UNIVERSIDAD DE GRANADA

PRÁCTICA EVALUABLE: APRENDIENDO DE LOS DATOS

ANA BUENDÍA RUIZ-AZUAGA

Correo electrónico
anabuenrua@correo.ugr.es
FACULTAD DE CIENCIAS

Granada, a 14 de enero de 2022

ÍNDICE GENERAL

1

PRÁCTICA

1.1 ABSTRACT

El problema a estudiar trata sobre información socioeconómica y cultural de treinta y cuatro países, recogida en once variables como número de libros publicados, tasa de consumo energético o densidad de población de cada uno de ellos. Para ello, se ha realizado un análisis exploratorio univariante, incluyendo estudio de los valores perdidos, un análisis descriptivo numérico clásico y tratado los outliers y comprobado el supuesto de normalidad de cada variable. Asimismo se ha realizado un análisis exploratorio multivariante comprobando los supuestos de correlación con el test de Barlett y los outliers. Además se ha estudiado la posibilidad de reducción de la dimensión mediante variables observables y latentes eligiendo el número óptimo de factores a considerar en cada caso. Finalmente se ha tratado de llevar a cabo varios métodos de aprendizaje supervisado, pero debido a la naturaleza de los datos no se han podido realizar, por lo que se ha usado en su lugar la técnica de aprendizaje no supervisado kmeans para ver similaridades entre países. Tras realizar el estudio se ha observado que los datos son insuficientes y por tanto no se ha obtenido ninguna información concluyente.

1.2 INTRODUCCIÓN

Se va a realizar un estudio sobre treinta y cuatro países distintos, de los cuales se conoce el número de libros publicado, cociente entre número de individuos en ejército de tierra y población total, cociente entre población activa y total, tasa de consumo energético, población del sector servicios, población del sector agrícola, tasa de médicos por habitante, esperanza de vida, tasa de mortalidad infantil, densidad de población y porcentaje de la población urbana de cada uno de ellos, haciendo un total de once variables a analizar.

Para comprender mejor los datos y la información que aportan, se realizará un análisis exploratorio univariante y multivariante.

El análisis exploratorio de los datos univariante incluye un estudio de los valores perdidos, un análisis descriptivo numérico clásico y tratado los outliers y comprobado el supuesto de normalidad de cada variable.

Asimismo se ha realizado un análisis exploratorio multivariante comprobando los supuestos de correlación con el test de Barlett y los outliers respecto a la distribución normal multivariante. Además se ha estudiado la posibilidad de reducción de la dimensión mediante variables observables y latentes intentado obtener el número óptimo de factores a considerar en cada caso.

Al comprender y estudiar la información proporcionada por todas las variables de cada país se quiere construir dos modelos de clasificación, uno lineal y otro cuadrático, para así poder predecir en qué continente se encuentra un país dados los datos ya mencionados. Nos encontraremos problemas al realizar esto por la poca cantidad de datos de la que disponemos, por lo que se hará un agrupamiento no supervisado usando kmeans con el fin de estudiar las similaridades y diferencias que encontramos entre los países, analizando el mejor número de clusters para realizar esta agrupación.

Así podremos estudiar las similitudes y diferencias entre los países, apreciando si hay diferencias notables entre países de distintos continentes, tales como indicadores de calidad de vida, como su esperanza de vida o tasa de mortalidad infantil.

De esta manera, podremos analizar cuáles son los factores más relevantes que afectan al desarrollo de un país, comprobar si hay relación entre la calidad de vida del país y el continente en el que se encuentra y solamente a partir de estos datos poder, por tanto, predecir a qué continente pertenece un país.

1.3 MATERIALES Y MÉTODOS

1.3.1 *Materiales*

La base de datos con la que vamos a trabajar consta de treinta y cuatro observaciones de países distintos, cada una con once variables describiendo la situación socioeconómica del país en cada caso.

Las variables son:

- Número de libros publicados (ZTLIBROP).
- Cociente entre número de individuos en ejército de tierra y población total del estado (ZTEJERCI).
- Cociente entre población activa y total (ZTPOBACT).
- Tasa de consumo energético (ZTENERGI).
- Población del sector servicios (ZPSERVI).
- Población del sector agrícola (ZPAGRICU).
- Tasa de médicos por habitante (ZTMEDICO).
- Esperanza de vida (ZESPVIDA).

