TREBALL DE FI DE GRAU

Grau en Estadística

Títol: COMPARACIÓ DE TÈCNIQUES DE CLUSTERING EN UNA BASE DE DADES DE SALUT

Autor: Xavier Ventayol Farras

Directors: Daniel Fernández Martínez i Albert Sanchez Niubó

Departament: Estadística i Investigació Operativa

Convocatòria: Juny 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

Les tècniques de *clustering* tenen l'objectiu de trobar patrons amagats dins de les dades i, particularment, dividir un conjunt d'observacions en grups acord a un conjunt de mesures. Els primers mètodes es van desenvolupar als anys 30 i 40 i avui dia n'existeixen més d'un centenar.

En aquest treball s'han estudiat tres tècniques de *Hard clustering*: K-means, *clustering* jeràrquic, K-medoids i una tècnica de *Soft clustering*: *Gaussian Mixture Models*. Addicionalment, s'han seleccionat aleatòriament dues mostres de 200 persones de l'estudi de salut ELSA amb els objectius d'il·lustrar aquests mètodes per descobrir quins s'adapten millor a aquestes dades i de determinar grups de persones, estratificats per sexe, amb perfils de salut comuns.

El K-means i el *clustering* jeràrquic aglomeratiu són les tècniques que han presentat els millors resultats. En canvi, els *Gaussian Mixture Models* és el mètode que pitjor s'ha adaptat a les dues mostres analitzades.

Paraules clau:

K-means, *clustering* jeràrquic, K-medoids, *Gaussian Mixture Models*, *Hard clustering*, *Soft clustering*, Índex de Rand Ajustat, *elbow method*, *average silhouette method*, *gap statistic*.

ABSTRACT

Clustering techniques aim to find hidden patterns within the data and to divide a set of observations into groups according to a set of measures. The first methods were developed in the 1930s and 1940s and today there are more than a hundred.

In this work, three Hard clustering techniques: K-means, hierarchical clustering, K-medoids and a Soft clustering technique: Gaussian Mixture Models have been studied. In addition, two samples of 200 people have been randomly selected from the ELSA health study with the objectives of illustrating these methods to discover which are best suited to these data and to determine groups of people, stratified by sex, with common health profiles.

K-means and agglomerative hierarchical clustering were the best performing techniques. On the other hand, the Gaussian Mixture Models is the method that has been the worst adapted to the two samples analyzed.

Key words:

K-means, hierarchical clustering, K-medoids, Gaussian Mixture Models, Hard clustering, Soft clustering, Adjusted Rand Index, elbow method, average silhouette method, gap statistic.

CLASSIFICACIÓ AMS

- 62H30 Classification and discrimination; cluster analysis
- 91C20 Clustering

Índex

I. INTRODUCCIÓ

Actualment, ens trobem en una època en què les empreses, els governs i les organitzacions tenen la necessitat d'obtenir la màxima informació possible i emmagatzemar-la en grans bases de dades. Malgrat això, tota aquesta informació necessita ser tractada per poder mostrar patrons d'utilitat de cara a extreure'n conclusions i, sobretot, prendre mesures enfocades a evolucionar, aprendre i millorar en el nostre dia a dia.

Normalment, les dades estan emmagatzemades com un conjunt d'observacions (que habitualment poden ser subjectives o individuals) a les que se'ls hi pren una sèrie de mesures. Per tractar aquestes dades, ens pot interessar classificar els individus en grups diferenciats, ja que es pot donar el cas que, en una mateixa base de dades, els individus es comportin de manera diferent en funció de les característiques que s'analitzen. A simple vista i sense tractar les dades, extreure patrons diferents en un *dataset* és pràcticament impossible, per això existeixen diferents tècniques de classificació, com per exemple les tècniques de *clustering*, que ens permeten assolir aquest objectiu.

En aquest treball s'estudiaran, per una banda, les tècniques de *clustering* més utilitzades avui dia: *clustering* jeràrquic, K-means i K-medoids, que són tècniques basades en dissimilaritats que, per a un conjunt d'atributs continus, calculen la similitud i la dissemblança entre les observacions d'una base de dades a través de distàncies matemàtiques (p.ex. la distància Euclidiana). D'altra banda, s'estudiarà una aproximació probabilística més actual de les tècniques de *clustering* basada en els *finite mixture models* (en català, models de barreges finites), el *clustering* probabilístic, en concret es tracta dels *Gaussian Mixture Models*.

1.1. Objectius del treball

Els objectius d'aquest treball es poden dividir en dues àrees ben diferenciades. El primer objectiu és estudiar i conèixer la metodologia dels diferents mètodes de *clustering* escollits en aquest treball: *clustering* jeràrquic, K-means, K-medoids i *Gaussian Mixture Models*. A més a més, es treballaran tres tècniques comunes de selecció de clústers òptims: l'*elbow method*, l'*average silhouette method* i el *gap statistic*, i una tècnica per comparar les estructures dels clústers entre diferents mètodes, l'Índex de Rand Ajustat (ARI). Assolir aquest objectiu em proporcionarà un *know-how* en l'àrea de *clustering* que no tenia al principi d'aquest treball.

El segon objectiu radica en, una vegada assolit el primer objectiu, aplicar els coneixements adquirits per una mostra procedent d'un estudi real de salut. Aquest objectiu em permetrà il·lustrar els coneixements adquirits i, empíricament, determinar quin mètode divideix millor les observacions per aquesta mostra i, a més d'això, si les quatre tècniques classifiquen les dades de manera similar o, per contra, succeeix el que en *Machine Learning* es coneix com el "*no free lunch methods*". És a dir, per a una mateixa base de dades no tots els mètodes extrauran les mateixes estructures de *clustering*, no perquè un mètode sigui millor o pitjor que un altre, sinó perquè depenent de les característiques que presenti la base de dades hi haurà algoritmes que s'adaptin millor al problema que es vol tractar.

1.2. Metodologia i estructura

El treball s'estructura en dues parts, la primera és la part purament teòrica on es presentaran tots els mètodes, i la segona és la part pràctica on s'aplicaran els mètodes mencionats anteriorment.

Tenint en consideració l'índex, la part teòrica cobreix del capítol II al capítol IV. Al capítol II s'introdueixen els mètodes de *clustering*, s'exposa cada tècnica amb les seves característiques i l'esquema que segueix l'algoritme. Al capítol III es presenten els tres mètodes de selecció de clústers òptims i al capítol IV s'explica com es compararan els mètodes de *clustering* i l'Índex de Rand Ajustat. Al final de cada mètode s'especifiquen quines funcions del software estadístic R són necessàries per dur-lo a terme.

La segona part del treball compren el capítol V i el capítol VI. En el capítol V es fa una breu anàlisi exploratòria de la base de dades de salut on s'apliquen els mètodes de *clustering*, s'estratifica la base de dades en funció del sexe, de manera que per a cada mètode s'acabaran analitzant dues mostres, una de dones i una altra d'homes, i s'especifiquen quines variables contínues s'utilitzen a l'hora de crear els clústers. Finalment, al capítol VI s'apliquen totes les tècniques introduïdes a la part teòrica per la mostra de dones i per la mostra d'homes, es fa el *profiling* dels clústers a través dels tests ANOVA i Kruskal-Wallis, per a les variables numèriques, i el test Chi-quadrat, per a les variables categòriques, i es fa una breu comparativa entre els diferents mètodes per veure si, la mostra de dones per una banda i la mostra d'homes per l'altre, perfilen les dades de manera similar i si les estructures de clústers que es creen per cada algoritme són similars entre elles.

En el darrer capítol s'exposa el que s'ha anat aprenent durant tot el treball i les conclusions a les quals s'ha arribat per la mostra de dones i per la mostra d'homes.

Pel que fa a la base de dades de salut amb les que s'ha treballat, s'han extret les dades de l'estudi de cohort *The English Longitudinal Study of Ageing* (ELSA) pels individus que van participar en l'estudi entre els anys 2008 i 2009.

Tota la informació sobre l'estudi ELSA i les dades que s'utilitzen es pot trobar al següent enllaç: https://www.elsa-project.ac.uk/.

1.3. Agraïments

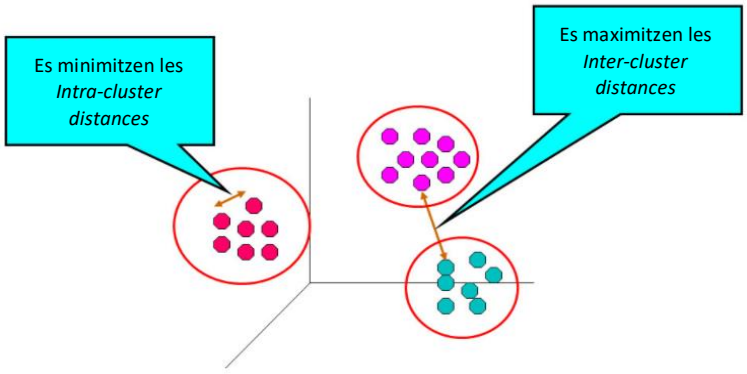
Gràcies als meus tutors, Daniel Fernández Martínez i Albert Sanchez Niubó, per tota l'ajuda proporcionada, per totes les reunions telemàtiques que hem pogut tenir i la predisposició que han mostrat a l'hora de resoldre tots els dubtes que m'han anat sorgint durant el treball de la manera més ràpida i eficient possible.

II. TÈCNIQUES DE CLUSTERING

El terme *clustering* es refereix a un gran nombre de tècniques no supervisades en l'anàlisi de dades multivariant amb la finalitat de dividir el conjunt d'observacions d'un *dataset* en diferents grups anomenats clústers, conglomerats, classes o simplement grups. L'objectiu d'aplicar el *clustering* és trobar estructures de clúster on les observacions dins de cada grup siguin similars entre elles i diferents de les observacions en els altres grups. Tenint en compte que la similitud entre observacions és una quantitat que reflecteix la força de la relació entre dos elements de dades i representa com són de similars dos patrons de dades. A més, el *clustering* és una eina que ajuda a determinar patrons amagats, és a dir, estructures que no es veuen ni es coneixen a simple vista.

En el cas particular d'aquest treball només s'aplicaran mètodes de *clustering* enfocats a variables contínues. Així, la similitud entre dades es calcularà a través de distàncies matemàtiques (es veuran en detall al subapartat 2.1.2). En altres paraules, aquests mètodes pretenen minimitzar les diferències/distàncies dins de cada clúster (*Intra-cluster distances*) i maximitzar les distàncies entre els diferents clústers (*Inter-cluster distances*) com es pot veure a la Figura 2.1.

Figura 2.1: Visualització de les distàncies dins dels clústers (Intra-cluster distances) i les distàncies entre clústers (Inter-cluster distances).



Font: <https://www.researchgate.net/figure/Intra-cluster-distances-vs-inter-cluster-distances_fig22_336111538>.

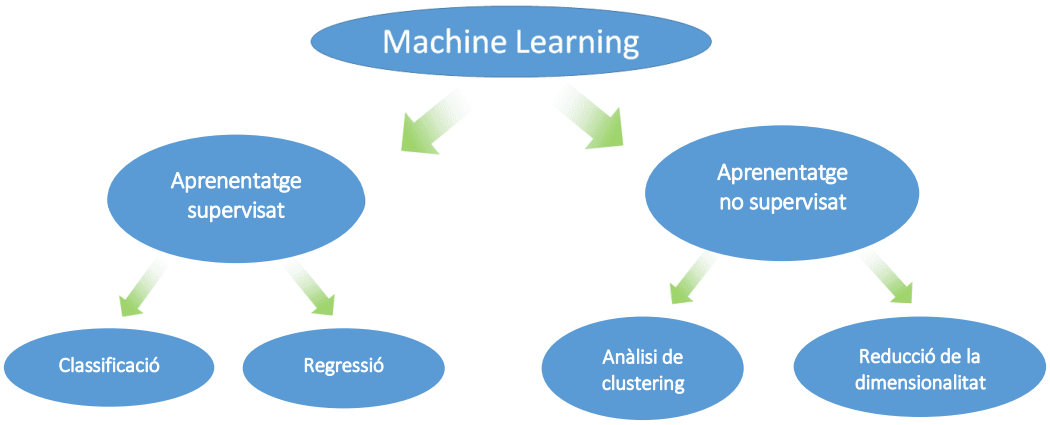
L'anàlisi de *clustering* forma part de l'àrea de *Machine Learning* (aprenentatge automàtic). El *Machine Learning* és una aplicació dins de la Intel·ligència Artificial (IA) que proporciona als sistemes informàtics la possibilitat d'aprendre i millorar automàticament a partir de l'experiència sense haver-se programat explícitament. L'aprenentatge automàtic se centra en el desenvolupament d'algoritmes matemàtics que poden accedir a les dades i utilitzar-les per aprendre per ells mateixos.

Els mètodes de *Machine Learning* es classifiquen, a grans trets, entre les tècniques d'aprenentatge supervisat i les tècniques d'aprenentatge no supervisat (veure Figura 2.2), tot i que també hi ha variants com l'aprenentatge semisupervisat, on es combina el coneixement a priori de l'usuari i les tècniques no supervisades en un mètode de dues fases.

Parlant de forma general, els mètodes classificats com aprenentatge supervisat són aquells on l'analista o usuari ja sap la variable resposta que es vol predir o classificar (depenent de l'objectiu a obtenir) i aquest només actua de guia per entrenar a l'algoritme d'aprenentatge les conclusions a les quals ha d'arribar. Per tant, requereix que els resultats de l'algoritme siguin coneguts amb anterioritat i que les dades utilitzades per entrenar-lo ja estiguin etiquetades amb la resposta correcta. La regressió lineal i logística o l'anàlisi discriminant són exemples molt utilitzats de tècniques d'aprenentatge supervisat.

En canvi, a les tècniques d'aprenentatge no supervisat no es disposa d'una variable resposta o una etiqueta de classe a predir o classificar coneguda. La idea general és que el mètode sigui capaç d'aprendre a identificar processos i patrons complexos durant el procés de l'algoritme d'aprenentatge. L'anàlisi de *clustering* està classificat dins de l'aprenentatge no supervisat.

Figura 2.2: Classificació del *Machine Learning* entre l'aprenentage supervisat i l'aprenentatge no supervisat.



Font: <https://www.diegocalvo.es/en/machine-learning-supervised-unsupervised/machine-learning-classification/>.

Les tècniques de *clustering* s'apliquen en moltes àrees d'investigació com la biologia, l'enginyeria, l'economia, la medicina, etc. donat que permet a l'analista dividir les dades per estudiar característiques i comportaments, a priori desconeguts, que són comuns dins de la mostra i actuar en funció dels resultats. Per exemple, en el món del màrqueting l'analista pot estar interessat a distingir perfils de compradors per vendre productes en funció dels gustos i les necessitats que tinguin.

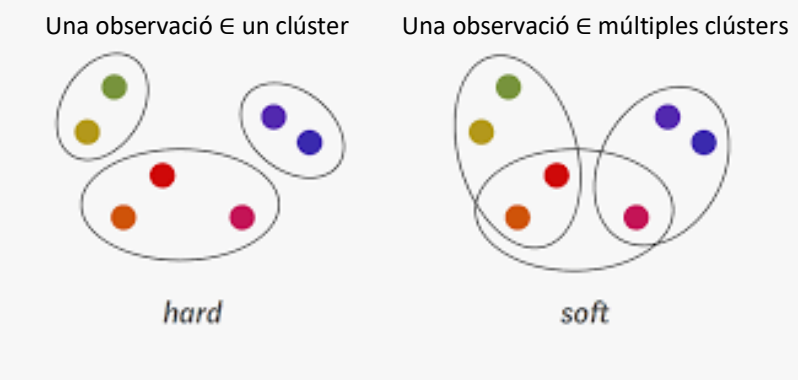
No obstant això, s'ha de tenir en compte que el *clustering* es tracta d'una tècnica descriptiva per explorar i descriure les dades, diferent de les tècniques inferencials o de predicció, com per exemple els models de regressió, on es busca predir una variable resposta en concret. De totes maneres, una vegada es té una estructura de *clustering* definida, es pot classificar a un nou objecte o individu si es disposa de les seves mesures corresponents a les variables utilitzades per definir el *clustering*.

Avui en dia, existeixen diferents metodologies per definir la similitud entre el conjunt d'observacions d'una base de dades i és una àrea en expansió, ja que actualment es coneixen més de 100 mètodes de *clustering* diferents. És important remarcar que les solucions que s'obtenen a l'hora d'aplicar el *clustering* no són úniques atès que, principalment, el resultat

pot variar depenent de les variables que es seleccionen per agrupar les dades i el mètode que s'utilitza. No tots els mètodes de *clustering* donaran les mateixes solucions per les mateixes dades (això és conegut en *Machine Learning* com el "*no free lunch methods*").

Existeixen diverses formes possibles per classificar els mètodes de *clustering*. Per aquest treball es dividiran les tècniques de *clustering* en dos subgrups: el ***clustering* no probabilístic** (*Hard clustering*) i **el *clustering* probabilístic** (*Soft clustering*). En el *clustering* no probabilístic, cada observació de la base de dades pertany a un sol clúster i no se sap quina és la probabilitat de pertànyer a un clúster diferent. D'altra banda, en el *clustering* probabilístic cada observació té associada una probabilitat de pertànyer a més d'un clúster a la vegada (veure Figura 2.3).

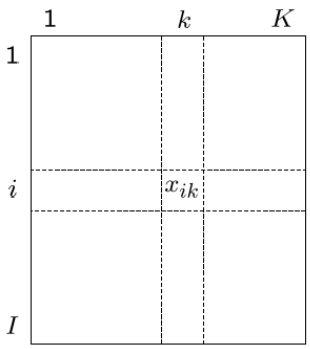
Figura 2.3: Exemple de com es classifiquen les dades en el *Hard clustering* (p.ex. el mètode jeràrquic, K-means i K-medoids) i en el *Soft clustering* (p.ex. els *Gaussian Mixture Models*).



Font: <https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04>

A grans trets, es pot definir que els mètodes de *clustering* probabilístics consideren que les dades provenen d'una distribució de probabilitat que és una barreja de dos o més clústers. Un dels mètodes més populars són els *Gaussian Mixture Models* (GMMs), que s'aplica per a variables contínues que han de seguir una distribució Gaussiana.

Figura 2.4: Matriu de dades pels mètodes de *clustering* amb I individus i K variables numèriques per crear els clústers.



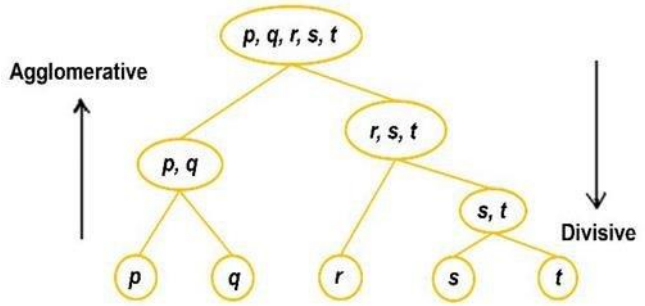
Font: Fernández, Daniel. Multivariate Analysis (MVA) Clustering [projecció visual]. Universitat Politècnica de Catalunya , 2021. 94 diapositives.

Dins del *clustering* no probabilístic es pot distingir entre dos grans grups de tècniques: els mètodes jeràrquics i els mètodes de partició. En aquest cas les variables contínues no han de seguir cap distribució de probabilitat, donat que són mètodes que depenen de distàncies matemàtiques i, per tant, no tenen associada cap distribució de probabilitat subjacent. Per aplicar els mètodes de *Hard clustering*, la matriu de dades amb la informació dels individus i les variables que s'utilitzen per crear els clústers pren la forma de la Figura 2.4.

Els mètodes jeràrquics consideren que les observacions més properes entre elles, a l'espai de les dades, presenten més similitud que les observacions que es troben més llunyanes entre si. En cada pas d'aquests mètodes es van generant grups fins que s'assoleix el nombre d'agrupacions òptimes considerades per l'algoritme.

El *clustering* jeràrquic es pot dividir entre les **tècniques d'aglomerat**, en què cada observació forma un clúster i es van unint fins a arribar a un únic clúster que conté totes les dades, i les **tècniques de divisió**, en aquest cas es parteix d'un sol clúster que agrupa totes les observacions i en cada pas es divideix en grups més heterogenis, acabant amb tants clústers com elements dins de la mostra.

Figura 2.5: Exemple del *clustering* aglomeratiu (*Agglomerative*) i divisiu (*Divisive*) per dividir les dades en clústers.



Font: <https://www.researchgate.net/figure/Conceptual-dendrogram-for-agglomerative-and-divisive-Hierarchical-based-clustering-19_fig2_321399805>.

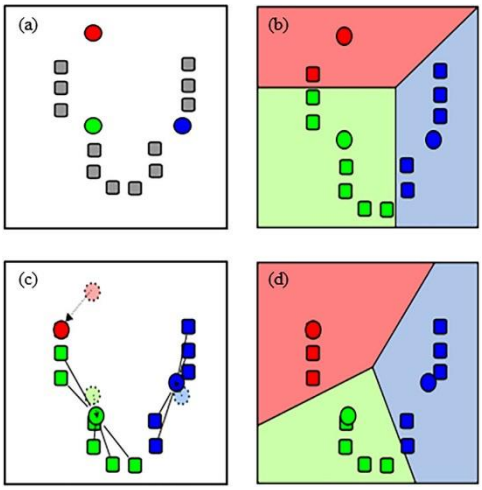
A la Figura 2.5 es pot observar un exemple de l'esquema de les tècniques d'aglomerat, a la part esquerra, i l'esquema de les tècniques de divisió, a la part dreta. En els subapartats 2.1.3 i 2.1.4 s'exposen detalladament els passos estàndard que segueix cada algoritme de *clustering* jeràrquic.

Per altra banda, els mètodes de partició són algoritmes de *clustering* iteratius que s'executen fins a trobar la millor partició de l'espai que es genera en la dimensió de les dades. Abans d'aplicar aquest tipus de mètodes és necessari que l'analista especifiqui el nombre de clústers que es volen crear. Per aquest motiu, és importat tenir coneixements a priori sobre les dades o utilitzar eines per calcular una aproximació del nombre òptim de particions com l'e*lbow method*, l'*everage silhouette method* i el *gap statistic* (veure capítol III). Dos dels desavantatges d'aquest tipus de mètodes són que l'algoritme pot acabar convergint a un

òptim local en lloc del màxim i, per això, es requereix executar els mètodes de partició molts cops des de diferents punts de partida de l'espai dimensional de les dades i veure a quin resultat convergeix totes les execucions de l'algoritme.

Els mètodes de partició més coneguts i utilitzats actualment són el K-means (veure Figura 2.6) i el K-medoids.

Figura 2.6: Exemple dels primers passos de l'algoritme de partició K-means (a l'apartat 2.2 es presenten els passos detalladament).



Font: *Wikipedia* < https://es.wikipedia.org/wiki/K-medias >

A continuació es descriurà les tècniques de *clustering* que s'aplicaran i estudiaran en aquest treball. Per tenir una idea inicial, a la Taula 2.1 es mostra una classificació dels mètodes estudiats segons si formen part del *clustering* no probabilístic o del *clustering* probabilístic.

Taula 2.1: Classificació de les tècniques del *clustering* no probabilístic i el *clustering* probabilístic.

CLUSTERING NO PROBABILÍSTIC *(Hard Clustering)*	
Mètodes Jeràrquics	Mètodes de partició
- *Clustering* Aglomeratiu - *Clustering* Divisiu	- K-means - K-medoids
CLUSTERING PROBABILÍSTIC *(Soft Clustering)*	
- *Gaussian Mixture Model (GMM)*	

2.1. Clustering Jeràrquic

El *clustering* Jeràrquic (*Ward, 1963*) és un mètode no probabilístic i desenvolupat per a variables numèriques. A més a més, es tracta de les tècniques de *clustering* més populars (junt

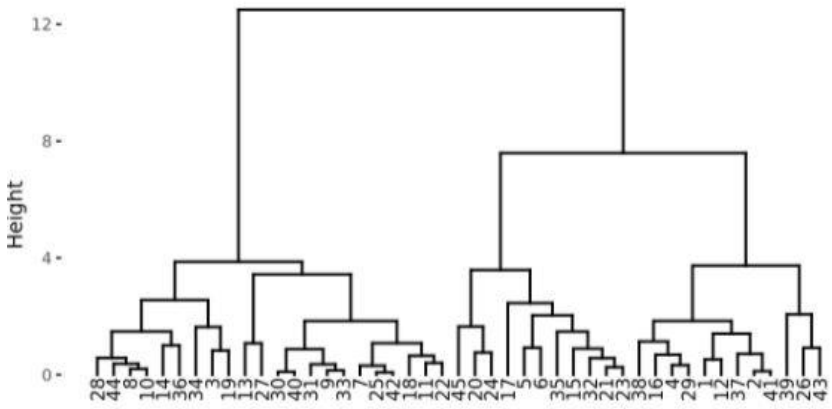
amb el mètode de partició K-means explicat a l'apartat 2.2). Com en tot anàlisi de *clustering* es necessita aplicar certes mesures de similitud i dissemblança i, al tractar-se d'un mètode numèric, aquestes mesures es representen com a distàncies, per exemple la distància Euclidiana, la distància de Manhattan i la distància Euclidiana al quadrat entre d'altres (veure Taula 2.2).

Dins d'aquest grup de mètodes existeixen dues estratègies diferents per agrupar les observacions en clústers. La primera estratègia és el *clustering* aglomeratiu (*bottom-up*) exposat al subapartat 2.1.3, el més utilitzat, i la segona estratègia és el *clustering* divisiu (*top-down*) exposat al subapartat 2.1.4, no tan utilitzat. Indistintament del tipus de *clustering* jeràrquic emprat, aquests mètodes creen els conglomerats de manera que existeixi un ordre predeterminat entre els clústers de la base de dades a estudiar, és a dir, una jerarquia. Aquesta jerarquia es mostra a través de dendrogrames, que es defineixen amb detall just a continuació.

2.1.1. Dendrograma

Un avantatge respecte a altres tècniques per dividir les dades en conglomerats, és que els mètodes jeràrquics permeten representar els resultats del *clustering* a través d'un diagrama d'arbre, que il·lustra les relacions entre clústers dins de la base de dades analitzada. Aquest diagrama pren el nom de dendrograma.

Figura 2.7: Visualització d'un dendrograma on es pot veure com les observacions es divideixen en branques.



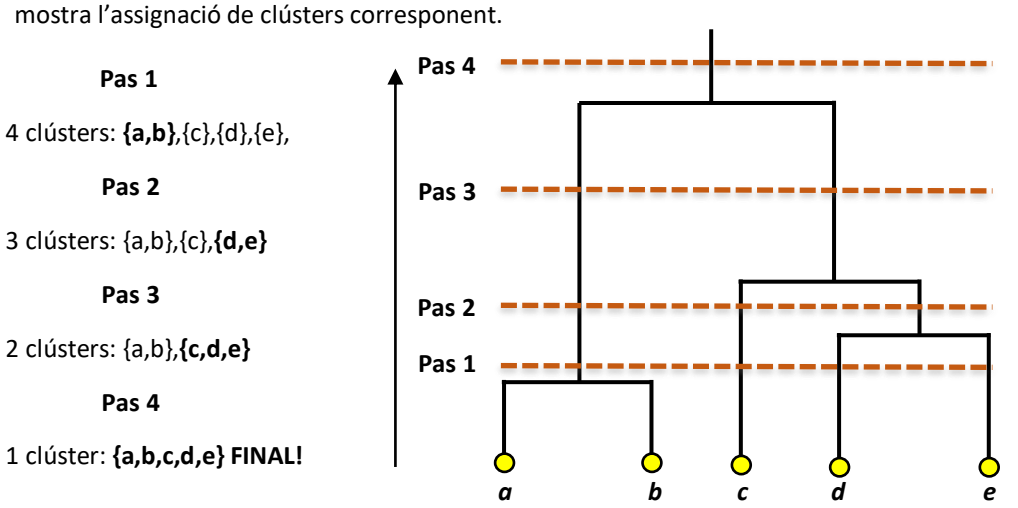
Font: *Clustering y heatmaps: aprendizaje no supervisado*
< https://rpubs.com/Joaquin_AR/310338 >

Agafant la Figura 2.7 com a exemple, en la base del dendrograma cada observació forma un clúster individual conegut com a fulla (*leaf*) de l'arbre. A mesura que s'ascendeix per l'estructura del diagrama, parells de fulles es fusionen formant les primeres branques, aquestes unions es coneixen com a nodes i representen els parells d'observacions més similars entre elles, més tard es van fusionant branques amb branques i també branques amb fulles.

Quan abans es produeix una fusió (més proper a la base del dendrograma), major és la similitud. Això significa que, per qualsevol parell d'observacions, es pot identificar el punt de

l'arbre on les branques es fusionen. L'altura (*Height*) on es produeix aquesta fusió, en l'eix vertical, indica com de similars o dissemblants són els parells d'objectes. L'altura dels dendrogrames normalment ve calculada a través de la deviancia o l'índex de GINI. Per tant, els dendrogrames s'han d'interpretar respecte a l'eix vertical (eix Y), en cap cas respecte a les posicions que ocupen les observacions en l'eix horitzontal (eix X), ja que no proporcionen informació sobre la similitud i la dissemblança entre parells i grups d'observacions.

Figura 2.8: Exemple d'un esquema aglomeratiu (*bottom-up*) amb 4 passos, on a cada pas es mostra l'assignació de clústers corresponent.



Font: Fernández, Daniel. Multivariate Analysis (MVA) Clustering [projecció visual]. Universitat Politècnica de Catalunya , 2021. 94 diapositives.

A la Figura 2.8 es pot observar la interpretació de l'altura dels dendrogrames a través d'un esquema aglomeratiu. En el pas 1, les observacions *a* i *b* s'uneixen formant un primer node i són les observacions més similars entre elles, ja que s'agrupen en la part més propera a la base del dendrograma. En el pas 2, s'uneixen les observacions *d* i *e* donat que són les segones més properes a la base del dendrograma. En el pas 3, l'observació *c* s'agrupa amb el node de les observacions *d* i *e* i, finalment, en el pas 4 totes les observacions formen un únic clúster.

A l'hora de decidir el nombre de clúster en què es dividirà les dades, hi ha vàries opcions. Una primera opció, és anar provant diferents talls horitzontals a les branques més llargues del dendrograma fins que es troba una bona diferenciació entre clústers. Una altra opció, és utilitzar el coneixement previ del context del problema que representen les dades i aplicar-ho per obtenir el tall que té més sentit, contextualment parlant. Una tercera opció, és aplicar mètodes matemàtics, com els que es plantegen al capítol III, per poder escollir el nombre òptim de clústers per la base de dades que s'estudia. Totes tres opcions són no excloents i, per tant, combinables.

2.1.2. Mesures de distància

Per tal d'aconseguir agrupar els objectes/observacions d'una base de dades de manera jeràrquica, tant pel *clustering* aglomeratiu com pel *clustering* divisiu, cal definir la mesura de

distància utilitzada per agrupar les observacions del *dataset* que s'estudia. Les distàncies més comunes a l'hora d'aplicar aquests mètodes es mostren a la Taula 2.2.

Taula 2.2: Nom, fórmula i notació de 4 de les mesures de distància més comunes en el *clustering* Jeràrquic.

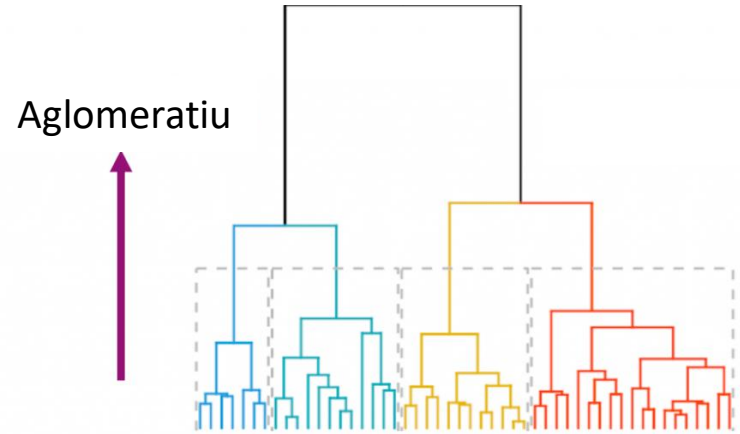
Distància	Fórmula	Notació
Distància Euclidiana	$\|\|a-b\|\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$	a i b representen dos clústers diferents i els parells a_i i b_i representen les observacions i-èssimes de cada clúster.
Distància Euclidiana al quadrat	$\|\|a-b\|\|_2^2 = \sum_i (a_i - b_i)^2$	a i b representen dos clústers diferents i els parells a_i i b_i representen les observacions i-èssimes de cada clúster.
Distància de *Manhattan*	$\|\|a-b\|\|_1 = \sum_i \|a_i - b_i\|$	a i b representen dos clústers diferents i els parells a_i i b_i representen les observacions i-èssimes de cada clúster.
Distància de *Mahalanobis*	$d(a,b) = \sqrt{(a-b)^T S^{-1} (a-b)}$	a i b representen dos clústers diferents. Es resta totes les observacions de cada un dels clústers entre ells. S^{-1} és la matriu de variàncies i covariàncies de les dades.

Per l'aplicació pràctica dels mètodes jeràrquics, es calcularà la similitud i la dissemblança entre els objectes de la base de dades a través de la distància Euclidiana, la més comuna a l'hora d'aplicar els mètodes de *clustering* per a dades de tipus numèric, que és la distància que s'implementa per defecte en la majoria de paquets estadístics en R.

2.1.3. Clustering aglomeratiu

En el *clustering* aglomeratiu, també denominat AGNES (***Ag**lomerative **nes**ting*), el mètode s'inicia des de la base del dendrograma, de manera que cada observació forma un clúster individual. A cada pas de l'algoritme, els dos clústers més similars es combinen per crear un

Figura 2.9: Exemple del *clustering* aglomeratiu (*bottom-up*).



Font: <https://www.mygreatlearning.com/blog/hierarchical-clustering/>

nou clúster, de tal manera que l'estructura del dendrograma creix de baix a dalt (*bottom-up*) fins a crear un únic conglomerat, la branca central (veure Figura 2.9).

Per aplicar aquest tipus de *clustering* jeràrquic cal definir com es quantifica la similitud entre dos clústers. És a dir, cal ampliar el concepte de distància entre parells d'observacions per aplicar-ho a parells de clústers formats per un o més objectes. Aquest procés es coneix com a mesures d'enllaç (*linkage*). A la Taula 2.3 es defineixen les cinc mesures d'enllaç més utilitzades: **l'enllaç complet o màxim**, **l'enllaç únic o mínim**, **l'enllaç mitjà**, **l'enllaç per centroides** i el **mètode de mínima variància de Ward**. També, a la Figura 2.10 es representen aquestes mesures gràficament.

Taula 2.3: Presentació de les 5 mesures d'enllaç més comunes amb la seva fórmula i definició.

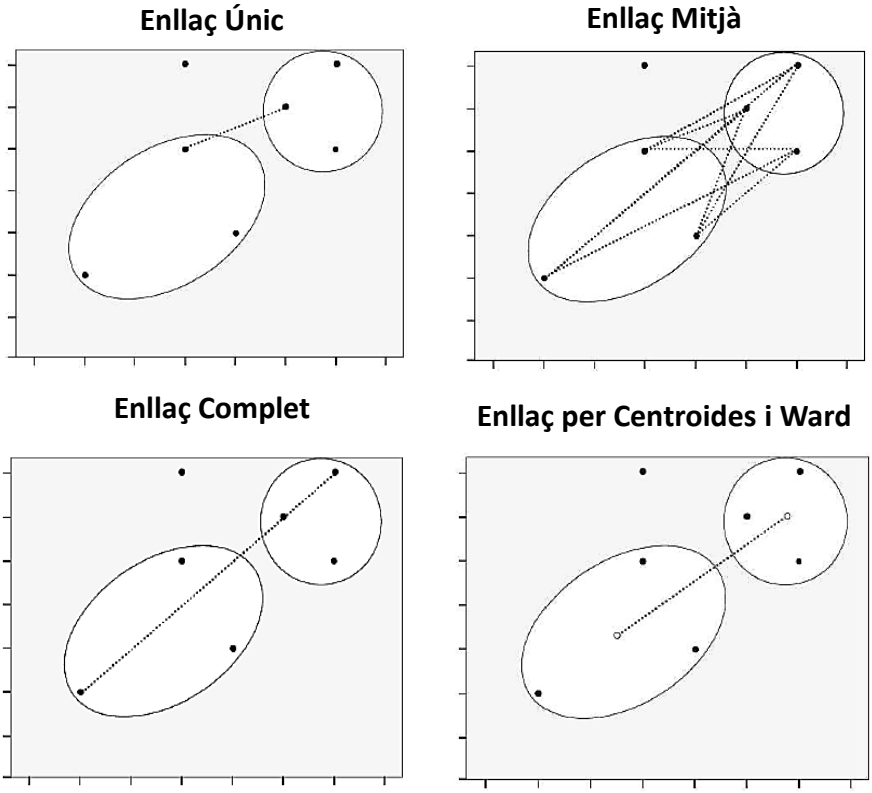
Enllaç	Fórmula	Definició
Complet o màxim	$D_{ab} = max_{i,j} d(x_i, y_j)$	Calcula les dissemblances entre tots els parells d'observacions del clúster a i el clúster b, de manera que el valor més gran passa a ser la distància entre aquests dos clústers. Es considera com la mesura més conservadora i tendeix a crear clústers més compactes.
Únic o mínim	$D_{ab} = min_{i,j} d(x_i, y_j)$	Calcula les dissemblances entre tots els parells d'observacions del clúster a i el clúster b, de manera que el valor més petit passa a ser la distància entre aquests dos clústers. Es considera com la mesura menys conservadora i tendeix a crear clústers més allargats.
Mitjà	$D_{ab} = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} d(x_i, y_j)$	Calcula les dissemblances entre tots els parells d'observacions del clúster a i el clúster b, la mitjana de les dissemblances es considera la distància entre aquests dos clústers. Com que es depèn del nombre d'observacions de cada conglomerat, la compacitat dels clústers que es creen pot variar.
Centroides	$D_{ab} = d\left(\underline{x}, \underline{y}\right)$	Calcula el centroide del clúster a i el clúster b, de manera que es considera la distància entre els centroides com la distància entre els dos clústers.
Mètode de variància mínima de Ward	$D_{ab} = \sqrt{\frac{2 \cdot \lvert n \rvert \cdot \lvert m \rvert}{\lvert n \rvert + \lvert m \rvert}} \cdot \lvert\lvert \overline{x} - \overline{y} \rvert\rvert$	Es minimitza la variància total dins de cada clúster. A cada pas de l'algoritme es combinen els grups d'observacions amb la mínima distància entre clústers. Tendeix a crear clústers més compactes.

La variable x_i es refereix a les observacions del clúster a, la variable y_j es refereix a les observacions del clúster b, n i m representen el total d'observacions del clúster a i b

respectivament, $\underline{x}$ i $\underline{y}$ són els centroides del clúster a i b, $d(x_i, y_j)$ és la distància entre les observacions dels dos clústers i $||\cdot||$ fa referència a la norma Euclidiana.

Els enllaços complet, mitjà i de variància mínima de Ward acostumen a ser els "preferits" entre els analistes de dades, donat que generen dendrogrames més compensats. Tot i això, no es pot determinar que un enllaç sigui millor que un altre perquè depèn del cas d'estudi en què es treballa. A la pràctica, la funció **agnes** del paquet **cluster** (*Maechler et. al, 2019*) permet calcular un coeficient per determinar quin es pot considerar el millor mètode d'enllaç dins d'una mateixa base de dades. Al subapartat 2.1.3.2 s'explicarà breument aquest coeficient.

Figura 2.10: Representació gràfica de les 5 mesures d'enllaç més comunes, amb l'enllaç per centroides i Ward junts donat a la seva similitud gràfica.



Font: *Modified from Prof. Christophe Biernacki's slides* < http://math.univ-lille1.fr/~biernack/ >

2.1.3.1. Esquema de l'algoritme

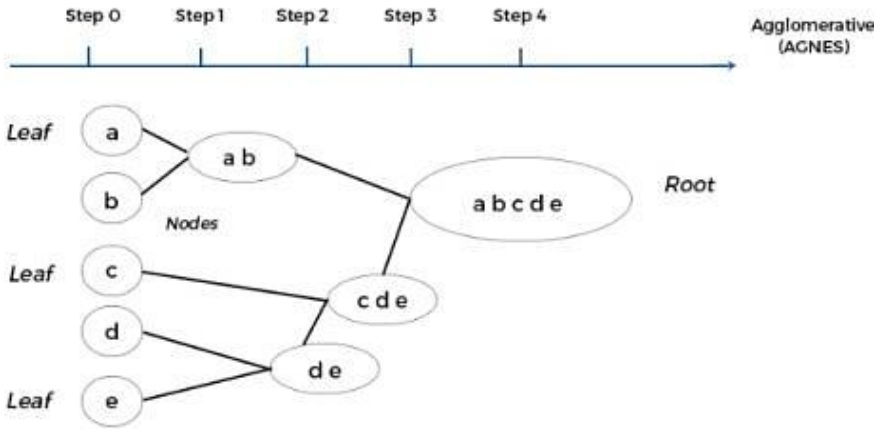
L'estructura final que presenta el *clustering* aglomeratiu s'obté a través d'un algoritme definit pels següents passos (veure Figura 2.11):

1. A l'inici, es considera cada una de les observacions del *dataset* com un clúster individual, formant d'aquesta manera la base del dendrograma.
2. S'inicia un procés iteratiu fins que totes les observacions formen part d'un únic clúster:
   a. Es calcula la distància entre cada possible parell dels k clústers. L'analista ha de decidir quina mesura de distància i enllaç (*linkage*) es vol utilitzar per quantificar la similitud i la dissemblança entre els clústers.

b. Els dos clústers més similars es combinen de manera que queden $k - 1$ clústers.

3. Determinar on tallar el dendrograma que es crea per escollir el nombre de clústers a analitzar.

Figura 2.11: Dendrograma en horitzontal amb els passos del *clustering* aglomeratiu (*Agglomerative*), començant des de la base del dendrograma (esquerra) i acabant amb un únic clúster (dreta).



Font: *Clustering Jerárquico en R* < RPubs - Clustering Jerárquico en R >

2.1.3.2. Aplicació

Pel *clustering* aglomeratiu hi ha diverses funcions a R que permeten dur a terme aquest algoritme. Entre elles hi ha la funció **hclust** del paquet **stats** (*R Core Team, 2020*) i la funció **agnes** del paquet **cluster.**

Les dues funcions operen de manera similar, tanmateix, la funció **agnes**, com a tret diferenciador, calcula el **coeficient d'aglomeració** (AC) que descriu la qualitat de la mesura d'enllaç que s'utilitza a l'hora d'aplicar l'algoritme.

$$AC = \frac{1}{n}\sum_{i} l(i)$$

On $l(i)$ representa l'altura (eix Y) de la unió entre l'observació i-èssima (eix X) del dendrograma. El valor del coeficient està comprès entre 0 i 1, de manera que quan és proper a 1 dona a entendre que s'està davant d'una estructura de *clustering* balancejada, en canvi, quan és proper a 0 dona a entendre que la formació dels clústers no està del tot balancejada i la mesura d'enllaç que s'utilitza no és del tot adient per les dades que s'estudien.

