Universitat Politècnica de Catalunya

Ciència i Enginyeria de Dades

Análisis de Datos Multivariante

Male fat body

Gerard Comas Quiles

June 2022

Índice

1. Introducción

Este proyecto se basa en aplicar los métodos que hemos ido aprendiendo en clase para analizar un conjunto de datos. Estas técnicas nos permiten resaltar la información útil para elaborar unas conclusiones sobre la base de datos inicial, así saber, por ejemplo, cuantos clusters diferentes hay en nuestra base de datos y que caracteriza a cada uno de ellos. En general, trabajaremos siempre con muchas variables y la representación gráfica de los datos no se podrá realizar, por lo que nos interesará sobre todo centrarnos en aquellos métodos que nos permitan reducir las dimensiones de nuestros datos.

Aplicar de esta forma nos será útil para un innumerable de casos en el futuro, pues para cualquier observación siempre contamos con muchísimas variables y nos puede costar entender las relaciones entre ellas. Por ejemplo, en nuestro caso estudiaremos el porcentaje de grasa en los hombres. Lo que podemos esperar del proyecto es entender que características de nuestro cuerpo están relacionadas con el porcentaje de grasa y poder llegar a hacer predicciones.

El estudio se basará en estudiar el porcentaje de grasa, pues realmente es una medida difícil de medir. Lo más preciso es actuar una densitometría ósea, también conocida como prueba DEXA, es un tipo de radiografía que permite medir la densidad de las partes de nuestro cuerpo, de esta manera se puede aproximar el porcentaje de grasa. Evidentemente, este proceso es bastante costoso y solo se puede proceder en clínicas especializadas. Una alternativa a esto serían las básculas de bioimpedancía, generan una pequeña corriente eléctrica imperceptible para que circule por nuestro cuerpo, luego el aparato vuelve a recuperar la señal que ha enviado. Se estudia la diferencia de intensidad entre la señal enviada y recibida para calcular la bioimpedancia hecha por nuestro cuerpo y aproximar la cantidad de grasa y músculo. Esta técnica, aún ser algo más simple, también suele limitarse al uso de profesionales, pues una buena báscula de bioimpedáncia es bastante cara y no es lo mejor para el uso cuotidiano. Por lo que si queremos medir nuestra grasa corporal desde casa usaremos un plícometro, para usarlo lo que tenemos que hacer es medir algunos pliegues de nuestro cuerpo y después usar alguna de las muchas fórmulas que hay para determinar el porcentaje de grasa. Nuestro trabajo se parecerá en cierto sentido a esta última técnica, pues unos de nuestros objetivos será aproximar el porcentaje de grasa a partir de medidas fáciles de tomar, como por ejemplo el peso, la altura o las medidas de nuestro abdomen.

Debemos comentar que este estudio se basará solo en los hombres, pues la diferencia entre el porcentaje de grasa de hombres y mujeres es significativa. Esto tiene una explicación biológica, teóricamente las mujeres acumulan más grasas para prepararse para la maternidad, un nivel de grasa saludable permitirá un ciclo menstrual regular y saludable y además una correcta nutrición del bebé en caso de embarazo. Lo idóneo en este caso sería estudiar tanto hombres como mujeres, pero como sería hacer dos veces el mismo trabajo, pero con resultados distintos y nos disponemos de estos datos, nos limitaremos al estudio representativo de la mitad de la población humana.

2. Descripción de la base de datos

La base de datos consiste en 252 muestras con 15 variables numéricas, creada por Roger Johnson, profesor del departamento de matemáticas e informática de Carleton College y se puede descargar aquí: http://www.statistics4u.com/fundstat_eng/data_bodyfat.html

Ahora haremos una explicación sobre que indica cada variable para una mayor comprensión de nuestra base de datos:

I) ***Density*:** equivale a la densidad del cuerpo en gramos por centímetro cúbico, comprende valores ligeramente superiores al uno. Su rango de valores en esta base de datos es de 0.9950 a 1.1089.

II) %***Fat*:** indica el porcentaje de grasa corporal, al tratar de un porcentaje tomará valores entre el 0 y el 100 %, en nuestro caso del 0 % hasta el 47.5 %.

III) ***Age*:** muestra la edad en años, por lo que serán valores positivos del orden de 10, esta muestra representa desde los 22 años hasta los 81.