Variable	Media	Desviación típica
ZTLIBROP	6.729069e-17	0.9847319
ZTEJERCI	1.859184e-16	1
ZTPOBACT	-2.952027e-16	1
ZTENERGI	3.586955e-17	1
ZPSERVI	-3.529078e-17	1
ZPAGRICU	6.355338e-17	1
ZTMEDICO	2.416495e-17	1
ZESPVIDA	-1.72884e-16	1
ZTMINFAN	8.982296e-17	1
ZPOBDENS	1.06373e-16	1
ZPOBURB	4.511206e-16	1

Cuadro 1: Variables con su media y desviación típica

- Tasa de mortalidad infantil (ZTMINFAN).
- Densidad de población (ZPOBDENS).
- Porcentaje de la población urbana (ZPOBURB).

Como primer contacto, en (1) se muestran las variables y su media y desviación típica.

1.3.2 *Métodos estadísticos*

Durante el estudio se han usado diversos métodos, entre los que se encuentran para el análisis de datos univariante la exploración descriptiva de cada variable (obteniendo así medidas de centralidad, desviación típica, rango intercuartílico, medidas de skewness, curtosis, asimetría de Yule, asimetría de Kelly y asimetría de Kelly adimensional), exploración gráfica con diagramas de cajas y bigotes para detectar los outliers, sustitución de valores perdidos y outliers por la media, qqplots para comprobar la normalidad y el test de Levene para comprobar la homocedasticidad.

Durante el análisis multivariante se ha usado el test de Barlett, análisis de componentes principales y la regla de Abdi para determinar el número óptimo de sus componentes, así como técnicas gráficas (fviz_pca_var, fviz_pca_ind y fviz_pca). También se ha usado el método del factor principal, método de máxima verosimilitud, scree_plots, técnicas gráficas para representar la asociación de factores y el test de factanal para comprobar que el número de factores elegido es suficiente. Además se ha usado el chi-square qq-plot para outliers, y para comprobar la normalidad multivariante el test de Royston y de Henze-Zirkler. Para comprobar la homogeneidad de la varianza se ha usado el test de Box-M y también se ha usado clasificación mediante discriminante lineal y cuadrático, así como agrupación por kmeans usando el método de la silueta para determinar el número óptimo de clusters.

1.4 RESULTADOS

Primero se ha comprobado la presencia de valores perdidos en el conjunto de datos, pero al no representar más del 5 % de datos totales se pueden sustituir por la media sin que esto afecte a los resultados.

No se ha necesitado recodificación ni agrupamiento, y los datos ya estaban estandarizados, luego no ha sido necesario realizar ninguno de estos métodos.

Ante la presencia de algunos outliers en cada variable se han reemplazado también por la media, y posteriormente mediante gráficos qqplot se ha comprobado que algunas variables no siguen una distribución normal, mientras otras sí.

Asimismo, también se ha comprobado que pese a que existe homocedasticidad en la mayoría de grupos, en algunos no se da.

Respecto al análisis exploratorio multivariante, al realizar el test de Barlett se rechaza la hipótesis nula, por lo que los datos resultan estar correlados.

Buscamos ahora reducir la dimensión del problema, y para ello se han representado distintas gráficas y usado la regla de Abdi, obteniendo como resultado escoger 3 componentes principales.