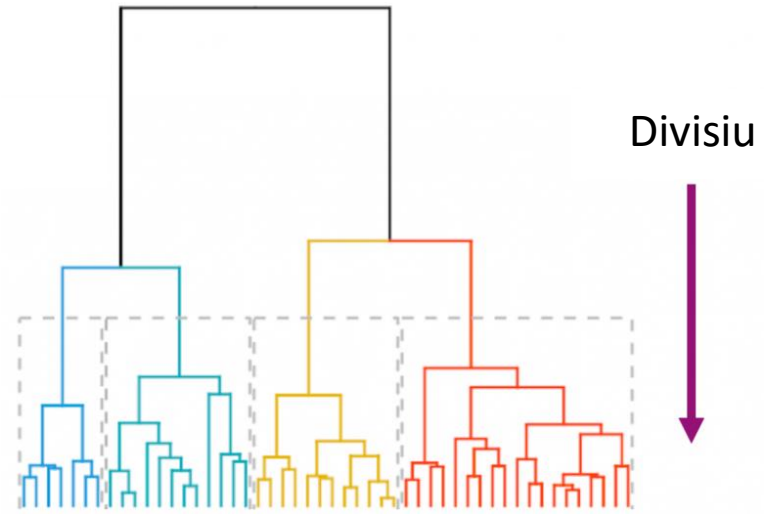
Aquest coeficient s'ha de calcular per una mateixa base de dades amb el mateix nombre d'observacions, ja que a mesura que el nombre d'objectes en la base de dades augmenta, el coeficient també tendeix a augmentar. De cara a la pràctica, el coeficient d'aglomeració de la funció **agnes** servirà com a guia per decidir quina és la mesura de *linkage* que s'adapta millor al *set* de dades que s'estudiarà.

D'altra banda, per tal de poder extreure la variable indicadora dels clústers, es farà ús de la funció **HCPC** del paquet **FactoMineR** (*Sebastien Le et. al, 2008*).

2.1.4. Clustering divisiu

L'algoritme més conegut del *clustering* divisiu és DIANA (**Di**visive **Ana**lysis), a diferència del *clustering* aglomeratiu, en aquest cas, s'inicia l'algoritme a través d'un únic clúster que conté totes les observacions de la base de dades i, a cada pas, es va dividint el conglomerat inicial fins a arribar a formar tants clústers com observacions dins del *dataset* (*top-down*; veure Figura 2.12), per tant, és el procediment a la inversa del *clustering* aglomeratiu.

Figura 2.12: Exemple de *clustering* divisiu (*top-down*).



Font: <https://www.mygreatlearning.com/blog/hierarchical-clustering/>

En cada iteració, es selecciona el clúster amb més diàmetre, entenent que el diàmetre d'un clúster és la màxima diferència entre dues de les seves observacions. Un cop seleccionat el clúster s'identifica l'observació més dispar, aquella que té una distància mitjana major que la resta d'objectes que formen el conglomerat, i aquesta observació inicia un nou clúster. Una vegada separat el clúster es reassignen els objectes amb els conglomerats que s'han obtingut i es torna a fer una nova partició.

Per aquest tipus de *clustering* Jeràrquic no s'utilitzen les mesures d'enllaç, només cal especificar quina mesura de distància es vol emprar per calcular la similitud i la dissemblança entre els nous clústers que es van formant.

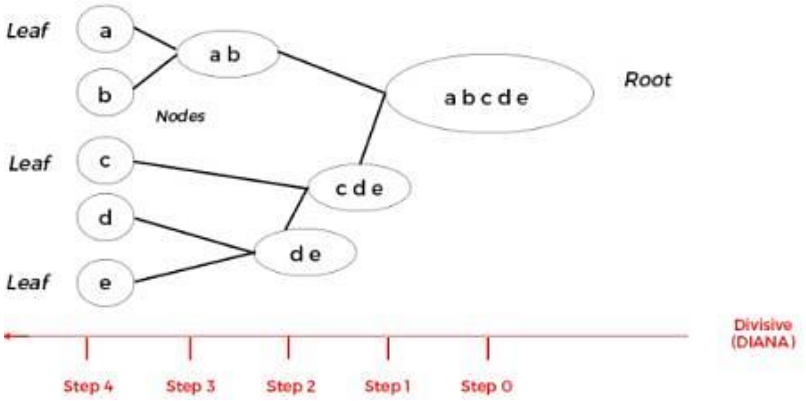
2.1.4.1. Esquema de l'algoritme

Els passos que segueix l'algoritme en el *clustering* divisiu són els següents (veure Figura 2.13):

1. Totes les observacions formen un únic clúster.
2. Es repeteix fins a obtenir un clúster per cada observació. Si es disposa d'N observacions, s'han d'obtenir N clústers:
   a. Per cada clúster es calcula la distància màxima entre parells d'observacions (diàmetre del clúster).

b. Se selecciona el clúster amb major diàmetre.
   i. Es calcula la distància mitjana de cada observació respecte a les altres.
   ii. L'observació més allunyada inicia un nou clúster.
   iii. Es reassignen les observacions entre el clúster nou i l'antic en funció de quin dels dos clústers estan més properes.

Figura 2.13: Dendrograma en horitzontal amb els passos del *clustering* divisiu (*Divisive*), començant amb un únic clúster (dreta) i acabant a la base del dendrograma (esquerra).



Font: *Clustering Jerárquico en R* < RPubs - Clustering Jerárquico en R >

2.1.4.2. Aplicació

En el software estadístic R la funció que aplica el *clustering* divisiu és **diana** del paquet **cluster**. En aquest cas també s'obté un coeficient que indica si la distinció entre clústers és bona, el **coeficient de divisió** (DC).

$$DC = \sum_i \frac{d(i)/d(BD)}{1 - d(i)}$$

On $d(i)$ fa referència al diàmetre de l'últim clúster al qual pertany l'observació i i $d(BD)$ és el diàmetre de tota la base de dades. El valor del coeficient està comprès entre 0 i 1. Valors propers a 1 indiquen que la distinció entre clústers és prou bona per la base de dades que s'analitza, valors propers a 0 indiquen tot el contrari.

Aquest coeficient, com passa amb el coeficient d'aglomeració, s'ha de calcular per una mateixa base de dades amb el mateix nombre d'observacions, ja que a mesura que el nombre d'objectes en la base de dades augmenta, el coeficient també tendeix a augmentar.

2.2. K-means

K-means (*MacQueen, 1967*; *Lloyd, 1982*) és un mètode de *clustering* no probabilístic, de partició i desenvolupat per a variables numèriques, per tant s'aplica només a variables contínues. És l'algoritme no supervisat més utilitzat per la seva poca complexitat i rapidesa. Es tracta d'un algoritme iteratiu que agrupa les observacions en k clústers diferents i troba la millor partició a través d'un òptim local. Abans d'aplicar el mètode, l'analista ha de decidir el nombre de clústers que es volen obtenir.

Com en tots els mètodes no probabilístics, cada observació de la base de dades pertany almenys a un clúster i no existeix solapament entre clústers. Explicat d'una altra manera, si es considera $C_1, \dots, C_k$ com els *sets* de dades que contenen els índexs de les observacions de cada un dels clústers, per exemple el *set* C_1 conté els índexs de les observacions que s'agrupen en el clúster 1, per aquest mètode tots els *sets* satisfan dues propietats:

a) $C_1 \cup C_2 \cup \dots \cup C_k = 1, \dots, n$ ⇒ Cada observació pertany al menys a un dels k clústers.
b) $C_1 \cap C_k = \emptyset$ per a tot $k \neq k'$ ⇒ No existeix solapament entre clústers, és a dir, cap observació pertany a més d'un clúster a la vegada.

2.2.1. Descripció de les mesures de variància

Com s'ha exposat al principi, K-means és un algoritme iteratiu amb l'objectiu de calcular un òptim local per trobar la millor partició possible de les dades. Per poder fer aquest càlcul, es necessita saber que la variància interna (*within-variation*) pel clúster k és una mesura $W(C_k)$ per quantificar la similitud i la dissemblança entre totes les observacions dins d'un clúster, per tant, el mètode K-means pretén minimitzar, el màxim possible, la suma de variàncies internes de tots els clústers:

$$\min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \quad (1)$$

Per poder aplicar l'algoritme és necessari especificar la variància interna $W(C_k)$. Com es tracta d'un mètode enfocat a variables contínues, el càlcul de $W(C_k)$ s'ajuda amb l'ús de distàncies. La mesura de distància més comuna per aquest mètode és la distància Euclidiana al quadrat, per tant $W(C_k)$ es defineix com:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=i}^{s} \left(x_{ij} - x_{i'j}\right)^2 \quad (2)$$

On $|C_k|$ representa el nombre d'observacions del clúster k-èssim, $i' \in C_k$ indica que l'observació i' pertany al clúster k i l'índex s fa referència al nombre total de parells d'observacions en que es calcula la distància pertinent dins de cada un dels k clústers. D'aquesta manera, la variància interna del k-èssim clúster es defineix com la suma de les distàncies euclidianes al quadrat de les observacions dins del clúster, dividit pel nombre total d'observacions del mateix clúster.

Quan s'uneixen les fórmules (1) i (2) s'obté el problema d'optimització que defineix el mètode K-means:

$$\min_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} \left(\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=i}^{s} (x_{ij} - x_{i'j})^2 \right) \right\} \quad (3)$$

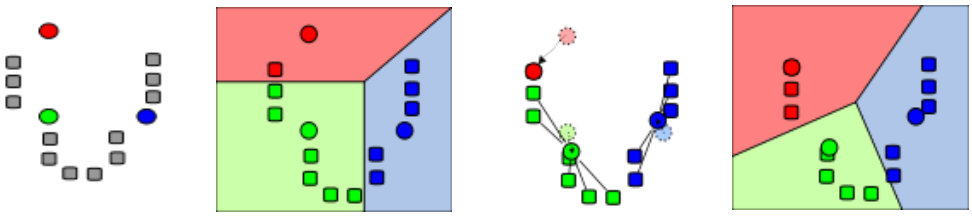
Una vegada definida la fórmula (3) a optimitzar per l'algoritme, s'aprecia que minimitzar la variància interna descrita de manera exacta és difícil, perquè hi ha pràcticament k^n maneres de dividir n observacions en k clústers, donat que, a excepció que el valor de k i n sigui extremadament petit, k^n és un nombre elevat. Per això, l'algoritme que s'aplica obté un òptim local per la suma de variàncies internes.

2.2.2. Esquema de l'algoritme

Tot seguit es presenten els passos que implementa l'algoritme K-means per agrupar les observacions d'un *set* de dades en clústers:

1. S'assigna de manera aleatòria k observacions que prenen el nom de centroides.
2. Es creen k clústers associant cada observació al centroide més proper. Representa l'assignació inicial de les observacions en els diferents clústers.
3. S'iteren els passos descrits a continuació fins que l'assignació dels clústers deixi de canviar, és a dir, s'arribi a un òptim local:
    a. Per a cada un dels k clústers es calcula el seu centroide. Entenent com a centroide, la posició definida per la mitjana de cada una de les dimensions (variables) de les observacions que formen el clúster. Encara que no sempre és equivalent, es pot entendre com el centre de gravetat.
    b. Assignar cada observació al clúster amb el centroide que es troba més pròxim.

Figura 2.14: Exemple gràfic i explicació dels passos que segueix l'algoritme K-means.



(1). Es generen *k* mitjanes aleatòries dins del conjunt de dades (en aquest cas 3 marcades en color).

(2). Es creen *k* clústers associant cada observació amb la mitjana més propera.

(3). El centroide de cada un dels *k* clústers es converteix en la nova mitjana.

(4). Es van repetint el pas (2) i (3) fins que es convergeix a l'òptim local.

Font: *Wikipedia* < https://es.wikipedia.org/wiki/K-medias >

Amb aquesta metodologia K-means garanteix, en cada pas, la reducció de la variància interna total de cada clúster.

Malgrat això, el mètode K-means presenta certes mancances. En primer lloc, és un mètode sensible a *outliers*, ja que es defineix el centre dels clústers a través de centroides, de manera que una dada extrema en un clúster pot fer variar notablement el càlcul del centroide. D'altra banda, K-means imposa una estructura esfèrica als clústers observats encara que els clústers "naturals" de les dades tinguin altres formes geomètriques, finalment, el mètode pot variar depenent de l'assignació inicial aleatòria de centroides, per tant es recomana iniciar el mètode per diferents assignacions aleatòries inicials.

De cara a l'aplicació pràctica del mètode, amb el software estadístic R, s'emprarà la funció **kmeans** del paquet **stats**.

2.3. K-medoids

K-medoids (*Kaufman i Rousseeuw, 1987*) està classificat dins del *clustering* no probabilístic, de partició i desenvolupat per variables numèriques. És molt similar al mètode K-means, ja que els dos són mètodes de partició, on cada observació pertany almenys a un dels k clústers i no existeix solapament, és a dir, cap observació pertany a més d'un clúster a la vegada. També, com en tots els mètodes de partició, abans d'aplicar el mètode, l'analista ha de determinar el nombre de clústers k que es volen obtenir.

La diferència entre els dos algoritmes radica que en el mètode K-medoids cada agrupació està representada per un element present dins del clúster, el que es coneix com a medoid. En concret, el terme medoid es refereix a una observació dins del clúster, en què la distància/dissemblança mitjana entre aquesta i totes les altres observacions del mateix grup és mínima. A més, cada medoid, un per conglomerat, correspon al punt més central de cada clúster i, per aquest motiu, es poden considerar com l'observació més representativa de cada un d'ells. Aquesta diferència entre els mètodes fa que K-medoids sigui una alternativa robusta a K-means, donat que s'utilitza medoids en comptes de centroides per representar els centres dels clústers, llavors el mètode K-medoids es veu menys afectat per *outliers* i soroll que el mètode K-means.

2.3.1. Esquema de l'algoritme

L'algoritme més utilitzat per aplicar el mètode K-medoids es coneix com a **PAM** (*Kaufman i Rousseeuw, 1990*) i les seves inicials fan referència a *Partitioning Around Medoids*.

Tenint en compte, que al començar l'algoritme, els objectes/observacions considerats com a medoids formen part del conjunt S d'objectes seleccionats i, considerant-se O com el conjunt de tots els objectes, $U = O - S$ representa el conjunt d'objectes no seleccionats. L'objectiu de l'algoritme PAM és minimitzar la diferència/dissemblança mitjana dels objectes no seleccionats amb els seus medoids més propers, és a dir, es tracta de minimitzar la suma de dissemblances entre totes les observacions del conjunt d'objectes U amb les observacions

més propers del conjunt d'objectes seleccionats S, els medoids. L'algoritme consta de dues fases:

i) En la primera fase, ***BUILD***, es selecciona una col·lecció de k objectes per inicialitzar el conjunt S.
ii) En la segona fase, ***SWAP***, es prova de millorar la qualitat del *clustering* intercanviant objectes seleccionats del conjunt S, amb objectes no seleccionats del conjunt U.

Per cada objecte p es guarden dos valors:

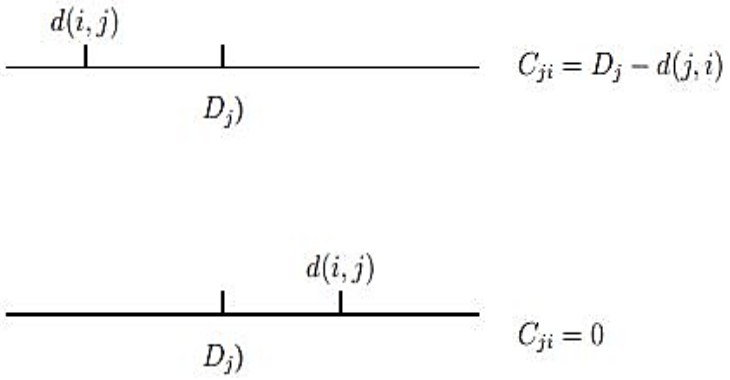
- D_p, la distància entre p i l'objecte més proper del conjunt S.
- E_p, la distància entre p i el segon objecte més proper del conjunt S.

A mesura que els conjunts S i U canvien, s'ha d'actualitzar el valor de D_p i E_p. Tenint en compte aquesta informació, es pot apreciar com $D_p \leq E_p$ i que l'objecte/observació $p \in S$ si i només sí $D_p = 0$.

La fase de construcció, *BUILD*, està formada pels següents passos:

1. S'inicia el conjunt S afegint un objecte, dins de les dades, que tingui la mínima suma de dissemblances respecte a tots els altres objectes.
2. Es considera un objecte $i \in U$ com a candidat per incloure'l dins del conjunt d'objectes seleccionats S.
3. Per un objecte $j \in U - \{i\}$ es calcula D_j, la dissemblança entre j i l'objecte més proper dins del conjunt S.
4. Si $D_j > d(i,j)$, la distància entre els objectes i i j, l'objecte j contribuirà en la decisió per seleccionar l'objecte i, atès que la qualitat del *clustering* podria millorar. Finalment, per aquest pas, es calcula la contribució $C_{ji} = max\{D_j - d(i,j), 0\}$.

Figura 2.15: Representació en el pla del càlcul de la contribució C_{ji}



Font: *The PAM Clustering Algorithm* < pam1.dvi (umb.edu) >

5. Un cop es disposa de totes les contribucions C_{ji} es calcula el guany (g_i) que s'obté en afegir l'objecte i dins del conjunt S com:

$$g_i = \sum_{j \in U} C_{ji}$$

6. Per concloure el *BUILD*, es selecciona l'objecte i que maximitza el guany g_i, com a conseqüència el conjunt S i U canvien: $S := S \cup \{i\}$ i $U := U - \{i\}$.
7. Aquests passos es repeteixen fins que es seleccionen k objectes per cada un dels k clústers.

Una vegada acabada la fase de construcció, entra en joc la fase d'intercanvi, *SWAP*, que pretén millorar el conjunt d'objectes seleccionats S i, a la vegada, millorar la qualitat del *clustering*. Per poder fer-ho cal considerar tots els parells d'objectes $(i, h) \in S \times U$, d'aquesta manera l'algoritme PAM calcula l'efecte T_{ih} sobre la suma de dissemblances entre els objectes U i els objectes S més propers a causa de l'intercanvi de les observacions i i h, en altres paraules, T_{ih} és l'efecte causat a l'hora de transferir l'objecte i del conjunt S al conjunt U i, a la vegada, transferir l'objecte h del conjunt U al conjunt S. És molt important el càlcul d'aquest efecte, ja que el valor de T_{ih} dictamina si l'algoritme ha de seguir o parar d'iterar.

El càlcul de l'efecte T_{ih} necessita, prèviament, calcular la contribució K_{jih} de cada una dels objectes $j \in U - \{h\}$ durant l'intercanvi dels parells d'objectes i i h. Com es veurà en el primer pas d'aquesta segona fase, només pot passar que la distància entre j i i sigui major o igual a D_j, la dissemblança entre j i l'objecte més proper dins del conjunt S:

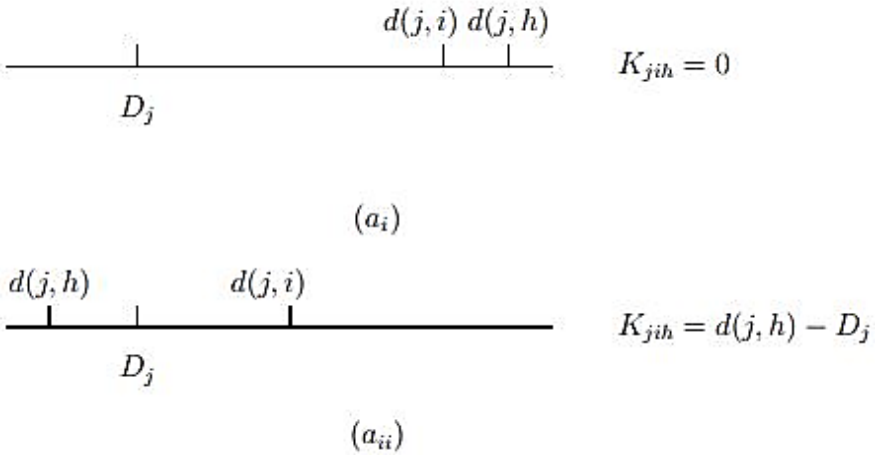
$d(j,i) > D_j$ o $d(j,i) = D_j$

La fase d'intercanvi, *SWAP*, consta dels següents passos:

1. Es calcula la contribució K_{jih} tenint en compte tot els casos possibles:
   a. Si $d(j,i) > D_j$, llavors hi ha dues opcions:
      i. Si $d(j,h) \geq D_j$, llavors $K_{jih} = 0$.
      ii. Si $d(j,h) < D_j$, llavors $K_{jih} = d(j,h) - D_j$.

      En els dos casos $K_{jih} = min\{d(j,h) - D_j, 0\}$.

Figura 2.16: Representació en el pla del càlcul de la contribució K_{jih} en el cas que $d(j,i) > D_j$



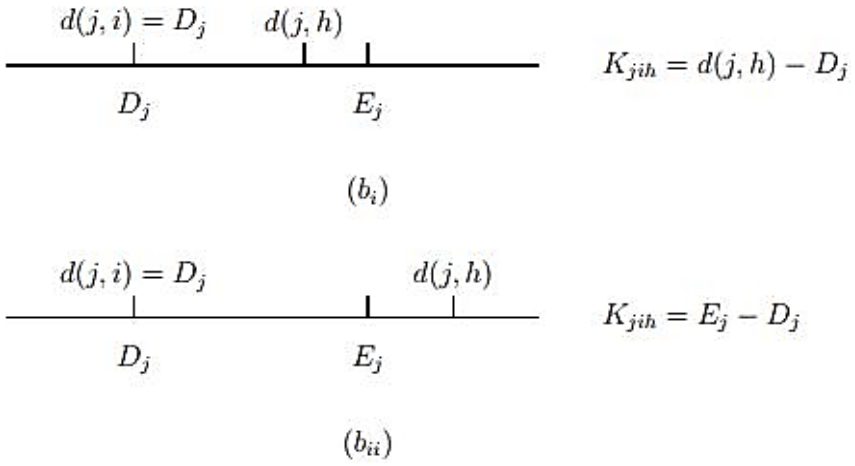
Font: *The PAM Clustering Algorithm* < pam1.dvi (umb.edu)

b. Si $d(j,i) = D_j$, també hi ha dues opcions:

i. Si $d(j,h) < E_j$, on E_j representa la distància entre j i el segon objecte més proper del conjunt S, llavors $K_{jih} = d(j,h) - D_j$. En aquest cas K_{jih} pot prendre valor positiu o negatiu.
ii. Si $d(j,h) \geq E_j$, llavors $K_{jih} = E_j - D_j$. En aquest cas $K_{jih} > 0$.

En els dos casos, $K_{jih} = min\{d(j,h), E_j\} - D_j$

Figura 2.17: Representació en el pla del càlcul de la contribució K_{jih} en el cas que $d(j,i) = D_j$



Font: *The PAM Clustering Algorithm* < pam1.dvi (umb.edu) >

2. Calcular l'efecte total de l'intercanvi com:

$$T_{ih} = \Sigma\{K_{jih} \mid j \in U\}$$

3. Es selecciona el parell $(i,h) \in S \times U$ que minimitzi T_{ih}.
4. Si $T_{ih} < 0$, es segueix amb els intercanvis, cal actualitzar el valor de D_p i E_p per a cada objecte p i es retorna al primer pas d'aquesta fase.

   Si el mínim de $T_{ih} > 0$, el valor objectiu de l'algoritme no pot decréixer més, per tant, es para d'iterar. Això passa quan tots els efectes T_{ih} són positius, que de fet, és la consigna de parada de l'algoritme.

Així doncs, aquests són els passos que segueix l'algoritme PAM a l'hora de dividir les dades en *k* conglomerats.

Durant els passos de l'algoritme s'ha exposat la necessitat de calcular les dissemblances/distàncies $d(i,j)$ entre els objectes del conjunt U i el conjunt S. Com que l'estudi de conglomerats se centra en variables de tipus continu, PAM pot utilitzar dues mètriques diferents per calcular les dissemblances, una és la distància Euclidiana i l'altre és la distància de Manhattan. En aquest estudi es farà el càlcul de les dissemblances entre objectes a través de la distància Euclidiana.

De cara a l'aplicació pràctica del mètode amb el software estadístic R s'emprarà la funció **pam** del paquet **cluster**.

2.4. Gaussian Mixture Models

Fins ara, totes les tècniques que s'han presentat estan classificades dins dels mètodes de *Hard clustering*, és a dir, el *clustering* no probabilístic, basats en mesures de distància per definir la similitud i la dissemblança entre clústers i on totes les observacions formen part d'un únic clúster.

Els *Gaussian Mixture Models* (*Banfield i Raftery, 1993*) estan classificats dins dels mètodes de clustering basats en models estadístics (*model-based clustering*) per a variables numèriques **Gaussianes**. Els *model-based clustering* són mètodes de *clustering* probabilístic, on es considera que les dades provenen d'una distribució de probabilitat que és una barreja de dos o més clústers (*Fraley i Raftery, 2002*).

Per aquest tipus de *clustering* les observacions tenen associada una probabilitat de pertànyer a diferents clústers, per tant, a diferència del *clustering* numèric, una observació pot estar associada a més d'un clúster a la vegada. Aquesta propietat del *clustering* probabilístic pot ser molt útil, donat que permet identificar observacions amb alt o baix nivell de pertinença en cada clúster i, potencialment, classificar-les de manera única o proporcionar solucions alternatives per a les observacions amb baix nivell de pertinença.

2.4.1. Finite mixture models

En el *model-based clustering* es considera que les dades proven d'un *finite mixture model* (model de barreja finit). Si es disposa de $x = \{x_1, x_2, \dots, x_i, \dots, x_n\}$ com a una mostra d'n observacions independents idènticament distribuïdes, la distribució de cada observació s'especifica mitjançant una funció de densitat a través d'un *finite mixture model* amb G components i pren la següent forma:

$$f(x_i; \Psi) = \sum_{g=1}^{G} \pi_g f_g(x_i; \theta_g)$$

on $\Psi = \{\pi_1, \dots, \pi_{G-1}, \theta_1, \dots, \theta_G\}$ són els paràmetres del *finite mixture model*, $f_g(x_i; \theta_g)$ és la densitat de barreja finita de la g-èssima component per l'observació x_i amb un vector de paràmetres θ_g, $(\pi_1, \dots, \pi_{G-1})$ són les probabilitats de barreja i G és el nombre de components de barreja.

Per tal que $f(x_i; \Psi)$ sigui una *finite mixture model* s'ha de complir:

- $\pi_g > 0$ i $\sum_{g=1}^{G} \pi_g = 1$, les probabilitats de barreja són majors a 0 i la suma de totes elles ha de donar 1.
- $f_g(x_i; \theta_g) \geq 0$ i $\int f_g(x_i; \theta_g)\, dx = 1$ per $g = 1, \dots, G$.

Si s'assumeix que G és un valor fix, els paràmetres del model de barreja Ψ normalment són desconeguts i s'han d'estimar. Per fer-ho s'ha de calcular la log-versemblança de $f(x_i;\Psi)$ que s'expressa com: $l(\Psi; x_1, \dots, x_n) = \sum_{i=1}^{n} log(\, f(x_i;\Psi))$. Intentar maximitzar la funció de la log-versemblança directament pot esdevenir complicat i, per aquest motiu, el càlcul de l'estimador de màxima versemblança d'un model de barreja finit normalment s'obté a través de l'algoritme *Expectation-Maximization* (*EM algorithm*), que es veurà en el subapartat 2.4.3.

Dins dels *model-based clustering*, cada component G d'una densitat de barreja finita queda associada a un clúster. Un dels models més populars dins del *model-based clustering* és els *Gaussian Mixture Models* (GMMs; *Banfield i Raftery, 1993*) que assumeix una distribució Gaussiana (multivariant) per cada component/clúster:

$$f_g(x_i;\theta_g) \sim N(\mu_g, \Sigma_g) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \cdot exp\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)$$

Per aquest mètode, els clústers són el·lipsoidals i centrats al vector de mitjanes μ_g, tanmateix, la matriu de variàncies i covariàncies Σ_g determina l'estructura geomètrica de la distribució i permet calcular diferents models de barreja Gaussiana que, a posteriori, es comparen per decidir el millor model possible per la base de dades analitzada.

2.4.2. Tipus de covariàncies

En els GMMs, la matriu de variàncies i covariàncies Σ_g determina la disposició geomètrica dels conglomerats i dóna informació sobre el volum, la forma i l'orientació dels clústers. Es pot obtenir una parametrització parsimoniosa de Σ_g per mitjà de la descomposició dels valors propis de la matriu amb la forma $\Sigma_g = \lambda_g D_g A_g D_g$, on λ_g representa un escalar que controla el volum de les el·lipses, A_g és una matriu diagonal que especifica la forma i els contorns de densitat dels clústers amb $det(A_g) = 1$, i D_g és una matriu ortogonal que determina l'orientació de les el·lipses/clústers (*Celeux i Govaert, 1995*).

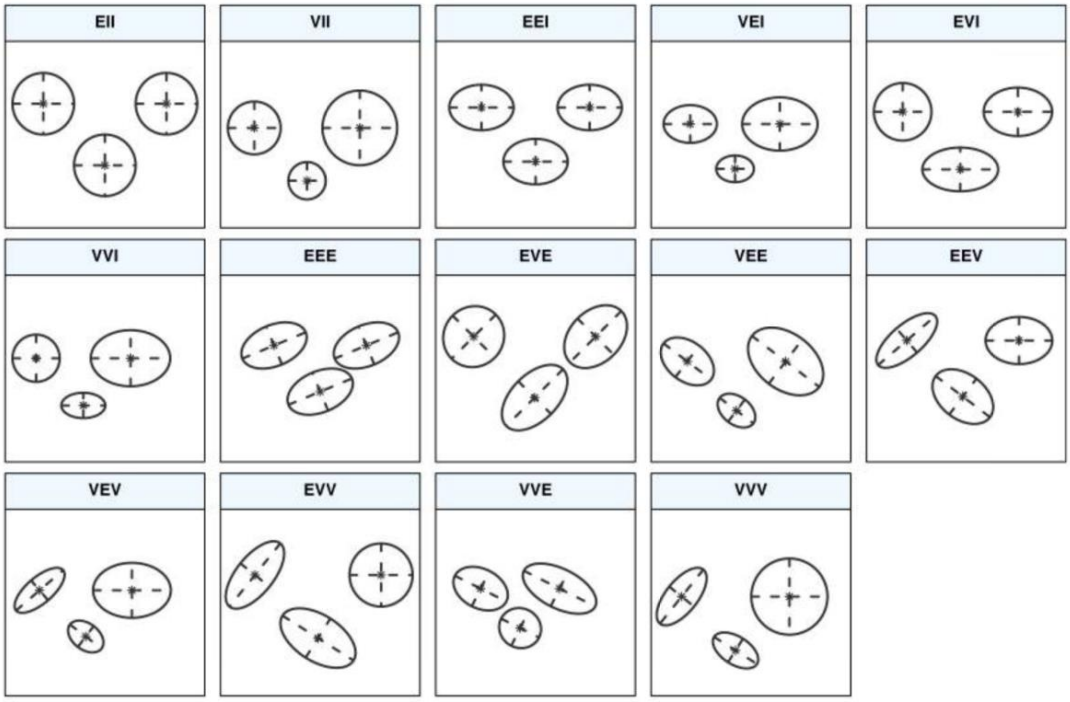
En una dimensió només existeixen dos models possibles, el model E per variàncies iguals i el model V per variàncies diferents. En un entorn multivariant, el volum, la forma i l'orientació de les covariàncies es poden limitar perquè siguin iguals o variables entre grups, llavors es poden especificar 14 models amb característiques geomètriques diferents. A la Taula 2.4 s'especifiquen tots els 14 models d'acord amb la seva matriu de variàncies i covariàncies per a un entorn multivariant. La Figura 2.18 representa de manera gràfica els 14 models possibles del GMMs en dues dimensions.

Taula 2.4: Parametrització de la matriu de variàncies i covariàncies Σ_g per un entorn multidimensional, amb el nom dels 14 possibles models i la seva disposició geomètrica.

Model	Σ_g	Distribució	Volum	Forma	Orientació
EII	λI	Esfèrica	Igual	Igual	-
VII	$\lambda_g I$	Esfèrica	Variable	Igual	-
EEI	λA	Diagonal	Igual	Igual	Eixos de coordenades
VEI	$\lambda_g A$	Diagonal	Variable	Igual	Eixos de coordenades
EVI	λA_g	Diagonal	Igual	Variable	Eixos de coordenades
VVI	$\lambda_g A_g$	Diagonal	Variable	Variable	Eixos de coordenades
EEE	λDAD^T	El·lipsoidal	Igual	Igual	Igual
EVE	$\lambda DA_g D^T$	El·lipsoidal	Igual	Variable	Igual
VEE	$\lambda_g DAD^T$	El·lipsoidal	Variable	Igual	Igual
VVE	$\lambda_g DA_g D^T$	El·lipsoidal	Variable	Variable	Igual
EEV	$\lambda D_g AD_g^T$	El·lipsoidal	Igual	Igual	Variable
VEV	$\lambda_g D_g AD_g^T$	El·lipsoidal	Variable	Igual	Variable
EVV	$\lambda D_g A_g D_g^T$	El·lipsoidal	Igual	Variable	Variable
VVV	$\lambda_g D_g A_g D_g^T$	El·lipsoidal	Variable	Variable	Variable

Font: Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal, 8(1), 289.

Figura 2.18: El·lipses de densitat constant per cada un dels 14 models obtinguts de la Taula 2.4 en el cas de tenir tres grups en dues dimensions.



Font: Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal, 8(1), 289.

2.4.3. Selecció del model

Una qüestió central en els *finite mixture models* és arribar a saber quants components/clústers G s'han d'incloure en la barreja. En els GMMs, a més, es necessita decidir quina parametrització de la matriu de variàncies i covariàncies Σ_g s'ha d'adoptar. Ambdues qüestions es poden resoldre per mitjà d'un criteri d'informació, com per exemple el BIC (*Schwartz, 1978; Fraley i Raftery, 1998*), criteri que s'aplica per defecte en els GMMs per seleccionar el model correcte, o l'*integrated complete-data likelihood criterion* (ICL; *Biernacki et. al, 2000*)

Aquest tipus de criteris es basen en penalitzacions de la funció de la log-versemblança. Mentre que la versemblança incrementa amb l'addició de més components, es resta la log-versemblança amb un terme que penalitza el nombre de paràmetres a estimar. El criteri del BIC és el més utilitzat pels GMMs i pren la següent forma:

$$BIC_{M,G} = 2l_{M,G}\left(x\middle|\widehat{\Psi}\right) - vlog(n)$$

on $l_{M,G}\left(x\middle|\widehat{\Psi}\right)$ és la funció de la log-versemblança per l'estimador de màxima versemblança $\widehat{\Psi}$ del model M amb G components, n és la mida de la mostra i v és el nombre estimat de paràmetres. Se selecciona el parell $\{M, G\}$ que maximitza el criteri del BIC.

A diferència dels models de *Hard clustering*, en què l'analista ha d'aplicar algun mètode de selecció de clústers òptims, els GMMs directament seleccionen el model que millor s'ajusta a les dades i el nombre de component/clústers òptims que s'han d'escollir, per tant, no cal recórrer a mètodes com l'*elbow method*, l'*average silhouette method* o el *gap statistic* per escollir el nombre de clústers que s'exposen al capítol III d'aquest treball.

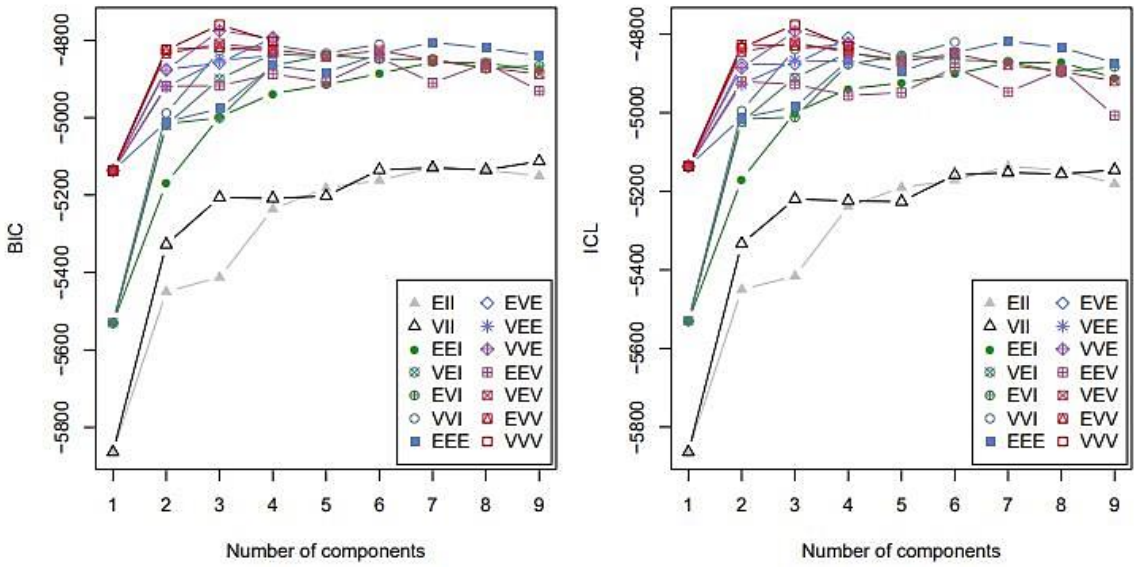
Per altra banda, el BIC tendeix a seleccionar el nombre de components de barreja que es necessiten per aproximar la densitat del model de manera raonable en comptes de seleccionar el nombre de clústers com a tal, mentre que l'ICL penalitza el BIC a través d'un terme d'entropia que mesura el solapament dels clústers. La forma de l'ICL és la següent:

$$ICL_{M,G} = BIC_{M,G} + 2\sum_{i=1}^{n}\sum_{g=1}^{G} c_{ig}\, log\left(z_{ig}\right)$$

on z_{ig} és la probabilitat condicionada referent a què x_i prové de la g-èssima component de barreja, quant a c_{ig} el seu valor assigna la pertinença de la i-èssima observació en el clúster/component g, si $c_{ig} = 1$ la i-èssima observació està assignada al clúster g, quan $c_{ig} = 0$ passa tot el contrari.

Altres criteris de selecció com l'AIC també es poden utilitzar per seleccionar el model i el nombre de clústers òptim pels models de barreja finits, tot i això per l'aplicació dels GMMs s'aplicarà el criteri del BIC, el que per defecte aplica el mètode a R.

Figura 2.19: Gràfics del BIC i l'ICL per seleccionar el millor model sobre una mateixa base de dades.



Font: Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal, 8(1), 289.

2.4.4. Esquema de l'algoritme

Per poder arribar a seleccionar el millor model dels GMMs, primer cal estimar els paràmetres del model de barreja Ψ a través de la funció de la log-versemblança de $f(x_i; \Psi)$.

Per a un model de barreja finit, la funció de versemblança pren la següent forma:

$$L(\Psi; x_1, \dots, x_n) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(x_i; \theta_g)$$

i la funció de la log-versemblança és:

$$l(\Psi; x_1, \dots, x_n) = lnL(\Psi; x_1, \dots, x_n) = \sum_{i=1}^{N} ln\left(\sum_{g=1}^{G} \pi_g f_g(x_i; \theta_g)\right)$$

Com que intentar maximitzar la funció de la log-versemblança directament pot esdevenir complicat, per procedir amb l'estimació dels paràmetres, s'utilitza l'algoritme EM (*Dempster et al., 1977*; *McLachlan i Peel, 2000*) que és un mètode iteratiu que permet calcular l'estimador de màxima versemblança d'un *finite mixture model*. Per tal de formular l'algoritme, es distingeixen les dades entre dades observades i dades mancants. Les dades observades són totes les observacions de la base de dades, mentre que les dades mancants són les assignacions dels components/clústers del model sobre les observacions: $z_g(x_i) \in \{0,1\}$.

Per cada iteració, l'algoritme EM presenta dos passos, el pas E (*Expectation*) i el pas M (*Maximization*):

a. **Pas E:** es calcula el valor esperat de les dades mancants respecte als paràmetres del model estimat, per a un model de barreja aquest càlcul pren el nom de probabilitat de propietat (*ownership probability*)

$$E\left[z_g(x_i)\right] = q_g(x_i)$$

b. **Pas M:** l'algoritme calcula l'estimador de màxima versemblança per cada paràmetre del model de barreja en funció de les dades observades i el valor esperat de les dades mancants, pels models de barreja això produeix un problema de regressió ponderada per a cada component del model:

$$\sum_{i=1}^{N} q_g(x_i) \frac{\partial}{\partial \theta_g} \, log \, f_g\left(x_i; \theta_g\right) = 0$$

on les probabilitats de barreja π_g són:

$$\pi_g = \frac{1}{N} \sum_{i=1}^{N} q_g(x_i)$$

I l'objectiu en aquest pas és trobar l'extrem de la funció log-versemblança en què les probabilitats de barreja π_g sumin 1, $\sum_g \pi_g = 1$.

Una de les propietats de l'algoritme EM és que a cada iteració, s'incrementa la versemblança de les dades observades donats els paràmetres dels models i s'actualitzen les estimacions d'aquests paràmetres fins que es detecta certa convergència en els resultats sota un criteri determinat per l'analista.

De cara a l'aplicació amb R s'emprarà la funció **Mclust** del paquet **mclust** (*Scrucca et. al, 2016*).

III. SELECCIÓ ÒPTIMA DE CLÚSTERS

Determinar el nombre òptim de clústers en una base de dades és un problema fonamental per aplicar tant als mètodes de *clustering* de partició com als mètodes de *clustering* Jeràrquic. Tot i disposar d'informació prèvia sobre les dades, no existeix una única resposta per aquest problema, ja que la quantitat òptima de clústers és, en certa manera, subjectiva i depèn del mètode de *clustering* utilitzat, donat que les mesures de similitud i els paràmetres aplicats per dividir les dades en clústers són diferents.

Excepte els *Gaussian Mixture Models* que, a través de l'algoritme EM i el criteri BIC o ICL, es troba el model amb el nombre de clústers òptim, per la resta d'algoritmes que s'apliquen en aquest estudi cal determinar prèviament, pels mètodes de partició, o posteriorment, pels mètodes jeràrquics, el nombre de clústers òptims per cada tècnica de *clustering*.

A continuació es presentaran tres mètodes per decidir la selecció òptima de clústers, en concret es tracta de l'***elbow method***, l'***average silhouette method*** i el ***gap statistic***.

