Técnicas de Clústering para datos longitudinales. Una aplicación al proyecto Aragón Worker's Health Study (AWHS)

Sara Castel Feced
Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Directores del trabajo:
Tomás Alcalá Nalvaiz
Lina Maldonado Guaje
26 de junio de 2020

Prólogo

Las enfermedades cardiovasculares (ECV) son la primera causa de mortalidad y morbilidad en todo el mundo. Aunque su incidencia varía según países, se calcula que en el año 2017 causaron mundialmente 17.8 millones de muertes [1]. En España la incidencia de ECV es menor que en otros países occidentales, a pesar de que la prevalencia de factores de riesgo cardiovascular (FRCV) tradicionales, como dislipemia, hipertensión, hábito tabáquico y diabetes es mayor [2]. Así, por ejemplo, en los últimos años, en España, ha aumentado la población con sobrepeso y obesidad, por lo que cabría esperar un aumento de ECV [3].

El estudio AWHS (Aragon Workers' Health Study) [3] fue diseñado para evaluar la evolución de FRCV y su asociación con la prevalencia de ECV, en una población de mediana edad en España. En este estudio se recopila anualmente, desde el año 2009, la información sobre factores de riesgo cardiovascular y datos clínicos en una cohorte de unos 5000 trabajadores de una fábrica de coches establecida en Figueruelas (Zaragoza).

La identificación de perfiles de pacientes, en función de la evolución de los factores de riesgo cardiovascular y del riesgo de sufrir ECV, puede ayudar al desarrollo de estrategias de prevención cardiovascular, además de aportar información útil para elaborar nuevas hipótesis de trabajo.

En relación con la frecuencia de FRCV, existen diferentes perfiles de pacientes que evolucionan a lo largo del tiempo. En la identificación de estos perfiles resultan útiles las técnicas de lo que se denomina aprendizaje no supervisado.

El denominado aprendizaje no supervisado es un tipo de aprendizaje automático en el que los datos no están en categorías o grupos definidos, y tiene como objetivo principal el estudio de la estructura intrínseca de los datos [4].

Dentro del aprendizaje no supervisado se encuentran las técnicas de clústering, las cuales tratan de agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. Las técnicas de clústering estudian tres tipos de problemas [5]. El primero es la partición de datos, en la que se trata de dividir individuos que sospechamos heterogéneos, de forma que cada elemento esté en un sólo grupo, todo elemento se asigne a algún grupo y que cada grupo sea internamente homogéneo. Otro problema es la construcción de jerarquías, en la que se pretende estructurar los elementos de forma jerárquica por su similitud. Y, finalmente, la clasificación de variables, ya que en estudios en los que se incluyen muchas variables es interesante dividirlas en grupos para disminuir así la dimensión del problema. Entre los métodos clásicos de partición se encuentra el algoritmo de k-medias, un método iterativo descendente que tiene cuatro etapas y se utiliza cuando todas las variables son cuantitativas. Este algoritmo puede ser aplicado a los datos del estudio AWHS, aunque este tiene ciertas particularidades que harán necesarias algunas modificaciones del algoritmo tradicional.

El estudio AWHS mencionado es un estudio de cohortes longitudinal en el que cada variable está medida en distintos instantes y, para cada individuo, estas variables evolucionan a lo largo del tiempo. El método estándar de trabajo con las trayectorias de las variables consiste en aplicar las técnicas de clústering a la trayectoria de cada variable por separado. No obstante, dado que la mayoría de estudios trabajan con múltiples variables, resulta interesante estudiar la evolución conjunta de las trayectorias de varias variables al mismo tiempo.

Tradicionalmente, el estudio de la evolución conjunta de las trayectorias de distintas variables se ha hecho agrupando las trayectorias de cada variable por separado y después combinando las particiones obtenidas, pero esta aproximación tiene ciertas limitaciones por dos razones importantes [6]. Una de

las utilidades de los métodos de clasificación es convertir datos continuos en categóricos para utilizar las categorías extraídas, por ejemplo, en modelos de regresión. Si dos variables están relacionadas en algún aspecto, las particiones obtenidas para cada variable por separado estarán correladas, y por tanto incluir ambas particiones en un modelo de regresión hará que éste sea inestable. Otra limitación es que la partición final obtenida con este método no es capaz de detectar grupos en los que la evolución conjunta de dos variables sea compleja.

Estas limitaciones evidencian la necesidad de la utilización de métodos de clústering en los que se tengan en cuenta las trayectorias de las distintas variables simultáneamente. Ante esta necesidad surge el algoritmo kml3d [7, 8], implementado en el paquete de R KmL3D, que es un algoritmo basado en la técnica de k-medias, que hace particiones teniendo en cuenta las trayectorias de varias variables al mismo tiempo. Esta técnica en R fue planteada por primera vez en el año 2013 por Genolini y Falissard en el artículo KmL3D: A non-parametric algorithm for clustering joint trajectories [6], hace tan sólo 7 años.

El presente trabajo trata de aplicar estas nuevas técnicas de clústering a los datos recogidos en tres momentos distintos del estudio AWHS. Su objetivo es la agrupación de individuos en función de la evolución de los factores de riesgo cardiovascular que presentan y del índice de riesgo de sufrir enfermedad cardiovascular.

El resto del trabajo se estructura en tres capítulos principales. El primer capítulo presenta el algoritmo de k-medias, tanto tradicional como el utilizado para datos longitudinales. El segundo capítulo es un análisis descriptivo de los datos que posteriormente se utilizarán para el análisis de clústeres. Finalmente, el tercer capítulo presenta los resultados obtenidos de la aplicación del algoritmo de k-medias para datos longitudinales presentado en el primer capítulo.

Summary

Cardiovascular disease (CVD) is the first cause of death and morbidity worldwide. Although its incidence is different depending on countries, it is estimated that cardiovascular diseases are responsible for an estimated 17.8 million deaths in 2017. In Spain, the incidence of CVD is lower than in other Western countries, in spite of the fact that the prevalence of traditional cardiovascular risk factors (CVRF) as hypertension, smoking and diabetes is higher. Furthermore, Spain has experienced an increase of overweight and obesity, so that the incidence of CVD could rise.

The Aragon Workers' Health Study (AWHS) was designed to evaluate the evolution of CVRF and its partnership with the prevalence of CVD in a population of middle-aged men and women in Spain. The study involves annual information about CVRF and clinical endpoints, from year 2009, of over 5000 workers of a car assembly plant in Figueruelas (Zaragoza).

The identification of patient's profiles can help to develop new strategies in cardiovascular prevention, depending on its CVRF evolution and its risk of suffer CVD. Moreover, it contributes to elaborate new work hypothesis.

Regarding the CVRF frequency, there are different profiles of patients that changes over the years. In the identification of these profiles are useful new techniques which are called unsupervised learning.

Unsupervised learning is a type of machine learning where data are not divided into defined categories or groups, and their main goal is the study of data's intrinsic structure.

Clustering techniques are a kind of unsupervised learning that try to gather elements in homogeneous groups according to the similarity between them. These techniques study three types of problems. The first is the partition of data, they try to divide individuals that are thought to be heterogeneous, in order to each element is in a single group, every element is in a group and every group is internally homogeneous. Other problem is the hierarchal construction, in which elements are structured hierarchically attending to its similarity. Finally, the variables classification, due to in some studies there are too many variables, it is useful divide them into groups to decrease the problem dimension. K-means is a classical clustering method, iterative and descendant that has four stages and it is used in studies where the variables are quantitative. This algorithm can be applied to AWHS's data, although this study has some particularities that require some changes in the traditional algorithm.

AWHS is a longitudinal, cohort study where each variable is measured in different times and, for every individual, these variables develop across the time. Normally, the standard method to cluster variable trajectories is to cluster each variable trajectory separately. As many studies use more than one variable, it is interesting to study the variables joint evolution.

Traditionally the way to cluster joint variable trajectories is to cluster each variable trajectory independently, then to consider the combination of the partitions obtained, but this approach is of limited value for two reasons. One advantage of classification methods is to enable the conversion of continuous data into categorical data, after which the categories obtained can be used, for instance in a regression model. If two variables are linked in some way, partitions obtained will be correlated. So the inclusion of both partitions in the regression will lead to instability of the model. Another weakness of the method is that the final partition does not enable detection of groups where the co-evolution of the two variables is complex.

These limitations highlight the need for a clustering method that considers several variable trajectories simultaneously. Because of that need kml3d algorithm was created, implemented in R package KmL3D, that is based on the k-means algorithm and works jointly on several variable trajectories.

In this work new clustering techniques are applied to AWHS data. The objective is to group individuals according to its CVRF and the risk of suffer CVD.

The results show that individuals divide in two or three groups according to three different quality criteria. Regarding the distribution in two groups, it shows that first group gather younger individuals with better CVRF and lower risk of CVD and, on the contrary, second group collect older individuals with worst CVRF and higher risk of CVD. The distribution in three groups shows that individuals in the first group are less than one year older than individuals in the second group, CVRF are worst in the second, but the risk of CVD is lower in this group than in the first one. Finally, the third group in this distribution gathers the youngest individuals and the ones with the best CVRF and lowest risk of CVD.

The memory of the work is organized in three main chapters. In the first one the k-means algorithm is developed, both classical and the one for longitudinal data. In the second one a descriptive analysis of AWHS data is presented. And finally, the third chapter presents the results obtained after application of k-means for longitudinal data.

Índice general

Índice de tablas

Índice de figuras

Lista de abreviaturas

AWHS Aragon Workers' Health Study.

EC Enfermedad Coronaria.

ECV Enfermedad Cadiovascular.

FRCV Factores de Riesgo Cardiovascular.

IMC Índice de Masa Corporal.

No-EC Enfermedad No Coronaria.

PC Perímetro de Cintura.

SALUD Sistema Aragonés de Salud.

TAD Tensión Arterial Diastólica.

TAS Tensión Arterial Sistólica.

Capítulo 1

Algoritmo de k-medias

El análisis de conglomerados o clúster es una técnica multivariante que tiene como objetivo agrupar o clasificar, una colección de elementos en subconjuntos o clústeres. Normalmente, se utiliza para agrupar observaciones pero también se pueden usar para agrupar variables [4]. Además, dentro de las técnicas de clústering se pueden encontrar tres grupos en función de sus objetivos [5]. El primer grupo busca la partición de datos tratando de dividir individuos que se sospechan heterogéneos, de forma que cada elemento esté en un sólo grupo, todo elemento se asigne a algún grupo y que cada grupo sea internamente homogéneo. Otro grupo busca la construcción de jerarquías, en la que se pretende estructurar los elementos de forma jerárquica por su similitud. Y finalmente, el tercer grupo trata de clasificar variables, ya que en problemas en los que se dispone de una gran cantidad de variables resulta interesante dividirlas en grupos para disminuir la dimensión del problema. El presente trabajo se centrará en el primer grupo, ya que para el conjunto de datos del que se dispone y el objetivo marcado resulta el más pertinente.

Como una buena partición se considera aquella que consigue agrupar a los elementos de forma que los que se encuentran dentro del mismo clúster están más próximos unos de otros que de los elementos que están en otro clúster. Es decir, una buena partición cumple el criterio de que los grupos son homogéneos internamente y a la vez son heterogéneos entre ellos. Así pues un punto fundamental en las técnicas de clústeres será el cálculo de distancias entre elementos, ya que las agrupaciones resultantes se basarán en estas medidas.

Las diferentes técnicas de clúster fueron desarrolladas principalmente a partir de los años 70. Este impulso y crecimiento se debe a la aparición de los ordenadores, que transformó radicalmente los métodos de análisis multivariante. MacQueen introdujo por primera vez en 1967 el algoritmo de las k-medias, un método de análisis de clúster que se ha convertido en uno de los más utilizados actualmente [5].

Como se ha dicho, el método de k- medias es uno de los más extendidos dentro de estos análisis de conglomerados o clúster y trata de asignar cada observación a un único grupo. Como técnica de agrupación de variables, se utiliza cuando todas las variables son cuantitativas y se enmarca dentro de los llamados métodos iterativos descendentes. Consta de cuatro etapas principales:

1. Selección de k puntos que se marcan como centros de los grupos.

2. Cálculo de las distancias de cada punto a los puntos seleccionados como centros.

3. Asignación de cada punto al grupo cuyo centro esté más cerca. Cada vez que un punto se asigna a un grupo se recalcula el centro de tal grupo.

4. Comprobación de un criterio de optimalidad, definido previamente. Si una nueva reasignación mejora este criterio se regresa al paso 3.

El proceso termina cuando el criterio de optimalidad no mejora. Para la aplicación de este algoritmo es necesario hacer un análisis previo del número de grupos en los que se va a dividir a la población.

1.1. Cálculo de distancias

Sea un conjunto de datos compuesto por $i = 1,...,n$ observaciones y $j = 1,...,p$ variables entonces la distancia entre las observaciones $x_{i.}$ y $x_{i\prime.}$, $d(x_{i.},x_{i\prime.})$, se define como sigue [4].

Sea *Dist* una función distancia y $\|\cdot\|$ la función norma, entonces $d_j(x_{ij},x_{i\prime j}) = Dist(x_{ij},x_{i\prime j})$ y el resultado será un vector de las distancias de cada variable entre los individuos i e i':

$$(d_1(x_{i1},x_{i\prime 1}),d_2(x_{i2},x_{i\prime 2}),...,d_p(x_{ip},x_{i\prime p})).$$

Finalmente, utilizaremos la función norma $\|\cdot\|$ para combinar estas p distancias:

$$d(x_{i.},x_{i\prime.}) = \|d_1(x_{i1},x_{i\prime 1}),d_2(x_{i2},x_{i\prime 2}),...,d_p(x_{ip},x_{i\prime p})\|.$$

Las funciones *Dist* y $\|\cdot\|$ más extendidas son la distancia Euclídea y la norma Euclídea al cuadrado, por lo que tendremos:

$$\begin{gathered} d_j(x_{ij},x_{i\prime j}) = |x_{ij}-x_{i\prime j}| \\ \Downarrow \\ d(x_{i.},x_{i\prime.}) = \sum_{j=1}^{p}(x_{ij}-x_{i\prime j})^2. \end{gathered}$$

Aunque estas son las más extendidas, hay otras que se utilizan para medir distancias y que pueden llevar a resultados diferentes. Para variables cuantitativas, que son las que utilizaremos en el análisis de k-medias, otras alternativas son el error absoluto o la correlación:

$$\rho(x_i,x_{i\prime}) = \frac{\sum_j(x_{ij}-\bar{x}_i)(x_{i'j}-\bar{x}_{i'})}{\sqrt{\sum_j(x_{ij}-\bar{x}_i)^2\sum_j(x_{i'j}-\bar{x}_{i'})^2}}$$

siendo $\bar{x}_i = \sum_j x_{ij}/p$. Si los datos han sido previamente estandarizados sabemos que $\sum_j(x_{ij}-x_{i'j})^2 = 2(1-\rho(x_i,x_{i'}))$, por lo tanto tenemos que los métodos de clúster basados en la correlación, son equivalentes a los basados en la distancia al cuadrado.

1.2. Implementación del algoritmo

Los métodos de clúster combinatorios, entre los que se encuentra el de las k-medias, precisan de un criterio de optimalidad. Para el método de las k-medias se utilizan distintos criterios entre los que están el minimizar la suma de las distancias entre los individuos dentro de los clúster y el criterio de la traza, los cuales son equivalentes.

El criterio que se utilizará en el presente trabajo es la estrategia de la distancia mínima o similitud máxima[4, 5], que busca minimizar la suma de distancias dentro de los clúster, ya que es el que mejor se ajusta a la notación que se está utilizando. La suma de estas distancias dentro de un clúster dado C se denotará por $W(C)$. Tener en cuenta que ahora las observaciones están divididas en K grupos y cada grupo tiene n_k observaciones, por lo que se trata de minimizar la suma de las distancias entre cada elemento y el centro del grupo. Es decir, en primer lugar se calcula para cada grupo la suma de las distancias entre cada elemento y su centro, y luego se suman los resultados de todos los grupos:

$$W(C) = \sum_{k=1}^{K}\sum_{i.k=1}^{n_k} d(x_{i.k},\bar{x}_{.k}) = \sum_{k=1}^{K}\sum_{i.k=1}^{n_k}\sum_{j=1}^{p}(x_{ijk}-\bar{x}_{jk})^2 \tag{1.1}$$

siendo $d(x_{i.k},\bar{x}_{.k})$ la distancia entre la observación i que está en el grupo k y el centro del grupo, $\bar{x}_{.k}$. Este centro se calcula como la media dentro del grupo de cada variable, es decir, $\bar{x}_{jk} = \sum_{i.k=1}^{n_k} x_{ij}/n_k$.

Así pues, el criterio de optimalidad vendrá dado por:

$$\min_{C} W(C) = \min_{C} \sum_{k=1}^{K} \sum_{i.k=1}^{n_k} d(x_{i.k}, \bar{x}_{.k}). \tag{1.2}$$

Dado que $\sum_{i.k=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2 = n_k s_{jk}^2$ con s_{jk}^2 la varianza de la variable j en el grupo k y que se puede cambiar el orden de los sumatorios, se puede poner la expresión anterior en función de la varianza:

$$W(C) = \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2.$$

La varianza de los grupos es una medida de heterogeneidad, por lo que con este criterio se está minimizando la heterogeneidad de los grupos, que, como se ha dicho antes, es uno de los objetivos en una buena partición.

Para minimizar $W(C)$ es necesario calcular (1.1) para todas las particiones posibles y si se pretende agrupar n observaciones en K grupos, el número de posibles particiones será [10]:

$$S(n,K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} i^n.$$

Demostración. Sea $S(n,K)$ el número de posibles particiones de n individuos en K grupos. No se tiene en cuenta el orden de los individuos dentro del grupo ni el orden de los grupos y no están permitidos los grupos vacíos. Suponiendo que se han obtenido las posibles particiones para $n-1$ individuos, al añadir el individuo n-ésimo, este puede:

1. Formar un grupo de un sólo individuo, el número de particiones entonces será el mismo que el de $n-1$ individuos en $K-1$ grupos.

2. Se puede añadir a cualquiera de las particiones en K grupos de $n-1$ individuos.

Entonces, $S(n,K)$ viene determinado por $S(n-1,K-1)$ y $S(n-1,K)$ de la siguiente forma:

$$S(n,K) = S(n-1,K-1) + KS(n-1,K-1).$$

Además, teniendo en cuenta que no se admiten grupos vacíos se tienen las siguientes condiciones de contorno:

$$S(n,1) = 1, S(1,K) = 1, S(n,K) = 0 \qquad \text{para } K > n$$

Para obtener el valor de $S(n,K)$ es necesario obtener previamente los valores de $\{S(j,p)\}$ para $1 \leq j \leq n-2$, $1 \leq p \leq K$. La solución a este problema son los llamados números de Stirling de segunda especie, los cuales se pueden calcular a partir de la siguiente fórmula explícita:

$$S(n,K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} i^n.$$

□

$S(n,K)$ crece muy rápido conforme aumeta n y para $n = 19$ y $K = 4$ tenemos $S(n,K) \simeq 10^{10}$, por lo que el algoritmo sólo serviría para n muy pequeños, situación que no ocurre en la práctica. Para resolver este inconveniente el algoritmo de las k-medias busca la partición óptima con la restricción de que en cada partición sólo se puede mover un único elemento de un grupo a otro.

El algoritmo de las k-medias es el siguiente:

1. Se parte de una asignación inicial.

2. Se comprueba si moviendo algún elemento de grupo se reduce $W(C)$.

3. Si $W(C)$ se puede reducir moviendo algún elemento, se mueve ese elemento y se vuelven a calcular las medias de los grupos afectados, volviendo al punto 2. Si $W(C)$ no se puede reducir, se termina.

Cada vez que se repiten los pasos 2 y 3 se reduce el valor de (1.2) por lo que la convergencia está asegurada. Además, el resultado del algoritmo puede depender de la asignación inicial dada, por lo que conviene hacer pruebas con distintas asignaciones iniciales y seleccionar con la que tenga como resultado el menor valor de la función objetivo.

Finalmente, este criterio de optimalidad tiene dos propiedades importantes:

- No es invariante a cambios de escala. Por ello, si las variables están medidas en distintas escalas y para evitar que el resultado del algoritmo de las k-medias dependa de cambios debidos a estas diferencias, conviene estandarizarlas. Por el contrario, si las variables están medidas en la misma escala, no es conveniente estandarizarlas ya que una varianza mucho mayor que las demás puede ser porque existan dos grupos en esa variable.
- Al minimizar la distancia euclídea este criterio produce grupos aproximadamente esféricos.

1.3. Determinación del número de grupos

Para aplicar el algoritmo de las k-medias es necesario definir previamente el número de clústeres. Los métodos utilizados para esto se basan normalmente en evaluar el valor de $W(C)$ obtenido en función del número de clústeres, es decir, evalúan la heterogeneidad dentro de los clúster en función del número de éstos [4].

En estos métodos se calcula $\min_C W(C)$ para $K \in \{1,2,\ldots,K_{max}\}$ grupos y para simplificar notación, a partir de ahora se denotará al valor de $W(C)$ para K grupos como W_K. De la forma que se ha definido W_K, cabe esperar que su valor disminuya conforme aumenta el valor de K. Si definimos K^* como el número de grupos en el que se divide nuestra población de estudio, entonces al tomar un $K < K^*$ los clústeres que resultan del algoritmo contienen subconjuntos de los grupos que hay realmente. Esto hace que el valor de W_K disminuya sustancialmente conforme aumenta el número de clústeres, es decir, $W_{K+1} \ll W_K$. Sin embargo, una vez se alcanza K^*, para $K > K^*$, al menos uno de los grupos subyacentes se repartirá en dos, lo que hará que el valor de W_K sufra un menor descenso al aumentar K.

Por lo tanto se puede decir que habrá un importante descenso en las sucesivas diferencias, $W_K - W_{K+1}$, para $K = K^*$. Es decir, se cumple que:

$$\{W_K - W_{K+1} | K < K^*\} \ll \{W_K - W_{K+1} | K \geq K^*\}.$$

Así pues, un estimador $\hat{K}^*$ de K^* se obtiene al identificar un *"codo"* en la gráfica de W_K en función de K. En estos principios se basa el *método del codo*, en el cual se representa gráficamente W_K en función de K y se obtiene como número de clústeres aquel en el que se identifica el *"codo"* mencionado anteriormente.

Recientemente se ha propuesto un nuevo método: *método gap* [11], el cual compara la curva $log(W_K)$ con la curva resultante de una distribución de datos aleatorios uniforme repartidos en el mismo rango que nuestros datos. En este método se estima el mejor número de clústeres como aquel en el que la diferencia entre ambas curvas es mayor. Una de las ventajas de este método es que es capaz de detectar incluso cuando los datos se agrupan en un sólo clúster, en tal caso indicará como mejor número de clústeres uno. En este punto la mayoría de métodos suelen fallar.

En el *método gap*, al representar gráficamente la diferencia entre ambas curvas en función del número de clústeres, aparecen también representados con unas barras verticales los rangos de error dados por:

$$s'_K = s_K \sqrt{1 + 1/N}$$

siendo s_K la desviación estándar de W_K tras N simulaciones de datos uniformemente distribuidos. Si $G(K)$ es la curva de Gap para K clústeres, la regla formal para estimar K^* es

$$K^* = \arg\min_K \{K | G(K) \geq G(K+1) - s'_{K+1}\}$$

Un último método que se utiliza también para determinar el número óptimo de clústeres es el *método de la silueta (Silhouette method)* [12]. Se basa en evaluar un indicador de la distancia entre clústeres y se obtiene calculando el siguiente coeficiente para cada punto i:

$$s(i) = \frac{B - A}{max(A, B)}$$

siendo A la media de la distancia del punto i a los demás puntos del clúster, es decir, una medida de cohesión del clúster, y B la media de la distancia del punto i a cada uno de los puntos del clúster más cercano, una medida de la separación entre clústeres. El coeficiente de la silueta se obtiene calculando el promedio de $s(i)$ para todas las observaciones del conjunto de datos y toma valores entre -1 y 1. Un valor mayor indica que ese número de clústeres es mejor que los demás.

Los coeficientes resultantes de estos métodos también son utilizados en el proceso de validación de un análisis de clústeres.

1.4. Adaptación algoritmo de k-medias para datos longitudinales

Hasta este punto se ha explicado el algoritmo de k-medias como técnica de análisis de clústeres o conglomerados para partir una colección de datos estáticos. En el presente estudio se tiene una colección de sujetos con la particularidad de que sus variables se miden en tres momentos de tiempo distintos, es decir, se trata de un estudio de cohorte longitudinal. Para el análisis de clústeres de conjuntos de datos con estas características se desarrolló el algoritmo kml3d, implementado en el paquete de R KmL3D [7, 8], ideado para conjuntos de datos como el que se presenta en este estudio en los que hay más de una variable medida varias veces.

El algoritmo kml3d constituye una implementación del algoritmo de k-medias diseñado para trabajar con varias variables medidas en distintos instantes. Además, el paquete de R KmL3D da distintas herramientas para trabajar con datos longitudinales como métodos de imputación, métodos para definir las condiciones de inicialización en k-medias y criterios de validez de los clústeres. También ofrece gráficos en 2-D y 3-D para representar resultados.

1.4.1. Notación

En esta sección se va a trabajar con un conjunto de datos que se corresponden con los registros de p distintas variables para n sujetos en t instantes de tiempo distintos. En la Tabla 1.1 se resume la notación que se utilizará a lo largo de la sección. Dado que en este estudio se cuenta con 4165 sujetos y 5 variables medidas en 3 instantes, se tendrán 4165 trayectorias conjuntas compuestas cada una de 5 trayectorias simples que serán vectores de 3 elementos. Es decir, se contará con 4165 matrices de dimensiones 5×3 cada una.

1.4.2. Cálculo de distancias con datos longitudinales

Como se ha visto previamente, para la implementación del algoritmo de k-medias es imprescindible el cálculo de distancias entre dos individuos, que en este caso se corresponde con el cálculo de la distancia entre trayectorias conjuntas. Esto se utilizará para determinar cuál es el clúster más cercano a un individuo. Para la determinación del centro de cada clúster, con este tipo de datos, se calculará la media de las trayectorias conjuntas de los individuos que se encuentran dentro del clúster.

En la primera parte de este capítulo se ha visto que, en el algoritmo de k-medias, dos puntos importantes son el cálculo de distancias entre un elemento y los centros de los clústeres y el cálculo de los

Notación	Denominación	Explicación	Tipo de dato
y_{ijX}	Registro.	Valor de la variable X para el sujeto i en el instante j.	Escalar.
$y_{i.X} = (y_{i1X}, y_{i2X}, ..., y_{itX})$	Trayectoria simple.	Valores de la variable X para el sujeto i en los distintos instantes.	Vector de t elementos.
$y_{i..} = \begin{pmatrix} y_{i11} & \cdots & y_{it1} \\ \vdots & & \\ y_{i1p} & \cdots & y_{itp} \end{pmatrix}$	Trayectoria conjunta.	Conjunto de trayectorias simples.	Matriz $p \times t$ cuyas filas son las trayectorias simples y las columnas corresponden al estado de un individuo i dado en el instante j.
En un conjunto S de n sujetos para los cuales se han registradas p variables medidas en t instantes de tiempo distintos entonces $X = 1, ..., p$, $i = 1, ..., n$ y $j = 1, ..., t$.			

Tabla 1.1: Resumen notación utilizada en esta sección.

centros de cada clúster. En el caso de datos longitudinales, los centros de los clústeres son las medias de las trayectorias de los individuos que están en cada grupo y para determinar el clúster más cercano a un elemento es necesario calcular la distancia entre la trayectoria del individuo y el centro del grupo. Es por esto que un concepto fundamental en la implementación del algoritmo de k-medias para datos logitudinales es la distancia entre dos trayectorias conjuntas que correspondan a dos individuos distintos.

Dadas las trayectorias conjuntas de dos individuos $y_{i..}$, $y_{i\prime..}$ se va a definir pues la distancia entre estas, $d(y_{i..}, y_{i\prime..})$. Como se ha visto en la Tabla 1.1 esta distancia entre dos trayectorias conjuntas equivale a buscar la distancia entre dos matrices. Una posibilidad para definir tal distancia es considerar las t columnas de ambas matrices por separado, calcular las t distancias entre las parejas de columnas y posteriormente combinar esas t distancias utilizando alguna función dada. Otra posibilidad sería replicar el mismo proceso pero separando las matrices en las p filas en lugar de por las columnas.

(a) **Cálculo de la distancia d entre dos trayectorias conjuntas $y_{i..}$, $y_{i\prime..}$ por columnas**: se toma una columna determinada j de ambas matrices o lo que es lo mismo, se toman los valores de las distintas variables de ambos individuos para un instante de tiempo j fijado. Así la distancia entre las dos columnas $y_{ij.}$ y $y_{i\prime j.}$ vendrá dada por $d_{j.}(y_{ij.}, y_{i\prime j.}) = Dist(y_{ij.}, y_{i\prime j.})$ y el resultado será un vector de distancias:

$$(d_{1.}(y_{i1.}, y_{i\prime 1.}), d_{2.}(y_{i2.}, y_{i\prime 2.}), ..., d_{t.}(y_{it.}, y_{i\prime t.})).$$

Para combinar estas t distancias se utilizará la función que algebraicamente se define como norma $\|\cdot\|$ del vector de distancias. Así pues, finalmente se tiene que la distancia entre dos trayectorias conjuntas $y_{i..}$ y $y_{i\prime..}$será:

$$d(y_{i..}, y_{i\prime..}) = \|(d_{1.}(y_{i1.}, y_{i\prime 1.}), d_{2.}(y_{i2.}, y_{i\prime 2.}), ..., d_{t.}(y_{it.}, y_{i\prime t.}))\|.$$

(b) **Cálculo de la distancia d entre dos trayectorias conjuntas $y_{i..}$, $y_{i\prime..}$ por filas**: Para el segundo método se tiene que tomar una fila X determinada de ambas matrices que se corresponden con las mediciones de la variable X en los t instantes de tiempo para los individuos i e $i\prime$. Ahora procedemos de la misma forma que antes y se define la distancia entre ambas filas como $d_{.X}(y_{i.X}, y_{i\prime.X}) = Dist(y_{i.X}, y_{i\prime.X})$, obteniendo así un vector de p distancias

$$(d_{.1}(y_{i.1}, y_{i\prime.1}), d_{.2}(y_{i.2}, y_{i\prime.2}), ..., d_{.p}(y_{i.p}, y_{i\prime.p})).$$

Se combinarán estas p distancias utilizando la función norma $\|\cdot\|$ como en el método anterior y así se obtiene una nueva definición de distancia

$$d'(y_{i..}, y_{i\prime..}) = \|(d_{.1}(y_{i.1}, y_{i\prime.1}), d_{.2}(y_{i.2}, y_{i\prime.2}), ..., d_{.p}(y_{i.p}, y_{i\prime.p}))\|.$$

De la elección de la norma $\|\cdot\|$ y de la distancia *Dist* se obtienen una amplia variedad de distancias entre dos trayectorias conjuntas. Sin embargo, si se toma como $\|\cdot\|$ la p- norma estándar y como la función *Dist* la distancia de Minkowski con parámetro p dada por:

$$Dist(y_{i..}, y_{i\prime..}) = \sqrt[p]{\sum_{j,X} |y_{ijX} - y_{i\prime jX}|^p}$$

se puede demostrar que las dos definiciones de distancia dan el mismo resultado.

Demostración:

$$\begin{aligned} d(y_{i..}, y_{i\prime..}) = \sqrt[p]{\sum_j (d_{j.}(y_{ij.}, y_{i\prime j.}))^p} = \sqrt[p]{\sum_j \left(\sqrt[p]{\sum_X |y_{ijX} - y_{i\prime jX}|^p} \right)^p} = \\ \sqrt[p]{\sum_j \sum_X |y_{ijX} - y_{i\prime jX}|^p} = \sqrt[p]{\sum_X \left(\sqrt[p]{\sum_j |y_{ijX} - y_{i\prime jX}|^p} \right)^p} = \\ \sqrt[p]{\sum_X (d_{.X}(y_{i.X}, y_{i\prime.X}))^p} = d'(y_{i..}, y_{i\prime..}) \end{aligned} \tag{1.3}$$

□

Como se ha dicho previamente, la distancia Euclídea, que es la definida como la de Minkowski con parámetro $p = 2$, es la más utilizada en el algoritmo de las k-medias y es la que utiliza el paquete KmL3D por defecto.

1.4.3. Estandarización

Un problema ampliamente discutido y documentado en el análisis de clústeres es la diferencia en las escalas de las variables que se utilizan y una solución es normalizar los datos de todas las variables. Este problema también se extiende a los datos de tipo longitudinal y en este caso se procede normalizando las trayectorias. La diferencia con los datos que no son de tipo longitudinal es que las trayectorias de cada variable no se normalizan en cada instante, sino en conjunto. Esto es, que si se llama $\overline{y_{..X}}$ a la media de $y_{..X}$ y $s_{..X}$ a su desviación típica, entonces la medida de y_{ijX} normalizada será:

$$y'_{ijX} = \frac{y_{ijX} - \overline{y_{..X}}}{s_{..X}}$$

y la trayectoria conjunta normalizada $y'_{i..}$ se obtendrá normalizando cada trayectoria individual, $y'_{i.1}, \ldots, y'_{i.p}$, por separado.

1.4.4. Determinación del número de clústeres

Al igual que en algoritmo de k-medias tradicional, se presentan diferentes índices que ayudan a determinar cuál es el mejor número de clústeres para dividir a la población. En esta sección se verán diferentes índices asociados con la calidad de cada partición que nos ayudarán a determinar el número de clústeres en datos longitudinales. Como ya se mostró, una buena partición se caracteriza porque sus clústeres sean simultáneamente homogéneos y estén bien separados unos de otros. Así pues, los distintos índices se calculan, en su mayoría, dividiendo otros dos estimados previamente: uno que mide la homogeneidad de los clústeres y otro índice que mide la separación entre los distintos clústeres.

Los distintos índices que nos ofrece este paquete son: Criterio Calinski & Harabasz (1974) con dos de sus variantes, la propuesta por Kryszczuk (2010) y la propuesta por Genolini; Criterio de Ray & Turi(1999) y Criterio Davies & Bouldin (1979). Estos cinco son criterios no paramétricos al poder ser calculados sin tener en cuenta ninguna hipótesis. Por otra parte, como criterios paramétricos es posible calcular los métodos BIC y AIC y algunas variantes de éstos. A diferencia de los anteriores, para que estos índices sean válidos las variables deben tener una distribución normal y tener homocedasticidad, es decir, la varianza debe ser igual para todos los instantes y todos los grupos. Otra limitación viene dada por la determinación del número de observaciones. En los estudios clásicos el número de observaciones independientes coincide con el tamaño de la población, pero en estudios longitudinales si el tamaño de la población es n y tenemos t mediciones de cada individuo el número de observaciones será $N = n \cdot t$. En este caso las N observaciones no serán medidas independientes. La decisión de tomar n o N como tamaño de la muestra lleva a la definición de distintos índices AIC y BIC.

Las fórmulas que se utilizan para el cálculo de estos criterios se encuentran en el Anexo B.

1.4.5. Inicializacion del k-medias

Como se se ha comentado previamente, la primera etapa en el método de las k-medias es la selección de k puntos como centro de los grupos para a partir de ellos comenzar el algoritmo. Esta selección es importante ya que determina la partición final resultante en el algoritmo y además de ella depende el tiempo de convergencia. Si un método es capaz de seleccionar una partición inicial cercana a la mejor partición, k-medias convergerá más rápido. Se han desarrollado distintas propuestas de métodos de inicialización que tratan de seleccionar como centros para la configuración inicial los puntos que están más separados unos de otros, al suponer que los puntos que están claramente separados pertenecerán a grupos diferentes. El paquete KmL3D ofrece siete posibles métodos de inicialización que se presentan en el Anexo C.

Capítulo 2

El estudio AWHS. Descripción de la muestra

La muestra para realizar el presente trabajo se obtuvo del AWHS, estudio de cohortes longitudinal y prospectivo, en el que se incluyeron los trabajadores de una fábrica de coches localizada en Figueruelas (Zaragoza) que aceptaron participar en el mismo y firmaron el consentimiento informado correspondiente. La incorporación de trabajadores al estudio se inició en el mes de Febrero de 2009 y finalizó en el mes de Diciembre de 2010. El objetivo de este estudio es, fundamentalmente, estudiar la asociación entre diferentes factores de riesgo y la aparición de ECV. De toda la cohorte se obtuvieron los datos del reconocimiento médico anual realizado en la empresa (N=5678). De una subcohorte se recopilaron datos de estilos de vida, mediante cuestionarios que cumplimentaron los trabajadores; analíticas de sangre y orina y pruebas radiológicas para evaluar la presencia de aterosclerosis subclínica (N=2667).

Para la realización del presente estudio, del total de trabajadores que entraron en la cohorte se seleccionaron sólo los hombres ya que el número de mujeres era muy inferior (N=380). Además se excluyeron aquellos que no tenían o dejaron de tener tarjeta sanitaria en Aragón entre su entrada y el 31 de mayo de 2019 y aquellos con algún evento cardiovascular previo a su entrada en el AWHS. Hubo ocho individuos de los que no se encontraron registros en el Sistema Aragonés de Salud (SALUD) y también se excluyeron.

El objetivo de este trabajo es agrupar a la población de estudio según los FRCV que presentan, el índice de riesgo de sufrir enfermedad cardiovascular en tres momentos de tiempo y analizar los cambios que se han producido a lo largo del periodo estudiado. De cada momento se tomaron los datos de los trabajadores correspondientes a sus reconocimientos médicos y sus analíticas. Dado que estos datos no estaban disponibles para la totalidad de los trabajadores en todos los años, el número de sujetos de cada momento fue diferente. Como primer momento se tomó el primer registro disponible, es decir los datos correspondientes a los años 2009, 2010 y 2011. Como segundo instante se tomó el año 2014 por ser un año intermedio entre el comienzo del estudio y el último año con datos disponibles y por tener la particularidad de que los datos registrados se encontraban bien cumplimentados. Como instante final se tomaron los datos de 2017 por ser el último año del que se disponen datos pero, por la cantidad de datos perdidos encontrados en este registro, se utilizaron también datos del año 2016. En resumen, para el primer corte se contó con la información de 5122 trabajadores recogidos en los años 2009, 2010 y 2011, para el segundo con los de 3891 trabajadores todos recogidos en 2014, y finalmente, se contó con los de 3891 individuos con datos procedentes del año 2016 y 2017. Por último, el criterio de selección utilizado para el estudio corresponde a trabajadores con mediciones en al menos dos de los tres momentos siendo finalmente 4165 los trabajadores seleccionados. La evolución del número de participantes se representa en la Figura 2.1.

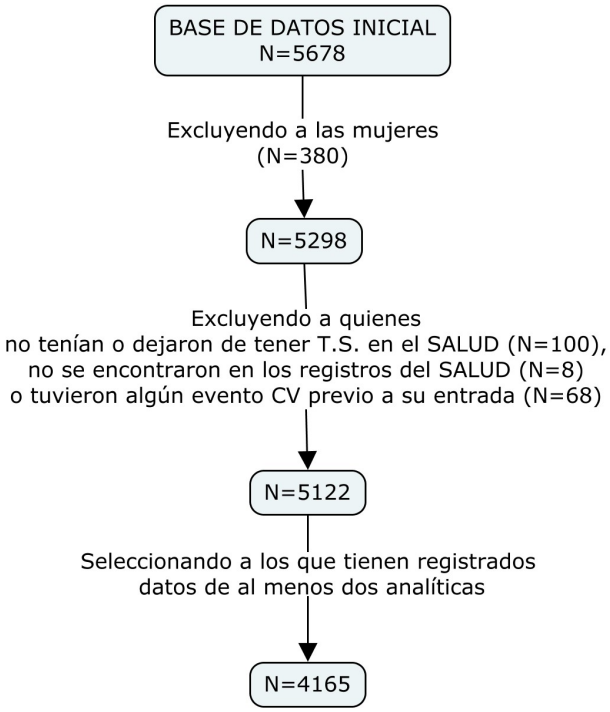


Figura 2.1: Evolución N.

2.1. Variables

Se consideran como principales factores de riesgo cardiovascular, tal y como está descrito en la literatura, el perímetro de cintura (PC), los niveles de colesterol HDL y glucosa en sangre y el índice de masa corporal (IMC) [2, 13]. Así pues en este trabajo se han tenido en cuenta estas 4 variables junto con el índice de riesgo a 10 años de enfermedad cardiovascular.

1. PERÍMETRO DE CINTURA: Variable cuantitativa recogida a partir de los datos registrados en el reconocimiento médico.

2. ÍNDICE DE MASA CORPORAL: Variable cuantitativa calculada a partir del peso en kilogramos y la altura en metros de cada individuo con la fórmula:

$$IMC = \frac{peso}{altura^2}.$$

   Ambas variables se obtuvieron a partir de los reconocimientos médicos, la altura se consideró la del primer reconocimiento para todos los años del estudio y el peso se tuvo en cuenta el registrado en cada año. Esta variable se convirtió también a categórica clasificando según el valor numérico obtenido. Los grupos en los que se dividió fueron:

   - Bajo peso si IMC $< 18{,}5$.
   - Normopeso si IMC $\geq 18{,}5$ y < 25.
   - Sobrepeso si IMC $\geq 25 < 30$.
   - Obesidad si IMC≥ 30.

3. GLUCEMIA Y COLESTEROL HDL: Niveles de glucosa y colesterol HDL en sangre, variables cuantitativas registradas en la analítica realizada anualmente medidas en mg/dL.

4. ÍNDICE DE RIESGO A 10 AÑOS DE ENFERMEDAD CARDIOVASCULAR: Variable cuantitativa calculada según el SCORE aplicable a poblaciones europeas con bajo riesgo cardiovascular [14]. Las variables que se tienen en cuenta son:

- Edad: Variable numérica calculada como la diferencia entre la fecha de nacimiento y la fecha de realización de la analítica de la que se han extraído las demás variables. Para quien faltó la fecha de la analítica, se calculó la edad como la diferencia entre el año de nacimiento y el año del estudio.
- Niveles de colesterol total: Variable numérica obtenida de la analítica anual medida de mg/dL. Dado que para el cálculo del SCORE esta medida debe estar en mmol/L fue necesario el cambio de unidades multiplicando el valor registrado por 0,02586.
- Tensión arterial sistólica (TAS): Variable numérica obtenida a partir del reconocimiento médico anual medida en mmHg.
- Hábito tabáquico: variable categórica, registrando con 0 a aquellos trabajadores no fumadores, con 1 a los fumadores y con 2 a los exfumadores. Para el cálculo del índice de riesgo de ECV se clasifica con 0 a los individuos no fumadores o exfumadores y con 1 a los fumadores por lo que se recodificó juntando las categorías iniciales 0 y 2 como 0. Esta variable sólo se encontraba disponible para los dos primeros cortes, al no tener registro para el tercero se realizó un paso previo para poder calcular el SCORE, en este caso se hizo una imputación con cadenas de Markov.

El cálculo del SCORE se realizó en seis pasos siguiendo la metodología propuesta por Conroya [14]. Los cinco primeros se calculan dos veces: una para calcular el riesgo de enfermedad coronaria (EC) y otra para el de enfermedad no coronaria (No-EC); finalmente se combinan ambos.

a) Cálculo del riesgo subyacente de enfermedad coronaria y no coronaria teniendo en cuenta la edad actual y la edad dentro de 10 años. Estos se calcularon con la fórmula:

$$\begin{aligned} S_0(edad) &= e^{-e^{\alpha}(edad-20)^p} \\ S_0(edad+10) &= e^{-e^{\alpha}(edad-10)^p}. \end{aligned} \tag{2.1}$$

Siendo los valores de beta los que aparecen en la siguiente tabla:

	EC		No-EC	
	α	p	α	p
Hombres	$-22{,}1$	4,71	$-26{,}7$	5,64
Mujeres	$-29{,}8$	6,36	-31	6,62

Tabla 2.1: Valores α y p en ecuación (2.1).

b) Cálculo de la suma ponderada, w, de los factores de riesgo colesterol, hábito tabáquico y TAS:

$$w = \beta_{col} * (colesterol - 6) + \beta_{TAS} * (TAS - 120) + \beta_{fumador} * (fumador). \tag{2.2}$$

Siendo los valores de β:

	EC	No-EC
β_{col}	0,71	0,63
β_{TAS}	0,24	0,02
$\beta_{fumador}$	0,018	0,022

Tabla 2.2: Valores de β en ecuación (2.2).

c) Cálculo de la probabilidad de supervivencia para cada edad y en cada caso, para esto se combinan los valores calculados en (2.1) y en (2.2) de la siguiente forma:

$$\begin{aligned} S(edad) &= S_0(edad)^{e^w} \\ S(edad+10) &= S_0(edad+10)^{e^w}. \end{aligned} \tag{2.3}$$

d) Para cada caso también se calcula la probabilidad de supervivencia en 10 años basada en la probabilidad de supervivencia para cada edad que hemos calculado en el paso *c)*:

$$S_{10}(edad) = \frac{S(edad+10)}{S(edad)}. \tag{2.4}$$

e) Cálculo del riesgo a 10 años para los dos casos:

$$Riesgo(edad)_{10} = 1 - S_{10}(edad). \tag{2.5}$$

f) Finalmente se suman los riesgos a 10 años del paso anterior para la enfermedad coronaria y la no coronaria:

$$RiesgoECV_{10}(edad) = [Riesgo_{10}EC(edad)] + [Riesgo_{10}No - EC(edad)]. \tag{2.6}$$

2.2. Análisis descriptivo

Como se ha explicado al principio de este capítulo de los 5122 trabajadores que quedaron después de aplicar los criterios de exclusión, 4165 contaban con al menos dos registros de los tres momentos. Del primer momento se contó con los datos de todos los trabajadores, del segundo momento se contó con los datos de 3891 y del tercero con los de 3545 trabajadores. Así pues para los individuos que faltaban los datos de uno de los momentos (274 en el segundo instante y 620 en el tercero) se realizaron imputaciones en las variables utilizadas para el análisis de clústeres.

En esta sección en primer lugar se presentará el análisis descriptivo de las variables utilizadas para los tres momentos por separado con los datos originales (sin imputaciones), posteriormente se presentará un breve análisis descriptivo de los datos obtenidos tras hacer las imputaciones, y finalmente se presentará un pequeño análisis descriptivo de los datos para los miembros de la subcohorte.

2.2.1. Primer momento de estudio

En el primer corte se obtuvieron los datos de los 4165 individuos incluidos en el estudio. El análisis descriptivo se puede ver en la Tabla 2.3. Cabe destacar que el peso medio de los individuos incluidos en el estudio fue 81,64 kg, más de la mitad tenían sobrepeso y dentro de los incluidos en el grupo de peso normal 3 casos (0.07 % del total) tenían un peso inferior al recomendado. El porcentaje de fumadores y exfumadores fue muy similar, mientras que el de no fumadores fue aproximadamente un 10 % inferior a los anteriores. Las medias de las variables procedentes de las analíticas de sangre estuvieron dentro de los valores normales. Finalmente, la media del índice del riesgo de sufrir ECV en 10 años fue de 1,56.

Variables cuantitativas	Media (sd)
Edad	48 (8.42)
TAS	126 (14.14)
TAD	83.44 (9.82)
Peso	81.64 (11.47)
PC	96.81 (9.61)
IMC	27.61 (3.54)
Colesterol HDL	52.45 (11)
Colesterol total	212.18 (37.62)
Glucosa	97.7 (18.75)
Índice RECV	1.56 (1.4)

Variables categóricas		% (N)
Hábito tabáquico	Fumador	36.82 (1488)
	No fumador	26.9 (1087)
	Exfumador	36.28 (1466)
IMC por grupos	Peso normal	23.05 (938)
	Sobrepeso	54.63 (2223)
	Obesidad	22.32 (908)

TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, IMC: Índice de Masa Corporal, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 2.3: Análisis descriptivo de las variables en el primer corte.

La media de los valores de estas variables según el hábito tabáquico de los sujetos estudiados la se puede ver en la Tabla 2.4. Cabe destacar que el perímetro de cintura y el IMC fueron mayores en los exfumadores que en los otros dos grupos, siendo entre estos prácticamente iguales. En cuanto a los valores de las analíticas, el colesterol total y la glucosa fueron mayores en los grupos de exfumadores y los valores de colesterol HDL fueron muy similares entre los grupos de no fumadores y exfumadores, siendo menores en el grupo de fumadores. Finalmente, el grupo con menor índice de riesgo de ECV fue el de no fumadores, seguido de el de exfumadores y el grupo de fumadores fue el que tuvo mayor índice. En cuanto al análisis descriptivo clasificando a los sujetos en función de IMC (Tabla 2.5), las

	No fumador N=1086	Fumador N=1488	Exfumador N=1466	p.overall
Edad	46.6 (9.24)	46.9 (9.19)	50.1 (6.12)	<0.001
IMC	27.4 (3.57)	27.1 (3.60)	28.3 (3.36)	<0.001
TAS	125 (13.2)	125 (14.7)	127 (14.2)	<0.001
TAD	82.9 (9.68)	82.3 (10.2)	85.1 (9.37)	<0.001
PC	95.9 (9.75)	95.6 (9.69)	98.8 (9.11)	<0.001
Colesterol HDL	53.6 (10.5)	50.3 (10.7)	53.8 (11.2)	<0.001
Colesterol total	209 (36.7)	211 (39.3)	216 (36.3)	<0.001
Glucosa	97.2 (16.8)	96.0 (18.2)	99.5 (18.9)	<0.001
Índice RECV	1.03 (0.87)	2.18 (1.84)	1.32 (0.88)	<0.001

IMC: Índice de Masa Corporal, TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 2.4: Análisis descriptivo según hábito tabáquico en el primer corte.

medias de los valores de TAS y tensión arterial diastólica (TAD), perímetro de cintura, colesterol total y glucosa fueron mayores en los grupos con un IMC más alto, la media de colesterol HDL siguió la tendencia inversa, siendo menor en los grupos con mayor valor de IMC y finalmente el índice de RECV fue menor en el grupo de trabajadores con normopeso que en los demás y mayor en el grupo de los que sufrían obesidad.

	Normopeso N=937	Sobrepeso N=2223	Obesidad N=908	p.overall
Edad	43.9 (10.6)	48.8 (7.42)	50.0 (6.48)	<0.001
TAS	122 (13.3)	126 (13.9)	130 (14.3)	<0.001
TAD	78.4 (9.27)	83.9 (9.23)	87.5 (9.63)	<0.001
PC	86.2 (5.74)	96.5 (5.66)	108 (7.02)	0.000
Colesterol HDL	56.0 (11.5)	52.2 (10.6)	49.5 (10.3)	<0.001
Colesterol total	201 (38.8)	216 (36.3)	216 (37.5)	<0.001
Glucosa	93.4 (17.6)	97.1 (15.2)	103 (23.6)	<0.001
Índice RECV	1.18 (1.39)	1.63 (1.41)	1.76 (1.32)	<0.001

TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, RECV: Riesgo Enfermedad Cardiovascular.

Tabla 2.5: Análisis descriptivo de las variables según IMC por grupos en el primer corte.

2.2.2. Segundo momento de estudio

En este segundo momento, del total de los trabajadores incluidos se obtuvieron datos completos de 3891. El análisis descriptivo de estos datos se encuentra en la Tabla 2.6. Comparando estos datos con los del corte anterior las medias de las variables perímetro de cintura, peso e IMC fueron similares en ambos cortes. El valor medio del colesterol total fue inferior en este corte, mientras que los de glucosa en sangre y colesterol HDL superiores. La distribución según los grupos de IMC fue prácticamente la

misma en ambos cortes y según el hábito tabáquico aumentó el número de exfumadores y disminuyó el de fumadores y el de no fumadores. Finalmente, la media del índice de RECV fue medio punto superior en este corte.

Variables cuantitativas	Media (sd)
Edad	51.49 (8.27)
TAS	124 (14.25)
TAD	79.8 (9.39)
Peso	82.1 (11.92)
PC	97.3 (10)
IMC	27.77 (3.67)
Colesterol HDL	54.07 (11.3)
Colesterol total	205.93 (34.75)
Glucosa	96.51 (19.46)
Índice RECV	2.05 (1.73)

Variables categóricas		% (N)
Hábito tabáquico	Fumador	32.19 (1235)
	No fumador	24.11 (925)
	Exfumador	43.71 (1677)
IMC por grupos	Peso normal	22.01(846)
	Sobrepeso	54.33 (2088)
	Obesidad	23.65 (909)

TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, IMC: Índice de Masa Corporal, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 2.6: Análisis descriptivo de las variables en el segundo corte.

El análisis descriptivo de estas variables según el hábito tabáquico (Tabla 2.7) siguió los mismos patrones que los descritos en el primer corte. Si comparamos los datos por grupos en ambos momentos, las medias de TAS y TAD fueron menores en este segundo momento en todos los grupos, las de perímetro de cintura e IMC se mantuvieron prácticamente iguales y, en cuanto a los valores de las analíticas, el colesterol HDL aumentó y el colesterol total y la glucosa disminuyeron en todos los casos. Finalmente, el índice de riesgo de enfermedad cardiovascular en 10 años aumentó en todos los grupos.

	No fumador N=925	Fumador N=1235	Exfumador N=1677	p.overall
Edad	49.9 (9.13)	50.0 (9.36)	53.4 (6.34)	<0.001
TAS	122 (14.0)	123 (14.6)	126 (13.9)	<0.001
TAD	78.7 (9.79)	78.6 (9.43)	81.3 (8.95)	<0.001
PC	95.5 (10.1)	95.8 (9.91)	99.4 (9.72)	<0.001
IMC	27.4 (3.67)	27.1 (3.66)	28.5 (3.53)	<0.001
Colesterol HDL	55.2 (10.9)	51.8 (10.8)	55.2 (11.6)	<0.001
Colesterol total	204 (33.5)	204 (34.5)	208 (35.5)	0.002
Glucosa	95.2 (16.7)	94.7 (18.7)	98.5 (21.2)	<0.001
Índice RECV	1.40 (1.14)	2.86 (2.27)	1.82 (1.25)	<0.001

IMC: Índice de Masa Corporal, TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 2.7: Análisis descriptivo en el segundo corte según hábito tabáquico.

En cuanto a los valores medios de las variables según el grupo de IMC (Tabla 2.8), también presentaron las mismas tendencias que en el corte anterior. En la comparación entre ambos cortes, dentro de los grupos, se observa que la TAS y TAD sufrieron un leve descenso entre ambos años y el perímetro de cintura se mantuvo. De los valores de la analítica, el colesterol HDL aumentó levemente, el colesterol total se mantuvo igual y los niveles de glucosa disminuyeron también muy levemente. El índice de riesgo de enfermedad también aumentó dentro de los grupos.

	Normopeso N=846	Sobrepeso N=2088	Obesidad N=909	p.overall
Edad	48.2 (10.2)	51.9 (7.83)	53.5 (6.18)	<0.001
TAS	118 (13.9)	124 (13.6)	130 (14.0)	<0.001
TAD	74.8 (9.43)	79.8 (8.69)	84.3 (8.52)	<0.001
PC	86.3 (5.85)	96.5 (5.69)	109 (7.54)	0.000
Colesterol HDL	57.5 (11.7)	54.2 (11.1)	50.9 (10.5)	<0.001
Colesterol total	202 (34.8)	207 (34.2)	207 (36.1)	0.002
Glucosa	91.8 (13.5)	95.8 (18.3)	102 (24.9)	<0.001
Índice RECV	1.67 (1.73)	2.09 (1.76)	2.33 (1.60)	<0.001

TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, RECV: Riesgo Enfermedad Cardiovascular.

Tabla 2.8: Análisis descriptivo según grupos de IMC en el segundo corte.

2.2.3. Tercer momento del estudio

En este corte se contó con las analíticas de 3545 trabajadores, el análisis descriptivo de estos datos se puede ver en la Tabla 2.9 y sigue una tendencia parecida a los anteriores cortes. Las medias de TAS, TAD, perímetro de cintura, IMC y peso fueron parecidas a las del corte anterior, sin embargo, las variables extraídas de las analíticas (colesterol HDL, total y glucosa) aumentaron levemente desde el corte anterior. La distribución en los grupos por IMC se mantuvo. Finalmente, la media del índice de riesgo de ECV aumentó muy levemente respecto al corte anterior.

Variables cuantitativas	Media(sd)
Edad	53 (8.25)
TAS	128.89 (15)
TAD	81.36 (9.68)
Peso	82.66 (12.38)
PC	97.73 (10.53)
IMC	27.84 (3.8)
Colesterol HDL	51 (12.4)
Colesterol total	187.96 (32.85)
Glucosa	88.06 (18.6)
Índice RECV	2.09 (1.74)

Variables categóricas		% (N)
IMC por grupos	Peso normal	22.38(762)
	Sobrepeso	54.25 (1813)
	Obesidad	24.38 (830)
Hábito tabáquico (Imputado)	Fumador	32.65 (1156)
	No fumador	24.09 (853)
	Exfumador	43.26 (1532)

TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, IMC: Índice de Masa Corporal, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 2.9: Análisis descriptivo de las variables en el tercer corte.

En cuanto al análisis descriptivo de las variables según el hábito tabáquico en este caso no resulta interesante ya que para este corte esta variable no estaba registrada y los datos utilizados para el cálculo del SCORE de riesgo de ECV se obtuvieron con una imputación.

Finalmente, el análisis descriptivo según el grupo de IMC se puede ver en la Tabla 2.10 y la tendencia de las variables dentro de los grupos es la misma que en los cortes anteriores. Comparando las medias obtenidas dentro de los grupos en este corte y en el anterior se puede ver que los valores medios de TAS y TAD aumentan levemente en los cuatro grupos, que el perímetro de cintura se mantuvo igual en todos los casos y que los valores de la analítica disminuyeron todos respecto al corte anterior. La media del índice de riesgo de ECV se mantuvo en todos los grupos.

	Normopeso N=762	Sobrepeso N=1813	Obesidad N=830	p.overall
Edad	50.0 (10.0)	53.6 (7.74)	54.8 (6.36)	<0.001
TAS	123 (14.5)	129 (14.5)	134 (15.1)	<0.001
TAD	76.7 (9.58)	81.5 (9.19)	85.4 (8.97)	<0.001
PC	86.5 (6.22)	97.0 (6.13)	110 (8.69)	0.000
Colesterol HDL	55.3 (14.1)	50.8 (11.8)	47.4 (10.7)	<0.001
Colesterol total	186 (33.9)	189 (32.8)	188 (32.3)	0.079
Glucosa	83.5 (15.7)	87.2 (16.7)	93.9 (23.0)	<0.001
Índice RECV	1.72 (1.71)	2.15 (1.76)	2.34 (1.67)	<0.001

TAS: Tensión Arterial Sistólica, TAD: Tensión Arterial Diastólica, PC: Perímetro de Cintura, RECV: Riesgo Enfermedad Cardiovascular.

Tabla 2.10: Análisis descriptivo según grupos de IMC en el último corte.

2.2.4. Imputaciones

Para el análisis de clústeres es necesario que no haya datos ausentes, por ello se realizó una imputación de datos. Para estas imputaciones se utilizó el paquete de R longitudinalData y su función imputation. Esta función permite hacer imputaciones a variables registradas en distintos instantes de tiempo por separado. De las distintas opciones que ofrece el paquete para la imputación de datos se utilizó la opción "linearInterpol.bisector", la cual actúa de forma distinta si los datos que faltan son intermitentes (falta un registro pero el anterior y el posterior sí que están disponibles) o son monótonos (no hay datos registrados hasta un momento dado o no hay datos a partir de un momento).

Para los datos faltantes intermitentes, esta opción une por una recta el dato previo y el posterior al dato que falta y de aquí toma el valor para imputar. En cuanto a los datos faltantes monótonos, toma la bisectriz de las dos rectas resultantes de unir, por un lado el primer y último dato disponibles, y por otro lado la resultante de unir o los dos primeros registros o los dos últimos disponibles, dependiendo de si los datos faltantes se agrupan al principio o al final de las mediciones respectivamente.

Este proceso es posible siempre que haya al menos un registro de todos los instantes, por ello hubo datos de algunas variables que no fue posible imputar. Así pues, dado que para la variable perímetro de cintura hubo 5 individuos que no tenían registro en ninguno de los instantes, 8 que no tenían ninguno para la variable IMC y 9 para el índice de RECV, estas variables tuvieron finalmente 5, 8 y 9 datos faltantes respectivamente.

El análisis descriptivo de las variables para los tres cortes (momentos del estudio) con imputaciones (Tabla 2.11) y las de los datos originales (Tablas 2.3, 2.6 y 2.9) fueron muy similares.

Variables	Primer corte	Segundo corte	Tercer corte	NA
PC	96.85 (9.7)	97.38 (10.06)	98.1 (10.87)	5
IMC	27.62 (3.56)	27.81 (3.69)	27.95 (3.88)	8
Colesterol HDL	52.44 (10.97)	53.83 (11.29)	51.5 (12.89)	0
Glucosa	97.71 (18.77)	98.44 (19.58)	89.9 (21.52)	0
Índice RECV	1.57 (1.43)	2.05 (1.72)	2.43 (2.2)	9

PC: Perímetro de Cintura, IMC: Índice de Masa Corporal, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 2.11: Análisis descriptivo de variables con imputaciones.

2.3. Análisis descriptivo subcohorte

El análisis descriptivo de los miembros de la subcohorte se muestra en la Tabla 2.12. Este fue similar al análisis de toda la cohorte y cabe destacar que la media de edad fue algo superior en los miembros

de la subcohorte, aunque su desviación típica fue la mitad que la de la cohorte en los tres instantes. Sin embargo, la media del índice de RECV fue menor en los tres momentos en la población de la subcohorte que en la de la cohorte.

	Primer momento	Segundo momento	Tercer momento
Edad	48.9 (4.05)	52.3 (4.07)	55.3 (4.06)
IMC	27.8 (3.34)	27.9 (3.51)	28.1 (3.63)
PC	97.3 (9.10)	97.9 (9.45)	98.7 (9.98)
Colesterol HDL	52.6 (10.9)	54.2 (11.3)	51.0 (12.3)
Glucosa	98.1 (18.0)	96.5 (18.8)	88.7 (18.1)
Índice RECV	1.43 (1.14)	1.87 (1.29)	2.22 (1.56)

PC: Perímetro de Cintura, IMC: Índice de Masa Corporal, RECV: Riesgo de Enfermedad Cardiovascular.

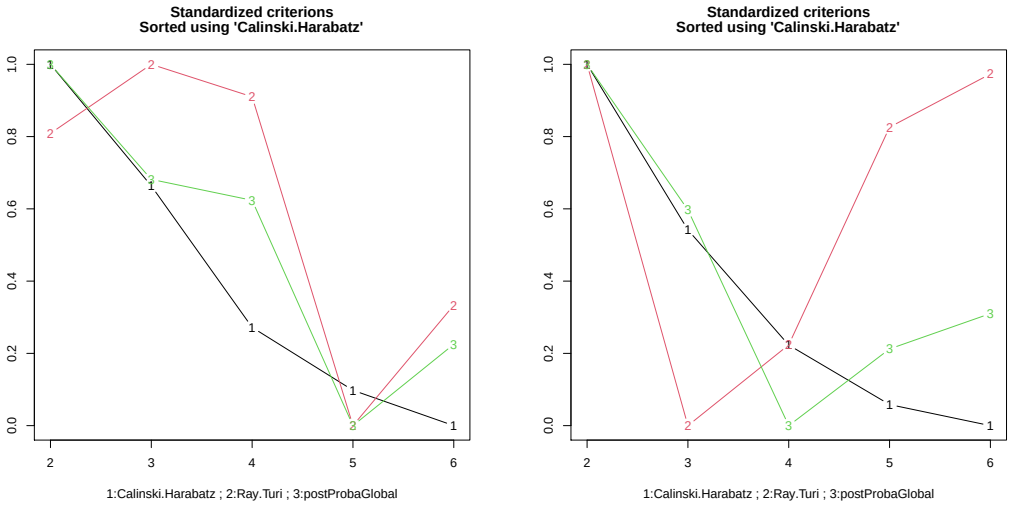
Tabla 2.12: Análisis descriptivo de variables en la subcohorte.

Capítulo 3

Resultados

El objetivo de este estudio fue agrupar a la población en función de la evolución de sus FRCV y el índice de riesgo de sufrir ECV en tres momentos de tiempo distintos. Para ello se realizó un análisis de clústeres para datos longitudinales en el que se estudió la evolución conjunta de las variables: edad, perímetro de cintura, IMC, niveles de glucosa y colesterol HDL en sangre y el índice de riesgo de sufrir ECV en 10 años. El análisis se realizó para los miembros de la cohorte y para los miembros de la subcohorte.

En primer lugar se determinó el mejor número de clústeres a partir de los índices de validez. En las Figuras 3.1a) y 3.1b) se muestran los índices que se tuvieron en cuenta para seleccionar el número de grupos óptimo para los estudios de la cohorte y de la subcohorte respectivamente. En estos gráficos dichos índices aparecen reescalados entre 0 y 1, lo que es conveniente para poder compararlos unos con otros.



(a) Para toda la cohorte

(b) Para la subcohorte

Figura 3.1: Índices de calidad según número de clústeres.

De acuerdo con estos resultados se realizó el análisis de clústeres para dos y tres grupos tanto para el análisis de los miembros de toda la cohorte como para el de los miembros de la subcohorte.

3.1. Análisis de toda la cohorte

3.1.1. División en dos grupos

En la Figura 3.2 se representan las trayectorias de los centros de cada grupo para cada variable por separado y en la Tabla 3.1 el análisis descriptivo de la cohorte según estos grupos.

En el primer clúster se agruparon aquellos individuos más jóvenes, con menores valores de perímetro de cintura, IMC, niveles de glucosa en sangre e índice de RECV y mayor valor de colesterol HDL en sangre. Atendiendo a la evolución a lo largo de los tres instantes de cada variable, todas siguen el mismo patrón dentro de cada grupo: las medias del PC e IMC experimentaron un leve aumento progresivo en cada instante, siendo este aumento un poco superior en el cluster 2; en los niveles de glucosa en sangre se observa una leve y progresiva disminución entre el primer instante y el tercero, siendo este descenso mayor en el primer grupo que en el segundo. Respecto a los niveles de colesterol HDL, en ambos grupos aumentan entre el primer y segundo corte y disminuyen entre el segundo y el tercero. Finalmente, en ambos grupos se observa un aumento del índice de RECV, siendo mayor en el caso del segundo clúster.

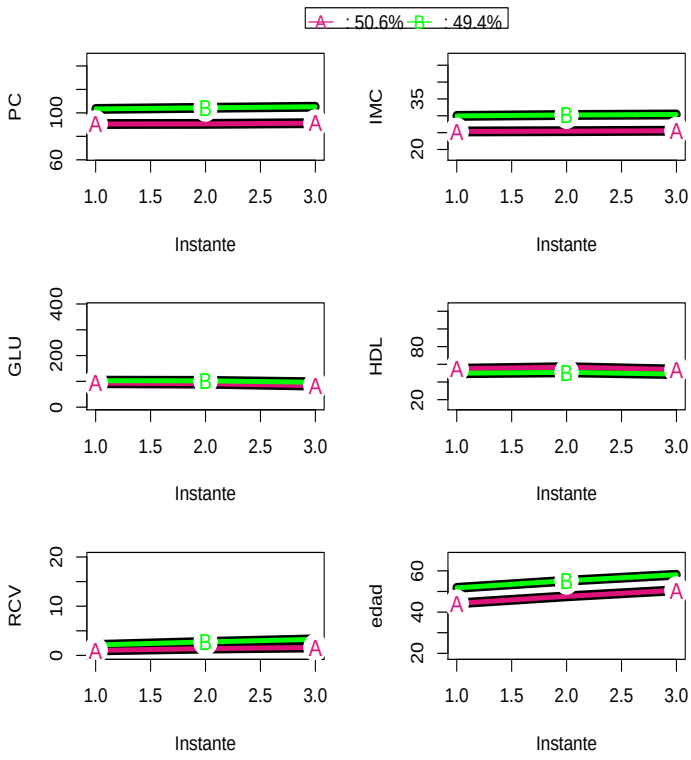


Figura 3.2: Centros de clústeres para 2 grupos en la cohorte.

		1 N=2099	2 N=2048	p.overall
Edad	Inicio	44.2 (9.58)	51.7 (4.59)	<0.001
	2014	47.6 (9.58)	55.2 (4.63)	<0.001
	2017	50.6 (9.55)	58.1 (4.58)	<0.001
PC	Inicio	90.5 (6.80)	103 (7.72)	0.000
	2014	90.6 (6.61)	104 (8.06)	0.000
	2017	91.2 (7.31)	105 (8.97)	0.000
IMC	Inicio	25.3 (2.30)	30.0 (3.04)	0.000
	2014	25.4 (2.25)	30.2 (3.22)	0.000
	2017	25.6 (2.41)	30.4 (3.49)	0.000
Glucemia	Inicio	92.9 (11.7)	103 (22.9)	<0.001
	2014	91.0 (11.5)	102 (24.1)	<0.001
	2017	83.4 (11.7)	96.5 (26.7)	<0.001
HDL	Inicio	55.0 (11.3)	49.9 (9.94)	<0.001
	2014	56.8 (11.8)	50.9 (9.96)	<0.001
	2017	54.1 (13.4)	48.9 (11.8)	<0.001
Índice RECV	Inicio	1.02 (1.03)	2.12 (1.55)	<0.001
	2014	1.38 (1.27)	2.74 (1.85)	<0.001
	2017	1.62 (1.49)	3.27 (2.49)	<0.001

PC: Perímetro de Cintura, IMC: Índice de Masa Muscular, HDL: Colesterol HDL, RECV: Riesgo de Enfermedad Cardiovascular

Tabla 3.1: Análisis descriptivo para 2 grupos en la cohorte.

3.1.2. División en tres grupos

En cuanto a la distribución en tres clústeres (Figura 3.3 y Tabla 3.2), en el primer clúster se agruparon los sujetos mayores y en el tercero los jóvenes, aunque la diferencia de edad entre el primero y el segundo fue de menos de un año y la de éstos con el tercero fue de 17 años. Así pues, el primer clúster recoge a los individuos de mayor edad y que tienen menor perímetro de cintura e IMC, menores niveles de glucosa en sangre y mayores niveles de colesterol HDL e índice de RECV en comparación a los individuos del segundo clúster, que tienen una edad parecida. En cuanto al tercer clúster, está formado

por los individuos más jóvenes y que tienen los valores más bajos de IMC, glucosa en sangre e índice de RECV, mientras que tienen los valores más altos de colesterol HDL. La diferencia en el índice de RECV es mucho mayor al comparar el primer o el segundo grupo con el tercero que entre el primer y segundo grupos.

Atendiendo a la evolución de cada variable dentro de los grupos, en este caso no se puede decir que sea la misma. En cuanto al PC e IMC, en el primer grupo se mantuvo prácticamente estable y en los otros dos grupos experimentó un leve incremento progresivo. Los niveles de glucosa disminuyeron progresivamente en los grupos 1 y 3, mientras que en el segundo grupo permanecieron estables entre los dos primeros instantes del estudio y disminuyeron en el tercero. En cuanto al colesterol HDL, en los tres grupos aumentó entre el primer y segundo instante y disminuyó en el tercero. El índice de RECV en los tres grupos sufrió un incremento progresivo con el paso del tiempo, siendo mayores estos incrementos en el primer y segundo grupo que en el tercero.

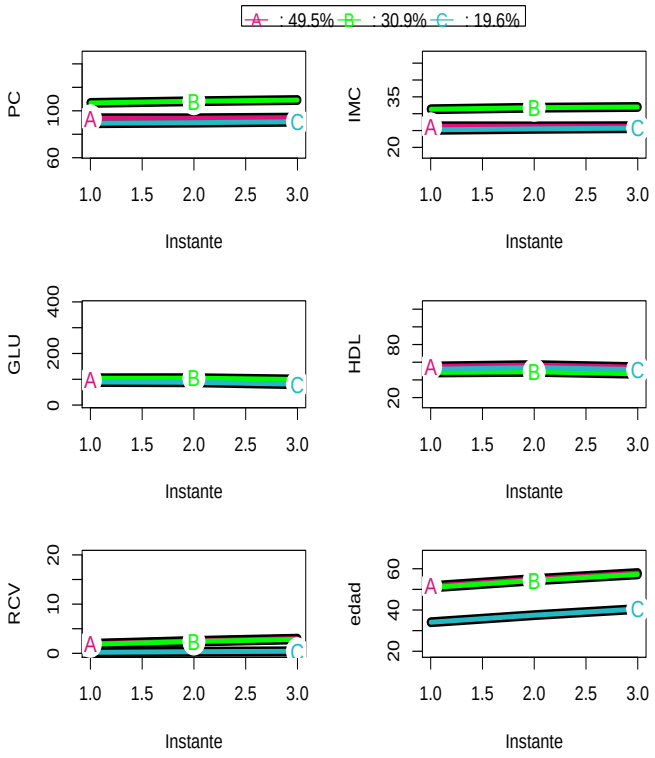


Figura 3.3: Centros de clústeres para 3 grupos en la cohorte.

		1 N=2051	2 N=1283	3 N=813	p.overall
Edad	Inicio	51.6 (3.57)	50.9 (4.75)	34.1 (7.43)	0.000
	2014	55.0 (3.62)	54.3 (4.77)	37.5 (7.43)	0.000
	2017	57.9 (3.58)	57.3 (4.71)	40.5 (7.46)	0.000
PC	Inicio	93.7 (6.18)	107 (7.21)	89.4 (8.17)	0.000
	2014	93.7 (6.10)	108 (7.29)	89.9 (7.95)	0.000
	2017	94.2 (6.91)	109 (8.35)	90.7 (8.52)	0.000
IMC	Inicio	26.3 (2.13)	31.3 (2.84)	25.2 (2.82)	0.000
	2014	26.2 (2.10)	31.8 (2.94)	25.5 (2.74)	0.000
	2017	26.3 (2.25)	32.0 (3.23)	25.7 (2.91)	0.000
Glucemia	Inicio	97.1 (15.7)	104 (24.7)	89.3 (8.92)	<0.001
	2014	94.7 (14.0)	104 (27.6)	88.3 (9.63)	<0.001
	2017	87.7 (15.4)	99.2 (30.1)	80.7 (10.4)	<0.001
HDL	Inicio	55.2 (11.2)	48.4 (9.30)	52.0 (10.8)	<0.001
	2014	56.8 (11.4)	49.1 (9.39)	54.0 (11.3)	<0.001
	2017	54.4 (13.3)	46.9 (10.9)	51.6 (12.8)	<0.001
Índice RECV	Inicio	1.93 (1.43)	1.85 (1.34)	0.20 (0.23)	<0.001
	2014	2.52 (1.71)	2.40 (1.58)	0.33 (0.34)	<0.001
	2017	2.98 (2.23)	2.84 (2.10)	0.42 (0.43)	<0.001

PC: Perímetro de Cintura, IMC: Índice de Masa Muscular,
HDL: Colesterol HDL, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 3.2: Análisis descriptivo para 3 grupos en la cohorte.

3.2. Análisis de la subcohorte

3.2.1. División en dos grupos

Las trayectorias de los centros de cada grupo y el análisis descriptivo de los miembros de la subcohorte por grupos se muestran en la Figura 3.4 y en la Tabla 3.3. Ambos grupos tuvieron las mismas características que los grupos de la cohorte. En este caso la media de edad del primer grupo, que corresponde a los individuos más jóvenes, es casi 4 años mayor que en el grupo correspondiente para el análisis de la cohorte y la del segundo grupo, que correspondería a los individuos mayores, fue casi un año menor. Las medias de PC e IMC fueron algo mayores en ambos grupos de la subcohorte, los niveles de glucemia y colesterol HDL fueron parecidos en los grupos de ambos estudios. Finalmente, el índice

de RECV fue mayor en el primer grupo de la subcohorte que en el de la cohorte y menor en el segundo grupo de la subcohorte. La evolución de cada variable en ambos grupos fue la misma que en el estudio de la cohorte.

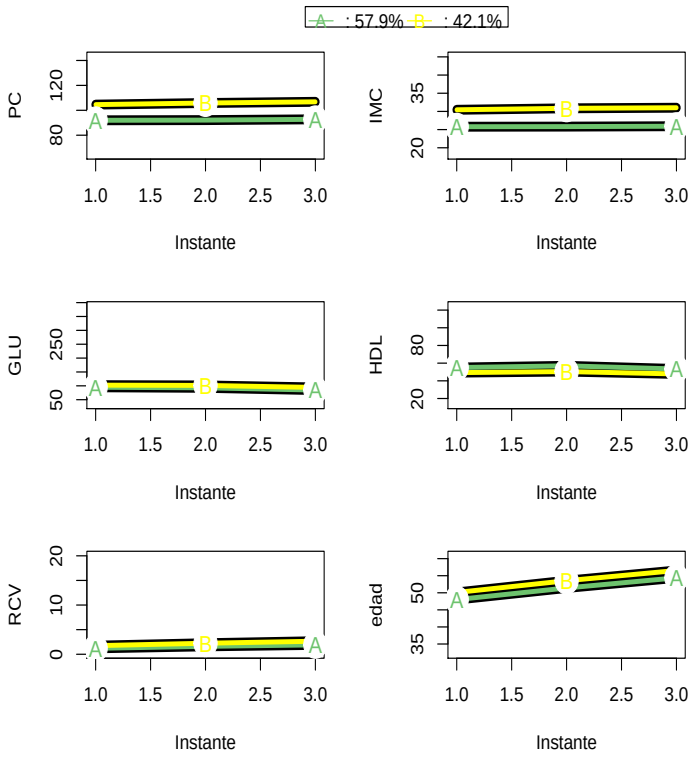


Figura 3.4: Centros de clústeres para 2 grupos en la subcohorte.

		1 N=1390	2 N=1011	p.overall
Edad	Inicio	48.0 (4.30)	50.1 (3.33)	<0.001
	2014	51.4 (4.31)	53.5 (3.34)	<0.001
	2017	54.4 (4.29)	56.5 (3.34)	<0.001
PC	Inicio	92.0 (6.14)	105 (7.29)	<0.001
	2014	92.1 (5.98)	106 (7.36)	0.000
	2017	92.8 (6.54)	107 (7.97)	<0.001
IMC	Inicio	25.8 (2.09)	30.4 (2.81)	<0.001
	2014	25.8 (2.05)	30.9 (2.96)	<0.001
	2017	25.9 (2.16)	31.0 (3.14)	<0.001
Glucemia	Inicio	94.8 (12.5)	103 (22.8)	<0.001
	2014	92.8 (12.0)	102 (24.5)	<0.001
	2017	84.7 (11.9)	94.2 (23.1)	<0.001
HDL	Inicio	55.0 (11.1)	49.3 (9.67)	<0.001
	2014	57.0 (11.6)	50.5 (9.60)	<0.001
	2017	53.4 (12.7)	47.8 (10.8)	<0.001
Índice ERCV	Inicio	1.20 (1.00)	1.74 (1.25)	<0.001
	2014	1.61 (1.08)	2.24 (1.46)	<0.001
	2017	1.91 (1.32)	2.65 (1.76)	<0.001

PC: Perímetro de cintura, IMC: Índice de Masa Muscular, HDL: Colesterol HDL, RECV: Riesgo de Enfermedad Cardiovascular

Tabla 3.3: Análisis descriptivo para dos grupos en la subcohorte.

3.2.2. División en tres grupos

La división en tres grupos de la subcohorte fue parecida a la de la cohorte (Figura 3.5 y Tabla 3.4). La media de edad en el primer grupo fue casi la misma en ambos estudios, en el segundo grupo de la subcohorte la edad fue un año inferior que en el de la cohorte y en el tercer grupo fue 10 años mayor en la subcohorte. En cuanto a las medias de las demás variables por grupos, en los clúster 1 y 2 fueron parecidas en los tres instantes en ambos estudios, salvo las medias del índice de RECV que fueron algo inferiores en la subcohorte. Respecto a las medias en el tercer grupo, las de la subcohorte fueron todas inferiores a las de la cohorte. Finalmente, la evolución de las medias de las variables dentro de cada grupo a lo largo del tiempo fue la misma en el estudio de la cohorte que en el de la subcohorte.

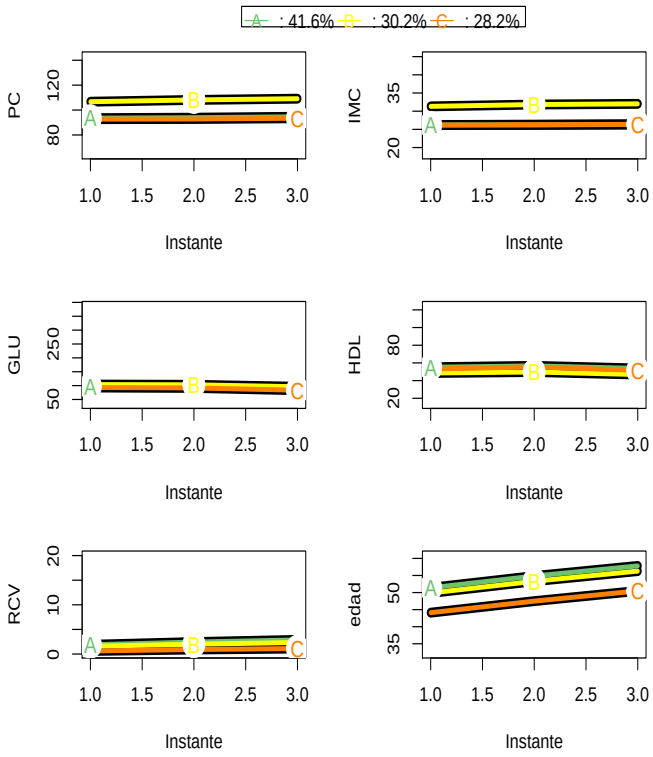


Figura 3.5: Centros de clústeres para 3 grupos en la subcohorte.

		1 N=1005	2 N=723	3 N=673	p.overall
Edad	Inicio	51.4 (1.71)	49.9 (3.26)	44.1 (3.07)	0.000
	2014	54.8 (1.73)	53.3 (3.27)	47.5 (3.04)	0.000
	2017	57.8 (1.72)	56.3 (3.26)	50.5 (3.04)	0.000
PC	Inicio	93.5 (6.25)	107 (7.03)	92.7 (6.79)	0.000
	2014	93.9 (6.13)	108 (7.01)	92.8 (6.85)	0.000
	2017	94.6 (6.73)	109 (7.83)	93.4 (7.20)	0.000
IMC	Inicio	26.2 (2.12)	31.3 (2.70)	26.2 (2.34)	0.000
	2014	26.2 (2.08)	31.8 (2.80)	26.3 (2.36)	0.000
	2017	26.3 (2.21)	32.0 (2.99)	26.4 (2.44)	0.000
Glucemia	Inicio	97.3 (14.5)	104 (24.6)	93.1 (11.4)	<0.001
	2014	95.1 (14.8)	103 (26.5)	91.6 (10.4)	<0.001
	2017	87.1 (13.5)	96.0 (25.5)	83.4 (10.5)	<0.001
HDL	Inicio	55.1 (11.1)	48.6 (9.76)	53.1 (10.6)	<0.001
	2014	56.6 (11.5)	49.7 (9.51)	55.5 (11.4)	<0.001
	2017	53.7 (12.9)	46.8 (10.4)	51.6 (12.0)	<0.001
Índice RECV	Inicio	1.88 (1.24)	1.57 (1.09)	0.62 (0.35)	<0.001
	2014	2.43 (1.44)	1.99 (1.09)	0.92 (0.45)	<0.001
	2017	2.87 (1.73)	2.35 (1.38)	1.11 (0.56)	<0.001

PC: Perímetro de Cintura, IMC: Índice de Masa Muscular,
HDL: Colesterol HDL, RECV: Riesgo de Enfermedad Cardiovascular.

Tabla 3.4: Análisis descriptivo para 3 grupos en la subcohorte.

Bibliografía

[1] S. KAPTOGE, L. PENNELLS, D. DE BACQUER, M. T. COONEY, M. KAVOUSI, G. STEVENS, ET AL, World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions, *Lancet Glob. Health* **7**(10) (2019), e1332-e1345.

[2] R.GABRIEL, M.ALONSO, A. SEGURA, M. J. TORMO, L. M. ARTIGAO, J. R. BANEGAS, ET AL, Prevalencia, distribución y variabilidad geográfica de los principales factores de riesgo cardiovascular en España. Análisis agrupado de datos individuales de estudios epidemiológicos poblacionales: estudio ERICE, *Rev. Esp. Cadiol* **61** (10) (2008), 1030-1040.

[3] J. A. CASASNOVAS, V. ALCAIDE, F.CIVEIRA,,E. GUALLAR, B. IBAÑEZ,J. J BORREGUERO, ET AL, Aragon workers' health study–design and cohort description, *BMC cardiovascular disorders* **12** (1) (2012), 45.

[4] T. HASTIE, R. TIBSHIRANI Y J.FRIEDMAN, *The elements os statistical learning: Data mining, inference and prediction*, Springer, Nueva York, 2001.

[5] D. PEÑA, *Análisis de datos multivariantes*, McGraw-Hill, Madrid, 2002.

[6] C. GENOLINI, J. B. PINGAULT, T. DRISS, S. CÔTÉ, R. E. TREMBLAY, F. VITARO, ET AL, KmL3D: a non-parametric algorithm for clustering joint trajectories, *Comput. Meth. Prog. Bio* **109** (1) (2013), 104-111.

[7] C. GENOLINI, X. ALACOQUE, M. SENTENAC, Y C. ARNAUD, kml and kml3d: R packages to cluster longitudinal data, *J. Stat. Softw* **65** (4) (2015), 1-34.

[8] C. GENOLINI, B. FALISSARD Y J.B. PINGAULT, *K-Means for Joint Longitudinal Data*, `https://cran.r-project.org/web/packages/kml3d/kml3d.pdf`, disponible en `http:www.r-project.org`.

[9] C. GENOLINI, B. FALISSARD, D. FANG Y L. TIERNEY, *Longitudinal Data*, `https://cran.r-project.org/web/packages/longitudinalData/longitudinalData.pdf`, disponible en `http:www.r-project.org`.

[10] A. JAIN Y R.DUBES, *Algorithms for Clustering Data*, Prentice-Hall, New Jersey, 1988.

[11] R. TIBSHIRANI, G. WALTHER Y T. HASTIE, Estimating the number of clusters in data set via the gap statistic, *J. R. Statistic. Soc. B* **63** (2) (2001), 411-423.

[12] P. ROUSSEEUW, Sihouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math* **20** (1987), 53-65.

[13] M. J. MEDRANO, E. CERRATO, R. BOIX Y M. DELGADO-RODRÍGUEZ, Factores de riesgo cardiovascular en la población española: metaanálisis de estudios transversales *Med Clin* **124** (16) (2005), 606-612.

[14] R.M. CONROYA, K.PYÖRÄLÄ Y A.P.FITZGERALD, Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project, *Eur Heart J* **24** (2003), 987-1003.

Anexo A

Paquete KmL3D

A.1. Preparación de datos

Para implementar el algoritmo del paquete KmL3D es necesario preparar los datos que se tienen disponibles para que éste funcione [7]. Es importante tener en cuenta que el algoritmo trabaja con datos de tipo "ClusterLongData3d", por ello la preparación de los datos consiste en transformar los datos disponibles en un objeto de este tipo. Esto se puede hacer con la función cld3d(), disponible en el mismo paquete y que transforma matrices o data frames en objetos "ClusterLongData3d".

Dentro de la función hay diferentes aspectos importantes que hay que definir con los argumentos que se especifican a continuación:

- Para especificar el objeto de clase matriz o data.frame que contiene los datos longitudinales, donde las filas son las trayectorias de los individuos y las columnas corresponden a las mediciones que se han hecho en cada instante de las variables, utilizaremos el argumento traj().

- Para especificar los tiempos en los que se han hecho las mediciones se hará con el argumento time() y se introducirán como un vector de números.

- Es necesario también especificar qué columnas contienen cada variable y los instantes en los que se midió. Para ello se utiliza el argumento timeInData() en el que se especificará con una lista las variables que tenemos y con vectores numéricos las columnas en las que se registra cada medición. Así pues, si queremos especificar que la variable perímetro abdominal se encuentra registrada en las columnas 2, 3 y 4 y la variable IMC en las columnas 5, 6 y 7, este argumento vendra dado por: timeInData = list (PC= c(2,3,4), IMC= c(5,6,7)).

- Finalmente hay que especificar el número máximo de datos faltantes en una trayectoria para que el individuo sea incluido en el estudio, esto se registra con el argumento maxNa().

  Dado que en este estudio las imputaciones se realizaron previamente, los individuos con datos faltantes en una variable estarán ausentes en todos los instantes.

A.2. Funcionamiento kml3d

Una vez están los datos preparados se puede llamar a la función kml3d() con la que se obtendrán las particiones. Esta función ejecuta el algoritmo de k-medias varias veces con diferentes condiciones de inicialización y para diferentes números de clústeres. Como condiciones de inicialización, además de las opciones descritas en el Anexo C, se pueden usar dos valores específicos: "nearlyAll", que utiliza primero como condición de inicialización el método "kmeans-" y continúa alternando los métodos "kmeans- -"y "randomK "; y "all", que comienza por "maxDist" y "kmeans-" y sigue alternando "kmeans- -" y "randomK". Por defecto, kml3d() ejecuta el algoritmo de k-medias para $k \in \{2,3,4,5,6\}$

número de clústeres 20 veces, utilizando la opción "nearlyAll" de inicialización y la distancia Euclídea para el cálculo de las distancias (esta se puede cambiar a otras distancias con la opción distance).

Los argumentos que admite la función kml3d() y que servirán para modificar algunas de las opciones que trae por defecto son:

- nbClusters() nos permite cambiar el número de clústeres para los que se harán las particiones. Como hemos dicho por defecto será 2 : 6.
- nbRedrawing() permite especificar el número de veces que se ejecuta el algoritmo de k-medias (por defecto es 20).
- toPlot() está relacionado con el gráfico que se muestra mientras la función kml3d() se está ejecutando y admite 4 opciones. Hay dos opciones de gráficos: el primero que se indica como "criterio" muestra los criterios de validez de todas las particiones ya encontradas y el segundo, indicado por "traj", representa la evolución del proceso de k-medias. Las otras dos opciones de este argumento son "none" si no queremos que muestre gráficos y "both" si queremos los dos.
- parAlgo() argumento que va asociado a un objeto de tipo "ParKml3d" y sirve para cambiar opciones más avanzadas como las condiciones de inicialización y el tipo de distancia a utilizar.

Tras ejecutar la función kml3d(), con la opción choice() se pueden visualizar los resultados obtenidos. A la izquierda de la imagen aparece un gráfico donde cada partición encontrada se muestra con el número de clúster y la altura vendrá dada por el valor del criterio de validez de esa partición. El programa nos permite explorar las distintas particiones encontradas y los resultados para diferentes criterios de validez utilizando nuestro teclado, así como seleccionar la partición que más nos interesa de todas las calculadas. A la derecha del gráfico nos mostrará los resultados de la partición seleccionada.

Anexo B

Formulario criterios de calidad

En este anexo se van a presentar las fórmulas utilizadas para los criterios explicados en 1.4.4 y que se pueden encontrar en [7].

- Criterios no paramétricos

  - Criterio Calinski & Harabasz (1974) se calcula con la fórmula:

    $$C(K) = \frac{tr(B)}{tr(W)} \frac{n-k}{k-1}.$$

    siendo B la matriz de covarianzas entre clústeres y W la matriz de covarianzas dentro de los clústeres. Es decir, valores altos de la traza de B indican clústeres bien separados y valores altos de la traza de W indican clústeres compactos.
    Este criterio tiene algunas variantes que utilizando esta misma notación vienen dadas por:

    - Variante de Kryszczuk (2010):

      $$C_K(K) = \frac{tr(B)}{tr(W)} \frac{n-1}{n-k}.$$

    - Variante de Genolini:

      $$C_G(K) = \frac{tr(B)}{tr(W)} \frac{n-k}{\sqrt{k-1}}.$$

  - Criterio de Ray & Turi (1999):

    $$R(K) = \frac{\sum_{i=1}^{K} \sum_{x \in C_i} (dist(x, \bar{y}_i))}{n \min_{C_i, C_j} (dist(\bar{y}_i, \bar{y}_j)^2)}$$

    siendo $\bar{y}_i$, $\bar{y}_j$ los centros de los clústeres C_i y C_j, con $i \neq j$.
  - Criterio Davies & Bouldin (1979):

    $$D(K) = media(P(C_i, C_j))$$

    siendo

    $$P(C_i, C_j) = \frac{DistInterna(C_i) + DistInterna(C_j)}{DistExterna(C_i, C_j)}$$

    donde C_i, C_j son los clústeres $i \neq j$.

- Criterios paramétricos: recordar que en este caso n es el número de individuos y $N = n \cdot t$, además p será el número de variables que estemos tomando.

  - $BIC = 2 \cdot \log(L) - p \cdot \log(n)$

- $BIC = 2 \cdot \log(L) - p \cdot \log(N)$
- $AIC = 2 \cdot \log(L) - 2 \cdot p$
- $AICc = AIC + \frac{2p(p+1)}{n-p-1}$
- $AICc = AIC + \frac{2p(p+1)}{N-p-1}$

L representa el máximo valor de la función de verosimilitud para el modelo estadístico con el que se está trabajando. Como se está suponiendo que las variables siguen una distribución normal y son homocedásticas, L es igual a la varianza del error, que se define de la siguiente como $L = \sum_i e_i^2/n$.

Anexo C

Inicialización k-medias

Las opciones que ofrece el paquete KmL3D para la inicialización del algoritmo de k-medias son las siguientes [7]:

1. **randomK:** Se seleccionan k individuos al azar como centros de los grupos.
2. **randomAll:** Todos los individuos se asignan a un clúster y se calcula el centro de cada clúster como su media.
3. **maxDist:** Es un método incremental. En primer lugar se calcula la matriz de distancias a todos los puntos y posteriormente se seleccionan los dos puntos más distantes como centro. Finalmente entra en un bucle que finaliza cuando se han elegido k centros. El bucle es el siguiente:
   I. Se considera la distancia de cada punto al centro más cercano.
   II. Se elige como un nuevo centro aquel punto cuya distancia es la mayor.

   Este método es más efectivo que los anteriores pero tiene un alto coste computacional $o(n^2)$.
4. **kmeans+:** Mejora el método anterior seleccionando los dos centros iniciales aleatoriamente y evitando así el cálculo de la matriz de distancias entre todos los individuos. Una vez elegidos los dos centros iniciales funciona de la misma manera y el coste computacional se reduce a $o(kn)$. La efectividad de este método es menor ya que la elección de los primeros centros puede no ser buena.
5. **kmeans-:** Trata de mejorar el método anterior solucionando el que los dos primeros centros pueden estar cerca. Para ello selecciona aleatoriamente un centro y calcula la distancia a este del resto de puntos, ahora selecciona como el segundo centro aquel que está más lejos del primero y elimina el primero como centro. Así se asegura que este segundo centro es adecuado ya que es el más distante de por lo menos a un punto. Con este nuevo centro empieza el bucle descrito en 3. Este método tiene las ventajas de los métodos maxDist y kmeans+ pero presenta el inconveniente de que, excepto para la primera elección, para las demás es determinista.
6. **kmeans- -:** Este método trata de mejorar el anterior eligiendo aleatoriamente los centros que se van añadiendo a la lista. Para que la probabilidad de que los centros sean distantes unos de otros, la probabilidad de que un individuo sea seleccionado como centro es proporcional al cuadrado de la distancia entre el individuo y los centros previamente seleccionados.

   Este método tiene todas las ventajas de los anteriores: los centros son distantes unos de otros, el primer centro no está mal seleccionado, el coste computacional no es excesivo, $o(kn)$, y el método no es determinista.
7. **kmeans++:** Es la versión no determinista del método kmeans+. Una vez elegido el primer centro, los demás son elegidos aleatoriamente con diferentes pesos de probabilidad en función de la distancia al cuadrado entre el individuo y los centros ya elegidos.