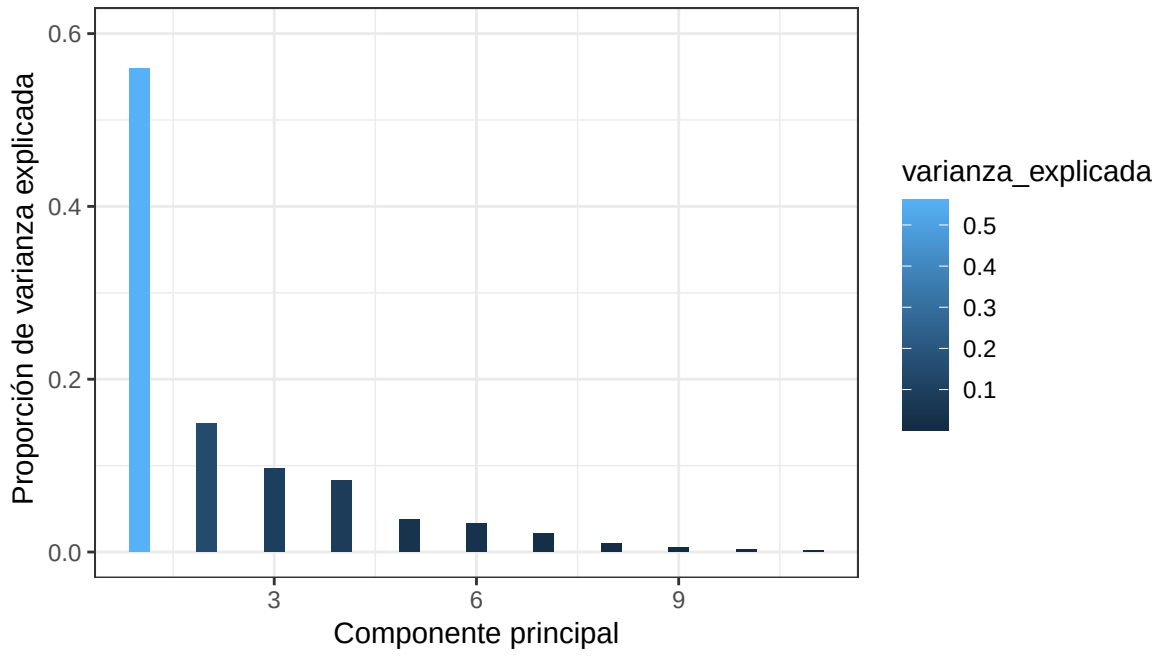

Figura 1: Proporción de varianza explicada y acumulada

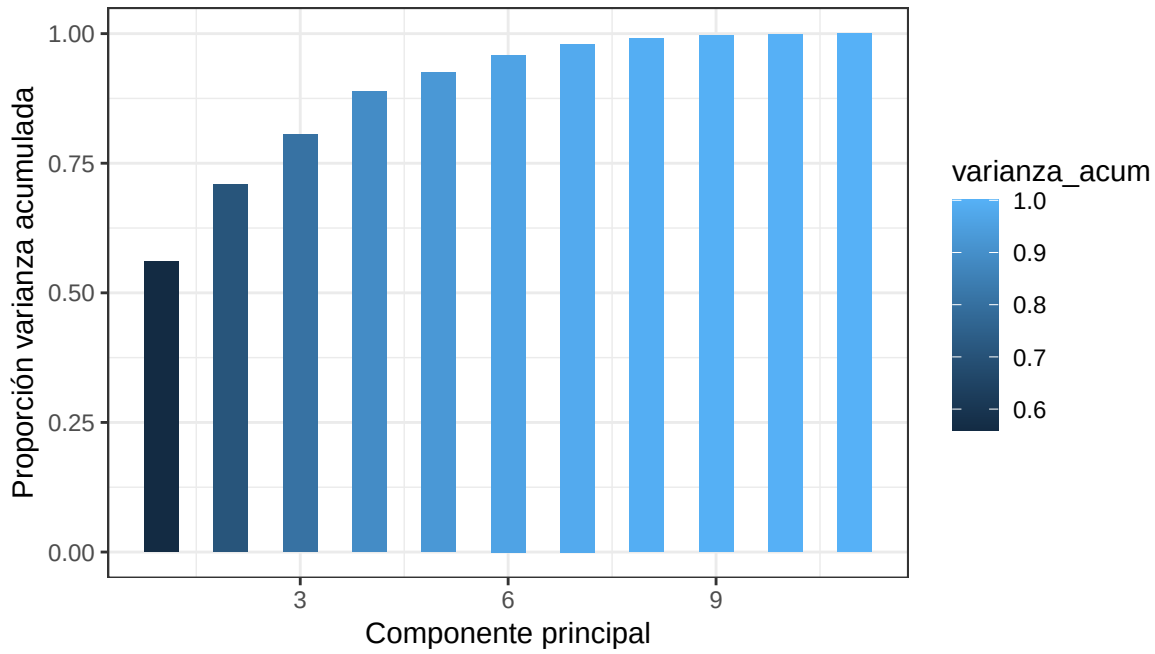

Figura 2: Proporción de varianza acumulada

En (1) se puede ver la proporción de la varianza explicada y en (2) la proporción de varianza acumulada.

Para estudiar qué variables son más relevantes en peso en cada una de las componentes principales se han representado cada una de ellas.