IV) ***Weight*:** denota el peso medido en libras, medida habitual para el peso utilizada en Estados Unidos. Los valores se moverán entre las 118.50 lb (53.75 Kg) hasta 363.15 lb (164.722 Kg)

V) ***Height*:** refleja la altura medida en pulgadas, o eso parece, como ya hemos visto el sistema de medidas americaso difiere del estándar internacional y podría equivaler a otra medida. Pero las pulgadas encajarian bastante con el conjunto de datos y su conversión a metros. Los datos comprenden valores entre 29.50 in (0.75 m) hasta 77.75 in (1.97 m).

VI) **Medidas del cuerpo:** tenemos 10 medidas de diferentes partes del cuerpo, estas son: cuello, pecho, abdomen, cadera, muslo, rodilla, tobillo, bíceps, antebrazo y muñeca. Las unidades que se emplean para las medidas no se especifican en la base de datos, podemos suponer que son cm, pues encaja bastante con lo que podríamos esperar. Pero tratándose de datos recogidos en Estados Unidos realmente las magnitudes que se han empleado son una incógnita.

Pero para nuestro estudio no nos hará falta tener en cuenta magnitudes, pues estandarizaremos los datos. No queremos que nuestro análisis dependa de las unidades que utilicemos y además no todas las variables se mueven en los mismos intervalos. Un ejemplo bastante claro es ***Density*** y ***Weight***, aunque en cierta manera estén relacionadas, pues la densidad es el peso entre el volumen, el peso es unas doscientas veces mayor. Esto puede provocar que los métodos intenten minimizar los errores sobre la variable peso y la variable densidad pase prácticamente desapercibida.

3. Descripción de los métodos empleados para analizar los datos

Una vez descrito el conjunto de datos que trataremos, vamos a explicar en que consisten los métodos que usaremos en este proyecto y que deben cumplir los datos para obtener un buen análisis.

3.1. Análisis de Componentes Principales: PCA

El primer método que estudiaremos será el Análisis de Componentes Principales, también conocido como PCA por sus siglas en inglés. La principal función de esta técnica es describir al conjunto de datos con variables no correlacionadas entre ellas, estas variables se conocen como *componentes* y están creados de tal forma para que estén ordenados según la variabilidad de los datos que explican. Esto es más fácil de entender con una imagen:

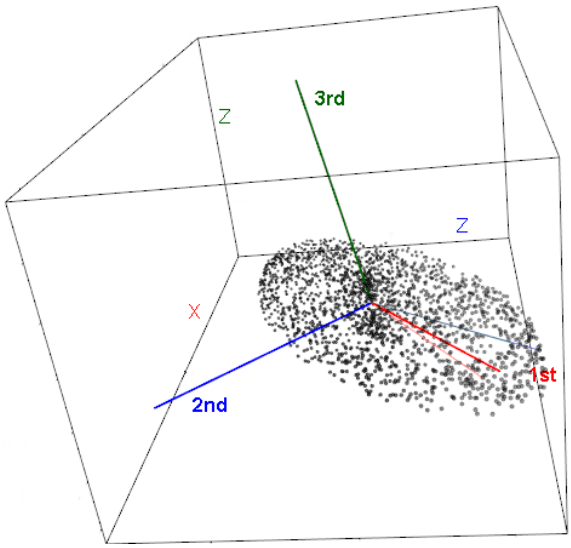

Figura 1: Representación del PCA en una nube de puntos elipsoidal en 3 dimensiones

El Primer Componente Principal, marcado en rojo en la *Figura 1*, captura la mayor parte de la variancia de los datos, pues es la dirección donde más crece el elipsoide. Las otras coordenadas explican en menor medida la variabilidad, siendo más importante la Componente azul.

Esto nos puede ayudar a reducir la dimensionalidad de los datos, ya que como hemos comentado, las primeras componentes explican más sobre el conjunto de datos que las otras, es decir, son más relevantes. Permitiéndonos entonces poder explicar características de los datos usando gráficos de bajas dimensiones, preferiblemente 2D aunque también podríamos estudiar gráficos 3D. Si la proporción de variabilidad de las primeras componentes es suficientemente grande y las proyecciones de los puntos en el plano formado por las primeras componentes son cercanas, entonces podemos saber que también están cerca en el espacio original del conjunto de datos.