3.1. Elbow Method

L'*elbow method* (*Thorndike, 1953*) és un mètode directe per trobar el nombre de clústers òptims per a una base de dades. El mètode pretén optimitzar la suma de variàncies internes de tots els clústers, també anomenat la suma total de quadrats dins dels clústers (*total within-cluster sum of square*), per tal de minimitzar el seu valor. L'*elbow method* contempla la suma total de variàncies internes dels clústers com una funció sobre el nombre de clústers:

$$Total\ withiness = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left(x_i - \underline{x}_k\right)^2$$

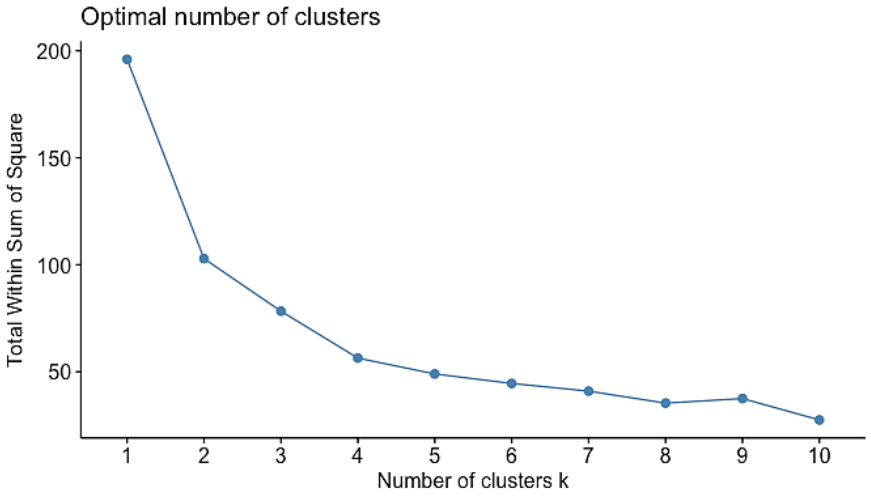
El valor de $\underline{x}_k$ es refereix al centroide del clúster k, $x_i \in C_k$ vol dir que la dada x_i pertanyen al *set* de dades C_k del clúster k.

Els passos que es segueix l'*elbow method* són els següents:

1. Calcular l'algoritme de *clustering* (K-means, K-medoids, Jeràrquic) per diferent nombre de clústers k. El valor de k varia de 1 a 10 clústers.
2. Per cada valor de k es calcula la suma de variàncies internes de tots els clústers.
3. Es fa el gràfic de la suma de variàncies internes de tots els clústers per cada valor de k (veure Figura 3.1).
4. On comença a corbar-se la línia del gràfic es considera, generalment, com a indicador del nombre adequat de clústers.

L'objectiu del mètode és escollir el nombre de clústers òptim de manera que quan s'afegeix un clúster nou, aquesta addició, no millori molt la funció a optimitzar, és a dir, que la suma de variàncies internes de tots els clústers no disminueixi considerablement del mètode aplicat sobre $k + 1$ clústers respecte al mètode aplicat sobre k clústers.

Figura 3.1: Gràfic exemple de l'*elbow method* per trobar el nombre òptim de clústers.



Font: <https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/>.

Es pot apreciar en la Figura 3.1 que, a vegades, l'*elbow method* pot arribar a ser ambigu, atès que pot ser difícil definir en quin punt la línia que uneix la suma total de variàncies internes dels clústers comença a corbar-se destacadament. És per aquest motiu que és recomanable basar-se en més d'un mètode de selecció de clústers.

Per aplicar l'*elbow method* es farà ús de la funció **fviz_nbclust** del paquet **factoextra** (*Kassambara i Mundt, 2020*) del software estadístic R. Com a afegit, s'aplicarà la funció **NbClust** del paquet **NbClust** (*Charrad et. al, 2014*), atès que permet calcular 30 mètodes de selecció de clústers diferents de manera simultània, per tant, aquesta funció permetrà corroborar els resultats del mètode a l'hora de decidir quin nombre de clústers és més òptim per a la base de dades que s'estudia depenent de la tècnica de *clustering* que s'apliqui.

3.2. Average Silhouette Method

L'*average silhouette method* (*Kaufman i Rousseeuw, 1990*) és un mètode directe per trobar el nombre de clústers òptims per a una base de dades. Aquest mètode determina la qualitat de classificació de cada observació dins del seu clúster, és a dir, la qualitat del *clustering*, i estima la distància mitjana entre clústers, l'*average silhouete*.

Per cada observació i, l'*average silhouete*, que mesura la proximitat de cada punt d'un clúster respecte als punts del clúster veí més proper, es calcula de la següent manera:

1. Per cada observació i es calcula la distància mitjana a_i entre l'observació i i la resta d'observacions del mateix clúster.
2. Per tots els clústers k que no pertanyen a l'observació i, es calcula la distància mitjana $d(i,k)$ entre l'observació i i totes les observacions del clúster k. La distància més petita es defineix com:

$$b_i = min_k d(i,k)$$

El valor de b_i representa la mínima distància entre l'observació i i les observacions del clúster veí més proper.

3. Finalment, l'*average silhouete* per l'observació i es defineix com:

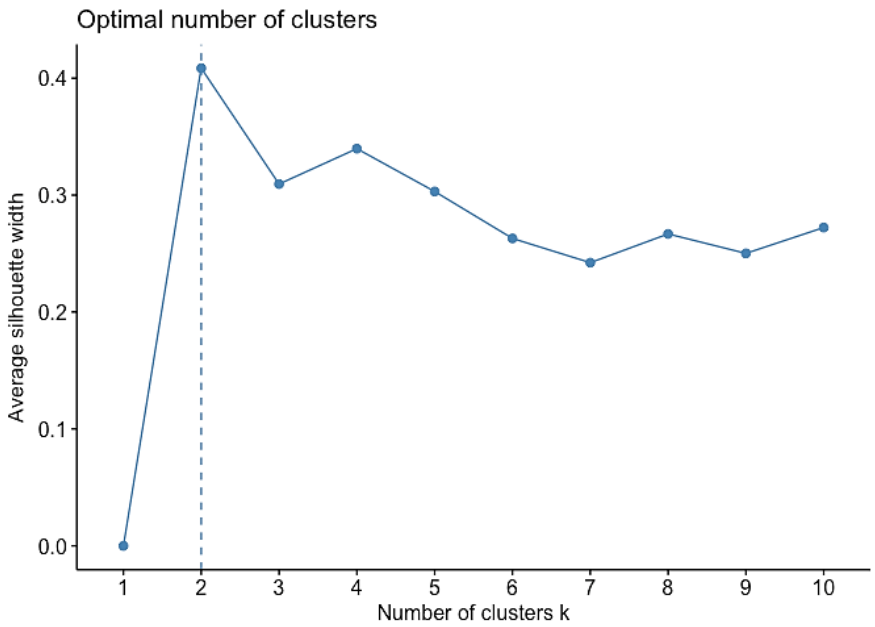
$$S_i = (b_i - a_i)/max(a_i, b_i)$$

Les observacions amb un valor de S_i proper a 1 es consideren ben classificades, quan el valor de S_i és proper a 0 significa que l'observació i es troba entre dos clústers i quan el valor de S_i és negatiu indica que l'observació i està classificada al clúster erroni.

El nombre òptim de clústers k és aquell que maximitza l'*average silhouette* (S) respecte a diferents valors de k. L'algoritme que segueix l'*average silhouette method* és similar a l'*elbow method* i els seus passos es mostren a continuació:

1. Calcular l'algoritme de *clustering* (K-means, K-medoids, Jeràrquic) per diferents nombres de clústers k. El valor de k varia de 1 a 10 clústers.
2. Per a cada valor de k es calcula l'*average silhouette* (S) de les observacions.
3. Es fa el gràfic dels valors d'S respecte al nombre de clústers k (veure Figura 3.2).
4. La posició del valor màxim d'S es considera el nombre de clústers adients.

Figura 3.2: Gràfic exemple de l'*average silhouette method* per trobar el nombre òptim de clústers.



Font: <https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/>.

Per aplicar l'*average solhouette method* es farà ús de la funció **fviz_nbclust** del paquet **factoextra** del software estadístic R. Com a afegit, també s'aplicarà la funció **NbClust** del paquet **NbClust**, per corroborar els resultats del mètode a l'hora de decidir quin nombre de clústers és més òptim per a la base de dades que s'estudia depenent de la tècnica de *clustering* que s'apliqui.

3.3. Gap Statistic

El *gap statistic* (*Tibshirani et. al, 2001*), a diferència dels dos mètodes directes anteriors on l'objectiu és optimitzar un criteri, es basa en un test estadístic per seleccionar el nombre de clústers òptims. El mètode compara la variància interna total per diferent nombre de clústers k amb els valors esperats de la variància interna total de les dades sota una distribució de referència. En altres paraules, compara els valors obtinguts amb els valors esperats de la hipòtesi nul·la del test estadístic.

Els passos que segueix el *gap statistic* són els següents:

1. Aplicar l'algoritme de *clustering* (K-means, K-medoids, Jeràrquic) variant el nombre de clústers k i calcular la variància interna total corresponent W_k. $k = 1, \dots, k_{max}$ on k_{max} representa el nombre màxim de clústers.
2. Generar B *datasets* de referència amb una distribució uniforme aleatòria. S'aplica a cada un dels *datasets* de referència l'algoritme de *clustering* variant el nombre de clústers k i es calcula la variància interna total corresponent W_{kb}. $k = 1, \dots, k_{max}$ on k_{max} representa el nombre màxim de clústers.
3. Es calcula l'estimació del *gap statistic* com la desviació de la variància interna total observada W_k respecte del seu valor esperat W_{kb} sota la hipòtesi nul·la:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B}\big(log(W_{kb}) - log(W_k)\big)$$

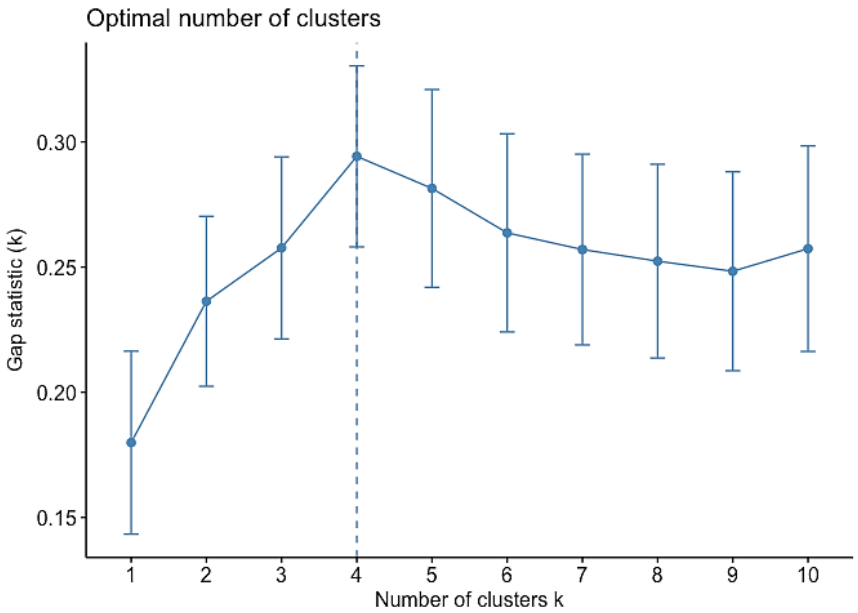
   Es calcula la desviació estàndard del *gap statistic* s_k a través de $sd(k) = \sqrt{Gap(k)}$ definint $s_k = sd(k) \cdot \sqrt{1 + 1/B}$.

4. S'escull el valor mínim de k com al nombre de clústers òptim quan es compleix:

$$Gap(k) \geq Gap(k+1) + s_{k+1}$$

Com es pot apreciar en els passos de l'algoritme, el *gap statistic* fa ús de la metodologia Bootstrap per calcular el $Gap(k)$, atès que es generen B remostres aleatòries de distribució uniforme per la base de dades, calculant la variància interna total nul·la W_{kb} per comparar-la amb la variància interna total calculada sobre la base de dades original W_k en cada remostra. De manera que es divideix el total de desviacions entre W_{kb} i W_k pel nombre total de remostres B obtenint així el valor de l'estadístic. A la Figura 3.3 es pot veure el gràfic que extreu R en aplicar aquest mètode de selecció de clústers òptim.

Figura 3.3: Gràfic exemple del *gap statistic* per trobar el nombre òptim de clústers.



Font: <https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/>.

Per aplicar el *gap statistic* es farà ús de la funció **fviz_nbclust** del paquet **factoextra** del software estadístic R. Com a afegit, també s'aplicarà la funció **NbClust** del paquet **NbClust**, per corroborar els resultats del mètode a l'hora de decidir quin nombre de clústers és més òptim per a la base de dades que s'estudia depenent de la tècnica de *clustering* que s'apliqui.

IV. MÈTODE PER COMPARAR TÈCNIQUES DE CLUSTERING

Un cop aplicats tots els mètodes de *clustering* que s'han presentat al capítol II d'aquest treball, caldrà comparar i avaluar els resultats entre les diferents tècniques per arribar a discernir quin mètode s'adapta millor a la base de dades estudiada.

Aquesta comparació es divideix en dues parts, per una banda, agafant tot el conjunt de variables de la base de dades, a través de taules de contingència, per a les variables categòriques, i estadístics descriptius com poden ser la mitjana i la mediana, per a les variables numèriques, es procedirà a especificar el perfil dels clústers, de manera que per tots els mètodes de *clustering*, a cada clúster se li atribuirà un nom amb la descripció de les seves característiques. Un cop definit el perfil de cada clúster, es compararà entre els diferents mètodes la caracterització dels clústers, per poder veure si els grups que es creen en cada mètode comparteixen les mateixes característiques o proporcionen informació complementaria.

D'altra banda, es compararà l'estructura dels clústers proporcionats per a cada mètode, en altres paraules, s'estudiarà si els diferents mètodes de *clustering* situen les observacions de manera similar a l'hora de dividir la base de dades en clústers. Per aquesta part s'utilitzarà l'Índex de Rand Ajustat (ARI; *Hubert i Arabie, 1985*) que es presenta a l'apartat 4.1.

4.1. Índex de Rand Ajustat (ARI)

Existeixen diversos índexs de rendiment per a avaluar els resultats del *clustering*. Els índexs són mesures de correspondència entre dues particions aplicades a les mateixes dades, i es calculen per mitjà de com es classifiquen els parells d'observacions entre les dues particions diferents en una taula de contingències (veure Taula 4.1). Per veure com es desenvolupa l'Índex de Rand Ajustat (ARI) és necessari introduir certa notació prèvia.

Si es considera un conjunt d' n objectes $S = \{O_1, O_2, \dots, O_n\}$ i es suposa que $U = \{u_1, u_2, \dots, u_R\}$ i $V = \{v_1, v_2, \dots, v_C\}$ representen dues particions diferents dels objectes S, de tal manera que $\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$ i $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ per $1 \leq i \neq i' \leq R$ i $1 \leq j \neq j' \leq C$. Tenint en compte les dues particions U i V amb R i C subconjunts respectivament, es pot crear una taula de contingències per detectar si hi ha solapament entre les dues particions com ve representat a la Taula 4.1.

Taula 4.1: Taula de contingències per comparar la partició U amb la partició V.

Partició		V				
	Grup	v_1	v_2	$\cdots$	v_C	Total
U	u_1	t_{11}	t_{12}	$\cdots$	t_{1C}	$t_{1.}$
	u_2	t_{21}	t_{22}	$\cdots$	t_{2C}	$t_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	u_R	t_{R1}	t_{R2}	$\cdots$	t_{RC}	$t_{R.}$
Total		$t_{.1}$	$t_{.2}$	$\cdots$	$t_{.C}$	$t_{..} = n$

Per la Taula 4.1, el valor genèric t_{rc} representa el nombre d'observacions que estan classificades en el subconjunt $r \in R$ de la partició U i en el subconjunt $c \in C$ de la partició V. Tenint en compte el nombre total de possibles combinacions de parells d'observacions entre particions $\binom{n}{2}$ per a una mateixa base de dades, es pot classificar els resultats de la Taula 4.1 en quatre tipus de parells d'observacions possibles:

a. Parells d'observacions localitzades en el mateix grup en U i en el mateix grup en V.
b. Parells d'observacions localitzades en el mateix grup en U i en diferents grups en V.
c. Parells d'observacions localitzades en diferents grups en U i en el mateix grup en V.
d. Parells d'observacions localitzades en diferents grups en U i en diferents grups en V.

Aquesta classificació permet simplificar els resultats de la Taula 4.1 per passar a tenir una taula de contingències de només dues files i dues columnes (veure Taula 4.2).

Taula 4.2: Taula de contingències simplificada 2x2 per comparar la partició U amb la partició V a través dels grups a, b, c i d.

Partició	V	
U	Parells en el mateix grup	Parells en grups diferents
Parells en el mateix grup	*a*	*b*
Parells en grups diferents	*c*	*d*

Utilitzant els resultats de la Taula 4.1 es poden calcular els valors de *a*, *b*, *c* i *d* de la Taula 4.2 de la següent manera:

$$a = \sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2} = \left(\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 - n\right)/2$$

$$b = \sum_{r=1}^{R}\binom{t_{r.}}{2} - a = \left(\sum_{r=1}^{R} t_{r.}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2\right)/2$$

$$c = \sum_{c=1}^{C}\binom{t_{.c}}{2} - a = \left(\sum_{c=1}^{C} t_{.c}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2\right)/2$$

$$b = \binom{n}{2} - a - b - c = \binom{n}{2} - \sum_{r=1}^{R}\binom{t_{r.}}{2} - \sum_{c=1}^{C}\binom{t_{.c}}{2} + a$$

$$= \left(\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 + n^2 - \sum_{r=1}^{R} t_{r.}^2 - \sum_{c=1}^{C} t_{.c}^2\right)/2$$

Amb aquests quatre valors ja és possible calcular diversos índexs de rendiment per avaluar els resultats del *clustering*. Un dels índexs més populars és l'Índex de Rand (RI; *Rand, 1971*) i es pot calcular fàcilment amb la següent fórmula:

$$RI = \frac{a+b}{a+b+c+d}$$

Bàsicament, aquest índex pondera aquelles observacions que estan classificades juntes i per separat tant en la partició U com en la partició V. Tot i això l'Índex de Rand presenta certs problemes, com ara el fet que el valor esperat del RI de dues particions aleatòries no pren un valor constant, per exemple zero, o que l'estadístic de Rand s'acosta al seu límit superior d'unitat a mesura que augmenta el nombre de clústers.

És per aquest motiu que s'han desenvolupat nous índexs de rendiment per poder corregir l'RI. Un dels exemples més clars és l'Índex de Rand Ajustat (ARI), que parteix de l'RI i és el que es farà servir en aquest treball. De fet, l'ARI s'ha convertit en un dels índexs de validació de *clustering* de més èxit, donat que en (*Milligran i Cooper, 1986*) es van avaluar diferents índexs per mesurar l'acord entre dues particions en l'anàlisi de *clustering*, i *Milligran* i *Cooper* van recomanar l'ARI com el millor índex per mesurar l'acord entre dues particions en l'anàlisi de *clustering* per a dues particions amb diferent nombre de clústers. Tenint en compte la Taula 4.2 l'ARI es pot calcular com:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

Els valors de l'ARI estan compresos entre 0 i 1, de manera que 0 significa que les dues particions classifiquen les dades completament diferent, en canvi, un valor d'1 significa que les dues particions avaluades classifiquen les dades exactament igual.

Per aplicar l'Índex de Rand ajustat en el software estadístic R es farà ús de la funció **adjustedRandIndex** del paquet **mclust**.

V. DADES

Les dades d'aquest treball provenen de l'estudi de cohort *The English Longitudinal Study of Ageing* (ELSA) dels individus que van participar en l'estudi durant els anys 2008 i 2009 (*Steptoe et. Al, 2013*).

Després d'aplicar un preanàlisi de *clustering* amb el mètode K-means, es va decidir separar per sexe la base de dades per evitar la seva influència a l'hora de crear els clústers. Per aquest motiu, per dur a terme l'anàlisi, s'ha decidit extreure dues mostres aleatòries de 200 observacions, de tal manera que l'anàlisi de *clustering* s'estratificarà en funció del sexe dels participants.

Cada mostra consta d'un total de 23 variables, de les quals 9 són de tipus categòric i 14 són de tipus continu. D'entre totes les variables de la base de dades de l'estudi ELSA, s'han extret aquests 23 atributs atès que, per una banda, són les variables sociodemogràfiques més representatives, i per l'altra banda, són totes aquelles variables contínues relacionades amb la salut. El conjunt de variables seleccionades per procedir amb l'anàlisi d'aquest treball es presenten a continuació a l'apartat 5.1.

Per evitar *missings* i fer imputacions a les dades, donat que l'objectiu del treball és l'aplicació de les tècniques de *clustering*, per a les mostres seleccionades es van agafar individus amb tots els valors observats.

5.1. Variables

Variables sociodemogràfiques:

- **age:** Variable numèrica. És l'edat dels individus seleccionats per la mostra.
- **sex:** variable categòrica sobre el sexe de l'individu.
  - *female*: dona
  - *male*: home
- **marital_status:** variable categòrica amb 4 nivells. Informa sobre l'estat civil dels participants.
  - *single*: solter/a
  - *married-cohabitating*: casat/da o vivint amb la parella
  - *divorced-separated*: divorciat/da o separat/da
  - *widow*: vidu o vídua
- **education:** variable categòrica amb 3 nivells. Marca el nivell d'estudis assolit.
  - *primary*: ensenyament primari o inferior
  - *secondary*: estudis secundaris
  - *tertiary*: estudis no obligatoris de grau o màster
- **wealth:** variable categòrica amb 5 nivells. Informa sobre la riquesa de la llar i està definida a través de quintils.
  - *1st Quintile (poorest)*: 1r quintil, els més pobres
  - *2nd Quintile*: 2n quintil
  - *3rd Quintile*: 3r quintil

- *4th Quintile*: 4t quintil
- *5th Quintile (wealthiest)*: 5è quintil, els més rics

Variables d'estil de vida i salut (categòriques):

- **loneliness:** variable categòrica sobre si l'individu s'ha sentit sol en algun moment.
  - *No*: no
  - *Yes*: sí
- **level_pa:** variable categòrica amb 4 nivells. Informa sobre el nivell d'activitat física i sedentarisme.
  - *Very/a lot/high*: molta activitat física, persona molt activa
  - *Fairly/moderate*: activitat física moderada
  - *Not very/low*: molt poca activitat física
  - *Not at all/inactive*: cap activitat física, persona inactiva
- **ah:** variable categòrica que marca si l'individu pateix hipertensió arterial.
  - *No*: no
  - *Yes*: sí
- **depression:** variable categòrica que informa sobre l'estat de depressió en el moment de l'enquesta.
  - *No*: no
  - *Yes*: sí
- **srh:** variable categòrica amb 3 nivells. Marca com s'autoavaluen els individus respecte a la seva salut.
  - *Good*: bona salut
  - *Average/Fair/Moderate*: estat de salut moderat
  - *Poor*: mal estat de salut

Variables sobre mesures físiques (contínues):

- **bmi:** variable numèrica. Informa sobre l'índex de massa corporal ($\mathrm{kg/m^2}$). Els valors de l'índex de massa corporal es poden classificar com:
  - Individus sota de pes: BMI $<$ 18,5
  - Individus amb pes normal: 18,5 $\leq$ BMI $\leq$ 24,9
  - Individus amb preobesitat: 25,0 $\leq$ BMI $\leq$ 29,9
  - Individus amb obesitat de classe I: 30,0 $\leq$ BMI $\leq$ 34,9
  - Individus amb obesitat de classe II: 35,0 $\leq$ BMI $\leq$ 39,9
  - Individus amb obesitat de classe III: BMI $\geq$ 40
- **waist:** variable numèrica. Informa sobre la circumferència de la cintura (cm).
- **hip:** variable numèrica. Informa sobre la circumferència dels malucs (cm).
- **grip:** variable numèrica. És la força d'adherència amb la mà dominant. Marca el màxim pes que s'ha pogut agafar amb la mà dominant (kg)

Variables de laboratori (contínues):

- **sbp:** variable numèrica. És la pressió sanguínia sistòlica (mmHg, mil·ligrams de mercuri) que marca quan la pressió arterial està en el seu punt més alt. Els valors normals en què es mou la pressió sistòlica són entre 90 i 120 mmHg. Sobrepassar aquests valors pot significar un risc important de patir malalties cardiovasculars. A mesura que passen els anys va augmentant.

- **dbp:** variable numèrica. És la pressió sanguínia diastòlica (mmHg, mil·ligrams de mercuri) que marca quan la pressió arterial està en el seu punt més baix. Els valors normals en què es mou la pressió diastòlica són entre 60 i 80 mmHg. Sobrepassar aquests valors pot significar un risc important de patir malalties cardiovasculars però de manera menys exagerada que amb la pressió sistòlica. A mesura que passen els anys va disminuint.

Els valors de la pressió sistòlica i diastòlica interpretats conjuntament marquen els nivells de la pressió arterial i es poden classificar quatre categories generals:

  - Pressió arterial normal: pressió sistòlica per sota de 120 mmHg i pressió diastòlica per sota de 80 mmHg.
  - Pressió arterial alta: pressió sistòlica entre 120 i 129 mmHg i pressió diastòlica per sota de 80 mmHg.
  - Hipertensió en etapa 1: pressió sistòlica entre 130 i 139 mmHg o pressió diastòlica entre 80 i 89 mmHg.
  - Hipertensió en etapa 2: pressió sistòlica igual o major a 140 mmHg o pressió diastòlica igual o major a 90 mmHg.

- **glucose:** variable numèrica. Marca els nivells de glucosa/sucre en sang (mg/dl). Sense haver ingerit cap aliment, els nivells de glucosa es poden classificar com:
  - Normal: menys de 100 mg/dl.
  - Prediabetis: entre 101 i 125 mg/dl.
  - Diabetis: Por sobre de 126 mg/dl.
- **triglycerides:** variable numèrica. Informa sobre el nivell de triglicèrids en sang (mg/dl). És un tipus de grassa que s'acumula a l'organisme i el cos l'utilitza quan necessita energia. Un alt nivell de triglicèrids pot augmentar el risc de patir malalties del cor com la malaltia de les artèries coronàries. Els nivells de triglicèrids es classifiquen en quatre grups:
  - Normal: menys de 150 mg/dl.
  - Límit alt: entre 150 i 199 mg/dl.
  - Alt: entre 200 i 499 mg/dl.
  - Molt alt: més de 500 mg/dl.
- **hdl_chol:** Variable numèrica. Marca el nivell de colesterol "bo" de cada individu (mg/dl). L'HDL absorbeix el colesterol i el torna al fetge, per després ser expulsat del cos. Nivells alts d'HDL poden disminuir el risc de patir malalties cardíaques i atacs de cor. Al Regne nit, el nivell desitjat d'aquest tipus de colesterol varia en funció del sexe:
  - Dones: menys de 46,40 mg/dl.
  - Homes: menys de 42,54 mg/dl.
- **ldl_chol:** Variable numèrica. Marca el nivell de colesterol "dolent" de cada individu (mg/dl). L'LDL produeix la majoria del colesterol en l'organisme. Nivells alts d'LDL poden augmentar el risc de patir malalties cardíaques i atacs de cor. Al Regne Unit, el nivell desitjat d'aquest tipus de colesterol es troba per sota dels 116 mg/dl.
- **total_chol:** Variable numèrica. Marca el nivell de colesterol total de cada individu (mg/dl). Nivells alts de colesterol total poden augmentar el risc de patir malalties cardíaques i atacs de cor. Al Regne Unit, el nivell desitjat de colesterol total es troba per sota dels 193,35 mg/dl.
- **crp:** variable numèrica. Informa sobre els nivells de la proteïna PCR en sang (mg/l). Aquesta proteïna a nivells alts pot augmentar el risc de patir un atac de cor. Segons un

estudi fet sobre la base de dades ELSA (*B Au et. al, 2014*) es considera com a nivell alt d'aquesta proteïna tenir més de 3 mg/l.

Variable sobre envelliment saludable (continua):

- **healthstatus:** variable numèrica. És un indicador de salut en quant a la capacitat funcional de les persones. Va ser creat en l'*Ageing Trajectories of Health-Longitudinal Opportunities and Synergies project* (projecte ATHLOS; *Sanchez-Niubo et. al, 2019*) a través de la teoria de resposta a l'ítem (*Sanchez-Niubo et. al, 2020*; *Rasch, 1960*). Engloba la informació sobre indicadors cognitius, psicològics, de vitalitat, de funcions sensorials, de mobilitat/locomoció, d'activitats de la vida diària i activitats instrumentals (p.ex. vestir-se, menjar, fer les feines de casa, etc.). L'escala es va construir segons una distribució $N \sim (\mu = 50, \sigma = 10)$. Puntuacions més altes de 50 indiquen millor salut funcional.

A l'hora d'aplicar els mètodes de *clustering* al capítol VI, es crearan els clústers tenint en compte totes les variables numèriques exceptuant l'edat dels enquestats. D'aquesta manera les variables que s'utilitzaran són: **bmi**, **waist**, **hip**, **sbp**, **dbp**, **glucose**, **triglycerides**, **hdl_chol**, **ldl_chol**, **total_chol**, **crp** i **healthstatus**.

5.2. Anàlisi univariant

A continuació es presenta l'anàlisi exploratòria de les dades per totes aquelles variables contínues i, seguidament, es presenta l'anàlisi exploratòria per a les variables de caràcter categòric.

5.2.1. Variables numèriques

Per a les mostres de dones i d'homes, s'ha calculat el rang, la mitjana i la desviació tipus de cada variable. També, tant per dones com per homes, s'ha aplicat el test de normalitat Shapiro-Wilk (*Shapiro i Wilk, 1965*) a través de la funció **shapiro.test** que té com a hipòtesis:

$$H_0 = s'asumeix\ normalitat$$

$$H_1 = es\ tenen\ evidències\ per\ rebutjar\ normalitat$$

per determinar quines variables caldrà normalitzar de cara a l'aplicació dels *Gaussian Mixture Models*. D'altra banda, s'ha aplicat un *t-Test* per veure si les variables amb les quals es construiran els clústers presenten diferències significatives entre dones i homes i corroborar la decisió d'estratificar l'anàlisi en funció del sexe.

A la Taula 5.1 es mostren totes aquestes mesures. Pel test de normalitat i pels *t-Test* s'ha tingut en compte una significació del 5% i a la taula apareix el p-valor que s'ha obtingut en cada cas.

Taula 5.1: Estadístics descriptius i test de normalitat per cada variable de la mostra de dones i d'homes, amb el càlcul del *t-Test* per determinar les variables estadísticament diferents entre sexe.

Variables Numèriques	Dones				Homes				t-Test
	Rang	Mitjana	Sd	Shapiro-Wilk	Rang	Mitjana	Sd	Shapiro-Wilk	
Edat	60,0 - 79,0	68,00	5,57	<0,001	60,0 - 80,0	68,46	5,21	<0,001	0,3996
BMI	16,9 - 50,6	27,75	5,28	<0,001	19,1 - 40,4	27,72	3,91	0,0059	0,9537
Circumferència cintura	65,0 - 131,8	91,46	12,35	0,1043	78,1 - 128,2	101,40	10,64	0,1604	<0,001
Circumferència maluc	83,0 - 147,2	107,20	11,61	<0,001	87,8 - 132,3	106,39	7,34	0,0524	0,4283
Força d'adherència	0 - 40,0	24,26	6,01	0,0215	20,0 - 67,0	39,57	7,76	0,279	<0,001
Pressió sistòlica	92,7 - 185,7	134,06	18,28	0,0526	101,0 - 204,7	135,60	17,91	<0,001	0,3948
Pressió diastòlica	52,5 - 97,0	75,40	9,72	0,2864	47,0 - 103,5	75,63	10,41	0,9966	0,8195
Glucosa	48,6 - 149,4	87,12	11,81	<0,001	63,0 - 243,0	91,62	15,54	<0,001	0,0012
Triglicèrids	53,1 - 389,7	132,32	62,56	<0,001	35,4 - 380,9	134,09	68,28	<0,001	0,7869
Colesterol HDL	30,9 - 119,9	66,88	15,44	<0,001	30,9 - 119,9	54,95	13,69	<0,001	<0,001
Colesterol LDL	54,1 - 243,6	132,50	37,32	0,1505	46,4 - 208,8	120,38	34,90	0,0481	<0,001
Colesterol total	146,9 - 371,2	225,72	42,88	0,0181	100,5 - 317,1	201,93	40,52	0,3185	<0,001
Proteïna PCR	0,3 - 41,8	3,87	5,41	<0,001	0,2 -79,5	3,31	6,96	<0,001	0,3704
Healthstatus	30,3 - 66,5	50,68	8,29	0,0068	31,1 - 66,5	51,96	7,85	0,001	0,1146

Per les dones no es tenen evidències suficients per rebutjar la hipòtesi de normalitat per a la **circumferència de la cintura**, la **pressió sistòlica** i **diastòlica** i el **colesterol LDL**. Pels homes, les variables que han presentat normalitat han estat la **circumferència de la cintura** i el **maluc**, la **força d'adherència**, la **pressió diastòlica** i el **colesterol total**.

Tanmateix, s'observen variables que presenten un comportament diferent entre sexe. En la **circumferència de la cintura**, els homes en mitjana presenten valors més alts, tot i que en comparació, les dones tenen associat un rang i una desviació tipus més gran. Quant a la **força d'adherència** els homes són capaços d'agafar pesos més grans que les dones, no obstant això presenten un rang i una desviació tipus superior al sexe femení, és a dir, entre homes els pesos que es poden agafar amb la mà dominant varien una mica més respecte a les dones. Pel que fa a la **glucosa** els homes presenten majors nivells de sucre en sang, més rang i també més dispersió, tanmateix, cap dels dos sexes es troba per sobre del nivell normal (glucosa < 100 mg/dl). Pel **colesterol HDL** (el "bo"), les dones en mitjana presenten valors superiors al seu llindar (HDL > 46,40 mg/dl), els homes en mitjana també presenten valors per sobre del seu llindar (HDL > 42,54 mg/dl), tot i això els homes mostren valors molt inferiors respecte a les dones i presenten menys dispersió, tenen una desviació tipus menor respecte el sexe femení. Per últim, en les variables sobre el **colesterol LDL** (el "dolent") i el **colesterol total**, les dones presenten valors més elevats tant pel rang, com la mitjana i la desviació tipus respecte els homes, tot i això els dos sexes es troben per sobre dels llindars recomanats (LDL > 116 mg/dl i total > 193,35 mg/dl).

De les 13 variables contínues que s'utilitzaran per crear els clústers, en total 6 mostren diferències entre homes i dones, per tant corroborem l'estratificació de l'anàlisi per poder trobar perfils diferenciats en funció del sexe.

5.2.2. Variables categòriques

A continuació es mostra una taula de freqüències completa per dones i homes.

Taula 5.2: Freqüències de les variables per dones i homes amb el test Chi-quadrat per determinar les variables diferents entre sexe.

Variables Categòriques			
Estat civil			
Nivells	Dones	Homes	Chi-quadrat
Solter/a	3,50%	5,50%	<0,001
Casat/da	62,50%	83,00%	
Divorciat/da	11,00%	4,50%	
Vidu/Vídua	23,00%	7,00%	
Educació			
Primària	41,50%	19,50%	<0,001
Secundària	43,00%	58,00%	
Estudis de grau	15,50%	22,50%	
Riquesa de la llar			
1r quintil (més pobres)	20,50%	12,00%	0,0272
2n quintil	25,50%	20,00%	
3r quintil	17,00%	19,00%	
4t quintil	16,00%	26,00%	
5è quintil (més rics)	21,00%	23,00%	
Soledat			
Sí	8,00%	6,00%	0,5566
No	92,00%	94,00%	
Activitat física			
Molta	15,50%	24,00%	0,0041
Moderada	54,50%	60,00%	
No molta	27,00%	13,50%	
Inactivitat	3,00%	2,50%	
Hipertensió arterial			
Sí	35,00%	37,50%	0,6774
No	65,00%	62,50%	
Depressió			
Sí	13,50%	7,00%	0,0479
No	86,50%	93,00%	
Autoavaluació de salut			
Bona	82,00%	77,50%	0,5294
Normal	15,00%	19,00%	
Dolenta	3,00%	3,50%	

A la Taula 5.2 es presenta els percentatges de les freqüències de les variables categòriques i, per veure si el sexe influeix en cada una de les variables, s'ha aplicat el test no paramètric d'independència Chi-quadrat. Les hipòtesis del test són:

$$H_0 = independència\ entre\ variables$$

$$H_1 = dependència\ entre\ variables$$

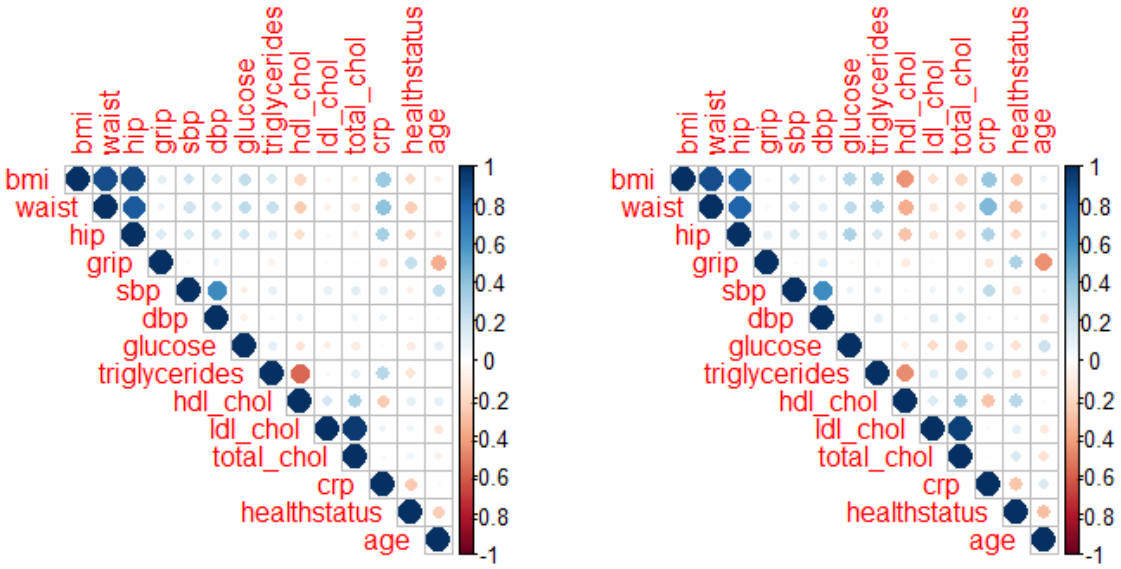
Amb una significació del 5%, les variables estadísticament dependents al sexe han estat **l'estat civil** dels enquestats, que es pot veure com els homes majoritàriament estan casats i hi ha més percentatge de dones divorciades o vídues. L**'educació dels pacients**, on les dones estan repartides entre l'educació primària o inferior i l'educació secundària i els homes estan repartits entre l'educació secundària i els estudis de grau. La **riquesa de la llar**, en què la majoria de dones es troben en el 2n quintil i la majoria d'homes es troba en el 4t quintil. El **nivell d'activitat física**, on es pot apreciar que el percentatge d'homes que es troben en els nivells "molta" i "moderada" (84%) és major al percentatge de dones en aquests dos nivells (70%) i, per concloure, la variable referent a la **depressió**, on les dones pràcticament presenten el doble de casos de depressió que els homes.

5.3. Anàlisi multivariant

5.3.1. Variables numèriques

Per fer una breu anàlisi multivariant de les variables numèriques, a continuació, a la Figura 5.1 es mostren els gràfics de les correlacions per cada sexe. Atès que molt poques variables segueixen una distribució normal, el càlcul de les correlacions s'ha fet d'acord amb la ρ d'Spearman.

Figura 5.1: Gràfic de correlacions per dones, esquerra, i gràfic de correlacions per homes, dreta.



Pels dos sexes, els nivells de colesterol totals estan molt correlacionats positivament amb els nivells de colesterol LDL. Les variables referents l'índex de massa corporal, la circumferència de la cintura i la circumferència dels malucs també estan fortament correlacionades de manera positiva pels dos sexes.

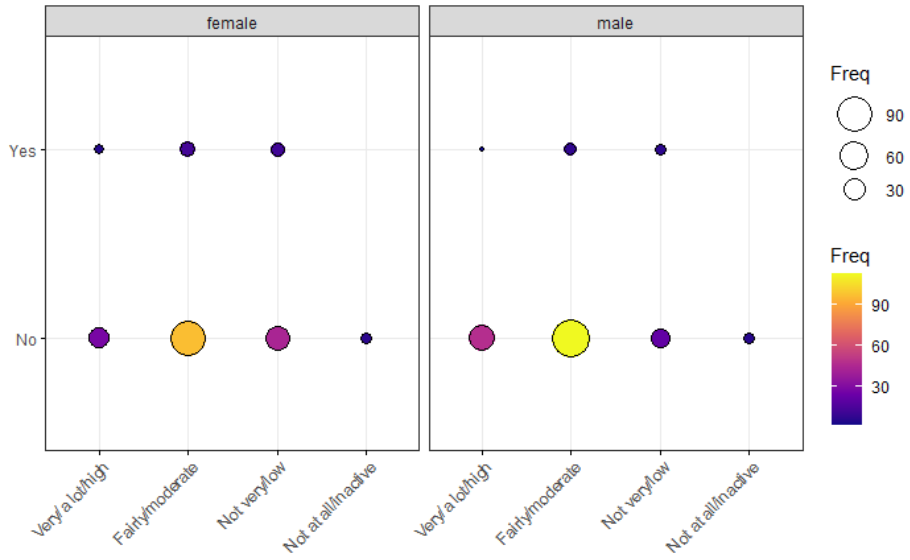
Sobre les correlacions negatives, es pot observar com en els dos sexes, l'edat i la força d'adherència estan bastant correlacionades, a més edat menys força d'adherència. Pel colesterol HDL i els nivells de triglicèrids passa el mateix, a més colesterol HDL (el "bo") menys grassa acumulada (triglicèrids) per gastar energia, aquesta relació sembla ser més forta pel sexe femení.

5.3.2. Variables categòriques

En l'anàlisi multivariant de les variables categòriques, s'han agafat quatre de les variables que, en l'anàlisi univariant, s'ha vist que depenien del sexe.

En el primer gràfic es mira la relació entre els nivells d'activitat física (eix X) i la variable referent a la depressió (eix Y) estratificat per sexe.

Figura 5.2: Gràfic de globus entre el nivell d'activitat física i la depressió estratificat per sexe.