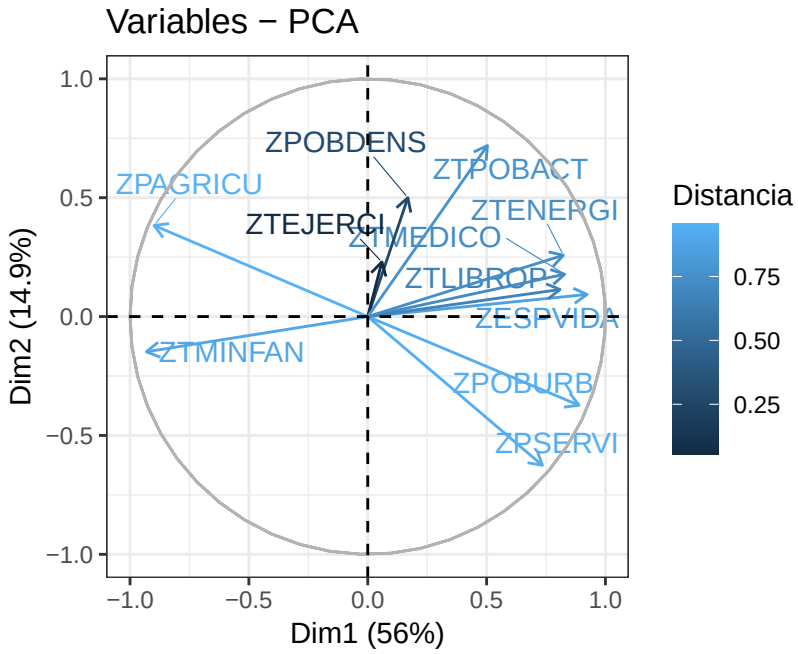

Figura 3: Variables PCA entre componentes 1 y 2

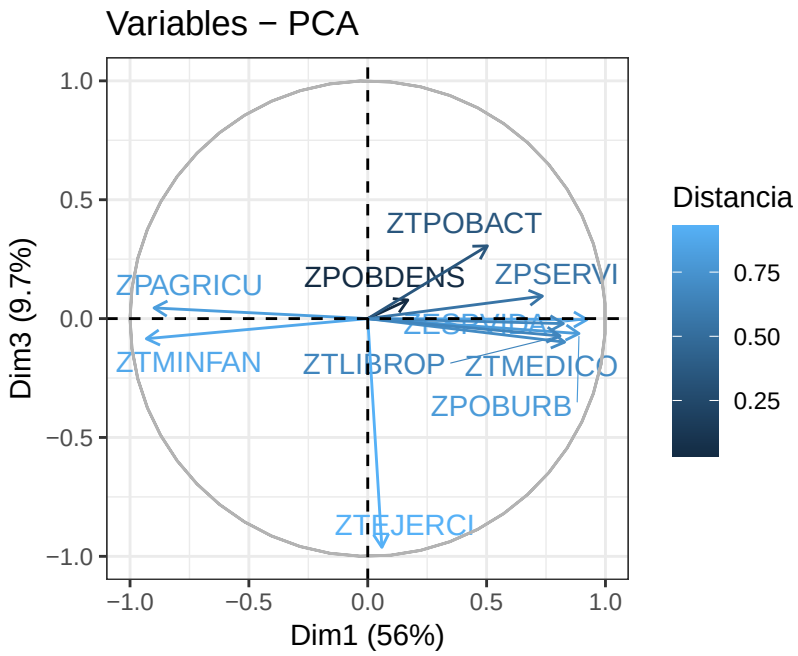

Figura 4: Variables PCA entre componentes 1 y 3

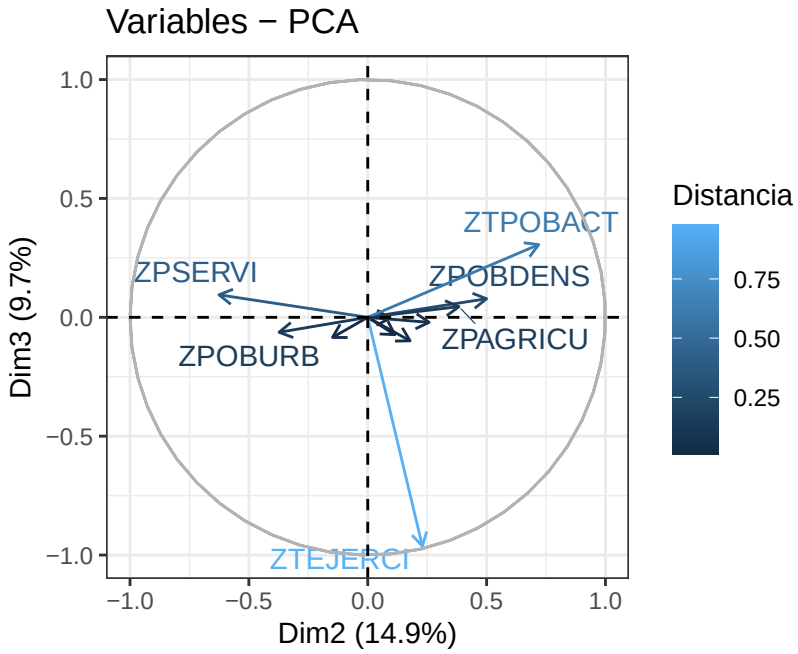

Figura 5: Variables PCA entre componentes 2 y 3

Como se aprecia en (3) las variables más relevantes para la primera componente son ZTMINFAN, ZESPVIDA, ZTMEDICO, ZTENERGI entre otras. Mientras en la segunda componente, como se ve en (3) y (5) es ZTPOBACT, y en la tercera componente, como también se refleja en (4) es ZTEJERCI.