3.2. Escalamiento Multidimensional: MDS

El Escalamiento Multidimensional, también conocido como MDS por sus siglas en inglés, es una técnica de análisis multivariante que produce una representación del conjunto de puntos en un espacio euclidiano de manera que las distancias entre los puntos sea prácticamente la misma que en el espacio original. El concepto de distancia puede variar dependiendo de lo que nos interese de nuestro conjunto de puntos. Considerando distancia como una función real que atribuye un número a cada par de puntos y cumple las siguientes propiedades:

(I) *Simetría*:$d(x, y) = d(y, x)$

(II) *Definida positiva*: $d(x, y) \geq 0$ y $d(x, y) = 0$ si y solo si $x = y$

(III) *Desigualdad triangular*: $d(x, y) \leq d(x, z) + d(z, y)$

La distancia más habitual es la distancia euclídea, que es la noción básica que nosotros entendemos sobre que es una distancia. Entonces se puede demostrar que utilizando la distancia euclídea entre las filas de nuestro Data Frame el MDS equivale al PCA. Es por eso que el MDS se conoce como una generalización del otro método. Además, nos es útil cuando la única información que tenemos sobre el conjunto de puntos es la distancia entre ellos.

En este caso nosotros no usaremos el MDS en las filas, lo usaremos en las columnas, para intentar entender como están relacionadas las variables y cuáles tienen más relevancia en cuanto a determinar el porcentaje de grasa.

3.3. Análisis Cluster

El análisis Cluster es la última técnica multivariante que trataremos en este proyecto. Esta es utilizada para clasificar el conjunto de datos en grupos homogéneos. La principal característica de estos métodos de clustering es que no sabemos cuantos grupos hay a priori y es precisamente lo que querremos determinar. Hay distintas formas de implementar el clustering, pero todas se basan en la misma idea: clasificar un conjunto de datos tales que dentro de cada grupo los datos sean los más similares que se pueda y entre grupos tan disimilares como sea posible.

Los distintos métodos de clustering se basarán en dos cosas: que criterio de similitud, usar y que algoritmo emplear para ir generando los clusters. En cuanto al criterio de similitud, usaremos la distancia euclidiana, de hecho la distancia euclidiana de los puntos proyectados a los componentes principales generados por el PCA. Esto es posible, ya que consideraremos que la variabilidad de los componentes principales representa

prácticamente toda la variabilidad del conjunto de dados

Estudiaremos dos algoritmos distintos en este trabajo, los cuales explicaremos a continuación.

3.3.1. Agrupación Jerárquica Aglomerativa

Este algoritmo se inicializa con n clusters distintos, donde n representa el número de puntos que tenemos. Por lo tanto, cada punto es su propio cluster. Entonces, se van mezclando los clusters más cercanos hasta que queden los k clusters que habremos definido previamente. La definición del valor k lo explicaremos más adelante. Ahora nos centraremos en determinar la noción de distancias entre clusters y tenemos varias opciones, nosotros usaremos el *Complete linkage*, equivale a la distancia entre los puntos que están más lejos entre sí.

El proceso de la combinación de clusters se puede representar en un dendrograma. Consiste en un árbol donde la altura representa cuando dos clusters se han agrupado.

3.3.2. K-means

A diferencia del anterior algoritmo, en este ya se establecen k clusters en un inicio. Primero se establecen k centroides y cada punto se agrupa con el centroide más cercano. Entonces se recalculan los centroides promediando los datos que pertenecen a cada cluster. Se vuelve a repetir este proceso hasta que los clusters no se modifican.

3.3.3. Método de Elbow

Este método es el que nos permitirá determinar que número k es óptimo, se basa en medir el *Total within sum of squares* (TWSS):

$$TWSS = \sum_{i=1}^{k} \sum_{x_i \in C_i} d(x_i, \bar{x}_{C_i})^2 \tag{1}$$

Donde $\bar{x}_{C_i}$ es la media de las observaciones en el cluster i.

Si la asignación de clusters es buena, entonces el TWSS será pequeño. Por lo tanto, el elbow method mide el TWSS para diferentes k y un algoritmo en concreto. Escogeremos el valor de k más pequeño posible y que proporcione un TWSS relativamente pequeño.

4. Resultados y su interpretación

Todos los métodos que hemos explicado anteriormente los usaremos a través de librerías y comandos en RStudio.