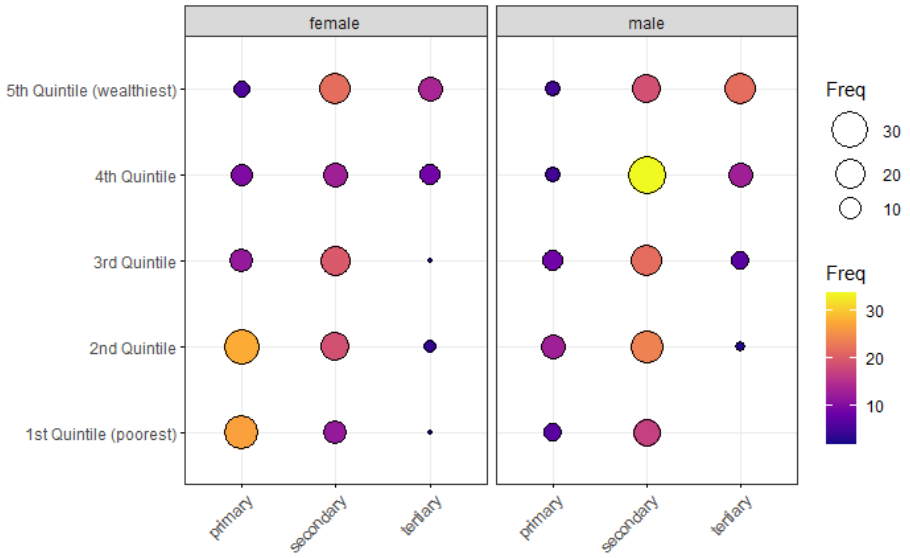


En el gràfic de globus de la Figura 5.2, s'observa que les dones que pateixen depressió practiquen més activitats físiques que els homes que pateixen depressió i que, com s'ha vist en l'anàlisi univariant, les donen presenten més casos de depressió que els homes. D'altra banda, quan ni dones ni homes pateixen depressió, són els homes els que practiquen més activitats físiques als nivells "molta" i "moderada". Tot i això, la majoria d'homes i dones fan alguna activitat física de forma moderada.

La freqüència de dones i homes que no practiquen cap activitat física és idèntica tant pels individus que pateixen depressió com pels individus que no en pateixen.

Al segon gràfic s'ha mirat la relació entre l'educació rebuda pels enquestats (eix X) i el nivell de riquesa de la llar (eix Y) estratificat per sexe.

Figura 5.3: Gràfic de globus entre l'educació i el nivell de riquesa a la llar estratificat per sexe.



Mirant el gràfic de globus de la Figura 5.3, s'observa com la majoria de dones es mouen entre l'educació primària o inferior i l'educació secundària, i que les freqüències més altes es troben entre les dones que s'han quedat a la primària o inferior i estan classificades en el 1r i el 2n quintil de riquesa (els més baixos). No obstant això, les dones que han cursat estudis secundaris es mouen entre el 2n i el 5è quintil, sent el 3r i el 5è quintil els que tenen una representació més elevada. Totes aquelles dones que han cursat estudis de grau, majoritàriament es troben entre el 4t i el 5è quintil.

Pel que fa als homes, la majoria s'han quedat als estudis secundaris i el quintil amb més freqüència és el 4t, tot i això es mouen entre el 2n i el 4t quintil. Molt pocs homes s'han quedat només en l'educació primària o inferior, però dels que es troben en aquest grup, la majoria està entre el 2n i el 3r quintil de riquesa. En contraposició a les dones, hi ha més homes que han cursat estudis de grau.

VI. APLICACIÓ DELS MÈTODES DE CLUSTERING SOBRE DADES DE SALUT

Per tal de poder aplicar les diferents tècniques de *clustering* presentades al capítol II per la mostra de dones i la mostra d'homes, en primer lloc, s'han escalat (estandarditzat) les dues mostres per poder comparar les dades i evitar que les diferents unitats de mesura entre variables influeixin en els resultats a l'hora d'aplicar l'anàlisi de *clustering*.

Tot i no ser estrictament necessari per als mètodes jeràrquics i de partició, a través de la funció **bestNormalize** del paquet **bestNormalize** (*Peterson, 2019*), s'han normalitzat totes les variables de les dues bases de dades per, d'aquesta manera, intentar aplicar els *Gaussian Mixture Models* (GMMs) sobre un conjunt de variables Gaussianes.

Abans de començar a aplicar les tècniques de *clustering*, per mitjà de l'estadístic de Hopkins (*Lawson i Jurs, 1990*) s'ha comprovat que les dues bases de dades fossin clusteritzables. Aquest estadístic s'utilitza per avaluar la tendència de *clustering* d'una base de dades mesurant la probabilitat que la base de dades en qüestió estigui generada per una distribució uniforme. En altres paraules, prova l'aleatorietat espacial de les dades.

Les hipòtesis que segueix aquest estadístic són les següents:

$$H_0 = la\ base\ de\ dades\ es\ distribueix\ uniformement, no\ és\ clusteritzable$$

$$H_1 = la\ base\ de\ dades\ no\ es\ distribueix\ uniformement, és\ clusteritzable$$

i els valors que pot prendre estan compresos entre 0 i 1. Quan el valor de l'estadístic és superior a 0,5 es considera que la base de dades és clusteritzable i, per tant, que les dades no es distribueixen de manera uniforme. El càlcul de l'estadístic de Hopkins s'ha fet a través de la funció **get_clust_tendency** del paquet **factoextra**.

A la Taula 6.1 es presenten els resultats de l'estadístic de Hopkins per la base de dades de dones i homes. El valor que s'ha obtingut de l'estadístic, ambdós casos, ha estat major a 0,5. Per tant, s'ha considerat que les dues bases de dades són clusteritzables i, conseqüentment, que es pot seguir amb l'anàlisi de *clustering*.

Taula 6.1: Valor de l'estadístic de Hopkins per a la base de dades de dones i homes.

Dones	Homes
0,69	0,66

L'aplicació de les tècniques de *clustering* comença amb l'anàlisi completa del mètode K-means i, durant tot el capítol, s'han analitzat la resta de mètodes de *clustering* segons el nombre de clústers que s'ha decidit escollir a l'hora d'aplicar el K-means, tant per la mostra de dones com per la mostra d'homes, atès que permetrà aplicar l'ARI sobre un mateix nombre de clústers i, a més a més, al fer el preanàlisi de *clustering* aquest mètode va ser el que va proporcionar una millor partició en l'espai de les dades per a les dues mostres. De manera que, a part de voler trobar estructures de clúster diferenciades, durant tot el capítol s'analitzarà si el mètode

jeràrquic, el K-medoids i el GMMs són capaços de trobar perfils similars per cada un dels *datasets* que s'estudien, prenent com a base de partida de l'anàlisi l'algoritme K-means.

A l'hora de fer el *profiling* dels mètodes, per a cada mostra (dones i homes), s'ha introduït a les mostres originals, sense escalar ni normalitzar les dades i tenint en compte totes les variables (numèriques i categòriques), la variable indicadora dels clústers. Per d'aquesta manera calcular, en primer lloc, les mitjanes de totes les variables numèriques en funció del clúster al qual pertanyen.

Tot seguit, per tal de discernir el comportament dels clústers en funció de les variables numèriques i comprovar quines variables presenten diferències significatives entre les mitjanes de cada clúster, per aquelles variables que han presentat les característiques adients, s'ha aplicat el test ANOVA d'un factor, tenint en compte els clústers com a factor, i per a les variables que no han complert els requisits necessaris se'ls hi ha aplicat el test no paramètric de Kruskal-Wallis. Ambdós tests s'ha tingut en compte una significació del 5%.

Els requisits per poder aplicar el test ANOVA d'un sol factor són els següents:

1) Independència entre els individus de la mostra
2) Cap *outlier*, d'entre els diferents nivells del factor, ha de ser significatiu. S'ha calculat a través de la funció **identify_outliers** del paquet **rstatix** (*Kassambara, 2020*).
3) S'ha de complir la hipòtesi de normalitat per cada nivell del factor analitzat. Per cada clúster s'ha aplicat el test de normalitat Shapiro-Wilk.
4) Les variables comparades han de ser homoscedàstiques. Les variàncies, per cada nivell del factor, han de ser homogènies entre elles. S'ha calculat el test de Levene (*Levene, 1960*) per comprovar-ho, que té les següents hipòtesis:

$$H_0 = variància\ entre\ grups\ igual, homoscedasticitat$$
$$H_1 = variància\ entre\ grups\ diferent, heteroscedasticitat$$

Després, a través de la funció **TukeyHSD** del paquet **stats**, pel test ANOVA, i la funció **dunn_tets** del paquet **rstatix**, pel test de Kruskal-Walis, s'ha fet l'anàlisi post-hoc del test de Tukey i del test de Dunn per determinar quins clústers són diferents i/o similars entre ells.

Finalment, per acabar de perfilar cada un dels clústers, s'ha estudiat quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat. Les variables que han indicat dependència en funció del clúster al qual pertanyen, se'ls hi ha calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells. Pel test de Chi-quadrat també s'ha considerat una significació del 5%.

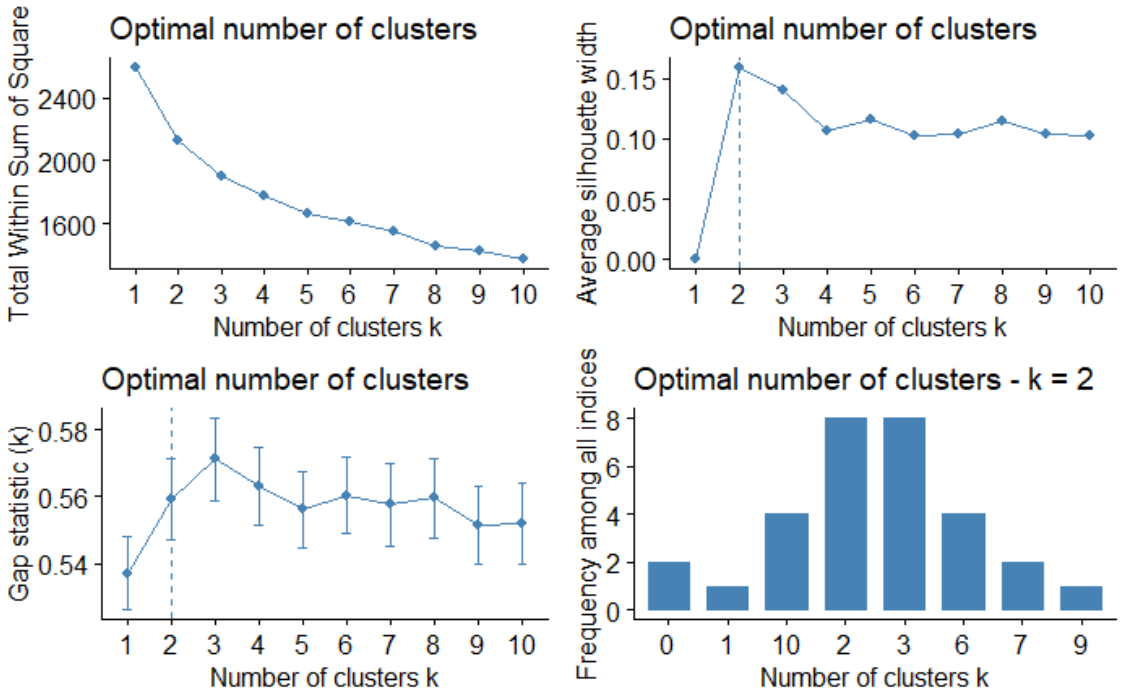
Tenint en compte la metodologia que s'ha emprat per dur a terme els *profilings* de cada mètode, els resultats es presenten just a continuació.

6.1. K-means

6.1.1. Dones

Al tractar-se d'un mètode de partició, en primer lloc cal escollir per quin nombre de clústers es vol partir la base de dades. Per això, s'ha calculat l'*elbow method*, l'*average silhouette method*, el *gap statistic* i la funció **NbClust** (veure Figura 6.1) que permet observar, d'entre 30 mètodes de selecció diferents, el nombre de clústers òptims d'acord amb la regla de la majoria (la partició que ha estat escollida més vegades d'entre els diferents mètodes de selecció).

Figura 6.1: Gràfics de l'*elbow method*, l'*average silhouette method*, el *gap statistic* i la funció NbClust.



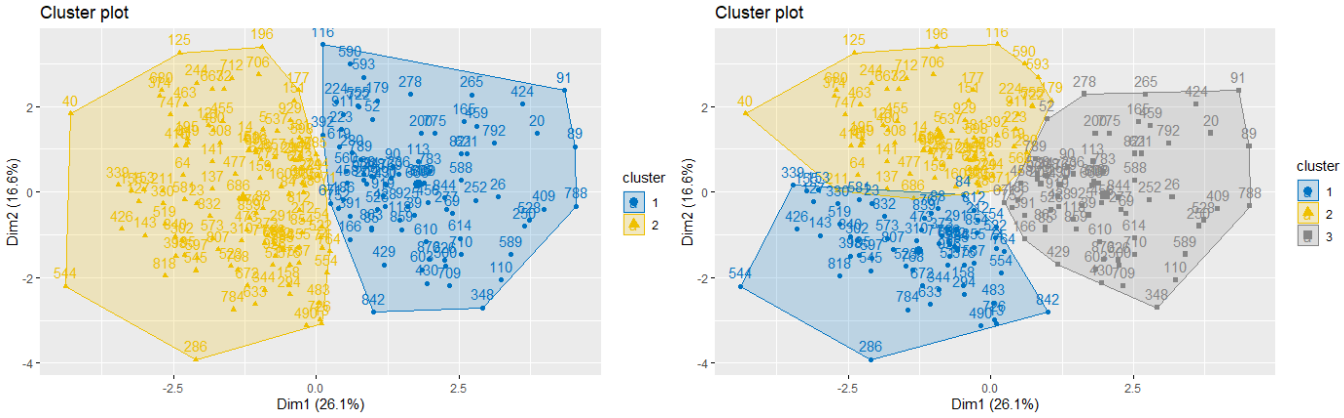
Els tres mètodes de selecció de clúster òptims indiquen, en primer lloc, que l'*elbow method* està entre 3 o 5 clústers, per l'*average silhouette method* marca que la millor partició seria considerar 2 clústers i pel *gap statistic* indica que la millor partició són 3 clústers. La funció **NbClust** , per la regla de la majoria, considera que el nombre de clústers òptim és 2, malgrat això, vuit mètodes han seleccionat tant la partició amb 2 clústers com la partició amb 3 clústers.

Per tal de decidir la millor divisió possible, s'ha aplicat el mètode K-means per 2 i 3 clústers, amb l'opció "nstart = 25" de la funció **kmeans** per compilar l'algoritme 25 vegades i extreure la divisió més òptima en l'espai de les dades. A la Figura 6.2 es mostra com es reparteixen les dades en un espai factorial bidimensional a través de la funció **fviz_cluster** del paquet **factoextra**.

Mirant la Figura 6.2, es pot observar com les dues particions divideixen la mostra de dones de manera diferenciada. No obstant això, per motius de caràcter contextual, s'ha decidit fer el

profiling en base a 3 clústers, atès que l'objectiu de l'anàlisi de *clustering* és poder trobar perfils amagats que no es veuen ni es coneixen a simple vista. D'aquesta manera, quants més perfils diferenciats es puguin distingir, posteriorment a través de l'estudi d'*outcomes* de caràcter longitudinal permetrà estudiar noves hipòtesis per millorar i desenvolupar tractaments innovadors en el camp de l'envelliment saludable (fora de l'*scope* d'aquest treball).

Figura 6.2: Visualització de les particions entre 2 i 3 clústers en un espai factorial bidimensional.



Així doncs, amb tres clústers l'algoritme K-means considera que les dones queden repartides de la següent manera: el clúster 1 compta amb 62 dones, el clúster 2 compta amb 68 dones i el clúster 3 compta amb 70 dones.

6.1.2. Profiling dones

A la Taula 6.1, es presenta el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original de dones.

Taula 6.1: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	24,82	84,47	101,15	24,48
2	25,17	85,43	101,60	23,56
3	32,85	103,53	117,88	24,74
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
138,84	76,98	83,64	125,28	77,78
124,93	72,10	86,03	113,45	64,43
138,68	77,21	91,26	156,90	59,61
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
166,71	269,31	2,57	52,40	68,19
105,38	192,55	2,03	53,69	67,91
128,55	219,31	6,79	46,24	67,91

Tot seguit, la Taula 6.2 mostra quines variables numèriques han presentat diferències significatives entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric en funció del comportament de cada variable, i el p-valor obtingut pels mètodes de comparació de mitjanes. A més, s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.2: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,8750	-	-
BMI	Kruskal-Wallis	< 0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	ANOVA	< 0,001	Tukey	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	< 0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	Kruskal-Wallis	0,4770	-	-
Pressió sistòlica	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Pressió diastòlica	ANOVA	0,0023	Tukey	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Glucosa	Kruskal-Wallis	< 0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Triglicèrids	Kruskal-Wallis	< 0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol HDL	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Colesterol LDL	ANOVA	< 0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol total	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Proteïna PCR	Kruskal-Wallis	< 0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Healthstatus	ANOVA	< 0,001	Tukey	C1 = C2 / C1 ≠ C3 / C2 ≠ C3

Per poder dur a terme el perfil dels clústers de manera completa, a la Taula 6.3 es mostra quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.3: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,6572	-
Educació	0,4705	-
Riquesa de la llar	0,1152	-
Soledat	0,4975	-
Activitat física	0,4005	-
Hipertensió arterial	<0,001	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Depressió	0,7834	-
Autoavaluació de salut	0,1643	-

En funció a les tres taules de l'anàlisi, tenint en compte que totes les dones presenten en mitjana una edat al voltant dels 68 anys, són capaces d'agafar pesos, aproximadament, d'entre 23 i 25 quilograms i tenen un nivell colesterol HDL superior a 46,40 mg/dl (veure Taula 6.1), els clústers es poden definir breument com:

- **Clúster 1:** Dones amb risc de partir malalties cardiovasculars però amb un alt índex de salut funcional.
- **Clúster 2:** Dones sanes amb un alt índex de salut funcional.
- **Clúster 3:** Dones amb alt risc de patir malalties cardiovasculars que presenten un índex de salut funcional baix.

L'anàlisi completa de cada clúster respecte de les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen dones amb un índex de massa corporal al límit del nivell normal ($18,5 \leq BMI \leq 24,9$) i la preobesitat ($25,0 \leq BMI \leq 29,9$), amb baix nivell de proteïna PCR (< 3 mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl). Presenten nivells de colesterol LDL i total molt per sobre dels llindars recomanats (LDL > 116 mg/dl, total > 193,35 mg/dl). Quant al colesterol HDL, és el clúster que presenta valors més alts. Pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i presenten un índex de salut funcional alt (healthstatus > 50), estadísticament iguals al clúster 2.

El **clúster 2** comparteix la majoria de característiques del clúster 1. El defineixen dones amb un índex de massa corporal al límit del nivell normal ($18,5 \leq BMI \leq 24,9$) i la preobesitat ($25,0 \leq BMI \leq 29,9$), amb baix nivell de proteïna PCR (< 3 mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl). És el clúster amb els nivells de colesterol LDL i total més baixos, que es troben per sota dels llindars establerts (LDL < 116 mg/dl, total < 193,35 mg/dl) i presenten el colesterol HDL més baix, ja que és estadísticament igual al clúster 3. Tenen la pressió arterial alta (pressió sistòlica entre 120 i 129 mmHg i pressió diastòlica per sota de 80 mmHg) i presenten un índex de salut funcional alt (healthstatus > 50), el més alt en mitjana d'entre tots els clústers però estadísticament igual al clúster 1.

El **clúster 3** és el més diferent de tots. El defineixen dones amb un índex de massa corporal classificat com obesitat de classe I ($30,0 \leq BMI \leq 34,9$), amb els nivells de glucosa en sang més alts respecte als altres clústers, tot i que encara es manté a un nivell normal (< 100 mg/dl), amb nivells preocupants de proteïna PCR, molt per sobre del límit establert (> 3 mg/dl) i amb un nivell de triglicèrids al límit alt (entre 150 i 199 mg/dl), el més alt entre tots els clústers. Presenten nivells de colesterol LDL i total per sobre dels llindars recomanats (LDL > 116 mg/dl, total > 193,35 mg/dl), no tan elevats com en el clúster 1, i amb el nivell de colesterol HDL en mitjana més baix, estadísticament igual al clúster 2. Pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i presenten un índex de salut funcional baix (healtstatus < 50), el més baix amb diferència d'entre tots els clústers.

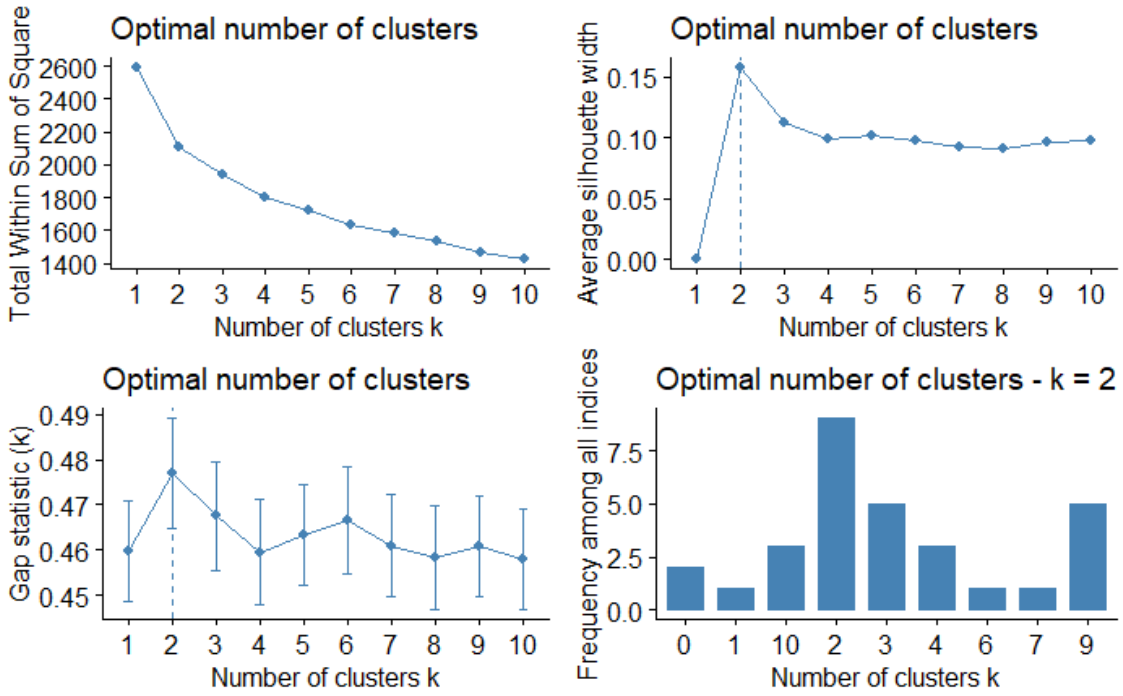
Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.1 de l'Annex referent al sexe femení.

6.1.3. Homes

Com s'ha fet amb la mostra de dones, en primer lloc s'ha d'escollir per quin nombre de clústers es vol dividir la base de dades. Per això, s'ha calculat l'*elbow method*, l'*average silhouette*

method, el *gap statistic* i la funció **NbClust** (veure Figura 6.3) per observar d'entre 30 mètodes de selecció diferents, el nombre de clústers òptims d'acord amb la regla de la majoria.

Figura 6.3: Gràfics de l'*elbow method*, l'*average silhouette method*, el *gap statistic* i la funció NbClust.

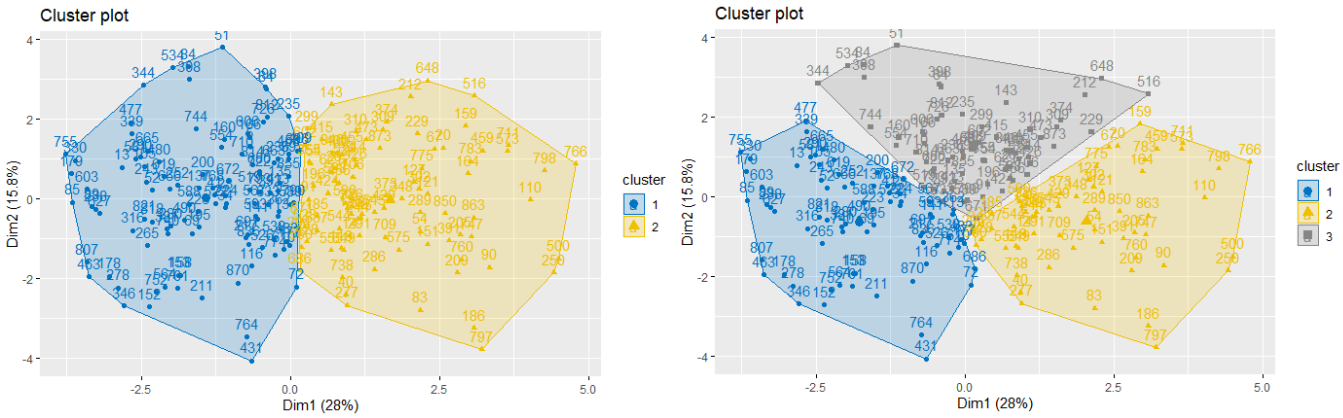


En aquest cas, observant la Figura 6.3, per la mostra d'homes els tres mètodes de selecció òptima de clústers arriben a la mateixa conclusió. Indiquen que el nombre de clústers òptim és 2. La funció **NbClust**, per la regla de la majoria, també marca que el nombre de clústers òptim és 2. Tot i això s'observa que cinc mètodes de selecció han escollit la partició per 3 i 9 clústers.

Tot i semblar que la decisió és clara, per assegurar-nos que els homes queden ben dividits, s'ha aplicat el mètode K-means per les particions de 2, 3 i 9 clústers. Quan s'han dividit les dades en 9 grups, la majoria de clústers es solapaven entre ells, per tant, aquesta partició ha quedat descartada.

En aplicar el mètode per 2 i 3 clústers amb l'opció "nstart = 25" de la funció **kmeans** per compilar l'algoritme 25 vegades i extreure la divisió més òptima en l'espai de les dades, la mostra d'homes ha presentat una divisió clara i sense solapament. A la Figura 6.4 es mostra com es reparteixen les dades en un espai factorial bidimensional a través de la funció **fviz_cluster**.

Figura 6.4: Visualització de les particions entre 2 i 3 clústers en un espai factorial bidimensional.



Les dues particions divideixen la mostra d'homes de manera diferenciada. No obstant això, en la línia del plantejament que s'ha aplicat per la mostra de dones, a causa de motius de caràcter contextual, s'ha decidit fer el *profiling* en base a 3 clústers, ja que l'algoritme produeix una bona divisió i es pot observar un patró en les dades que no es veu a simple vista si no s'aplica el mètode. D'aquesta manera, quants més perfils diferenciats es puguin distingir, posteriorment per mitjà de l'estudi d'*outcomes* de caràcter longitudinal permetrà estudiar noves hipòtesis per millorar i desenvolupar nous tractaments en el camp de l'envelliment saludable (fora de l'*scope* d'aquest treball).

Així doncs, amb tres clústers, l'algoritme K-means considera que els homes queden repartits de la següent manera: el clúster 1 compta amb 74 homes, el clúster 2 compta amb 60 homes i el clúster 3 compta amb 66 homes.

6.1.4. Porfiling homes

El càlcul de les mitjanes de totes les variables numèriques de la base de dades, introduint la variable indicadora dels clústers a la mostra original d'homes, es presenten a la Taula 6.4.

Taula 6.4: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	24,71	93,09	101,14	40,18
2	31,87	113,03	113,63	39,18
3	27,32	100,15	105,69	39,23
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
133,43	75,51	85,45	124,72	62,45
140,62	77,20	95,70	173,89	48,79
133,47	74,34	94,83	108,43	52,15
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
144,07	231,13	1,69	54,98	67,39
120,84	204,18	4,55	48,97	68,62
93,39	167,16	3,97	51,28	69,50

Tot seguit, a la Taula 6.5 es mostra quines variables numèriques han presentat diferències significatives entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric en funció del comportament de cada variable, i el p-valor obtingut pels mètodes de comparació de mitjanes. A més, s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.5: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,0344	Dunn	C1 = C2 / C1 ≠ C3 / C2 = C3
BMI	ANOVA	< 0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	ANOVA	0,6970	-	-
Pressió sistòlica	Kruskal-Wallis	0,0699	-	-
Pressió diastòlica	ANOVA	0,3040	-	-
Glucosa	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Triglicèrids	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Colesterol HDL	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Colesterol LDL	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol total	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Proteïna PCR	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Healthstatus	Kruskal-Wallis	< 0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3

Per poder dur a terme el perfil dels clústers de manera completa, a la Taula 6.6 es mostra quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.6: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,6241	-
Educació	0,4143	-
Riquesa de la llar	0,1181	-
Soledat	0,2490	-
Activitat física	0,0176	C1 ≠ C2 / C1 = C3 / C2 = C3
Hipertensió arterial	0,09829	-
Depressió	0,2337	-
Autoavaluació de salut	<0,001	C1 ≠ C2 / C1 = C3 / C2 ≠ C3

En funció a les tres taules de l'anàlisi, tenint en compte que tots els homes presenten en mitjana hipertensió arterial entre l'etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i l'etapa 2 (pressió sistòlica major a 140 mmHg), poden agafar, aproximadament, pesos fins a 40

quilograms i tenen un nivell colesterol HDL superior a 42,54 mg/dl (veure Taula 6.4), els clústers es poden definir de manera breu com:

- **Clúster 1**: Homes d'uns 67 anys, actius físicament, amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt
- **Clúster 2**: Homes d'entre 67 i 70 anys, poc actius físicament, amb alt risc de patir malalties cardiovasculars i un índex de salut funcional baix.
- **Clúster 3**: Homes d'uns 70 anys, moderats quant a l'activitat física, amb un risc mitjà de patir malalties cardiovasculars i un índex de salut funcional baix.

L'anàlisi completa de cada clúster respecte a les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen homes d'uns 67 anys, amb un índex de massa corporal normal ($18,5 \leq BMI \leq 24,9$), amb baix nivell de proteïna PCR (< 3 mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl). Presenten els nivells més alts de colesterol LDL i total, per sobre dels llindars recomanats (LDL > 116 mg/dl, total > 193,35 mg/dl), i són homes amb el nivell més alt de colesterol HDL. Practiquen molta activitat física en comparació als altres clústers (cap home està inactiu), s'autoavaluen amb un bon nivell de salut (cap home s'autoavalua amb un nivell de salut pobre) i presenten un índex de salut funcional alt (healthstatus > 50), el més alt entre tots els clústers.

El **clúster 2** el defineixen homes d'entre 67 i 70 anys, amb un índex de massa corporal classificat com a obesitat de classe I ($30,0 \leq BMI \leq 34,9$), amb nivells de glucosa en sang alts, estadísticament iguals al clúster 3, però que encara es manté a un nivell normal (< 100 mg/dl), amb un nivell de proteïna PCR per sobre del límit recomanat (> 3 mg/l), els més alt en tots els clústers, i amb un nivell de triglicèrids al límit alt (valor entre 150 i 199 mg/dl). Presenten nivells de colesterol LDL i total per sobre del llindar recomanat, però no tan elevat com els homes del clúster 1, i són els homes amb el nivell de colesterol HDL més baix. Practiquen poca activitat física, sent el clúster on més homes s'autoavalua amb un nivell de salut dolent i presenten un índex de salut funciona baix (healthstatus < 50), estadísticament igual al clúster 3.

El **clúster 3** el defineixen homes d'uns 70 anys, amb un índex de massa corporal classificat com a preobesitat ($25,0 \leq BMI \leq 29,9$), amb nivells de glucosa en sang alts, estadísticament iguals al clúster 2, però que encara es manté a un nivell normal (< 100 mg/dl), amb nivells de proteïna PCR baixos, estadísticament iguals al clúster 1 (el valor mitjà és elevat a causa de l'existència d'*outliers*), i amb un nivell de triglicèrids normal (< 150 mg/dl), també estadísticament igual al clúster 1. Presenten els nivells de colesterol LDL i total més baixos, per sota dels llindars establerts (LDL < 116 mg/dl, total < 193,35 mg/dl). Practiquen activitats físiques de manera moderada, són homes que s'autoavaluen amb un nivell de salut bo o normal i presenten un índex de salut funcional estadísticament igual al clúster 2, però que en mitjana es troba per sobre de 50.

Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.1 de l'Annex referent al sexe masculí.

6.2. Clustering Jeràrquic

Per dur a terme el *clustering* jeràrquic, s'ha optat per analitzar les dues mostres en funció de l'esquema d'aglomerat. En primer lloc perquè es tracta del mètode jeràrquic més utilitzat, per altra banda, perquè és l'esquema que ha proporcionat una partició dels individus més clara i per últim, donat que presenta resultats equiparables amb els altres mètodes.

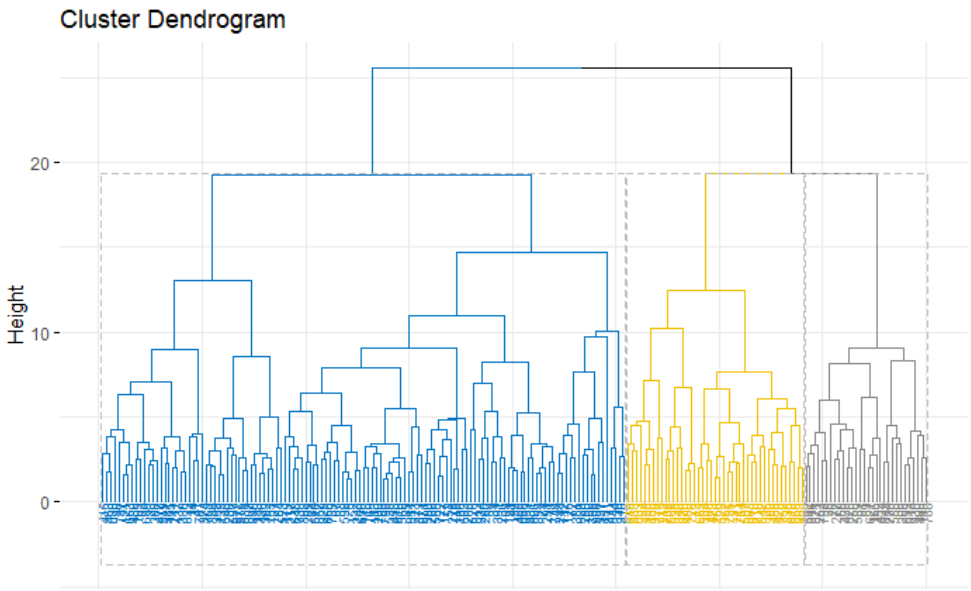
Els dendrogrames de la mostra de dones i la mostra d'homes del *clustering* jeràrquic divisiu apareixen al principi del subapartat 9.1.2 de l'Annex.

6.2.1. Dones

Pel *clustering* jeràrquic aglomeratiu la distància matemàtica que s'ha escollit ha estat l'Euclidiana. Pel que fa a la mesura d'enllaç, a través del coeficient aglomeratiu (AC, veure subapartat 2.1.3.2) calculat amb la funció **agnes**, amb un valor de 0,9020 (molt proper a 1), s'ha escollit l'enllaç de Ward.

A la Figura 6.5 es mostra el dendrograma del *clustering* jeràrquic aglomeratiu dividit en 3 clústers. A simple vista, pot donar a entendre que la millor partició possible podrien ser dos o quatre clústers, tot i això partint del mètode K-means com a base de l'anàlisi, s'han dividit les dades en 3 clústers. El clúster 1 compta amb 65 dones, el clúster 2 compta amb 68 dones i el clúster 3 compta amb 67 dones.

Figura 6.5: Dendrograma del *clustering* jeràrquic aglomeratiu dividit en 3 clústers.



6.2.2. Profiling dones

A la Taula 6.7 es presenta el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original de dones.

Taula 6.7: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	24,95	84,90	101,36	24,54
2	25,17	85,43	101,60	23,56
3	33,09	103,97	118,42	24,70
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
138,50	77,09	84,41	127,00	76,80
124,93	72,10	86,03	113,45	64,43
139,01	77,11	90,86	156,65	59,74
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
166,22	268,19	2,63	52,64	67,94
105,38	192,55	2,03	53,69	67,91
127,32	218,17	6,93	45,73	68,15

Tot seguit, la Taula 6.8 mostra quines variables numèriques han presentat diferències significatives entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric, i el p-valor obtingut pels mètodes de comparació de mitjanes. A més, s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.8: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,9490	-	-
BMI	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	ANOVA	<0,001	Tukey	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	Kruskal-Wallis	0,4720	-	-
Pressió sistòlica	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Pressió diastòlica	ANOVA	0.0023	Tukey	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Glucosa	Kruskal-Wallis	0.0028	Dunn	C1 = C2 / C1 ≠ C3 / C2 = C3
Triglicèrids	Kruskal-Wallis	0.0011	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol HDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Colesterol LDL	ANOVA	<0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol total	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Proteïna PCR	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Healthstatus	ANOVA	<0,001	Tukey	C1 = C2 / C1 ≠ C3 / C2 ≠ C3

A continuació, per tal de dur a terme els perfils dels clústers de manera completa, a la Taula 6.9 es mostra quines variables categòriques presenten freqüències diferents en funció dels

clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.9: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,8294	-
Educació	0,2161	-
Riquesa de la llar	0,1151	-
Soledat	0,5718	-
Activitat física	0,3156	-
Hipertensió arterial	<0,001	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Depressió	0,8613	-
Autoavaluació de salut	0,1050	-

En funció de les tres taules de l'anàlisi, tenint en compte que totes les dones presenten en mitjana una edat al voltant dels 68 anys, són capaces d'agafar pesos, aproximadament, d'entre 23 i 25 quilograms i tenen un nivell colesterol HDL superior a 46,40 mg/dl (veure Taula 6.7), els clústers es poden definir breument com:

- **Clúster 1:** Dones amb risc de partir malalties cardiovasculars però amb un alt índex de salut funcional.
- **Clúster 2:** Dones sanes amb un alt índex de salut funcional.
- **Clúster 3:** Dones amb alt risc de patir malalties cardiovasculars que presenten un índex de salut funcional baix.

L'anàlisi completa de cada clúster respecte a les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen dones amb un índex de massa corporal al límit del nivell normal (18,5 ≤ BMI ≤ 24,9) i la preobesitat (25,0 ≤ BMI ≤ 29,9), amb baix nivell de proteïna PCR (< 3mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl). Presenten els nivells de colesterol LDL i total més elevats, molt per sobre dels llindars recomanats (LDL > 116 mg/dl, total > 193,35 mg/dl). Per altra banda, és el clúster que presenta nivells de colesterol HDL més alts. Pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i presenten un índex de salut funcional alt (healthstatus > 50), estadísticament igual al clúster 2.

El **clúster 2** comparteix moltes de les característiques del clúster 1. El defineix dones amb un índex de massa corporal al límit del nivell normal (18,5 ≤ BMI ≤ 24,9) i la preobesitat (25,0 ≤ BMI ≤ 29,9), amb baix nivell de proteïna PCR (< 3 mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl). Es tracta del clúster amb els nivells de colesterol LDL i total més baixos, per sota del llindar establert LDL < 116 mg/dl, total < 193,35 mg/dl), i presenten el colesterol HDL més baix, ja que és estadísticament igual al clúster 3. Tenen la pressió arterial alta (pressió sistòlica entre 120 i 129 mmHg i pressió diastòlica per sota de 80 mmHg) i presenten un índex de salut funcional alt (healthstatus > 50), el més alt en mitjana d'entre tots els clústers i estadísticament igual al clúster 1.

El **clúster 3** és el més diferent de tots. El defineixen dones amb un índex de massa corporal classificat com obesitat de classe I (30,0 ≤ BMI ≤ 34,9), amb els nivells de glucosa en sang més

alts respecte als altres clústers, tot i que encara es manté a un nivell normal (< 100 mg/dl), amb nivells preocupants de proteïna PCR, molt per sobre del límit establert (> 3 mg/dl) i amb un nivell de triglicèrids al límit alt (entre 150 i 199 mg/dl), el més alt entre tots els clústers. Tenen els nivells de colesterol LDL i total per sobre dels llindars recomanats (LDL > 116 mg/dl, total > 193,35 mg/dl), tot i això no són tan elevats com en el clúster 1, i amb el nivell de colesterol HDL en mitjana més baix, estadísticament igual al clúster 2. Pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i presenten l'índex de salut funcional més baix (healthstatus < 50) respecte als altres clústers.

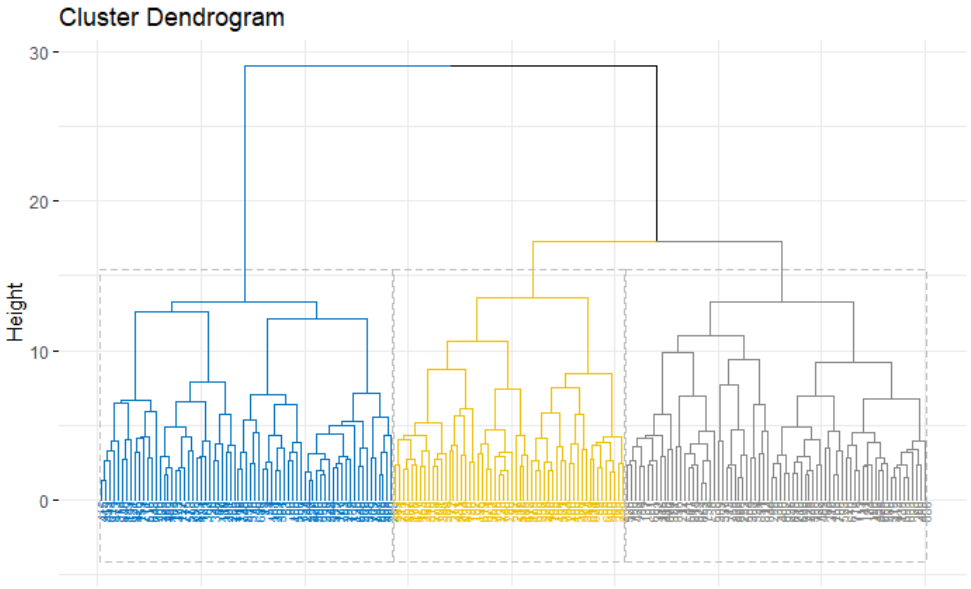
Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.2 de l'Annex referent al sexe femení.

6.2.3. Homes

Per construir el dendrograma de la mostra d'homes s'ha escollit la distància Euclidiana. Quant a la mesura d'enllaç, a través del coeficient aglomeratiu (AC) calculat amb la funció **agnes**, s'ha determinat que l'enllaç de Ward és el millor per aquesta mostra, amb un valor de l'AC de 0,9108.

A la Figura 6.6 està representat el dendrograma del *clustering* jeràrquic aglomeratiu dividit en tres clústers. Aquesta divisió es fonamenta en funció del mètode K-means, donat que és la base d'on parteix tota l'anàlisi, per veure si els resultats que s'obtenen a l'hora de fer el *profiling* són similars entre mètodes.

Figura 6.6: Dendrograma del *clustering* jeràrquic aglomeratiu dividit en 3 clústers.