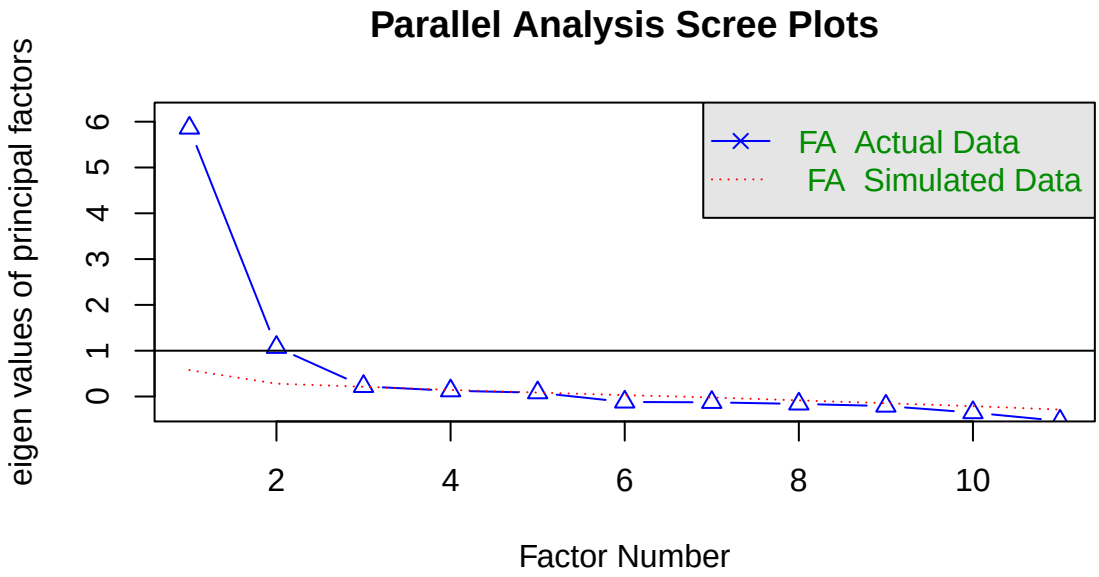

Figura 6: Scree plot de análisis factorial

Al realizar la reducción de dimensionalidad mediante variables latentes se ha calculado el número óptimo de componentes mediante scree plots, como (6), resultando este ser 3 y comprobado usando factanal que ese número de componentes es suficiente y 2 no, ya que las funciones usadas para obtener los screeplots a veces devolvían 2 o 3 indistintamente.