4.1. Análisis de componentes principales: PCA

Lo primero que haremos será hacer el PCA de nuestros datos a través del comando *prcomp*, además especificaremos que queremos estandarizar las columnas de nuestros datos. De esta forma obtenemos los 15 componentes principales que describen a nuestro conjunto de datos. Cada uno de estos están descritos por 15 coeficientes, pues cada componente principal es la combinación lineal de todas las variables originales.

Una vez obtenidos los *loadings*, así se conocen a los coeficientes que caracterizan a cada componente principal, entonces estudiaremos la variabilidad explicada por cada componente. Esto lo haremos a través del comando *fviz_eig* de la librearia *factoextra*, obteniendo la siguiente gráfica:

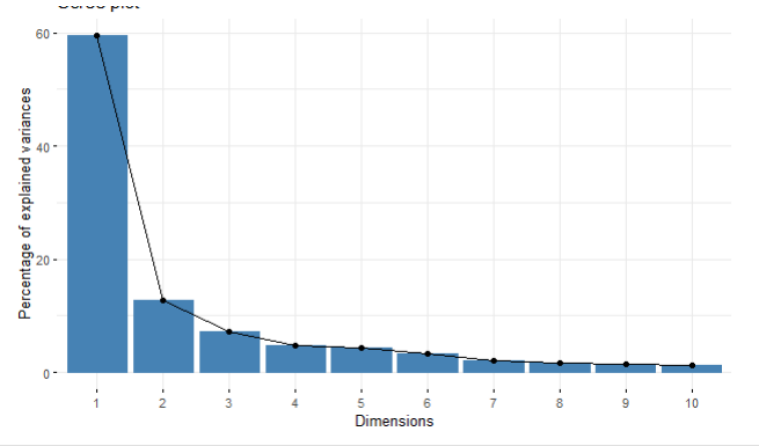

Figura 2: Porcentaje de la variabilidad explicada por cada componente principal

Como vemos, el primer componente explica prácticamente el 60 % de la variabilidad, mientras que el segundo componente explica aproximadamente el 12 %. Podríamos considerar también al tercer componente, pues corresponde al 8 % de la variabilidad, pero esto conllevaría hacer análisis en 3D y se alejaría de lo que hemos ido estudiando en clase. Aunque realmente los resultados parecen indicar que sería una opción muy viable. Aun así, consideraremos que los dos primeros componentes ya explican suficiente variabilidad y serán los que consideraremos.

Ahora representaremos las proyecciones de los puntos del conjunto de datos sobre el plano generado por las dos principales componentes, juntamente con las proyecciones de las variables originales en ese mismo plano. Esto será realizado con un simple comando de R llamado *fvizpcabiplot*, también se encuentra dentro de la

libreria *factoextra*:

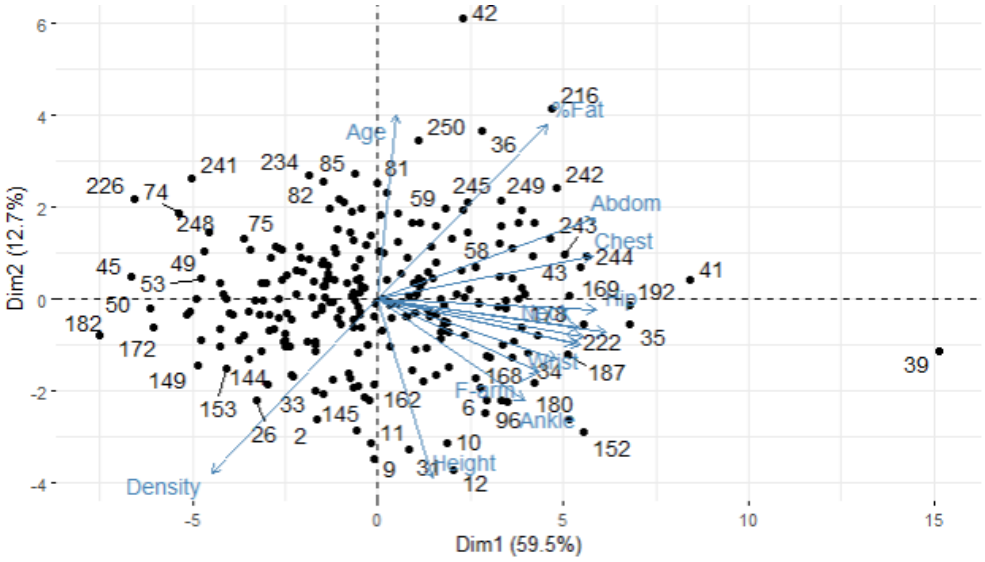

Figura 3: Representación de las proyecciones de los puntos y las variables sobre el plano de los primeros componentes principales