A simple vista sembla que, per tres clústers, la mostra d'homes presenta una bona divisió de les dades. Per aquesta partició, el clúster 1 compta amb 70 homes, el clúster 2 compta amb 61 homes i el clúster 3 compta amb 69 homes.

6.2.4. Profiling homes

A la Taula 6.10 es mostra el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original d'homes.

Taula 6.10: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	24,56	92,71	100,73	39,69
2	31,90	112,94	113,76	39,72
3	27,24	100,01	105,61	39,30
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
133,17	75,32	85,35	123,48	63,14
139,04	76,74	95,31	174,53	48,62
135,02	74,96	94,72	109,11	52,23
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
144,02	231,47	1,69	55,24	67,39
121,84	205,14	4,45	49,22	68,41
95,11	169,14	3,93	51,05	69,58

Seguidament, a la Taula 6.11 es presenta quines variables numèriques han resultat estadísticament diferents entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric, i el p-valor obtingut pels mètodes de comparació de mitjanes. També s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.11: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,0332	Dunn	C1 = C2 / C1 ≠ C3 / C2 = C3
BMI	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	ANOVA	0,4920	-	-
Pressió sistòlica	Kruskal-Wallis	0,1790	-	-
Pressió diastòlica	ANOVA	0,5980	-	-
Glucosa	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Triglicèrids	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Colesterol HDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Colesterol LDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol total	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Proteïna PCR	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Healthstatus	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3

Finalment, per tal de dur a terme els perfils dels clústers de manera completa, a través del test d'independència Chi-quadrat, a la Taula 6.12 s'ha estudiat quines variables categòriques

presenten freqüències diferents en funció dels clústers i, per les variables en qüestió, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.12: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,5854	-
Educació	0,3340	-
Riquesa de la llar	0,1621	-
Soledat	0,2493	-
Activitat física	0,0282	C1 ≠ C2 / C1 = C3 / C2 = C3
Hipertensió arterial	0,2649	-
Depressió	0,2466	-
Autoavaluació de salut	<0,001	C1 ≠ C2 / C1 = C3 / C2 ≠ C3

En funció de les tres taules de l'anàlisi, tenint en compte que la majoria d'homes en el conjunt de clústers presenten hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg), poden agafar, aproximadament, pesos fins a 40 quilograms i tenen un nivell colesterol HDL superior a 42,54 mg/dl (veure Taula 6.10), els clústers es poden definir breument com:

- **Clúster 1**: Homes d'uns 67 anys, actius físicament, amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt.
- **Clúster 2**: Homes d'entre 67 i 70 anys, poc actius físicament, amb alt risc de patir malalties cardiovasculars i un índex de salut funcional baix.
- **Clúster 3**: Homes d'uns 70 anys, moderats quant a l'activitat física, amb un risc mitjà de patir malalties cardiovasculars i un índex de salut funcional baix.

L'anàlisi completa de cada clúster respecte a les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen home d'uns 67 anys, amb un índex de massa corporal normal ($18,5 \leq BMI \leq 24,9$), amb baix nivell de proteïna PCR (< 3 mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl). Tenen els nivells més alts de colesterol LDL i total, per sobre dels llindars recomanats (LDL > 116 mg/dl, total > 193,35 mg/dl), i són homes amb el nivell més alt de colesterol HDL. Practiquen molta activitat física, s'autoavaluen amb un bon nivell de salut (cap home considera que tingui un nivell de salut pobre) i presenten un índex de salut funcional alt (healthstatus > 50), el més alt entre tots els clústers.

El **clúster 2** el defineixen homes d'entre 67 i 70 anys, amb un índex de massa corporal classificat com obesitat de classe I ($30,0 \leq BMI \leq 34,9$), amb els nivells de glucosa en sang alts, estadísticament iguals al clúster 3, però que encara es manté a un nivell normal (< 100 mg/dl), amb un nivell de proteïna PCR per sobre del llindar recomanat (> 3 mg/l), els més alt en tots els clústers, i amb un nivell de triglicèrids al límit alt (valor entre 150 i 199 mg/dl). Presenten nivells de colesterol LDL i total per sobre del llindar recomanat, tot i que no tant com en el clúster 1, i són els homes amb el nivell de colesterol HDL més baix. Practiquen poca activitat física en comparació als altres grups, és el clúster on més homes s'autoavalua amb un nivell

de salut pobre i presenten un índex de salut funcional baix (healthstatus < 50), estadísticament igual al clúster 3.

El **clúster 3** el defineixen homes d'uns 70 anys, amb un índex de massa corporal classificat com preobesitat ($25{,}0 \leq \text{BMI} \leq 29{,}9$), amb nivells alts de glucosa en sang, estadísticament iguals al clúster 2, però que encara es manté a un nivell normal (< 100 mg/dl), amb nivells de proteïna PCR baixos, estadísticament iguals al clúster 1 (el valor mitjà és elevat a causa de l'existència d'*outliers*) i amb un nivell de triglicèrids normal (< 150 mg/dl), estadísticament igual al clúster 1. Presenten els nivells de colesterol LDL i total més baixos, per sota dels llindars establerts (LDL < 116 mg/dl, total < 193,35 mg/dl). Practiquen activitats físiques de manera més moderada, s'autoavaluen amb un nivell de salut bo o normal i presenten un índex de salut funcional estadísticament igual al clúster 2, però que en mitjana es troba per sobre de 50.

Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.2 de l'Annex referent al sexe masculí.

6.3. K-medoids

6.3.1. Dones

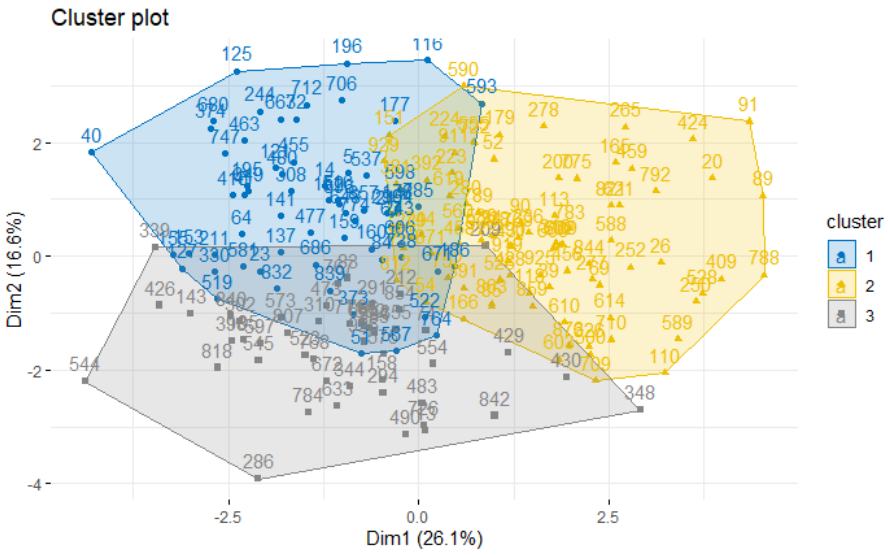
Per aplicar el mètode de partició K-medoids en la mostra de dones, tenint en compte el mètode K-means com a base de l'anàlisi, s'han dividit les dades en tres clústers per veure si la partició es distribueix de manera similar al K-means. A la Figura 6.7 es mostra com es reparteixen les dades en un espai factorial bidimensional a través de la funció **fviz_cluster**.

Pel mètode K-medoids els 3 clústers se solapen entre ells. Això és degut a les diferències entre els algoritmes del K-means i el K-medoids. També es deu per l'ús de medoids com a centres dels clústers en comptes de centroides.

Tot i no haver obtingut una partició tan diferenciada com al mètode K-means, a continuació s'ha procedit a fer l'anàlisi de *clustering* per veure si els perfils que s'extreuen són similars o equiparables entre els dos mètodes de partició. Per aquesta divisió de les dades, el clúster 1 compta amb 67 dones, el clúster 2 compta amb 49 dones i el clúster 3 compta amb 84 dones.

A causa del "*label switching problem*" (*M. Stephens, 2000*), a la Figura 6.7 el clúster 2 i el clúster 3 es troben intercanviats. Per fer més fàcil l'anàlisi, en les taules s'ha canviat els nivells d'aquests clústers per poder comparar-los de manera més clara a l'apartat 6.5.

Faula 6.7: Visualització del mètode K-medoids dividit en 3 clústers en un espai factorial bidimensional.



6.3.2. Profiling dones

A la Taula 6.13 es presenta el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original de dones.

Taula 6.13: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	24,13	82,95	99,51	24,37
2	25,61	86,41	102,22	23,63
3	31,88	101,20	116,13	24,53
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
124,97	70,98	85,00	120,83	64,30
142,99	78,99	82,14	122,37	78,84
136,10	76,83	91,71	147,30	61,96
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
121,90	210,43	2,17	55,65	67,09
169,36	272,51	3,51	50,52	68,59
119,46	210,61	5,43	46,81	68,38

A continuació, la Taula 6.14 mostra quines variables numèriques han presentat diferències significatives entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric, i el p-valor obtingut pels mètodes de comparació de mitjanes. També s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.14: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,2160	-	-
BMI	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	Kruskal-Wallis	0,6220	-	-
Pressió sistòlica	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Pressió diastòlica	ANOVA	<0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Glucosa	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Triglicèrids	Kruskal-Wallis	0,0082	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol HDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Colesterol LDL	ANOVA	<0,001	Tukey	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Colesterol total	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Proteïna PCR	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Healthstatus	ANOVA	<0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3

Tot seguit, per poder dur a terme el perfil dels clústers de manera completa, a la Taula 6.15 es mostra quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ell.

Taula 6.15: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,4209	-
Educació	0,0360	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Riquesa de la llar	0,0099	C1 ≠ C2 / C1 = C3 / C2 = C3
Soledat	0,7008	-
Activitat física	0,1722	-
Hipertensió arterial	<0,001	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Depressió	0,4063	-
Autoavaluació de salut	0,2014	-

En funció de les tres taules de l'anàlisi, tenint en compte que totes les dones presenten en mitjana una edat al voltant dels 68 anys, poden agafar pesos, aproximadament, d'entre 23 a 25 quilograms i tenen un nivell colesterol HDL superior a 46,40 mg/dl (veure Taula 6.13), els clústers es poden definir com:

- **Clúster 1:** Dones amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt, nivell adquisitiu mitjà-alt i estudis mitjans.
- **Clúster 2:** Dones amb un risc mitjà de patir malalties cardiovasculars amb un índex de salut funcional i nivell adquisitiu mitjà i estudis mitjans.

- **Clúster 3:** Dones amb alt risc de patir malalties cardiovasculars amb un índex de salut funcional baix, nivell adquisitiu baix i estudis primaris o inferiors.

L'anàlisi completa de cada clúster respecte a les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen dones amb un índex de massa corporal al límit del nivell normal (18,5 $\leq$ BMI $\leq$ 24,9) i la preobesitat (25,0 $\leq$ BMI $\leq$ 29,9), donat que són estadísticament iguals al clúster 2, amb el nivell més baix de proteïna PCR (< 3mg/l) i amb nivells de triglicèrids normals (< 150 mg/dl). Presenten els nivells de colesterol LDL i total per sobre del llindar establert (LDL > 116 mg/dl, total > 193,35 mg/dl). Són dones amb la pressió arterial alta (pressió sistòlica entre 120 i 129 mmHg i pressió diastòlica per sota de 80 mmHg), que majoritàriament s'han quedat en estudis secundaris, tenen una riquesa de la llar repartida de manera general entre el 2n, el 3r i el 5è quintil i presenten un índex de salut funcional alt (healthstatus > 50), el més alt d'entre tots els clústers.

El **clúster 2** presenta certes similituds amb el clúster 1 i el defineixen dones amb un índex de massa corporal al límit del nivell normal (18,5 $\leq$ BMI $\leq$ 24,9) i la preobesitat (25,0 $\leq$ BMI $\leq$ 29,9), donat que són estadísticament iguals al clúster 1, amb un nivell baix de proteïna PCR (< 3 mg/l), estadísticament igual al clúster 1, i amb un nivell de triglicèrids normal (< 150 mg/dl). És el clúster amb els nivells de colesterol LDL i total més elevats, també amb el nivell de colesterol HDL més gran. Són dones que pateixen hipertensió arterial entre l'etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i l'etapa 2 (pressió sistòlica igual o major a 140 mmHg), atès que presenten valors estadísticament iguals al clúster 3, generalment s'han quedat en els estudis secundaris o primaris, tenen una riquesa de la llar majoritàriament repartida entre el 1r i en el 4t quintil i presenten un índex de salut funcional mitjà (healthstatus $\cong$ 50).

El **clúster 3** el defineix dones amb un índex de massa corporal classificat com a obesitat de classe I (30,0 $\leq$ BMI $\leq$ 34,9), amb els nivells de glucosa en sang alts en comparació als altres clústers, però que encara es manté a un nivell normal (< 100 mg/dl), amb un nivell alt de proteïna PCR (> 3 mg/l) i amb un nivell de triglicèrids normal proper amb la frontera amb el límit alt ($\cong$ 150 mg/dl). Com en la resta de grups, presenten nivells de colesterol LDL i total per sobre del llindar recomanat. Es tracta de dones amb hipertensió arterial entre l'etapa 1 (pressió sistòlica entre 130 i 139 mmHg) i l'etapa 2 (pressió sistòlica igual o major a 140 mmHg), donat que presenten valors estadísticament iguals al clúster 2, que en general s'han quedat en l'educació primària o inferior, tenen una riquesa de la llar repartida, majoritàriament, entre el 1r i el 2n quintil i presenten un índex de salut funcional baix (healthstatus < 50), el més vaig d'entre tots els clústers.

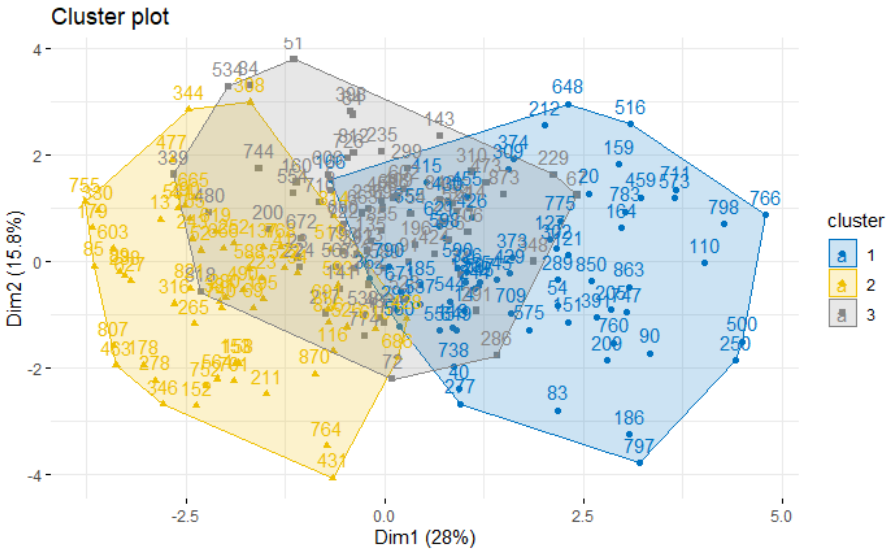
Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.3 de l'Annex referent al sexe femení.

6.3.3. Homes

En la mateixa línia que per la mostra de dones, per aplicar el mètode de partició K-medoids en la mostra d'homes s'han dividit les dades en tres clústers per veure si la partició es distribueix

de manera similar al mètode K-means. A la Figura 6.8 es veu com es reparteixen les dades en un espai factorial bidimensional a través de la funció **fviz_cluster**.

Faula 6.8: Visualització del mètode K-medoids dividit en 3 clústers en un espai factorial bidimensional.



Pel mètode K-medoids els 3 clústers se solapen entre ells, sobretot el tercer clúster. Com en la mostra de dones, això és degut a les diferències entre els algoritmes K-means i K-medoids i a l'ús de medoids com a centres dels clústers en comptes de centroides.

Tot i no haver obtingut una partició tan diferenciada com al mètode K-means, a continuació s'ha procedit a fer l'anàlisi de *clustering* per veure si els perfils que s'extreuen són similars o equiparables entre els dos mètodes de partició. Per aquesta divisió de les dades, el clúster 1 compta amb 62 homes, el clúster 2 compta amb 69 homes i el clúster 3 compta amb 69 homes.

Com ha passat amb la mostra de dones, a causa del "*label switching problem*", a la Figura 6.8 el clúster 1 i el clúster 2 es troben intercanviats. Per fer més fàcil l'anàlisi, en les taules s'ha canviat els nivells d'aquests clústers per poder comparar-los de manera més clara a l'apartat 6.5.

6.3.4. Profiling homes

A la Taula 6.16 es presenta el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original d'homes.

Seguidament, a la Taula 6.17 es presenta quines variables numèriques han resultat estadísticament diferents entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric, i el p-valor obtingut pels mètodes de comparació de mitjanes. També s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.16: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	24,36	91,96	100,39	38,40
2	31,26	111,54	113,01	38,01
3	27,20	99,74	105,15	42,16
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
132,06	74,62	87,59	115,86	65,30
142,45	77,14	97,10	151,72	51,00
131,93	75,02	89,77	132,86	49,60
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
145,95	234,01	1,76	54,54	67,97
115,00	196,21	6,21	48,05	69,80
102,78	178,83	1,79	53,54	67,55

Taula 6.17: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,0180	Dunn	C1 = C2 / C1 = C3 / C2 ≠ C3
BMI	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	Kruskal-Wallis	0,0014	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Pressió sistòlica	Kruskal-Wallis	0,0014	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Pressió diastòlica	ANOVA	0,3210	-	-
Glucosa	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Triglicèrids	Kruskal-Wallis	0,0067	Dunn	C1 ≠ C2 / C1 = C3 / C2 = C3
Colesterol HDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Colesterol LDL	ANOVA	<0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol total	ANOVA	<0,001	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Proteïna PCR	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Healthstatus	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3

Finalment, per poder dur a terme el perfil dels clústers de manera completa, a la Taula 6.18 es mostra quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.18: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,3676	-
Educació	0,0417	C1 ≠ C2 / C1 = C3 / C2 = C3
Riquesa de la llar	0,2700	-
Soledat	0,0974	-
Activitat física	0,0516	-
Hipertensió arterial	0,0196	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Depressió	0,4020	-
Autoavaluació de salut	0,0034	C1 ≠ C2 / C1 = C3 / C2 = C3

En funció de les tres taules de l'anàlisi, tenint en compte que tots els homes presenten un nivell colesterol HDL superior a 42,54 mg/dl (veure Taula 6.16), es poden definir els clústers breument com:

- **Clúster 1**: Homes d'entre 67 i 70 anys, amb risc de patir malalties cardiovasculars, amb alt nivell educatiu i que presenten un índex de salut funcional alt.
- **Clúster 2**: Homes d'uns 70 anys, amb alt risc de patir malalties cardiovasculars, amb baix nivell educatiu i que presenten un índex de salut funcional baix.
- **Clúster 3**: Homes d'uns 67 anys, amb risc de patir malalties cardiovasculars, amb un nivell mitjà d'educació i que presenten un índex de salut funcional alt.

L'anàlisi completa de cada clúster respecte a les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen homes d'entre 67 i 70 anys, amb un índex de massa corporal normal ($18,5 \leq BMI \leq 24,9$), amb un nivell de proteïna PCR baix (< 3 mg/l) i amb el nivell de triglicèrids més baix (< 150 mmHg). És el clúster amb els nivells de colesterol LDL i total més alts, per sobre dels llindars recomanats (LDL > 116 mg/dl i total > 193,35 mg/dl), també es tracta dels homes amb el nivell de colesterol HDL més alt. Poden agafar pesos d'aproximadament 38 kg i pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg). Majoritàriament s'han quedat en estudis secundaris o superiors, s'han autoavaluat amb bon nivell de salut, cap home s'ha autoavaluat amb un nivell de salut dolent, i presenten un índex de salut funcional alt (healtstatus > 50), el més alt d'entre tots els clústers però estadísticament igual al clúster 3.

El **clúster 2** el defineixen homes d'uns 70 anys, amb un índex de massa corporal classificat com obesitat de classe I ($30,0 \leq BMI \leq 34,9$), amb els nivells de glucosa en sang alts en comparació als altres clústers, però que encara es manté a un nivell normal (< 100 mg/dl), amb nivells de proteïna PCR molt per sobre del nivell recomanat (> 3 mg/l) i amb un nivell de triglicèrids pel límit alt (valors entre 150 i 199 mmHg). Tenen els nivells de colesterol LDL molt propers al límit i els nivells de colesterol total per sobre del llindar establert (LDL $\cong$ 116 mg/dl, total > 193,35 mg/dl). Són homes que poden agafar pesos d'aproximadament 38 quilograms, pateixen hipertensió arterial en etapa 2 (pressió sistòlica igual o major a 140 mmHg), majoritàriament s'han quedat en estudis primaris i secundaris, sent el clúster que pitjor s'ha autoavaluat respecte al seu estat de salut i presenten un índex de salut funcional baix (healtstatus < 50), el més baix d'entre tots els clústers.

El **clúster 3** el defineixen homes d'uns 67 anys, amb un índex de massa corporal classificat com a preobesitat (24,5 ≤ BMI ≤ 29,9), amb baix nivell de proteïna PCR (< 3 mg/l) i amb un nivell de triglicèrids normal (< 150 mg/dl) que es troba entre el clúster 1 i el clúster 2. Tenen els nivells de colesterol LDL i total per sota dels llindars establerts (LDL < 116 mg/dl i total < 193,35 mg/dl), ambdós colesterols representen els nivells més baixos d'entre tots els clústers, i presenten el nivell més baix de colesterol HDL. Poden aixecar pesos d'aproximadament 42 kg, pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg), en general s'han quedat als estudis secundaris, s'han autoavaluat entre un nivell bo i mitja de salut i presenten un índex de salut funcional alt (healtstatus > 50) estadísticament iguals al clúster 1.

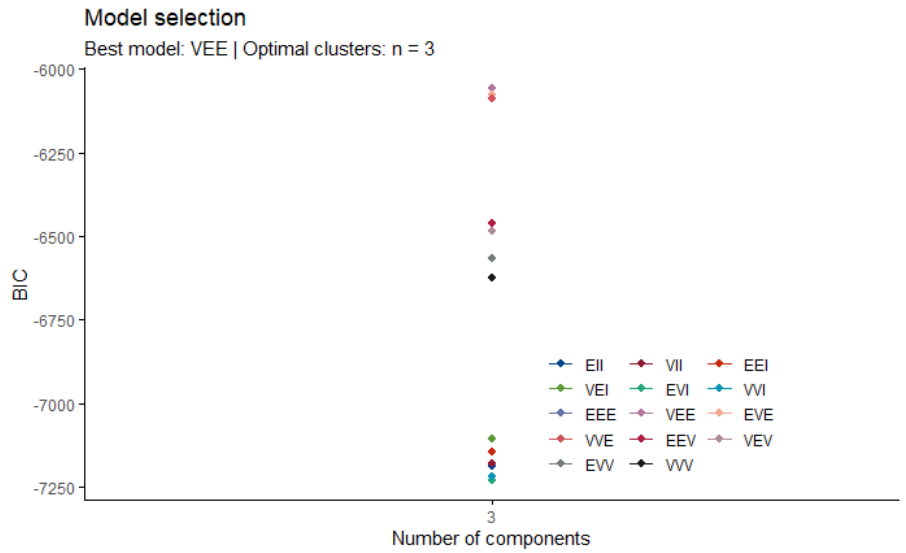
Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.3 de l'Annex referent al sexe masculí.

6.4. Gaussian Mixture Models

6.4.1. Dones

Per dur a terme l'anàlisi de la mostra de dones, s'ha forçat al mètode dels *Gaussian Mixture Models* (GMMs) a seleccionar el millor model possible per tres clústers a través de l'opció "G = 3" de la funció **Mclust**. Pel criteri del BIC, el millor model que ha extret l'algoritme ha estat el VEE (veure subapartat 2.4.1), el model amb distribució el·lipsoidal, volum variable, forma igual i orientació igual pel conjunt de clústers (veure Figura 6.9).

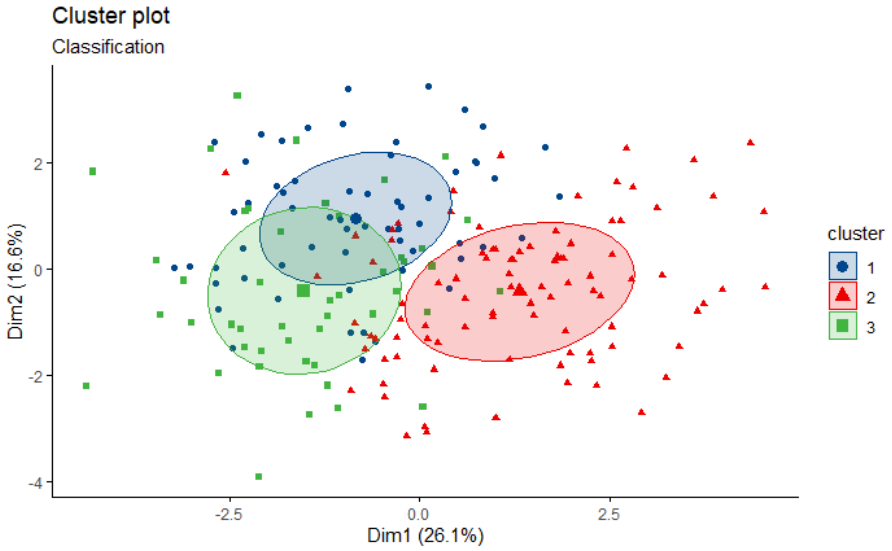
Figura 6.9: Gràfic del criteri BIC per la divisió en 3 clústers on s'indica quin és el model més òptim.



A la figura 6.10 es pot observar la classificació dels tres clústers en un espai factorial bidimensional. En el gràfic es veu la representació de les observacions i la silueta de cada

clúster. A causa del "*label switching problem*", a les figures 6.10 i 6.11 el clúster 2 i el clúster 3 es troben intercanviats. Per fer més fàcil l'anàlisi, en les taules s'ha canviat els nivells d'aquests clústers per poder comparar-los de manera més clara a l'apartat 6.5.

Figura 6.10: Representació dels clústers i les seves observacions en un espai factorial bidimensional.



Els clústers, sobretot el clúster 1 i el clúster 3 (clúster 2), estan molt solapats entre ells i es pot observar com les dades han quedat barrejades entre elles, això es deu al fet que tot i haver normalitzat les variables amb les quals s'han aplicat tots els mètodes de *clustering* a través de la funció **bestNormalize**, per la mostra de dones, les variables de la **força d'adherència**, la **glucosa**, els **triglicèrids**, el **colesterol HDL** i l'índex de salut **healthstatus** no s'han acabat de normalitzar i no presenten una distribució Gaussiana. Tot i això s'ha decidit aplicar el mètode dels GMMs per veure si els clústers que s'obtenen presenten un perfil similar al mètode K-means.

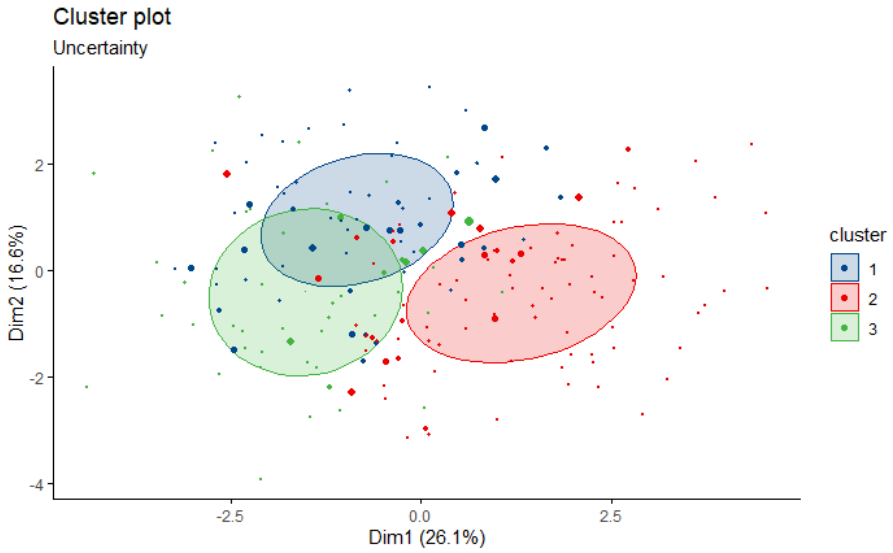
A la Figura 6.11 es mostra la incertesa de les observacions per a cada clúster. Aquelles observacions on els punts són més grans determinen les dones amb un alt grau d'incertesa a l'hora de pertànyer al clúster al qual han quedat assignades.

Es pot apreciar que els clústers que presenten observacions amb una incertesa elevada són el clúster 1, molt solapat amb el clúster 3 (clúster 2), i el clúster 2 (clúster 3), ja que és el que presenta més individus.

Tenint en compte l'etiquetatge dels clústers correctament, el clúster 1 està format per 61 dones, el clúster 2 per 46 dones i el clúster 3 per 93 dones (veure Figura 6.11).

S'ha calculat que existeixen vuit observacions amb una probabilitat d'incertesa per sobre de 0,4, d'aquestes només tres tenen una probabilitat d'incertesa per sobre de 0,5 i l'observació més incerta té associada una probabilitat de 0,57.

Figura 6.11: Representació del grau d'incertesa de cada observació en funció del clústers al qual pertanyen.



6.4.2. Porfiling dones

A la Taula 6.19 es presenta el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original de dones.

Taula 6.19: Mitjanes de les variables numèriques originals separades per

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	25,39	85,08	101,30	24,70
2	24,61	83,45	100,73	21,57
3	30,85	99,62	114,18	25,30
Pressió sistòlica	Pressió diastòlica	Glucosa	Triglicèrids	Colesterol HDL
128,30	71,97	88,41	127,34	61,05
130,30	72,97	85,15	104,94	87,09
139,69	78,85	87,25	149,14	60,71
Colesterol LDL	Colesterol total	Proteïna PCR	Healthstatus	Edat
119,43	205,90	1,17	52,74	67,41
138,54	246,40	2,68	50,25	70,24
138,09	228,48	6,22	49,55	67,28

A continuació, per comprovar que existeixen diferències significatives entre les mitjanes de cada clúster, a la Taula 6.20 es presenten les variables que han resultat estadísticament diferents, el mètode que s'ha aplicat, paramètric o no paramètric, i el p-valor obtingut pels mètodes de comparació de mitjanes. També s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.20: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,0039	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
BMI	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència cintura	ANOVA	<0,001	Tukey	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Força d'adherència	Kruskal-Wallis	0,0028	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Pressió sistòlica	Kruskal-Wallis	<0,001	Dunn	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Pressió diastòlica	ANOVA	<0,001	Tukey	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Glucosa	Kruskal-Wallis	0,0466	Dunn	C1 ≠ C2 / C1 = C3 / C2 = C3
Triglicèrids	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Colesterol HDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Colesterol LDL	Kruskal-Wallis	0,0040	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Colesterol total	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Proteïna PCR	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Healthstatus	Kruskal-Wallis	0,1050	-	-

Tot seguit, per tal de dur a terme els perfils dels clústers de manera completa, a la Taula 6.21 es mostra quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.21: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,0612	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Educació	0,3615	-
Riquesa de la llar	0,3573	-
Soledat	0,0246	C1 = C2 / C1 = C3 / C2 ≠ C3
Activitat física	0,7619	-
Hipertensió arterial	<0,001	C1 = C2 / C1 ≠ C3 / C2 ≠ C3
Depressió	0,8128	-
Autoavaluació de salut	0,0830	-

En funció de les tres taules de l'anàlisi, tenint en compte que presenten un índex de salut mitjà (healthstatus ≅ 50) en el conjunt dels clústers, tot i que observant la Taula 6.19 sembla que les dones del clúster 1 presenten el millor índex de salut i les dones del clúster 2 el pitjor índex de salut, i tenen uns nivells de colesterol HDL per sobre de 46,40 mg/dl (veure Taula 6.19), els clústers es poden definir breument com:

- **Clúster 1:** Dones d'uns 67 anys, amb risc de patir malalties cardiovasculars i que en mitjana presenten l'índex de salut més alt.
- **Clúster 2:** Dones d'uns 70 anys, amb un risc mitjà de patir malalties cardiovasculars, un índex de salut mitjà i que s'han sentit més soles.
- **Clúster 3:** Dones d'uns 67 anys, amb alt risc de patir malalties cardiovasculars i un índex de salut mitjà.

L'anàlisi completa de cada clúster respecte a les tres taules que s'han presentat és el següent:

El **clúster 1** el defineixen dones d'uns 67 anys, amb un índex de massa corporal al límit del nivell normal (18,5 $\leq$ BMI $\leq$ 24,9) i la preobesitat (25,0 $\leq$ BMI $\leq$ 29,9), donat que són estadísticament iguals al clúster 2, amb els nivells de glucosa en sang alts en comparació als altres clústers, però que encara es manté a un nivell normal (< 100 mg/dl), amb un nivell baix de proteïna PCR (< 3 mg/l) i nivells de triglicèrids normals (< 150 mg/dl). Són el clúster amb els nivells de colesterol LDL i total mes baixos, tot i això, es troben per sobre dels llindars establerts (LDL > 116 mg/dl i total > 193,35 mg/dl). Es tracta de dones al límit entre la pressió arterial alta (pressió sistòlica entre 120 i 129 mmHg i pressió diastòlica per sota de 80 mmHg) i la hipertensió en etapa 1 (pressió sistòlica entre 130 i 139 mmHg), ja que presenten valors estadísticament iguals al clúster 2, poden agafar un pes d'uns 25 quilograms, majoritàriament estan casades i que en general no s'han sentit soles.

El **clúster 2** el defineixen dones d'uns 70 anys, amb un índex de massa corporal al límit del nivell normal (18,5 $\leq$ BMI $\leq$ 24,9) i la preobesitat (25,0 $\leq$ BMI $\leq$ 29,9), donat que són estadísticament iguals al clúster 1, amb nivells de proteïna PCR menor al llindar recomanat (< 3 mg/l) tot i que es troben molt properes al límit i amb nivells de triglicèrids normal (< 150 mg/dl), el més baix d'entre tots els clústers. Presenten els nivells de colesterol LDL i total més elevats, per sobre dels llindars recomanats i estadísticament iguals al clúster 2, tanmateix també presenten els nivells més alts amb diferència de colesterol HDL. Són dones que poden agafar uns 22 kg, el pes més petit respecte els altres clústers, es troben al límit entre la pressió arterial alta (pressió sistòlica entre 120 i 129 mmHg i pressió diastòlica per sota de 80 mmHg) i la hipertensió en etapa 1 (pressió sistòlica entre 130 i 139 mmHg), ja que presenten valors estadísticament iguals al clúster 1. Es tracta de dones que majoritàriament estan divorciades o vídues i són el clúster on se senten més soles.

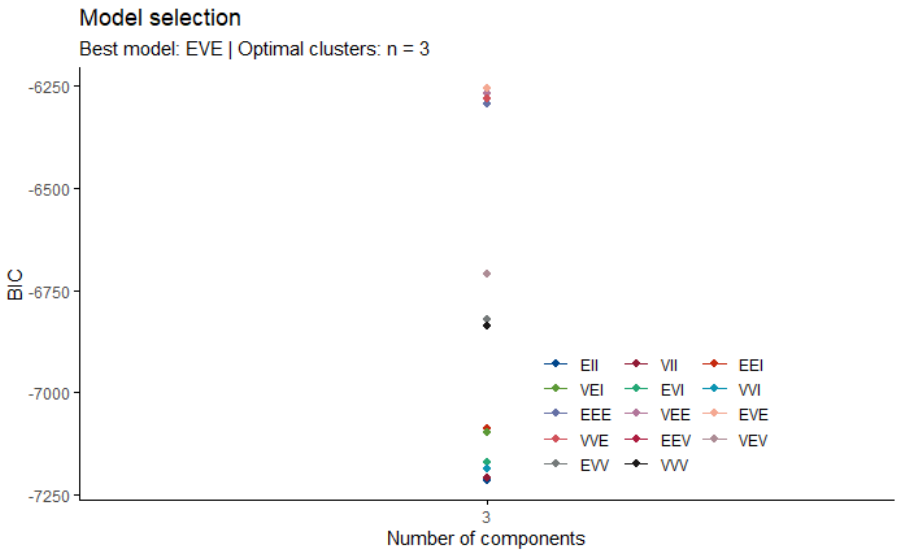
El **clúster 3** el defineixen dones d'uns 67 anys, amb un índex de massa corporal classificat com a obesitat de classe I (30,0 $\leq$ BMI $\leq$ 34,9), amb nivells de proteïna PCR preocupants, molt per sobre del llindar recomanat (< 3 mg/l), i amb nivells de triglicèrids normals (< 150 mg/dl) estadísticament iguals al clúster 1, que tot i presentar una mitjana propera al límit alt això es deu al fet que la majoria d'*outliers* del nivell de triglicèrids estan situats en aquest clúster. Presenten nivells de colesterol LDL i total per sobre dels llindars recomanats, amb uns valors una mica més alts que en el clúster 1 i estadísticament iguals al clúster 3. Són dones que poden agafar uns 25 kg i que pateixen hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg) al límit de l'etapa 2 i poden agafar pesos d'aproximadament 25 quilograms. Es tracta de dones que majoritàriament estan casades i que en general no s'han sentit soles.

Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.4 de l'Annex referent al sexe femení.

6.4.3. Homes

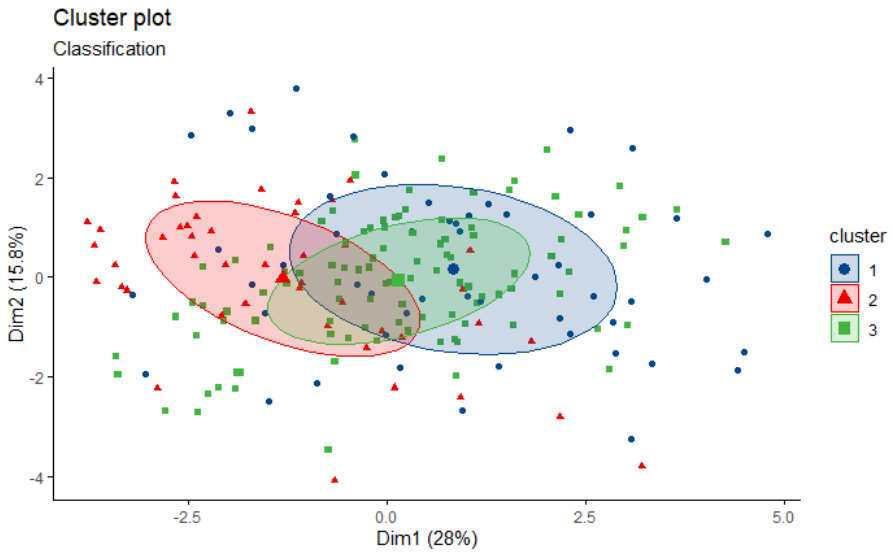
Per dur a terme l'anàlisi per la mostra d'homes, també s'ha forçat al mètode dels *Gaussian Mixture Models* (GMMs) a seleccionar el millor model possible per tres clústers a través de l'opció "G = 3" de la funció **Mclust**. Pel criteri del BIC, el millor model que ha extret l'algoritme ha estat l'EVE (veure subapartat 2.4.1), el model amb distribució el·lipsoidal, volum igual, forma variable i orientació igual pel conjunt de clústers (veure Figura 6.12)

Figura 6.12: Gràfic del criteri BIC per la divisió en 3 clústers on s'indica quin és el model més òptim.



A la Figura 6.13 es pot observar la classificació dels tres clústers en un espai factorial bidimensional. En el gràfic es veu la representació de les observacions i la silueta de cada clúster.

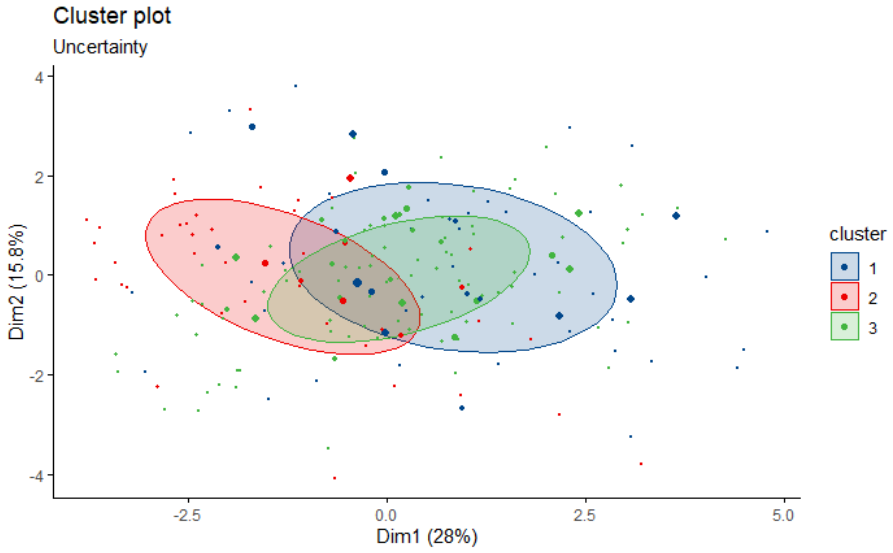
Figura 6.13: Representació dels clústers i les seves observacions en un espai factorial bidimensional.



Per la mostra d'homes, tots els clústers estan solapats entre ells, sobretot el clúster 3 que es troba just al mig entre el clúster 1 i el clúster 2, i es pot observar com les dades han quedat barrejades entre elles. Això, com ha succeït amb la mostra de dones, és degut al fet que tot i haver normalitzat les variables amb les quals s'han aplicat tots els mètodes de *clustering* a través de la funció **bestNormalize**, per la mostra d'homes, les variables de l'**índex de massa corporal**, la **pressió diastòlica**, la **glucosa**, el **colesterol HDL**, la **proteïna PCR** i l'índex de salut **healthstatus** no s'han acabat de normalitzar i no presenten una distribució Gaussiana. No obstant això, sense tenir una mostra amb el conjunt de variables Gaussianes, s'ha decidit aplicar el mètode dels GMMs per veure si els clústers que s'obtenen presenten un perfil similar al mètode K-means.

La Figura 6.14 mostra la incertesa de les observacions per a cada clúster. Aquelles observacions on els punts són més grans determinen els homes amb un alt grau d'incertesa a l'hora de pertànyer al clúster al qual han quedat assignats.

Figura 6.14: Representació del grau d'incertesa de cada observació en funció del clústers al qual pertanyen.



En aquest cas, tots els clústers presenten observacions amb una incertesa elevada, sobretot en el clúster 1 i en el clúster 3 que és el que està més solapat de tots i és el que conté més individus. En concret, el clúster 1 està format per 52 homes, el clúster 2 per 45 homes i el clúster 3 per 103 homes.

S'ha calculat que existeixen sis observacions amb una probabilitat d'incertesa per sobre de 0,4, d'aquestes només una observació té una probabilitat d'incertesa per sobre de 0,5, l'observació més incerta, que pren una probabilitat de 0,52.

6.4.4. Profiling homes

A la Taula 6.22 es presenta el càlcul de les mitjanes de totes les variables numèriques de la base de dades introduint la variable indicadora dels clústers a la mostra original d'homes.

Taula 6.22: Mitjanes de les variables numèriques originals separades per clústers.

Clústers	BMI	Circumferència cintura	Circumferència maluc	Força d'adherència
1	29,43	104,48	108,37	36,48
2	25,14	95,33	102,66	41,64
3	27,99	102,50	107,02	40,21
Pressió sistòlica	**Pressió diastòlica**	**Glucosa**	**Triglicèrids**	**Colesterol HDL**
130,46	74,33	94,60	142,22	47,82
134,93	76,77	84,76	147,22	65,82
138,48	75,79	93,11	124,26	53,80
Colesterol LDL	**Colesterol total**	**Proteïna PCR**	**Healthstatus**	**Edat**
116,46	192,68	4,65	50,48	69,92
117,98	213,29	2,18	52,20	66,53
123,41	201,65	3,11	52,60	68,55

Seguidament, a la Taula 6.23 es presenta quines variables numèriques han resultat estadísticament diferents entre clústers, el mètode que s'ha aplicat, paramètric o no paramètric, i el p-valor obtingut pels mètodes de comparació de mitjanes. També s'hi presenta l'anàlisi post-hoc amb el test que s'ha aplicat per determinar, per cada una de les variables significatives, quins clústers són diferents i/o similars entre ells.

Taula 6.23: Anàlisi de quines variables numèriques són estadísticament significatives entre clústers i, per les variables en qüestió, anàlisi post-hoc.

Diferència de mitjanes			Anàlisi post-hoc	
Variable	Test	P-Valor	Test	Clústers
Edat	Kruskal-Wallis	0,0028	Dunn	C1 ≠ C2 / C1 = C3 / C2 = C3
BMI	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Circumferència cintura	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Circumferència maluc	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Força d'adherència	ANOVA	0,0020	Tukey	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Pressió sistòlica	Kruskal-Wallis	0,0570	-	-
Pressió diastòlica	ANOVA	0,5050	-	-
Glucosa	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 = C3 / C2 ≠ C3
Triglicèrids	Kruskal-Wallis	0,7840	-	-
Colesterol HDL	Kruskal-Wallis	<0,001	Dunn	C1 ≠ C2 / C1 ≠ C3 / C2 ≠ C3
Colesterol LDL	ANOVA	0,4420	-	-
Colesterol total	ANOVA	0,0432	Tukey	C1 ≠ C2 / C1 = C3 / C2 = C3
Proteïna PCR	Kruskal-Wallis	0,0038	Dunn	C1 ≠ C2 / C1 = C3 / C2 = C3
Healthstatus	Kruskal-Wallis	0,1990	-	-

Finalment, per poder dur a terme el perfil dels clústers de manera completa, a la Taula 6.24 es mostra quines variables categòriques presenten freqüències diferents en funció dels clústers a través del test d'independència Chi-quadrat i, per aquelles variables que han indicat dependència, s'han calculat les diferències dos a dos entre cada un dels clústers per determinar quins grups són similars i/o diferents entre ells.

Taula 6.24: Anàlisi de les variables categòriques estadísticament diferents entre clústers i, per les variables en qüestió, anàlisi de quins clústers son similars i/o diferents entre ells.

Variable	P-Valor	Clústers
Estat civil	0,6892	-
Educació	0,0989	-
Riquesa de la llar	0,7137	-
Soledat	0,1467	-
Activitat física	0,6566	-
Hipertensió arterial	0,0421	C1 ≠ C2 / C1 ≠ C3 / C2 = C3
Depressió	0,4673	-
Autoavaluació de salut	0,6468	-

En funció de les tres taules de l'anàlisi, tenint en compte que tots els clústers presenten un índex de salut mitjà tirant cap a alt (healthstatus ≥ 50), la majoria d'homes pateix hipertensió arterial en etapa 1 (pressió sistòlica entre 130 i 139 mmHg), tenen un nivell de triglicèrids normal (< 150 mg/dl), els nivells de colesterol LDL per sobre del llindar recomanat (> 116 mg/dl) i presenten un nivell colesterol HDL superior a 42,54 mg/dl (veure Taula 6.22), els clústers es poden definir breument com:

- **Clúster 1:** Homes d'uns 70 anys, amb alt risc de patir malalties cardiovasculars i una força d'adherència menor als altres clústers.
- **Clúster 2:** Homes d'uns 67 anys amb risc de patir malalties cardiovasculars i força d'adherència elevada.
- **Clúster 3:** Homes d'entre 67 i 70 anys amb risc mitjà de patir malalties cardiovasculars i força d'adherència elevada.

L'anàlisi completa de cada clúster respecte a les taules que s'han presentat és el següent:

El **clúster 1** el defineixen homes d'uns 70 anys, amb un índex de massa corporal classificat com a preobesitat ($25{,}0 \leq BMI \leq 29{,}9$), amb el nivell de glucosa en sang alt, estadísticament iguals al clúster 3, però que encara es manté a un nivell normal (< 100 mg/dl), i amb nivells de proteïna PCR per sobre del llindar màxim recomanat (> 3 mg/l). Presenten el nivell de colesterol total més baix, gairebé igual al llindar establert ($\cong$ 193,35 mg/dl), amb el nivell més baix de colesterol HDL i poden agafar pesos d'aproximadament 37 quilograms.

El **clúster 2** el defineixen homes d'uns 67 anys, amb un índex de massa corporal al límit del nivell normal ($18{,}5 \leq BMI \leq 24{,}9$) i la preobesitat ($25{,}0 \leq BMI \leq 29{,}9$) i amb un nivell de proteïna PCR baix (< 3 mg/l). Tenen els nivells de colesterol total per sobre del llindar recomanat (> 193,35 mg/dl), els més grans d'entre tots els clústers, tot i això presenten els valors més alts de colesterol HDL i poden agafar pesos d'aproximadament 41 quilograms.

El **clúster 3** el defineixen homes d'entre 67 i 70 anys aproximadament, amb un índex de massa corporal classificat com a preobesitat ($25{,}0 \leq BMI \leq 29{,}9$), amb el nivell de glucosa en sang alt, estadísticament igual al clúster 1, però que encara es manté a un nivell normal (< 100 mg/dl), i amb nivells de proteïna PCR per sobre del límit recomanat (> 3 mg/l) entremig dels valors del clúster 1 i el clúster 2. Presenten uns nivells de colesterol total per sobre del llindar establert i poden aixecar pesos d'aproximadament 41 quilograms.

Els gràfics de caixa i els gràfics de barres de les variables estadísticament diferents entre clústers es troben al subapartat 9.1.4 de l'Annex referent al sexe masculí.

6.5. Comparació profilings

6.5.1. Dones

A continuació, es presenta una taula resum dels quatre mètodes de *clustering* aplicats respecte a les taules 6.1, 6.7, 6.13 i 6.19 de l'anàlisi. A la taula es presenten les mitjanes de les variables numèriques en funció dels clústers.

Taula 6.25: Taula resum amb les mitjanes de les variables numèriques en funció dels clúster per cada mètode aplicat.

Mètodes	K-means			Jeràrquic			K-medoids			GMMs		
Clústers	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3
BMI	24,8	25,2	32,9	25,0	25,2	33,1	24,1	25,6	31,9	25,4	24,6	30,9
Circumferència cintura	84,5	85,4	103,5	84,9	85,4	104,0	83,0	86,4	101,2	85,1	83,5	99,6
Circumferència maluc	101,2	101,6	117,9	101,4	101,6	118,4	99,5	102,2	116,1	101,3	100,7	114,2
Força d'adherència	24,5	23,6	24,7	24,5	23,6	24,7	24,4	23,6	24,5	24,7	21,6	25,3
Pressió sistòlica	138,8	124,9	138,7	138,5	124,9	139,0	125,0	143,0	136,1	128,3	130,3	139,7
Pressió diastòlica	77,0	72,1	77,2	77,1	72,1	77,1	71,0	79,0	76,8	72,0	73,0	78,9
Glucosa	83,6	86,0	91,3	84,4	86,0	90,9	85,0	82,1	91,7	88,4	85,2	87,3
Triglicèrids	125,3	113,5	156,9	127,0	113,5	156,7	120,8	122,4	147,3	127,3	104,9	149,1
Colesterol HDL	77,8	64,4	59,6	76,8	64,4	59,7	64,3	78,8	62,0	61,1	87,1	60,7
Colesterol LDL	166,7	105,4	128,6	166,2	105,4	127,3	121,9	169,4	119,5	119,4	138,5	138,1
Colesterol total	269,3	192,6	219,3	268,2	192,6	218,2	210,4	272,5	210,6	205,9	246,4	228,5
Proteïna PCR	2,57	2,03	6,79	2,63	2,03	6,93	2,17	3,51	5,43	1,17	2,68	6,22
Healthstatus	52,4	53,7	46,2	52,6	53,7	45,7	55,7	50,5	46,8	52,7	50,3	49,6
Edat	68,2	67,9	67,9	67,9	67,9	68,2	67,1	68,6	68,4	67,4	70,2	67,3

Un cop fets tots els *profilings* per la mostra de dones, s'ha pogut veure que el mètode K-means i el *clustering* jeràrquic aglomeratiu s'han definit de manera idèntica. Donat que presenten les mateixes característiques per a cada clúster. En aquests dos mètodes, totes les dones presenten una edat mitjana d'uns 68 anys, poden agafar pesos d'entre 23 a 25 quilograms i els perfils s'han definit breument com:

- **Clúster 1:** Dones amb risc de partir malalties cardiovasculars però amb un alt índex de salut funcional.
- **Clúster 2:** Dones sanes amb un alt índex de salut funcional.
- **Clúster 3:** Dones amb alt risc de patir malalties cardiovasculars que presenten un índex de salut funcional baix.

La diferència més clara es troba entre el clúster 2 i el clúster 3, ja que es tracta dels clústers més polaritzats. Per una banda, les dones del clúster 2, com a tret negatiu, només presenten una pressió arterial alta. Per altra banda, les dones del clúster 3 pateixen hipertensió arterial en etapa 1, tenen uns nivells de proteïna PCR elevats, presenten obesitat de classe I, es troben

al nivell més alt de triglicèrids en comparació als altres clústers i se'ls hi associa un índex de salut funcional baix.

En canvi, el clúster 1 és molt similar al clúster 2. Tot i això, com que es tracta de dones amb un colesterol elevat i que pateixen hipertensió arterial en etapa 1, característiques comunes amb el clúster 3, fa que existeixi certa diferència entre aquests dos grups. Per aquest motiu, tot hi compartir un índex de salut funcional alt, les dones del clúster 1 presenten un cert risc de partir malalties cardiovasculars, cosa que a les dones del clúster 2 no els hi passa.

Pel mètode K-medoids els perfils dels clústers són similars al K-means i al *clustering* jeràrquic aglomeratiu, però no acaben de ser el mateix. Igual que els dos mètodes anteriors, totes les dones presenten una edat mitjana d'uns 68 anys i poden agafar pesos d'entre 23 a 25 quilograms. Els perfils dels clústers s'han definit breument com:

- **Clúster 1:** Dones amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt, nivell adquisitiu mitjà-alt i estudis mitjans.
- **Clúster 2:** Dones amb un risc mitjà de patir malalties cardiovasculars, amb un índex de salut funcional i nivell adquisitiu mitjà i estudis mitjans.
- **Clúster 3:** Dones amb alt risc de patir malalties cardiovasculars, amb un índex de salut funcional baix, nivell adquisitiu baix i estudis primaris o inferiors.

A diferència del mètode K-means i el mètode jeràrquic, entren en joc dues variables categòriques que són diferents entre clústers i proporcionen informació sociodemogràfica per a cada grup. També es pot veure que, pel mètode K-medodis, totes les dones presenten un cert risc de patir malalties cardiovasculars i, a diferència del K-means i el *clustering* jeràrquic, el clúster de dones sanes no queda definit. Tot i això, sense tenir en compte les característiques sociodemogràfiques, el clúster 1 pel K-medoids s'equipara al clúster 1 dels mètodes K-means i jeràrquic, mentre que el clúster 3 d'aquest mètode s'equipara al clúster 3 dels altres dos mètodes, que representen dos dels clústers més polaritzats.

Així doncs, el mètode K-medoids proporciona informació addicional en comparació al mètode K-means i jeràrquic. Dona informació sociodemogràfica sobre el nivell educatiu i el poder adquisitiu de cada clúster. A més a més, troba clústers equiparables (clúster 1 i clúster 3) entre les diferents tècniques de *clustering*, però a diferència dels dos primers mètodes exposats, no distingeix el clúster de dones sanes. En canvi defineix el clúster 2 que es troba a mig camí entre el clúster 1 i el clúster 3, ja que per una banda són dones que pateixen hipertensió arterial en etapa 1 (com en el clúster 3), tenen valors elevats de colesterol (com en el clúster 1 i 3), presenten nivells de proteïna PCR baixos (com en el clúster 1) i com a tret diferenciador, se'ls hi associa un índex de salut funcional mitjà.

Per últim, els perfils del mètode dels *Gaussian Mixture Models* són els que presenten més diferències respecte a la resta de mètodes. Per aquest cas, totes les dones presenten un índex de salut funcional mitjà, estadísticament igual per a tots els grups, i els perfils dels clústers s'han definit breument com:

- **Clúster 1:** Dones d'uns 67 anys, amb risc de patir malalties cardiovasculars i que en mitjana presenten l'índex de salut funcional més alt.

- **Clúster 2:** Dones d'uns 70 anys, amb un risc mitjà de patir malalties cardiovasculars, un índex de salut funcional mitjà i que s'han sentit més soles.
- **Clúster 3:** Dones d'uns 67 anys, amb alt risc de patir malalties cardiovasculars i un índex de salut funcional mitjà.

Com en el mètode K-medoids, entren en joc dues variables categòriques que presenten diferències entre grups i proporcionen informació sobre l'estat civil de les dones, variable sociodemogràfica, i sobre si les dones s'han sentit soles, variable d'estil de vida i salut. També, es pot apreciar que totes les dones presenten un cert risc de patir malalties cardiovasculars i, a diferència dels altres mètodes, es distingeixen les dones per franges d'edat.

L'únic clúster que sembla ser equiparable amb la resta de mètodes és el clúster 3, el que presenta pitjors característiques, atès que s'assembla al clúster 3 dels mètodes K-means, jeràrquic i K-medoids. El clúster 2 presenta les dones que s'han sentit més soles, això es deu al fet que és el grup, que per aquest mètode, classifica les dones divorciades i vídues.

En definitiva, considerant la partició de la mostra de dones en tres grups, el mètode K-means i el jeràrquic perfilen els clústers de manera idèntica. El mètode K-medoids és equiparable als clústers del K-means i el mètode Jeràrquic, però no arriba a discernir el clúster de dones sanes, i el mètode dels GMMs és el que presenta els perfils més diferents.

Tenint en compte aquesta informació i observant les figures 6.2, 6.5, 6.7 i 6.10 on es mostra la divisió dels clústers per a cada mètode, els millors mètodes per trobar perfils diferenciats per aquesta mostra són el K-means i el *clustering* jeràrquic aglomeratiu. Pel mètode K-medoids i els GMMs els clústers presenten solapament entre ells i per aquest motiu, tot i trobar informació addicional respecte als altres dos mètodes, no acaben d'extreure els mateixos perfils i la informació analitzada pel mètode K-means i jeràrquic queda, en certa manera, difuminada a causa de la barreja que presenten els clústers a l'espai de les dades.

6.5.2. Homes

Seguint la línia del subapartat anterior, es presenta una taula resum dels quatre mètodes de *clustering* aplicats respecte a les taules 6.4, 6.10, 6.16 i 6.22 de l'anàlisi. A la taula es presenten les mitjanes de les variables numèriques en funció dels clústers.

Taula 6.26: Taula resum amb les mitjanes de les variables numèriques en funció dels clúster per cada mètode aplicat.

Mètodes	K-means			Jeràrquc			K-medoids			GMMs		
Clústers	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3
BMI	24,7	31,9	27,3	24,6	31,9	27,2	24,4	31,3	27,2	29,4	25,1	28,0
Circumferència cintura	93,1	113,0	100,2	92,7	112,9	100,0	92,0	111,5	99,7	104,5	95,3	102,5
Circumferència maluc	101,1	113,6	105,7	100,7	113,8	105,6	100,4	113,0	105,2	108,4	102,7	107,0
Força d'adherència	40,2	39,2	39,2	39,7	39,7	39,3	38,4	38,0	42,2	36,5	41,6	40,2
Pressió sistòlica	133,4	140,6	133,5	133,2	139,0	135,0	132,1	142,5	131,9	130,5	134,9	138,5
Pressió diastòlica	75,5	77,2	74,3	75,3	76,7	75,0	74,6	77,1	75,0	74,3	76,8	75,8
Glucosa	85,5	95,7	94,8	85,4	95,3	94,7	87,6	97,1	89,8	94,6	84,8	93,1
Triglicèrids	124,7	173,9	108,4	123,5	174,5	109,1	115,9	151,7	132,9	142,2	147,2	124,3
Colesterol HDL	62,5	48,8	52,2	63,1	48,6	52,2	65,3	51,0	49,6	47,8	65,8	53,8
Colesterol LDL	144,1	120,8	93,4	144,0	121,8	95,1	146,0	115,0	102,8	116,5	118,0	123,4
Colesterol total	231,1	204,2	167,2	231,5	205,1	169,1	234,0	169,2	178,8	192,7	213,3	201,7
Proteïna PCR	1,69	4,55	3,97	1,69	4,45	3,93	1,76	6,21	1,79	4,65	2,18	3,11
Healthstatus	55,0	49,0	51,3	55,2	49,2	51,1	54,5	48,1	53,5	50,5	52,2	52,6
Edat	67,4	68,6	69,5	67,4	68,4	69,6	68,0	69,8	67,6	69,9	66,5	68,6

Un cop analitzats tots els mètodes, s'ha pogut veure que per la mostra d'homes, com ha passat amb la mostra de dones, el mètode K-means i el *clustering* jeràrquic aglomeratiu s'han definit pràcticament igual, atès que presenten les mateixes característiques per a cada clúster. L'única diferència entre els mètodes són les assumpcions generals dels *profilings*. El mètode K-means considera que la majoria d'homes presenten hipertensió arterial entre l'etapa 1 i l'etapa 2 i poden agafar pesos d'uns 40 quilograms. En canvi, el *clustering* jeràrquic aglomeratiu considera que la majoria d'homes presenten només hipertensió arterial en etapa 1 i poden agafar pesos d'uns 40 quilograms. Tot i aquesta diferència entre mètodes els perfils s'han definit breument com:

- **Clúster 1**: Homes d'uns 67 anys, actius físicament, amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt.
- **Clúster 2**: Homes d'entre 67 i 70 anys, poc actius físicament, amb alt risc de patir malalties cardiovasculars i un índex de salut funcional baix.
- **Clúster 3**: Homes d'uns 70 anys, moderats quant a l'activitat física, amb un risc mitjà de patir malalties cardiovasculars i un índex de salut funcional baix.

Per aquesta mostra, partint les dades en tres grups, el K-means i el *clustering* jeràrquic distingeixen tres clústers diferenciats. Per una banda el clúster 1 són homes d'uns 67 anys amb un índex de salut funcional elevat i que, com a única característica negativa, presenten nivells de colesterol molt elevats. El clúster 2 són homes d'entre 67 i 70 anys amb l'índex de salut funcional més baix, que presenten obesitat de classe I i alts nivells de proteïna PCR. També es troben al nivell més alt de triglicèrids en comparació als altres clústers, presenten nivells de colesterol alts i practiquen poca activitat física. Finalment, el clúster 3 són homes d'uns 70 anys, amb un índex de salut baix, però que en mitjana es troba per sobre del clúster 2, que presenten preobesitat, practiquen una activitat mitjana entre el clúster 1 i el clúster 2, però que a diferència del clúster 2 tenen uns nivells de proteïna PCR, triglicèrids i colesterol baixos.

D'aquesta manera, el clúster 1 i el clúster 2 són els que es diferencien més entre ells, tenint el clúster 3 entremig dels dos. El que distingeix als clústers en primer lloc és l'índex de salut

funcional que presenten i, en segon lloc, els nivells d'índex de massa corporal, donat que els homes del clúster 1 presenten nivells normals, els homes del clúster 2 presenten obesitat de classe I i els homes del clúster 3 presenten preobesitat.

Pel mètode K-medoids, els perfils dels clústers són equiparables amb el K-means i el *clustering* jeràrquic aglomeratiu, però en certs aspectes presenten algunes característiques diferents. El mètode K-medoids no considera que els homes tinguin cap característica comuna, cosa que si ha passat pel mètode K-means i el mètode jeràrquic. Els perfils dels clústers s'han definit breument com:

- **Clúster 1**: Homes d'entre 67 i 70 anys, amb risc de patir malalties cardiovasculars, amb alt nivell educatiu i que presenten un índex de salut funcional alt.
- **Clúster 2**: Homes d'uns 70 anys, amb alt risc de patir malalties cardiovasculars, amb baix nivell educatiu i que presenten un índex de salut funcional baix.
- **Clúster 3**: Homes d'uns 67 anys, amb risc de patir malalties cardiovasculars, amb un nivell mitjà d'educació i que presenten un índex de salut funcional alt.

Per una banda, entra en joc la variable educació, i desapareix la variable activitat física. Per altra banda, les franges d'edat s'han vist afectades tenint en compte les característiques que presenta cada clúster. Per últim, els índexs de salut funcional han canviat, on abans es tenia un índex de salut elevat i dos índexs de salut entre un nivell baix i un nivell mitjà, ara pel mètode K-medoids es presenten dos clústers amb un índex de salut elevat i un clúster amb un índex de salut baix. No obstant això, per aquesta partició de les dades, l'índex de massa corporal segueix presentant diferències entre cada clúster.

Sense tenir en compte les franges d'edat i les variables categòriques es pot apreciar el següent. El clúster 1 del mètode K-medoids presenta característiques similars al clúster 1 del K-means i el *clustering* jeràrquic, ja que són homes que només tenen un nivell molt elevat de colesterol i pateixen hipertensió arterial en etapa 1. El clúster 2 del mètode K-medoids presenta característiques similars al clúster 2 dels mètodes K-means i jeràrquic. Es tracta d'homes amb obesitat de classe I, amb alts nivells de proteïna PCR, que es troben al nivell més alt de triglicèrids, pateixen hipertensió arterial en etapa 1 i tenen un índex de salut funcional baix. Per concloure, el clúster 3 del mètode K-medoids és equiparables al clúster 3 dels mètodes K-means i jeràrquic, atès que són homes amb preobesitat i hipertensió arterial en etapa 1, malgrat això el K-medoids, en contraposició als dos mètodes anteriors, considera que són homes amb un índex de salut funcional alt.

Així doncs, el mètode K-medoids troba certes característiques similars amb el mètode K-means i el *clustering* jeràrquic aglomeratiu. Segueix distingint els clústers entre homes amb obesitat de classe I, homes amb preobesitat i homes amb un pes normal. Tanmateix, troba que el conjunt d'homes presenten hipertensió arterial i, tot i no distingir les franges d'edat d'igual manera, troba els dos clústers més polaritzats de manera similar als dos primers mètodes de *clustering*. Els clúster polaritzats pel K-medoids són el clúster 1 i el clúster 2.

Per últim, com ha passat amb la mostra de dones, els perfils pels *Gaussian Mixture Models* són els més diferents d'entre tots els mètodes aplicats. En aquest cas, tots els homes presenten un índex de salut mitjà tirant cap a alt, en mitjana pateixen hipertensió arterial en

etapa 1, tenen un nivell de triglicèrids normal i nivells de colesterol LDL per sobre del llindar establert. Per aquest mètode els perfils dels clústers s'han definit breument com:

- **Clúster 1:** Homes d'uns 70 anys, amb alt risc de patir malalties cardiovasculars i una força d'adherència menor als altres clústers.
- **Clúster 2:** Homes d'uns 67 anys amb risc de patir malalties cardiovasculars i força d'adherència elevada.
- **Clúster 3:** Homes d'entre 67 i 70 anys amb risc mitjà de patir malalties cardiovasculars i força d'adherència elevada.

Per aquest cas, ha estat difícil diferenciar els tres clústers. Al contrari que als altres mètodes, no s'han respectat ni els nivells d'índex de massa corporal que diferenciava les característiques dels tres clústers, atès que el clúster 1, en teoria, són homes amb preobesitat, el clúster 2 són homes entre un nivell normal i la preobesitat i el clúster 3 també presenta preobesitat com el clúster 1. Per aquest mètode cap clúster és equiparable amb els mètodes anteriors. Només el clúster 1, pels nivells elevats de proteïna PCR en comparació als altres clústers, es podria considerar com el grup més extrem. Per això s'ha classificat com els homes amb alt risc de patir malalties cardiovasculars.

En definitiva, com a passat per la mostra de dones, per la mostra d'homes i considerant la divisió de les dades en tres clústers, el mètode K-means i el jeràrquic perfilen les dades pràcticament igual. El mètode K-medoids és equiparable als clústers del K-means i el *clustering* Jeràrquic aglomeratiu, però moltes característiques com l'edat i l'índex de salut funcional no s'assemblen. Finalment, el mètode dels GMMs no és equiparable amb cap clúster dels mètodes anteriors i és el més diferent respecte als altres mètodes de *clustering*.

Tenint en compte aquesta informació i observant les figures 6.4, 6.6, 6.8 i 6.13 on es mostra la divisió dels clústers per a cada algoritme de *clustering*, els millors mètodes per trobar perfils diferenciats per aquesta mostra són el K-means i el *clustering* jeràrquic aglomeratiu. El mètode K-medoids és molt capaç de trobar els clústers més polaritzats, però donat al solapament del clúster 3 no ha acabat extraient els mateixos resultats que els dos mètodes anteriors. Els GMMs ha estat el mètode que ha presentat una pitjor divisió de les dades, ja que els tres clústers s'han solapat entre ells i la informació que s'ha pogut extreure és més que dubtosa.

6.6. Comparació ARI

Durant tot el capítol VI, s'han estat estudiant els perfils dels clústers de la mostra de dones i la mostra d'homes per cada un dels mètodes estudiats en el capítol III. Per acabar l'anàlisi de *clustering*, en aquest apartat es compararan les estructures que s'han anat obtenint dels clústers per a les dues mostres analitzades. És a dir, s'estudiarà a través de l'Índex de Rand Ajustat (ARI) si els diferents mètodes de *clustering* classifiquen, per una banda les dones i per altra banda els homes, les observacions de manera similar a l'hora de dividir les bases de dades en clústers. L'ARI s'ha calculat gràcies a la funció **adjustedRandIndex** d'R (per a més informació consultar el capítol V).

6.6.1. Dones

En primer lloc, s'ha calculat l'ARI entre totes les particions del mètode jeràrquic, el K-means, el K-medoids i els *Gaussian Mixture Models* (GMMs).

Taula 6.25: Comparació de les estructures de *clustering* entre tots els mètodes.

Mètodes	ARI
K-means vs. Jeràrquic	0,956
K-means vs. K-medoids	0,499
K-means vs. GMMs	0,279
Jeràrquic vs K-medoids	0,477
Jeràrquic vs GMMs	0,265
K-medoids vs GMMs	0,181

A la Taula 6.25 es veu clarament que per la mostra de dones, el K-means i el *clustering* jeràrquic són els dos mètodes de *clustering* que classifiquen de manera pràcticament idèntica les observacions de la base de dades, atès que el valor de l'ARI és molt proper a 1.

També es pot observar com la partició del mètode K-medoids i els GMMs és la que menys s'assembla. En fixar-se amb els GMMs, cap tècnica de divisió presenta una partició equiparable a aquest mètode de *clustering* probabilístic. El valor de l'ARI pren valors baixos, per cap comparació supera el 0,3. Això és normal, donat que la idiosincràsia dels mètodes de *Soft clustering* és molt diferent de la de *Hard clustering* com s'ha explicat detalladament al capítol II.

Taula 6.26: Comparació de l'ARI pel *clustering* jeràrquic i el K-means.

Mètodes	ARI	
K-means vs. Jeràrquic	0,956	
K-means vs. K-medoids	0,499	52,2%
K-means vs. GMMs	0,279	29,2%
Jeràrquic vs. K-medoids	0,477	49,9%
Jeràrquic vs. GMMs	0,265	27,7%

Finalment, a la Taula 6.26 es calcula el percentatge de similitud pel mètode K-means i el *clustering* jeràrquic enfront del valor de l'ARI que han obtingut entre ells, per determinar quin dels dos mètodes s'assembla més en estructura al mètode K-medoids i als GMMs.

Es pot observar que la partició del mètode K-means enfront del K-medoids i els GMMs és més semblant que no pas la partició del *clustering* jeràrquic. En concret, l'estructura del K-means presenta un percentatge de similitud del 52,2% pel mètode K-medoids i del 29,2% respecte dels GMMs, en canvi, el *clustering* jeràrquic presenta una similitud del 49,9% amb l'estructura del mètode K-medoids i del 27,7% amb els GMMs.

6.6.2. Homes

Per la mostra d'homes, també s'ha calculat l'ARI entre totes les particions del mètode jeràrquic, el K-means, el K-medoids i els *Gaussian Mixture Models* (GMMs).

Taula 6.27: Comparació de les estructures de *clustering* entre tots els mètodes.

Mètodes	ARI
K-means vs. Jeràrquic	0,925
K-means vs. K-medoids	0,479
K-means vs. GMMs	0,035
Jeràrquic vs K-medoids	0,485
Jeràrquic vs GMMs	0,040
K-medoids vs GMMs	0,021

Igual que ha passat en la mostra de dones, a la Taula 6.27 es pot veure que per la mostra d'homes el mètode K-means i el *clustering* jeràrquic són els que s'assemblen més pel que fa a estructura. A més, classifiquen les observacions de la base de dades de manera pràcticament idèntica, donat que el valor de l'ARI és molt proper a 1.

Per aquesta mostra, l'estructura de *clustering* dels GMMs no és similar amb cap de les tècniques de partició analitzades. Els valors de l'ARI es troba molt proper a 0 quan es compara els GMMs amb els altres mètodes. Com en el cas de les dones, això es deu a les diferències intrínseques dels GMMs amb la resta de mètodes basats en dissimilaritats.

Taula 6.28: Comparació de l'ARI pel *clustering* jeràrquic i el K-means.

Mètodes	ARI	
K-means vs. Jeràrquic	0,925	
K-means vs. K-medoids	0,479	51,8%
K-means vs. GMMs	0,035	3,7%
Jeràrquic vs K-medoids	0,485	52,4%
Jeràrquic vs GMMs	0,040	4,4%

Per concloure la comparativa entre les estructures de *clustering*, a la Taula 6.28 es calcula el percentatge de similitud pel mètode K-means i el *clustering* jeràrquic enfront del valor de l'ARI que han obtingut entre ells, per determinar quin dels dos mètodes s'assembla més en estructura al mètode K-medoids. A la taula apareix també la comparació amb els GMMs, però com es pot apreciar representa un nivell de similitud molt baix tant pel K-means com pel *clustering* jeràrquic.

A diferència de la mostra de dones, per la mostra d'homes és el *clustering* jeràrquic el que s'assembla més a nivell d'estructura al mètode K-medoids. En concret, el *clustering* jeràrquic presenta un percentatge de similitud del 52,4% enfront del 51,8% de similitud del mètode K-means.

VII. CONCLUSIONS

Un cop fet l'anàlisi en el software estadístic R pels quatre mètodes de *clustering* estudiats: K-means, *clustering* jeràrquic, K-medoids i *Gussian Mixture Models*, he pogut determinar de forma empírica quins mètodes s'adapten millor per la mostra de dones i la mostra d'homes de la base de dades de l'estudi de salut ELSA. A més, també he tingut l'oportunitat d'estudiar si les diferents tècniques de *clustering* classifiquen les dades de manera similar.

Ambdues mostres s'ha arribat a la mateixa conclusió. Per una banda, un cop aplicats tots els passos previs per poder executar els mètodes de *clustering*, per mitjà dels mètodes de selecció de clústers òptims estudiats al capítol III i a través d'informació contextual sobre la base de dades original, les dues mostres s'han dividit en tres clústers. D'altra banda, els mètodes que divideixen millor les dades de dones i homes i troben perfils de salut més diferenciats són el K-means i el *clustering* jeràrquic aglomeratiu. En primer lloc, ja que han estat els mètodes que han classificat les mostres de manera més diferenciada respecte a l'espai de les dades, i en segon lloc, perquè han extret el mateix perfil dels clústers (és a dir, estructures de *clustering* similars) pels dos mètodes. A part, quan s'ha aplicat l'Índex de Rand Ajustat (ARI), el K-means i el *clustering* jeràrquic aglomeratiu han classificat als individus de les dues mostres de la mateixa manera. Els dos sexes han donat un resultat de l'ARI per sobre de 0,9, els més alts d'entre totes les tècniques de *clustering* aplicades.

En concret, la mostra de dones ha classificat, pel clúster 1, les dones amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt. Pel clúster 2, les dones sanes amb alt índex de salut funcional. I pel clúster 3, les dones amb alt risc de patir malalties cardiovasculars i amb un índex de salut funcional baix. La diferència entre els clústers 1 i 2 radica en el fet que les dones sanes no pateixen ni hipertensió arterial en fase 1, ni presenten alts nivells de colesterol. Sobre els clústers 2 i 3 representen els grups més polaritzats, atès que les dones del clúster 3 són les que presenten pitjors indicadors de salut.

Per la mostra d'homes, el clúster 1 l'han definit homes d'uns 67 anys, actius físicament i amb risc de patir malalties cardiovasculars però amb un índex de salut funcional alt. El clúster 2 consta d'homes entre 67 i 70 anys, poc actius físicament, amb alt risc de patir malalties cardiovasculars i un índex de salut funcional baix. I el clúster 3, ha quedat definit per homes d'uns 70 anys, moderats respecte a l'activitat física, amb un risc mitjà de patir malalties cardiovasculars i un índex de salut funcional baix. Els clústers 1 i 2 han resultat ser els més polaritzats, ja que al clúster 2 es troben els homes amb més indicadors de salut negatius i al clúster 1 es troben els homes amb l'índex de salut funcional més alt i, com a únic tret negatiu, amb un colesterol elevat, mentre que el clúster 3 es troba entremig d'aquests dos grups. El que més ha diferenciat als homes ha estat el seu índex de massa corporal: pel clúster 1 els homes tenen un nivell normal, pel clúster 2 els homes tenen obesitat de classe I, i finalment, pel clúster 3 els homes tenen preobesitat.

Pels altres dos mètodes s'ha produït el que s'ha anat parlant en tot aquest treball: el "*no free lunch methods*", és a dir, que a causa de les característiques de les mostres estudiades, no tots els mètodes han pogut donar la mateixa solució. Tot i això, el K-medoids s'ha apropat als perfils de les dues mostres i ha estat capaç de trobar els clústers més polaritzats tant per dones com per homes. Per tant, ens ha donat a entendre que si s'hagués dividit el mètode per 2 clústers, hauria estat capaç de trobar els grups més polaritzats per a cada una de les mostres.

Pel que fa als *Gaussian Mixture Models*, ha estat el mètode que ha presentat resultats més diferents, amb poca distinció entre grups i més solapament entre clústers, sobretot en la mostra d'homes. Aquest fet s'ha produït, per una banda, ja que tot i haver normalitzat les dades en les dues mostres, un gran nombre de variables no han acabat de normalitzar-se, llavors no s'ha pogut aplicar el mètode sobre un conjunt de variables Gaussianes. D'altra banda, en estimar el conjunt de paràmetres del model de mixtura a través de l'algoritme EM, es pot haver quedat estancat en un màxim local de la funció de versemblança, que és comú que aquesta presenti multimodalitat en el context dels *mixture models*. Addicionalment, els GMMs no són tan estrictes, com per exemple el K-means, en el fet que hi hagi solapament entre clústers. Això pot provocar que els GMMs ens estiguin identificant menys clústers. També, com a última possibilitat, pot haver estat un problema de mida mostral, ja que quan es té un nombre insuficient de punts per mixtura, l'estimació de les matrius de covariàncies és complexa, llavors l'algoritme pot divergir i troba una solució amb probabilitat infinita, a no ser que es regularitzin les covariàncies de manera artificial.

Personalment, tot i que els GMMs no han acabat de funcionar en les mostres analitzades per aquest treball, trobo que és un mètode molt potent donat que assigna la probabilitat de pertinença en els clústers per a cada individu, fet que no es produeix pels altres mètodes de *clustering* que he estat estudiant. A part, per les dues mostres analitzades, el fet de no haver pogut aplicar el mètode sobre un conjunt en què totes les variables fossin Gaussianes, crec que ha influït molt a l'hora d'obtenir els resultats. En primer lloc, perquè en la mostra de dones, la que ha presentat més variables Gaussianes, el mètode dels GMMs ha estat capaç de dividir les dades de manera més diferenciada que per la mostra d'homes, que ha presentat menys variables Gaussianes i un grau de solapament entre clústers molt més elevat. I en segon lloc, perquè els valors de l'ARI per la mostra de dones, han estat molt més elevats que per la mostra d'homes. És a dir, la mostra amb més variables de tipus gaussià ha presentat una estructura de *clustering* molt més similar respecte els mètodes que s'han adaptat millor a les dades (K-means, *clustering* jeràrquic). Tanmateix, un altre problema que crec que pot haver influït perquè els GMMs hagi estat el mètode més diferent, és que com s'ha vist en l'anàlisi exploratòria de les dades, a les dues mostres, moltes variables han presentat un alt nivell de correlació entre elles (p. ex. l'índex de massa corporal amb la circumferència de la cintura i el maluc, la pressió sistòlica amb la pressió diastòlica i el colesterol LDL amb el colesterol total). Aquest fet pot haver influït de cara al solapament entre clústers que ha presentat el mètode. Per futurs anàlisis, caldria provar d'aplicar els GMMs sobre un conjunt de variables no correlacionades per tal que els individus de cada mostra fossin estrictament independents entre ells.

Com a línies de futur (fora de l'*scope* d'aquest treball), seria molt interessant, en primer lloc, comprovar que els perfils de dones i homes que s'han extret en aquest treball es compleixin per altres mostres aleatòries de la cohort ELSA, agafant el mateix conjunt de variables contínues amb les quals s'han creat els clústers. Per d'aquesta manera, poder validar internament els perfils obtinguts. En segon lloc, seguint amb la validació interna, estudiar a fons les variables més explicatives a l'hora de perfilar els clústers. De manera que no existís correlació entre variables i, d'una banda, comprovar si amb menys variables s'obtenen les mateixes estructures de *clustering*, i d'altra banda, si amb l'addició de noves variables independents entre elles s'obtenen més característiques influents en el camp de l'envelliment saludable. Finalment, tenint en compte que en el projecte ATHLOS, d'on sorgeix la variable healthstatus (l'indicador de salut funcional), compta amb altres cohorts d'arreu del món, un

cop feta la validació interna dels resultats, caldria validar si per diferents cohorts es mantenen els perfils dels clústers. Així doncs, en general, es tractaria d'unir la validació interna dels resultats amb la validació externa en altres cohorts d'arreu del món, per individus de les mateixes franges d'edat que s'han analitzat al llarg d'aquest treball. Per d'aquesta manera, trobar clústers de dones i homes diferenciats i poder aplicar tractaments personalitzats en funció de les característiques de cada grup, com també les condicions de salut òptimes per cada sexe en l'àrea de l'envelliment saludable.

VIII. BIBLIOGRAFIA

1) Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.
2) Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.
3) R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
4) Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01
5) MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).
6) Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.
7) Kaufman, L., & Rousseeuw, P. (1987). Statistical data analysis based on the L1-norm and related methods. Clustering by means of medoids. North-Holland, 405-416.
8) Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis, 344, 68-125.
9) Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 803-821.
10) Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association, 97(458), 611-631.
11) Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern recognition, 28(5), 781-793.
12) Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal, 8(1), 289.
13) Schwarz, G. (1978). Estimating the dimension of a model. Annals of statistics, 6(2), 461-464.
14) Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal, 41(8), 578-588.
15) Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence, 22(7), 719-725.
16) Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.
17) Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and computing, 10(4), 339-348.
18) Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 289-317
19) Thorndike, R. L. (1953). Who belongs in the family?. Psychometrika, 18(4), 267-276.

20) Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.
21) Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.
22) Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.
23) Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.
24) Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), 846-850.
25) Milligan, G. W. and Cooper, M. C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21, 441–458.
26) Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: the English longitudinal study of ageing. International journal of epidemiology, 42(6), 1640-1648.
27) Au, B., Smith, K. J., Gariépy, G., & Schmitz, N. (2014). C-reactive protein, depressive symptoms, and risk of diabetes: results from the English Longitudinal Study of Ageing (ELSA). Journal of psychosomatic research, 77(3), 180-186.
28) Sanchez-Niubo, A., Egea-Cortés, L., Olaya, B., Caballero, F. F., Ayuso-Mateos, J. L., Prina, M., ... & ATHLOS Consortium. (2019). Cohort profile: the ageing trajectories of Health longitudinal opportunities and synergies (ATHLOS) project. International journal of epidemiology, 48(4), 1052-1053i.
29) Sanchez-Niubo, A., Forero, C. G., Wu, Y. T., Giné-Vázquez, L., Prina, M., De La Fuente, J., ... & Haro, J. M. (2020). Development of a common scale for measuring healthy ageing across the world: results from the ATHLOS consortium. International Journal of Epidemiology.
30) Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.
31) Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611.
32) Ryan A. Peterson (2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. Journal of Applied Statistics, 1-16.
33) Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. Journal of chemical information and computer sciences, 30(1), 36-41.
34) Alboukadel Kassambara (2020). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.5.0. https://CRAN.R-project.org/package=rstatix
35) Levene, H. (1960). Robust tests for equality of variances. In 'Contributions to probability and statistics: essays in honor of Harold Hotelling'.(Eds I Olkin, SG Ghurye, W Hoeffding, WG Madow, HB Mann) pp. 278–292.
36) Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(4), 795-809.

Matteucci, Matteo. *A Tutorial on Clustering Algorithm* [en línia]. matteucci.faculty, 2008, [consulta: 20 de febrer de 2021]. Disponible a: <https://matteucci.faculty.polimi.it/Clustering/tutorial_html/index.html>.

Amat Rodrigo, Joaquín. *Clustering y heatmaps: aprendizaje no supervisado* [en línia]. RPubs, Setembre 2017, [consulta: 26 de febrer de 2021]. Disponible a: <https://rpubs.com/Joaquin_AR/310338>.

Packt. *Introduction to Clustering and Unsupervised Learning* [en línia]. Packt>, febrer 2016 [consulta: 1 de març de 2021]. Disponible a <https://hub.packtpub.com/introduction-clustering-and-unsupervised-learning/>.

Pedamkar, Priya. *Clustering in Machine Learning* [en linia]. EDUCBA, -, [consulta: 6 de març de 2021]. Disponible a: < https://www.educba.com/clustering-in-machine-learning/>.

Kassambara, Alboukadel. *Cluster Analysis in R: Practical Guide* [en línia]. Datanova, novembre 2019, [consulta: 8 de març de 2021]. Disponible a: <https://www.datanovia.com/en/blog/cluster-analysis-in-r-practical-guide/>.

Yobero, Czar. *K-Means Clustering Tutotial* [en línia]. RPubs, febrer 2018, [consulta: 9 de març de 2021]. Disponible a: <https://rpubs.com/cyobero/k-means>.

University of Massachusetts Boston. *The PAM Clustering Algorithm* [en línia]. College of Science and Mathematics, Department of Computer Science, -, [consulta: 25 de març de 2021]. Disponible a: <https://www.cs.umb.edu/cs738/pam1.pdf>.

Wikipedia. *K-media* [en línia]. Novembre 2020, [consulta: 21 d'abril de 2021]. Disponible a: <https://es.wikipedia.org/wiki/K-medias>.

Jiménez Cuadrillero, Miguel Ángel. *Clustering Jerárquico en R* [en línia]. RPubs, maig 2018, [consulta: 30 d'abril de 2021]. Disponible a: <https://rpubs.com/mjimcua/clustering-jerarquico-en-r>

Boehmke, Bradley i Greenwell, Brandon. *Hands-On Machine Learning with R*[en línia].Github. Section IV. Clustering. [consulta: 30 d'abril de 2021]. Disponible a: <https://bradleyboehmke.github.io/HOML/kmeans.html>.

Fernández, Daniel. *Multivariate Analysis (MVA) Clustering* [projecció visual]. Universitat Politècnica de Catalunya , 2021. 94 diapositives.

Grosse, Roger i Srivastava, Nitich. *Lecture 16: Mixture models*[en línia]. University of Toronto, Department of Computer Science, -, [consulta: 5 de maig de 2021]. Disponible a: <http://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf>.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.

Kassambara, Alboukadel. *Determining The Optimal Number Of Clusters: 3 Must Know* [en línia]. Datanova, desembre 2018, [consulta: 10 de maig de 2021]. Disponible a: <https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/>.

Santos, J. M., & Embrechts, M. (2009, September). On the use of the adjusted rand index as a metric for evaluating supervised classification. In International conference on artificial neural networks (pp. 175-184). Springer, Berlin, Heidelberg.

Yeung, K. Y., & Ruzzo, W. L. (2001). Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics, 17(9), 763-774.

IBM Cloud Education. *Machine Learning* [en línia]. IBM, juny 2020, [consulta: 30 de maig de 2021]. Disponible a: <https://www.ibm.com/cloud/learn/machine-learning>.

Stack Exchange. *Hout to intèrpret agglomerative coefficient agnes() function?* [en línia].Data Science, Questions, maig 2017, [consulta: 31 de maig de 2021]. Disponible a: <https://datascience.stackexchange.com/questions/18860/how-to-interpret-agglomerative-coefficient-agnes-function>.

Kassambara, Alboukadel. *Assessing Clustering Tendency* [en línia]. Datanova, 2017, [consulta: 7 de juny de 2021]. Disponible a: < https://www.datanovia.com/en/lessons/assessing-clustering-tendency/>.

Daniels, Luke. *Cluster Analyses Lecture* [en línia]. Github, abril 2018, [consulta: 7 de juny de 2021]. Disponible a: <https://lukedaniels1.github.io/Bio381_2018/Daniels_Cluster_Analysis_Lecture.html>.

World Health Organization. *Body mass index – BMI* [en línia]. Regional officer for Europe, 2021, [consulta: 8 juny de 2021]. Disponible a: <https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi?source=post_page---------->.

Actualidad sanitaria. *Tensión sistòlica y diastòlica: valores normales* [en línia]. 2021, [consulta: 8 de juny de 2021]. Disponible a: <https://actualidadsanitaria.com/vida-saludable/tension-sistolica-y-diastolica-valores-normales/>.

Health Checkup. *Systolic And Diastolic Blood Pressure: Know The Difference* [en línia]. General, Heart, gener 2020, [consulta: 8 de juny de 2021]. Disponible a: <https://www.healthcheckup.com/general/systolic-vs-diastolic-blood-pressure/>.

RN, Debra Manzelly. *What to Know About Blood Glucose Levels* [en línia]. verywellhealth, maig 2021, [consulta: 8 de juny de 2021]. Disponible a: <https://www.verywellhealth.com/recommended-blood-glucose-levels-for-diabetes-1087681#:~:text=Target%20Pre-exercise%20Blood%20Glucose%20Levels%20%20%20,anaerobic%20%20...%20%201%20more%20rows%20>

Rojas, María. *Niveles de glucosa saludables ¿Cuál es el rol del MCG?* [en línia]. Health Sensor, -, [consulta: 8 de juny de 2021]. Disponible a: <https://healthsensor.es/glucosa-saludable-monitoreo-continuo-glucosa/>.

HEART UK - *The colesterol charity* [en línia]. -, [consulta: 8 de juny de 2021]. Disponible a: <https://www.heartuk.org.uk/>.

IX. ANNEX

9.1. Aplicació dels mètodes de clustering sobre dades de salut

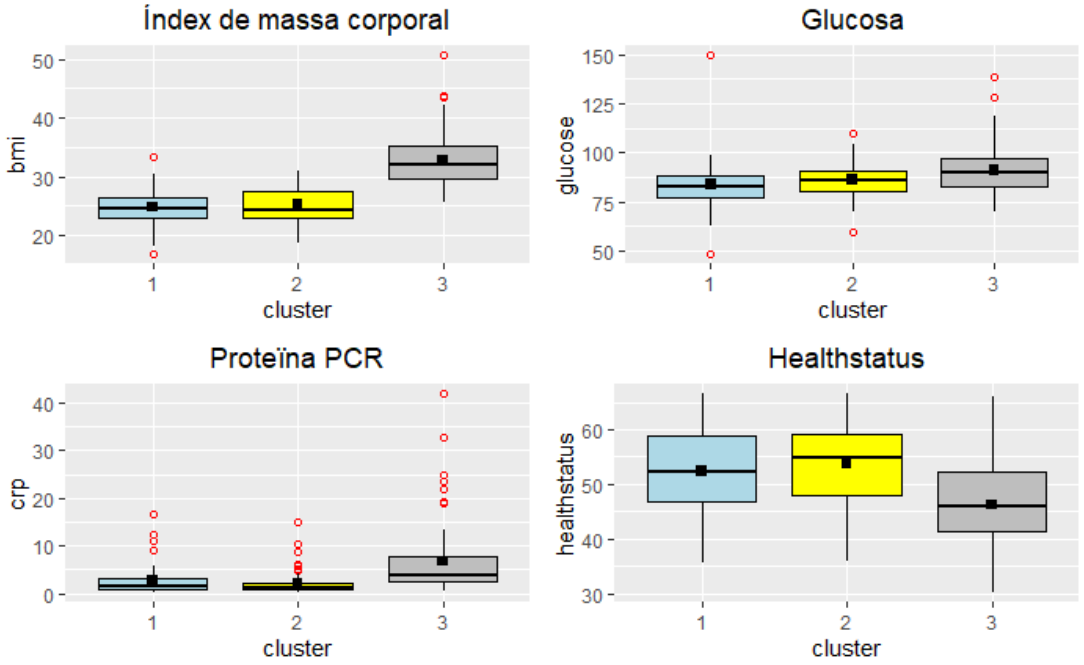
9.1.1. K-means

- Dones:

A continuació es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.2, referent a les variables numèriques, i la Taula 6.3, referent a les variables categòriques.

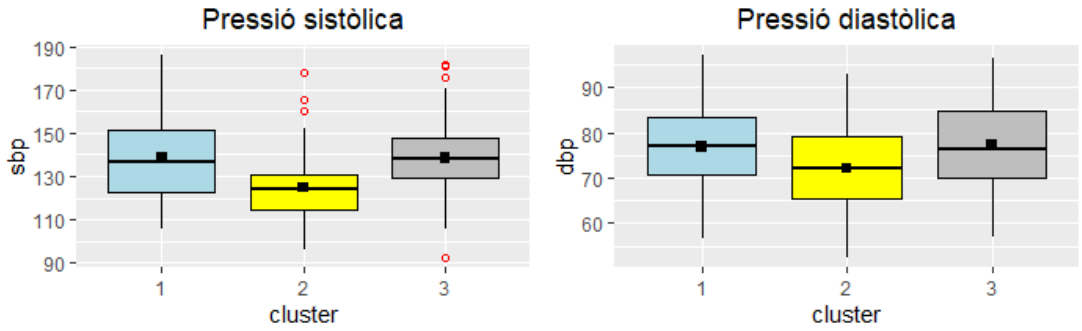
Per les variables numèriques, es pot apreciar que per les variables de l'índex de massa corporal, la circumferència de la cintura i el maluc, la glucosa, els triglicèrids, la proteïna PCR i el healthstatus, els clústers 1 i 2 presenten valors estadísticament iguals, mentre que el clúster 3 és diferent (veure Figura 9.1).

Figura 9.1: Representació de 4 gràfics de caixa on es pot veure les diferències dels clústers 1 i 2 (blau i groc respectivament) amb el clúster 3 (gris).



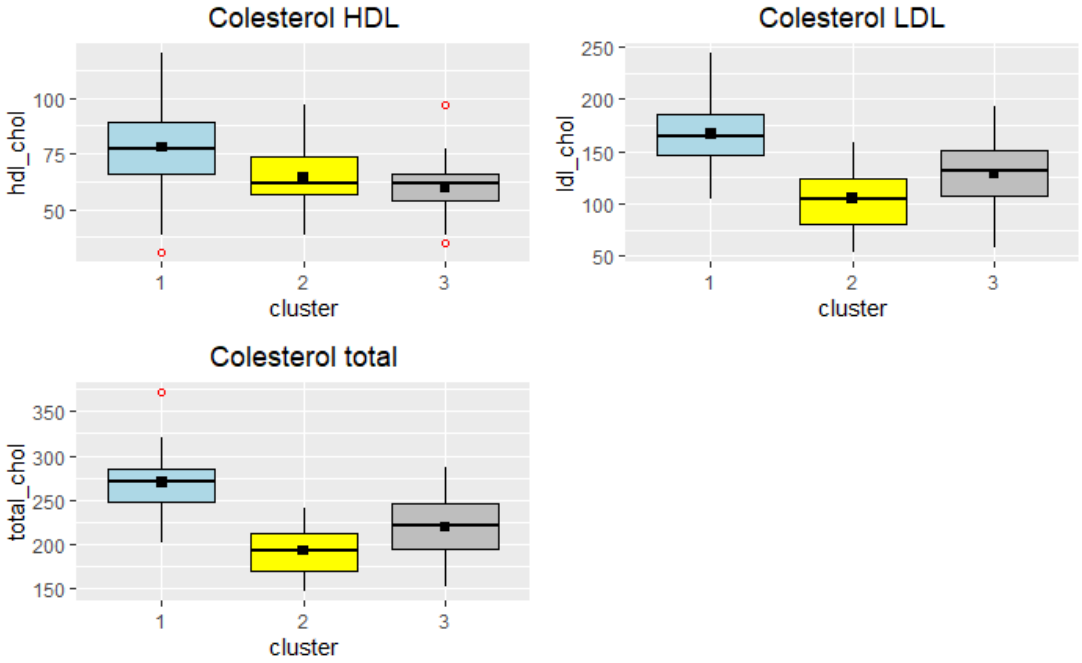
Per la pressió sistòlica i diastòlica, els clústers 1 i 3 tenen valors estadísticament iguals i el clúster 2 és el que es diferencia entre ell (veure Figura 9.2).

Figura 9.2: Gràfics de caixa de la pressió sistòlica i diastòlica on es pot veure les diferències dels clústers 1 i 3 (blau i gris respectivament) amb el clúster 2 (groc).



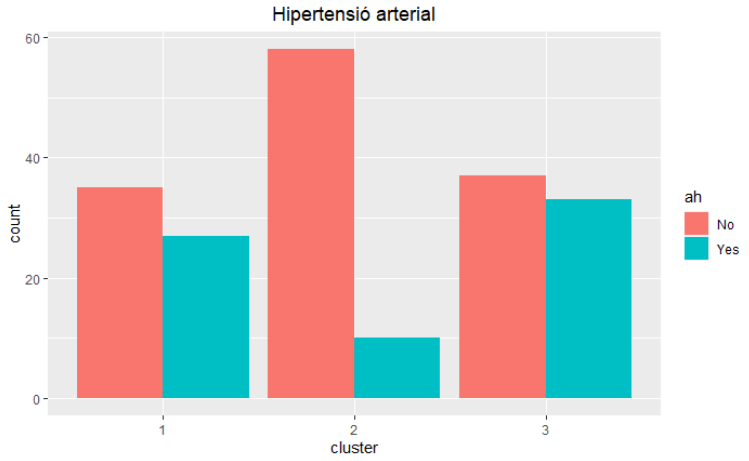
Pel colesterol HDL, els clústers 2 i 3 presenten valors estadísticament iguals, mentre que pels valors del colesterol LDL i el colesterol total, els tres clústers són diferents entre ells (veure Figura 9.3).

Figura 9.3: Gràfics de caixa dels tres tipus de colesterol. Clúster 1 representat en blau, clúster 2 en groc i clúster 3 en gris.



Respecte a les variables categòriques, l'única variable que presenta diferencies estadísticament significatives, amb un nivell de significació del 5%, és la variable de la hipertensió arterial, on els clústers 1 i 3 presenten valors iguals i el clúster 2 és estadísticament diferent als altes (veure Figura 9.4).

Figura 9.4: Gràfic de barres per la Hipertensió arterial.

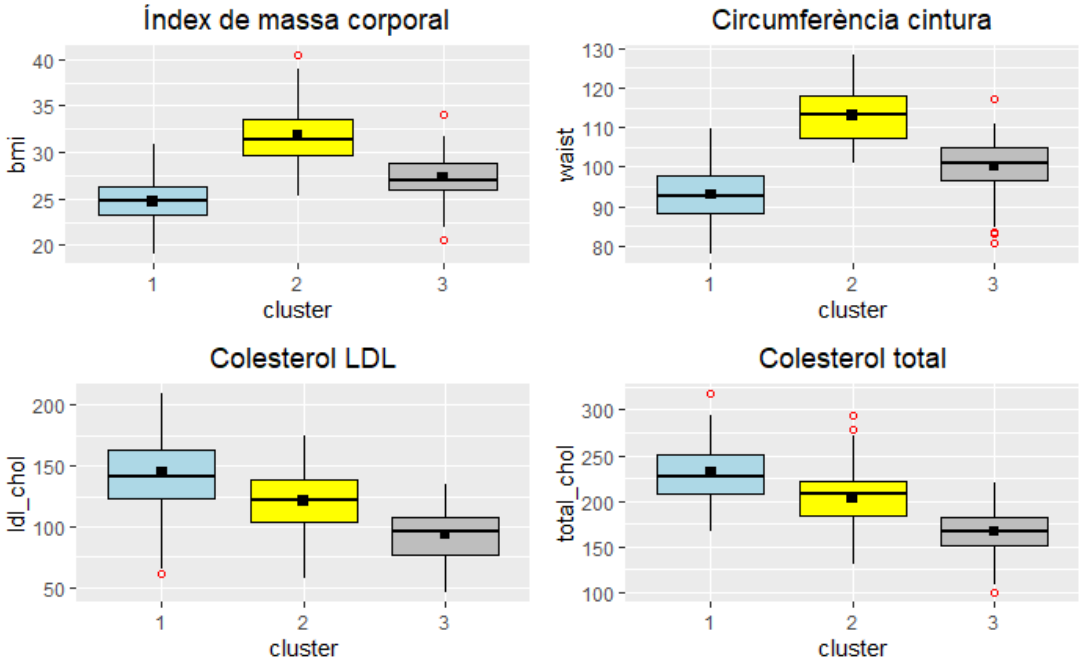


- Homes:

A continuació es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.5, referent a les variables numèriques, i la Taula 6.6, referent a les variables categòriques.

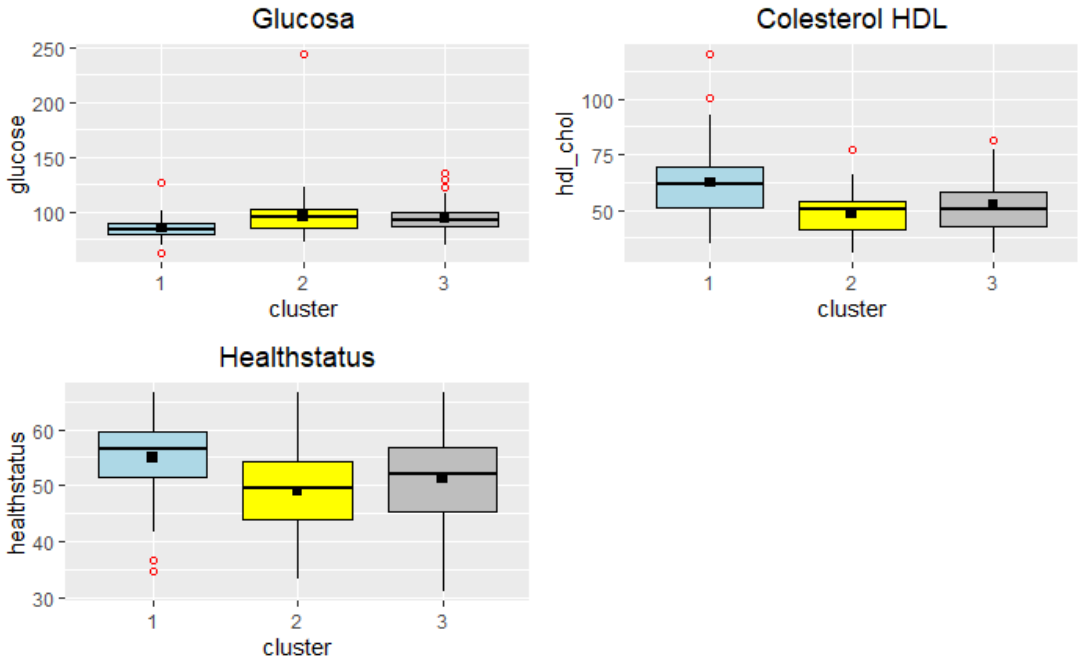
Per les variables numèriques, es pot apreciar que els 3 clústers són diferents pel que fa a l'índex de massa corporal, la circumferència de la cintura, la circumferència del maluc, el colesterol LDL i el colesterol total (veure Figura 9.5).

Figura 9.5: Gràfics de caixa de 4 variables on es pot veure les diferències entre tots els clústers (clúster 1 blau, clúster 2 groc i clúster 3 gris).



Es pot veure que el clúster 2 i el clúster 3 es comporten estadísticament igual pels nivells de glucosa en sang, el colesterol HDL i l'índex de salut referent a la variable healthstatus (veure Figura 9.6).

Figura 9.6: Gràfics de caixa per veure la igualtat entre els clústers 2 i 3 (groc i gris respectivament).



També, el clúster 1 i el clúster 3 es comporten de manera similar pels nivells de triglicèrids i proteïna PCR i, en canvi, l'edat del clúster 1 és estadísticament diferent de l'edat del clúster 3, deixant el clúster 2 al mig dels dos grups (veure Figura 9.7).

Respecte a les variables categòriques, s'aprecia que els nivells d'activitat física són diferents pel clúster 1 i pel clúster 2 amb el clúster 3 es troba entremig dels dos. El clúster 1 i el clúster 3 s'autoavaluen de manera similar (veure Figura 9.8).

Figura 9.7: Gràfics de caixa per veure igualtats i diferències entre els clústers 1 i 3 (blau i gris respectivament).

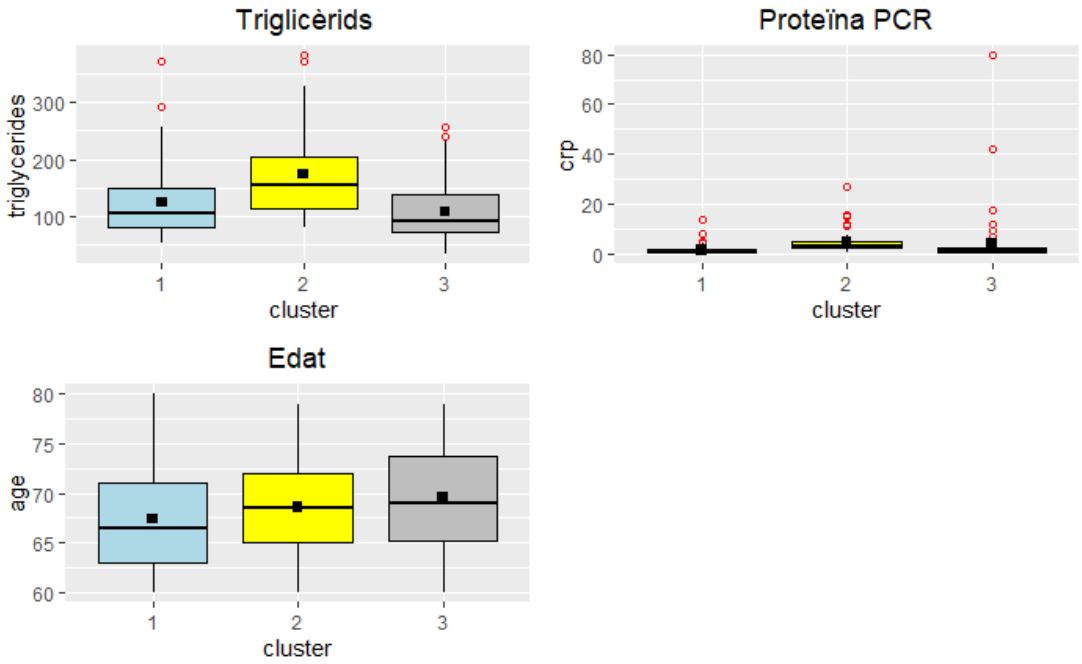
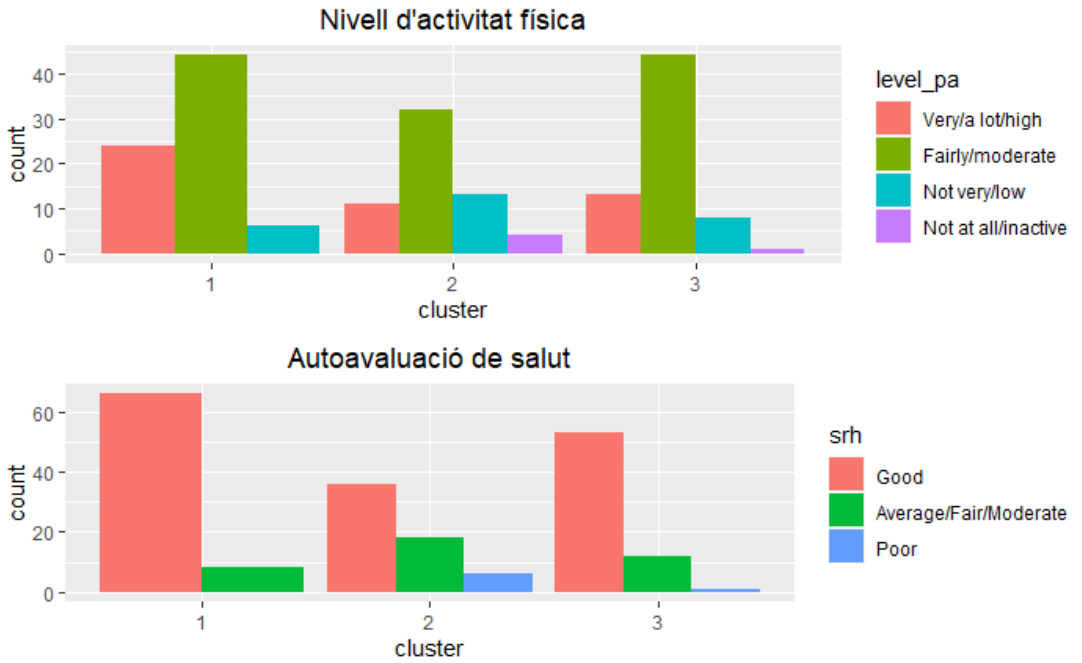


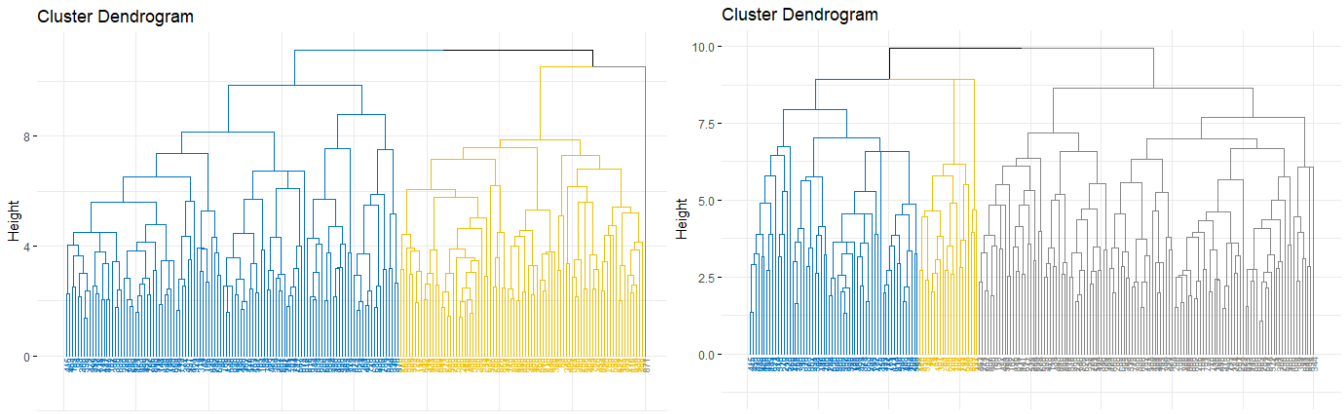
Figura 9.8: Gràfics de barres per les dues variables categòriques presentades a la taula 6.6.



9.1.2. Clustering Jeràrquic

En primer lloc, a la Figura 9.9 es presenten els dendrogrames del *clustering* jeràrquic divisiu

Figura 9.9: Dendrograma del *clustering* jeràrquic divisiu dividit en 3 clústers per la mostra de dones, esquerra, i la mostra d'homes, dreta.



- Dones:

Tot seguit es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.8, referent a les variables numèriques, i la Taula 6.9, referent a les variables categòriques.

Per les variables numèriques, es pot apreciar que les variables de l'índex de massa corporal, la circumferència de la cintura i el maluc, els triglicèrids, la proteïna PCR i el healthstatus als clústers 1 i 2 presenten valors estadísticament iguals, mentre que el clúster 3 és diferent (veure Figura 9.10).

Pels nivells de pressió sistòlica i diastòlica el clúster 1 presenta valors estadísticament iguals al clúster 3 (veure Figura 9.11).

Quant al colesterol LDL i el colesterol total, els tres clústers són diferents entre ells (veure Figura 9.12).

Per últim, el colesterol HDL presenta valors estadísticament iguals pels clústers 2 i 3, mentre que la glucosa és diferent en els clústers 1 i 3 i el clúster 2 presenta un valor intermedi entre aquests dos grups(veure Figura 9.13)

Figura 9.10: Representació de 4 gràfics de caixa on es pot veure les diferències dels clústers 1 i 2 (blau i groc respectivament) amb el clúster 3 (gris).

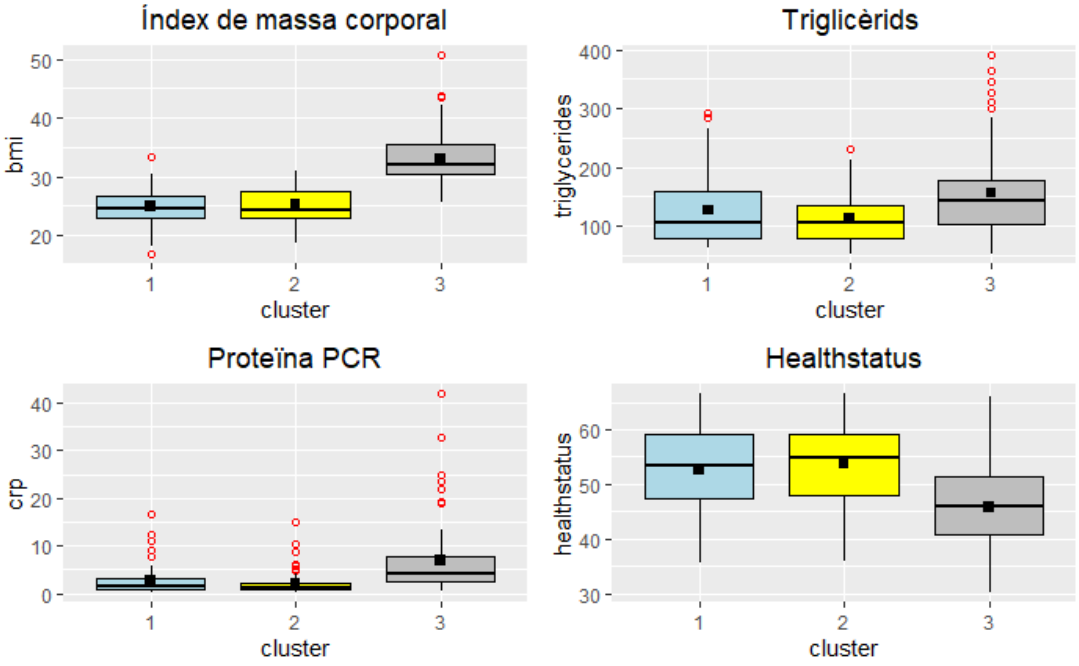


Figura 9.11: Gràfics de caixa de la pressió sistòlica i diastòlica on es pot veure les diferències dels clústers 1 i 3 (blau i gris respectivament) amb el clúster 2 (groc).

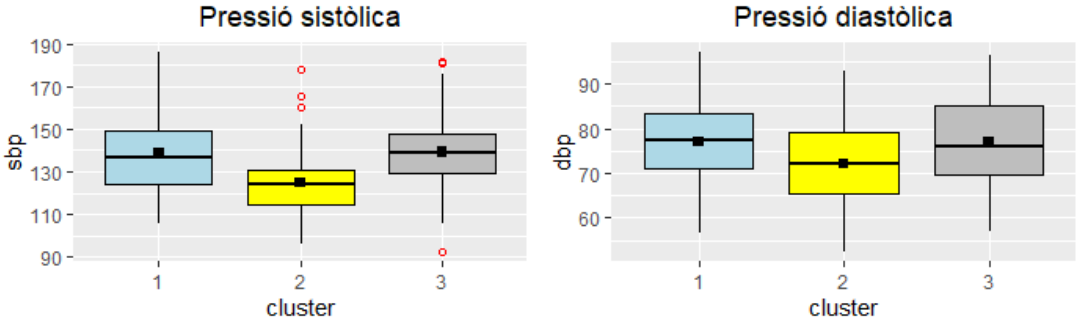


Figura 9.12: Gràfics de caixa pel colesterol LDL i el colesterol total on tots els clústers són estadísticament diferents entre ells.

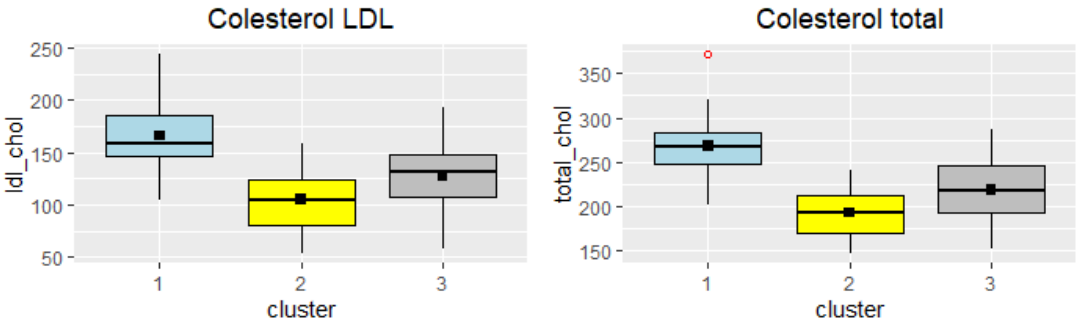
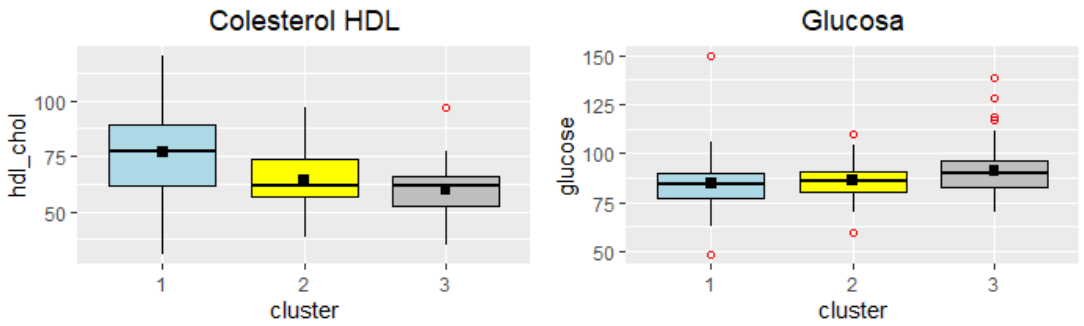
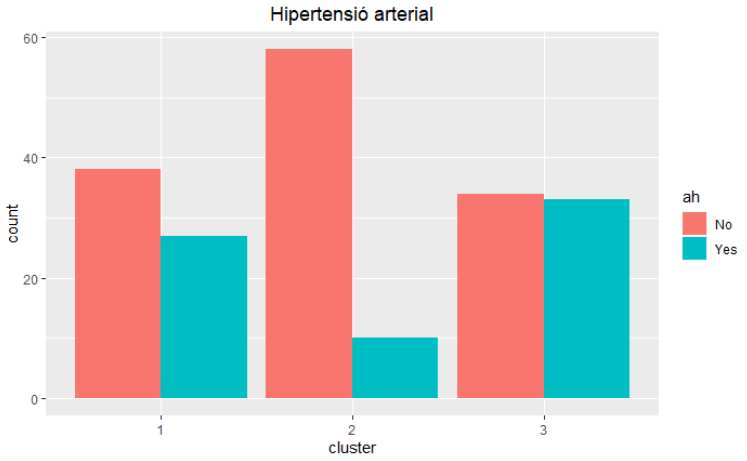


Figura 9.13: Gràfics de caixa del colesterol HDL i la glucosa.



Referent a les variables categòriques, l'única variable que presenta diferencies estadísticament significatives amb un nivell de significació del 5% és la variable de la hipertensió arterial, on els clústers 1 i 3 presenten valors iguals i el clúster 2 és estadísticament diferent dels altres (veure Figura 9.14).

Figura 9.14: Gràfic de barres per la Hipertensió arterial.



- Homes:

Tot seguit es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.11, referent a les variables numèriques, i la Taula 6.12, referent a les variables categòriques.

Per les variables numèriques, s'aprecia que els tres clústers són diferents per les variables de l'índex de massa, la circumferència de la cintura i el maluc, el colesterol LDL i el colesterol total (veure Figura 9.15).

Es pot veure que el clúster 2 i el clúster 3 es comporten estadísticament igual pels nivells de glucosa, el colesterol HDL i l'índex de salut referent a la variable healthstatus (veure Figura 9.16).

Figura 9.15: Representació de gràfics de caixa de 4 variables on es pot veure les diferències entre tots els clústers (clúster 1 blau, clúster 2 groc i clúster 3 gris).

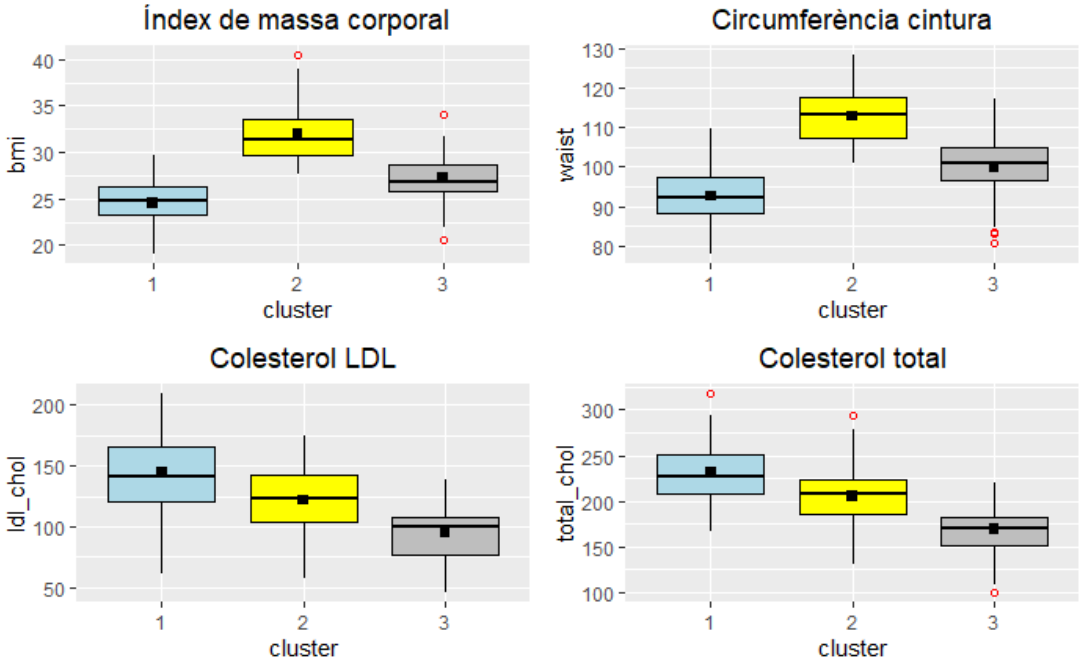
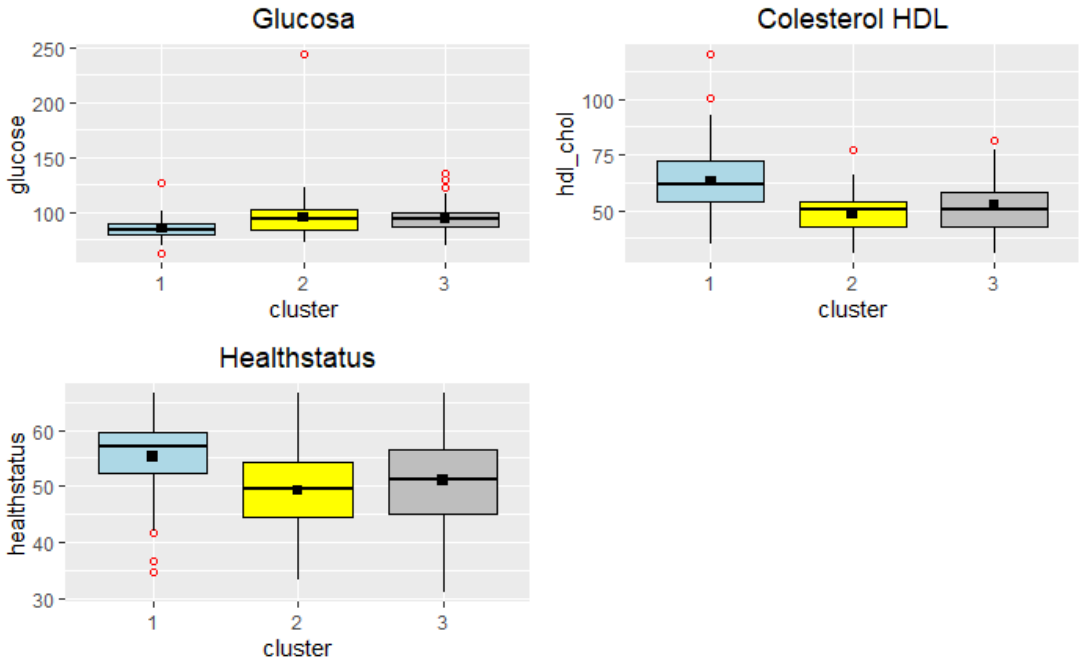
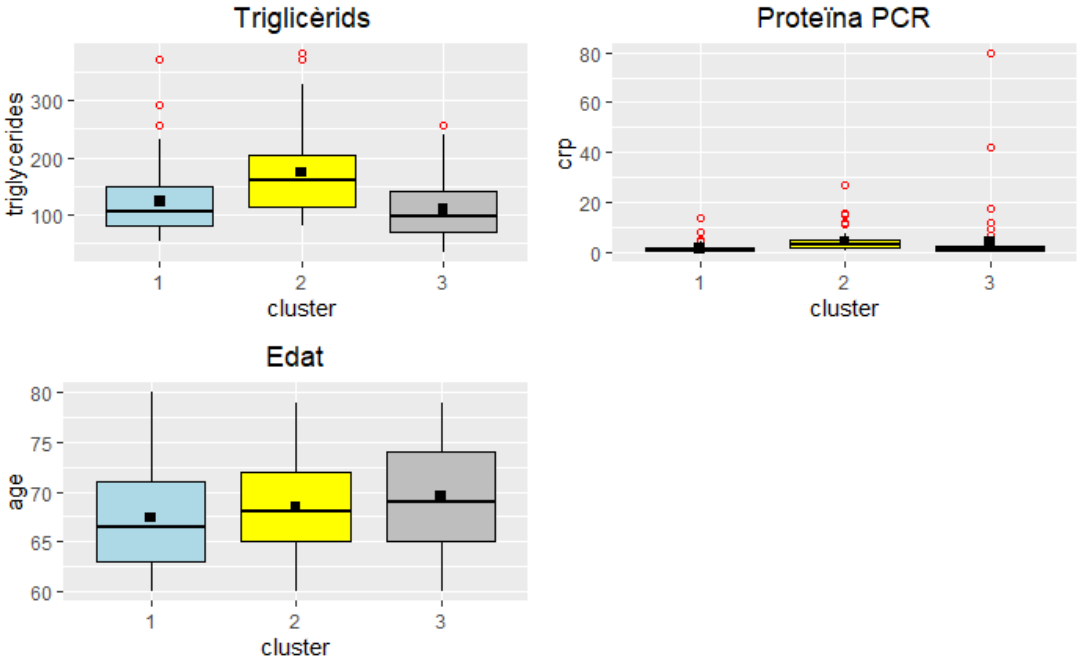


Figura 9.16: Gràfics de caixa on es mostra el comportament similar dels clústers 2 i 3.