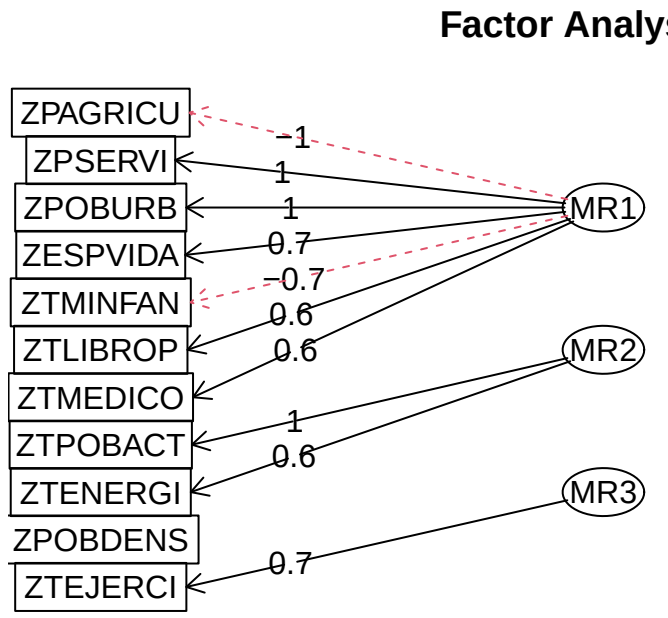

Figura 7: Análisis factorial

En (7) se puede ver como la mayoría de variables las explica el primer factor, y destaca como ZTPOBDENS no está asociado a ninguno de los factores.

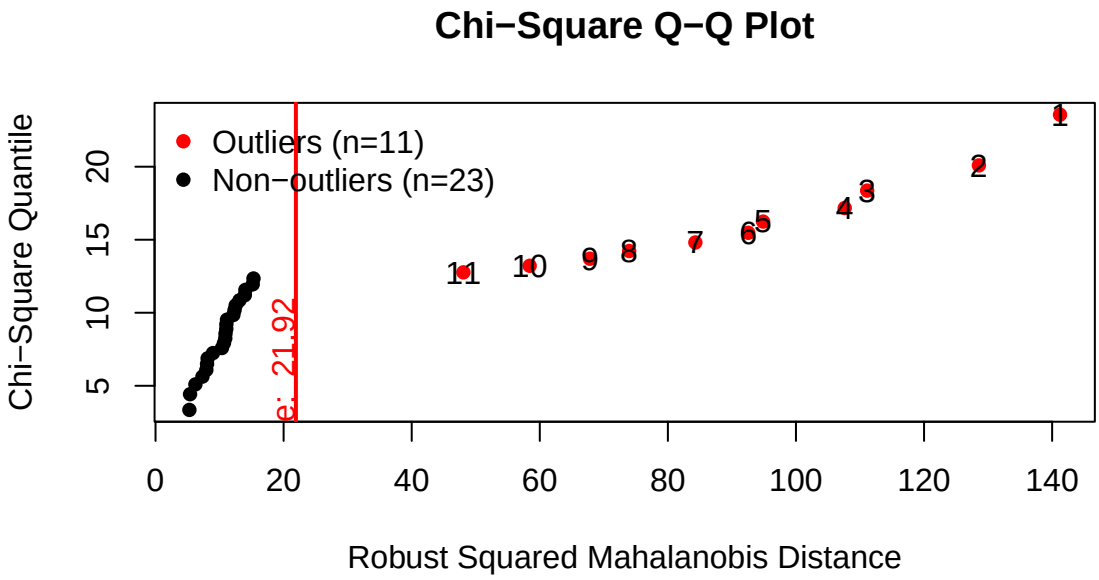

Figura 8: Estudio de outliers

A continuación se han estudiado los outliers, como se ve en (8) ha detectado 11 puntos como outliers. Y, al comprobar la normalidad con el test de Royston y el test de Henze-Zirkler el test de Royston rechaza que la distribución sea normal, mientras que el segundo sí acepta la hipótesis de normalidad.

Finalmente con el fin de construir un clasificador se ha intentado comprobar la homogeneidad de la varianza mediante un test de chi cuadrado, pero no se ha podido al no haber suficientes datos.

Pese a esto, se ha construido un clasificador mediante discriminante lineal que en validación ha proporcionado un 53 % de error en el conjunto de entrenamiento, mientras el clasificador mediante discriminante cuadrático no ha podido construirse por falta de datos.