Si nos fijamos en la nube de puntos esta parece formar una elipse centrada en el origen, pero podemos destacar dos observaciones la *42* y la *39* que se diferencian claramente de esta nube de puntos.

Si ahora prestamos atención a las variables, vemos a primera vista que la densidad y el porcentaje de grasa son completamente opuestos. Esto tiene mucho sentido, pues la densidad de la grasa es menos que la del músculo y la de otros tejidos de nuestro organismo, por lo tanto, cuanto más porcentaje de grasa, más densidad. Esto nos indica que si sabemos la densidad de un organismo podemos estimar su porcentaje de grasa de forma lineal. El problema es que la densidad es complicada de obtener, especialmente el volumen de nuestro cuerpo. La única manera que se me ocurre sería hacer una prueba de Arquímedes, es decir, adentrarnos en un recipiente lleno de agua y ver la diferencia de volumen de este recipiente cuando estamos adentro o fuera. Podríamos plantearnos entonces extraer la variable densidad de este conjunto, pues está demasiado relacionada con la variable que intentamos predecir y además es complicada de obtener.

Siguiendo con las variables, si hacemos un estudio de los ejes, podemos ver que los puntos situados a la derecha del eje Y tienen medidas superiores a la media en todas las variables, a excepción de la densidad. En cambio, a la izquierda encontraremos personas con medidas inferiores a la media de la muestra, de nuevo con la excepción de la densidad.

Volviendo de nuevo a los puntos atípicos previamente mencionados, podemos ver que el *42* destaca por tener una altura sorprendentemente baja, que encaja bastante con el gráfico, pues está prácticamente en

sentido opuesto que la variable ***Height***. Podría padecer de enanismo o simplemente ser un error tipográfico. En cuanto a la observación *39* podemos ver que es la que presenta un mayor peso y además tiene los récords de prácticamente todas las medidas, es por eso que se encuentra tan a la derecha.

4.2. Escalamiento multidimensional: MDS

Después de hacer el PCA, estudiaremos el escalamiento multidimensional. Para ello primero escalaremos la base de datos con el comando de R *scale(df)* y calcularemos la distancia entre columnas con el comando *dist(t(dfscaled))*. Entonces aplicamos el MDS a través del comando *cmdscale* y habremos obtenido las distancias entre columnas. Ahora solo falta hacer la representación gráfica:

