TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Desenvolupament i comparació de recomanadors d'articles tèxtils i estudi de regles d'associació

Autor: Clara Tarragó Grau

Director: Sergi Ramírez Mitjans

Departament: Dept. Estadística i Investigació Operativa

Convocatòria: 2024 - 2025

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

Aquest treball consisteix en el desenvolupament d'un sistema recomanador d'articles de roba i l'anàlisi de regles d'associació en les preferències de compra dels usuaris.

La principal motivació darrere d'aquest estudi és proporcionar un camí a la transformació digital a l'abast de les petites i mitjanes empreses, amb un enfocament social i d'accessibilitat tecnològica. La implementació d'un sistema de recomanació pot ajudar a gestionar estocs, millorar l'experiència de compra dels usuaris i augmentar les vendes.

Els principals objectius d'aquest treball consisteixen en: desenvolupar un model de recomanació d'articles de roba basat en preferències d'usuaris, aplicar tècniques de *machine learning* per a la recomanació de productes, optimitzar el sistema per oferir recomanacions visuals precises i rellevants, desenvolupar una interfície d'usuari intuïtiva per a la implementació del recomanador i aplicar regles d'associació per estudiar patrons en les preferències de roba dels usuaris. També s'han plantejat alguns objectius transversals: el desenvolupament d'una enquesta interactiva de classificació d'imatges mitjançant *Python* i *Streamlit*, l'aprenentatge de nocions bàsiques de *Python* i de *Latex* mitjançant *Overleaf*.

Hem conseguit desenvolupar una enquesta interactiva on es recollien variables demogràfiques dels usuaris, i se'ls demanava que fessin un rànquing de 10 imatges de peces de roba. Cada usuari veia 10 imatges diferents, escollides de manera aleatòria d'una base de dades comú de 100 imatges. Un cop recollides les dades, hem començat treballant amb 4 subtipus de recomanadors dins el filtratge col · laboratiu: enfocament *user-based*, enfocament *item-based*, *Singular Value Decomposition* i *Alternating Least Squares*. Pels dos primers enfocaments s'ha començat amb un plantejament senzill, treballant només amb les dades dels rànquings. Posteriorment, s'han complementat amb característiques demogràfiques dels usuaris, pel cas *user-based*, i amb característiques de les peces de roba i similitud d'imatges, pel cas *item-based*. Comparant el plantejament senzill versus el plantejament híbrid en cada cas, hem pogut verificar mitjançant testos estadístics que els plantejaments híbrids proporcionen recomanacions més precises. Així doncs, els quatre sistemes recomanadors finals amb els que hem treballat han estat: enfocament *user-based* híbrid, enfocament *item-based* híbrid, *Singular Value Decomposition* i *Alternating Least Squares*. No s'han trobat diferències estadísticament significatives entre aquests quatre models, però el que presenta recomanacions més acurades és l'enfocament *user-based* híbrid. S'ha conduït una prova àcida en una mostra reduïda d'individus per verificar aquest model: se'ls feia respondre l'enquesta inicial i, un cop enviats els rànquings, les dades eren processades pel recomanador *user-based* híbrid i retornava les top-3 recomanacions, que l'usuari havia de puntuar. La puntuació mitjana obtiguda de les recomanacions ha estat un 8,2/10, consolidant la seva precisió. Finalment, s'han estudiat possibles regles d'associació en les preferències de roba dels usuaris enquestats, però no s'han detectat regles rellevants per l'estudi.

Paraules clau: Sistemes recomanadors, Filtratge col · laboratiu, Aprenentatge automàtic, Similitud, Regles d'associació.

Abstract

This work consists of the development of a clothing article recommendation system and the analysis of association rules in users' shopping preferences.

The main motivation behind this study is to provide a path to digital transformation within reach of small and medium-sized enterprises, with a focus on social and technological accessibility. The implementation of a recommendation system can help manage stock, improve the shopping experience of users and increase sales.

The main goals of this work are: developing a clothing article recommendation system based on user preferences, applying machine learning techniques for product recommendation, optimizing the system to offer accurate and relevant visual recommendations, developing an intuitive user interface for the implementation of the recommender, and applying association rules to study patterns in users' clothing preferences. Some cross-curricular objectives have also been proposed: the development of an interactive image classification survey using Python and Streamlit, learning basic notions of Python, and Latex using Overleaf.

We have managed to develop an interactive survey in which demographic variables of the users were collected, and they were asked to rank 10 images of clothing items. Each user saw 10 different images, randomly chosen from a common database of 100 images. Once the data was collected, we began working with 4 subtypes of recommenders within collaborative filtering: user-based approach, item-based approach, Singular Value Decomposition and Alternating Least Squares. For the first two methods, we started with a simple approach, working only with the ranking data. Later, they were complemented with demographic characteristics of the users, in the user-based approach, and with characteristics of the clothing items and image similarity, in the item-based approach. Comparing the simple approach versus the hybrid approach in each case, we were able to verify using statistical tests that the hybrid approaches provide more accurate recommendations. Thus, the four final recommendation systems we worked with were: hybrid user-based approach, hybrid item-based approach, Singular Value Decomposition and Alternating Least Squares. No statistically significant differences were found between these four models, but the one that provides the most accurate recommendations is the hybrid user-based approach. A stress test was conducted on a small sample of individuals to verify this model: they were asked to complete the initial survey and, once the rankings were submitted, the data was processed by the hybrid user-based recommender and returned the top-3 recommendations, which the users had to rate. The average score obtained from the recommendations was 8,2/10, consolidating the recommender's accuracy. Finally, possible association rules in the clothing preferences of the surveyed users were studied, but no relevant rules were detected for the study.

Keywords: Recommender systems, Collaborative filtering, Machine learning, Similarity, Association rules.

Classificació AMS (MSC2010)

62-XX STATISTICS

- **62-07** Data analysis
- **62Hxx** Multivariate analysis
  - **62H20** Measures of association (correlation, canonical correlation, etc)
  - **62H35** Image analysis

68-XX COMPUTER SCIENCE

- **68-04** Explicit machine computation and programs (not the theory of computation or programming)
- **68Uxx** Computing methodologies and applications
  - **68U10** Image processing

Índex

Resum i
Abstract ii
Classificació AMS iii
Índex de figures v
Índex de taules vii
Abreviatures i acrònims ix

1 Introducció **1**

2 Estat de l'art **3**
2.1 Recomanadors d'articles tèxtils basats en imatges 3
2.2 Regles d'associació en comerços electrònics 7

3 Marc teòric **10**
3.1 Similitud d'imatges 11
3.2 Normalització de les dades 12
3.3 Imputació de valors faltants 13
3.4 Proves estadístiques 14
3.5 Mètriques de l'error 15
3.6 Sistemes recomanadors 15
3.6.1 Filtratge col · laboratiu 16
3.7 Proves àcides 20
3.8 Regles d'associació 20

4 Treball de camp **23**
4.1 Fonts de dades 23
4.2 Preprocessament 25
4.3 Anàlisi descriptiva exploratòria 26
4.3.1 Anàlisi univariant 26
4.3.2 Anàlisi bivariant 33
4.4 Tractament d'imatges 34
4.5 Desenvolupament dels recomanadors 36
4.5.1 *User-based* 37
4.5.2 *Item-based* 39
4.5.3 *Singular Value Decomposition* (SVD) 41

4.5.4 *Alternating Least Squares* (ALS) 42
4.5.5 Comparació de models 42
4.6 Regles d'associació 44

5 Conclusions **47**
5.1 Línies futures 48

Annex **53**
A.1 Gràfics complementaris per l'anàlisi descriptiva 53

Índex de figures

3.1 Esquema del procediment a seguir per la realització d'aquest Treball Final de Grau, segons la metodologia CRISP-DM. 10
3.2 Esquema del procediment que segueix l'algoritme d'imputació *MICE*. *Medium*, 2024. Disponible a: https://medium.com/aimonks/imputation-by-chained-equations-mice-bridging-the-gap-in-incomplete-data-analysis-38a4bfcff88e (consultat: 17jun2025). 13
3.3 Exemple de matriu usuari-article. *IBM*, 2024. Disponible a: https://www.ibm.com/think/topics/collaborative-filtering (consultat: 21abr2025). . 16
3.4 Exemple de matriu usuari-article. *IBM*, 2024. Disponible a: https://www.ibm.com/think/topics/collaborative-filtering (consultat: 21abr2025). . 17
3.5 Mostra d'una taula de transaccions. *Medium*, 2024. Disponible a: https://medium.com/image-processing-with-python/fundaments-of-associate-rule-mining-468801ec0a29 (consultat: 22abr2025). 21

4.1 Distribució dels colors. 27
4.2 Distribució del tipus d'estampat. 27
4.3 Distribució de la presència de marca. 28
4.4 Distribució del gènere dels models. 28
4.5 Distribució de l'edat dels models. 29
4.6 Distribució del gènere dels enquestats. 29
4.7 Distribució de l'edat dels enquestats. 30
4.8 Distribució de la preferència de compra dels enquestats. 31
4.9 Distribució dels colors de les peces de roba a la primera posició dels rànquings. 31
4.10 Distribució del tipus d'estampat de les peces de roba a la primera posició dels rànquings. 32
4.11 Distribució de la presència de marca de les peces de roba a la primera posició dels rànquings. 33
4.12 Imatges seleccionades de peces de roba similars. 35
4.13 Imatges seleccionades de peces de roba diferents. 35
4.14 Ítems ordenats segons la seva freqüència. 45

A.1 Distribució dels colors de les peces de roba a la segona posició dels rànquings. 53
A.2 Distribució dels colors de les peces de roba a la tercera posició dels rànquings. 54
A.3 Distribució dels colors de les peces de roba a l'última (desena) posició dels rànquings. 54

A.4 Distribució dels colors de les peces de roba a la penúltima (novena) posició dels rànquings. 55
A.5 Distribució dels colors de les peces de roba a l'antepenúltima (vuitena) posició dels rànquings. 55
A.6 Distribució del tipus d'estampat a les peces de roba a la segona posició dels rànquings. 56
A.7 Distribució del tipus d'estampat a les peces de roba a la tercera posició dels rànquings. 56
A.8 Distribució del tipus d'estampat a les peces de roba a l'última (desena) posició dels rànquings. 57
A.9 Distribució del tipus d'estampat a les peces de roba a la penúltima (novena) posició dels rànquings. 57
A.10 Distribució del tipus d'estampat a les peces de roba a l'antepenúltima (vuitena) posició dels rànquings. 58
A.11 Distribució de la presència de marca de les peces de roba a la segona posició dels rànquings. 58
A.12 Distribució de la presència de marca de les peces de roba a la tercera posició dels rànquings. 59
A.13 Distribució de la presència de marca de les peces de roba a l'última (desena) posició dels rànquings. 59
A.14 Distribució de la presència de marca de les peces de roba a la penúltima (novena) posició dels rànquings. 60
A.15 Distribució de la presència de marca de les peces de roba a l'antepenúltima (vuitena) posició dels rànquings. 60
A.16 Per les tres primeres posicions dels rànquings, distribució dels gèneres respecte els colors. 61
A.17 Distribució de l'edat de l'individu enquestat versus l'edat del model de la imatge que ha posat a la primera posició. 61
A.18 Distribució de l'edat de l'individu enquestat versus l'edat del model de la imatge que ha posat a l'última posició. 62
A.19 Gènere dels individus enquestats (eix de les X) versus gènere del model de la imatge que han posat a la primera posició. 62
A.20 Gènere dels individus enquestats (eix de les X) versus gènere del model de la imatge que han posat a l'última posició. 63
A.21 Nombre de vegades que cada imatge ha aparegut als rànquings. La línia blava fosca indica la mitjana, que es troba al voltant d'11 aparicions per imatge. . 63

Índex de taules

4.1 Mesures descriptives de l'edat dels enquestats. 30
4.2 Resultats del test χ^2 entre el color de les peces de roba i el gènere dels enquestats. 33
4.3 Resultats del test χ^2 entre el gènere dels enquestats i el gènere dels models. . 34
4.4 Resultats de la similitud del cosinus per cada casuística. 35
4.5 Resultats de la similitud entre les top-4 imatges reals i les top-4 predites pels 15 usuaris escollits, segons cada mètode. 43
4.6 Resultats del *t-test* entre *user-based* senzill i *user-based* híbrid. 43
4.7 Resultats del *t-test* entre *item-based* senzill i *item-based* híbrid. 43
4.8 Resultats de la similitud entre les top-4 imatges reals i les top-4 predites pels 15 usuaris escollits, segons cada mètode. 43

Abreviatures i acrònims

PIMES Petites i Mitjanes Empreses

CRISP-DM *Cross-Industry Standard Process for Data Mining*

MSE *Mean Squared Error*

MICE *Multiple Imputation by Chained Equations*

k-NN *k-Nearest Neighbours*

RMSE *Root Mean Squared Error*

CCP Coeficient de Correlació de Pearson

SVD *Singular Value Decomposition*

ALS *Alternating Least Squares*

Capítol 1

Introducció

Les empreses del sector tèxtil estan experimentant una tendència creixent a incorporar procediments impulsats per *machine learning* per optimitzar processos, millorar la presa de decisions, adaptar-se millor a les necessitats del mercat i potenciar la innovació de la seva oferta. Un exemple podria ser el cas d'Stradivarius, que el març de 2023 va llançar una col · lecció de roba "reinterpretada" amb intel · ligència artificial [1]. La campanya realitzada per promocionar aquesta nova col · lecció també va ser dissenyada amb intel · ligència artificial.

Aquest Treball Final de Grau consisteix en el desenvolupament d'un algoritme recomanador d'articles tèxtils mitjançant tècniques de similitud i d'aprenentatge automàtic, així com en l'estudi de regles d'associació entre les preferències de peces de roba de diferents usuaris. A partir d'una enquesta on cada individu ordena unes imatges d'articles de roba segons la seva preferència, es desenvolupen diversos models de *machine learning* que ofereixen recomanacions de peces de roba similars segons els gustos, i s'estudien les relacions entre les decisions preses pels individus.

Plantejant-ho de manera més concreta al nostre cas, les pàgines web de molts comerços electrònics ja presenten algoritmes recomanadors. Un exemple podria ser el web de Zara, que dins la pestanya d'un article determinat ofereix un subapartat anomenat "Et pot interessar". Recomana o bé articles similars, o bé articles per acabar de completar el nostre conjunt.

Tot i la importància de la transformació digital, els costos tant econòmics com humans per desenvolupar i mantenir aquestes infraestructures tecnològiques sovint resulten massa elevats per a les petites i mitjanes empreses [2]. Aquest tipus de sistemes solen estar restringits a grans empreses amb equips tècnics especialitzats i pressupostos elevats. Això impedeix a les PIMES competir en igualtat de condicions en un entorn cada vegada més digitalitzat.

La motivació per escollir aquesta temàtica ve donada per l'interès en desenvolupar una solució útil i demandada a un problema existent: si la peça que volem no es troba disponible, quines altres opcions similars hi ha, que als usuaris semblants a nosaltres també els agradin? Donant resposta a aquesta pregunta, amb un enfocament social i d'accessibilitat tecnològica, podem proporcionar un camí a la transformació digital a l'abast de les petites i mitjanes empreses.

D'aquesta manera poden millorar la seva estratègia de negoci, aconseguir augmentar les vendes i millorar l'experiència de compra dels usuaris. A més, aquest enfocament també pot ajudar a les empreses a optimitzar la gestió d'estocs, contribuint a fer que la indústria de la moda sigui un sector més eficient i sostenible. Sorgeix així el plantejament d'aquest Treball de Final de Grau d'Estadística.

Els principals objectius que pretén assolir aquest treball es poden dividir en dues categories: principals i transversals. Els objectius principals estan directament relacionats amb la temàtica del treball, mentres que els objectius transversals són aquells relacionats amb l'adquisició en segon pla de nous coneixements per tal de poder realitzar el treball.

Els objectius principals són:

- Desenvolupar un model de recomanació d'articles de roba basat en preferències d'usuaris.
- Aplicar tècniques de *machine learning* per a la recomanació de productes en estoc.
- Optimitzar el sistema per oferir recomanacions visuals precises i rellevants.
- Desenvolupar una interfície d'usuari intuïtiva per a la implementació del recomanador.
- Aplicar regles d'associació per estudiar patrons en les preferències de roba dels usuaris.

Els objectius transversals són:

- Aprendre nocions bàsiques del llenguatge de programació *Python*.
- Desenvolupar una enquesta interactiva de classificació d'imatges mitjançant *Python*.
- Aprendre a escriure amb *Latex* i a treballar amb la plataforma *Overleaf*.

Aquesta memòria es divideix en 5 apartats:

1. Introducció, l'apartat actual. S'estableix una breu base del tema del treball i s'expliquen les motivacions, objectius i apartats de la memòria.
2. Estat de l'art. Es contextualitza al lector sobre la investigació duta a terme prèviament per altres autors respecte la matèria.
3. Marc teòric. S'expliquen detalladament les eines i algoritmes usats al llarg del treball per assolir cadascun dels objectius.
4. Treball de camp. Aquesta secció es divideix en els subapartats corresponents, segons els diferents aspectes que es treballen durant el procés (les fonts de dades, el preprocessament, l'anàlisi descriptiva prèvia, el desenvolupament dels recomanadors i l'estudi de les regles d'associació). Cada subapartat explica i discuteix el procediment i els resultats obtinguts.
5. Conclusions. Es presenta un resum dels resultats més destacats, donant resposta als objectius plantejats a l'introducció. S'introdueixen línies futures per a l'estudi.
6. Bibliografia. Informació bibliogràfica de totes les obres emprades per a l'elaboració del treball.

Capítol 2

Estat de l'art

2.1 Recomanadors d'articles tèxtils basats en imatges

Tot comença amb l'obra de David Goldberg, David Nichols, Brian M. Oki i Douglas Terry: *Using collaborative filtering to weave an information tapestry* [3], l'any 1992. Aquest article explica el desenvolupament del primer sistema recomanador. Era un sistema de filtratge de documents en llistes de correus, basat en el contingut dels documents.

Aquesta obra presenta per primera vegada el terme **filtratge col · laboratiu**, que fa referència a que s'involucra a les persones en el procés de filtratge. Per al cas específic d'aquest article, significa que les persones col · laboren per ajudar-se els uns als altres a realitzar el filtratge, registrant les seves reaccions als documents llegits.

El terme "filtratge col · laboratiu" passarà a la història i serà adoptat per pràcticament tota la resta d'autors futurs que desenvoluparan sistemes de recomanació d'aquest tipus.

La publicació de *Recommender systems*, per Paul Resnick i Hal R. Varian el 1997 [4], és la primera a plantejar una sèrie de qüestions importants a tenir en compte a l'hora de crear un recomanador.

Es parla sobre l'estructura del cost de les decisions que pren la gent sobre els ítems: el cost de fallar i no presentar un bon ítem com a recomanació, el cost de presentar un mal ítem i la comparació d'aquests dos costos.

Es posa sobre la taula el tema de la privacitat de les dades, l'oferta d'incentius per tal d'obtenir recomanacions, i parlen sobre qui pot oferir recomanacions: podria passar que els creadors de contingut generessin bones recomanacions pels seus productes i males recomanacions per als seus competidors, esbiaixant així el sistema.

Relacionat amb això, plantegen 3 possibles escenaris futurs de models de negoci: un on els usuaris paguin per rebre recomanacions, un on s'obtinguin ingressos gràcies a anunciants que es beneficien de les dades recollides pel sistema, i un últim on els creadors de contingut paguen per a que els seus ítems siguin puntuats. Els dos últims models poden fer pensar als clients

que el sistema està esbiaixat i que les recomanacions no són veritablement personalitzades.

Es conclou que com més usuaris té el sistema de recomanació, més fàcil és trobar "algú com tu". Assenyalen que això portarà a una competició intensa entre sistemes de recomanació, fent que probablement només sobrevisquin un o dos competidors en cada àrea de contingut (llibres, pel · lícules, etcètera). Aquests sistemes dominants poden concedir llicències dels seus serveis de recomanació als venedors de productes, com un servei de valor afegit.

Un aspecte que comença a cobrar cada vegada més importància en els comerços, sobretot en els comerços electrònics, és la personalització massiva. Consisteix en un procés basat en proporcionar béns i serveis personalitzats, que satisfacin més adequadament les necessitats específiques de cada client. Mitjançant l'estandardització i economies d'escala, es subministren productes adaptats als requisits específics del client en volums grans i a un cost raonablement baix [5].

L'obra de J. Ben Schafer, Joseph Konstan i John Riedl [6] planteja que per aconseguir aquesta personalització massiva els e-comerços augmenten la informació que els clients han de processar a l'hora d'escollir quins productes adquireixen. Una solució a aquest increment d'informació és l'ús de sistemes de recomanació. Els productes es poden recomanar en funció dels *best-sellers* del comerç, les característiques demogràfiques del client o en funció d'un anàlisi del comportament de compra en ocasions prèvies del client, com a predicció del comportament de compra futur. Els sistemes de recomanació permeten automatitzar tota aquesta personalització al web.

Aquesta obra posa sobre la taula tres maneres en què els sistemes recomanadors milloren les vendes dels e-comerços: transformen els usuaris "navegadors" (els que només naveguen pel web) en compradors, augmenten les vendes a través de *cross-sell* (es recomanen productes addicionals al client) i afavoreixen la lleialtat dels clients, ja que els e-comerços inverteixen temps i recursos en conèixer els gustos dels usuaris i presenten interfícies personalitzades, fent que els clients acabin tornant a aquests webs que són capaços de captar millor les seves necessitats.

També s'estableix que hi ha diferents maneres de proporcionar les recomanacions al client. Les principals són: navegació pel web, ítems similars, recomanacions per correu electrònic, comentaris de text d'altres clients, puntuació mitjana d'un producte per part dels altres clients, top-N ítems i resultats de búsqueda ordenats.

Es parla de quatre tecnologies de recomanació: els recomanadors no-personalitzats (tots els clients reben les mateixes recomanacions, basades en comentaris de text d'altres clients o en la puntuació mitjana d'un producte), els sistemes de recomanació basats en atributs (recomanacions segons les propietats sintàctiques de la búsqueda realitzada pel client), els sistemes de recomanació de correlació ítem a ítem (recomanacions en funció d'ítems en els que el client ha mostrat interès) i els sistemes de recomanació de correlació persona a persona (recomanacions basades en la correlació entre el client i altres clients del comerç). Aquestes dues darreres tecnologies de recomanació són les incloses en l'enfocament popularment conegut com a **filtratge col · laboratiu**.

Finalment, proposa diverses oportunitats pels comerços electrònics per l'expansió dels sistemes de recomanació. La primera oportunitat comporta incloure tant les qualificacions positives com les negatives. Molts webs usen les dades de vendes com una qualificació implícita positiva, i no sempre haver comprat un producte és sinònim d'una bona valoració d'aquest producte. Plantegen que seria interessant incloure les dades de les devolucions, com a qualificacions implícites negatives, ja que molts algoritmes de recomanació funcionen millor amb valoracions positives i negatives.

La següent oportunitat consisteix en combinar diferents fonts d'informació. No només fer servir un petit subconjunt de la informació disponible, sinó combinar-la tota per proporcionar recomanacions més robustes i en temps real: usar simultàniament dades demogràfiques del client, dades de compres, valoracions explícites, etcètera.

Una limitació important dels sistemes de recomanació és la recopilació d'informació suficient per a realitzar recomanacions adients als nous clients. Una proposta seria que els webs compartissin la informació sobre els seus usuaris, però això disminueix els beneficis dels comerços individuals. Com que la informació és poder, els comerços tenen pocs incentius per compartir-la. Es planteja la possibilitat de formar consorcis de comerços no competidors per compartir les dades i augmentar el seu valor, però apareix de nou el problema de la privacitat de les dades dels usuaris.

Preveuen que els sistemes de recomanació aniran més enllà i s'utilitzaran per predir la demanda dels productes, i seran un factor clau per als comerços electrònics per aconseguir l'esmentat a l'inici: la customització massiva.

Més endavant, l'any 2017, les quatre tecnologies de recomanació proposades per J. Ben Schafer, Joseph Konstan i John Riedl [6] evolucionen per donar pas als diferents tipus de sistemes de recomanació que podem trobar: *content-based* (basats en contingut), filtratge col · laboratiu, sistema recomanador híbrid, sistema recomanador demogràfic i *keyword-based* (basat en paraules clau) [7].

Tot seguit, l'any 2018 es publica una obra destacada en l'àmbit dels sistemes de recomanació aplicats a comerços electrònics, concretament de la indústria tèxtil. L'estudi presenta un recomanador de filtratge col · laboratiu implementat en una companyia de moda coreana, que ven productes tant de manera *online* com en botigues físiques [8]. Aquest estudi pioner incorpora tres factors significatius en el desenvolupament del recomanador. Primerament, integra les dades tant de la part *online* del comerç com de la part de venda en botiga física. En segon lloc, té en compte que els productes acostumen a ser estacionals, pel que la preferència del client pot canviar segons l'època de l'any. Darrerament, considera que els clients normalment compren productes per substituir altres productes previs o per complementar altres productes que ja posseeixen. Proposen un nou sistema recomanador anomenat *K-RecSys*, que és una extensió dels algoritmes de filtratge col · laboratiu tradicionals incorporant els tres punts mencionats anteriorment. Aquest nou sistema combina dades de *clicks* en productes *online* i dades de vendes de productes en tendes físiques de manera ponderada, adopta una funció de desgast de les preferències per reflexar els canvis al llarg del temps, i recomana productes

substitutius i complementaris.

Un estudi l'any 2020 apunta que les plataformes d'*e-commerce* comencen a utilitzar el filtratge col · laboratiu per segmentar el mercat en funció dels comportaments dels clients, juntament amb mètriques demogràfiques o psicogràfiques [9].

A continuació entrem en detall en el cas particular de sistemes de recomanació incorporant imatges. Tot i que es publiquen obres de classificació d'imatges de peces de roba [10] [11], el primer treball que aplica el tractament d'imatges per desenvolupar un recomanador d'articles tèxtils data l'any 2016 [12]. Assenyala que les recomanacions dels recomanadors desenvolupats fins al moment es basaven principalment en una cerca de similitud de les etiquetes dels productes. Fins i tot nota que alguns webs recomanen els productes basats en la seva popularitat, considerant que el client voldrà comprar allò que molts altres han comprat. Aquesta obra també incorpora una perspectiva novedosa: considera que una imatge d'una peça de roba pot resultar interessant al client per diversos factors: el patró, el color, el coll o el tipus de màniga. Per afrontar aquest problema, proposa un sistema de recomanació basat en la classificació de múltiples atributs d'una imatge, en comptes de focalitzar-se en un únic atribut.

Uns anys després, el 2020, l'aplicació d'imatges per a la creació d'un sistema de recomanació es segueix desenvolupant a l'obra de Ssvr Kumar Addagarla i Anthoniraj Amalanathan [13]. Plantegen un enfocament basat en aprenentatge automàtic per crear un sistema de recomanació d'articles similars basat en imatges. Apliquen tècniques de reducció de dimensionalitat mitjançant ACP (Anàlisi de Components Principals), a través de la descomposició de valors singulars (SVD). A més, apliquen el mètode de *clustering K-Means++* per agrupar possibles grups de productes similars. Calculen la distància de Manhattan entre la imatge d'entrada i el conjunt de clústers per trobar els N productes més similars amb una distància menor. Van comparar el seu plantejament amb els resultats obtinguts a l'utilitzar 5 mètodes d'agrupació diferents, i van demostrar que el seu enfocament presentava un rendiment superior per recomanar imatges de productes similars.

Amb la pandèmia de la COVID-19, molts usuaris van canviar el seu comportament respecte les compres i es va viure una transició important d'un model més tradicional i físic, en botigues, a compra online. Per afrontar aquest creixement de demanda del comerç electrònic, Ssvr Kumar Addagarla i Anthoniraj Amalanathan proposen un model de recomanació amb imatges basat en *Deep Learning* per comerços electrònics [14]. El seu enfocament aplica Xarxes Neuronals Convolucionals. Primerament, realitzen una extracció de característiques i normalització mitjançant Xarxes Neuronals Convolucionals. Posteriorment, construeixen un índex usant l'algoritme ANNOY (*Approximate Nearest Neighbour Oh Yeah*), per finalment obtenir el llistat dels top-N ítems més propers usant la distància euclidiana. Avaluen el model amb mètriques diverses, com ara *accuracy*, *recall*, precisió, *F1 score* i *error rate*.

2.2 Regles d'associació en comerços electrònics

Cap als anys 90 es comença a estendre la famosa història de la cervesa i els bolquers: una cadena de supermercats dels Estats Units, a l'analitzar les dades de vendes mitjançant tècniques de *market basket analysis*, van descobrir que els homes que compraven bolquers entre setmana també acostumaven a comprar cervesa. Van decidir col · locar aquests dos ítems a prop, aconseguint augmentar les vendes d'ambdós productes. Aquest cas va posar en manifest com les regles d'associació poden descobrir patrons no evidents en el comportament de compra dels consumidors. Tot i que és un exemple molt utilitzat en àrees com la mineria de dades o el màrqueting, no hi ha una font acadèmica oficial que documenti aquest cas.

El treball clàssic que introdueix formalment el tema de les regles d'associació data l'any 1993 [15]. A partir d'una base de dades de transaccions de clients, on cada transacció conté els articles comprats per un client en una visita, presenten un algoritme que genera totes les regles d'associació significatives entre els articles de la base de dades, incorporant tècniques d'estimació i poda.
L'algoritme apriori, que es convertirà en l'algoritme principal per trobar regles d'associació, es presenta un any després pel mateix autor [16].

El 2007 es presenta el primer treball que combina els sistemes de recomanació amb les regles d'associació [17]. Com es comenta a l'apartat anterior, en aquesta època comença a cobrar importància el tema de la personalització o customització massiva. Els enfocaments tradicionals a aquest problema adopten formats predefinits per descriure els requisits del client, cosa que porta a recomanacions errònies [17]. Per solventar aquest problema, es proposa l'addició de regles d'associació mitjançant l'algoritme apriori als sistemes de recomanació.

L'any 2010 es publica el primer treball que combina els sistemes de recomanació amb les regles d'associació aplicat concretament al cas de la indústria tèxtil i la roba [18]. Discuteix que, a diferència d'altres camps d'aplicació (com ara els llibres), és més difícil classificar sistemàticament els diferents atributs de la roba i construir un sistema que analitzi les fluctuacions del comportament de compra en funció de les tendències. Per abordar-ho, en comptes d'utilitzar l'algoritme apriori per generar les regles d'associació utilitza l'algoritme onto-apriori. Aquest algoritme es caracteritza per estar basat en ontologies, i es diferencia de l'algortime apriori tradicional en què utilitza l'heurística per generar les regles d'associació. Bàsicament, extreu regles d'associació de transaccions ontològiques i realitza inferència per proporcionar recomanacions de roba basades en aquestes regles.

L'obra de Ozgur Cakir i Murat Efe Aras, l'any 2012, també incorpora regles d'associació pel desenvolupament d'un motor de recomanació per un comerç electrònic [19]. Les regles d'associació es generen estudiant la relació entre les paraules claus que el client busca al web i els ítems que afegeix al seu carro de la compra virtual. En aquest estudi, s'utilitza el filtratge col · laboratiu com a enfocament de personalització. La particularitat d'aquest estudi es troba en la manera d'avaluar el rendiment del recomanador i les regles d'associació. Com a mesura d'èxit s'utilitza el *basket ratio*, que és la ràtio entre el nombre de productes afegits al carro

electrònic i el nombre de paraules claus buscades per l'usuari. Aquesta avaluació va durar tres setmanes. La primera setmana, el *software* desenvolupat estava inactiu i no es proporcionaven recomanacions. La segona setmana, el *software* es va activar i s'oferien les recomanacions en funció de les regles d'associació trobades. L'última setmana es va tornar a desactivar el *software*, però s'oferien recomanacions de manera aleatòria. Assumint que el tamany de la mostra era suficient, apliquen el test de la distribució normal estàndard (Z) per comparar els *basket ratios* de les 3 setmanes, realitzant comparacions 2 a 2. Amb una significació del 5%, determinen que no hi ha diferències entre les setmanes 1 i 3, però sí entre l'1 i la 2, i entre la 2 i la 3. Això ve a dir que, quan s'utilitza el recomanador amb regles d'associació, el *basket ratio* augmenta significativament.

Més endavant, l'any 2016, un estudi posa sobre la taula un inconvenient de les regles d'associació tradicionals: els algoritmes usats fins al moment, com ara l'apriori, consideren només el volum de vendes dels productes, sense tenir en compte els beneficis d'aquests productes [20]. Aquest esquema no permet descobrir els ítems que tenen un volum de vendes escàs, tot i ser molt rentables. Determina que els estudis previs assumeixen que tots els ítems de la base de dades són de la mateixa natura: s'assumeix que cada producte té el mateix benefici, i que tots els productes apareixen amb la mateixa (o similar) freqüència a la base de dades. La proposta de Jiabei Dai i Bin Zeng consisteix en un nou algoritme de regles d'associació de benefici-suport per sistemes de recomanació *e-commerce*. Té en compte la relació entre el volum de vendes i la rendibilitat de cada ítem. S'utilitza el benefici com a criteri per establir el suport mínim de cada ítem. A l'apartat de Marc teòric es proporciona una explicació més detalla de què és el suport en el context de les regles d'associació.

Una nova aproximació a l'estudi de l'ús de regles d'associació en comerços electrònics consisteix en la implementació de la teoria de conjunts difusos [21]. Aquesta teoria, en contraposició amb la lògica clàssica, planteja que alguna cosa pot ser parcialment verdadera. Es treballa amb valors imprecisos o parcials, en lloc d'únicament valors vertaders o falsos, per poder modelar l'ambigüitat i la gradació de les dades. Seguint en aquesta línia, un conjunt difús defineix com es relaciona un valor continu a un grau de pertinença, cosa que es realitza amb funcions de pertinença. Aquest treball busca estudiar el potencial d'utilitzar aquestes regles d'associació difuses amb l'algoritme apriori. Cerca recomanar productes rellevants a clients considerant no només els productes venuts, sinó també l'import de les vendes, de manera similar a l'enfocament de Jiabei Dai i Bin Zeng [20]. No només considera si el producte ha estat venut o no (binari), sinó que també té en compte l'import de les vendes per crear les regles difuses. Contextualitzem aquest plantejament amb un exemple: les regles d'associació tradicionals ens indiquen que si es compren els productes A i B, aleshores el producte C també es compra. Tanmateix, les regles d'associació difuses indiquen que si es compren més de 8 unitats del producte A, i fins a 8 unitats del producte B, aleshores es compren fins a 8 unitats del producte C. Aquesta aproximació genera regles amb més informació que l'enfocament tradicional. Tota aquesta informació addicional pot ajudar a l'empresa a decidir quins descomptes aplicar per quins productes i en quines quantitats.

Hi ha altres obres que es limiten a l'ús de regles d'associació aplicades a dades de la indústria tèxtil [22], [23]. Afirmen que les conclusions obtingudes d'estudis d'aquestes característiques poden facilitar la presa d'un gran nombre de decisions estratègiques, com ara la segmentació de mercat, la determinació dels preus dels productes, la proposta de descomptes i promocions o la gestió d'estocs, entre d'altres. L'última obra publicada relacionada amb l'ús de regles d'associació en comerços electrònics, que data l'any 2024, també presenta aquest enfoc [24]. L'objectiu és realitzar una anàlisi exhaustiva de l'algoritme apriori aplicat a varis conjunts de dades, per trobar metodologies que puguin afavorir la creació d'un sistema de recomanació amb major flexibilitat. Es busca descobrir regles d'associació valuoses entre els articles amb una freqüència de compra més elevada. L'anàlisi comença examinant la distribució de vendes entre diferents municipis de la República Kirguís, un país d'Àsia central. Es destaca la importància d'adaptar les estratègies de màrqueting a les dinàmiques de cada regió. Addicionalment, aprofunditza en la distribució de vendes segons el tipus de producte, remarcant així també la importància de posar el foc sobre els articles més freqüents a l'hora de crear les regles d'associació. S'emfatitza que així es millora la pertinença i l'eficàcia de les recomanacions dels productes.

Capítol 3

Marc teòric

Iniciem el Marc teòric mostrant l'esquema del procediment que seguirem en aquest treball. Aquest esquema ha estat fet seguint la metodologia CRISP-DM, que se sol usar per marcar les fases a seguir durant un projecte de ciència de dades.

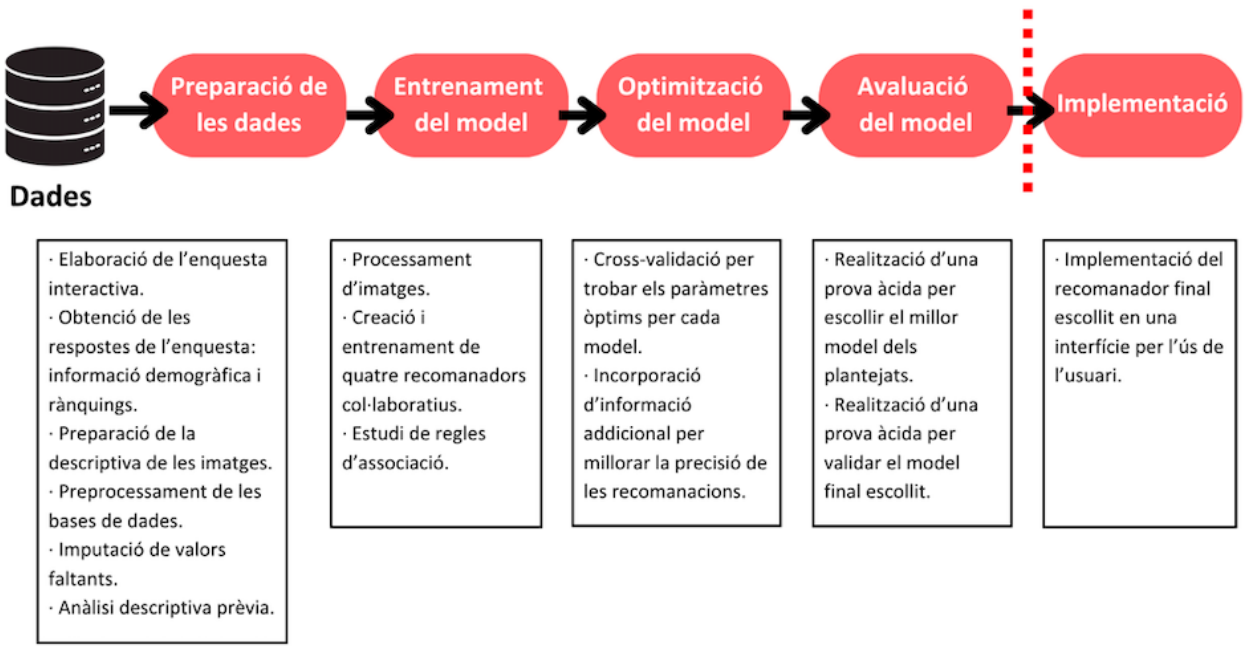

Figura 3.1: Esquema del procediment a seguir per la realització d'aquest Treball Final de Grau, segons la metodologia CRISP-DM.

A continuació, explicarem tots aquells conceptes teòrics i pràctics que s'aplicaran al llarg del cos del treball per assolir els objectius.

3.1 Similitud d'imatges

L'aprenentatge automàtic (o *machine learning*) és una branca de la intel · ligència artificial que permet que les màquines aprenguin determinades tasques sense ser programades específicament per a dites tasques, utilitzant algoritmes i models estadístics [25].

L'aprenentatge automàtic basat en imatges fa referència al procés mitjançant el qual s'entrenen les màquines per a processar, analitzar i interpretar imatges i dades visuals. Aquest reconeixement d'imatges consisteix en tasques com classificar imatges en categories, identificar-ne objectes o detectar la presència de característiques específiques [26].

Actualment, l'ús d'algoritmes de *machine learning* basats en imatges està vivint un gran creixement en disciplines com ara la ciència, la tecnologia o l'enginyeria. A l'àmbit mèdic és on hi veiem un ús més estès en estudis de detecció de tumors, en anàlisis d'imatges de cèl · lules o en classificacions d'imatges radiològiques.

La similitud visual fa referència al procés mitjançant el qual un algoritme computacional quantifica el grau de semblança entre dos o més objectes visuals [27]. Aquesta quantificació es duu a terme a través de l'extracció i comparació de característiques visuals. Per fer-ho existeixen diversos mètodes, com ara la semblança basada en píxels, la semblança basada en histogrames o arquitectures de xarxes neuronals, però en aquest treball ens centrarem en el següent:

- Semblança basada en característiques
  Aquest mètode extreu i compara característiques de la imatge de nivell superior, com ara els marges, les cantonades o les formes. En aquest enfocament es treballa amb vectors de característiques, que són representacions numèriques del contingut de la imatge. Aquests vectors de característiques s'anomenen *embeddings*.

  Per extreure els *embeddings* podem utilitzar models de *deep learning* preentrenats, com ara *VGG16* [28] [29], *ResNet50* [30] o *EfficientNet* [31], entre d'altres. Aquests models estan originalment dissenyats per classificar imatges. Per adaptar-ho a les nostres necessitats, en tots tres models mencionats s'exclouen les capes finals de les xarxes neuronals d'aquests models, anomenades "capes totalment connectades", que són usades per classificar. A l'excloure aquestes capes, ens quedem amb el vector de característiques de les imatges.

  Una altra alternativa són els codificadors automàtics, popularment coneguts com *autoencoders*. Es tracta d'una arquitectura de xarxa neuronal entrenada per reconstruir les imatges d'entrada. L'*autoencoder* està format per dues parts: un *encoder* i un *decoder*. L'*encoder* comprimeix els aspectes més rellevants de la imatge en un espai de menor dimensionalitat. Aquest vector latent resultant és l'*embedding*. El *decoder* intenta reconstruir la imatge a partir de l'*embedding*. L'objectiu de l'*autoencoder* és minimitzar la diferència entre la imatge d'entrada i la imatge reconstruïda. Un cop entrenat, es pot usar només l'*encoder* per convertir qualsevol imatge en un vector de característiques [32].

També es podrien emprar tècniques tradicionals, com ara SIFT (*Scale-Invariant Feature Transform*), SURF (*Speeded-Up Robust Features*) o ORB (*Oriented FAST and Rotated BRIEF*) [33] [34], però en el nostre cas ens centrarem en els models preentrenats mencionats i els codificadors automàtics.

Un cop s'han extret les característiques, es pot mesurar la similitud entre les dues imatges calculant la distància entre els dos vectors de característiques (*embeddings*). Les mètriques de distància més usades són la distància euclidiana o la similitud del cosinus [34]. A continuació les seves respectives fórmules, per a dos vectors x i y:

- Distància euclidiana
  Distància en línia recta entre dos punts de l'espai de característiques.

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$$

  Com més proper a 0 sigui el valor obtingut, més semblants es consideren els vectors.

- Similitud del cosinus
  Mesura l'angle entre dos vectors.

$$Cosinus(x, y) = \frac{xy}{||x||\ ||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

  Com més proper a 1 sigui el valor obtingut, més semblants es consideren els vectors.

3.2 Normalització de les dades

La normalització de les dades és especialment necessària quan es treballa amb algoritmes d'aprenentatge automàtic que es basen en distàncies [35]. En termes generals, la normalització fa referència al procés mitjançant el qual s'ajusten valors mesurats en diferents escales a una escala comú. El mètode més comú per fer-ho per a variables numèriques és la normalització *Z-score* [36]:

$$Z = \frac{X - \mu}{\sigma}$$

μ i σ són la mitjana i la desviació estàndard de les dades, respectivament. Aquest procediment transforma tots els valors de la variable per a que tingui mitjana igual a 0 i desviació estàndard igual a 1.

Per normalitzar variables categòriques podem usar la codificació *one-hot*, consistent en crear tantes variables binàries com categories tinguem. Per exemple, si tenim una variable que defineix la preferència de compra dels usuaris, amb 3 categories (físicament en botiga, *online* o ambdues opcions), crearem una variable per cada categoria. Assignarem 1 a la variable corresponent segons la resposta de l'usuari, i 0 a la resta. Aquest procediment també es coneix com a creació de variables *dummy*.

3.3 Imputació de valors faltants

Un altre aspecte important a considerar és la imputació de valors faltants. Un dels possibles mètodes per realitzar aquesta imputació és el MICE (*Multiple Imputation by Chained Equations*). De manera general, l'esquema de funcionament és el següent: agafa les variables amb valors faltants i els hi assigna com a valor la mitjana, si són numèriques, o la moda, si són categòriques. Un cop la taula de dades està completa, torna a declarar com a *missing* el primer valor faltant. Aplica un model de regressió on la variable resposta és aquella amb el valor faltant, i la resta de variables completes són les predictores. Obté una predicció per a aquest valor faltant, i li assigna aquesta predicció com a nou valor. Es repeteix aquest procés per cada valor faltant que haguem imputat amb la mitjana o la moda a l'inici. Un cop acabat aquest procés, podem dir que el model ha fet una iteració. El model realitza iteracions fins que les diferències entre la taula de dades imputada amb la mitjana o la moda i la taula de dades a l'acabar la iteració siguin properes a 0.

Adjuntem a continuació un esquema per facilitar la comprensió del procediment explicat [37].

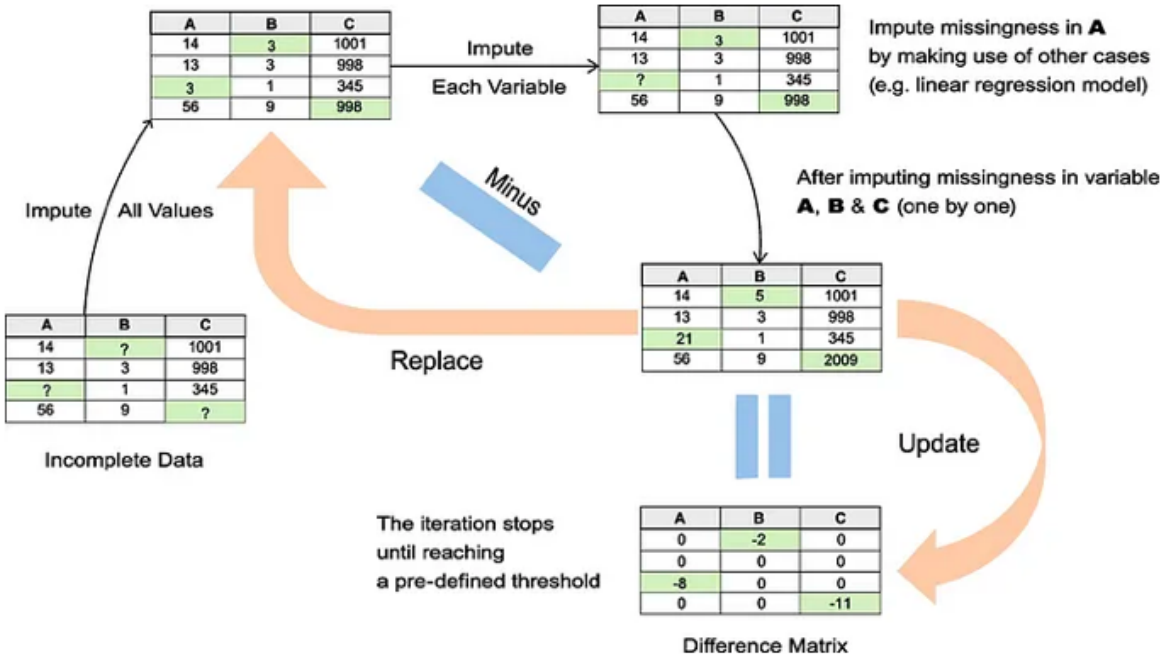

Figura 3.2: Esquema del procediment que segueix l'algoritme d'imputació *MICE*. *Medium*, 2024. Disponible a: https://medium.com/aimonks/imputation-by-chained-equations-mice-bridging-the-gap-in-incomplete-data-analysis-38a4bfcff88e (consultat: 17jun2025).

L'elecció d'aquest algoritme com a mètode d'imputació ve donada per la seva flexibilitat vers el k-NN (*K-Nearest Neighbours*), l'altre principal mètode d'imputació. El MICE funciona millor per *datasets* petits (com és el nostre cas, ja que només tindrem 108 registres) i permet agafar com a context per realitzar les imputacions tant variables numèriques com categòriques,

sense haver-les de preprocessar.

3.4 Proves estadístiques

La primera prova que usarem en aquest treball, a l'anàlisi descriptiva bivariant de les dades, és la prova Xi-quadrat (χ^2). És una prova estadística usada per determinar si existeix associació significativa entre variables categòriques. Compara les freqüències observades d'ocurrències en diferents categories amb les freqüències esperades si no hi hagués associació entre les variables. Les hipòtesis són [38]:

$$\begin{cases} H_0 : P(A,B) = P(A) \cdot P(B) \\ H_1 : P(A,B) \neq P(A) \cdot P(B) \end{cases}$$

L'estadístic de test és:

$$\chi^2 = \sum \frac{(Observat - Esperat)^2}{Esperat}$$

La hipòtesi nul · la estableix que no hi ha diferències significatives entre les dues variables. Si rebutgem la hipòtesi nul · la, afirmem que existeix una associació significativa entre les dues variables, i que les diferències observades no es deuen a l'atzar [39].

El segon test estadístic que usarem per comparar models és la prova t de *Student* (o *t-test*). Aquesta prova es pot fer servir per comparar si les mitjanes de dos grups de dades són iguals, o per comparar la mitjana d'un grup a un valor determinat [40]. Particular al nostre cas, aplicarem aquest test per comparar les mitjanes de dos grups de dades aparellades, i voldrem veure si la mitjana del segon grup de dades és superior. Per tant, les hipòtesis seran les següents:

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d < 0 \end{cases}$$

μ_d és el promig de les diferències entre les dades del primer grup i les del segon grup. Pel nostre cas, la hipòtesi alternativa indicaria que les dades del segon grup són superiors. L'estadístic de test és [40]:

$$t = \frac{\bar{d} - 0}{\bar{S}_d / \sqrt{n}}$$

, on $\bar{d}$ i $\bar{S}_d$ són:

$$\bar{d} = \frac{\sum_{i=1}^{n} x_{i1} - x_{i2}}{n}$$

$$\bar{S}_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

La prova assumeix que les dades segueixen una distribució Normal. Per tant, caldrà comprovar-ho abans d'aplicar-la [41]. La prova es realitzarà amb la funció d'R *t.test()*.

La darrera prova que aplicarem serà una prova ANOVA de blocs aleatorietzats. Aquesta prova l'utilitzarem per comparar els quatre models de recomanador finals, però tenint en

compte l'individu com a bloc, per controlar-ne l'efecte. L'objectiu és estudiar si existeix una diferència significativa entre les mitjanes dels quatre grups. La hipòtesi nul · la i l'alternativa són [42]:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \exists\ i, j \quad tq \quad \mu_i \neq \mu_j\ ,\ i \neq j \end{cases}$$

En cas de rebutjar la hipòtesi nul · la, aplicarem el test de *Tukey* per realitzar les comparacions 2 a 2 i detectar on es troben les diferències. Aquest test s'utilitza després d'un ANOVA per detectar on es troben les diferències quan comparem més de dues categories. De nou, la prova assumeix que les dades segueixen una distribució Normal. Per tant, caldrà comprovar-ho abans d'aplicar-la. També s'assumirà homocedasticitat i que les dades són independents entre blocs. Tant l'ANOVA com el test de *Tukey* també s'aplicaran a través de funcions d'R: *anova()* i *HSD.test()*.

3.5 Mètriques de l'error

Per a determinats processos dins el desenvolupament del recomanador serà interessant dividir les dades en *train* i *test* i avaluar les diferències entre els rànquings predits i els rànquings reals per un determinat usuari o imatge. Per avaluar aquesta diferència, calcularem l'RMSE (*Root Mean Squared Error*):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

on $\hat{y}_i$ representa el valor predit i y_i el valor real.

Un cop tenim totes aquestes bases, podem procedir a desenvolupar el recomanador de peces de roba.

3.6 Sistemes recomanadors

Un sistema recomanador és un sistema d'aprenentatge automàtic que suggereix als usuaris nous productes o serveis. Aquestes suggerències poden estar basades en diversos factors, com les compres anteriors, els *best-sellers* del comerç o les característiques demogràfiques del client, entre d'altres [43].

Tot i que la literatura presenta fins a 5 tipus de sistemes de recomanació (*content-based* (basats en contingut), filtratge col · laboratiu, sistema recomanador híbrid, sistema recomanador demogràfic i *keyword-based* (basat en paraules clau)) [7], podem acotar-ho i quedar-nos amb els dos tipus principals: filtratge per contingut i filtratge col · laboratiu [43].

El filtratge per contingut utilitza els atributs d'un article per recomanar altres articles similars a aquells amb els que el client ha interactuat positivament. Aquest tipus de sistema recomanador es centra única i exclusivament en les característiques dels productes [44].

Particular al cas del nostre projecte, ens centrarem sobretot en el filtratge col · laboratiu.

3.6.1 Filtratge col · laboratiu

Tenim un conjunt d'n individus i un conjunt d'x articles. Representem la preferència de cada individu per cada article en una matriu usuari-article, de mida n x x. Les files representen els individus i les columnes els articles, generalment, tot i que es pot invertir l'ordre. Aquesta matriu també es pot denominar matriu de valoració de l'usuari. Un valor dins la matriu representa el comportament de l'usuari u vers l'article a. En el nostre cas, aquests valors seran el rànquing establert per cada individu a cada article de roba visualitzat a l'enquesta interactiva. A la figura següent podem veure un exemple de matriu usuari-article, per al cas d'un exemple realitzat amb valoracions de llibres. A l'identificar usuaris o articles similars basant-nos en les valoracions que tenim, podem predir valoracions d'articles que un usuari determinat no hagi vist, i recomanar-li o no els articles. Aquests valors són els representats amb un *null* a la matriu.

	Northanger Abby	Wuthering Heights	Oroonoko	Bondswoman's Narrative
Alex	5	4	3	4
Loren	1	2	4	5
Taylor	1	2	3	*null*
Ainsley	*null*	4	3	1

Figura 3.3: Exemple de matriu usuari-article. *IBM*, 2024. Disponible a: https://www.ibm.com/think/topics/collaborative-filtering (consultat: 21abr2025).

Es disposa de dues maneres principals d'implementar filtratge col · laboratiu: *memory-based approach* (enfocament basat en memòria) i *model-based approach* (enfocament basat en un model) [45].

Enfocament basat en memòria

L'enfocament basat en memòria generalment aplica algoritmes de veïnat per predir la valoració que farà un usuari d'un ítem en funció d'usuaris o ítems similars, basant-se en les preferències dels usuaris. Es pot subdividir en dos tipus: *user-based* (basat en usuaris) i *item-based* (basat en ítems).

L'enfocament **basat en usuaris** recomana articles a un usuari determinat basant-se en les preferències d'usuaris amb un comportament similar. Es compara el comportament de l'usuari

objectiu amb el comportament de la resta d'usuaris, i es calcula la similitud entre usuaris. Aquells usuaris amb una similitud més gran amb l'usuari objectiu s'anomenen "veïns". Se seleccionen els n usuaris més similars a l'usuari objectiu i es prediuen les valoracions que faria sobre els articles que no ha vist. Cada una d'aquestes prediccions es calcula com la mitjana ponderada de les valoracions que han fet els usuaris similars que sí han vist l'article. El sistema recomanarà aquells articles amb una puntuació més elevada [44].

En línies generals, es busquen usuaris similars al client i es recomanen articles que aquests usuaris similars hagin valorat positivament. Es basa en el principi que els usuaris amb característiques i comportaments similars comparteixen interessos semblants [44]. Si dos usuaris tenen gustos similars, el que li agrada a un usuari probablement també li agradarà a l'altre.

L'enfocament **basat en ítems** recomana articles a un usuari determinat basant-se en la semblança entre articles segons les valoracions dels usuaris. Es recomanen ítems similars als que l'usuari ha valorat positivament.

Per cada un dels articles que l'usuari ha valorat, es calcula la similitud amb la resta d'articles que no ha valorat, en funció de les valoracions fetes per altres usuaris. D'entre els ítems que l'usuari objectiu sí ha vist, s'agafen els n ítems més similars a cada article no vist i es prediuen les valoracions que faria sobre els articles no vistos. Cada una d'aquestes prediccions es calcula com la mitjana ponderada de les valoracions que ha fet l'usuari objectiu dels n ítems similars que sí ha vist.

Ara bé, és important remarcar la diferència entre *item-based* i el filtratge basat en contingut. Aquest darrer es basa en les característiques dels articles, mentres que l'enfocament *item-based* es basa en les interaccions dels usuaris amb els ítems, a través de les valoracions [44].

Il · lustrem la diferència entre els plantejaments *user-based* i *item-based* a través de dos exemples, amb les dades de la figura anterior. Tornem a adjuntar la figura a sota per poder seguir l'exemple amb més claredat. Les puntuacions es troben en una escala de l'1 al 5, sent 5 la puntuació màxima.

	Northanger Abby	Wuthering Heights	Oroonoko	Bondswoman's Narrative
Alex	5	4	3	4
Loren	1	2	4	5
Taylor	1	2	3	*null*
Ainsley	*null*	4	3	1

Figura 3.4: Exemple de matriu usuari-article. *IBM*, 2024. Disponible a: https://www.ibm.com/think/topics/collaborative-filtering (consultat: 21abr2025).

Per al cas *user-based*: l'usuari objectiu és la Taylor, que ha valorat el llibre Northanger Abbey amb un 1, Wuthering Heights amb un 2 i Oroonoko amb un 3, però no ha valorat

Bondswoman's Narrative. Comparant les valoracions, veiem que la Loren té un patró semblant: té les mateixes puntuacions per Northanger Abbey i Wuthering Heights, i només difereixen en un punt per Oroonoko. Això indica que la Taylor i la Loren tenen gustos similars. La Loren ha valorat la novel · la Bondswoman's Narrative amb un 5, que és una valoració alta. Per tant, com que la Loren i la Taylor tenen una similitud elevada, el sistema recomanarà Bondswoman's Narrative a la Taylor.

Per al cas *item-based*: l'usuari objectiu és l'Ainsley, que ha valorat Wuthering Heights amb un 4, Oroonoko amb un 3 i Bondswoman's Narrative amb un 1, però no ha valorat Northanger Abbey. Analitzant les valoracions, veiem que Wuthering Heights i Northanger Abbey tenen un patró semblant entre els usuaris: les seves valoracions difereixen només en un punt per l'Alex, la Loren i la Taylor. Quan la valoració de Wuthering Heights és alta, la de Northanger Abbey també ho és. De la mateixa manera, quan la puntuació de Wuthering Heights és baixa, la de Northanger Abbey també ho és. Això suggereix que són llibres similars en termes de preferències. Com que l'Ainsley ha valorat Wuthering Heights amb una puntuació alta, el sistema prediu que li pot agradar Northanger Abbey i l'hi recomana.

Així doncs, en ambdós casos s'utilitzen les valoracions de la matriu usuari-ítem per trobar les similituds, però amb enfocaments diferents: en el cas *user-based* les similituds es calculen entre usuaris, mentres que en el cas *item-based* les similituds es calculen entre articles.

Per determinar la similitud entre usuaris o articles es calcula la distància a l'espai vectorial. El mètode més usat per calcular aquesta proximitat és el coeficient de correlació de Pearson [44].

- Coeficient de correlació de Pearson (CCP)
  Es calcula la correlació entre les valoracions de dos individus o articles.

$$CCP = \frac{\sum(r_{ai} - \bar{r}_a)(r_{bi} - \bar{r}_b)}{\sqrt{\sum(r_{ai} - \bar{r}_a)^2}\sqrt{\sum(r_{bi} - \bar{r}_b)^2}}$$

  r_{ai} i r_{bi} són les puntuacions de l'usuari A o B a l'article i, respectivament.

També es podria fer servir la similitud del cosinus, explicada a l'apartat de "Similitud d'imatges", a l'inici del Marc teòric.

Per tant, l'esquema general del procediment de l'enfocament basat en memòria queda de la següent manera: s'utilitza una matriu per mapar el comportament de cada individu (files) vers cada article (columnes). El sistema extreu aquests valors de la matriu com a punts de dades en un espai vectorial, i es calcula la distància entre els punts per obtenir la similitud individu-individu o article-article. Amb els individus o articles més similars, es calcula la predicció de la valoració dels articles no vistos, i es recomanen aquells amb prediccions bones.

Enfocament basat en un model

L'enfocament basat en un model usa tècniques d'aprenentatge automàtic per trobar patrons a la base de dades i aprendre d'aquests patrons, generant un model amb el que oferir les recomanacions. Aquest enfocament és especialment útil quan la base de dades conté un gran nombre d'usuaris i ítems, ja que els càlculs de similitud de l'enfocament basat en memòria seran molt costosos a nivell computacional.

Hi ha diversos models per plantejar aquest enfocament. En el nostre treball, ens centrarem únicament en dos: SVD (*Singular Value Decomposition*) i ALS (*Alternating Least Squares*). Tots dos mètodes busquen factoritzar la matriu de valoracions dels usuaris, reduint-ne la dimensionalitat, tot i que de maneres diferents.

SVD descompon la matriu usuari-ítem en tres matrius [46]:

$$R \approx U \cdot S \cdot V^T$$

R és la matriu usuari-ítem, U és una matriu amb els usuaris com a files i els factors latents com a columnes, S és una matriu diagonal amb els valors singulars, que determinen els factors latents, i V és una matriu amb els ítems com a files i els factors latents com a columnes [47].

Es conserven només els *top-k* valors singulars i els seus vectors corresponents. D'aquesta manera es redueix el cost compuatacional sense perdre massa informació.

Un cop tenim les matrius, podem trobar usuaris similars mesurant la similitud de les seves representacions latents respectives a U. Per fer el mateix per a ítems, calculem la similitud de les represenatcions latents a V. També podem obtenir recomanacions predint com valoraria un usuari determinat un article que no ha vist: multiplicant la representació latent a U de l'usuari per la representació latent a V de l'article [47].

ALS descompon la matriu usuari-ítem en dues matrius [48]:

$$R \approx U \cdot P^T$$

R és la matriu usuari-ítem, U és una matriu dels factors latents pels usuaris i P és una matriu dels factors latents pels ítems.

L'algoritme funciona alternant l'optimització de U i de P. Primer fixa P i optimitza U minimitzant l'error entre la matriu original i la reconstruida. Després, fa el mateix pas però a l'inversa: fixa U i optimitza P, minimitzant el mateix error [49]. Les matrius òptimes són aquelles que minimitzen l'error de mínims quadrats (diferències entre els valors observats i els predits al quadrat).

Per obtenir les prediccions es multipliquen les dues matrius U i P^T per a cada parell usuari-ítem per al que volguem predir la valoració.

3.7 Proves àcides

Un altre pas clau que haurem de realitzar en el nostre projecte és una prova àcida. Es tracta d'un experiment breu però d'elevada exigència que es duu a terme abans de desenvolupar un projecte de ciència de dades complet, per comprovar que el plantejament i els supòsits són vàlids.

Per al nostre cas particular, la primera prova àcida que aplicarem consisteix en entrenar varis prototips de sistemes recomanadors amb un subconjunt de les dades i verificar si se supera un llindar objectiu per les mètriques d'èxit que escollim, així com determinar quin prototip presenta millors mètriques. Si la prova àcida fallés, seria necessari replantejar l'enfocament del projecte.

La segona prova àcida que aplicarem consisteix en provar el sistema recomanador escollit a la prova anterior i usar-lo per proveïr recomanacions a una mostra reduïda d'individus, que hauran de puntuar les recomanacions. De nou, caldrà verificar si se supera un llindar objectiu per les mètriques d'èxit que escollim, i si la prova àcida fallés, seria necessari replantejar l'enfocament del projecte.

3.8 Regles d'associació

Per entendre els hàbits de consum dels usuaris aplicarem regles d'associació. Es tracta d'una tècnica de mineria de dades que permet descobrir relacions, patrons, associacions o correlacions en un conjunt de dades [50]. En aquest conjunt de dades, cada observació (cada fila) s'anomena transacció.

Les regles d'associació són implicacions de la forma $X \rightarrow Y$, on X i Y són dos subconjunts de tots els elements disponibles, anomenats *itemsets* [51]. Un *itemset* és una col · lecció d'un o més ítems trobats al conjunt de dades. Per al nostre cas, serà una col · lecció d'un o més articles de roba o característiques demogràfiques dels usuaris. L'*itemset* X s'anomena antecedent, i l'*itemset* Y és el consegüent. La regla ens indica que quan apareix X, sol aparèixer també Y.

Per posar un exemple pràctic: la figura següent mostra una taula amb les dades de cinc transaccions en un supermercat [50].

Transaction ID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Diapers, Beer
T3	Milk, Diapers, Beer, Cola
T4	Bread, Milk, Diapers, Beer
T5	Bread, Milk, Diapers, Cola

Figura 3.5: Mostra d'una taula de transaccions. *Medium*, 2024. Disponible a: https://medium.com/image-processing-with-python/fundaments-of-associate-rule-mining-468801ec0a29 (consultat: 22abr2025).

Cada fila representa una transacció (la compra d'un client). La columna *Items Purchased* (articles comprats) detalla els productes adquirits a cada transacció. En aquest cas, una possible regla d'associació seria *Bread*→*Milk* (Pa→Llet). És probable que els clients que compren pa també comprin llet.

Les regles d'associació que es troben han de satisfer unes restriccions definides per l'usuari en tres mètriques clau: el suport, la confiança i el *lift* [50].

- Suport
  El suport mesura la freqüència en què apareixen tant l'element X com l'Y junts al conjunt de dades. És la proporció de transaccions (observacions) que contenen tant X com Y respecte al total de transaccions.

$$Suport = \frac{\text{Freqüència(X,Y)}}{\text{Total transaccions}} = P(X,Y)$$

  Aquest valor oscil · la entre 0 i 1. Un suport elevat indica que els *itemsets* són comuns a les transaccions.

- Confiança
  La confiança mesura la probabilitat de trobar el consegüent (Y) en transaccions sota la condició de que les transaccions també continguin l'antecedent (X).

$$Confiança = \frac{\text{Freqüència(X,Y)}}{\text{Freqüència(X)}} = \frac{\text{Suport(X,Y)}}{\text{Suport(X)}} = \frac{P(X,Y)}{P(X)} = P(Y/X)$$

  Aquest valor també oscil · la entre 0 i 1. Una confiança elevada indica que el consegüent i l'antecedent tendeixen a aparèixer junts.

- *Lift*
  El *lift* mostra amb quanta més freqüència trobem l'antecedent i el consegüent de la

regla junts que si fossin estadísticament independents. Això ens ajuda a mesurar la força d'associació entre X i Y.

$$Lift = \frac{\text{Suport(X,Y)}}{\text{Suport(X)Suport(Y)}} = \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y/X)}{P(Y)}$$

Aquest valor com a mínim pot ser 0 i no existeix un valor màxim. Un *lift* menor a 1 ens pot indicar que la relació és a l'inversa: en comptes de tenir $X \rightarrow Y$, tenim que $Y \rightarrow X$. Un *lift* igual a 1 indica que X i Y són independents. Un *lift* superior a 1 indica que X i Y són dependents. Com més gran sigui el *lift*, més valor té i més forta és la regla d'associació.

L'algoritme usat per trobar les regles d'associació és l'algoritme apriori. Primerament, l'algoritme compta el nombre d'ocurrències per cada *itemset* de només un element. Tots aquells que no superin el suport mínim especificat queden descartats. A continuació, combina els *itemsets* que no han quedat descartats en *itemsets* de 2 ítems. Torna a fer-ne un recompte i descarta aquells que no assoleixin el suport mínim establert. L'algoritme va repetint aquest esquema, combinant *itemsets* cada vegada de major tamany i descartant aquells que són infreqüents, fins que es generen tots els *itemsets* freqüents que compleixen el suport mínim [52]. Cada iteració de l'algoritme genera associacions més complexes. L'algoritme es basa en el principi que qualsevol subconjunt d'un *itemset* freqüent també ha de ser freqüent.

Capítol 4

Treball de camp

4.1 Fonts de dades

Per iniciar el treball hem partit d'una base de dades d'imatges extreta de Kaggle [1]. Aquesta base de dades inicial conté 16.186 imatges de peces de roba, dividides en sis carpetes, que corresponen a sis tipus de peces: dessuadores per homes (2.144 imatges), dessuadores per dones (1.210 imatges), samarretes de màniga llarga (1.700 imatges), samarretes de màniga curta (4.790 imatges), dessuadores sense caputxa per homes (3.886 imatges) i dessuadores sense caputxa per dones (2.456 imatges). Es tracta de fotografies de peces de roba de l'*e-commerce* Zalando. Algunes de les imatges presenten un model amb la peça, i d'altres són únicament de la peça de roba.

A continuació, s'han seleccionat de manera aleatòria 100 imatges d'entre les 16.186 totals. La clau d'aquesta selecció aleatòria ha estat mantenir la distribució de la base de dades original, seleccionant les imatges de manera proporcional a la mida de cada una de les sis carpetes sobre el total, usant el següent esquema:

$$\text{Nº imatges a seleccionar de la carpeta } i = \frac{\text{nº imatges carpeta } i}{\text{total imatges}} \cdot 100$$

Per exemple, per al cas de dessuadores per a homes tindrem unes 13 ($2.144/16.186 \cdot 100$) imatges. D'aquesta manera assegurem un mostreig adequat de les imatges, i no perdem la proporció de cada tipus de peces present al *dataset* original.

D'altra banda, cal esmentar que l'elecció de només treballar amb 100 imatges ve donada pel disseny de l'enquesta que s'explicarà a continuació. Com que la previsió de respostes rondava els 100 usuaris, treballar amb 100 imatges és idoni per a poder desenvolupar correctament el recomanador.

Aquesta selecció aleatòria d'imatges respectant les proporcions de les carpetes s'ha fet mitjan-

[1] https://www.kaggle.com/datasets/dqmonn/zalando-store-crawl/data

çant un script amb *RStudio*.

El següent pas ha estat crear una enquesta dinàmica que permeti als usuaris ordenar 10 peces de roba segons els seus gustos. D'aquesta manera podem obtenir les dades de preferències dels usuaris i desenvolupar, més endavant, el recomanador, així com trobar regles d'associació. La clau de l'enquesta és que ha estat programada de manera que cada cop que s'obre apareixen 10 imatges diferents, agafades de manera aletòria d'entre les 100 imatges que hem mostrejat anteriorment. Tot i que les imatges que veu cada persona són diferents, les peces es van repetint en diferents enquestes. Això ens permet tenir múltiples opinions sobre cada peça, però en contextos diferents. Addicionalment, a l'enquesta s'han recollit dades demogràfiques dels individus, que es detallen més endavant.

Aquesta enquesta ha estat desenvolupada programant amb *Python*, usant l'editor de codi *Visual Studio Code* i treballant principalment amb la llibreria *Streamlit*. Per a aprendre a programar en aquest llenguatge s'ha comptat amb el suport de companys del grau de Ciència i Enginyeria de Dades.

Al treballar d'aquesta manera l'enquesta només era accesible de manera local, és a dir, només s'hi podia accedir des de l'ordinador en el qual s'executava el codi. Per tal de fer-la accessible per a tothom, s'ha penjat el codi usat i la carpeta amb les 100 imatges mostrejades anteriorment a la plataforma *Github*, i s'ha usat la plataforma *Streamlit Community Cloud* per a desplegar l'enquesta. D'aquesta manera hem obtingut un enllaç [2] a l'aplicació d'*Streamlit Community Cloud*, fent que qualsevol persona amb l'enllaç pugui respondre el qüestionari. Finalment, s'ha programat de manera que les respostes es vagin emmagatzemant en un únic fitxer a *Dropbox*.

Un cop explicats els aspectes tècnics de la creació i desplegament de l'enquesta, passem a comentar l'estructura i les preguntes de la pròpia enquesta. L'enquesta consta de dues parts: un primer apartat amb tres preguntes de caràcter demogràfic i una segona part per ordenar les 10 peces de roba. Les preguntes demogràfiques són:

- Gènere. Pregunta de resposta única, amb un desplegable amb 3 opcions: Home, Dona o Altres.
- Edat. Pregunta de resposta única, que permet introduir un número de l'1 al 99.
- Pregunta de preferència de compra de roba. Pregunta de resposta única, on es preguntava als usuaris si preferien comprar roba físicament en botiga, de manera *online*, o d'ambdues formes per igual.

La segona part, consistent en ordenar les 10 imatges de peces de roba, ofereix una primera secció on ordenar les etiquetes de les imatges, i una segona secció on es mostren les imatges segons el rànquing actual. Cada cop que es canvia l'ordre de les etiquetes, l'ordre de les imatges s'actualitza d'acord amb el rànquing. Les respostes de l'enquesta interactiva es guarden en un fitxer que conté 13 columnes: les deu primeres corresponen als rànquings de l'enquesta (de l'1 al 10), i les 3 darreres corresponen a les tres preguntes demogràfiques. Les variables de les 10

[2] https://clothing-rank-formapp-bm4aqc3r2pwxqqcfvwewxu.streamlit.app

posicions del rànquing emmagatzemen el nom de la imatge colocada en la posició corresponent. Per exemple: si una usuària ha posat la imatge "UR622J001-K13@8.jpg" a la posició 6, a la fila corresponent a les seves dades i a la columna 6 hi haurà el valor "UR622J001-K13@8.jpg".

Finalment, l'últim dels preparatius per poder passar a desenvolupar el recomanador i estudiar les regles d'associació consisteix en realitzar una descriptiva de les 100 imatges amb les que treballem. S'ha recollit informació sobre el color principal de la peça, el tipus d'estampat (llis o estampat) i el tipus de peça (cadascuna de les 6 carpetes mencionades a l'inici). A més, en cas de tractar-se d'una imatge on la peça de roba ve modelada per un model, s'ha anotat el gènere i l'edat aproximada de la persona, per estudiar si suposa alguna mena de biaix pels usuaris a l'hora d'ordenar les peces de roba. També s'ha anotat, en cas de veure's a la fotografia, la marca de la peça de roba.

A mode de resum, comptem amb 3 fonts de dades principals per desenvolupar el recomanador i estudiar les regles d'associació:

- La base de dades inicial, a partir de la qual ens quedem amb una mostra de 100 imatges.
- Les respostes de l'enquesta interactiva.
- La descriptiva de les 100 imatges.

Tot el codi usat per obtenir la mostra de 100 imatges, desenvolupar l'enquesta interactiva, desenvolupar els recomanadors i analitzar les regles d'associació es pot trobar al següent enllaç a Github: https://github.com/claratg15/clothing-rank-form_STREAMLIT.

4.2 Preprocessament

Per poder tenir totes les dades tabulars necessàries en un format que ens permeti treballar-les de manera adient, duem a terme un preprocessament adequat. Dividim el preprocessament en dues parts:

- Creació d'un *dataset* combinant les respostes de l'enquesta amb les dades de la descriptiva de les imatges, per realitzar una anàlisi exploratòria apropiada.
- Adaptació de les respostes de l'enquesta i de la descriptiva de les imatges a un format adient per a poder desenvolupar el recomanador.

Al *dataset* resultant de la primera part del preprocessament cada fila correspon a una resposta de l'enquesta. Com a columnes, tenim les 3 variables demogràfiques (gènere, edat i preferència de compra). També tenim la marca, el color, el tipus d'estampat, el gènere del model, l'edat del model i el tipus de peça de roba per les imatges que ha col · locat cada usuari en cada posició del rànquing. Així doncs, cadascuna d'aquestes variables mencionades té 10 columnes al dataset, una per cada posició del rànquing. S'indica l'atribut de la imatge col · locada en el rànquing corresponent. Per exemple, l'atribut "Color" té 10 columnes anomenades "Color Rànking1", "Color Rànking2", fins a "Color Rànking10". Cada

columna indica el color principal de la peça de roba que l'usuari ha col · locat en la posició corresponent del rànquing. Finalment, tenim 100 columnes més, cadascuna amb el nom d'una imatge. Les cel · les d'aquestes columnes indiquen el rànquing en què cada usuari ha posat la imatge corresponent. Per aquelles imatges que l'usuari no ha vist, hi consta un *NA*.

Addicionalment, hem realitzat la imputació de tres valors erronis detectats: la variable "Edat", recollida com a pregunta demogràfica de l'enquesta, tenia tres registres amb edat igual a 1. Com que no és possible que un individu que hagi respost l'enquesta tingui tan sols 1 any d'edat, s'ha determinat que es tracta d'un error. Per tant, hem establert aquests valors com a *NA* i els hem imputat amb el mètode MICE. L'algoritme seguit per aquest mètode està detallat al Marc teòric, al subapartat "Imputació de valors faltants".

Pel que fa a la segona part del preprocessament, partim de la base de dades preparada a la primera part, que conté els valors de la variable Edat imputats. Creem una nova taula on, de nou, cada fila representa una resposta de l'enquesta. El nombre de columnes, però, es redueix: per cada atribut de les imatges, eliminem les 10 columnes corresponents a cada posició del rànquing. Ens quedem només amb les variables demogràfiques (gènere, edat i preferència de compra) i les 100 columnes indicant el rànquing en què cada usuari ha col · locat la imatge corresponent. Aquesta base de dades rebrà el nom "valoracions", per a identificar-la més endavant.

Finalment, adaptem la base de dades de la descriptiva d'imatges pel seu ús en la creació del recomanador: normalitzem totes les seves variables, excepte l'Edat del model, mitjançant codificació *one hot*, explicada al Marc teòric. Aquesta base de dades rebrà el nom "atributs", per a identificar-la més endavant.

4.3 Anàlisi descriptiva exploratòria

Un cop preprocessades les dades, passem a realizar un estudi descriptiu de les nostres variables. Aquesta anàlisi descriptiva es dividirà de la següent manera: una anàlisi univariant de la descriptiva de les imatges, una anàlisi univariant de les respostes de l'enquesta i una anàlisi bivariant creuant la descriptiva de les imatges i les respostes de l'enquesta. Totes aquestes anàlisis es realitzaran a partir de la base de dades treballada a la primera part del preprocessament.

4.3.1 Anàlisi univariant

Descriptiva de les imatges

Presentem una breu descriptiva de les característiques principals de les 100 imatges amb les que treballem.

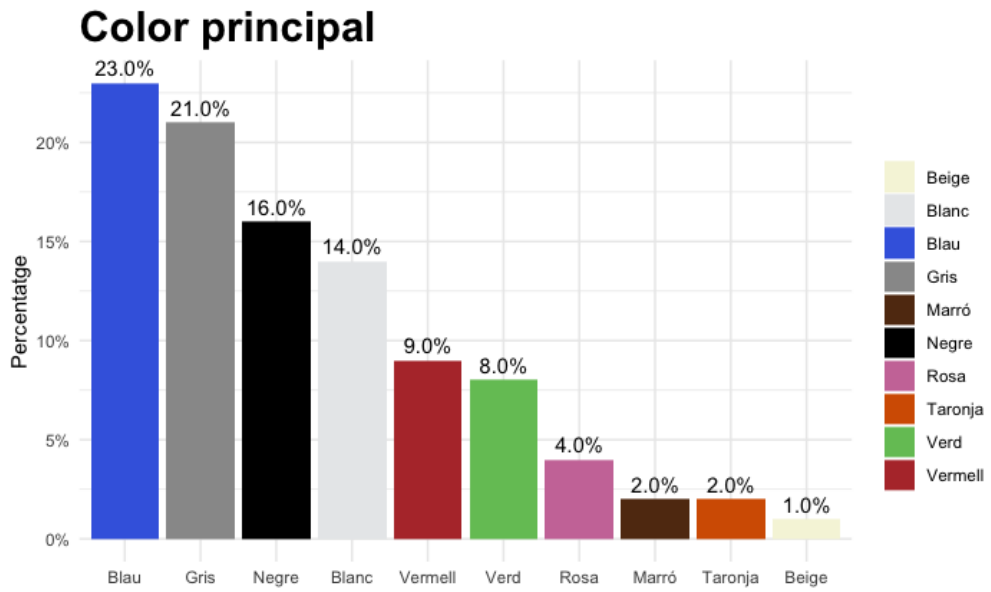

Figura 4.1: Distribució dels colors.

Observem que els colors predominants a les imatges són el blau, el gris, el negre i el blanc. Els colors més llampants, com el rosa, el taronja o el verd, es troben poc representats.

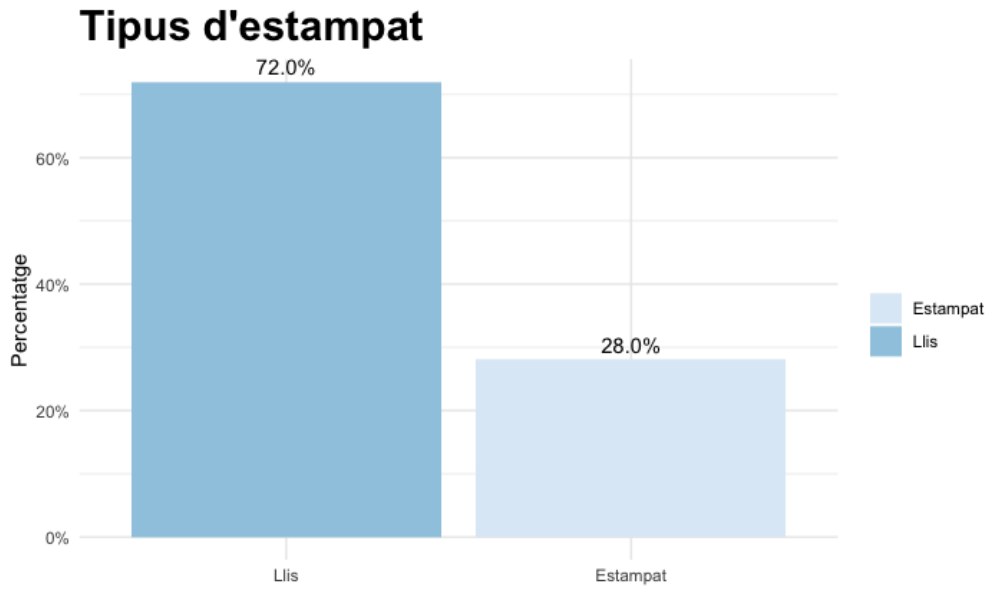

Figura 4.2: Distribució del tipus d'estampat.

La majoria de les peces de roba presenten un estampat llis.

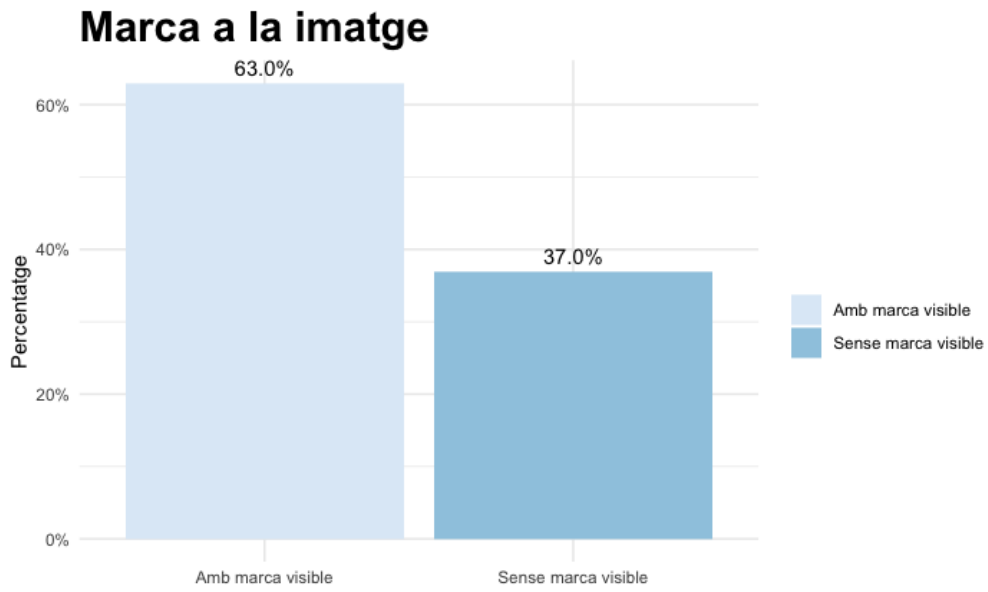

Figura 4.3: Distribució de la presència de marca.

Per analitzar l'efecte de la marca s'ha dicotomitzat la variable, segons si hi ha una marca visible a la imatge o no. Com s'observa al gràfic, la majoria de les peces de roba contenen el logotip d'alguna marca.

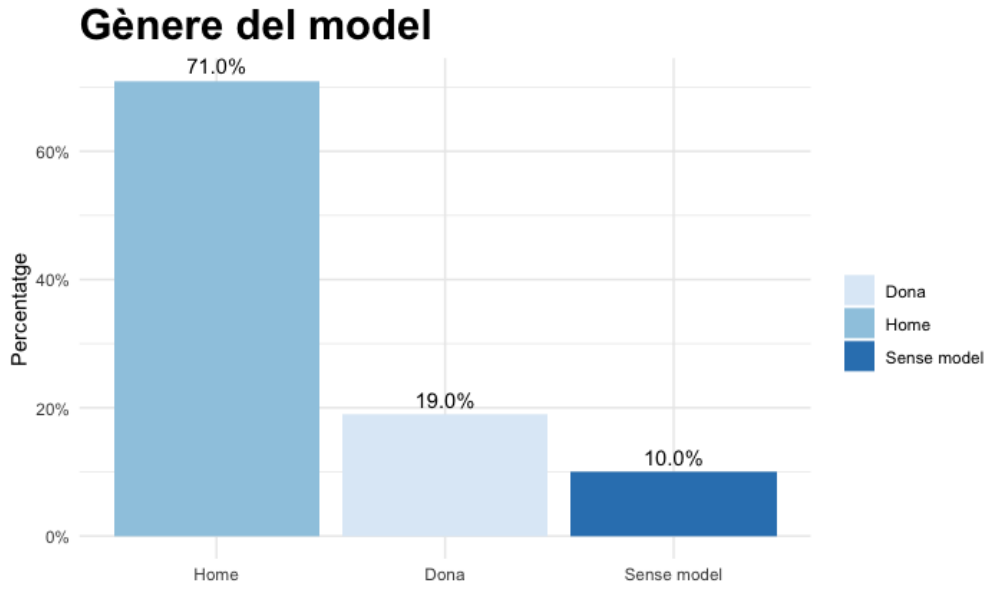

Figura 4.4: Distribució del gènere dels models.

Veiem com les imatges de la base de dades que no tenen model són poques, tan sols un 10%. De les que sí tenen model, la gran majoria tenen un model home. Més endavant estudiarem si això pot suposar alguna mena de biaix a l'hora d'ordenar les peces de roba, sobretot per les dones.

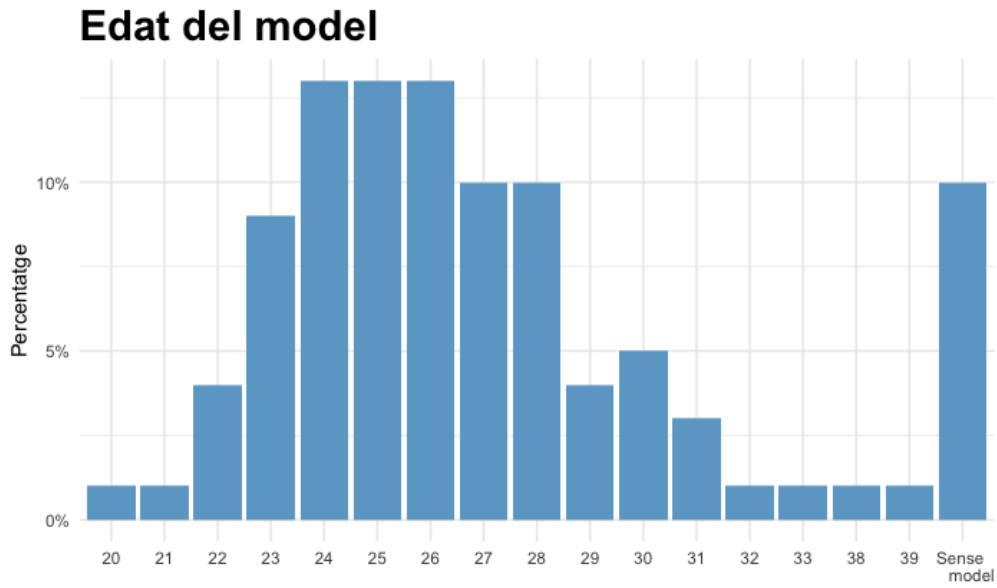

Figura 4.5: Distribució de l'edat dels models.

Finalment, pel que fa a l'edat, veiem que la majoria dels models de les imatges semblen tenir una edat compresa entre 23 i 28 anys.

Respostes de l'enquesta

Iniciem aquesta anàlisi exploratòria amb les dades demogràfiques dels enquestats: gènere, edat i preferència de compra.

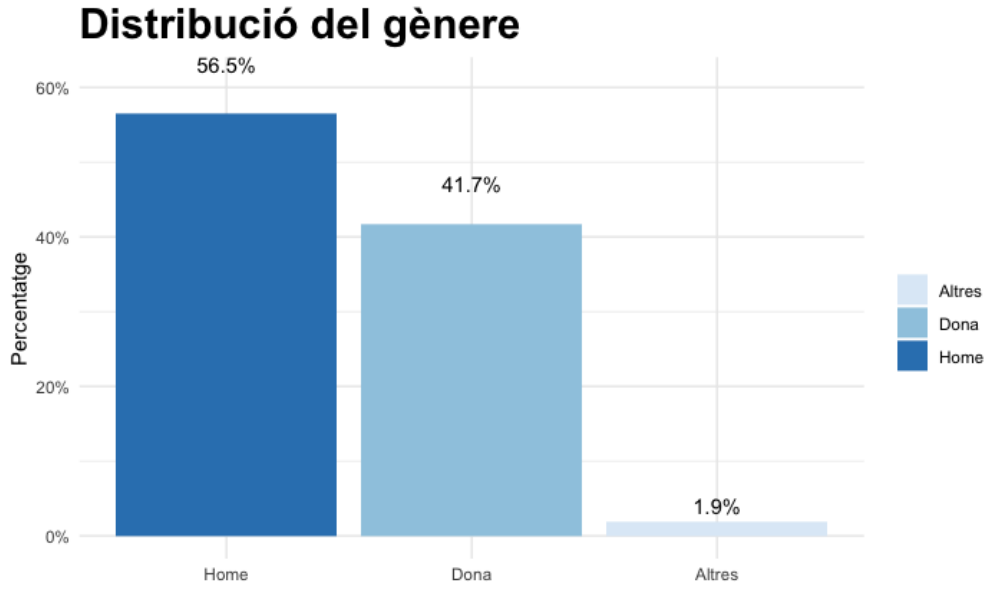

Figura 4.6: Distribució del gènere dels enquestats.

Observem que la distribució entre homes i dones està bastant equilibrada: un 42% de dones i un 57% d'homes, aproximadament. En canvi, els gèneres agrupats sota la categoria "Altres" representen un percentatge molt baix dels enquestats.

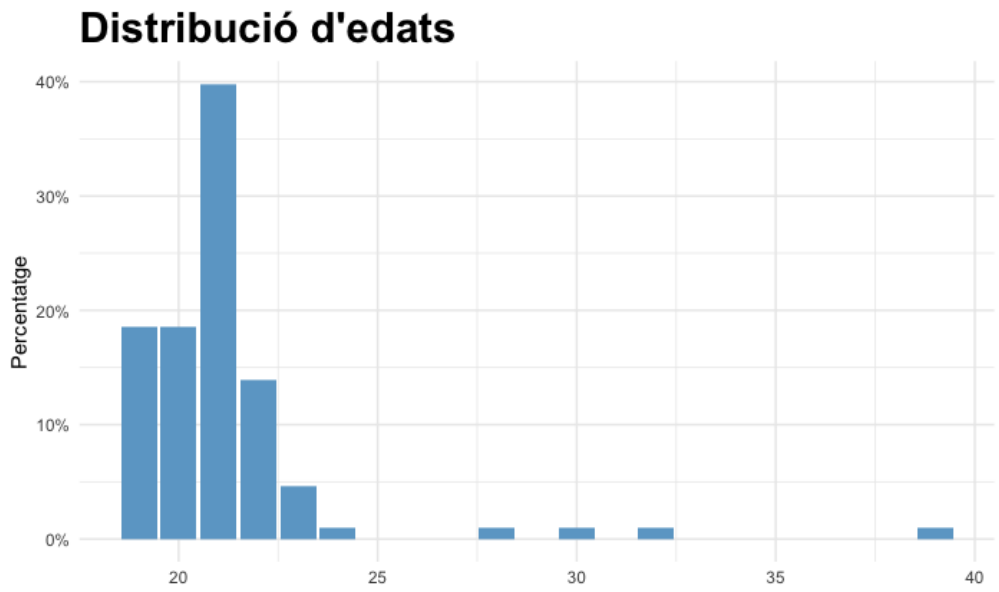

Figura 4.7: Distribució de l'edat dels enquestats.

Mínim	Mitjana	Mediana	Màxim
19	21'12	21	39

Taula 4.1: Mesures descriptives de l'edat dels enquestats.

Pel que fa a l'edat, veiem que la majoria dels enquestats tenen entre 19 i 22 ays, concentrant-se la majoria en els 21 anys. L'edat mínima és 19 i la màxima 39.

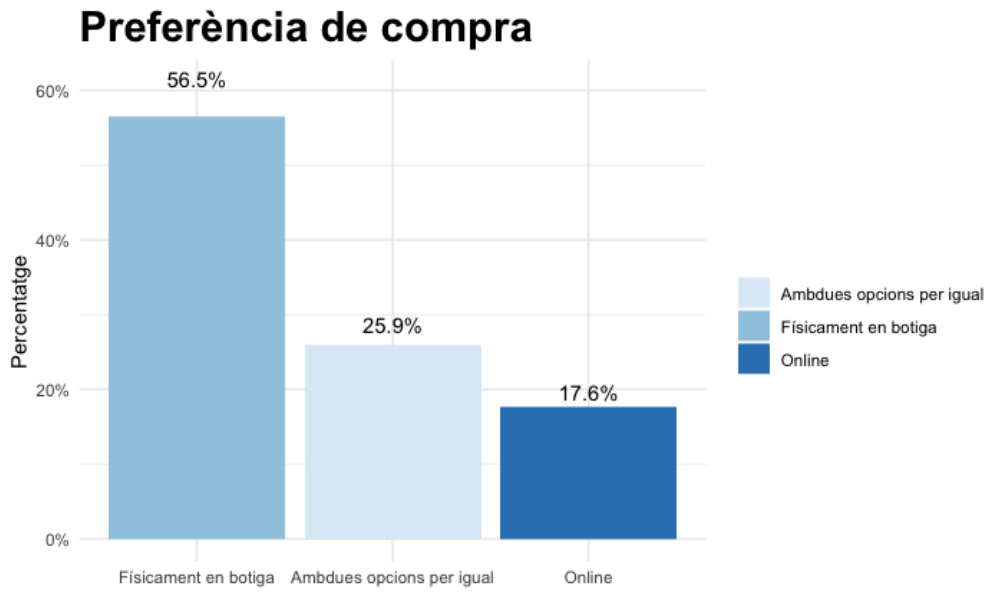

Figura 4.8: Distribució de la preferència de compra dels enquestats.

Respecte la preferència de compra, veiem una clara guanyadora: més de la meitat dels enquestats prefereixen comprar físicament en botiga. Un quart dels enquestats els agraden les dues opcions per igual, i tan sols un 18%, aproximadament, prefereix exclusivament *online*.

Tot seguit, hem analitzat els colors, el tipus d'estampat i la presència de marca per les tres primeres posicions dels rànquings i per les tres últimes. Només adjuntarem el gràfic per la primera posició en cada cas, degut a la presència d'un patró repetitiu entre gràfics. La resta de gràfics es poden trobar a l'annex.

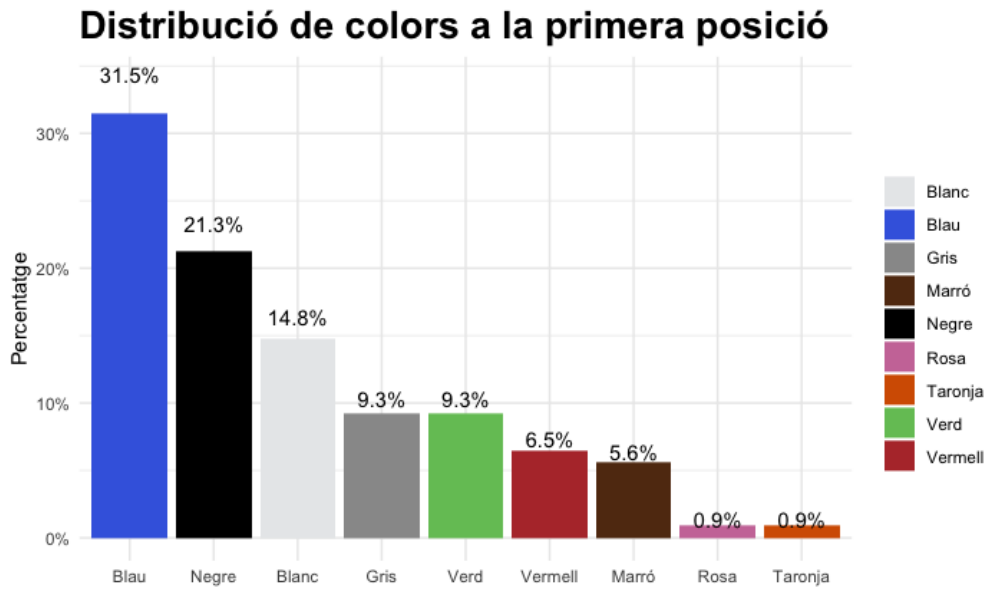

Figura 4.9: Distribució dels colors de les peces de roba a la primera posició dels rànquings.

Els colors de les peces de roba situades en la primera posició dels rànquings mantenen el patró vist a la descriptiva de les imatges: com que la majoria de les peces són de color blau, blanc, negre o gris, la majoria de les peces de roba col · locades en aquest rànquing també ho son.

Tot i així, veiem que les proporcions dels colors blau, negre i blanc viu un cert augment. Per exemple: el blau representa el 23% a la descriptiva, però a la posició 1 del rànquing representa un 31'5%.

En canvi, colors més vibrants com ara el taronja, el vermell o el rosa, pateixen un descens: tot i ja estar poc representats a la mostra, encara ho estan menys a la primera posició del rànquing. Per exemple: el rosa representa un 4% a la descriptiva, però a la posició 1 del rànquing representa un 0'9%.

Tot i que tinguem més imatges de peces de roba de colors blau, negre i blanc, i això es vegui representat als rànquings, els usuaris s'inclinen cap aquests colors de totes maneres.

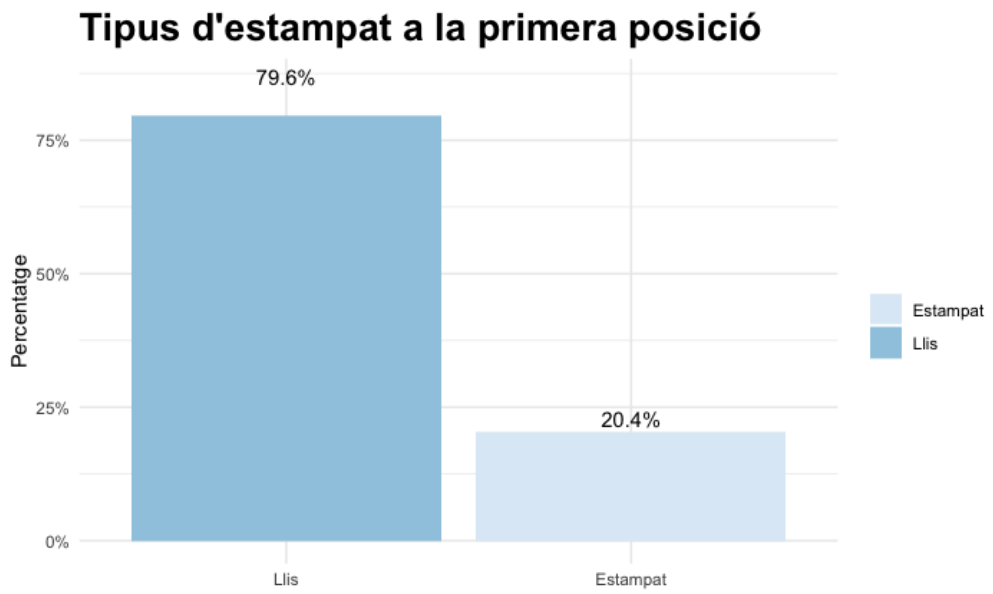

Figura 4.10: Distribució del tipus d'estampat de les peces de roba a la primera posició dels rànquings.

Pel que fa a l'estampat, el patró vist a la descriptiva es manté: la majoria de les peces de roba són llises, sense un estampat clar. A la primera posició del rànquing, la gran majoria de les peces també son llises.

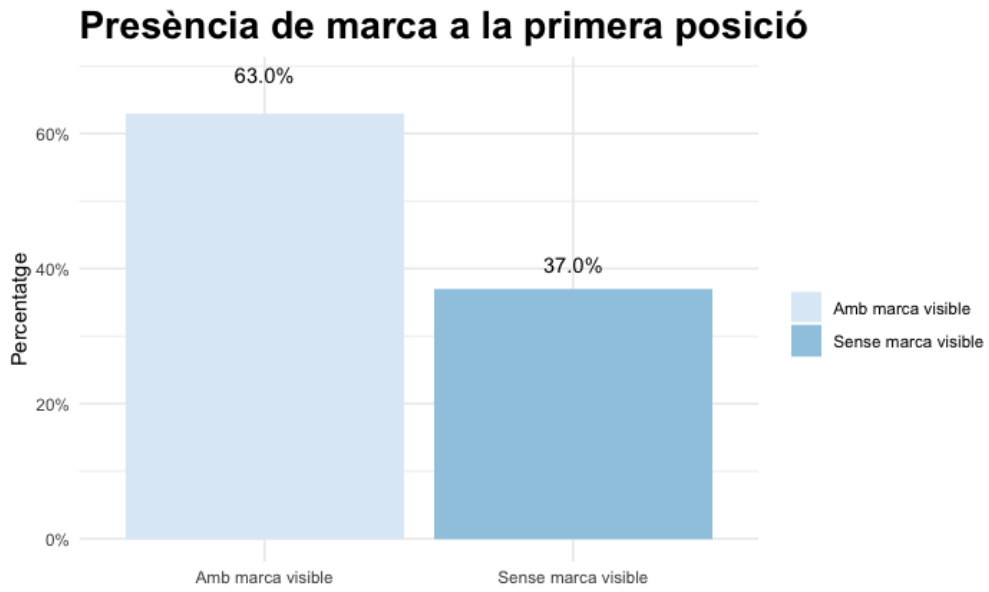

Figura 4.11: Distribució de la presència de marca de les peces de roba a la primera posició dels rànquings.

De nou, el patró vist a la descriptiva es manté: la majoria de les peces de roba tenen una marca visible, i a la primera posició del rànquing també.

Finalment, tot i que es pugui sobreentendre a partir de l'explicació del procediment i l'estructura de l'enquesta, cal mencionar que tenim un gran escasseig de dades a la part dels rànquings: com que cada usuari només veu 10 imatges de les 100 totals, tenim un 90% de valors faltants a la part de la matriu usuari-ítem.

4.3.2 Anàlisi bivariant

Iniciem l'anàlisi bivariant creuant els colors a les tres primeres posicions dels rànquings i el gènere dels enquestats. Per a estudiar la possible relació entre les dades, com que es tracta de dues variables categòriques, realitzem un test χ^2, obtenint els següents resultats:

Estadístic	Graus de llibertat	P-valor
15'90	16	0'46

Taula 4.2: Resultats del test χ^2 entre el color de les peces de roba i el gènere dels enquestats.

Obtenim un p-valor molt superior al nivell de significació (0'46>0'05). Per tant, no tenim prou evidència per afirmar que hi ha alguna relació entre els colors de les peces de roba classificades a les tres primeres posicions i el gènere dels enquestats.

En segon lloc, estudiem si hi ha alguna correlació entre l'edat dels enquestats i l'edat del

model de les imatges. Podria ser que els usuaris més joves tendissin a col · locar en posicions més baixes del rànquings aquelles peces amb un model més gran. Per a estudiar-ho, excloem les dades de les imatges sense model i calculem el coeficient de correlació de Pearson entre tots dos grups de dades, ja que es tracta de dues variables numèriques. Obtenim una correlació igual a -0'017. Al tractar-se d'una correlació molt propera a 0, determinen que no hi ha correlació entre l'edat dels enquestats i l'edat del model de la imatge.

Finalment, examinem si hi ha alguna relació entre el gènere dels enquestats i el gènere dels models de les imatges. Podria passar que les dones tendissin a posicionar millor al rànquing les peces de roba modelades per dones, i que els homes fessin el mateix amb les peces de roba modelades per homes. Com que es tracta de dues variables categòriques, realitzem un test χ^2, obtenint els següents resultats:

Estadístic	Graus de llibertat	P-valor
1'35	4	0'85

Taula 4.3: Resultats del test χ^2 entre el gènere dels enquestats i el gènere dels models.

Obtenim un p-valor molt superior al nivell de significació (0'85>0'05). Per tant, no tenim prou evidència per afirmar que hi ha alguna relació entre el gènere dels enquestats i el gènere dels models.

Es poden trobar gràfics complementaris per a aquests tres casos d'estudi de l'anàlisi bivariant a l'annex.

4.4 Tractament d'imatges

El primer pas més directament relacionat amb el desenvolupament del recomanador és el càlcul de la similitud d'imatges. Per fer-ho, s'ha seguit el mètode de semblança basada en característiques, detallat al Marc teòric. Per l'extracció dels vectors de característiques de les imatges s'han aplicat 4 mètodes diferents: tres models preentrenats (*VGG16*, *ResNet50* i *EfficientNet*) i *autoencoders*.

Per valorar quin mètode extreu els vectors de característiques més acurats, s'han entrenat els models i extret els vectors de característiques de totes les 100 imatges segons cada mètode. S'han seleccionat dues imatges de la base de dades considerades similars i dues imatges considerades diferents, i s'ha calculat la similitud del cosinus (explicada al Marc teòric) per cada cas. A continuació es presenten les imatges seleccionades i una taula amb els resultats de similitud obtinguts.

Figura 4.12: Imatges seleccionades de peces de roba similars.

Figura 4.13: Imatges seleccionades de peces de roba diferents.

Mètode	*VGG16*	*ResNet50*	*Autoencoders*	*EfficientNet*
Similitud imatges similars	0.499	0'492	0'85	0'74
Similitud imatges diferents	0.500	0'493	0'53	0'66

Taula 4.4: Resultats de la similitud del cosinus per cada casuística.

El mètode que presenta una major diferència entre la similitud de les peces de roba semblants i les diferents és *autoencoders*. Per tant, per desenvolupar el recomanador usarem els vectors de característiques de les imatges extrets mitjançant *autoencoders*. Concretament, entrenarem un model amb *autoencoders* però ens quedarem només amb la part de l'*encoder*, tal com s'explica al Marc teòric. Pel que fa als tres models preentrenats, com que no han estat seleccionats, a l'apartat de Marc teòric no s'entra en detall en l'explicació del seu funcionament. Es citen els webs on es poden trobar explicats amb detall, en cas de ser d'interès.

4.5 Desenvolupament dels recomanadors

Primerament, es va plantejar el desenvolupament dels següents quatre sistemes de recomanació, corresponents als 4 subtipus de filtratge col · laboratiu mencionats al Marc teòric: enfocament *user-based*, enfocament *item-based*, SVD i ALS.

Inicialment, els casos *user-based* i *item-based* es van treballar simplement amb la matriu usuari-ítem amb els rànquings de les imatges, sense tenir en compte informació addicional per calcular la similitud entre usuaris o ítems. Posteriorment, es va decidir enriquir aquests dos mètodes afegint les característiques demogràfiques dels usuaris, per al cas *user-based*, i afegint els atributs de les peces de roba i la similitud d'imatges, per al cas *item-based*. Aquests plantejaments híbrids sorgeixen com a proposta de millora i d'aprofitament de tota la informació disponible, donat el percentatge elevat de valors faltants a la matriu usuari-ítem. Al final de l'apartat 4.5 es comparen els enfocaments inicials i els híbrids, per demostrar que els enfocaments híbrids ofereixen millors resultats. Per tant, als subapartats corresponents no s'explicaran els enfocaments inicials, sinó directament els híbrids.

Per aquests dos enfocaments híbrids, partim de les dues bases de dades preparades a la segona part del preprocessament: "valoracions" i "atributs". Treballarem amb les dades normalitzades. A la taula "valoracions", les columnes corresponents a les variables Gènere i Preferència de compra es normalitzen aplicant la codificació *one hot*, explicada al Marc teòric. La variable Edat i els rànquings es normalitzen mitjançant el mètode *Z-score*, explicat també al Marc teòric. Pel que fa a la taula "atributs", l'única variable que queda per normalitzar és l'Edat del model. Es normalitza a través del mètode *Z-score*, també.

Per als enfocaments SVD i ALS no s'ha treballat amb més informació que la matriu usuari-ítem, ja que la implementació d'aquests models està feta amb la llibreria *recommenderlab* d'R, dissenyada principalment per treballar únicament amb matrius de valoracions. Es podrien afegir les variables demogràfiques, però els models no intepretarien explícitament aquestes columnes com a dades demogràfiques i serien tractades com a part de la matriu, cosa que pot diluir el focus en les valoracions de les imatges. Per tant, es treballarà únicament amb els rànquings del *dataset* "valoracions", excloent les variables demogràfiques.

Així doncs, procedim a explicar el procediment seguit per als següents quatre sistemes de recomanació: *user-based* híbrid, *item-based* híbrid, SVD i ALS. Al final de l'apartat 4.5 es triarà el millor model d'entre aquests 4. També, com s'ha mencionat prèviament, es

compararan els resultats obtinguts en el plantejament inicial i el plantejament híbrid per als casos *user-based* i *item-based*.

4.5.1 *User-based*

Fonamentalment és un sistema de filtratge col · laboratiu basat en usuaris, millorat amb característiques del perfil de l'usuari: el gènere, l'edat i la preferència de compra. Aquesta millora fa que el sistema sigui més robust, sobretot tenint en compte l'escasseig de dades, ja que tenim un 90% de valors faltants.

El primer pas consisteix en calcular la similitud entre l'usuari escollit (l'usuari pel que volem obtenir les recomanacions) i per cadascun dels usuaris restants. Per fer-ho, plantegem un càlcul combinat de la similitud entre l'usuari x i l'usuari y:

- Similitud per rànquings. Calculem el coeficient de correlació de Pearson entre els vectors formats pels seus rànquings, tal com s'explica al Marc teòric.
- Similitud per edat. Calculem la similitud entre les seves edats:

  $$\text{Similitud per edat} = 1 - |edat_x - edat_y|$$

  Com més proper a 1 sigui aquest valor, més similars són les edats. Si les edats són molt distants, la similitud total es veurà penalitzada.
- Similitud per altres variables categòriques. Calculem la similitud segons el gènere i la preferència de compra, per cada variable *dummy*. Per exemple, pel cas de la variable *dummy* Dona, que és 1 si el gènere és dona i 0 si no:

  $$\text{Similitud}_{Dona} = \begin{cases} 1 & \text{si } \ Dona_x = Dona_y \\ 0 & \text{si } \ Dona_x \neq Dona_y \end{cases}$$

  S'adjudica 1 punt per cada variable *dummy* en la que tinguin el mateix valor.

Per agrupar la similitud de les variables demogràfiques en un únic valor, es calcula la mitjana de les similituds de cada variable *dummy* i de l'edat. A aquesta mitjana l'anomenarem similitud demogràfica.

Finalment, el valor final de similitud entre l'usuari x i l'y prové d'una mitjana ponderada:

$$\text{Similitud final}(x, y) = 0{,}6 \cdot \text{Similitud per rànquings } + 0{,}4 \cdot \text{Similitud demogràfica}$$

S'otorguen aquests pesos per mantenir l'essència de l'enfocament *user-based*, donant més importància als rànquings.

El segon pas es basa en identificar les 90 imatges de peces de roba que l'usuari escollit no ha vist a l'enquesta interactiva, ja que no volem recomanar-li peces que ja hagi vist.

El tercer pas consisteix en seleccionar els top-k individus més semblants a l'usuari escollit, mitjançant validació creuada, que tinguin una similitud positiva (una negativa indicaria que

tenen preferències oposades). Es duu a terme *5-fold cross validation*, agafant valors de k des d'1 fins a 15. Agafem les dades dels rànquings de l'usuari escollit, i creem una nova variable anomenada "fold" en la que assignem aleatòriament un valor d'1 a 5 a cada dada, que indica el *fold* de *cross*-validació corresponent. Per cada k d'1 a 15, i per cada *fold* d'1 a 5:

1. El grup de *test* seran les dades de l'usuari escollit i del *fold* corresponent.
2. El grup de *train* seran les dades de l'usuari escollit que no són del *fold* corresponent i les dades de la resta d'usuaris.
3. Agafem les dades d'una de les imatges del conjunt de test que tingui assignat el *fold* corresponent. Anomenarem a aquesta imatge t. Es tracta d'una imatge vista per l'usuari escollit.
4. Busquem els usuaris del conjunt de *train* que també han vist la imatge t. D'entre aquests usuaris, agafem els top-k amb una similitud positiva amb l'usuari escollit.
5. Agafem els rànquings que aquests top-k usuaris han donat a la imatge t.
6. Predim el rànquing en el que hauria posat la imatge t l'usuari escollit si no l'hagués vist. Per fer-ho, calculem una mitjana ponderada dels rànquings en què l'han posat els top-k usuaris més semblants que també l'han vist, donant més pes als rànquings dels usuaris més semblants.

$$\text{Predicció} = \frac{\sum \text{Rànquing} \cdot \text{Similitud}}{\sum \text{Similitud}}$$

7. Calculem l'RMSE entre el rànquing real i el predit. El càlcul de l'RMSE està detallat al Marc teòric.

Un cop fetes totes les iteracions (una per cada k i per cada *fold*), calculem la mitjana de l'RMSE dels *folds* de cada k. Ens quedarem amb aquella k que tingui un valor d'RMSE menor.

Cal mencionar que per assegurar certa robustesa en els resultats, només treballem amb aquelles imatges que hagin estat vistes per, com a mínim, 5 usuaris. Dins de les imatges que compleixin aquesta condició, volem tenir com a mínim 3 usuaris semblants a l'escollit que hagin vist la imatge. Si no es compleix alguna d'aquestes dues condicions, es descarta la imatge no vista amb la que s'està treballant.

Aleshores, un cop trobada la k òptima, per cada una de les 90 imatges trobades al segon pas seleccionem els top-k individus més semblants i que hagin vist la imatge de la peça de roba p, i calculem la mitjana ponderada dels rànquings en què cadascun d'aquests top-k individus ha posat la peça de roba p. Així obtenim una estimació del rànquing que hauria donat l'usuari escollit a la peça de roba p. L'esquema seguit és pràcticament igual al seguit quan s'ha fet la validació creuada.

El quart i últim pas consisteix en agafar les 10 peces de roba amb millor predicció. Degut a la normalització prèvia dels rànquings, les posicions més elevades al rànquing tenen valors més petits (fins i tot negatius). Per tant, busquem aquelles imatges no vistes que tinguin valors de predicció petits, cosa que es tradueix en posicions elevades al rànquing.

En el cas del desenvolupament del recomanador *user-based* incial, el procediment és exactament el mateix. L'únic que canvia és el càlcul de la similitud al primer pas: en comptes de computar una similitud combinada, el 100% del pes recau sobre la similitud per rànquings.

4.5.2 *Item-based*

En el cas del mètode *item-based*, a l'afegir dades de les característiques de les peces de roba ens trobem davant un enfocament híbrid que barreja el mètode *item-based* amb el filtratge basat en contingut. Recordem que l'enfocament basat en contingut utilitza els atributs d'un article per recomanar altres articles similars a aquells amb els que el client ha interactuat positivament.

El primer pas es basa en identificar les 90 imatges de peces de roba que l'usuari escollit no ha vist a l'enquesta interactiva, ja que no volem recomanar-li peces que ja hagi vist.

Per cadascuna de les 90 peces de roba trobades al primer pas, calculem la similitud entre la imatge p (p dins les 90) i cadascuna de les 10 imatges que sí ha vist l'usuari. Per fer-ho, plantegem un càlcul combinat de la similitud entre l'ítem x i l'ítem y:

- Similitud per rànquings. Calculem el coeficient de correlació de Pearson entre els vectors formats pels seus rànquings.
- Similitud d'imatges. Es calcula la similitud del cosinus entre els vectors d'ambdues imatges.
- Similitud per l'edat del model de la imatge. Calculem la similitud entre les seves edats:

  $$\text{Similitud per edat} = 1 - |edat_x - edat_y|$$

  Com més proper a 1 sigui aquest valor, més similars són les edats. Si les edats són molt distants, la similitud total es veurà penalitzada. En cas de tractar-se d'una imatge sense model, el valor d'aquesta similiud seria *NA*. S'ha adaptat el codi per aquesta casuística: si totes dues edats del model que s'estan comparant són *NA*, la similitud per edat val 1. Si només una és *NA*, la similitud per edat val 0.
- Similitud per altres variables categòriques. Calculem la similitud segons si hi ha marca visible o no, el color principal, el tipus d'estampat, el gènere de model i el tipus de peça de roba, per cada variable *dummy*. Per exemple, pel cas de la variable *dummy* Llis, que és 1 si l'estampat és llis i 0 si no:

  $$\text{Similitud}_{Llis} = \begin{cases} 1 & \text{si } \ Llis_x = Llis_y \\ 0 & \text{si } \ Llis_x \neq Llis_y \end{cases}$$

  S'adjudica 1 punt per cada variable *dummy* en la que tinguin el mateix valor.

Per agrupar la similitud de les variables relatives a les característiques dels ítems en un únic valor, es calcula la mitjana de les similituds de cada variable *dummy* i de l'edat del model de la imatge. A aquesta mitjana l'anomenarem similitud de característiques.

Finalment, el valor final de similitud entre l'ítem x i l'y prové d'una mitjana ponderada:

$$\begin{aligned}\text{Similitud final}(x, y) = 0{,}5 \cdot \text{Similitud per rànquings} + 0{,}25 \cdot \text{Similitud de característiques} \\ + 0{,}25 \cdot \text{Similitud d'imatges}\end{aligned}$$

Otorgant aquests pesos reflexem l'enfocament híbrid: el 50% del pes resideix en l'enfocament *item-based*, amb els rànquings, i el 50% restant resideix en l'enfocament basat en contingut, amb les característiques de les peces i la similitud d'imatges.

A continuació, d'entre les 10 peces de roba vistes per l'usuari escollit, cal trobar el nombre òptim k a agafar segons la seva similitud amb cada imatge no vista, per tal de generar les prediccions dels rànquings.

La idea és la següent: a partir de les 10 peces de roba vistes per l'usuari escollit, seleccionar les top-k més similars a cada imatge p no vista, i calcular la mitjana ponderada dels seus rànquings. Aquesta mitjana ponderada representarà la predicció de la posició del rànquing en què l'usuari hauria col · locat la imatge p si l'hagués vist. Trobem aquest valor òptim de k mitjançant validació creuada, avaluant valors de k des d'1 fins a 10. Els valors de k només poden oscil · lar entre aquests valors perquè cap usuari ha vist més de 10 imatges. La validació creuada es durà a terme amb les 10 imatges vistes per l'usuari, dividides un 70% a *train* i un 30% a *test*. Per cada k d'1 a 10, i per cada imatge dins el conjunt de *test*:

1. Calculem la similitud entre la imatge actual de *test* amb la que estem treballant i la resta d'imatges de *train*.
2. Agafem les top-k imatges de *train* amb una similitud positiva.
3. Agafem els rànquings reals en què l'usuari escollit va col · locar les top-k imatges seleccionades al pas anterior.
4. Calculem la mitjana ponderada d'aquests rànquings en funció de la similitud de cada imatge amb la imatge de test amb la que estem treballant, per predir el rànquing que en faria l'usuari escollit.

Un cop fetes les 3 iteracions corresponents a cada imatge del grup de *test*, calculem l'RMSE entre els rànquings reals d'aquestes imatges i els predits. El càlcul de l'RMSE està detallat al Marc teòric. Repetim aquest procés per cada valor de k. Ens quedarem amb aquella k que tingui un valor d'RMSE menor.

Cal mencionar que per assegurar certa robustesa en els resultats, només treballem amb aquelles imatges per les que haguem trobat com a mínim dues imatges de les vistes per l'usuari amb una similitud superior a 0. Si no es compleix aquesta condició, es descarta la imatge no vista amb la que s'està treballant, ja que no tenim prou dades per oferir una predicció mínimament fiable del seu rànquing.

Un cop escollida la k òptima per validació creuada, procedim a calcular el rànquing predit per cada una de les 90 imatges no vistes. Seleccionem, per cada cas, les top-k imatges sí vistes més similars a cada imatge no vista. Calculem la mitjana ponderada dels rànquings d'aquestes top-k imatges, en funció de la similitud, per predir el rànquing de cada imatge p no vista.

L'últim pas consisteix en agafar les 10 peces de roba amb millor predicció. Degut a la normalització prèvia dels rànquings, les posicions més elevades al rànquing tenen valors més petits (fins i tot negatius). Per tant, busquem aquelles imatges no vistes que tinguin valors de predicció petits, cosa que es tradueix en posicions elevades al rànquing.

En el cas del desenvolupament del recomanador *item-based* incial, el procediment és exactament el mateix. L'únic que canvia és el càlcul de la similitud al primer pas: en comptes de computar una similitud combinada, el 100% del pes recau sobre la similitud per rànquings.

Tant els mètodes *user-based* com *item-based* són enfocaments basats en memòria. Aquests enfocaments estan basats en el mètode *k-Nearest Neighbours*, ja que estan alineats amb la definició estàndard de k-NN en sistemes de recomanació: se seleccionen els k veïns més propers per predir la puntuació o rànquing que s'otorgaria als ítems no vistos. Així doncs, tant per *user-based* com per *item-based* hem implementat filtratge col · laboratiu k-NN, amb la millora afegida de característiques demogràfiques, pel cas *user-based*, i característiques del ítems i similitud d'imatges, pel cas *item-based*, fent sistemes híbrids. D'aquesta manera s'aprofita la simplicitat del k-NN, però abordant l'escasseig de dades del nostre conjunt de dades (90% de valors faltants), a través de mètriques de similitud híbrides i validació creuada.

4.5.3 *Singular Value Decomposition* (SVD)

El tercer enfocament plantejat és l'SVD, un dels possibles models que trobem dins el paraigües dels enfocaments basats en models, en el filtratge col · laboratiu.

Aquest model l'abordem a partir de la matriu usuari-ítem, amb els rànquings. Com s'ha mencionat a l'inici d'aquest apartat, l'implementació d'aquests models està feta amb la llibreria *recommenderlab* d'R, dissenyada principalment per treballar únicament amb matrius de valoracions. Per tant, a part de no incloure característiques demogràfiques o altra informació addicional, cal que transformem els valors de la matriu de rànquings a valoracions. Per fer-ho, apliquem una transformació senzilla:

$$\text{Valoració} = 11 - \text{Rànquing}$$

D'aquesta manera, les posicions al capdavant del rànquing es converteixen en puntuacions elevades, i les posicions al capdavall es converteixen en puntuacions baixes. Un cop tenim les dades transformades, les normalitzem mitjançant el procés *Z-Score*.

El model SVD requereix un únic paràmetre k, que representa el número de factors latents (dimensions) a la factorització matricial. En altres paraules, indica quantes dimensions usem per representar els usuaris i els ítems en un espai dimensional menor. Per determinar el millor valor de k realitzem *5-fold cross validation*, fent ús de funcions pròpies de la llibreria *recommenderlab*. Com a valors de k treballem amb $k = 5, 10, 15, 20$. Obtenim la mitjana dels RMSE per cada *fold* per cada valor de k, i ens quedem amb aquella k amb un RMSE menor.

Un cop escollida la k, dividim les dades en un 80% *train* i un 20% *test*, i entrenem el model amb les dades de *train*. Amb el model entrenat, ja podem realitzar prediccions per un o varis usuaris determinats i obtenir les top-10 recomanacions per cadascú.

4.5.4 *Alternating Least Squares* (ALS)

El quart i últim enfocament plantejat és l'ALS, un altre dels possibles models que trobem dins el paraigües dels enfocaments basats en models, en el filtratge col · laboratiu.

Aquest model l'abordem de la mateixa manera que l'SVD: treballem únicament amb la matriu usuari-ítem, transformem els valors dels rànquings en puntuacions, aplicant la mateixa transformació que la vista a l'apartat d'SVD, i normalitzem aquestes puntuacions mitjançant el procés *Z-Score*.

Ara bé, el model ALS requereix tres paràmetres:

- k, que indica quantes dimensions usem per representar els usuaris i els ítems en un espai dimensional menor, igual que a l'SVD.
- λ, que és un paràmetre de regularització que penalitza la magnitud dels factors latents (k), per prevenir *overfitting*.
- Nº màxim d'iteracions, per trobar un equilibri entre convergència i cost computacional. En el nostre cas, el fixem directament a 10, ja que sol ser entre 5 i 20, i un valor de 10 permet assegurar-nos l'assoliment de la convergència sense suposar un cost computacional excessiu.

Per determinar la millor combinació de k i λ realitzem, de nou, *5-fold cross validation*, fent ús de funcions pròpies de la llibreria *recommenderlab*. Com a valors de k treballem amb $k = 5, 10, 15, 20$ i com a valors de λ treballem amb $\lambda = 0.01, 0.1, 1$. Obtenim la mitjana dels RMSE per cada *fold* per cada valor de k i λ, i ens quedem amb aquella combinació de paràmetres amb un RMSE menor.

Un cop escollida la k i la λ, el procediment és el mateix que en el model SVD: dividim les dades en un 80% *train* i un 20% *test*, i entrenem el model amb les dades de *train*. Amb el model entrenat, ja podem realitzar prediccions per un o varis usuaris determinats i obtenir les top-10 recomanacions per cadascú.

4.5.5 Comparació de models

Primerament mostrarem els resultats de la comparació entre els enfocaments senzills dels mètodes *user-based* i *item-based* versus els enfocaments híbrids respectius. Després, mostrarem els resultats de la comparació entre els 4 mètodes finals: l'enfocament híbrid per *user-based*, l'enfocament híbrid per *item-based*, SVD i ALS.

Per cada cas, la comparació dels models s'ha dut a terme de la següent manera: s'han escollit quinze individus aleatoris de la base de dades, i els hem generat les top-4 recomanacions segons cada model. Per cada individu, hem calculat la similitud del cosinus entre les top-4 recomanacions de cada model i les imatges que va col · locar a les 4 primeres posicions de rànquing quan va respondre l'enquesta interactiva. Així obtenim 16 valors de similitud per cada individu. Per cada model, doncs, tindrem 240 (16 valors x 15 individus) valors de similitud. Per obtenir una mètrica final per cada model, calculem la mitjana d'aquests 240 valors.

Aquesta estratègia de comparació és la primera prova àcida que realitzem en el nostre treball.

L'elecció de treballar només amb el top-4 ve donada pel factor d'aleatorietat de l'enquesta. Recordem que cada individu només veu 10 imatges aleatòries. Per tant, és lògic pensar que aquelles a les darreres posicions del rànquing no són d'interès per l'individu, i només les primeres li poden agradar. Per tant, decidim només treballar amb les 4 primeres imatges.

Un cop explicat el procediment de comparació, presentem la primera taula de resultats:

Mètode	*User-based* senzill	*User-based* híbrid	*Item-based* senzill	*Item-based* híbrid
Similitud	0,707	0,735	0,683	0,727

Taula 4.5: Resultats de la similitud entre les top-4 imatges reals i les top-4 predites pels 15 usuaris escollits, segons cada mètode.

Veiem com en ambdós casos el valor de la similitud entre les top-4 imatges reals i les predites augmenta en el cas del sistema híbrid. Per acabar de validar si les diferències entre els plantejaments senzills i els híbrids són estadísticament significatives, realitzem una prova t de Student per dades aparellades, ja que a cada individu li oferim recomanacions segons el plantejament senzill i l'híbrid. Les bases d'aquesta prova es troben detallades al Marc teòric. La duem a terme amb els valors de similitud calculats anteriorment i, donat el tamany elevat de la mostra, assumim normalitat. Obtenim els següents resultats:

Estadístic	**P-valor**
-2,432	0,00788

Taula 4.6: Resultats del *t-test* entre *user-based* senzill i *user-based* híbrid.

Estadístic	**P-valor**
-3,573	0,0002

Taula 4.7: Resultats del *t-test* entre *item-based* senzill i *item-based* híbrid.

En ambdós casos obtenim p-valors per sota del nivell de significació del 5%, permetent-nos rebutjar les hipòtesis nul · les respectives: les mitjanes de cada parella de models comparada no són iguals. Per tant, podem afirmar que la inclusió de les característiques del perfil de l'usuari, en el cas *user-based*, o de les característiques de les peces de roba i la similitud d'imatges, en el cas *item-based*, suposen un augment de la precisió de les recomanacions.

A continuació, la segona taula de resultats, per escollir el model òptim:

Mètode	*User-based* híbrid	*Item-based* híbrid	SVD	ALS
Similitud	0'735	0'727	0'729	0'704

Taula 4.8: Resultats de la similitud entre les top-4 imatges reals i les top-4 predites pels 15 usuaris escollits, segons cada mètode.

Novament, per validar si les diferències entre els quatre models són estadísticament significatives, realitzem una prova estadística. En aquest cas, un ANOVA de blocs aleatorietzats, tractant els individus com a blocs. Les bases d'aquesta prova es troben detallades al Marc teòric. La duem a terme amb els valors de similitud calculats anteriorment i, donat el tamany elevat de la mostra, tornem a assumir normalitat.

Segons l'ANOVA, ni el factor del model ni el bloc de l'individu són significatius. Per tant, no tenim prou evidència per afirmar que hi hagi diferències significatives entre els quatre models finals.

Per tal d'implementar el recomanador en una interfície per a l'ús dels usuaris, hem d'escollir un dels quatre models. Tot i que tots presenten resultats prou similars i acabem de demostrar que no són estadísticament diferents, observem que el que presenta una similitud superior entre les imatges reals i les predites és el model *user-based* híbrid. Així doncs, el model final amb el que ens quedem és l'*user-based* híbrid.

Per acabar de validar el model escollit hem realitzat una segona prova àcida, consistent en una segona enquesta interactiva [3]. L'estructura és igual que la de la primera enquesta, però un cop l'usuari envia les seves respostes, les seves dades són processades pel recomanador *user-based* i es retornen les top-3 recomanacions per l'usuari. Aleshores, l'individu ha de puntuar de l'1 al 10 cadascuna de les 3 recomanacions fetes. Aquesta segona enquesta s'efectua sobre una petita mostra de 6 individus que no responguessin la primera enquesta, per assegurar una major fiabilitat de les dades. Com a mètrica de validació, computem la mitjana de totes les 18 puntuacions rebudes (6 usuaris x 3 puntuacions cadascun), i obtenim una puntuació mitjana de 8,2/10 i una mediana de 8. Tot i que el criteri de puntuació alta pot ser subjectiu, considerem que una mitjana de 8,2 representa resultats favorables i podem afirmar que el recomanador encerta amb bastanta fiabilitat les recomanacions que proporciona.

4.6 Regles d'associació

Aquest últim apartat del cos del treball consisteix en l'estudi de regles d'associació per trobar patrons de preferències dels usuaris. Aquesta anàlisi es duu a terme a partir de la taula "valoracions", preparada a la segona part del preprocessament, que conté les característiques demogràfiques dels enquestats i els rànquings assignats a cada imatge.

L'anàlisi comença amb un breu estudi previ de les transaccions presents a la base de dades. Recordem que cada fila de la nostra base de dades, és a dir, les dades de cada usuari, formen una transacció. Amb aquest breu estudi el que es pretén és determinar el valor amb el qual ens haurem de basar per definir el suport mínim de les regles. Normalment, aquest valor mínim es determina mirant el següent gràfic, que mostra la freqüència dels top-100 ítems presents a la base de dades, ordenats de major a menor freqüència.

[3] https://clothing-rank-formapp-ukgse2cpyl22kpicqbpijj.streamlit.app

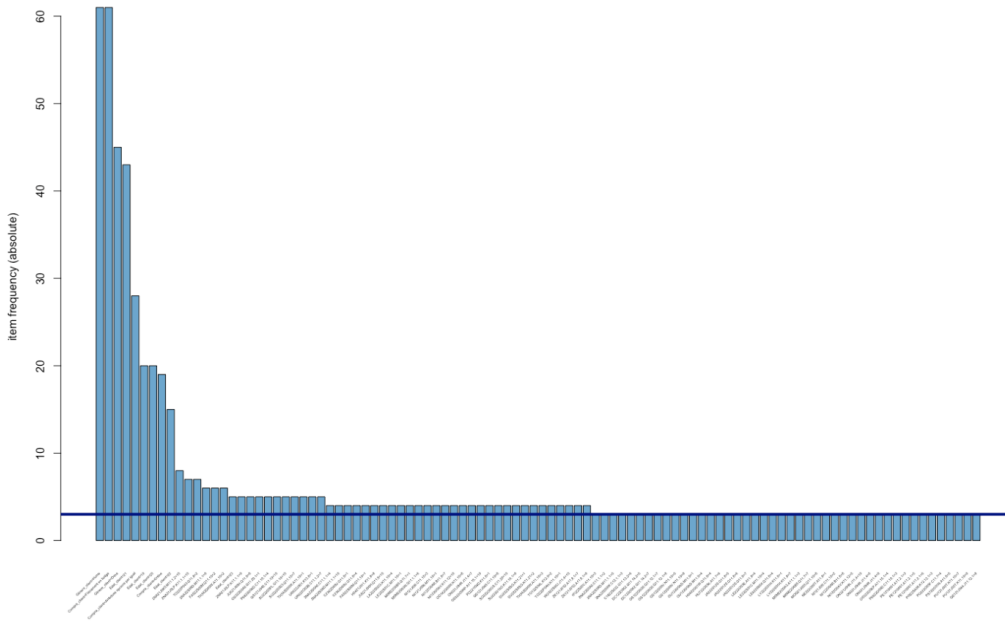

Figura 4.14: Ítems ordenats segons la seva freqüència.

El valor mínim de suport el trobem a la cua de la dreta del gràfic, on es troben els ítems amb menor freqüència. La línia blava més fosca indica la freqüència mínima a basar-nos per determinar el valor mínim de suport.

Donat que la visió de la freqüència mínima és complicada degut al gran nombre d'ítems presents al gràfic, ajudant-nos de codi hem pogut determinar que aquest valor mínim és 3. Dividint 3 entre el total de transaccions que tenim (108), obtenim un valor mínim de suport a establir igual a 0,0278, que arrodonim 0,03.

Un cop definit el suport mínim, passem a estudiar els *itemsets* freqüents. Aquest segon estudi és un pas intermig abans d'arribar a les regles d'associació, que ens permet identificar patrons comuns. Obtenim 116 *itemsets* amb un suport superior al mínim establert, dels quals 56, gairebé la meitat, només contenen un únic ítem. Ens centren en analitzar aquells *itemsets* formats per més d'un ítem. D'entre aquests *itemsets*, destaquem els següents:

- {Gènere_client=Home, Compra_client=Físicament en botiga}, amb un suport igual a 0,32.
- {Gènere_client=Dona, Compra_client=Físicament en botiga}, amb un suport igual a 0,24.
- {Edat_client=21, Compra_client=Físicament en botiga}, amb un suport igual a 0,22.

Aquests *itemsets* destaquen perquè reflexen característiques interessants dels usuaris i, a més, són els tres que presenten valors de suport superiors. D'aquí podem extreure que, en general, tant els homes com les dones prefereixen comprar físicament en botiga abans que de manera *online*, i que els usuaris de 21 anys d'edat també prefereixen comprar de manera

física. Aquestes conclusions ens permeten entendre els patrons de preferència de compra de la gent jove.

Finalment, només queda estudiar les regles d'associació. Amb el suport mínim determinat anteriorment, obtenim tan sols 20 regles no redundants. Els valors de suport d'aquestes regles són, en general, bastant baixos: el valor de suport més elevat és de 0,06, cosa que indica que tan sols apareix en unes 7 transaccions. Pel que fa a la mètrica de confiança, totes les regles presenten valors que oscil · len entre 0,8 i 1, valors molt elevats: per aquests casos, el consegüent i l'antecedent tendeixen a aparèixer junts. En última instància, analitzem el *lift*. Totes les 20 regles tenen valors de *lift* superiors a 1, cosa que reforça la dependència entre l'antecedent i el consegüent. Passant a analitzar les regles pròpiament, veiem que totes tenen una estructura similar: a l'antecedent (X) tenim el nom d'una imatge, i al consegüent (Y) hi tenim una característica demogràfica (ja sigui l'edat, el gènere o la preferència de compra). A partir d'aquest tipus de regles, creiem que no podem obtenir conclusions d'interès per l'estudi.

Capítol 5

Conclusions

En aquest darrer apartat de la memòria parlarem sobre el compliment dels objectius, tant principals com transversals, definits a l'inici del treball, a l'apartat d'Introducció. Prosseguirem a fer un breu resum dels resultats més importants extrets de l'anàlisi, i es comentaran possibles línies futures de treball en cas de voler seguir investigant aquesta matèria.

Tots els cinc objectius principals plantejats s'han complert degudament. S'ha conseguit desenvolupar varis models de recomanació d'articles de roba basats en preferències d'usuaris, i s'han optimitzat afegint informació addicional per oferir recomanacions més precises. A més, en el desenvolupament d'aquests sistemes recomanadors s'han implementat tècniques d'aprenentatge automàtic, com ara *autoencoders*, o els propis mètodes de recomanació SVD i ALS. També s'ha desenvolupat una interfície d'usuari on s'ha implementat el millor recomanador dels treballats, permetent als usuaris interactuar amb el recomanador en temps real. Finalment, hem estudiat les regles d'associació per intentar esbrinar patrons en les preferències de roba dels usuaris, tot i que de manera poc satisfactòria. Amb l'assoliment de tots els objectius principals, doncs, presentem un treball que permet a les petites i mitjanes empreses adaptar-se a la transformació tecnològica de manera accessible. Aquest sistema recomanador les pot ajudar a optimitzar la gestió d'estocs, millorar l'estratègia de negoci i conseguir augmentar el volum de vendes.

Pel que fa als objectius transversals, hem après nocions bàsiques del llenguatge de programació *Python*. Això ens ha permès desenvolupar l'enquesta interactiva de classificació d'imatges mitjançant *Python* i *Streamlit*. Addicionalment, s'ha redactat tota la memòria a través de la plataforma *Overleaf*, cosa que ha propiciat l'aprenentatge de l'ús de *Latex*.

Com a principal síntesi de l'estudi, concloem que els enfocaments *user-based* i *item-based* híbrids milloren la qualitat de les recomanacions, oferint suggeriments més precisos que els seus enfocaments senzills, respectivament. Pel que fa als 4 sistemes recomanadors finals amb els que hem treballat (*user-based* híbrid, *item-based* híbrid, SVD i ALS), no hem trobat diferències prou significatives per determinar que un model sigui millor que un altre. Tot i

això, optem per quedar-nos amb l'enfocament *user-based* híbrid, que és el que proporciona recomanacions més acurades.

Totes aquestes comparacions de models suposen la primera prova àcida del treball, i s'han dut a terme emprant proves estadístiques. Fins ara, a la literatura prèvia no s'havien comparat els diferents mètodes entre sí mitjançant testos.

Addicionalment, un cop realitzades les comparacions dels models, hem acabat de validar el model escollit sotmetent-lo a una altra prova àcida, on una mostra d'usuaris ha puntuat les recomanacions obtingudes amb el mètode *user-based* híbrid. En aquesta prova àcida els usuaris qualifiquen les recomanacions del model amb un 8,2 de mitjana, cosa que sustenta que el model desenvolupat ofereix bons suggeriments.

5.1 Línies futures

Com a línies futures d'estudi, pensem que seria interessant intentar adaptar els enfocaments basats en models (SVD i ALS) per a que incorporessin informació addicional, com ara característiques demogràfiques o altra informació del perfil de l'usuari. Aquests dos enfocaments, que només contenen les dades dels rànquings, no són estadísticament diferents a l'*user-based* i l'*item-based* híbrids, que sí que tenen més informació addicional. Per tant, creiem que els models SVD i ALS tenen potencial per oferir recomanacions encara més acurades si s'optimitzessin amb més dades.

Seguint en aquesta línia, es podria estudiar com afecten les ponderacions otorgades al càlcul de similitud als casos *user-based* i *item-based* híbrids, per investigar si hi ha una combinació òptima que ofereixi millors resultats.

També creiem que endinsar-se més en l'estudi de les preferències de compra dels usuaris a través de regles d'associació és una via d'estudi interessant. A través d'una enquesta més completa que demani més informació de preferències i gustos als usuaris es podrien treure conclusions interessants respecte els comportaments de compra, informació valuable per complementar les estratègies de negoci de les empreses.

Bibliografia

[1] T. Alonso, Stradivarius presenta su primera colección reinterpretada con inteligencia artificial (2023).

[2] J. A. Clemente-Almendros, D. Nicoara-Popescu, I. Pastor-Sanz, Digital transformation in SMEs: Understanding its determinants and size heterogeneity, Technology in Society 77 (2024) 102483.

[3] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, Communications of the ACM 35 (1992) 61–70.

[4] P. Resnick, H. R. Varian, Recommender systems, Communications of the ACM 40 (1997) 56–58.

[5] C. Team, Mass customization, 2023. URL: https://corporatefinanceinstitute.com/resources/management/mass-customization/#:~:text=Mass%20customization%20refers%20to%20a,at%20near%20mass%20production%20efficiency.

[6] J. B. Schafer, J. Konstan, J. Riedl, Recommender systems in e-commerce, ACM Digital Library (1999) 158–166.

[7] Recommender systems-the need of the ecommerce ERA, 2017. URL: https://ieeexplore.ieee.org/abstract/document/8282616.

[8] H. Hwangbo, Y. S. Kim, K. J. Cha, Recommendation system development for fashion retail e-commerce, Electronic Commerce Research and Applications 28 (????) 94–101.

[9] A Systematic Study on the Recommender Systems in the E-Commerce, 2020. URL: https://ieeexplore.ieee.org/abstract/document/9118884.

[10] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, T. L. Berg, Retrieving Similar Styles to Parse Clothing, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2014) 1028–1040.

[11] K. Yamazaki, M. Inaba, Clothing classification using image features derived from clothing fabrics, wrinkles and cloth overlaps, 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (2013) 2710–2717.

[12] D. Sha, D. Wang, X. Zhou, S. Feng, Y. Zhang, G. Yu, An Approach for Clothing Recommendation Based on Multiple Image Attributes, 2016. URL: https://doi.org/10.1007/978-3-319-39937-9_21. doi:10.1007/978-3-319-39937-9\{_21.

[13] S. K. Addagarla, A. Amalanathan, Probabilistic Unsupervised Machine Learning Approach for a Similar Image Recommender System for E-Commerce, Symmetry 12 (2020) 1783.

[14] S. K. Addagarla, A. Amalanathan, e-SimNet: a visual similar product recommender system for E-commerce, Indonesian Journal of Electrical Engineering and Computer Science 22 (2021) 563.

[15] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, ACM SIGMOD Record 22 (1993) 207–216.

[16] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Very Large Data Bases (1994) 487–499.

[17] Building Personalized Recommendation System in E-Commerce using Association Rule-Based Mining and Classification, 2007. URL: https://ieeexplore.ieee.org/abstract/document/4370866/references#references.

[18] J.-W. D.-K. Lee, Chong-Hyeon;Lee Suk-Hoon;Kim, A Personalized Clothing Recommender System Based on the Algorithm for Mining Association Rules, koreascience.kr (2010).

[19] O. Cakir, M. E. Aras, A Recommendation Engine by Using Association Rules, Procedia - Social and Behavioral Sciences 62 (2012) 452–456.

[20] J. Dai, B. Zeng, An Association Rule Algorithm for Online e-Commerce Recommendation Service, Journal of Economics Business and Management 4 (2016) 573–576.

[21] O. Dogan, F. C. Kem, B. Oztaysi, Fuzzy association rule mining approach to identify e-commerce product association considering sales amount, Complex Intelligent Systems 8 (2022) 1551–1560.

[22] M. Y. Avcilar, E. Yakut, Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store, Canadian social science 10 (2014) 75–83.

[23] A. Jain, S. Jain, N. Merh, Application of Association Rule Mining in a Clothing Retail Store, 2021. URL: https://doi.org/10.1007/978-981-33-6656-5_9. doi:10.1007/978-981-33-6656-5\{_9.

[24] N. Wahidi, R. Ismailova, Association rule mining algorithm implementation for e-commerce in the retail sector, Journal of Applied Research in Technology Engineering 5 (2024) 63–68.

[25] Communications, 'Machine learning': ¿qué es y cómo funciona el maestro en reconocer patrones? (2024).

[26] Image Recognition Machine Learning: Use Cases and Common Algorithms | Cloudinary, 2025. URL: https://cloudinary.com/guides/front-end-development/image-recognition-machine-learning-use-cases-and-common-algorithms.

[27] L. S. Easily, Qué es: Semejanza visual explicada en detalle, 2024. URL: https://es.statisticseasily.com/glosario/%C2%BFQu%C3%A9-es-la-similitud-visual-expli

cada-en-detalle%3F/#:~:text=La%20similitud%20visual%20se%20refiere,humano%20o%20a%20los%20algoritmos%20computacionales.

[28] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014. URL: https://arxiv.org/abs/1409.1556.

[29] GeeksforGeeks, VGG16 | CNN model, 2024. URL: https://www.geeksforgeeks.org/vgg-16-cnn-model/.

[30] N. Kundu, Exploring ResNet50: An In-Depth Look at the Model Architecture and Code Implementation (2023).

[31] GeeksforGeeks, EfficientNet Architecture, 2024. URL: https://www.geeksforgeeks.org/efficientnet-architecture/.

[32] A. Brital, AutoEncoders explained - Anas BRITAL - medium (2022).

[33] GeeksforGeeks, Image Similarity: Understanding and Implementing Methods, 2024. URL: https://www.geeksforgeeks.org/image-similarity-understanding-and-implementing-methods/?utm%20https://medium.com/@aleksej.gudkov/understanding-image-similarity-with-machine-learning-c8680c24dd56.

[34] UATeam, Understanding Image Similarity with Machine Learning (2024).

[35] Normalización VS estandardización, 2024. URL: https://www.datacamp.com/es/tutorial/normalization-vs-standardization.

[36] L. S. Easily, Qué es: Normalización de Zscore explicada, 2024. URL: https://es.statisticseasily.com/glosario/%C2%BFQu%C3%A9-es-la-normalizaci%C3%B3n-de-zscore%3F/.

[37] E. Gomede, PhD, Imputation by Chained Equations (MICE): Bridging the Gap in Incomplete Data Analysis (2024).

[38] E. McClenaghan, The Chi-Squared Test, 2024. URL: https://www.technologynetworks.com/informatics/articles/the-chi-squared-test-368882#:~:text=The%20null%20hypothesis%20(H0)%20is,an%20association%20of%20any%20kind.

[39] A. Thevapalan, Prueba Chi-Cuadrado en R: Guía completa, 2025. URL: https://www.datacamp.com/es/tutorial/chi-square-test-r.

[40] J. Amat, RPubs - T-test de medias independientes y dependientes con R, ???? URL: https://rpubs.com/Joaquin_AR/218467.

[41] T-Test, Chi-Square, ANOVA, Regression, Correlation..., ???? URL: https://datatab.es/tutorial/paired-t-test.

[42] D. SALGADO, RPubs - ANOVA diseño de bloque completo al azar, 2024. URL: https://rpubs.com/deliosalgado/1168360.

[43] M. Rawal, Unveiling Visual Horizons: An Introduction to Image Recommendation Systems (2023).

[44] Ibm, Collaborative Filtering, 2025. URL: https://www.ibm.com/think/topics/collaborative-filtering.

[45] S. Sharma, A. Mahajan, Suggestive Approaches to Create a Recommender System for GitHub, International Journal of Information Technology and Computer Science 9 (2017) 48–55.

[46] How does singular value decomposition (SVD) work in recommender systems?, ???? URL: https://milvus.io/ai-quick-reference/how-does-singular-value-decomposition-svd-work-in-recommender-systems.

[47] R. Gupta, How Singular Value Decomposition (SVD) is used in Recommendation Systems, "Clearly Explained". (2023).

[48] A. Mishra, Movie recommendation system using alternating least squares (2024).

[49] A gentle introduction to Alternating Least Squares, 2018. URL: https://sophwats.github.io/2018-04-05-gentle-als.html.

[50] S. Herath, Fundamentals of Associate Rule mining - Data Science and Machine Learning - Medium, Medium (2024).

[51] Association rules, 2022. URL: https://www.ibm.com/docs/en/db2-warehouse?topic=procedures-association-rules.

[52] Ibm, Apriori Algorithm, 2025. URL: https://www.ibm.com/think/topics/apriori-algorithm#:~:text=The%20Apriori%20algorithm%20is%20an,items%20called%20itemsets%20in%20data.

Annex

A.1 Gràfics complementaris per l'anàlisi descriptiva

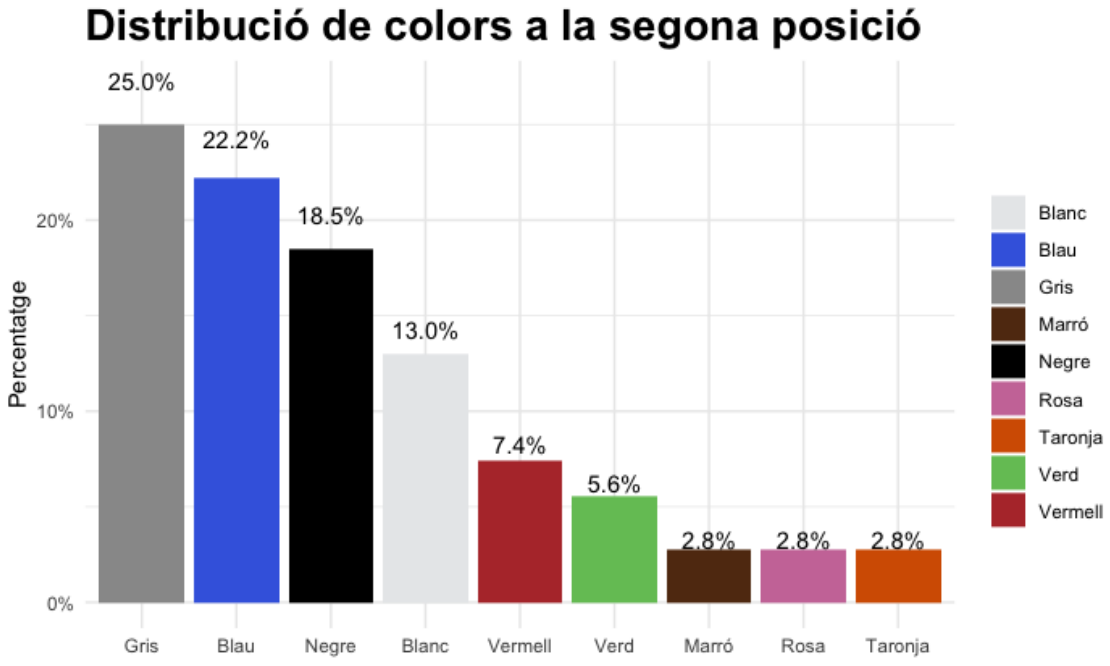

Figura A.1: Distribució dels colors de les peces de roba a la segona posició dels rànquings.

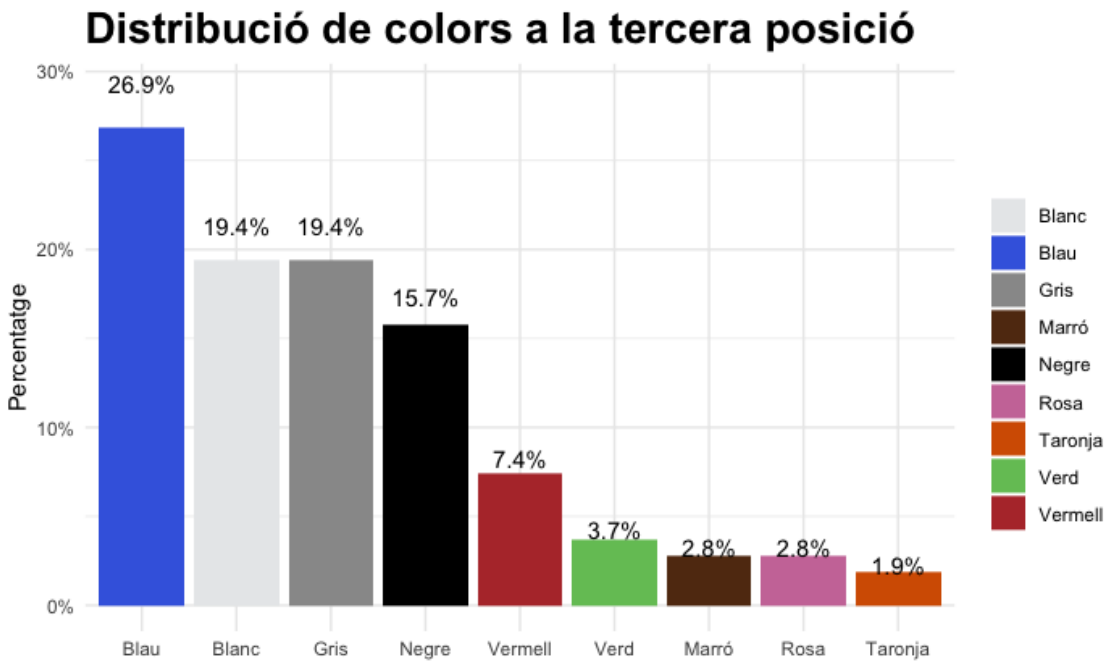

Figura A.2: Distribució dels colors de les peces de roba a la tercera posició dels rànquings.

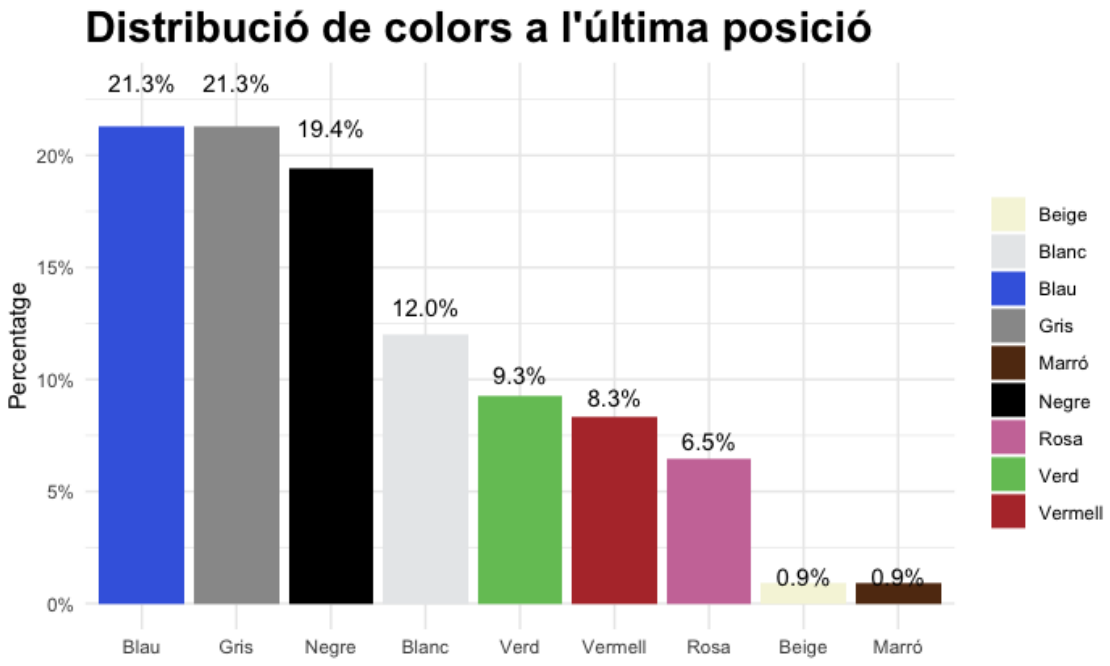

Figura A.3: Distribució dels colors de les peces de roba a l'última (desena) posició dels rànquings.

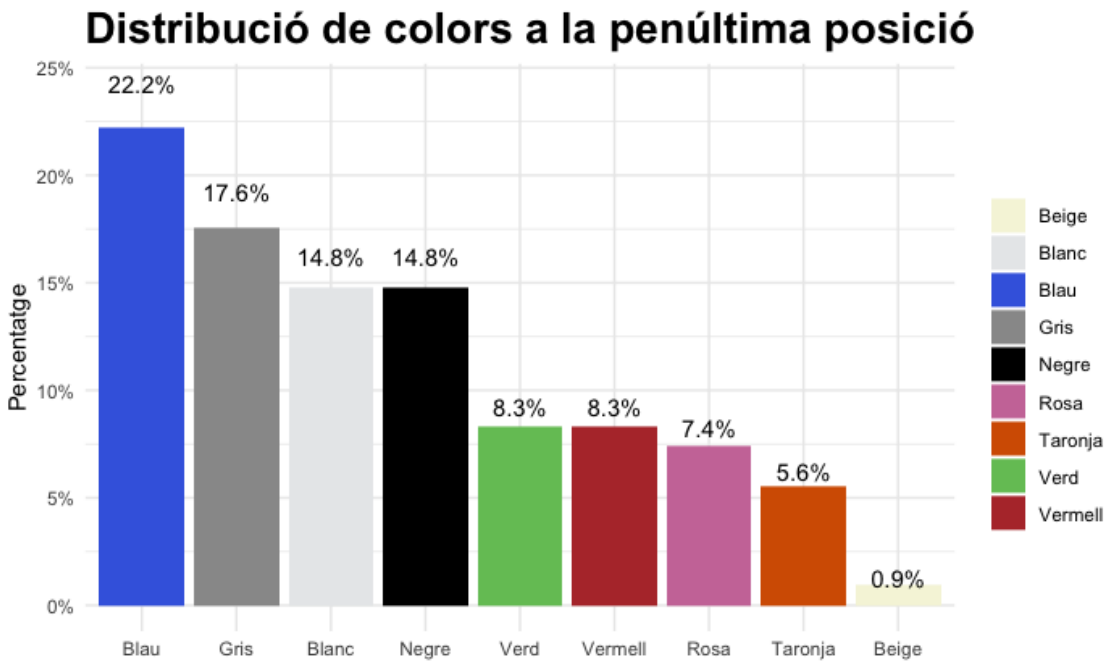

Figura A.4: Distribució dels colors de les peces de roba a la penúltima (novena) posició dels rànquings.

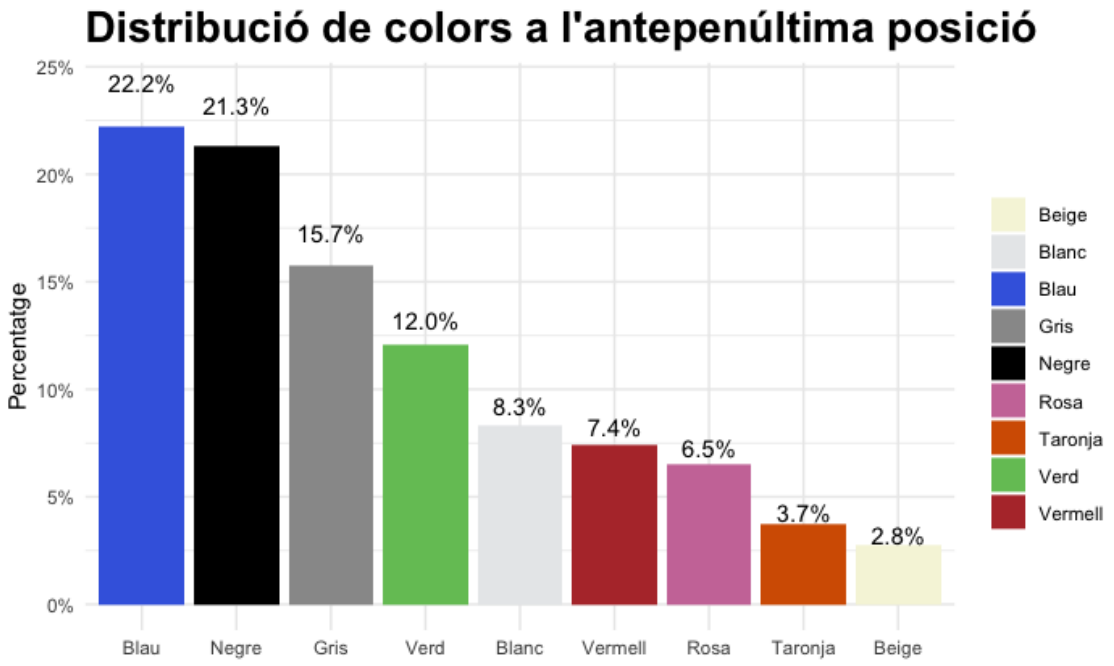

Figura A.5: Distribució dels colors de les peces de roba a l'antepenúltima (vuitena) posició dels rànquings.

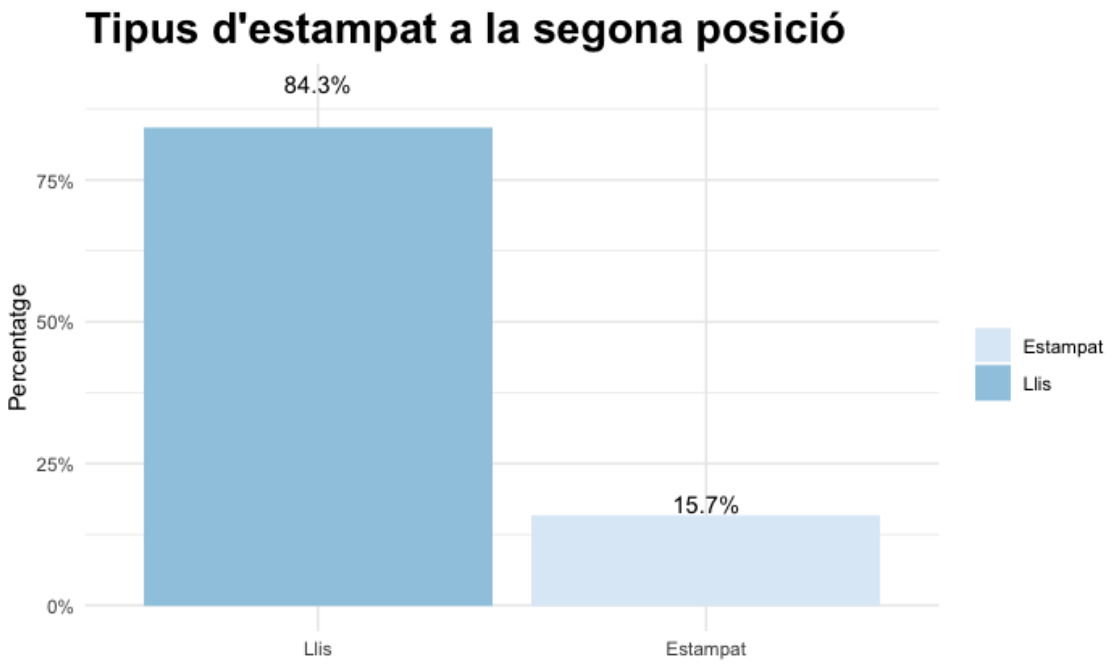

Figura A.6: Distribució del tipus d'estampat a les peces de roba a la segona posició dels rànquings.

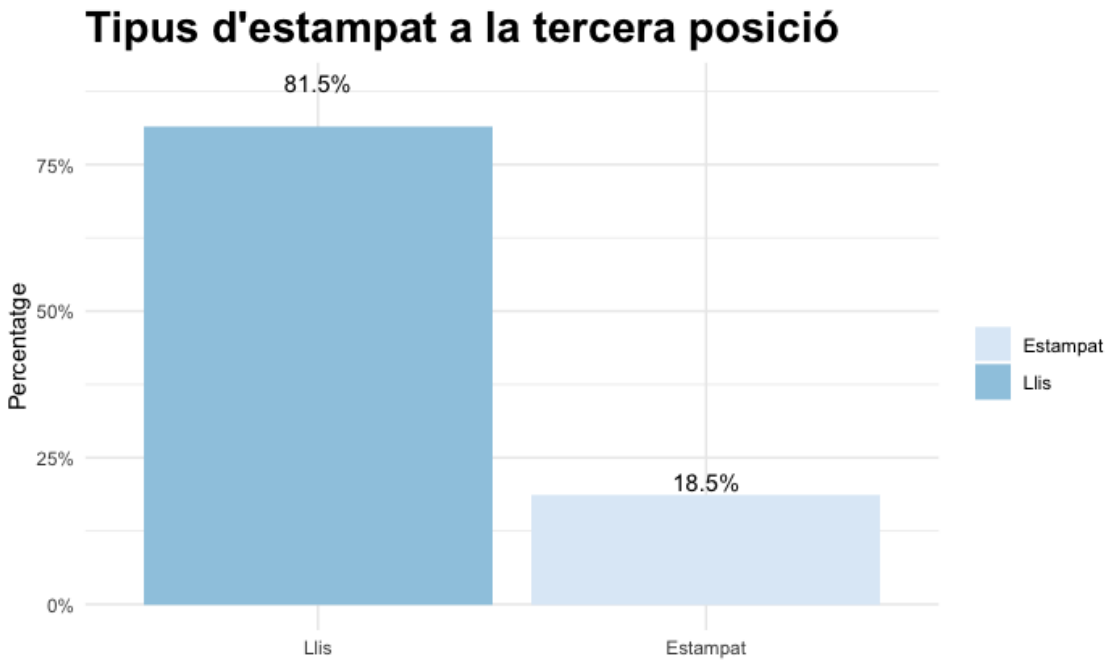

Figura A.7: Distribució del tipus d'estampat a les peces de roba a la tercera posició dels rànquings.

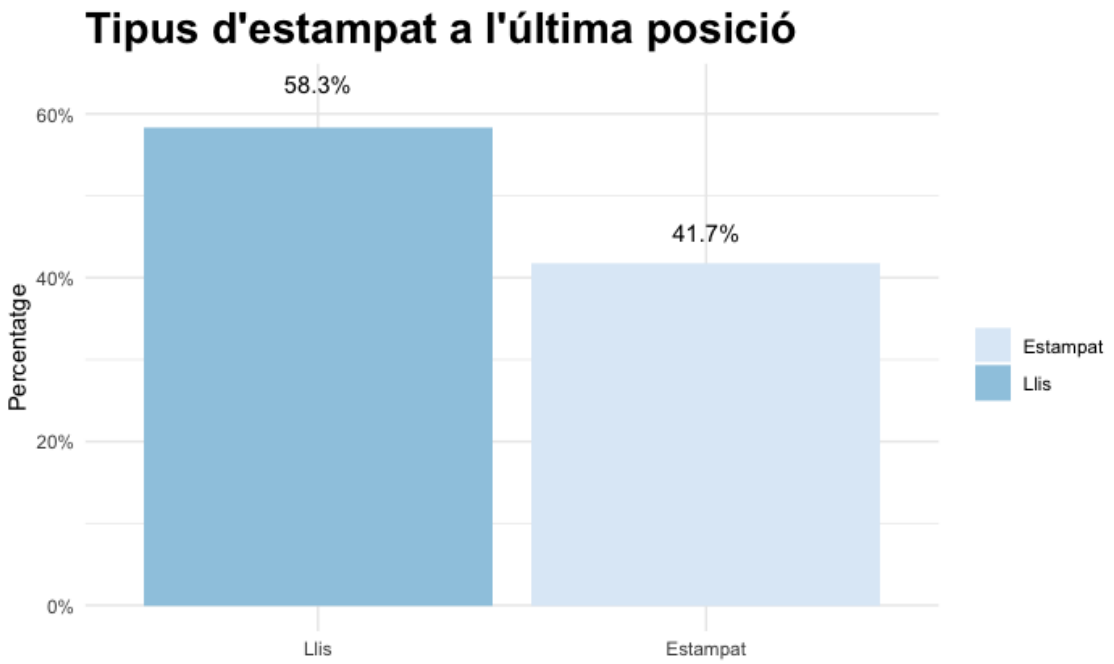

Figura A.8: Distribució del tipus d'estampat a les peces de roba a l'última (desena) posició dels rànquings.

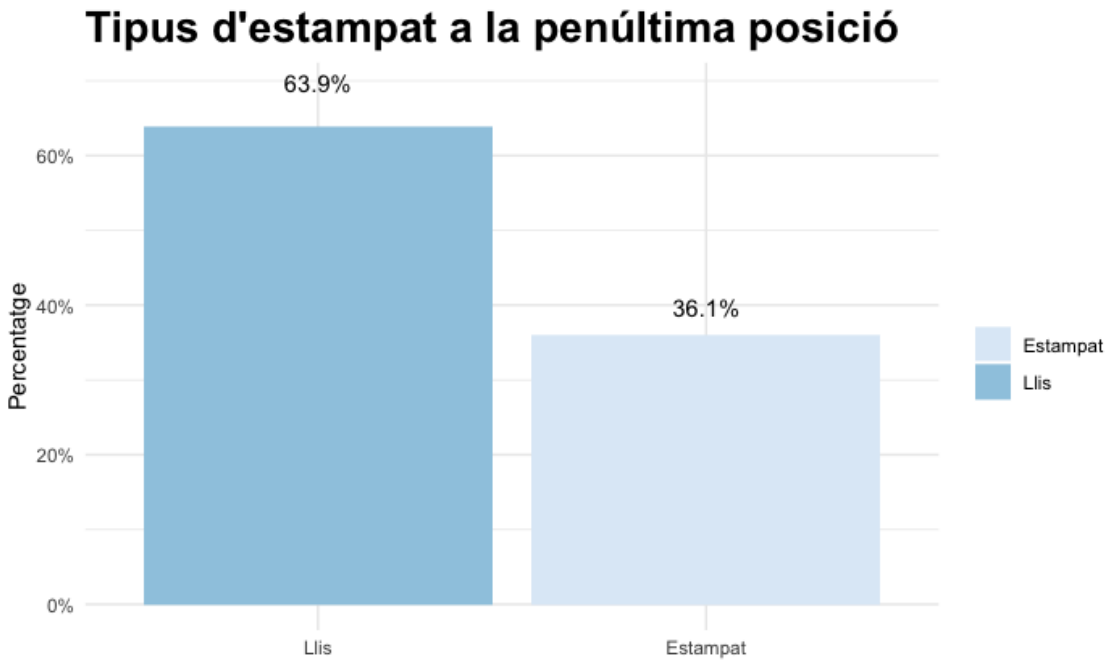

Figura A.9: Distribució del tipus d'estampat a les peces de roba a la penúltima (novena) posició dels rànquings.

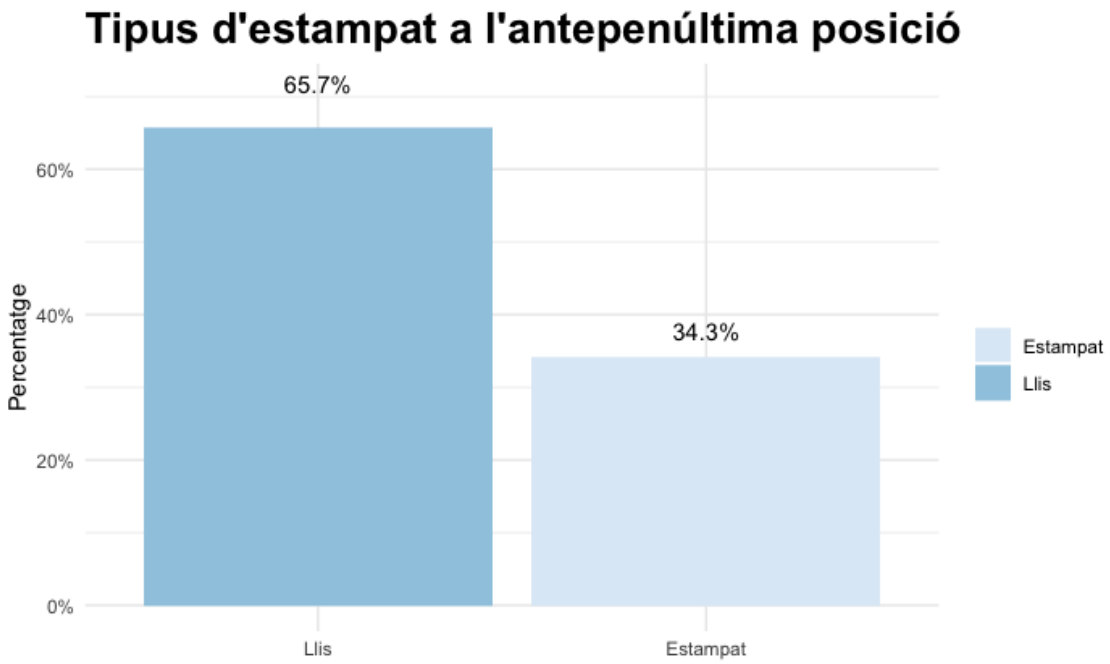

Figura A.10: Distribució del tipus d'estampat a les peces de roba a l'antepenúltima (vuitena) posició dels rànquings.

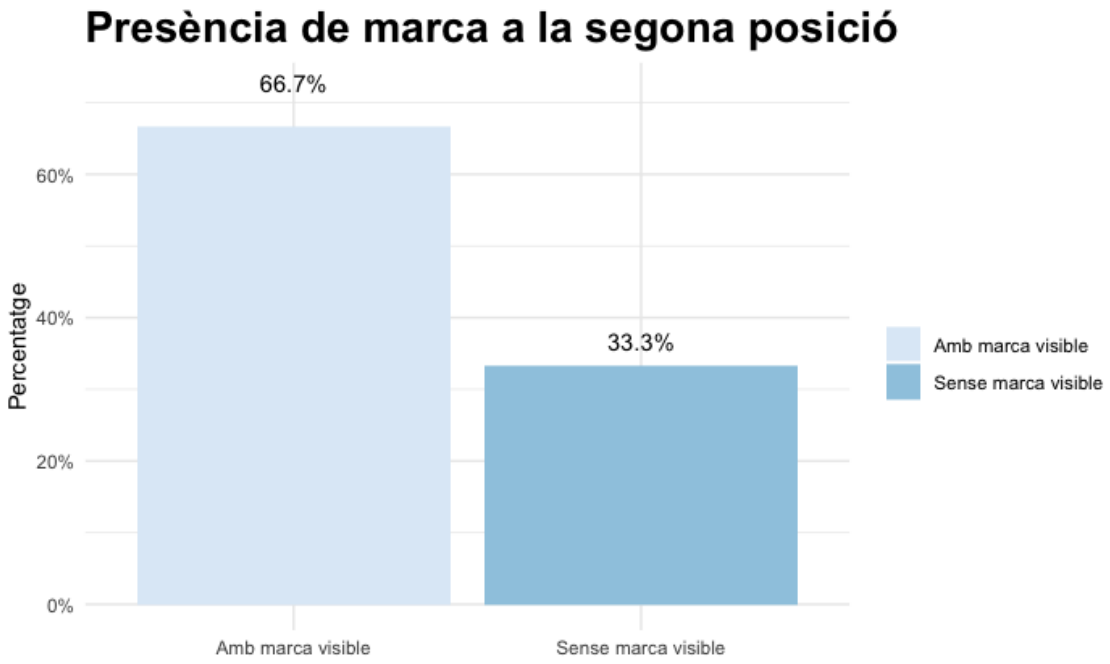

Figura A.11: Distribució de la presència de marca de les peces de roba a la segona posició dels rànquings.

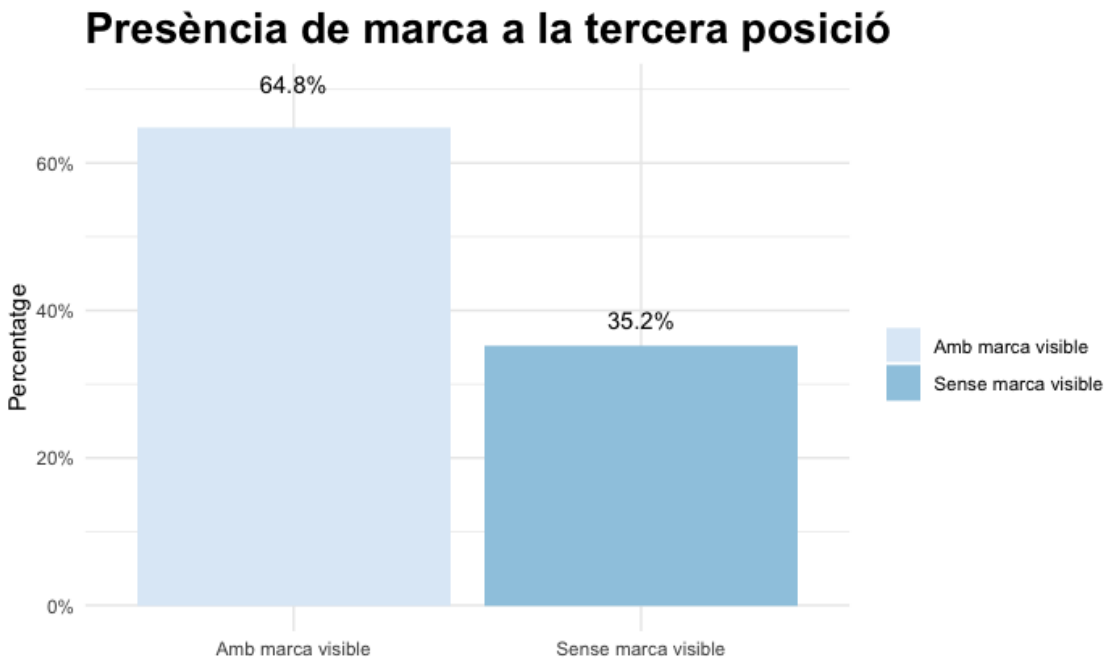

Figura A.12: Distribució de la presència de marca de les peces de roba a la tercera posició dels rànquings.

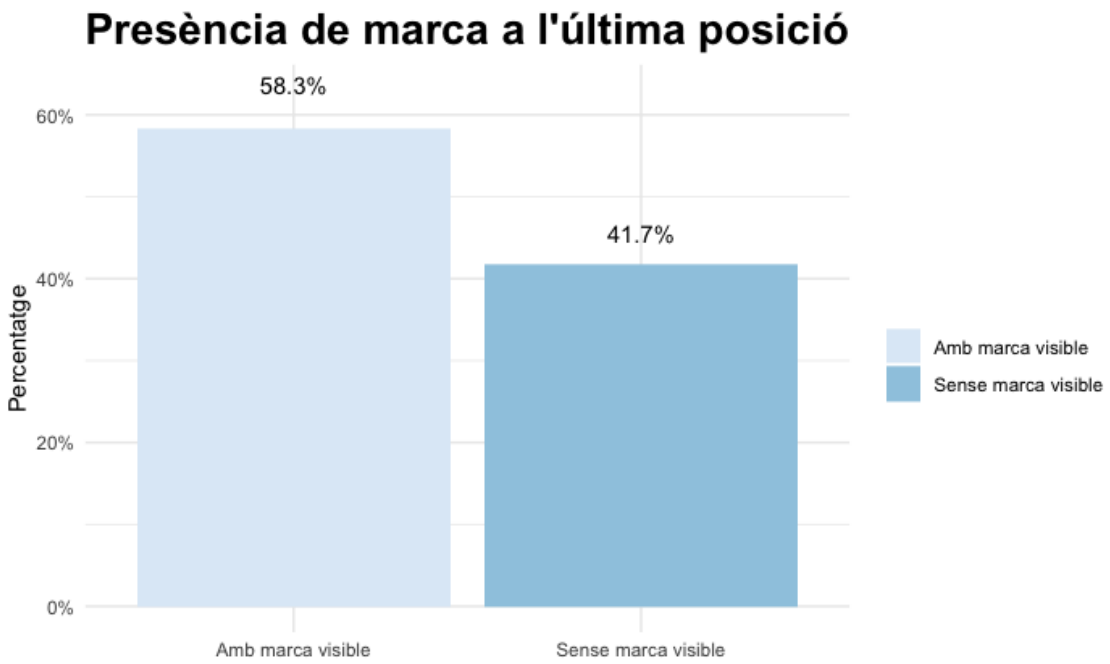

Figura A.13: Distribució de la presència de marca de les peces de roba a l'última (desena) posició dels rànquings.

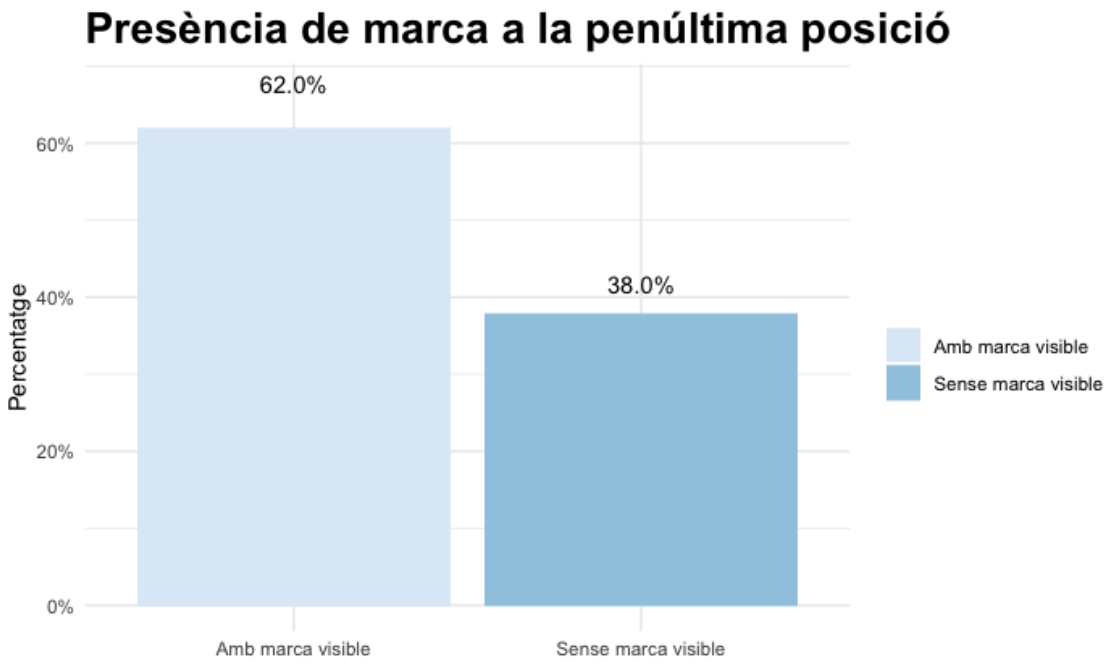

Figura A.14: Distribució de la presència de marca de les peces de roba a la penúltima (novena) posició dels rànquings.

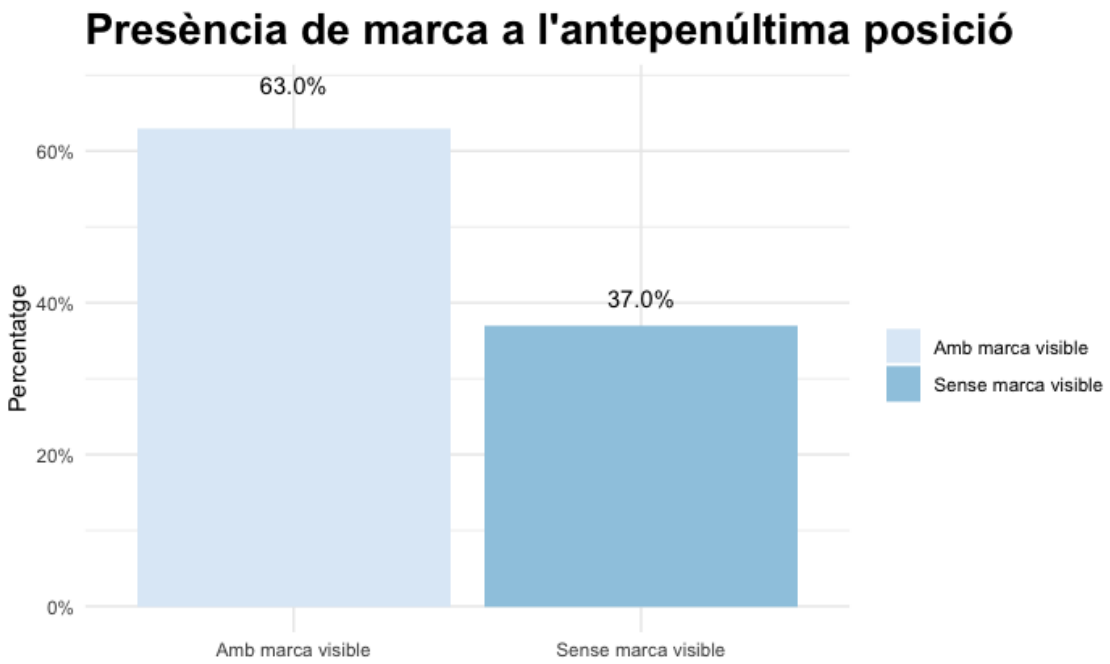

Figura A.15: Distribució de la presència de marca de les peces de roba a l'antepenúltima (vuitena) posició dels rànquings.

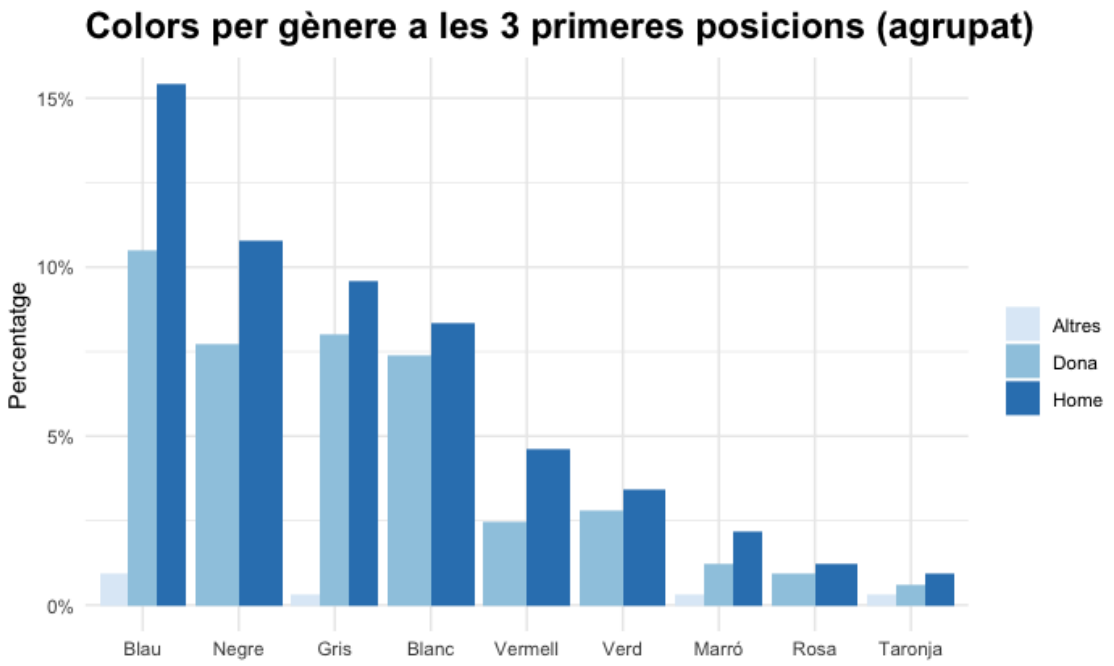

Figura A.16: Per les tres primeres posicions dels rànquings, distribució dels gèneres respecte els colors.

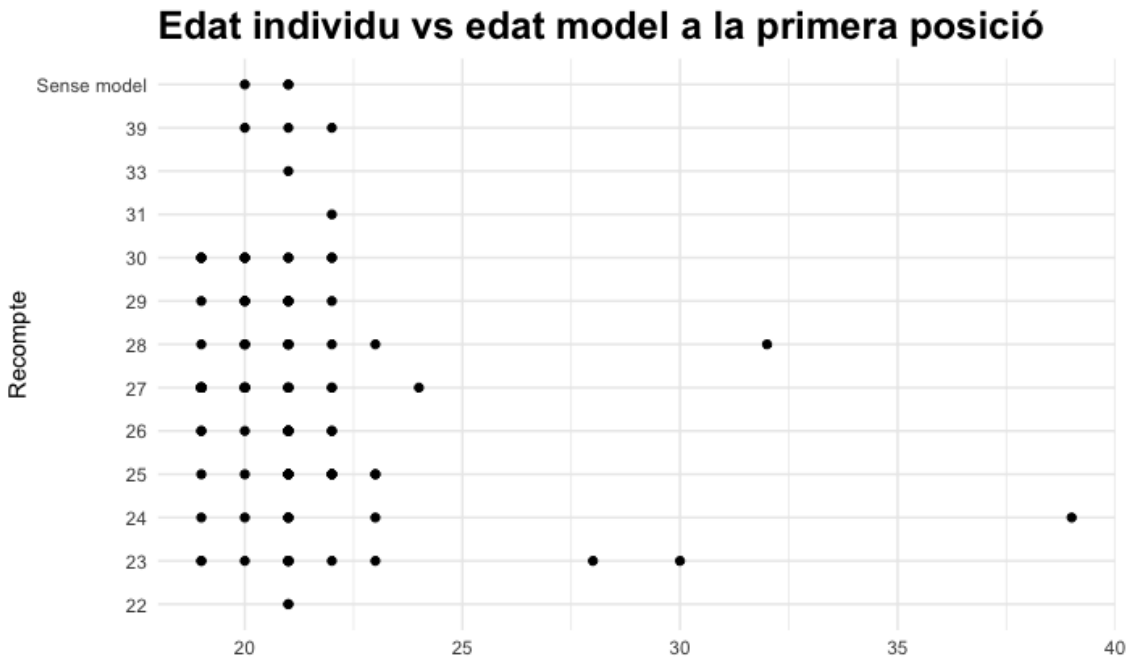

Figura A.17: Distribució de l'edat de l'individu enquestat versus l'edat del model de la imatge que ha posat a la primera posició.

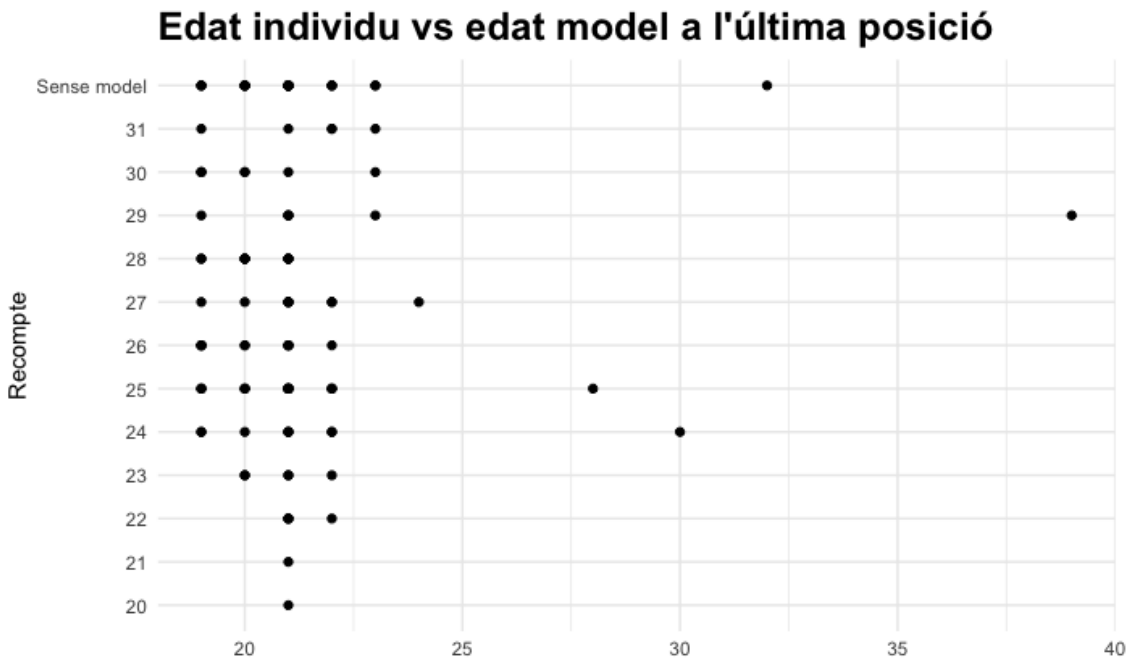

Figura A.18: Distribució de l'edat de l'individu enquestat versus l'edat del model de la imatge que ha posat a l'última posició.

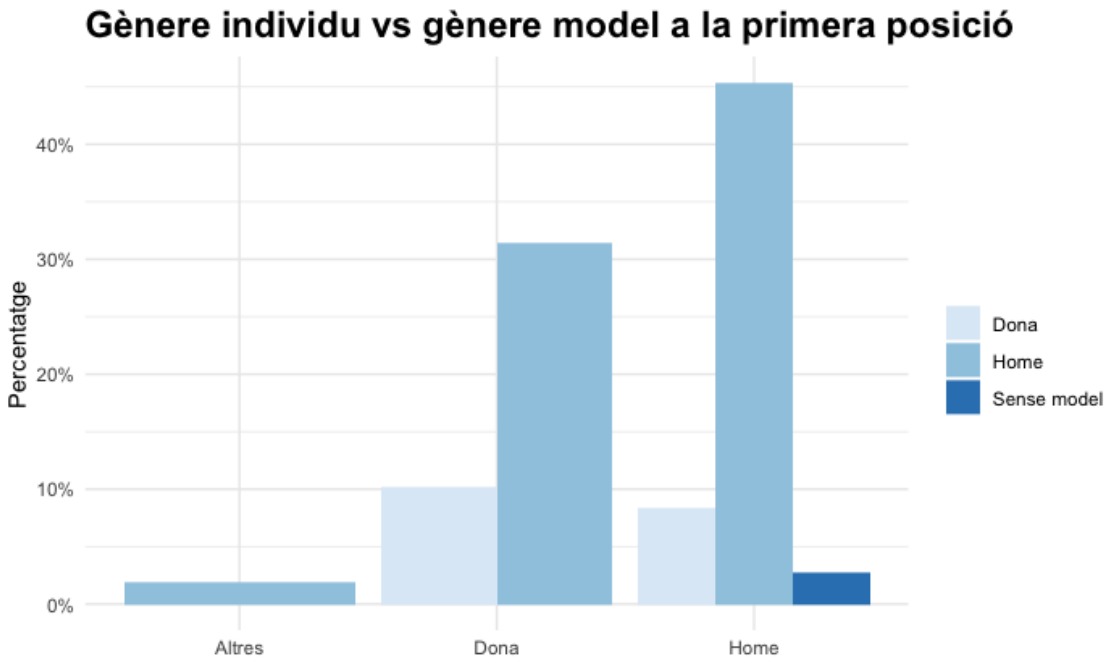

Figura A.19: Gènere dels individus enquestats (eix de les X) versus gènere del model de la imatge que han posat a la primera posició.

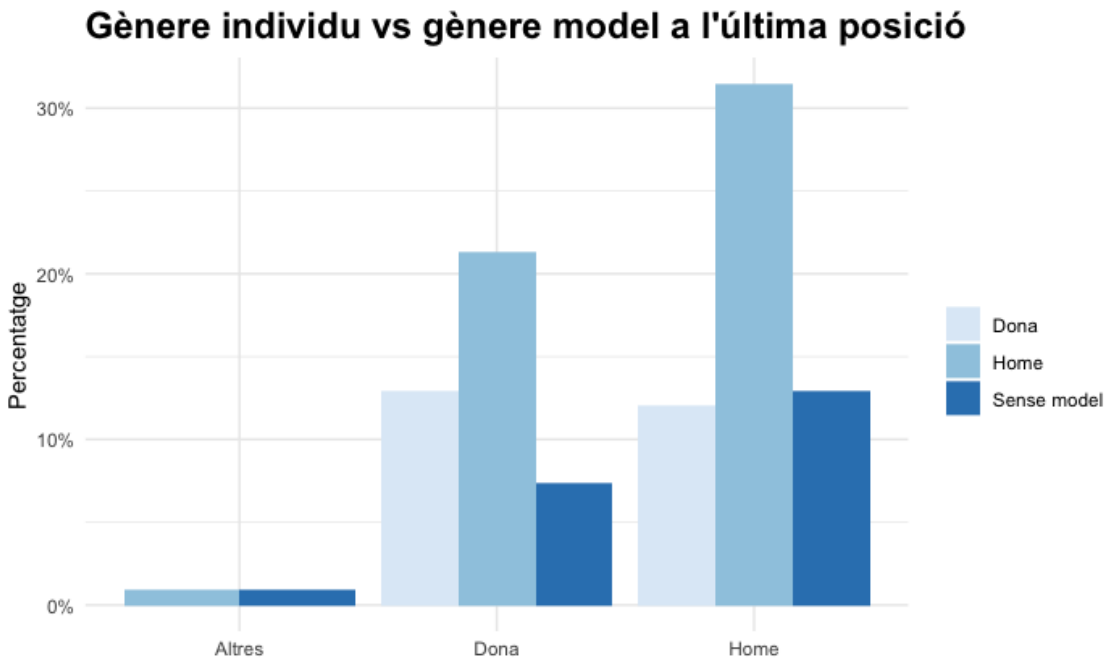

Figura A.20: Gènere dels individus enquestats (eix de les X) versus gènere del model de la imatge que han posat a l'última posició.

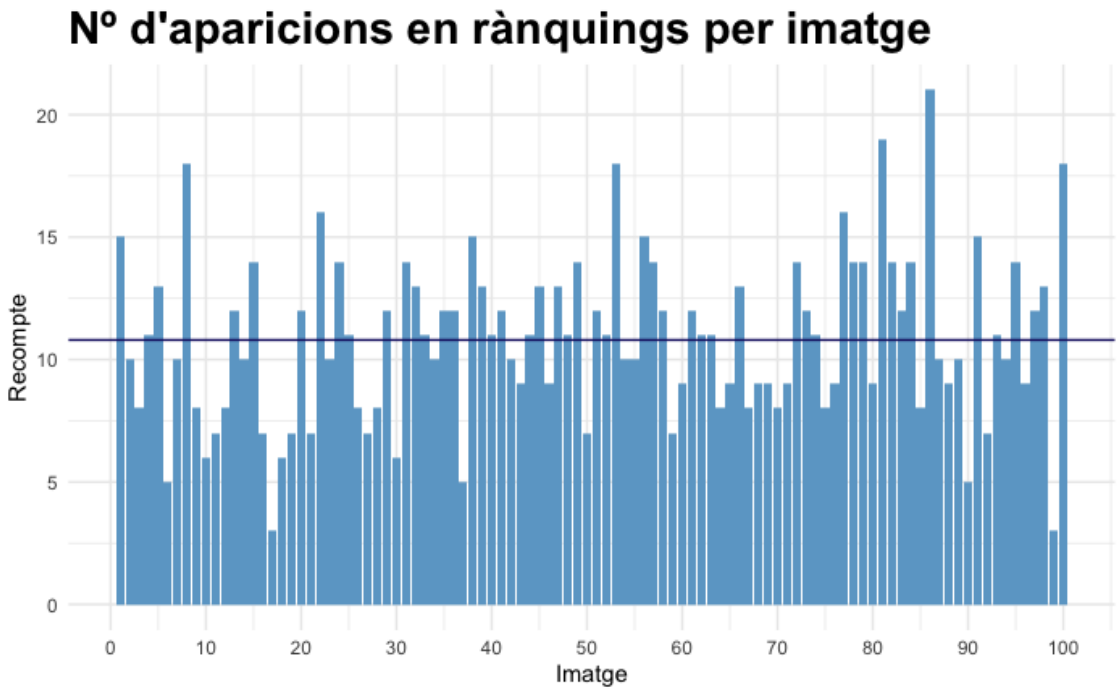

Figura A.21: Nombre de vegades que cada imatge ha aparegut als rànquings. La línia blava fosca indica la mitjana, que es troba al voltant d'11 aparicions per imatge.