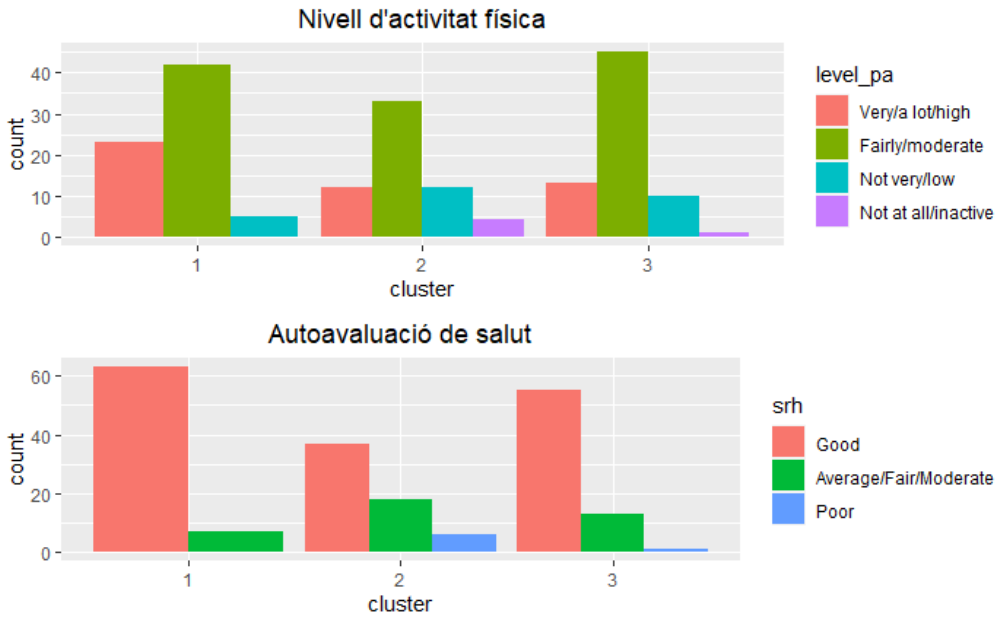
Tanmateix, el clúster 1 i el clúster 3 es comporten de manera similar pels nivells de triglicèrids i proteïna PCR, en canvi, l'edat del clúster 1 és estadísticament diferent de l'edat del clúster 3, on el clúster 2 presenta valors entremig dels dos grups (veure Figura 9.17).

Figura 9.17: Gràfics de caixa on es mostra el comportament similar dels clústers 1 i3 i la variable diferent entre aquests dos clústers.



Basant-se en les variables categòriques, s'aprecia que els nivells d'activitat física són diferents pel clúster 1 i pel clúster 2, el clúster 3 es troba entremig dels dos. Sobre l'autoavaluació de salut, el clúster 1 i el clúster 3 s'autoavaluen de manera similar (veure Figura 9.18).

Figura 9.18: Gràfics de barres per les tres variables categòriques presentades a la taula 6.12.



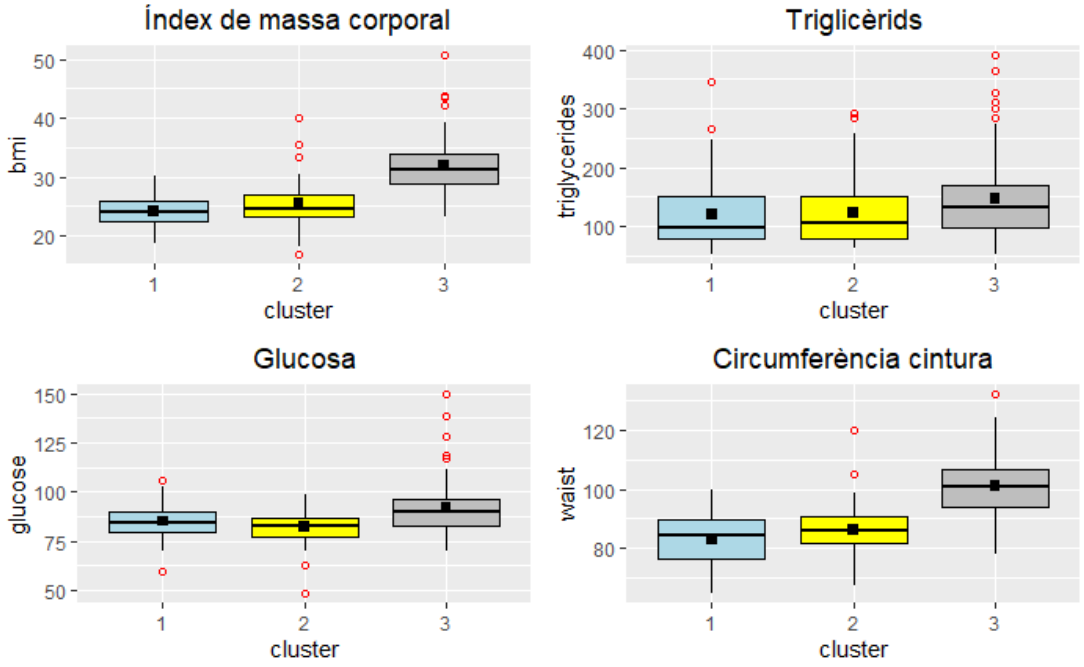
9.1.3. K-medoids

- Dones:

A continuació es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.14, referent a les variables numèriques, i la Taula 6.15, referent a les variables categòriques.

Respecte a les variables numèriques, es pot apreciar que existeixen forces diferències entre el clúster 3 i els clústers 1 i 2 (veure Figura 9.19).

Figura 9.19: Representació de 4 gràfics de caixa on es pot veure les diferències dels clústers 1 i 2 (blau i groc respectivament) amb el clúster 3 (gris).



Els clústers 2 i 3 presenten valors estadísticament iguals per a la pressió sistòlica, la pressió diastòlica i els nivells de proteïna PCR (veure Figura 9.20).

Per altra banda, els clústers 1 i 3 són similars per a totes les variables referents al colesterol i la variable que és estadísticament diferent per a tots els clústers és healthstatus, que es refereix al l'índex de salut funcional (veure Figura 9.21).

Figura 9.20: Gràfics de caixa de les variables estadísticament iguals pels clústers 2 i 3 (groc i gris respectivament).

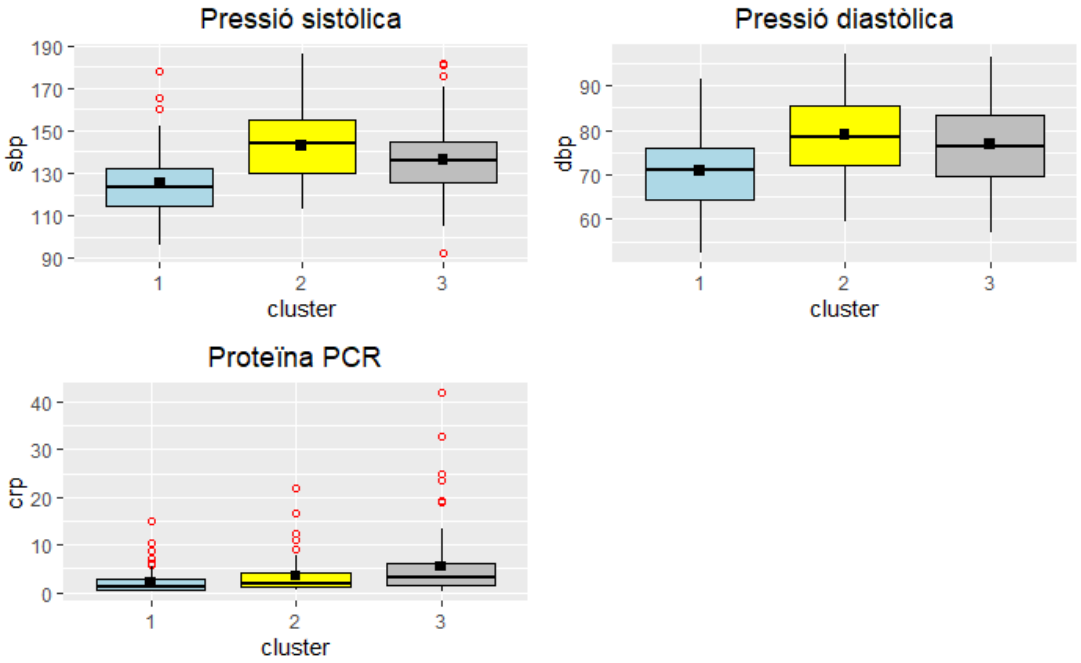
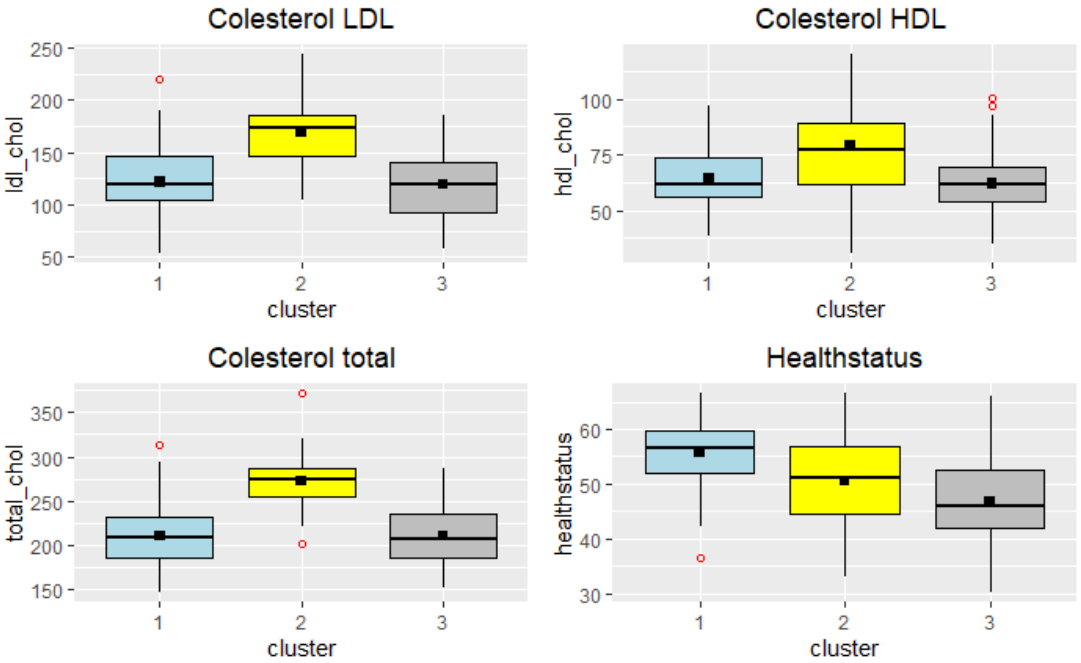
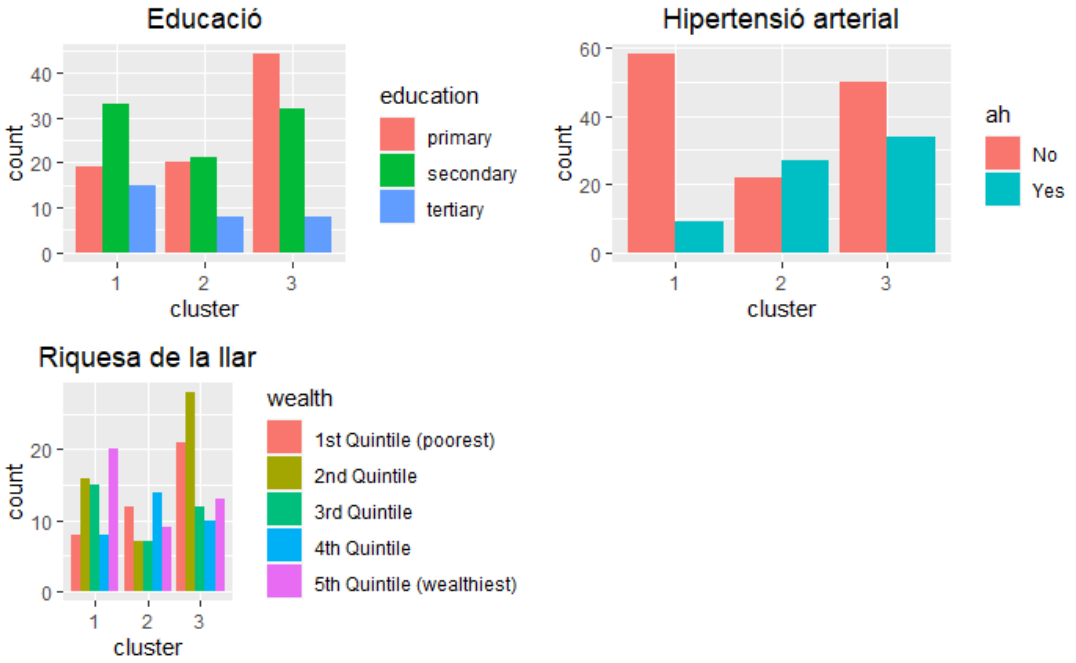


Figura 9.21: Gràfics de caixa de les variables estadísticament iguals pels clústers 1 i 3 (blau i gris respectivament), junt amb la variable healthstatus, estadísticament diferent en tos els clústers.



Per a les variables categòriques, s'observa que el clúster 1 i el clúster 2 presenten valors similars quant al nivell d'educació i els clústers 2 i 3 presenten valors similars per la hipertensió arterial. Finalment, per la riquesa de la llar el clúster 1 és estadísticament diferent del clúster 2, on el clúster 3 es troba entremig dels clústers 1 i 2 (veure Figura 9.22).

Figura 9.22: Gràfics de barres per les tres variables categòriques presentades a la taula 6.15.



- <u>Homes</u>:

A continuació es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.16, referent a les variables numèriques, i la Taula 6.17, referent a les variables categòriques.

Referent a les variables numèriques, es pot veure que els tres clústers són diferents per les variables referents a l'índex de massa, la circumferència de la cintura i el maluc, el colesterol LDL i el colesterol total (veure Figura 9.23).

El clúster 1 i el clúster 3 es comporten estadísticament igual per la pressió sistòlica, els nivells de glucosa, els nivells de proteïna PCR i l'índex de salut referent a la variable healthstatus (veure Figura 9.24).

Pel que fa a la força d'adherència dels individus, s'observa que el clúster 1 i el clúster 2 es comporten de manera similar. Respecte al colesterol HDL succeeix el mateix però pels clústers 2 i 3 (veure Figura 9.25).

L'edat del clúster 2 és estadísticament diferent de l'edat del clúster 3, tenint el clúster 1 al mig d'aquests dos clústers. Pels nivells de triglicèrids són els clústers 1 i 2 els que presenten valors diferents, tenint el clúster 3 entremig d'aquests dos grups(veure Figura 9.26).

Figura 9.23: Gràfics de caixa per 4 de les variables estadísticament diferents en tot els clústers.

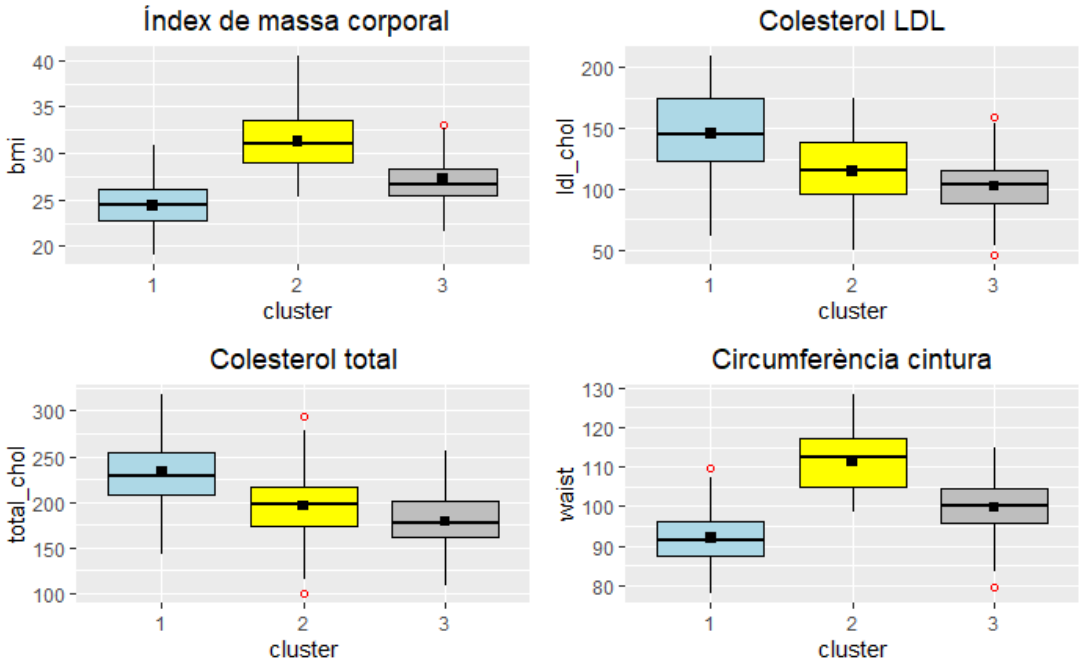


Figura 9.24: Gràfics de caixa de les variables estadísticament iguals pels clústers 1 i 3 (blau i gris respectivament).

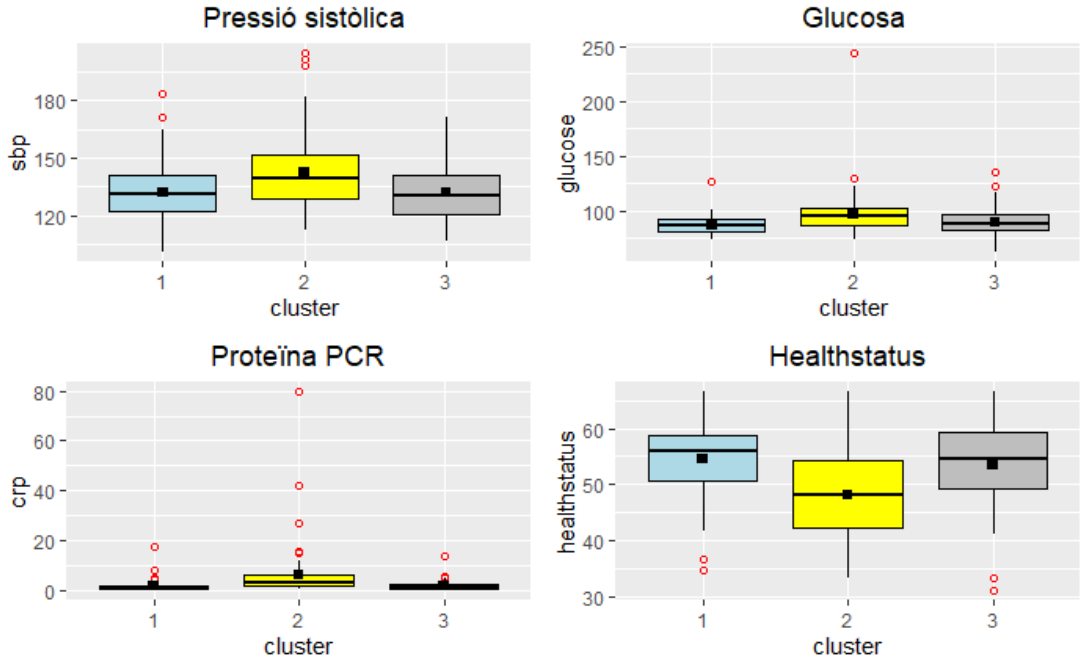


Figura 9.25: Gràfics de caixa de la força d'adherència i el colesterol HDL.

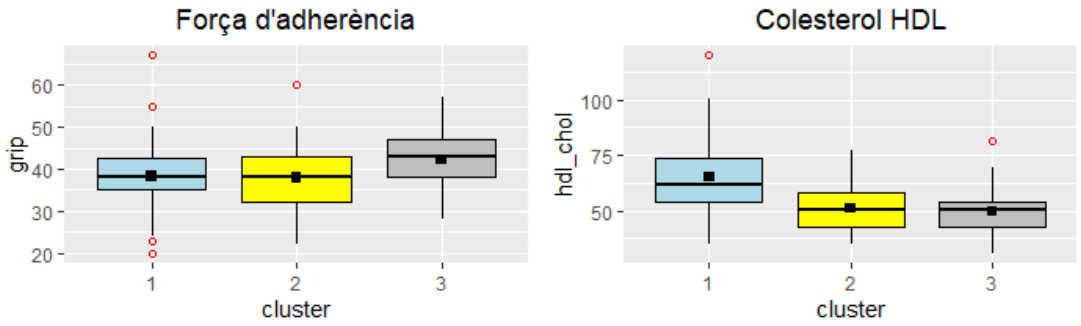
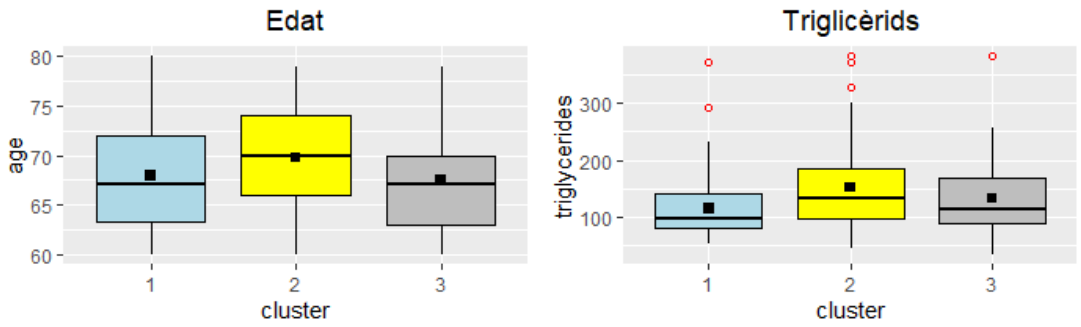
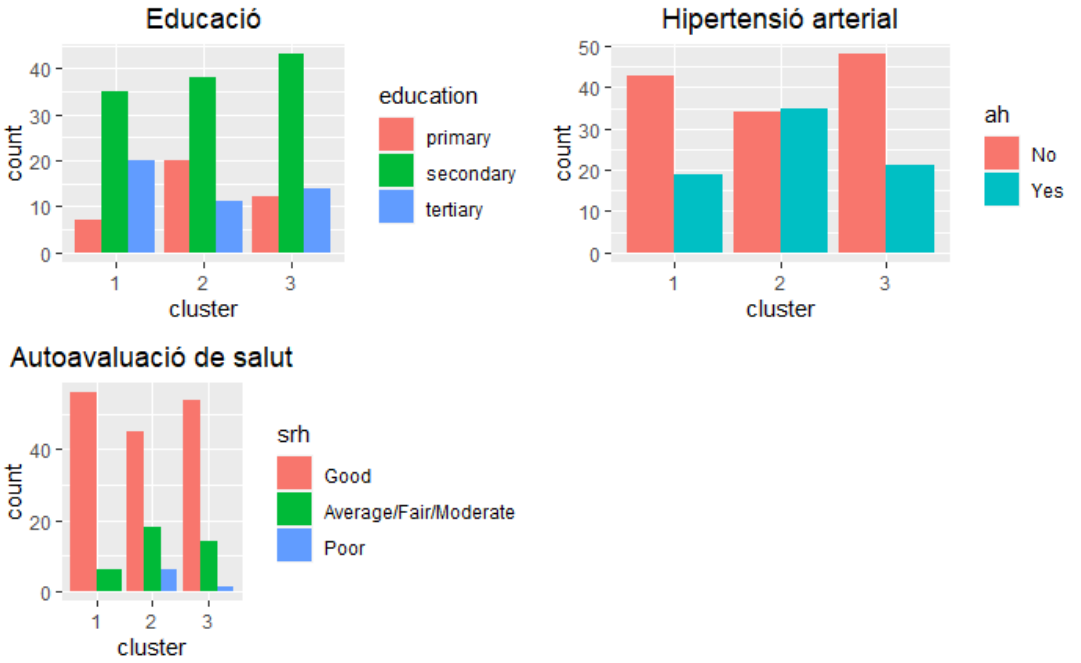


Figura 9.26: Gràfics de caixa de l'edat i els nivells de triglicèrids.



Quant a les variables categòriques, s'aprecia que pels clústers 1 i 2 el nivell d'educació i l'autoavaluació de salut són diferents i el clúster 3 es troba en un punt mitjà. Per la hipertensió arterial els clústers 1 i 3 mostra valors iguals (veure Figura 9.27).

Figura 9.27: Gràfics de barres per les variables que apareixen a la Taula 6.17.



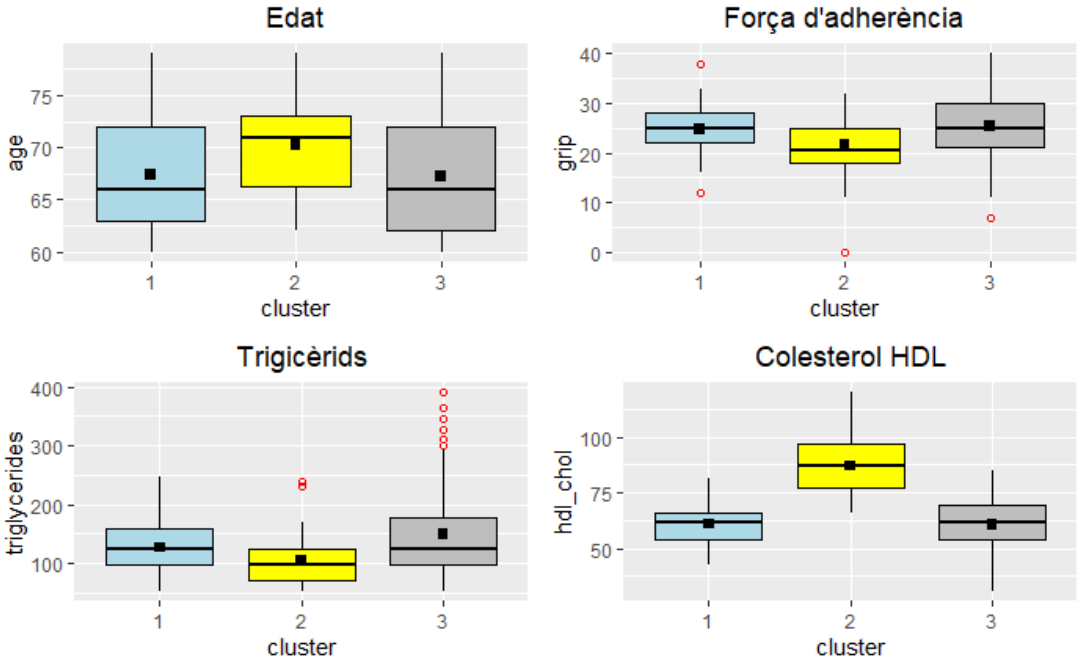
9.1.4. Gaussian Mixture Models

- Dones:

Tot seguit, es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.20, referent a les variables numèriques, i la Taula 6.21, referent a les variables categòriques.

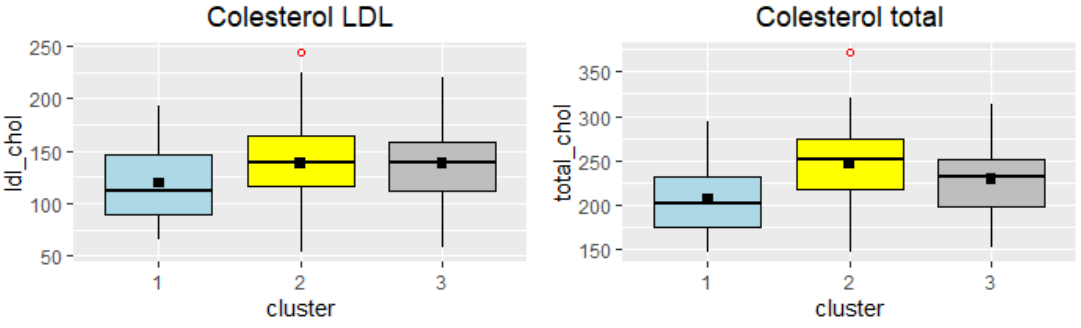
Per les variables numèriques s'observa que els clústers 1 i 3 són iguals per l'edat, la força d'adherència, els triglicèrids i el colesterol HDL (veure Figura 9.28).

Figura 9.28: Gràfics de les variables iguals pel clúster 1 i 3 (blau i gris).



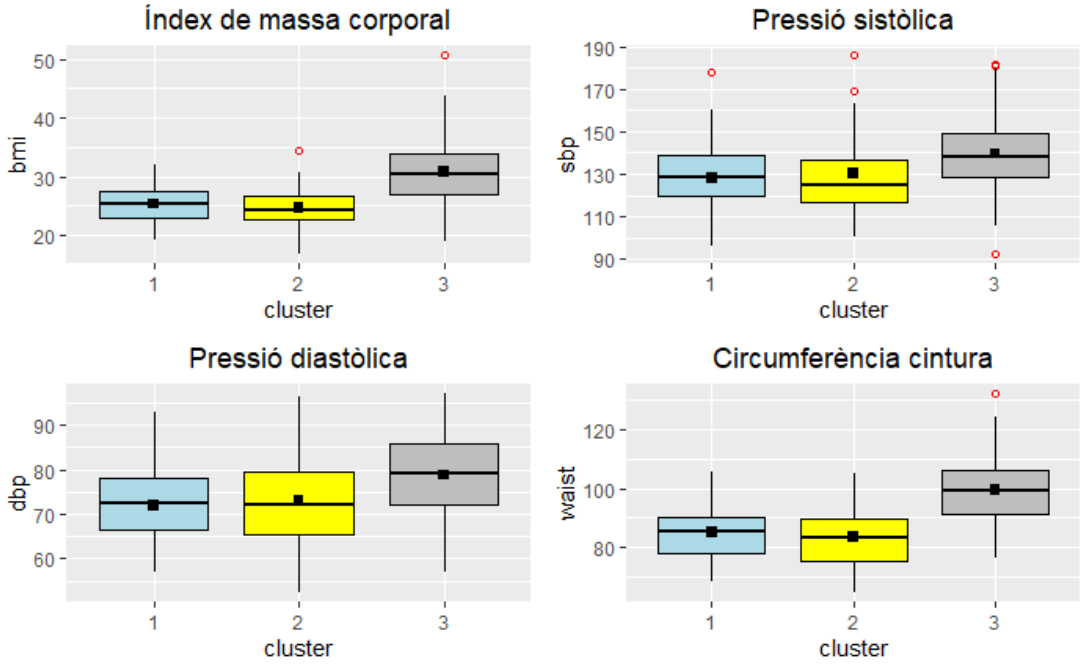
Els clústers 2 i 3 presenten valors estadísticament iguals pel colesterol LDL i el colesterol total (veure Figura 9.29).

Figura 9.29: Gràfics de les variables iguals pel clúster 2 i 3 (groc i gris).



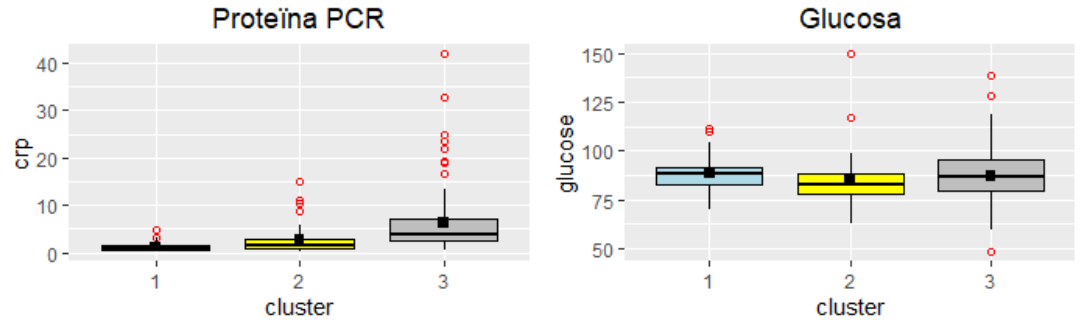
Per altra banda, els clústers 1 i 2 són iguals per l'índex de massa corporal, les circumferències del maluc i la cintura, la pressió sistòlica i la pressió diastòlica (veure Figura 9.30).

Figura 9.30: Gràfics de les variables iguals pel clúster 1 i 2 (blau i groc).



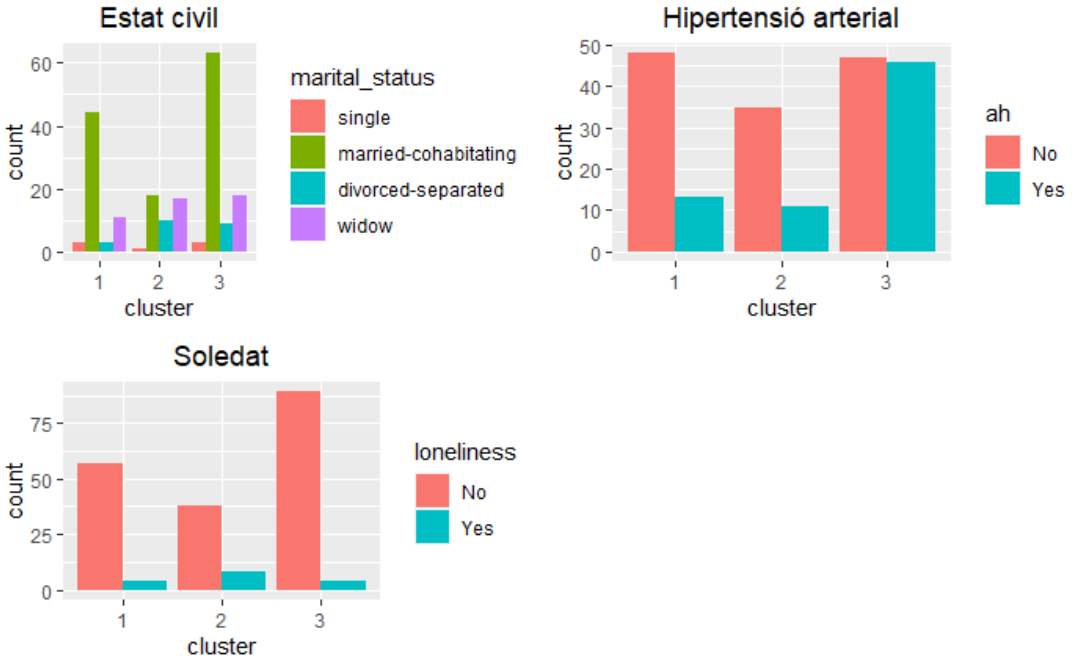
I la variable que és estadísticament diferent per a tots els clústers és la proteïna PCR, mentre que el nivell de glucosa en sang és diferent pels clústers 1 i 2, deixant el clúster 3 amb un valor entremig d'aquests dos clústers (veure Figura 9.31).

Figura 9.31: Gràfics de la proteïna PCR i la glucosa.



Respecte a les variables categòriques, s'aprecia que els clústers 1 i 3 prenen valors iguals d'estat civil, el clúster 1 és igual al clúster 2 per la hipertensió arterial i quant a la soledat els clústers 2 i 3 presenten valors diferents entre ells (veure Figura 9.32).

Figura 9.32: Gràfics de barres per les variables que apareixen a la Taula 6.21.

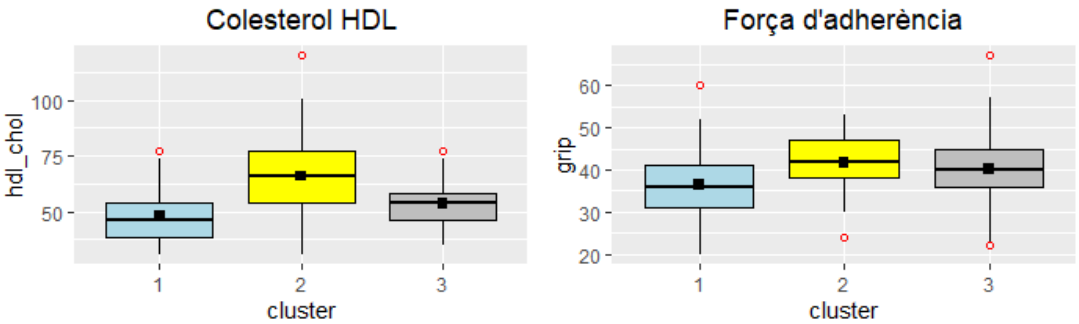


- Homes:

Finalment, es presenten els gràfics de les variables estadísticament diferents entre clústers de la Taula 6.23, referent a les variables numèriques, i la Taula 6.24, referent a les variables categòriques.

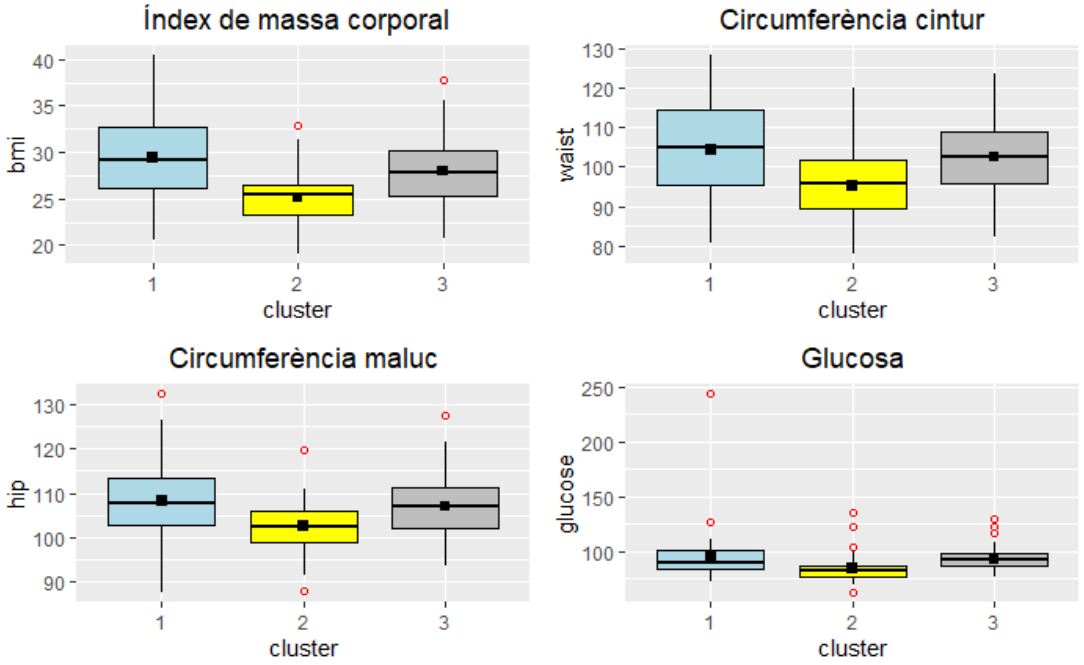
Pel que fa a les variables numèriques, es pot veure que els tres clústers són diferents pels nivells de colesterol HDL, mentre que els clústers 2 i 3 presenten valors estadísticament iguals per la variable referent a la força d'adherència (veure Figura 9.33).

Figura 9.33: Gràfics del colesterol HDL i la força d'adherència.



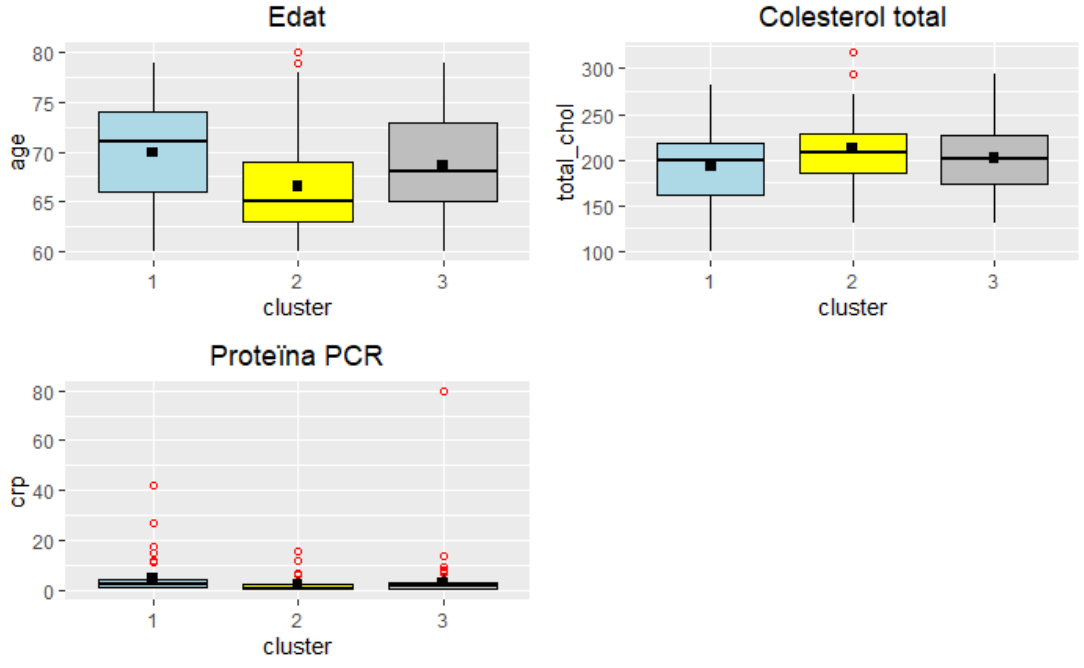
En les variables de l'índex de massa corporal, circumferència de la cintura i del maluc i els nivells de glucosa, els clústers 1 i 3 mostren valors estadísticament iguals (veure Figura 9.34).

Figura 9.34: Gràfics de les variables iguals pel clúster 1 i 3 (blau i gris).



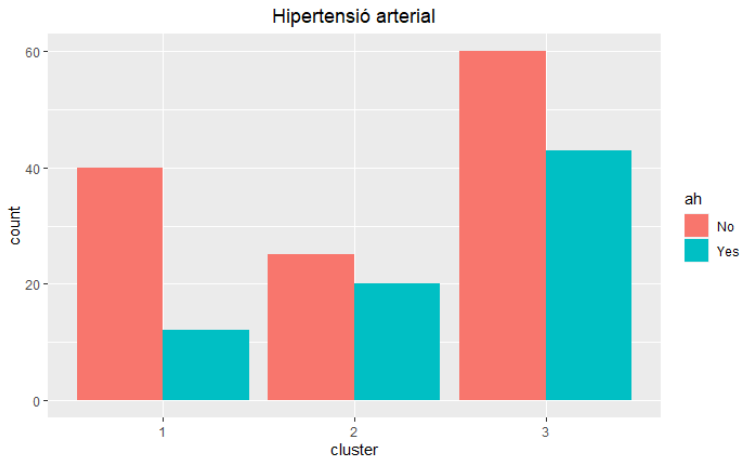
En l'edat, el colesterol total i els nivells de proteïna PCR, els clústers 1 i 2 són estadísticament diferents entre ells, mentre que el clúster 3 presenta valors entremig d'aquests dos grups (veure Figura 9.35).

Figura 9.35: Gràfics de les variables diferents pel clúster 1 i 2 (blau i groc).



Finalment, per les variables categòriques es pot veure que l'única variable significativa és la hipertensió arterial, que presenta valors iguals pels clústers 2 i 3 (veure Figura 9.36).

Figura 9.36: Gràfic de barres per la Hipertensió arterial.



9.2. Codi R

A causa de l'extensió del codi emprat per dur a terme l'anàlisi pràctic del treball, el codi i la base de dades s'ha penjat a un repositori públic de github.

El repositori es pot trobar al següent enllaç: https://github.com/Xventafa7/TFG-CLUTERING-Xavier-Ventayol