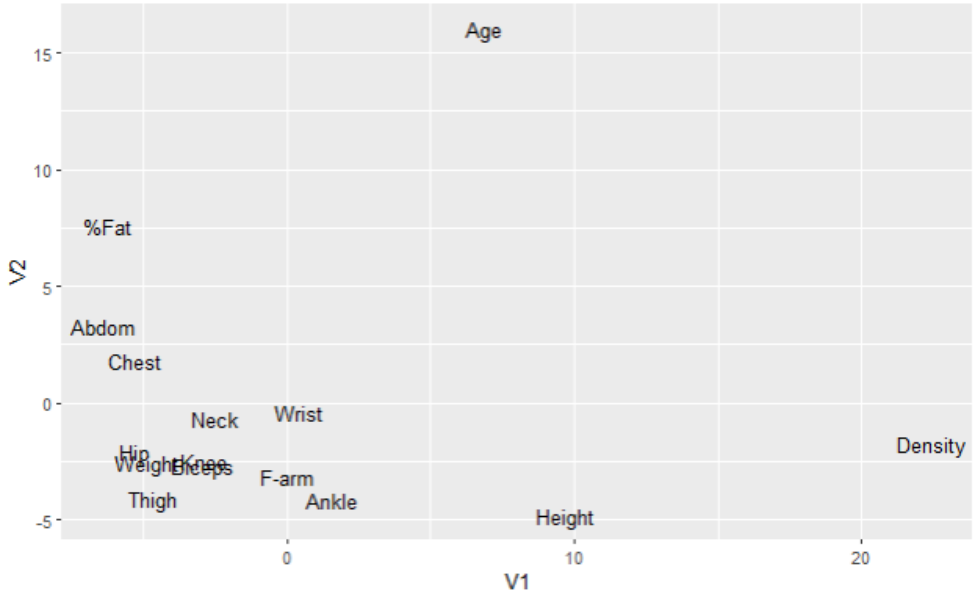

Figura 4: Representación de las distancias entre las variables

Podemos ver que las variables más cercanas a la variable % ***Fat*** son las variables de medidas del cuerpo, siendo la más cercana *Abdomn*. Podemos concluir entonces que el abdomen es la medida que más marca la diferencia en la acumulación de grasa y tiene bastante sentido. Otras medidas que se toman, por ejemplo el cuello o el bíceps, pueden tener grandes medidas sin tener grandes porcentajes de grasa, normalmente por un alto nivel de músculo. En cambio, los abdominales son un músculo plano y por lo que aún estando muy desarrollado, este no aumenta considerablemente su medida. El abdomen aumenta prácticamente con la cantidad de grasa.

De nuevo podemos ver como la densidad está muy alejada de las otras variables, por lo que ya habíamos comentado. También hace falta apreciar como la altura y la edad no están demasiado relacionadas con el porcentaje de grasa. Si bien es cierto que se tiene la creencia que a medida que pasan los años tiende a aumentar el porcentaje de grasa corporal, este cambio no es tan importante.

4.3. Clustering

Por último, analizaremos el resultado de aplicar Clustering a nuestros datos. Podríamos también aplicar clustering para agrupar variables que mantengan relaciones, pero no parece que tenga tanto potencial explicativo, pues ya nos lo podemos imaginar después de analizar el MDS del apartado anterior.

Empezaremos aplicando el método de la agrupación jerárquica aglomerativa, para ello tendremos que emplear el método de Elbow para determinar el número de clusters que tendremos. Empleando el comando *fviz_nbclust* sobre la base de datos escalada y especificando que usaremos el método de *Complete linkage*, obtenemos:

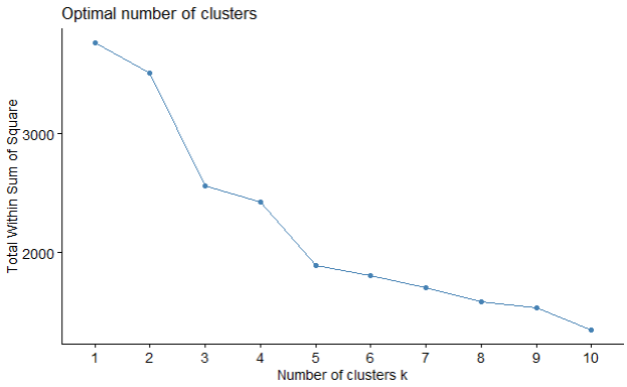

Figura 5: Representación del error generado para cada número de clusters según el método Elbow

Podemos determinar que la k óptima será $k = 5$, pues ahí es donde se produce el último cambio drástico en la disminución del TWSS.

Una vez determinado el número de clusters, solo hace falta aplicar el algoritmo que utilizaremos el comando *fviz_cluster*, asegurándonos de especificar el número de clusters que habíamos acordado. En la Figura 6, podemos apreciar como se distribuyen los clusters y es curioso como existen dos clusters independientes que equivalen a los dos valores atípicos que habíamos identificado en el PCA. Lo mejor sería eliminarlos para no sobre ajustar nuestros posibles modelos de predicción y entonces nos quedarían tres clusters.

En el cluster rojo encontraríamos las personas con un índice de grasa baja, pues si nos acordamos del PCA todas las componentes relacionadas con el porcentaje de grasa tenían una dirección hacia la derecha. El grupo amarillo tomaría valores normales de porcentaje graso y por último el cluster verde con niveles altos en grasa. No sería una clasificación del todo correcta, pues encontraríamos muchos atípicos, por ejemplo el caso de la observación *12*, que se encuentra en el grupo verde aun teniendo un 7.8 % de grasa corporal, eso es debido a que es muy alto y, por lo tanto, también tendrá valores altos en todas las demás variables sobre medidas del cuerpo.