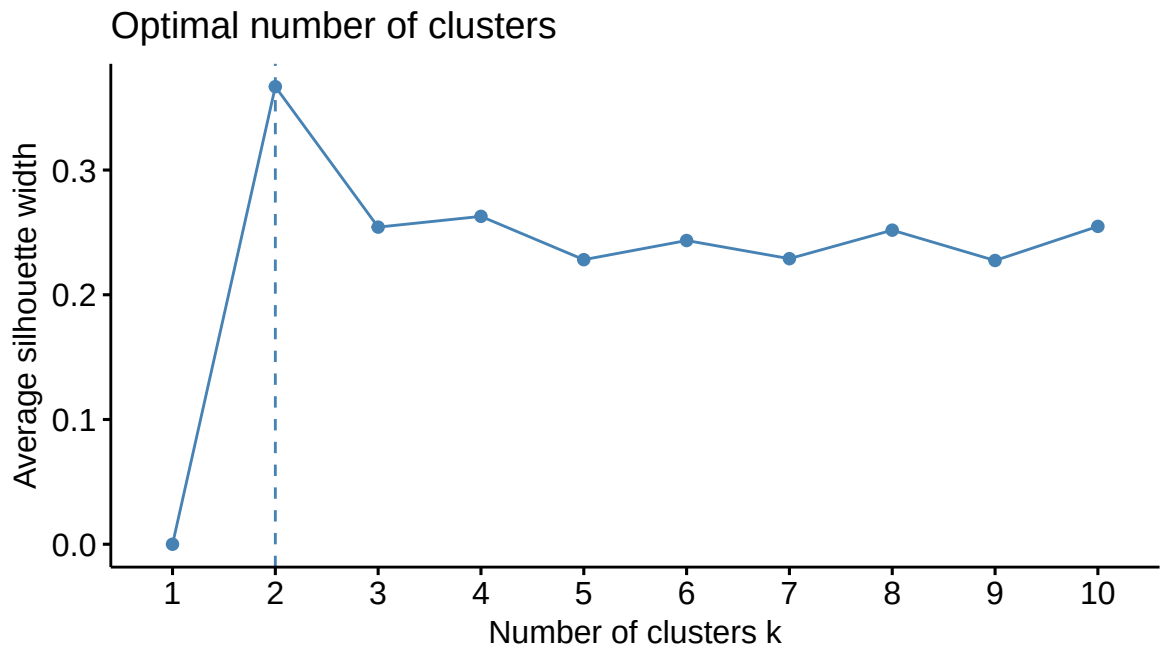

Figura 9: Método de la Silueta para obtener el número óptimo de clusters

Por ello se ha realizado un agrupamiento mediante kmeans usando el método de silhouette para obtener que el número óptimo de clusters es 2, como se ve en (9).

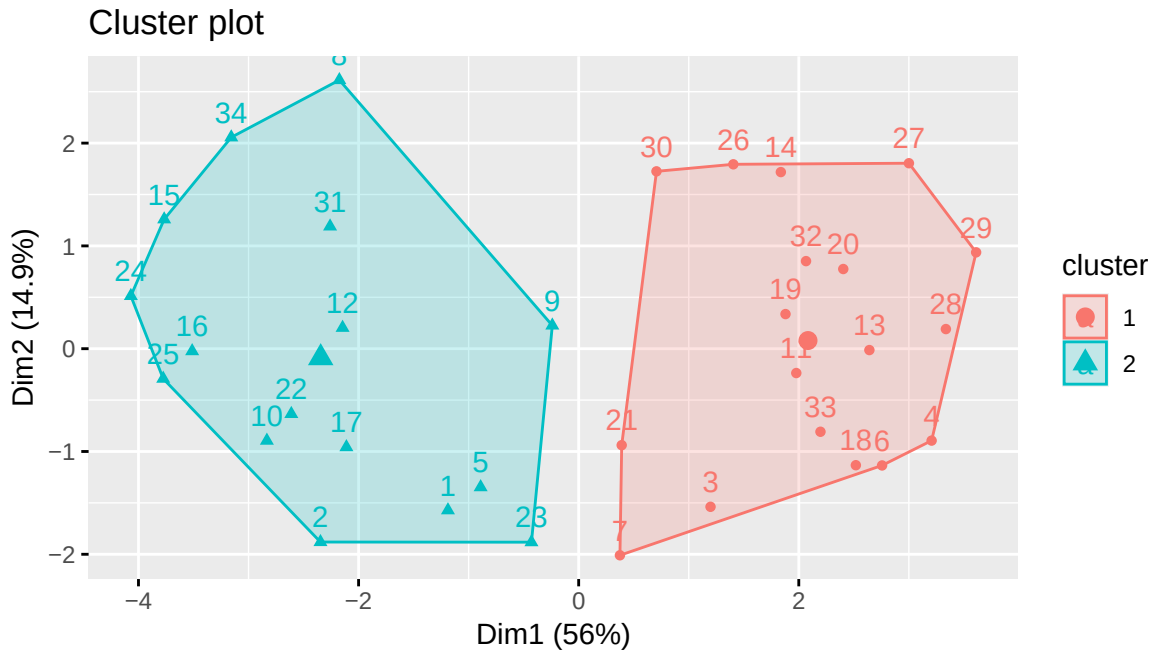

Figura 10: Clustering mediante kmeans

En (10) se ve el resultado de aplicar clustering con 2 clusters usando el método kmeans.

