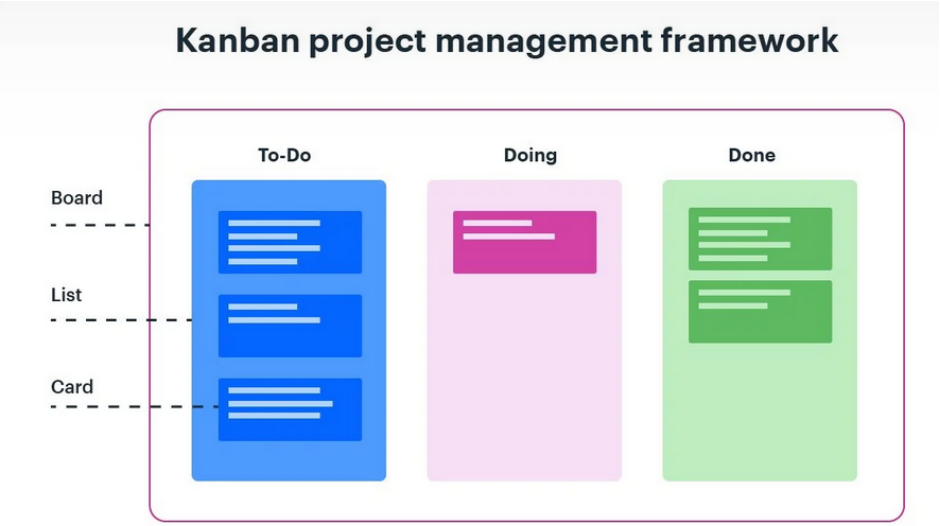

Figura 1.2: Diagrama il·lustratiu del flux de treball segons la metodologia Kanban [9]

A la figura 1.2 es mostren tres columnes principals: *To Do* (Per fer), *Doing* (En procés) i *Done* (Fet). Aquestes columnes representen les fases de treball en què es troben les tasques. Els membres de l'equip traslladen les tasques des de la columna de *To Do* cap a *In Progress* quan comencen a treballar-hi i, finalment, a *Done* quan s'acaben, seguint els límits de treball en curs (WIP) per garantir que el flux de treball es mantingui eficaç.

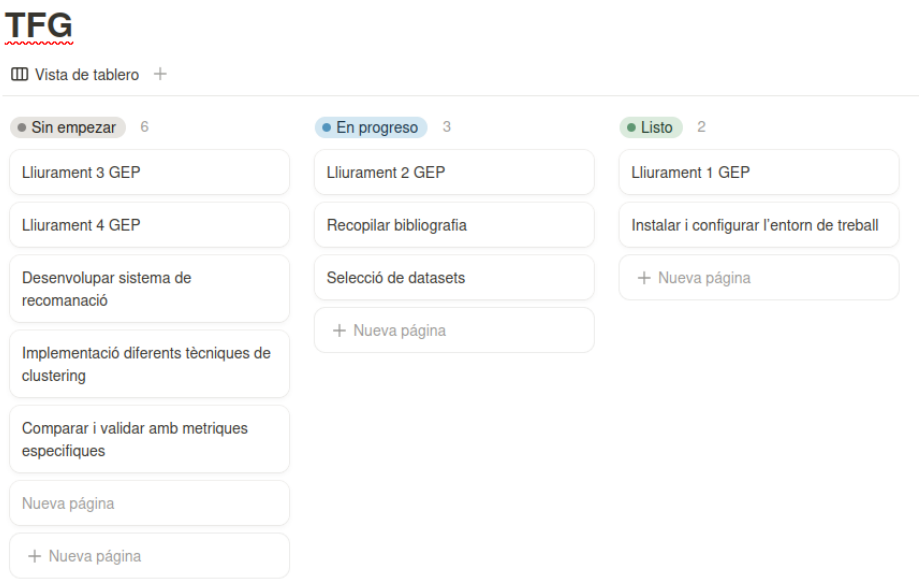

Figura 1.3: Captura de pantalla de la plataforma Notion

Per a la implementació d'aquesta metodologia s'utilitzarà la web de Notion. Notion és una plataforma col·laborativa i de gestió de projectes que facilita l'organització de tasques de manera visual i flexible. Com es veu a la figura 1.3, permet dividir les tasques en els tres grups esmentats, cosa que s'alinea perfectament amb l'enfocament

Kanban, assegurant una gestió eficient, una comunicació fluida i una transparència en el seguiment del progrés del projecte.

1.4.2 Seguiment

Per tal de verificar que el desenvolupament del projecte s'ajusta als requeriments establerts i per corregir sense dilació qualsevol desviació que pugui sorgir, es realitzarà el seguiment del projecte mitjançant reunions presencials cada dues setmanes. En aquestes trobades es procedirà a:

- **Analitzar el progrés global:** Revisarem l'evolució tant del programari com de la documentació, identificant els avanços assolits i les possibles àrees de millora.
- **Detectar dificultats i incidències:** S'exposaran les dificultats trobades durant el desenvolupament, així com qualsevol desviació respecte als requisits inicials.
- **Establir mesures correctores:** Es discutiran i acordaran les millores i ajustaments necessaris per garantir l'alineació amb els objectius del projecte.
- **Comprovar el compliment dels requisits:** Amb l'ajuda d'indicadors clars, es verificarà el grau de compliment dels requisits i subrequisits, facilitant així una gestió proactiva de les incidències.

Aquest sistema de seguiment, basat en reunions presencials periòdiques, permetrà una presa de decisions àgil i una coordinació efectiva entre els actors implicats, assegurant que el projecte es desenvolupi de manera coherent i amb el nivell de qualitat desitjat.

Capítol 2

Planificació temporal

En aquesta secció es realitzarà una planificació temporal del projecte definint tasques específiques i blocs de tasques, per tal d'obtenir una estimació realista de la seva durada. El treball s'inicia el 27 de gener i la presentació està prevista per al 25 de juny, de manera que aquesta seria la data màxima de finalització.

El projecte es desenvoluparà durant aproximadament 140 dies, amb una durada estimada de 450 hores, equivalent als 18 crèdits ECTS que conformen el treball. Aquesta quantitat d'hores es té en compte per calcular el temps que s'ha de dedicar a cada activitat, tal com es mostra a la Taula 2.1. La dedicació diària de dilluns a divendres serà aproximadament de 3-4 hores, i els caps de setmana es podrà dedicar més temps. Es disposarà d'una gran flexibilitat ja que el projecte es realitza íntegrament des de casa. Aquestes hores es podran destinar tant a la documentació com a la programació segons la necessitat.

2.1 Descripció de les tasques

En aquesta secció es descriuran totes les tasques que conformen el projecte. Les tasques més grans es dividiran en subtasques per facilitar la comprensió de les diferents fases. Finalment, es presentarà una taula resum de totes les activitats amb la seva descripció, durada, dependències, recursos i rols assignats, així com el diagrama de Gantt.

2.1.1 GP - Gestió del Projecte

Una gestió eficient és essencial per assegurar-ne l'èxit. Aquesta proporciona un marc estructurat que facilita l'assoliment dels objectius de manera organitzada, dins dels límits establerts de temps, recursos i pressupost. A més, una bona gestió permet minimitzar riscos, garantir la qualitat i adaptar-se a possibles imprevistos. La duració total estimada per a la gestió del projecte és de **65 hores**. Aquest apartat es divideix en diverses fases clau, que es detallen a continuació.

- **GP1 - Contextualització i Abast del projecte**

  Una bona definició de l'abast del projecte és fonamental per garantir-ne l'èxit. En aquesta fase inicial es determinen els objectius, els requisits, les possibles dificultats i els riscos associats. A més, cal haver recopilat informació prèvia sobre les tecnologies i metodologies que es volen aplicar. Es calcula que aquesta tasca requerirà aproximadament **15 hores**.

- **GP2 - Planificació temporal**

  Per tal de garantir una execució òptima, cal establir una planificació temporal adequada, identificant les tasques específiques, assignant responsabilitats i establint terminis clars. Això permet una millor organització del treball, minimitzant possibles retards i optimitzant els recursos. Es calcula que aquesta tasca requerirà aproximadament **15 hores**.

- **GP3 - Gestió Econòmica i sostenibilitat**

  Es durà a terme una anàlisi detallada dels costos associats al projecte, incloent tant les despeses de personal com els materials necessaris. A més, es tindran en compte possibles imprevistos i costos indirectes. També es realitzarà un informe sobre l'impacte ambiental, social i econòmic del projecte per assegurar-ne la sostenibilitat. En total, es preveu una dedicació de **15 hores** per aquesta tasca.

- **GP4 - Seguiment del projecte**

  El seguiment del projecte és essencial per garantir que es compleixen els objectius i els terminis establerts. Es realitzaran reunions quinzenals amb el supervisor per revisar l'evolució del projecte, resoldre possibles problemes i ajustar la planificació si cal. Aquestes trobades també permetran rebre orientació i millorar la presa de decisions. S'estima que es destinaran **20 hores** a aquesta fase.

2.1.2 TP - Treball Previ

El Treball Previ constitueix la base teòrica i tècnica del projecte. Aquesta fase s'articula en tres parts fonamentals:

- **TP1 - Estudi sobre sistemes de recomanació**

  Es realitzarà una revisió exhaustiva de la literatura existent per comprendre els principis bàsics, les metodologies actuals i les tendències en el desenvolupament de sistemes de recomanació. Aquesta anàlisi permetrà identificar els punts forts i les limitacions dels enfocaments tradicionals. Es calcula que aquesta tasca requerirà aproximadament **25 hores**.

- **TP2 - Estudi sobre clustering en sistemes de recomanació**

  S'analitzaran les diferents tècniques de clustering aplicades en entorns de recomanació, valorant com la seva implementació pot millorar la densitat de dades i reduir la complexitat computacional. L'objectiu és entendre l'impacte d'aquestes tècniques en la precisió i eficiència del sistema. Es calcula que aquesta tasca requerirà aproximadament **25 hores**.

- **TP3 - Preparació de l'entorn de treball**

  Consistirà en la configuració i comprovació de totes les eines i llibreries necessàries (com Python, Jupyter Notebook, Git, LaTeX, etc.) per assegurar que el desenvolupament posterior es realitzi en un entorn òptim i estable. Es calcula que aquesta tasca requerirà aproximadament **5 hores**.

2.1.3 SD - Selecció i tractament del conjunt de dades

Aquest apartat es centra en la gestió dels conjunts de dades que serviran de base per als experiments:

- **SD1 - Recol·lecció de datasets**

  S'identificaran i obtindran conjunts de dades reals que representin diversos escenaris en sistemes de recomanació, tenint en compte diferents volums i nivells de dispersió. Aquesta tasca és clau per garantir que els experiments reflecteixin situacions reals. Es calcula que aquesta tasca requerirà aproximadament **15 hores**.

- **SD2 - Creació de datasets sintètics**

  Paral·lelament, es generaran conjunts de dades sintètics que permetin simular escenaris específics. D'aquesta manera, es podran avaluar els algorismes en entorns controlats, facilitant la comparació dels resultats i l'anàlisi dels comportaments davant diferents configuracions. Es calcula que aquesta tasca requerirà aproximadament **25 hores**.

2.1.4 SR - Desenvolupament del sistema de recomanació

En aquesta fase es durà a terme la implementació del sistema de recomanació mitjançant diferents mètodes de clustering. Aquesta etapa constitueix el nucli del projecte, ja que el correcte desenvolupament de les tècniques determinarà la qualitat dels resultats obtinguts.

A cada mètode se li assignarà un temps estimat de desenvolupament segons la seva complexitat i la necessitat d'ajustament de paràmetres. Es seguirà un esquema comú per a totes les implementacions, per garantir la comparabilitat dels resultats i la coherència en el codi.

- **SR1 - Implementació amb clustering dur**

  Aquesta fase consistirà en la implementació amb clustering dur amb un algoritme com K-Means, un mètode de clustering dur on cada element es classifica de manera exclusiva en un clúster. Es realitzarà una anàlisi prèvia per determinar el nombre òptim de clústers utilitzant mètriques com l'Elbow Method o la Silhouette Score. Tindrà una duració estimada de **30 hores**.

- **SR2 - Implementació amb clustering difús**

  Es desenvoluparà la implementació amb clustering difús amb un mètode com Fuzzy C-Means, que permet una assignació parcial d'un element a diferents clústers amb un cert grau de pertinença. Aquest enfocament és especialment útil per a dades amb transicions difuses entre categories. S'estima una duració de **40 hores**.

- **SR3 - Implementació amb clustering jeràrquic**

  El clustering jeràrquic permet estructurar les dades en una jerarquia de clústers mitjançant l'agrupació ascendent (agglomerative) o descendent (divisive). Aquesta implementació inclourà l'elecció de la mesura de distància i el mètode d'enllaç per a la creació dels clústers. Tindrà una duració estimada de **40 hores**.

- **SR4 - Implementació amb clustering basat en densitat**

  Amb algoritmes com DBSCAN identificarà regions d'alta densitat de punts, permetent detectar patrons sense necessitat de definir prèviament el nombre de clústers. Això el fa ideal per a conjunts de dades amb estructures no lineals. Tindrà una duració estimada de **25 hores**.

- **SR5 - Implementació amb clustering basat en models**

  Es desenvoluparà una implementació amb un algoritme com el de Mixtures Gaussiana (GMM), que assumeix que les dades provenen de diverses distribucions gaussianes. Aquesta tècnica proporcionarà una estimació probabilística de l'assignació de cada punt a un clúster. S'estima una duració de **40 hores**.

2.1.5 AR - Comparació de resultats i anàlisi del rendiment

Aquesta fase està dedicada a l'avaluació i comparació dels algorismes implementats. Amb índexs com el MAE (Mean Absolute Error) o el RMSE (Root Mean Square Error) per quantificar la precisió de les recomanacions generades. I també mesurar l'eficiència computacional i el temps d'execució de cada tècnica, contrastant els resultats obtinguts. Aquesta anàlisi permetrà identificar l'algorisme que ofereix el millor equilibri entre precisió i rendiment en diferents escenaris. Es preveu dedicar **40 hores** a aquesta tasca.

2.1.6 D - Documentació

Per a l'avaluació del treball és necessària la entrega d'una memòria final perquè aquest sigui avaluat. La documentació s'anirà redactant durant tot el procés del treball. Aquest recull tots els aspectes del projecte, incloent-hi la introducció, els objectius, la metodologia, els resultats obtinguts i les conclusions. Es calcula que aquesta tasca requerirà aproximadament **60 hores** degut a que és bastant extensa.

2.1.7 DT - Defensa del treball

Per a defensar el treball davant del tribunal és necessari preparar unes diapositives i un guió que recullin tots els punts claus treballats. A més a més, caldrà realitzar una sèrie d'assaigs. S'estima una duració de **15 hores**.

2.2 Recursos

2.2.1 Recursos humans

En aquest projecte es distingiran quatre rols clau: Cap del projecte, Investigador, Desenvolupador i Analista. Tenint en compte que el projecte es realitza de manera individual, l'autor del treball assumirà els diferents rols en funció de les tasques a realitzar. A la taula 2.1 es detallen els rols assignats a cada tasca.

- **Cap del projecte (J):** Responsable de la gestió global del projecte, incloent-hi la planificació, el seguiment i la coordinació de les tasques. També s'encarregarà de l'elaboració de la documentació final i de la defensa del treball davant del tribunal.
- **Investigador (I):** Encarregat de la recerca bibliogràfica, de l'estudi de les tècniques de clustering i de la seva aplicació en sistemes de recomanació.
- **Desenvolupador (D):** Responsable de la implementació del sistema de recomanació amb els diferents algorismes de clustering. Aquest rol se centrarà en la programació.
- **Analista (A):** Encarregat de l'anàlisi dels resultats obtinguts, incloent-hi la comparació de les diferents tècniques de clustering i l'avaluació de la precisió i de l'eficiència computacional de cada mètode.

A més a més, es comptarà amb el suport del director del projecte, que actuarà com a guia i consultor durant el desenvolupament del treball, i amb el tutor de GEP, que s'encarregarà de donar suport en la gestió del projecte durant el primer mes.

2.2.2 Recursos materials

Els recursos materials inclouen totes les eines, tant de programari com de maquinari, necessàries per al desenvolupament del projecte.

- **Recursos de maquinari:** S'utilitzarà un ordinador portàtil amb una capacitat de processament suficient per dur a terme les tasques de programació i d'anàlisi de dades, així com per a la investigació i la redacció de la documentació.

- **Recursos de programari:** Es farà servir un conjunt d'eines com Python, Jupyter Notebook, Matplotlib, NumPy, Pandas, Scikit-learn, Git, LaTeX i Notion. Aquestes eines permetran la programació, la visualització de dades, la gestió de versions i la redacció de la documentació.

2.3 Taula resum de les tasques

A continuació, es presenta una taula resum 2.1 de totes les tasques amb la seva descripció, durada, dependències, recursos i rols assignats.

ID	Tasca	Temps (h)	Recursos	Dep.	Rols
GP	**Gestió del Projecte**	**65**	**-**	**-**	**-**
GP1	Contextualització	15	Portàtil, LaTeX	TP	J
GP2	Planificació temporal	15	Portàtil, LaTeX	GP1	J
GP3	Sostenibilitat	15	Portàtil, LaTeX	GP2	J
GP4	Seguiment del projecte	20	-	-	J, I, D, A
TP	**Treball Previ**	**55**	**-**	**-**	**-**
TP1	Sis. de Recomanació	25	Portàtil	-	I
TP2	Clustering	25	Portàtil	TP1	I
TP3	Entorn de treball	5	Portàtil, Python	TP2	I
SD	**Selecció dades**	**40**	**-**	**-**	**-**
SD1	Recol·lecció de datasets	15	Portàtil, Python	GP3	A
SD2	Creació de datasets	25	Portàtil, Python	SD1	A
SR	**Sis. de recomanació**	**175**	**-**	**-**	**-**
SR1	Clustering dur	30	Portàtil, Python	SD	D
SR2	Clustering difús	40	Portàtil, Python	SR1	D
SR3	Clustering jeràrquic	40	Portàtil, Python	SR2	D
SR4	Clustering densitat	25	Portàtil, Python	SR3	D
SR5	Clustering models	40	Portàtil, Python	SR4	D
AR	**Anàlisi resultats**	**40**	**Portàtil, Python**	**SR**	**A**
D	**Documentació**	**60**	**Portàtil, LaTeX**	**-**	**J, I, D, A**
DT	**Defensa del treball**	**15**	**Portàtil, LaTeX**	**D**	**J**
-	**Total**	**450**	**-**	**-**	**-**

Taula 2.1: Taula resum de les tasques del projecte
Llegenda de rols: J - Cap del projecte, I - Investigador, D - Desenvolupador, A - Analista

2.4 Diagrama de Gantt

A continuació, es mostra la figura 2.1 amb el diagrama de Gantt de la planificació temporal del projecte. Aquest inclou totes les tasques definides anteriorment, agrupades per colors, juntament amb les seves dependències i durades.

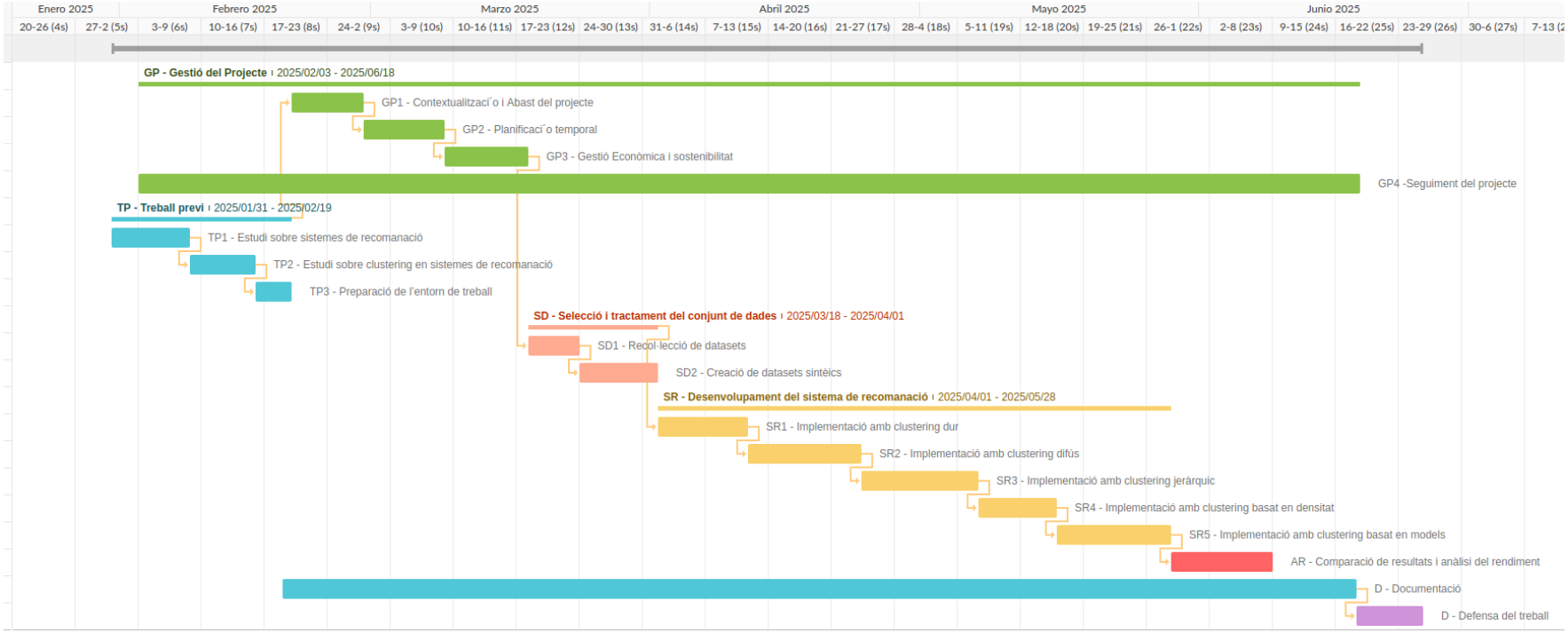

Figura 2.1: Diagrama de Gantt creat amb GanttPro [10]

2.5 Gestió del risc

Tots els projectes impliquen certs riscos que cal avaluar per garantir una cobertura eficient. La gestió de riscos és un element clau en qualsevol iniciativa, ja que permet identificar, analitzar i classificar els riscos per reduir-ne els possibles efectes negatius. En aquest sentit, es detallen els principals riscos que podrien afectar el projecte i les estratègies de mitigació. A la taula 2.2 es resumeix la probabilitat d'ocurrència i una valoració qualitativa sobre com cada risc podria influir en el desenvolupament del projecte.

Risc	**Probabilitat**	**Impacte**
Falta de coneixement	Alta	Mitjà
Retards en la planificació	Mitjana	Elevat
Dificultats tècniques	Alta	Mitjà
Problemes amb els equips	Baixa	Elevat
Complexitat computacional	Mitjana	Mitjà

Taula 2.2: Taula de riscos del projecte

- **Falta de coneixement:** Aquest risc fa referència a la manca de coneixements previs en sistemes de recomanació i clustering, que podria dificultar el desenvolupament del projecte. Per mitigar aquest risc, s'ha reservat temps per a la recerca bibliogràfica i l'aprenentatge de les tècniques necessàries durant la planificació del projecte.
- **Retards en la planificació:** Aquest risc es refereix a la possibilitat que les tasques no s'executin segons el calendari establert, provocant endarreriments en la finalització del projecte. Per evitar aquest risc, es programaran reunions de seguiment periòdiques per avaluar l'estat d'execució de les tasques, es revisarà periòdicament el diagrama de Gantt i es reservaran marges temporals per absorbir possibles retards.
- **Dificultats tècniques:** En desenvolupar codi, és molt probable que sorgeixin imprevistos, com ara errors inesperats, incompatibilitats de llibreries o problemes de rendiment. Per resoldre aquests problemes, es reservarà temps extra per al desenvolupament de cada tècnica.
- **Problemes amb els equips:** Aquest risc contempla la possibilitat de fallades o inestabilitat en el maquinari utilitzat durant el desenvolupament i els experiments. Si fos necessari, es disposarà d'equips de recanvi per minimitzar la interrupció de la feina.
- **Complexitat computacional:** Algunes tècniques de clustering, especialment aquelles basades en models o clustering difús, poden requerir un alt consum de recursos computacionals, afectant el temps d'execució i la resposta del sistema. Per mitigar aquest risc, es poden optimitzar les implementacions o emprar hardware més potent, sol·licitant-lo a la universitat si cal.

Capítol 3

Gestió Econòmica i sostenibilitat

3.1 Gestió Econòmica

Després d'establir la planificació temporal de la iniciativa, es procedeix a l'estimació dels costos requerits per al seu desenvolupament. S'identifiquen diverses categories de despeses, incloent-hi les associades al personal, l'espai de treball i les eines i dispositius emprats. A més, per afrontar possibles contratemps i cobrir despeses no previstes, s'elabora un pla de contingència, es defineix una partida per a imprevistos i s'estableixen mecanismes per al control pressupostari.

3.1.1 Costos de personal

A partir de la planificació de tasques, es calcula el cost de personal, tenint en compte els quatre rols definits anteriorment: Cap de projecte, Investigador, Desenvolupador i Analista. Els costos anuals s'obtenen de la web de PayScale, que ofereix informació sobre salaris en la indústria, suposant menys d'un any d'experiència. El salari anual de cada rol es divideix pel nombre d'hores anuals de treball per obtenir el cost per hora, suposant 1.760 hores en total, tenint en compte vacances i festius. Aquestes dades es mostren a la taula 3.1.

Rol	**Cost anual**	**Cost per hora**
Cap del projecte	46.303€[11]	46.303/1.760 = 26,30€/h
Investigador	34.000€[12]	34.000/1.760 = 19,32€/h
Desenvolupador	21.237€[13]	21.237/1.760 = 12,07€/h
Analista	24.439€[14]	24.439/1.760 = 13,89€/h

Taula 3.1: Costos de personal

ID	Tasca	Temps (h)	Rols	Cost (€)	Cost SS (€)
GP	**Gestió del Projecte**	**65**	**-**	**1541,4**	**2003,82**
GP1	Contextualització	15	J	394,5	512,85
GP2	Planificació temporal	15	J	394,5	512,85
GP3	Sostenibilitat	15	J	394,5	512,85
GP4	Seguiment del projecte	20	J, I, D, A	357,9	465,27
TP	**Treball Previ**	**55**	**-**	**1062,6**	**1381,38**
TP1	Sis. de recomanació	25	I	483	627,9
TP2	Clustering	25	I	483	627,9
TP3	Entorn de treball	5	I	96,6	125,58
SD	**Selecció dades**	**40**	**-**	**555,6**	**722,28**
SD1	Recol·lecció de datasets	15	A	208,35	270,855
SD2	Creació de datasets	25	A	347,25	451,425
SR	**Sis. de recomanació**	**175**	**-**	**2112,25**	**2745,925**
SR1	Clustering dur	30	D	362,1	470,73
SR2	Clustering difús	40	D	482,8	627,64
SR3	Clustering jeràrquic	40	D	482,8	627,64
SR4	Clustering densitat	25	D	301,75	392,275
SR5	Clustering models	40	D	482,8	627,64
AR	**Anàlisi resultats**	**40**	**A**	**555,6**	**722,28**
D	**Documentació**	**60**	**J, I, D, A**	**1073,7**	**1395,81**
DT	**Defensa del treball**	**15**	**J**	**394,5**	**512,85**
-	**Total**	**450**	**-**	**7295,65**	**9484,345**

Taula 3.2: Costos de personal per tasca.
Llegenda de rols: J - Cap del projecte, I - Investigador, D - Desenvolupador, A - Analista

A la taula 3.2 es mostren els costos de personal per tasca, on s'ha calculat el cost total de cada tasca en funció dels rols que hi intervenen i el temps que hi dediquen. En les tasques on el rol que la realitzi podria ser qualsevol, s'ha suposat el cost mitjà per hora dels rols.

El cost total de personal és de 7.295,65 €, mentre que el cost total de personal amb seguretat social, suposant un 30 %, és de 9.484,35 €.

3.1.2 Costos genèrics

Tot el software utilitzat en el projecte és de codi obert, per tant, no hi ha cap cost associat a l'adquisició de llicències. El cost total de software és de 0 €.

Per realitzar el projecte, s'utilitza un ordinador portàtil ASUS amb un preu de 699€. Es calcula l'amortització tenint en compte que a l'any hi ha 220 dies laborables i 8 hores laborables al dia, segons la següent fórmula:

Es considera una vida útil de 5 anys.

$$\text{Amortització} = \frac{\text{Preu ordinador}}{\text{Dies laborables} \times \text{Hores laborables} \times \text{Vida útil}} \times \text{Hores d'ús} \quad (3.1)$$

$$\text{Amortització} = \frac{699}{220 \times 8 \times 5} \times 450 = 35,74€ \quad (3.2)$$

El projecte també té associats una sèrie de costos indirectes:

- L'electricitat té un cost mitjà de 0,215€/kWh. El portàtil té un consum de 45W. Suposant que el portàtil estarà encès 430 hores, el cost de l'electricitat és de 0,215€/kWh * 0,045kW * 430h = 4,16€.
- Internet té un cost de 40€/mes. Suposant que el projecte duri 5 mesos, el cost total és de 40€/mes * 5 mesos = 200€.
- El desplaçament per a reunions de seguiment i altres té un cost de 44€ de manera trimestral, així que el cost total és de 44€ * 2 = 88€.
- El projecte es desenvolupa íntegrament a l'habitatge propi de l'autor, però si hagués de llogar l'habitació, el preu aproximat seria de 400€/mes. Suposant que el projecte duri 5 mesos, el cost total és de 400€/mes * 5 mesos = 2000€.

A la taula 3.3 es mostra un resum dels costos genèrics del projecte. El cost total de costos genèrics és de 2.277,9 €.

Concepte	**Cost (€)**
Amortització	35,74
Electricitat	4,16
Internet	150
Desplaçament	88
Lloc de treball	2000
Total	**2277,9**

Taula 3.3: Costos genèrics

3.1.3 Contingències

En qualsevol iniciativa, és essencial incorporar un marge addicional per afrontar possibles contratemps i imprevistos. En aquest context, atès que es tracta d'un estudi basat en tecnologies emergents, hi ha una alta probabilitat d'enfrontar-se a desafiaments durant la seva execució. Per això, s'ha determinat establir un marge del 15 % per cobrir aquests costos addicionals.

El cost de contingència es calcula amb la fórmula següent:

$$\text{Cost contingència} = (\text{Cost total de personal} + \text{Cost total genèric}) \times 0,15 \tag{3.3}$$

$$\text{Cost contingència} = (9484,35 + 2277,9) \times 0,15 = 1764,34€ \tag{3.4}$$

3.1.4 Imprevistos

És fonamental tenir en compte les despeses derivades dels eventuals contratemps que puguin sorgir al llarg de l'execució d'aquest projecte. Com s'ha indicat prèviament, és important establir mesures per mitigar els possibles riscos associats a la realització de la tasca.

Per determinar el cost dels imprevistos, s'aplica la fórmula següent:

$$\text{Cost imprevist} = \text{Cost} \times \frac{\text{Probabilitat (\%)}}{100} \tag{3.5}$$

Imprevist	**Cost (€)**	**Prob. (%)**	**Rol**	**Cost imprevist (€)**
Increment desenvolupament (25 h)	392,275	20	D	78,455
Nou portàtil	699	5	-	34,95
Total	**-**	**-**	**-**	**113,405**

Taula 3.4: Costos imprevistos. Llegenda de rols: D - Desenvolupador

A la taula 3.4 es mostren els costos dels imprevistos, on s'ha calculat el cost total de cada imprevist en funció de la probabilitat que succeeixi.

El risc de necessitar més temps de desenvolupament és alt, per la qual cosa s'estima una probabilitat del 20% i un cost de 25 hores de desenvolupament. El risc de necessitar un nou portàtil és baix, per la qual cosa s'estima una probabilitat del 5% i el cost d'un portàtil.

3.1.5 Cost total

A continuació es presenta una taula resum 3.5 amb el cost total del projecte. El cost total del projecte és de 13.639,99 €.

Tipus	**Cost**
Cost personal	9484,345
Cost genèrics	2277,9
Cost de contingència	1764,34
Cost imprevistos	113,405
Cost total	**13639,99**

Taula 3.5: Cost total del projecte

3.1.6 Control de gestió

El control de gestió té com a objectiu principal supervisar, avaluar i ajustar de manera contínua totes les operacions del projecte per tal d'assegurar el compliment dels objectius establerts. Per aconseguir-ho, és imprescindible definir indicadors quantitatius que permetin mesurar el rendiment i detectar desviacions respecte al pressupost, el cronograma i els recursos planificats.

A continuació, es presenten alguns dels indicadors i mètriques que s'utilitzaran:

1. Desviació cost personal per tasca:

$$(\text{cost\_estimat} - \text{cost\_real}) * \text{hores\_reals}$$

2. Desviació realització tasques:

$$(\text{hores\_estimades} - \text{hores\_reals}) * \text{cost\_real}$$

3. Desviació total realització tasca:

$$\text{cost\_estimat} - \text{cost\_real}$$

4. Desviació cost genèric:

$$\text{cost\_estimat\_genèric} - \text{cost\_real\_genèric}$$

5. Desviació cost imprevistos:

$$\text{cost\_imprevistos\_estimat} - \text{cost\_imprevistos\_real}$$

6. Desviació hores del projecte:

$$\text{hores\_estimades} - \text{hores\_reals}$$

Aquest control es durà a terme de manera periòdica mitjançant reunions bisetmanals, on es revisaran i actualitzaran els indicadors mitjançant una fulla de càlcul. En acabar cada tasca, es compararan les dades reals amb les estimades. A més, es registraran de forma detallada tots els despeses extra derivats d'imprevistos i contingències, de manera que qualsevol desviació, especialment aquella superior al 5 pugui ser identificada i corregida a temps.

En cas de detectar desviacions negatives (per exemple, un cost superior al previst o retards en el cronograma), es prendran mesures correctives com ara:

- Augmentar el pressupost mitjançant el fons reservat per a contingències.
- Ajustar el temps assignat o reassignar tasques per optimitzar el consum d'hores.
- Reduir o modificar parts del projecte per minimitzar els efectes dels desajustos detectats.

Aquest enfocament proactiu permetrà no només identificar ràpidament qualsevol desviació, sinó també actuar de manera eficaç per mantenir el projecte dins dels paràmetres planificats i assegurar-ne l'èxit global.

3.2 Sostenibilitat

Al llarg dels meus estudis en el Grau d'Enginyeria Informàtica a la FIB, he tingut l'oportunitat d'aprofundir en els aspectes clau de la sostenibilitat a través de diverses assignatures, reconeixent la seva importància durant el desenvolupament del TFG.

La sostenibilitat, entesa des d'una perspectiva integral que inclou les dimensions econòmica, social i ambiental, és un factor essencial en qualsevol projecte, tant si és de gran com de petita escala. A nivell personal, sempre he considerat fonamental l'anàlisi econòmica per garantir la viabilitat d'un projecte, ja que sovint es busca millorar l'eficiència econòmica d'una solució existent.

A més, aquest treball m'ha permès reflexionar sobre la importància d'abordar la sostenibilitat de manera global. He observat que, en moltes ocasions, es dona més pes als beneficis econòmics en detriment de les dimensions social i ambiental, un fet que considero crucial per assolir un impacte positiu i sostenible en el temps.

Finalment, crec que tots els projectes s'haurien de desenvolupar amb l'objectiu de cobrir una necessitat real de la societat.

3.2.1 Dimensió econòmica

El projecte s'ha plantejat amb una acurada estimació de costos, on s'han identificat i analitzat de manera detallada les diferents parts de personal, despeses genèriques, contingències i imprevistos. Així, en el PPP s'estableix un pressupost realista (aproximadament 13.640 €) que permet debatre tant la viabilitat econòmica del projecte com l'adequada assignació de recursos.

Pel que fa a la vida útil i a la resolució actual dels aspectes de costos del problema, l'estat de l'art aposta pel ús de tecnologies de codi obert i models d'anàlisi que optimitzen la gestió de dades. No obstant això, moltes de les solucions vigents no aborden completament el repte del cost computacional derivat del processament de grans volums de dades, aspecte que es compensa en el nostre enfocament mitjançant l'aplicació de tècniques de clustering que redueixen la dispersió i milloren l'eficiència.

Finalment, la solució proposada ofereix una millora econòmica significativa en comparació amb els mètodes existents. En incorporar el clustering com a pas previ al filtratge col·laboratiu, s'optimitzen els temps de processament i es redueix el consum de recursos computacionals, cosa que es tradueix en menys costos operatius i en una major rendibilitat i sostenibilitat a llarg termini.

3.2.2 Dimensió ambiental

Atès que es tracta d'un projecte de recerca i investigació dels diferents mètodes de clustering en sistemes recomanadors, l'impacte ambiental associat prové principalment del consum d'electricitat, sobretot si aquesta electricitat es genera a partir de fonts d'energia no renovables. El càlcul del consum elèctric que s'ha determinat correspon a l'electricitat consumida pel portàtil.

Al reutilitzar el portàtil, s'evita la producció de residus electrònics i es redueix l'impacte ambiental associat a la fabricació de nous dispositius. A més, la durada de vida útil del portàtil es veu ampliada gràcies a la seva correcta gestió i manteniment, la qual cosa contribueix a minimitzar el consum de recursos naturals i a reduir l'emissió de gasos contaminants.

En l'estat de l'art actual les solucions aborden el problema principal des d'un punt de vista tecnològic, però sovint no integren de manera completa criteris ambientals. La nostra proposta millora ambientalment les solucions existents, ja que no només optimitza els processos i redueix el consum de recursos, sinó que també incorpora mesures específiques per disminuir l'emissió de residus i l'ús energètic, contribuint així a una reducció global de l'impacte ambiental.

3.2.3 Dimensió social

La realització del projecte posat en producció suposa, a nivell personal, una gran oportunitat per créixer tant professionalment com humanament. A més d'ampliar els meus coneixements en àrees com la intel·ligència artificial i l'anàlisi de dades, em permetrà desenvolupar habilitats en la gestió i coordinació de projectes reals, fet que enriquirà el meu perfil professional.

D'altra banda, actualment el problema que es vol abordar es resol amb sistemes de recomanació basats en mètodes tradicionals, com el filtratge col·laboratiu, que tot i ser útils, presenten limitacions pel que fa a la personalització i l'eficiència. La solució que proposo pretén millorar socialment la qualitat de vida dels usuaris, oferint recomanacions més ajustades a les seves necessitats i interessos, fet que afavoreix una interacció més rellevant i satisfactòria.

inalment, existeix una necessitat real d'aquest projecte, ja que l'increment exponencial de la informació digital exigeix noves eines que superin les limitacions dels mètodes actuals. Això no sols permetrà una millor gestió dels recursos, sinó que també contribuirà a millorar la qualitat de vida social a través d'una experiència d'usuari més adaptada i eficient.

Capítol 4

Conjunt de dades

Una selecció acurada dels conjunts de dades és fonamental per a l'avaluació rigorosa i la comparació de les tècniques de clustering aplicades als sistemes de recomanació. Aquest capítol detalla els datasets emprats en aquest estudi, escollits per la seva rellevància en la comunitat d'investigació de sistemes de recomanació i per representar una varietat de dominis, mides i nivells de dispersió. L'ús de múltiples datasets permet analitzar la robustesa i la capacitat de generalització dels mètodes de clustering en diferents contextos.

4.1 Restriccions i requisits dels conjunts de dades

Amb l'objectiu de facilitar la comparació i la reproduïbilitat dels experiments, s'ha definit una sèrie de restriccions i requisits que han de complir els conjunts de dades utilitzats. Aquestes condicions tenen com a finalitat garantir una certa homogeneïtat entre els conjunts i, alhora, millorar la qualitat i la interpretabilitat dels resultats obtinguts.

En primer lloc, s'exigeix que les dades estiguin emmagatzemades en format CSV. Aquest format ha estat escollit perquè és àmpliament utilitzat, fàcilment llegible tant per humans com per màquines, i compatible amb la majoria de biblioteques i entorns de programació emprats habitualment en l'anàlisi de dades. A més, el format CSV afavoreix una estructura tabular clara, la qual facilita el processament i la validació automàtica dels conjunts de dades.

Aquest arxiu contindrà exactament tres columnes amb els següents noms, que inclouran la informació següent:

- **userId:** Identificador únic de cada usuari del sistema de recomanació.
- **itemId:** Identificador únic de cada element o producte que es recomana.
- **rating:** Valoració atorgada per l'usuari a l'element.

Tant **userId** com **itemId** es representaran com a enters de 64 bits, i **rating** com un valor de tipus float de 64 bits.

Cap de les tres columnes pot contenir valors no numèrics o NaN; en cas que se'n detectin, es descartaran.

No es permetrà que hi hagi parells **(userId, itemId)** repetits. Si se'n troben, es mantindrà l'última valoració afegida per aquest parell.

Com a restricció tècnica, s'imposarà un màxim d'uns 10.000 usuaris. Aquesta limitació es deu a l'ús de matrius per emmagatzemar la similitud entre usuaris, amb un creixement quadràtic en funció del nombre d'usuaris. L'espai necessari ve donat per la fórmula següent:

$$\text{Nombre d'usuaris}^2 \times 8 \text{ bytes}$$

Els 8 bytes corresponen a la representació de nombres amb comes flotants de 64 bits per expressar la similitud. Assumint 10.000 usuaris, això suposa un total de 10^8 posicions, amb un requeriment d'uns 800 MB, la qual cosa comença a representar un desafiament per executar els experiments en local.

Una solució senzilla seria calcular la similitud entre usuaris de manera dinàmica, només quan sigui necessària, en lloc de calcular-les totes prèviament i emmagatzemar-les. Tanmateix, aquesta estratègia implica un increment significatiu del temps de càlcul, que pot resultar inviable per a una execució local. Per aquest motiu, s'ha descartat aquesta opció.

De manera similar a l'anterior, s'imposarà un màxim d'uns 10.000 elements. Aquesta limitació també es deu a l'ús de matrius per emmagatzemar la valoració entre usuari i element, la qual cosa implica un creixement proporcional al producte entre el nombre d'usuaris i el nombre d'elements.

En aquest cas, l'espai necessari ve donat per:

$$\text{Nombre d'usuaris} \times \text{Nombre d'elements} \times 8 \text{ bytes}$$

Els 8 bytes corresponen a la representació de nombres amb comes flotants de 64 bits per expressar la valoració. De la mateixa manera que amb la matriu de similituds, l'espai comença a ser significatiu i pot comprometre el rendiment en entorns amb recursos limitats.

Una manera de reduir dràsticament l'espai utilitzat per aquesta matriu és fer ús de matrius disperses. El motiu d'això és el fet que la majoria d'elements no solen estar valorats per l'usuari i, per tant, la gran majoria de les posicions de la matriu contindran valors nuls o buits.

L'ús de matrius disperses permet emmagatzemar només les posicions que contenen informació rellevant (és a dir, valoracions efectives), reduint dràsticament l'espai requerit.

Tot i això, finalment s'ha descartat l'ús de matrius disperses, ja que, malgrat l'estalvi d'espai, aquest enfocament incrementa considerablement la complexitat del codi i el temps de càlcul. La gestió de matrius disperses implica estructures de dades més sofisticades i operacions més costoses, especialment en operacions massives o en entorns on es requereix un accés ràpid i freqüent a les valoracions. Per aquest motiu, s'ha optat per mantenir l'enfocament basat en matrius densament poblades, tot i la limitació del nombre màxim d'elements.

4.2 Conjunt de dades utilitzades

4.2.1 Conjunt de dades MovieLens

El conjunt de dades MovieLens [15] és un dels més utilitzats en la recerca sobre sistemes de recomanació. Proporcionat pel GroupLens Research Project, aquest conjunt inclou valoracions d'usuaris sobre pel·lícules i s'ha convertit en un estàndard de facto per a l'avaluació d'algorismes de filtratge col·laboratiu.

Aquest conjunt recull valoracions amb una escala de fins a 5 estrelles del servei de recomanació de pel·lícules MovieLens.

En aquest estudi s'han emprat dos subconjunts del conjunt de dades MovieLens: el MovieLens Latest Small i el MovieLens 1M.

El MovieLens Latest Small és un subconjunt recent que conté 100.836 valoracions sobre 9.724 pel·lícules, realitzades per 610 usuaris entre el 29 de març de 1996 i el 24 de setembre de 2018.

A partir d'una anàlisi exploratòria de les dades (Figura 4.1), s'observa que aquest conjunt presenta una densitat de l'1,7 %, fet que indica que només l'1,7 % de les possibles valoracions estan registrades al conjunt.

$$\text{Densitat} = \frac{\text{Nombre de valoracions}}{\text{Nombre d'usuaris} \times \text{Nombre de pel·lícules}} \times 100$$

Les valoracions són predominantment positives, amb una mitjana de 3,5 estrelles i una desviació estàndard d'1,04. A més, les valoracions per usuari i pel·lícula tenen una distribució fortament sesgada a la dreta. El sistema té una majoria d'usuaris i pel·lícules amb poques valoracions, mentre que una petita fracció d'usuaris i pel·lícules concentra una gran quantitat de valoracions.

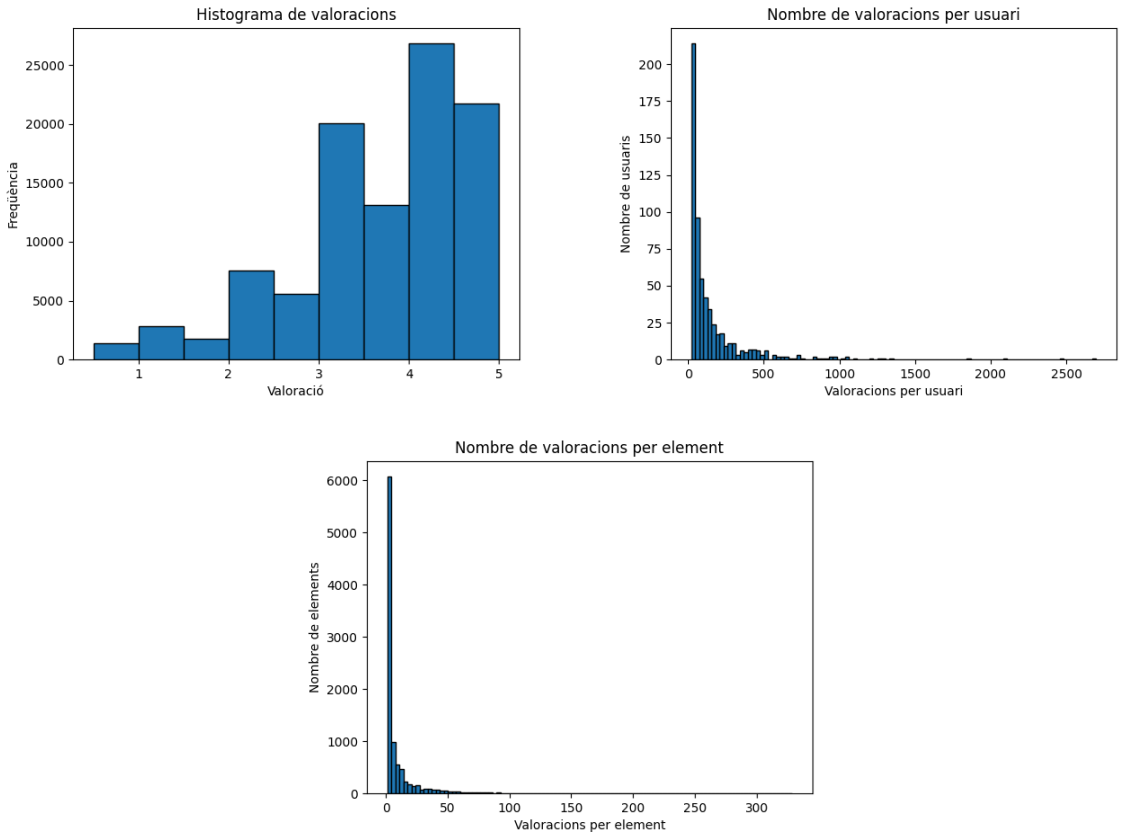

Figura 4.1: Anàlisi exploratòria del conjunt MovieLens Latest Small

D'altra banda, el MovieLens 1M és un conjunt de dades més gran que conté 1.000.209 valoracions de 6.040 usuaris sobre 3.706 pel·lícules, recollides entre els anys 2000 i 2003.

A partir de l'anàlisi exploratòria (Figura 4.2), s'observa una densitat del 4,5 %, és a dir, només el 4,5 % de les possibles valoracions estan recollides en el conjunt.

També en aquest cas, les valoracions són majoritàriament positives, amb una mitjana de 3,58 estrelles i una desviació estàndard d'1,12. De manera similar al conjunt MovieLens Latest Small, les valoracions per usuari i pel·lícula presenten una distribució fortament sesgada a la dreta, però amb una densitat més alta que en el conjunt anterior.

Cal destacar que les valoracions en aquest conjunt no inclouen puntuacions fraccionàries (com 0,5, 1,5, 2,5, etc.), ja que el sistema original de valoració utilitzava només valors enters de 1 a 5 estrelles.

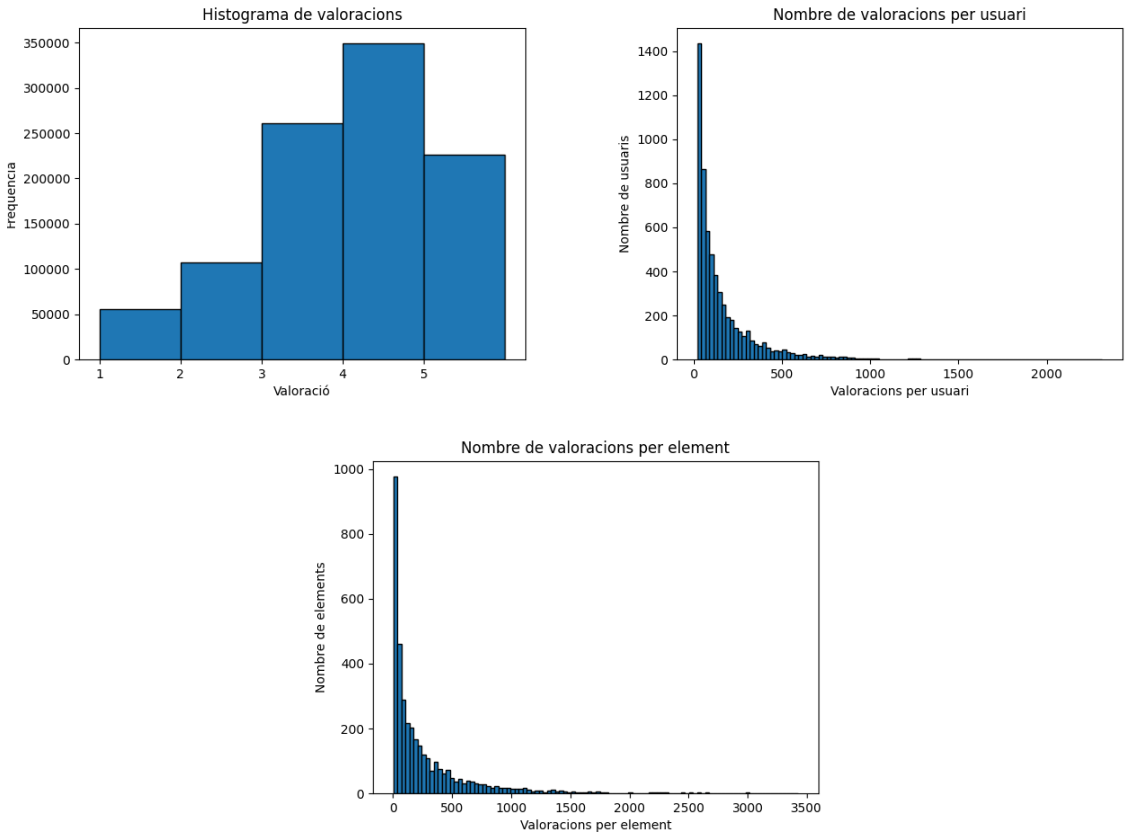

Figura 4.2: Anàlisi exploratòria del conjunt MovieLens 1M

4.2.2 Conjunt de dades Book-Crossing

El conjunt de dades *Book-Crossing* [16] prové de la comunitat en línia *BookCrossing.com*, una plataforma destinada a amants de la lectura d'arreu del món que intercanvien llibres i comparteixen experiències. Aquest conjunt ha estat àmpliament utilitzat en estudis sobre sistemes de recomanació, especialment en escenaris que combinen valoracions explícites i implícites.

Les dades es van recollir durant un període de quatre setmanes (agost/setembre de 2004), recopilant informació sobre 278.858 usuaris i 1.157.112 valoracions corresponents a 271.379 llibres diferents (identificats mitjançant ISBN).

Les valoracions s'expressen en una escala de fins a 10 estrelles, segons el sistema de recomanació del servei *Book-Crossing*.

Dels 278.858 usuaris inicials, només una fracció significativa ha realitzat valoracions:

- 99.053 usuaris han valorat almenys un llibre.
- 43.385 usuaris han valorat almenys dos llibres.
- 12.306 usuaris han valorat almenys deu llibres.

Pel que fa als 271.379 llibres presents al conjunt:

- 270.171 llibres han estat valorats almenys una vegada.
- 124.513 llibres han rebut almenys dues valoracions.
- 17.480 llibres han rebut almenys deu valoracions.

Degut a l'elevada dispersió del conjunt original, s'ha realitzat un procés de filtratge per obtenir resultats més significatius en l'avaluació d'algorismes de filtratge col·laboratiu. Primer, s'han seleccionat els usuaris que han valorat almenys 10 llibres; a continuació, s'ha mantingut només aquells llibres que han rebut un mínim de 10 valoracions dins aquest subconjunt.

Aquest procés ha generat un conjunt de dades més compacte i manejable, compost per 74.907 valoracions, 6.570 usuaris i 3.411 llibres.

A partir d'una anàlisi exploratòria de les dades (Figura 4.3), s'observa que aquest conjunt presenta una densitat del 0,33 %, la qual cosa indica que només el 0,33 % de les valoracions possibles estan registrades.

Les valoracions són predominantment positives, amb una mitjana de 7,87 estrelles i una desviació estàndard d'1,77. A més, com ja es detectava en conjunts com *MovieLens*, tant les valoracions per usuari com per llibre mostren una distribució amb una llarga cua a la dreta: la majoria d'usuaris i llibres tenen poques valoracions, mentre que una petita fracció concentra un nombre elevat de puntuacions.

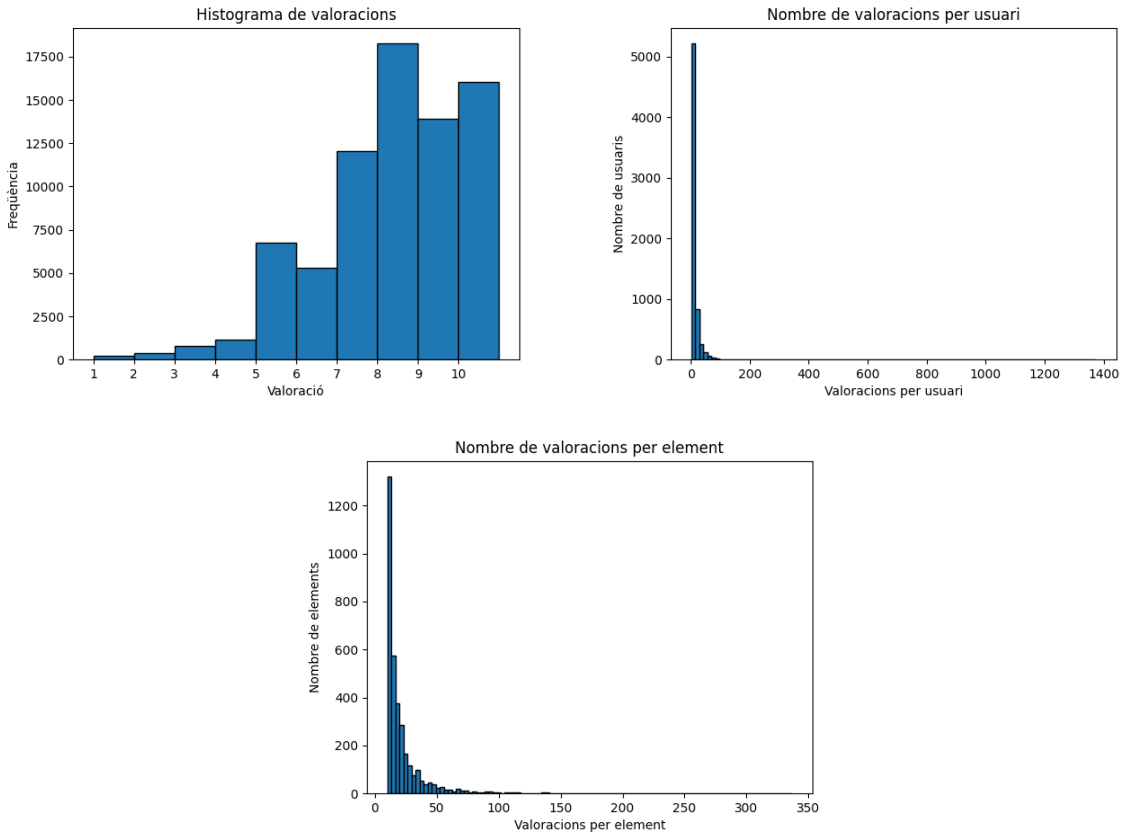

Figura 4.3: Anàlisi exploratòria del conjunt Book-Crossing

4.2.3 Conjunt de dades Jester

El conjunt de dades Jester [17, 18] prové d'un sistema de recomanació d'acudits, en què els usuaris valoren els acudits en una escala contínua de -10 a +10. Tot i el seu caràcter informal, aquest conjunt ha estat àmpliament utilitzat per la seva densitat i per una estructura peculiar que el fa especialment útil en l'àmbit de la recerca en sistemes de recomanació.

El conjunt original conté 100 acudits i 73.421 usuaris, amb un total de 4,1 milions de valoracions recollides entre abril de 1999 i maig de 2003.

Per reduir la mida del conjunt i facilitar l'anàlisi, s'ha aplicat un filtre per conservar únicament 10.000 usuaris seleccionats aleatòriament.

Una de les característiques més destacades d'aquest conjunt és la seva elevada densitat: cada usuari ha valorat un nombre considerable d'ítems, cosa que permet observar amb més claredat el comportament dels mètodes de clustering en entorns amb una gran quantitat d'informació disponible.

A partir d'una anàlisi exploratòria de les dades (Figura 4.4), s'observa una densitat del 56,27 %, és a dir, s'han registrat més de la meitat de les valoracions possibles. Es tracta d'un valor molt elevat en comparació amb altres conjunts de dades de sistemes de recomanació, que habitualment presenten una cobertura molt més escassa.

Les valoracions mostren una tendència generalment positiva, amb una mitjana de 0,72 i una desviació estàndard de 5,3. Tot i que la majoria d'usuaris només valoren una part dels acudits, s'observa un pic significatiu d'usuaris que han valorat gairebé la totalitat dels acudits disponibles. Pel que fa als acudits, les valoracions es distribueixen de manera relativament uniforme, amb dos pics: alguns acudits han estat valorats per la majoria d'usuaris, mentre que d'altres només han rebut valoracions d'una minoria.

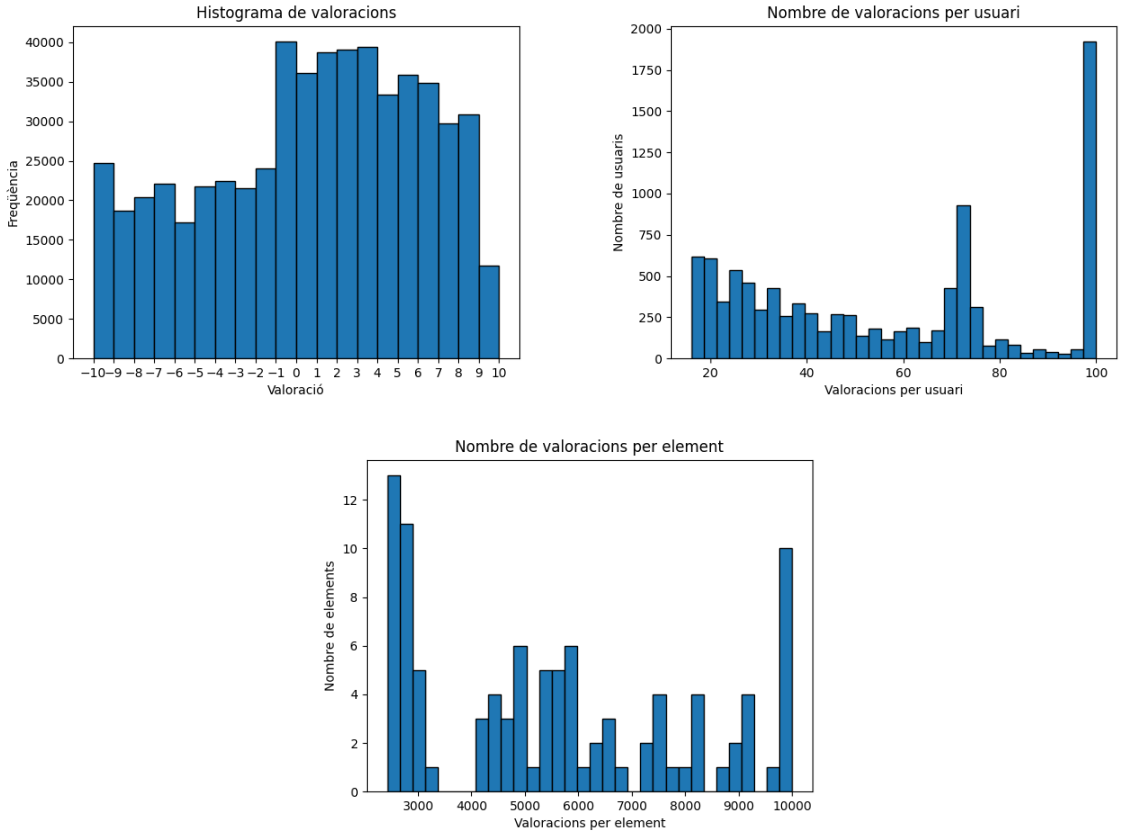

Figura 4.4: Anàlisi exploratòria del conjunt Jester

A continuació es presenta una taula resum amb les característiques dels conjunts de dades utilitzats en aquest estudi (Taula 4.1).

Conjunt	**Usuaris**	**Elements**	**Valoracions**	**Densitat (%)**	**Escala**
MovieLens Small	610	9.724	100.836	1,7	0.5-5
MovieLens 1M	6.040	3.706	1.000.209	4,5	1-5
Book-Crossing	6.570	3.411	74.907	0,33	1-10
Jester	10.000	100	4.134.000	56,27	-10-10

Taula 4.1: Característiques dels conjunts de dades utilitzats

4.3 Conjunts de dades sintètiques

El nombre i la varietat de conjunts de dades de valoracions disponibles públicament són sovint limitats, especialment en àmbits menys convencionals.

Les empreses que poden recopilar conjunts de dades de valoracions solen ser reticents a compartir-los, per por de vulnerar la privadesa dels seus usuaris o d'exposar informació comercialment sensible als seus competidors.

A causa d'aquesta escassetat de dades públiques, els professionals han començat a dependre de valoracions sintètiques.

Tanmateix, els resultats obtinguts d'aquests experiments poden ser qüestionables, ja que els conjunts de dades generats sovint no són capaços de captar adequadament les característiques pròpies d'un domini d'interès concret.

En aquest estudi, ens proposem analitzar la capacitat de generalització dels mètodes de clustering en situacions especialment extremes pel que fa a la densitat de les valoracions.

Per aconseguir aquests conjunts de dades peculiars, utilitzem un conjunt de dades de referència com a punt de partida, el qual codifica les particularitats d'un domini específic. Aquest mètode generatiu permet crear diversos conjunts de dades de valoracions amb usuaris que mostren comportaments similars als dels usuaris del conjunt de dades original. Tanmateix, aquests usuaris sintètics no tenen cap relació directa amb persones reals, de manera que no es compromet cap informació privada ni comercialment sensible.

Sostenim que confiar únicament en unes poques distribucions estadístiques calculades empíricament a partir d'un conjunt de dades existent, o definides per un investigador, no és suficient per simular de manera realista els gustos individuals dels éssers humans. Mètodes d'aquest tipus generen conjunts de dades amb usuaris sense preferències pròpies, fet que dificulta enormement el funcionament efectiu de qualsevol sistema de recomanació.

Així doncs, utilitzarem un enfoc similar al de l'artice [19] amb un enfocament basat en clustering per generar conjunts de dades sintètiques a partir de conjunts de dades reals, amb l'objectiu d'obtenir conjunts de dades.

Aquesta generació de conjunts sintètics es basa en un procés en dues fases: una fase d'anàlisi del conjunt de referència i una fase de mostreig de valoracions.

Fase 1: Anàlisi del conjunt de dades

En primer lloc, s'aplica una anàlisi detallada del conjunt de dades de referència per obtenir una representació fidel del domini. L'element clau d'aquesta fase és la identificació de comunitats d'usuaris mitjançant el mètode de clustering K-means. Per capturar millor la variabilitat de les valoracions, en lloc d'utilitzar una representació binària (valoració positiva o negativa), s'ha optat per una aproximació que considera cada valor de puntuació per separat. Això vol dir que el procés es repeteix per a cada possible valor de valoració (per exemple, de 1 a 5), generant representacions específiques

per a cada cas.

Per a cada valor de valoració, es representa cada usuari com un vector binari d'interaccions amb els ítems: un 1 indica que l'usuari ha assignat aquell valor concret a l'ítem, i un 0 indica absència d'aquesta valoració. Aquesta representació permet aplicar K-means per agrupar usuaris amb patrons similars de puntuació en cada nivell.

A partir d'aquesta agrupació, es construeixen tres distribucions empíriques per a cada valor de valoració:

- $\mathbf{P_C^r}$: Distribució de probabilitat que un usuari pertanyi a un determinat clúster per a la valoració r.
- $\mathbf{P_U^{k,r}}$: Distribució del nombre de valoracions amb valor r per usuari dins del clúster k.
- $\mathbf{P_I^{k,r}}$: Distribució de probabilitat d'interaccions amb valoració r per ítem dins del clúster k.

Aquestes distribucions es construeixen comptant les freqüències observades dins cada comunitat i per cada valor de valoració, reflectint així el comportament col·lectiu detallat dels usuaris.

Fase 2: Generació de valoracions

Un cop obtingudes les distribucions, es procedeix a generar el nou conjunt de dades mitjançant un procés de mostreig estocàstic, repetit per a cada valor de valoració r.

Per cada valor r del conjunt de puntuacions:

1. Per cada usuari sintètic, se l'assigna a un clúster k segons la distribució $\mathbf{P_C^r}$.
2. Es determina el nombre de valoracions amb valor r mitjançant un mostreig de $\mathbf{P_U^{k,r}}$.
3. Es seleccionen ítems a valorar amb puntuació r, mitjançant un mostreig sense reemplaçament de la distribució $\mathbf{P_I^{k,r}}$.

Després de repetir aquest procés per a tots els valors possibles, es fusionen les valoracions generades per construir el conjunt de dades final, que conté valoracions multinivell i reflecteix una estructura rica i fidel al conjunt original.

Aquesta metodologia assegura que els conjunts sintètics reflecteixin una estructura latent realista, en què els usuaris tenen gustos diferenciats i les seves interaccions segueixen patrons similars als del conjunt de referència. Això fa que els conjunts generats siguin adequats per a l'avaluació comparativa d'algorismes de clustering i sistemes de recomanació.

Un cop definit i validat el mètode de generació de conjunts de dades sintètiques, el següent pas és escollir el conjunt de dades de referència a partir del qual es generarà. En aquest estudi, s'ha seleccionat el conjunt MovieLens Latest Small, un dels

més utilitzats en la recerca sobre sistemes de recomanació, conegut per la seva estructura rica i diversa. A partir d'aquest conjunt, es generaran tres conjunts de dades sintètiques. El primer actuarà com a conjunt base o de control, reproduint de manera fidel les característiques del conjunt original. En canvi, en el segon i tercer conjunts es modificarà manualment la distribució $\mathbf{P_U^{k,r}}$ per tal de crear situacions amb densitats de valoració extremadament baixes i altes, respectivament. Aquest disseny experimental permet avaluar la robustesa i la capacitat de generalització dels mètodes de clustering en escenaris amb densitats diverses.

4.3.1 MovieLens sintètic (base)

El conjunt de dades MovieLens sintètic (base) es genera a partir del conjunt MovieLens Latest Small, amb l'objectiu de reproduir les mateixes característiques i patrons de valoració que el conjunt original. Això permet establir un punt de referència sòlid per a la comparació amb altres conjunts generats.

A partir de l'anàlisi exploratòria del conjunt MovieLens sintètic (base) (Figura 4.5), s'observa que aquest presenta una densitat del 2,38 %, similar a la del conjunt original.

La corba de distribució de les valoracions mostra una tendència comparable a la del conjunt original, tot i que amb una distribució més uniforme. La mitjana de les valoracions és de 3,3 estrelles, amb una desviació estàndard d'1,16.

A més, la distribució de valoracions per usuari i per pel·lícula és també semblant a la del conjunt original, tot i que amb una pèrdua notable de la seva llarga cua característica.

En resum, degut a la naturalesa estocàstica del procés de generació, el conjunt sintètic manté l'estructura general del conjunt original, però amb una variabilitat lleugerament més baixa.

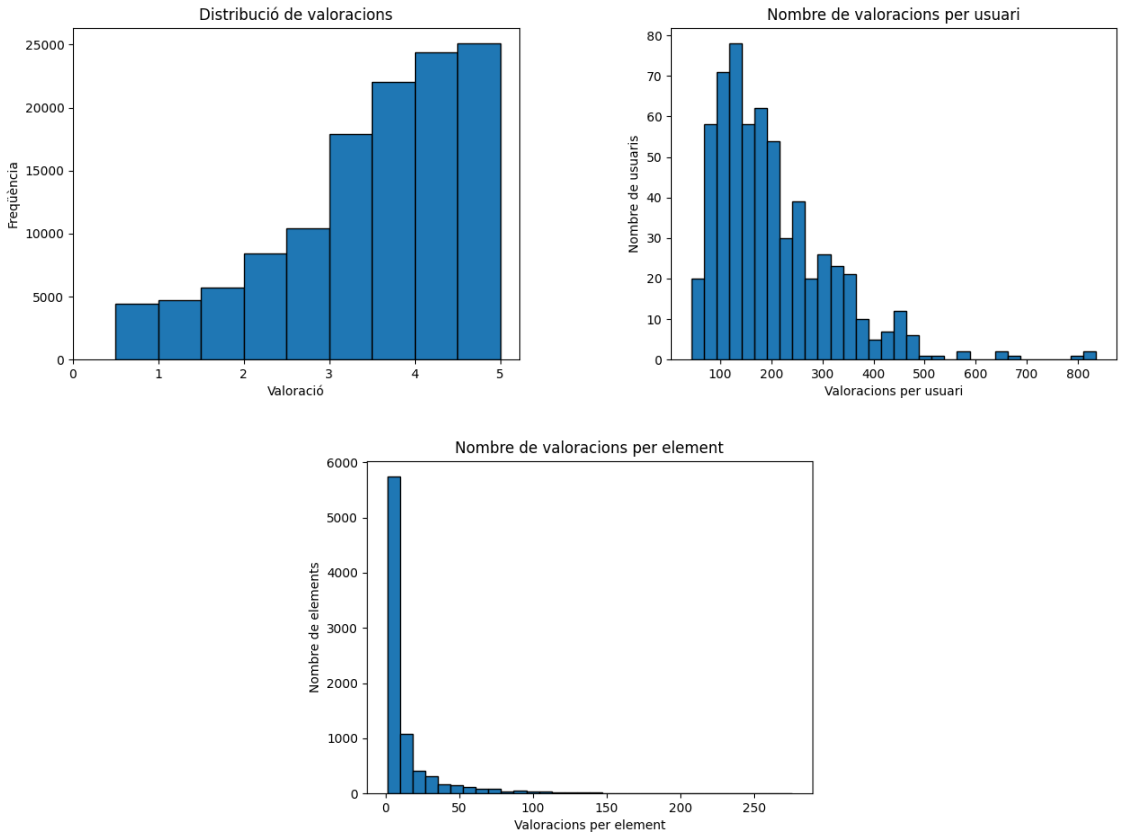

Figura 4.5: Anàlisi exploratòria del conjunt MovieLens sintètic (base)

4.3.2 MovieLens sintètic (baixa densitat)

Per generar aquest conjunt, s'ha aplicat una reducció del 98 % a la distribució $\mathbf{P}_{\mathbf{U}}^{\mathbf{k,r}}$ amb l'objectiu de disminuir la quantitat de valoracions per usuari. Aquesta reducció dràstica provoca una disminució significativa de la densitat del conjunt, que passa a ser del 0,29 %.

En l'anàlisi (Figura 4.6), s'observen propietats similars, amb una reducció evident en la quantitat de valoracions per usuari i per pel·lícula. La mitjana de les valoracions és de 3,4 estrelles, amb una desviació estàndard d'0,92. Es pot observar una clara diferenciació entre la distribució de les valoracions en valoracions majors que 3 estrelles i les inferiors. També s'observa una reducció en el nombre de valoracions per pel·lícula i per usuari.

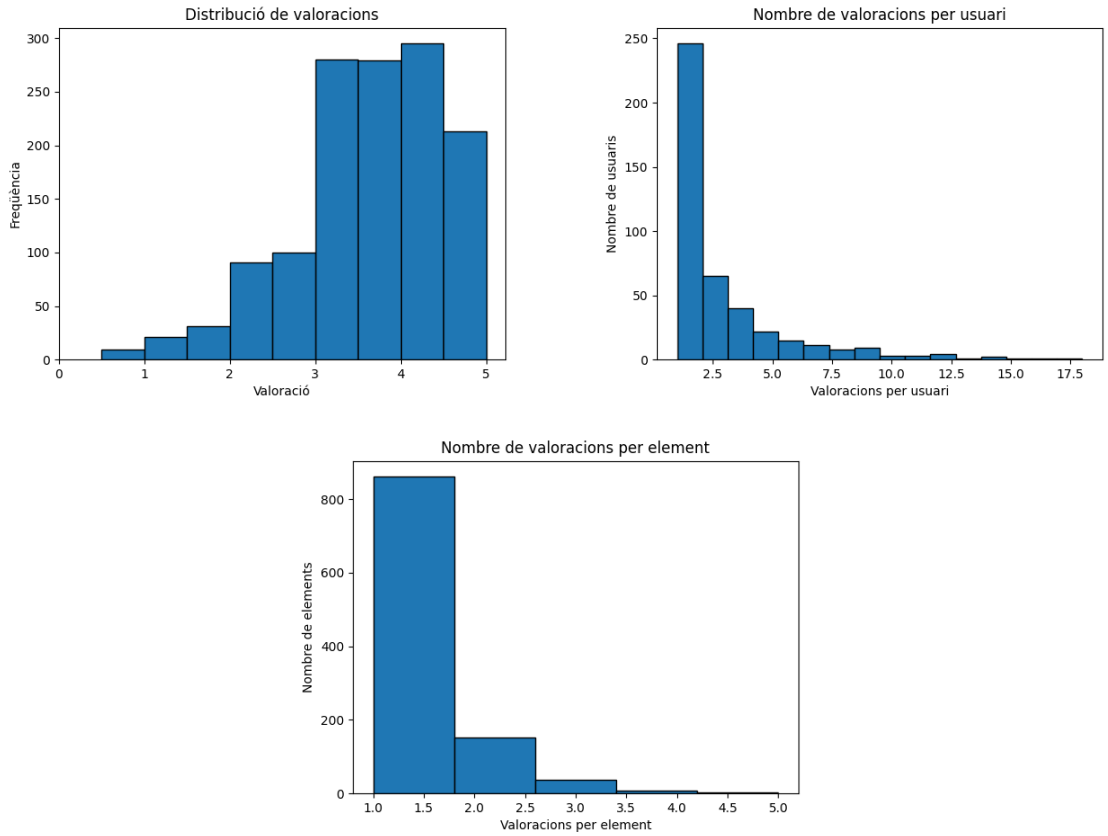

Figura 4.6: Anàlisi exploratòria del conjunt MovieLens sintètic (baixa densitat)

4.3.3 MovieLens sintètic (alta densitat)

En generar aquest conjunt, s'ha assolit una densitat del 83,9 %, fet que representa un increment significatiu respecte al conjunt original.

En l'anàlisi (Figura 4.7), es poden observar canvis notables en la distribució de les valoracions. Aquest fenomen es deu al fet que s'ha forçat que cada usuari sintètic valori un gran nombre d'ítems, incloent-hi molts que probablement no li agradin. Això contradiu la naturalesa del conjunt original, en què els usuaris tendeixen a valorar només aquells ítems que els resulten d'interès.

A conseqüència d'això, la mitjana de les valoracions se situa en 2,39 estrelles, amb una desviació estàndard d'1,21. A més, la distribució de les valoracions per usuari i per pel·lícula s'inverteix respecte a les anteriors, mostrant una gran concentració de valoracions per cada usuari i per cada pel·lícula.

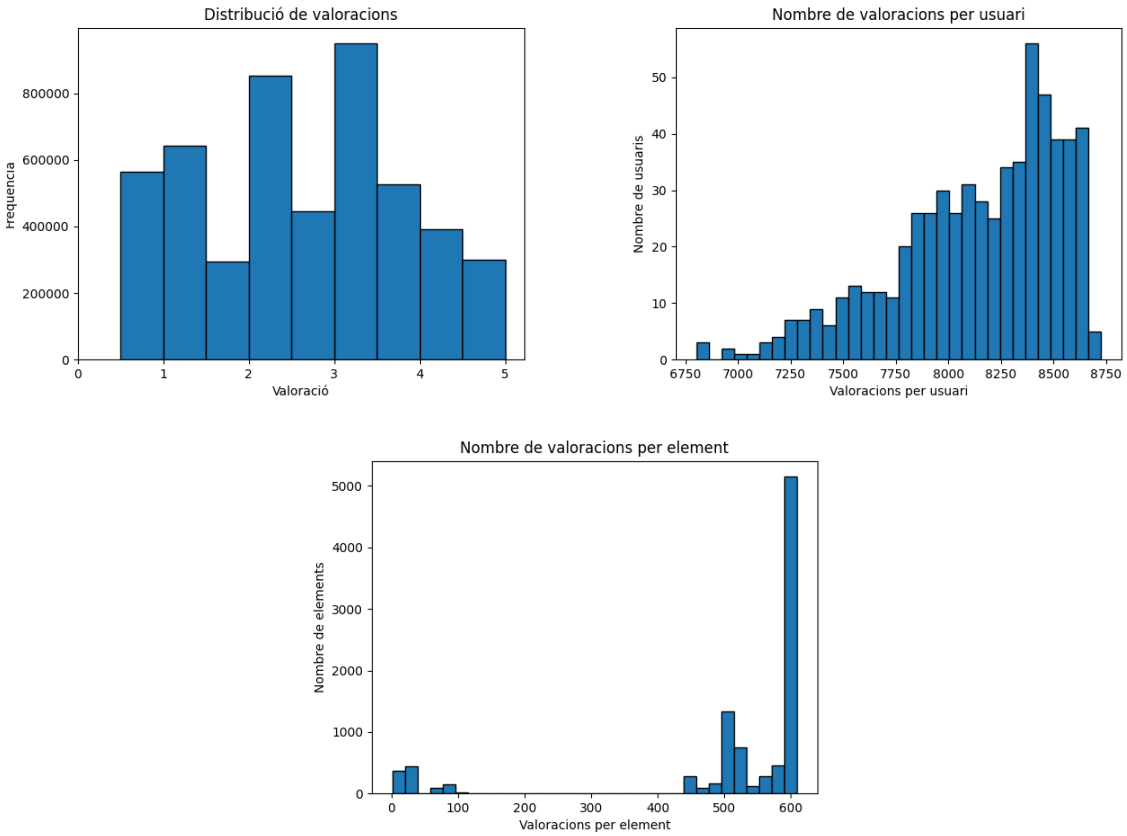

Figura 4.7: Anàlisi exploratòria del conjunt MovieLens sintètic (alta densitat)

A continuació es presenta una taula resum amb les característiques dels conjunts de dades sintètiques generats (Taula 4.2).

Conjunt	**Usuaris**	**Elements**	**Valoracions**	**Densitat (%)**	**Escala**
Sintètic (base)	610	8.481	123.033	2,38	0.5-5
Sintètic (baixa densitat)	432	1.060	1.319	0,29	0.5-5
Sintètic (alta densitat)	610	9.694	4.963.951	83,9	0.5-5

Taula 4.2: Característiques dels conjunts de dades sintètiques generats

Encara que s'utilitzin dades sintètiques, es donarà més importància a les que provenen d'interaccions reals d'usuaris. Tot i així, aquestes dades s'empraran per analitzar el comportament de les implementacions en situacions anòmales.

Capítol 5

Sistema de recomanació

Arquitectura comuna del sistema de recomanació

Aquest capítol descriu el sistema de recomanació implementat per a l'avaluació dels diferents mètodes de clustering considerats en aquest estudi. Amb l'objectiu de garantir una comparació justa entre les diferents variants, s'ha establert una arquitectura comuna que comparteix tant la funció de similitud com la metodologia de predicció. Això permet que qualsevol diferència en el rendiment dels sistemes sigui atribuïble únicament a la tècnica de clustering emprada.

Per mesurar la similitud entre usuaris, s'ha optat per la correlació de Pearson. Aquesta és una mesura de similitud fonamental en els sistemes de recomanació i compta amb una àmplia literatura que en recolza l'eficàcia [20].

Aquesta mesura és àmpliament utilitzada en sistemes de recomanació basats en el filtratge col·laboratiu, ja que captura la correlació lineal entre els patrons de valoració dels usuaris, tot tenint en compte les diferències en les seves escales personals de valoració. Donats dos usuaris u i v, i el conjunt d'elements que ambdós han valorat I_{uv}, la correlació de Pearson s'expressa de la forma següent:

$$\text{sim}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \bar{r}_u)^2} \cdot \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \bar{r}_v)^2}}$$

on $r_{u,i}$ és la valoració de l'usuari u sobre l'element i, i $\bar{r}_u$ és la mitjana de les valoracions de l'usuari u. Aquesta mesura ofereix una normalització implícita respecte a les mitjanes individuals i afavoreix els casos on els usuaris valoren de forma coherent, encara que utilitzin escales diferents.

La correlació de Pearson oscil·la típicament entre -1 i +1, on -1 indica una correlació negativa forta, +1 una correlació positiva forta, i 0 implica absència d'associació. Visualment, una correlació positiva es manifesta quan els valors augmenten conjuntament, mentre que una correlació negativa apareix quan disminueixen conjuntament.

Per a la predicció de valoracions, s'utilitza la tècnica coneguda com a predicció centrada en la mitjana o basada en la desviació respecte a la mitjana. Aquesta meto-

dologia es basa en ajustar la predicció a partir de la mitjana de l'usuari i les desviacions observades en altres usuaris similars. La predicció $\hat{r}_{u,i}$ que fa l'usuari u sobre l'element i es calcula com:

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u)} \text{sim}(u,v) \cdot (r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u)} |\text{sim}(u,v)|}$$

on $N(u)$ és el conjunt d'usuaris més similars a u que han valorat l'element i, i $\text{sim}(u,v)$ és la similitud entre els usuaris u i v, segons la correlació de Pearson. De manera similar a la mesura de similitud, la predicció centrada en la mitjana també té en compte les escales personals de valoració, cosa que permet obtenir estimacions més precises i comparables entre usuaris.

Segons l'anàlisi empírica presentada a [21], la formulació basada en la desviació respecte a la mitjana resulta ser una de les estratègies de predicció no personalitzada més precises, superant altres enfocaments habituals. Això reforça l'ús d'aquest model com a base de comparació robusta en aquest estudi.

L'ús conjunt de la similitud de Pearson i de la predicció centrada en la mitjana com a base comuna respon a diversos motius. En primer lloc, són tècniques consolidades i àmpliament adoptades en la literatura sobre sistemes de recomanació. A més, proporcionen una base robusta per analitzar l'impacte específic del clustering en el rendiment del sistema, i faciliten la comparació directa entre mètodes, evitant que les diferències puguin atribuir-se a variacions en la funció de similitud o en la fórmula de predicció.

Per dur a terme el procés de clustering és imprescindible definir prèviament una funció de distància que mesuri quant "allunyats" estan dos punts (en el nostre cas, dos usuaris) dins l'espai de característiques. Aquesta funció de distància actua com a nucli de qualsevol mètode de partició ja que depenen íntegrament de la mètrica que comparem.

Per seguir garantint que qualsevol variació en el rendiment obtingut es degui només a la mecànica pròpia de l'algorisme de clustering hem optat per experimentar de forma consistent les mateixes dues mesures de distància en tots els procediments d'agrupament. D'aquesta manera, la comparació entre tècniques és estrictament "com a resultat de l'estructura" que el clustering imposa, i no d'algun component extern a l'algorisme.

La primera mètrica de distància que emprem és la distància euclidiana, probablement la més clàssica i intuïtiva en l'anàlisi de dades. Donats dos vectors de característiques $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$, la seva distància euclidiana ve donada per:

$$d_{Euc}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Aquesta fórmula mesura directament la separació geomètrica entre punts en un espai n-dimensional, ponderant igualment totes les dimensions.

Els seus punts forts són la seva simplicitat i la seva interpretabilitat, que reflecteix clarament la magnitud total del desplaçament.

Per incorporar la noció de "coherència de patró" en les valoracions, utilitzarem a més una distància basada en la similitud de Pearson. Aquesta segona mètrica permet agrupar usuaris que, tot i fer servir escales de puntuació diferents, mostren un comportament sincronitzat (pujades i baixades conjuntes). Hipotetitzem que aquesta donarà millors resultats, ja que esta relacionada amb la funció de predicció utilitzada.

En una primera instancia es va implementar la mesura de distància proposada a [22]:

$$d_{Pearson}(x, y) = \frac{1}{1 + \text{sim}(x, y)}$$

Tanmateix, es va acabar descartant aquesta mesura perquè no compleix les propietats d'una mètrica de distància (identitat, simetria i desigualtat triangular). Tot i que es va considerar utilitzar la distància angular, finalment es va optar per implementar la transformació:

$$d_{Pearson}(x, y) = \frac{1 - \text{sim}(x, y)}{2}$$

Ja que és més senzilla de calcular i compleix les propietats requerides d'una mètrica

Algorismes de clustering emprats

A grans trets, els algorismes de clustering més comuns, i que tractarem en aquest treball, es poden agrupar en cinc famílies principals segons com entenen i construeixen els grups. [23]

En primer lloc, tenim el clustering dur, que divideix el conjunt de dades en k grups mútuament excloents. Entre els algorismes més coneguts d'aquesta família hi ha el K-means, K-means++, Mini-Batch K-means i K-medoids, entre d'altres. En aquest treball ens centrarem en el K-means, ja que, a més de ser àmpliament utilitzat en entorns pràctics, destaca per la seva simplicitat conceptual i eficiència computacional, la qual cosa el converteix en un bon punt de partida per introduir el clustering dur.

A continuació, trobem el clustering difús, que permet que un mateix element pugui pertànyer a més d'un clúster amb un grau de pertinença associat. Els algorismes més representatius d'aquesta categoria són el Fuzzy C-means (FCM), Possibilistic C-means (PCM) i Gustafson–Kessel (GK). En aquest treball ens centrarem en el FCM, atès que és àmpliament emprat en problemes on es busca gestionar la incertesa o l'ambigüitat en la classificació, com ara en el reconeixement de patrons.

El tercer tipus és el clustering jeràrquic, que construeix una jerarquia d'agrupacions mitjançant un arbre o dendrograma. Els mètodes més coneguts d'aquesta família són els algorismes aglomeratius i divisoris. En aquest treball ens centrarem en l'aglomeratiu, ja que ofereix una major flexibilitat en l'elecció de la mesura de distància i permet

visualitzar fàcilment l'estructura de les dades, cosa que el fa especialment útil per a anàlisis exploratòries.

El quart enfocament és el clustering basat en densitat, que identifica agrupacions com a regions de major densitat separades per àrees de baixa densitat. Els algorismes més coneguts d'aquesta família són DBSCAN, OPTICS i HDBSCAN. En aquest treball posarem l'accent en DBSCAN, perquè destaca per la seva capacitat per detectar formes arbitràries de clústers i gestionar bé el soroll, cosa que el fa adequat per a dades reals amb estructures no lineals.

Finalment, trobem el clustering basat en models, que assumeix que les dades provenen d'una combinació de distribucions estadístiques subjacents. Els algorismes més habituals en aquesta categoria són el Gaussian Mixture Model (GMM) i el Dirichlet Process Mixture Model (DPMM). En aquest treball analitzarem el GMM, ja que ofereix una aproximació probabilística robusta i flexible que permet modelar clústers amb formes el·líptiques i proporciona una estimació explícita de la probabilitat de pertinença de cada punt.

A continuació es presenta una taula resum amb els algorismes de clustering analitzats (Taula 5.1).

Algorisme	**Tipus**	**Complexitat**
K-means	Dur	$O(n)$
Fuzzy C-means	Difús	$O(n)$
Aglomeratiu	Jeràrquic	$O(n^2)$
DBSCAN	Densitat	$O(n \cdot \log n)$
GMM	Model	$O(n)$

Taula 5.1: Algorismes de clustering analitzats

5.1 Clustering dur

Tal com es fa en altres treballs [24], els usuaris s'agrupen exclusivament en un únic clúster mitjançant l'algorisme K-means. Aquesta tècnica de partició busca dividir l'espai d'usuaris en K grups disjunts, de manera que cada usuari pertanyi al clúster amb el centre (centroid) més proper.

L'algorisme K-means funciona de la manera següent:

1. **Inicialització:** Es trien K punts inicials com a centroids, habitualment de manera aleatòria o mitjançant una estratègia com K-means++ per millorar la convergència.
2. **Assignació d'usuaris:** Cada usuari, representat com un vector de característiques (per exemple, el vector de valoracions normalitzat), s'assigna al clúster amb el centroid més proper, mesurat normalment amb la distància euclidiana.
3. **Reajustament de centroids:** Per a cada clúster, es recalcula el centroid com la mitjana de tots els vectors d'usuaris assignats a aquell clúster.

4. **Iteració:** Els passos d'assignació i reajustament es repeteixen fins que els centroids deixen de moure's de manera significativa (convergeixen) o s'arriba a un nombre màxim d'iteracions.

Aquest procediment assegura que la suma dels quadrats de les distàncies internes als clústers (inertia) es minimitza localment, tot i que K-means no garanteix trobar el mínim global.

Un cop determinats els clústers, es calcula la predicció de les valoracions considerant només els veïns dins del mateix clúster que han valorat l'element corresponent.

Per a la implementació amb distància euclidiana s'ha fet ús de la biblioteca Python `scikit-learn` [25], que ofereix la classe `KMeans` només amb aquesta mesura. Com que no permet canviar la funció de distància, s'ha desenvolupat una implementació pròpia de l'algorisme K-means que utilitza la distància de Pearson definida prèviament. Aquesta versió s'ha optimitzat amb la biblioteca `numba` per a una major eficiència i ha estat paral·lelitzada per accelerar els càlculs.

5.2 Clustering difús

S'ha descartat la implementació proposada per [26], que feia ús del grau de pertinença i dels centroids per calcular les prediccions. Tot i que aquesta proposta aprofita millor les propietats del clustering difús, contradiu la nostra premissa de mantenir una fórmula de predicció comuna entre totes les variants. A més, en una anàlisi preliminar s'ha detectat un increment significatiu del temps de còmput i un empitjorament considerable dels resultats en comparació amb altres alternatives.

Seguint la proposta de [27], s'ha implementat un mètode de clustering difús per a la creació de clústers d'usuaris. Aquest mètode permet que cada usuari pugui pertànyer a més d'un clúster, amb un grau de pertinença que es determina mitjançant la distància entre l'usuari i els centroids dels clústers.

L'algorisme utilitzat és el Fuzzy C-Means (FCM), una extensió del K-means que introdueix la possibilitat que un mateix usuari tingui diversos graus de pertinença a diferents clústers, en lloc de ser assignat exclusivament a un sol grup. Aquesta característica reflecteix millor la naturalesa difusa de les preferències dels usuaris en sistemes de recomanació, on sovint comparteixen interessos amb múltiples comunitats.

El funcionament de l'algorisme FCM es pot descriure amb els passos següents:

1. **Inicialització:** Es tria un nombre de clústers K i un paràmetre de difusió $m > 1$ (habitualment $m = 2$), que controla el grau de "fuzziness". També s'inicialitza la matriu de pertinences U, on u_{ij} representa el grau de pertinença de l'usuari i al clúster j, amb valors entre 0 i 1 i la restricció que $\sum_{j=1}^{K} u_{ij} = 1$ per a tot i.
2. **Càlcul dels centroids:** Per a cada clúster j, es calcula el centroid c_j com una mitjana ponderada dels vectors d'usuaris, utilitzant els graus de pertinença

elevats a la potència m:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

on x_i és el vector de característiques de l'usuari i.

3. **Actualització dels graus de pertinença:** Es recalculen els valors de u_{ij} a partir de la distància entre l'usuari i i cada centroid c_j:

$$u_{ij} = \left(\sum_{k=1}^{K} \left(\frac{||x_i - c_j||}{||x_i - c_k||} \right)^{\frac{2}{m-1}} \right)^{-1}$$

   Aquest pas assegura que els usuaris més propers a un centroid tinguin un grau de pertinença més elevat.

4. **Iteració:** Es repeteixen els passos de càlcul de centroids i actualització de pertinences fins que es compleixi un criteri de convergència, com ara un canvi mínim en U o un nombre màxim d'iteracions.

Seguint la proposta adoptada, una vegada obtinguts els graus de pertinença dels usuaris a cada clúster, per obtenir una assignació final i discreta de cada usuari a un únic clúster, s'aplica la tècnica del *Centre de Gravetat (Center of Gravity)* sobre els graus de pertinença:

$$\text{COG}_i = \sum_{j=1}^{K} u_{ij} \cdot j, \quad \text{cluster}_i = \text{round}(\text{COG}_i)$$

Això permet capturar la distribució global de la pertinença de l'usuari i assignar-lo al clúster més representatiu segons la seva posició mitjana entre els clústers.

Un cop determinada l'assignació, es calculen les prediccions de valoracions considerant només els veïns del mateix clúster.

Per a la implementació de l'algorisme Fuzzy C-Means s'ha fet ús de la biblioteca Python `scikit-fuzzy` (`skfuzzy`). Com que aquesta implementació no permet definir una funció de distància pròpia, s'ha modificat manualment la llibreria per afegir l'opció d'utilitzar la distància de Pearson en el càlcul dels graus de pertinença i dels centroides. Aquestes adaptacions s'han integrat de manera nadiua en les funcions `cmeans` i `cmeans_predict` de `skfuzzy`, afegint un nou paràmetre `metric='pearson'`.

5.3 Clustering jeràrquic

El clustering jeràrquic és una tècnica d'agrupament que construeix una jerarquia de clústers, en lloc de generar una partició plana com fan algorismes com K-means o Fuzzy C-means.

Seguint altres treballs [28] i inspirant-nos en una proposta similar, en aquesta secció es descriu la implementació de l'algorisme d'Agrupament Jeràrquic Aglomeratiu (HAC), utilitzant el mètode d'enllaç promig (*average linkage*).

L'algorisme HAC parteix del supòsit que cada usuari constitueix un clúster individual. A cada iteració, es fusionen els dos clústers amb major similitud.

El mètode de fusió utilitzat ha estat l'*average linkage*, o Mètode No Ponderat de Grups amb Mitjanes Aritmètiques (UPGMA, per les seves sigles en anglès). Aquest enfocament calcula la distància entre dos clústers com la mitjana de les distàncies entre totes les parelles d'usuaris que els componen:

$$d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{u \in A} \sum_{v \in B} d(u, v)$$

Aquest mètode presenta avantatges respecte a altres com l'enllaç simple o complet, ja que produeix clústers amb formes més equilibrades i és menys sensible als valors atípics.

Així, en cada pas del procés d'agrupament aglomeratiu:

1. S'identifica la parella de clústers amb la distància d mínima.
2. Es fusionen en un nou clúster i es recalculen les distàncies amb la resta segons la fórmula anterior.
3. El procés es repeteix fins a obtenir el nombre desitjat de clústers K.

Un cop determinada l'assignació de clústers, les prediccions de valoracions es calculen considerant exclusivament els veïns que pertanyen al mateix clúster.

Per a la implementació de l'algorisme d'Agrupament Jeràrquic Aglomeratiu (HAC) s'ha fet ús de les funcions `linkage` i `fcluster` del mòdul `scipy.cluster.hierarchy`. Abans d'invocar `linkage`, les mesures de distància entre usuaris es precalculen en funció de si s'ha utilitzat la distància euclidiana o la distància de Pearson, i es transmeten a la funció mitjançant el paràmetre `metric='precomputed'`. Això permet aplicar el mateix procés de clustering sense dependre de la distància interna per defecte.

5.4 Clustering per densitat

Per a l'agrupament basat en densitat s'ha utilitzat l'algorisme DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), com també s'ha fet en altres treballs [29].

El funcionament de DBSCAN es basa en dos paràmetres principals:

- ε: radi del veïnat considerat per avaluar la densitat al voltant d'un punt.

- **MinPts**: nombre mínim de punts que han d'existir dins del radi ε perquè un punt pugui iniciar la formació d'un grup.

El funcionament de l'algorisme DBSCAN es pot resumir en els següents passos:

1. **Inicialització:** Es fixen els valors dels paràmetres ε (distància màxima per considerar dos punts com a veïns) i *MinPts* (nombre mínim de veïns per formar un grup).
2. **Avaluació dels punts:** Per a cada usuari, es compta quants altres usuaris es troben dins la distància ε.
   - Si un usuari té com a mínim *MinPts* veïns, es considera apte per iniciar un grup i unir-hi els punts propers.
   - Si en té menys, només podrà afegir-se a un grup si es troba dins del veïnat d'un altre usuari que sí compleixi el criteri anterior.
   - Els usuaris que no compleixen cap dels dos casos anteriors no s'inclouen en cap grup.
3. **Formació dels grups:** Per a cada usuari que pot formar grup:
   (a) Es crea un nou grup i s'hi afegeix aquest usuari.
   (b) Es continua afegint al grup qualsevol usuari proper que compleixi els criteris de densitat, així com els seus veïns, de manera recursiva.
4. **Finalització:** Quan tots els grups han estat formats, l'algorisme conclou. Els usuaris que no han estat assignats a cap grup en formen un de nou.

Un cop formats els grups, la predicció de valoracions es calcula seguint la mateixa lògica que en les tècniques anteriors: per a cada parella usuari–ítem, només es tenen en compte les valoracions d'altres usuaris que formen part del mateix grup.

Per a l'agrupament basat en densitat s'ha fet ús de la classe `DBSCAN` de la biblioteca `scikit-learn` [25]. Igual que en el cas jeràrquic, les distàncies entre usuaris es precalculen prèviament (euclidiana o Pearson) i es passen a l'algorisme mitjançant el paràmetre `metric='precomputed'`. D'aquesta manera es manté consistència en la mesura de similitud utilitzada.

5.5 Clustering per model

Per al clustering per model s'ha escollit l'ús del *Gaussian Mixture Model* (GMM), una tècnica probabilística que modela la distribució dels usuaris com una combinació de K distribucions gaussianes. A diferència de K-means, que realitza una partició dura, GMM permet un modelatge de pertinença suau, obtenint per a cada usuari responsabilitats (*responsibilities*) associades a cada component gaussià. Aquesta aproximació ja ha estat utilitzada en sistemes de recomanació col·laboratius, com es mostra a [30].

L'algorisme GMM mitjançant l'EM (*Expectation–Maximization*) es desenvolupa de la següent manera:

1. **Inicialització:** Es defineix el nombre de components K i s'inicialitzen els paràmetres del model:
$$\Theta^{(0)} = \{\pi_j^{(0)}, \mu_j^{(0)}, \Sigma_j^{(0)}\}_{j=1}^{K},$$
on π_j són els pesos de barreja (satisfent $\sum_j \pi_j = 1$), μ_j els vectors de mitjana i Σ_j les matrius de covariància diagonals de cada component.

2. **E–step (Expectació):** Per a cada usuari i (representat pel vector de característiques x_i), es calculen les responsabilitats γ_{ij}, és a dir, la probabilitat de pertinença de x_i al component j:
$$\gamma_{ij} = \frac{\pi_j \, \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{l=1}^{K} \pi_l \, \mathcal{N}(x_i \mid \mu_l, \Sigma_l)},$$
on $\mathcal{N}(x \mid \mu, \Sigma)$ és la densitat de la distribució normal multivariante.

3. **M–step (Maximització):** Es reestimen els paràmetres del model utilitzant les responsabilitats calculades:
$$N_j = \sum_{i=1}^{N} \gamma_{ij},$$
$$\pi_j^{\text{new}} = \frac{N_j}{N}, \quad \mu_j^{\text{new}} = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_{ij} \, x_i, \quad \Sigma_j^{\text{new}} = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_{ij} \, (x_i - \mu_j^{\text{new}})(x_i - \mu_j^{\text{new}})^{\top}.$$

4. **Iteració:** Els passos d'*E–step* i *M–step* es repeteixen fins que la variació en el log-vèrtex o en els paràmetres sigui inferior a un llindar predeterminat, o s'arribi al nombre màxim d'iteracions.

Un cop convergit el model, es pot assignar cada usuari i al clúster $\hat{j}$ per al qual $\gamma_{i\hat{j}}$ és màxima (assignació dura) o bé mantenir les responsabilitats per a un càlcul de predicció difús. En la nostra implementació hem optat per l'assignació dura:

$$\hat{j}_i = \arg\max_j \ \gamma_{ij}.$$

Finalment, la predicció de la valoració de l'usuari u per l'ítem v es calcula considerant exclusivament les valoracions aportades pels usuaris que han estat assignats al mateix component gaussià que u.

Per a la implementació del *Gaussian Mixture Model* s'ha fet ús de la classe `GaussianMixture` del mòdul `sklearn.mixture` de la biblioteca `scikit-learn` [25].

Capítol 6

Anàlisi de resultats

En aquest capítol es presenten i s'analitzen els resultats obtinguts en l'avaluació dels diferents mètodes de clustering aplicats al sistema de recomanació. Per tal d'assegurar la fiabilitat i generalitzabilitat de les conclusions, s'ha emprat una metodologia de validació creuada (cross-validation). Concretament, s'ha aplicat una validació creuada de 5 particions (5-fold cross-validation), on el conjunt de dades es divideix en 5 subconjunts o *folds*. En cada iteració, un dels *folds* s'utilitza com a conjunt de prova, mentre que els 4 restants s'empren com a conjunt d'entrenament. Aquest procés es repeteix 5 vegades, de manera que cada *fold* actua com a conjunt de prova exactament una vegada. Les mètriques de rendiment obtingudes en cada una de les 5 iteracions es promedian per obtenir una estimació més robusta i menys susceptible a la particularitat d'una única partició de les dades.

Per avaluar el rendiment i les característiques dels clústers formats, s'empraran diverses mètriques quantitatives, els valors de les quals seran, per tant, la mitjana obtinguda després del procés de validació creuada.

Primerament, per avaluar la precisió de les prediccions de valoracions generades pel sistema, s'utilitzaran l'Error Absolut Mitjà Normalitzat (NMAE) i l'Error Quadràtic Mitjà Normalitzat (NRMSE). Aquestes mètriques es calculen de la següent manera:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{r}_{u,i} - r_{u,i}|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{r}_{u,i} - r_{u,i})^2}$$

on N és el nombre total de valoracions predites en un *fold* de prova, $\hat{r}_{u,i}$ és la valoració predita per a l'usuari u sobre l'ítem i, i $r_{u,i}$ és la valoració real. Per obtenir les versions normalitzades, NMAE i NRMSE, aquests valors es divideixen pel rang de les valoracions presents al conjunt de dades ($\text{rating}_{\text{max}} - \text{rating}_{\text{min}}$):

$$\text{NMAE} = \frac{\text{MAE}}{\text{rating}_{\text{max}} - \text{rating}_{\text{min}}}$$

$$\text{NRMSE} = \frac{\text{RMSE}}{\text{rating}_{\text{max}} - \text{rating}_{\text{min}}}$$

La normalització d'aquestes mètriques d'error és particularment útil en aquest estudi, ja que permet comparar de manera més equitativa el rendiment dels algorismes a través de diferents conjunts de dades que puguin tenir escales de valoració distintes. Valors més baixos de la mitjana d'NMAE i NRMSE obtinguts a través de la validació creuada indiquen una major precisió generalitzada en les prediccions.

En segon lloc, per avaluar la qualitat de la distribució dels usuaris entre els clústers formats, s'utilitzarà l'entropia normalitzada dels clústers. Aquesta mètrica mesura el grau de desequilibri en l'assignació dels usuaris als diferents grups. Una situació ideal és aquella on els usuaris es distribueixen de manera relativament uniforme entre els clústers disponibles, evitant que la majoria dels usuaris siguin assignats a un sol clúster o a un nombre molt reduït de clústers. Si la majoria d'usuaris són agrupats en un únic clúster, l'estratègia de clustering no aporta un valor significatiu, ja que el sistema es comportaria de manera similar a un enfocament sense clustering. L'entropia normalitzada es calcula com:

$$H_{\text{norm}} = \frac{-\sum_{j=1}^{K} p_j \log_2(p_j)}{\log_2(K)}$$

on K és el nombre de clústers i p_j és la proporció d'usuaris assignats al clúster j en un *fold* particular. Un valor de la mitjana d'entropia normalitzada (calculada a partir dels valors obtinguts en cada *fold*) proper a 1 indica una distribució equilibrada dels usuaris de manera consistent, mentre que un valor proper a 0 suggereix que la majoria dels usuaris estan concentrats en pocs clústers de forma recurrent. Per tant, es cercaran valors mitjans d'entropia més elevats, ja que reflecteixen una millor utilització de la capacitat de segmentació del mètode de clustering en diferents particions de les dades.

Bibliografia

[1] Hassaan Idrees. *Recommender systems: Personalizing the digital experience*. Juny de 2024. URL: https://medium.com/@hassaanidrees7/recommender-systems-personalizing-the-digital-experience-d2e2ba3d3250.

[2] Paul B. Kantor et al. *Recommender Systems Handbook*. Springer US, 2011.

[3] Ian MacKenzie, Chris Meyer i Steve Noble. *How retailers can keep up with consumers*. Oct. de 2013. URL: https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers.

[4] Badrul Sarwar et al. "Item-based collaborative filtering recommendation algorithms". A: *Proceedings of the 10th international conference on World Wide Web*. 2001.

[5] Gediminas Adomavicius i Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions". A: *IEEE transactions on knowledge and data engineering* 17.6 (2005).

[6] Anil K Jain. "Data clustering: 50 years beyond k-means". A: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2008.

[7] Kim Falk. *Practical recommender systems*. Simon i Schuster, 2019.

[8] Dietmar Jannach et al. *Recommender Systems: An Introduction*. New York, NY, USA: Cambridge University Press, 2011.

[9] Person. *What is kanban methodology: Introduction to kanban framework*. Febr. de 2025. URL: https://kissflow.com/project/agile/kanban-methodology/.

[10] *Online gantt chart maker for Project Planning*. URL: https://app.ganttpro.com/.

[11] *Project Manager, Information Technology (IT) Salary in Spain in 2025 – PayScale*. URL: https://www.payscale.com/research/ES/Job=Project_Manager,_Information_Technology_(IT)/Salary.

[12] *Researcher, Scientific Salary in Spain in 2025 – PayScale*. URL: https://www.payscale.com/research/ES/Job=Research_Scientist/Salary.

[13] *Software Developer Salary in Spain in 2025 – PayScale*. URL: https://www.payscale.com/research/ES/Job=Software_Developer/Salary.

[14] *Data Analyst Salary in Spain in 2025 – PayScale*. URL: https://www.payscale.com/research/ES/Job=Data_Analyst/Salary.

[15] F. Maxwell Harper i Joseph A. Konstan. "The MovieLens Datasets: History and Context". A: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5.4 (2015), 19:1 - 19:19. DOI: 10.1145/2827872. URL: https://doi.org/10.1145/2827872.

[16] Cai-Nicolas Ziegler et al. "Improving Recommendation Lists Through Topic Diversification". A: *Proceedings of the 14th International World Wide Web Conference (WWW '05)*. Chiba, Japan: ACM, maig de 2005, pàg. 22 - 32. DOI: 10.1145/1060745.1060754.

[17] Ken Goldberg. *The Jester Dataset*. http://goldberg.berkeley.edu/jester-data/. Accessed: 2025-05-06. 2001.

[18] Ken Goldberg et al. "Eigentaste: A Constant Time Collaborative Filtering Algorithm". A: *Information Retrieval* 4.2 (jul. de 2001), pàg. 133 - 151. DOI: 10.1023/A:1011419012201.

[19] Diego Monti, Giuseppe Rizzo i Maurizio Morisio. "All you need is ratings: A clustering approach to synthetic rating datasets generation". A: *arXiv preprint arXiv:1909.00687* (2019).

[20] Pradipto Chowdhury i Bam Bahadur Sinha. "Evaluating the Effectiveness of Collaborative Filtering Similarity Measures: A Comprehensive Review". A: *Procedia Computer Science* 235 (2024), pàg. 2641 - 2650.

[21] Jon Herlocker, Joseph A Konstan i John Riedl. "An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms". A: *Information retrieval* 5 (2002), pàg. 287 - 310.

[22] Sobia Zahra et al. "Novel centroid selection approaches for KMeans-clustering based recommender systems". A: *Information sciences* 320 (2015), pàg. 156 - 189.

[23] Absalom E Ezugwu et al. "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects". A: *Engineering Applications of Artificial Intelligence* 110 (2022), pàg. 104743.

[24] Gilda Moradi Dakhel i Mehregan Mahdavi. "A new collaborative filtering algorithm using K-means clustering and neighbors' voting". A: *2011 11th International conference on hybrid intelligent systems (HIS)*. IEEE. 2011, pàg. 179 - 184.

[25] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". A: *the Journal of machine Learning research* 12 (2011), pàg. 2825 - 2830.

[26] Kiatichai Treerattanapitak i Chuleerat Jaruskulchai. "Items based fuzzy C-mean clustering for collaborative filtering". A: *Information Technology Journal KMUTNB* 5.2 (2009), pàg. 30 - 34.

[27] Hamidreza Koohi i Kourosh Kiani. "User based collaborative filtering using fuzzy C-means". A: *Measurement* 91 (2016), pàg. 134 - 139.

[28] César Inga Chalco, Rodolfo Bojorque Chasi i Remigio Hurtado Ortiz. "Hierarchical clustering for collaborative filtering recommender systems". A: *Advances in Artificial Intelligence, Software and Systems Engineering: Joint Proceedings of the AHFE 2018 International Conference*. Springer. 2019, pàg. 346 - 356.

[29] Anna Satsiou, Stefanos Vrochidis i Ioannis Kompatsiaris. “A Hybrid Recommendation System Based on Density-Based Clustering”. A: *LNCS 10750 Internet Science* (2017), pàg. 49.

[30] Hangyu Yan i Yan Tang. “Collaborative filtering based on Gaussian mixture model and improved Jaccard similarity”. A: *Ieee Access* 7 (2019), pàg. 118690-118701.