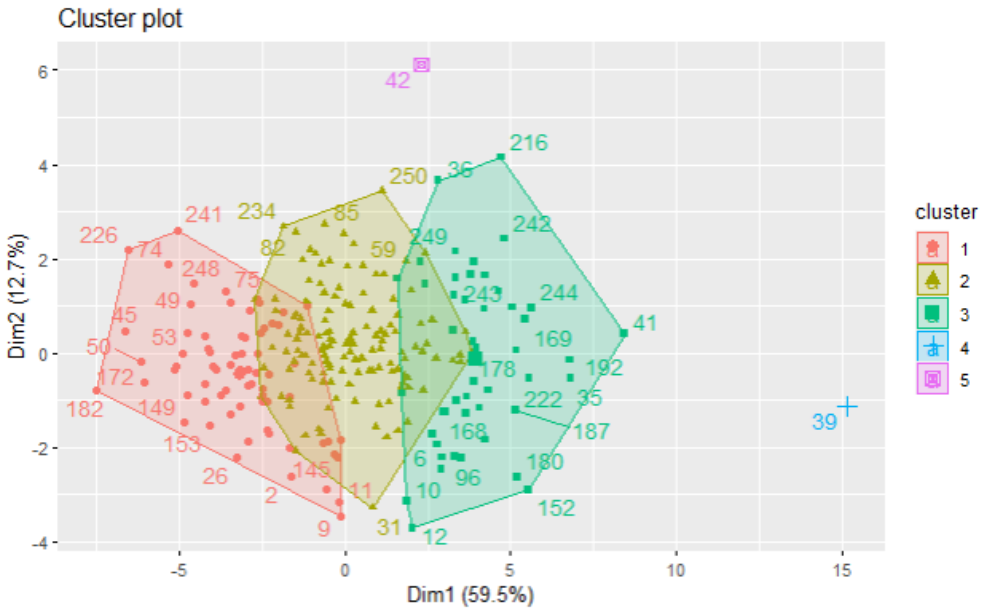

Figura 6: Representación del clustering utilizando el método de agrupación jerárquica aglomerativa para k = 5

Si pasamos al método de K-means, tendremos que volver a mirar el gráfico de Elbow para determinar el valor óptimo de K, pues no siempre coinciden entre los dos algoritmos. Como veremos en este caso encontramos que $k = 3$, un valor mucho más lógico por lo que comentábamos anteriormente.

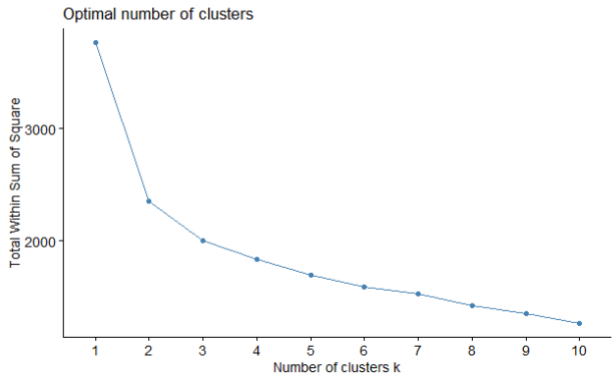

Figura 7: Representación del error generado para cada número de clusters según el método Elbow

Si representamos ahora los diferentes clusters, observamos que coincide mucho mejor con la descripción que habíamos mencionado para cada cluster. Donde ahora el grupo verde representaba a los individuos con baja cantidad de grasa, de color rojo a los que presentaban porcentajes intermedios y por último, de color azul, los que presentan índices de grasa elevada e incluso peligrosos para la salud.