1.5 DISCUSIÓN

A la vista de los resultados anteriores queremos averiguar qué variables son más relevantes en el desarrollo de un país y comprobar si hay relación entre los factores socioeconómicos de un país y su continente, tratando de predecir según solamente estos datos a qué continente pertenece cada país.

Al realizar el análisis de componentes principales se han obtenido 3 componentes, de las cuáles la primera es la que más varianza explica. Por tanto, las variables más relevantes que estamos buscando serán las que tengan más peso en esta componente. Por ello, como vimos en (3) y (4) las variables más importantes para los países son ZTMINFAN, ZESPVIDA, ZTMEDICO y ZTENERGI.

Otro resultado llamativo ha sido el de análisis factorial, ya que en (7) la variable ZTPOBDENS no está asociada a ningún factor, lo que puede ser resultado de que su correlación con todas las demás variables y por tanto los factores es muy baja.

Asimismo, al estudiar los outliers como vimos en (8) muchos datos se consideran outliers, pero esto no parece razonable ya que los outliers de cada variable se sustituyeron previamente por la media y considera casi un tercio de los datos outliers. Por tanto, parece razonable asumir que los datos en realidad no son outliers y este resultado atípico, así como que el test de Royston rechace que los datos siguen una distribución normal multivariante mientras el test de Henze-Zirkler asume normalidad se pueden deber a que la muestra de datos es muy pequeña. Esto también se ha podido apreciar cuando no se pudo comprobar la homogeneidad de la varianza o construir el modelo de clasificación mediante discriminante cuadrático por el mismo motivo.

Otra posible causa de considerar tantas instancias como outlier es el rango de las variables, pues todas se encuentran cerca del intervalo [0,1].

Respecto al clasificador mediante discriminante lineal entrenado, dado que su error en training es del 53 % no es fiable para usarlo, pues su error en test será aún mayor. Un error así de alto puede deberse a varios factores, como ruido, que todas las variables al estar previamente estandarizadas estaban en un rango muy pequeño, outliers no tratados o a la falta de instancias del conjunto de entrenamiento, que producen sobreajuste o pueden no ser representativas de ejemplos reales.

Dado que el clasificador mediante discriminante lineal no ha proporcionado buen resultado y no se ha podido construir el clasificador mediante discriminante cuadrático finalmente se ha optado por realizar clustering para ver si se aprecian diferencias significativas entre los grupos.

Al estudiar el número óptimo de clusters (9) se aprecia que la máxima separabilidad se encuentra en 2 clusters, mientras que hay 5 continentes, lo que podría implicar que realmente no afecta tanto el continente al que se pertenezca, si no otros factores.

Al realizar el agrupamiento como se ve en (10) los clusters resultantes no se superponen en su proyección y las diferencias entre las medias de cada una de las variables estudiadas para cada cluster son muy distintas. Aún así, dada la aleatoriedad del algoritmo kmeans y los pocos datos de los que se dispone, no se podría asegurar que los clusters realizados sean representativos, pues se necesitarían más datos.

1.6 CONCLUSIÓN

Se ha observado que algunas variables dependen significativamente de otras, y que por lo tanto las más significativas en las que los países deberían concentrar esfuerzos en mejorar son tasa de mortalidad infantil, esperanza de vida, ratio de médicos por habitantes y consumo energético, así como la población activa y el ratio de población en el ejército. Mejorando algunas o todas estas características el resto deberían variar de acuerdo a ellas.

También parece que el continente en el que se encuentra cada país no se puede predecir con las variables dadas, pero no es posible concluir si esto se debe a falta de variables contempladas, las variables elegidas no son buenas predictoras del continente o a falta de ejemplos representativos de cada continente en el conjunto de datos.

Finalmente, y dado que muchas hipótesis importantes tales como la normalidad univariante o multivariante, tratamiento de outliers o la homogeneidad de la varianza, al igual que el clasificador mediante discriminante cuadrático no se han podido comprobar o construir debido al tamaño de la muestra sobre la que se ha trabajado, se deben reunir más datos para repetir el estudio sobre un conjunto lo bastante grande como para poder comprobar las hipótesis requeridas y tener seguridad tanto en los clasificadores como en el agrupamiento.

2

BIBLIOGRAFÍA

- Código y apuntes dados en clase.