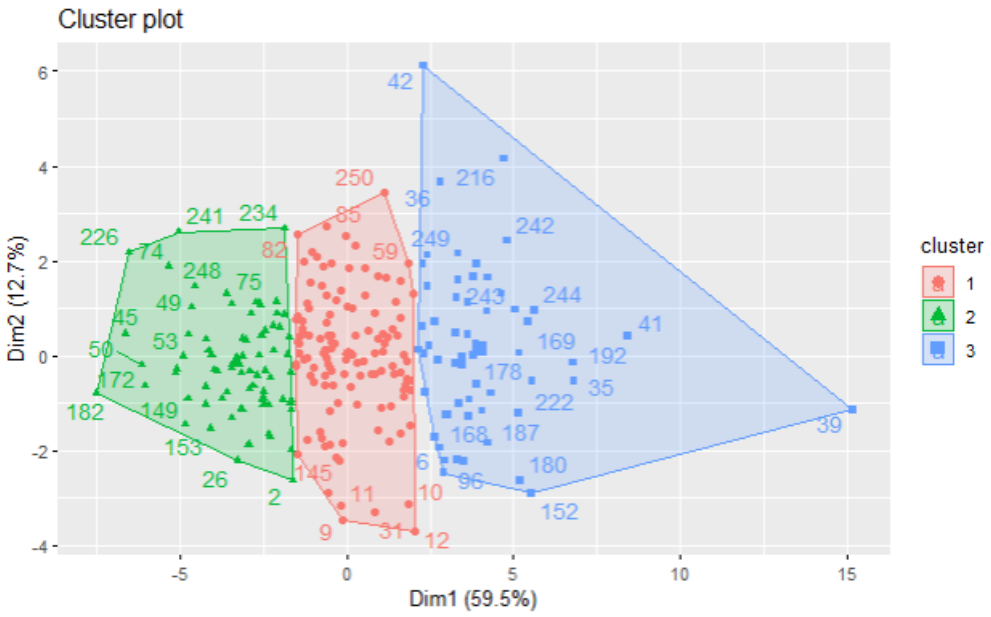

Figura 8: Representación del error generado para cada número de clusters según el método Elbow

5. Conclusiones y discusiones

Este trabajo nos ha ayudado a entender como se relaciona el porcentaje de grasa con otras variables de nuestro cuerpo, pero no solo eso, sino que además hemos podido a aprender a trabajar con técnicas de análisis de datos y extraer conclusiones a partir de ellas. Este ha sido un simple ejemplo de una posible aplicación de estos métodos, pero como hemos explicado en la introducción, realmente hay infinitas aplicaciones de estos. Por lo que seguramente en un futuro acabemos haciendo estudios similares a los que hemos hecho aquí para alguna empresa o para estudiar algo de nuestro interés.

En la introducción había propuesto determinar que variables eran las más importantes para calcular el porcentaje de grasa y partiendo del MDS podemos deducir que estas son las medidas del abdomen, del pecho y del cuello. Lo que podemos hacer para generar unas predicciones es construir un modelo lineal que nos aproxime el porcentaje de grasa en función de estas 3 variables explicativas. En R utilizaremos el comando *lm*. Obteniendo la siguiente ecuación:

$$\%Fat = 0{,}862 \times Abdomn - 0{,}110 \times Chest - 0{,}897 \times Neck - 15{,}530 \tag{2}$$

Tal y como habíamos comentado, un abdomen grande contribuye a altas tasas de grasa, mientras que tener un pecho grande o un cuello grande beneficia a tener menos grasa, pues en esas zonas se acumula más músculo que grasa. Aplicando esta función a mis medidas se obtiene que tengo un porcentaje de grasa del 7.1 %. Concuerda con mi constitución, aunque posiblemente un poco más bajo del real.

Para acabar este trabajo, comentaremos algunas posibles ampliaciones, la principal sería aumentar este trabajo con un estudio en el caso de las mujeres. Una posible hipótesis seria que la grasa está relacionada con las medidas de glúteos, abdomen y muslos, que es donde las mujeres suelen acumular más grasas. Otra posible ampliación podría comparar mi predicción sobre mi porcentaje de grasa de alguna forma con el valor real. Para de esta manera ver el error asumido por el modelo lineal. Podríamos esperar un error significativo por la simplicidad del modelo, además este solo explica un 70 % de la variabilidad.

6. Webgrafía

Información sobre la grasa corporal y formas de medirla:
- http://www.statistics4u.com/fundstat_eng/data_bodyfat.html
- https://www.runtastic.com/blog/es/calculo-del-porcentaje-de-grasa-corporal/
- https://medlineplus.gov/spanish/pruebas-de-laboratorio/densitometria-osea/
- https://www.quironsalud.es/es/comunicacion/notas-prensa/funciona-bascula-bioimpedancia

Información sobre el PCA:
- https://www.joyofdata.de/blog/illustration-of-principal-component-analysis-pca/

Información sobre el clustering:
- https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm