Análisis DALY

MDP I

Álvaro Prado Expósito

Óscar García Martínez

Víctor Mañez Poveda

Aleixandre Tarrasó Sorní

Índice de contenidos:

1. Introducción

Tras la búsqueda de diferentes bases de datos que cumplieran con las exigencias y necesidades que nos plantea este proyecto, finalmente encontramos una que nos ofrecía un gran número de variables y de observaciones, aspecto fundamental para el desarrollo del mismo.

Esta base de datos trata de mostrar información acerca del DALYs en diferentes poblaciones entre los años 1990 y 2019. Esta abreviatura hace referencia a "Disability-Adjusted Life Years" la cual es una medida utilizada en el campo epidemiológico y de la salud que trata de cuantificar la carga de una enfermedad en una población teniendo en cuenta tanto la mortalidad como la morbilidad causada por enfermedades o lesiones.

En nuestra base de datos contamos con un total de 6150 observaciones, donde cada observación corresponde a un país en un año determinado. Cada país está representado 30 veces, debido a que se tienen datos desde 1990 hasta 2020 para cada uno de ellos. Por otra parte, nuestra BBDD está compuesta por 28 variables, 25 de ellas hacen referencia al DALY por el tipo de enfermedad, mientras que las de Entity, Code y Year hacen referencia al país, el código de dicho país y al año respectivamente.

En la base de datos original había datos faltantes en el atributo 'Code' porque se medían valores de ciertas regiones que no tenían ningún código asignado. Sin embargo no planteamos realizar un análisis de los datos por regiones geográficas, realizaremos el análisis por países.
Tras esta explicación pasaremos a mostrar cada una de sus variables, con sus significado y el tipo que es.

VARIABLES	SIGNIFICADO	TIPO
Entity	País	Categórica Nominal
Code	Código identificativo del país	Categórica Nominal
Year	Año de los datos	Fecha
Self-harm	DALY por autolesiones	Numérico Continuo
Exposure to forces of nature	DALY por fuerzas de la naturaleza	Numérico Continuo
Conflict and terrorism	DALY por conflictos armados	Numérico Continuo
Interpersonal violence	DALY por violencia interpersonal	Numérico Continuo
Neglected tropical diseases and malaria	DALY por enfermedades tropicales	Numérico Continuo
Substance use disorders	DALY por abuso de substancias	Numérico Continuo
Skin and subcutaneous diseases	DALY por enfermedades cutáneas y subcutáneas	Numérico Continuo
Enteric infections	DALY por infecciones intestinales	Numérico Continuo
Diabetes and kidney diseases	DALY por diabetes y enfermedades renales	Numérico Continuo

Cardiovascular diseases	DALY por enfermedades cardiovasculares	Numérico Continuo
Digestive diseases	DALY por enfermedades digestivas	Numérico Continuo
Nutritional deficiencies	DALY por deficiencias nutricionales	Numérico Continuo
Respiratory infections and tuberculosis	DALY por infecciones respiratorias y tuberculosis	Numérico Continuo
Neonatal disorders	DALY por enfermedades de nacimiento	Numérico Continuo
Chronic respiratory diseases	DALY por enfermedades respiratorias	Numérico Continuo
Other non-communicable diseases	DALY por enfermedades no comunicadas	Numérico Continuo
Maternal disorders	DALY por enfermedades en el embarazo	Numérico Continuo
Unintentional injuries	DALY por lesiones no intencionadas	Numérico Continuo
Musculoskeletal disorders	DALY por enfermedades musculoesqueléticos	Numérico Continuo
Neoplasms	DALY por neoplasmas	Numérico Continuo
Mental disorders	DALY por enfermedades mentales	Numérico Continuo
Neurological disorders	DALY por enfermedades neurológicos	Numérico Continuo
HIV/AIDS and sexually transmitted infections	DALY por VIH y enfermedades de transmisión sexual	Numérico Continuo
Transport injuries	DALY por lesiones en el transporte	Numérico Continuo
Sense organ diseases	DALY por enfermedades en los órganos sensoriales	Numérico Continuo

Además, para el objetivo del PLS trabajamos con una base de datos que contiene valores del HDI para cada país desde 1990 a 2019.

VARIABLES	**SIGNIFICADO**	**TIPO**
HDI	Índice de Desarrollo Humano	Numérico Continuo

Ahora pasaremos a plantear los objetivos. En un primer momento, no teníamos los conocimientos suficientes para poder plantear adecuadamente los objetivos a seguir, es por ello que en la primera entrega de la base de datos no pudimos plantearlo adecuadamente. Sin embargo, según han ido

avanzando las clases y hemos dado diferentes herramientas hemos podido plantear los siguientes objetivos:

- Mediante un PCA reducir la dimensionalidad y comprender la naturaleza de nuestra base de datos. Entendiendo las diferencias entre países según los valores DALY de distintas afecciones.

- Agrupar las clases a las que pertenecen los distintos países mediante grupos formados por variables con un comportamiento parecido mediante clustering y poder agrupar los países en función del tipo que sean.

- Confirmar los resultados creados por el clustering mediante un análisis discriminante de Fisher.

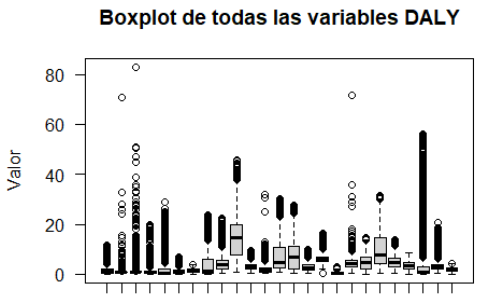
- Clasificar por nivel de desarrollo según el índice de desarrollo humano (HDI), mediante un PLS.

2. Análisis Exploratorio y preprocesamiento de los datos.

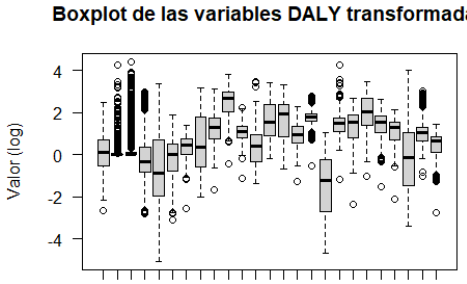
Ninguna variable ha sido eliminada del modelo. Consideramos que todas son interesantes para el estudio que vamos a realizar, y ninguna tiene valores tan anómalos como para tener que descartarla.

El único registro que hemos eliminado es el de Ruanda en el año 1994. Como se puede ver en el Anexo 2, esta observación tiene un valor de T^2 de Hotelling significativamente anómalo. Por ende, con tal de que no influya a las componentes principales y a futuros análisis, la eliminamos de la base de datos.

Realizando el análisis exploratorio, creamos un boxplot que muestra la variabilidad de cada variable. Como se puede ver en el gráfico, todas las variables tienen una pronunciada asimetría positiva.



Para solucionar este problema, tras centrar las variables las logaritmizamos. Como se puede ver en el boxplot del para las variables logaritmizadas, corregimos considerablemente la asimetría. Este análisis se ha realizado con detalle en el Anexo 1.

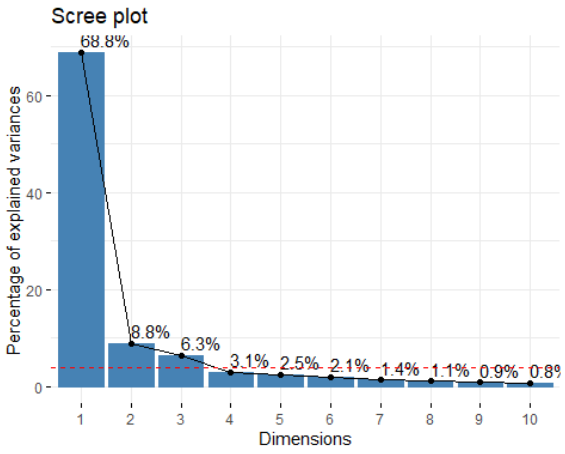


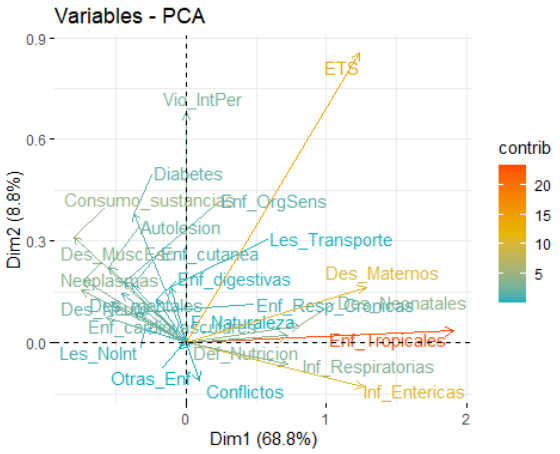
No escalamos las variables porque todas las variables numéricas están en las mismas unidades, y no queremos perder las diferencias de variabilidad entre variables.

En la base de DALY no hay datos faltantes. En la base de datos del Índice de Desarrollo Humano que hemos usado para el PLS sí que hay datos faltantes. Cómo la tratamos para cruzarla con la base de datos de DALY, la añadimos antes de realizar el PLS y eliminamos todas las observaciones con datos faltantes.

3. Análisis PCA

Iniciamos el Análisis de Componentes Principales, tomando a las tres primeras variables (Código, País y Año) como suplementarias para el modelo.





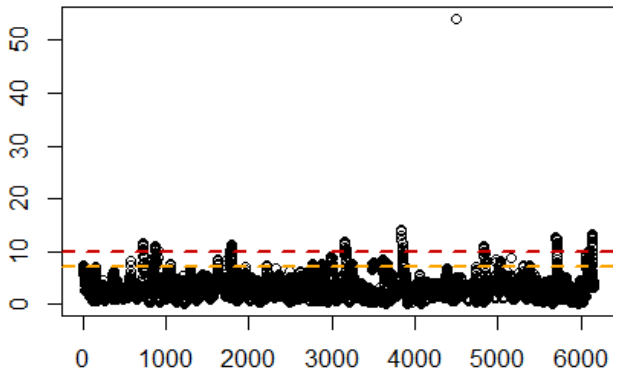
Según el criterio del codo y de la media, escogemos tres componentes principales para explicar la variabilidad de nuestro modelo.

La primera dimensión contribuye considerablemente más que el resto de dimensiones. A grandes rasgos, vemos que la primera dimensión está influenciada principalmente por enfermedades como las ETS, Desordenes Maternos, Infecciones Entéricas y Enfermedades Tropicales.

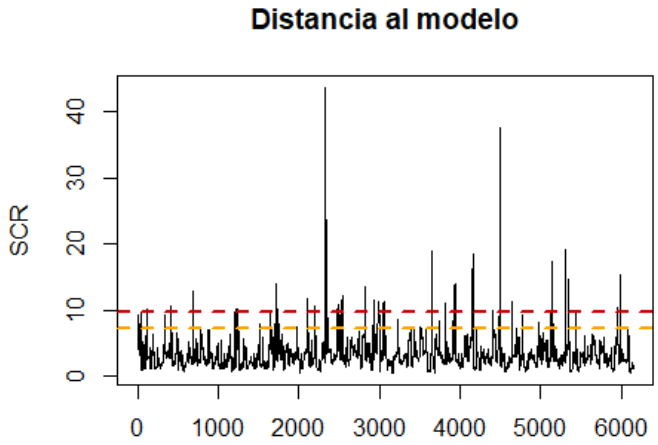
En cambio, otras variables como el consumo de sustancias y los desórdenes neuronales toman valores negativos en dichas componentes.

Antes de realizar el análisis discriminante o el PLS, podemos suponer que esta componente diferencia entre enfermedades más propias de países desarrollados o subdesarrollados.

El gráfico de scores [Anexo 2] no aporta información relevante dado el elevado número de observaciones.



Vemos que la observación 4505 es un outlier severo debido a su valor en la T2 de Hotelling. Eso quiere decir que el modelo está siendo modifcado significativamente para poder incluir a está observación. Revisando en nuestra base de datos apreciamos que está observación pertenece a Ruanda en el año 1994. Por lo tanto está observación se ve afectada por el genocidio de Ruanda, una masacre que aniquiló al 70% de la etnia tutsi y que no tiene comparación con ningún hecho ocurrido en los 30 años que abarca el dataset. Por lo tanto, consideramos que es más sensato eliminar está observación del modelo y volver a realizar el PCA.



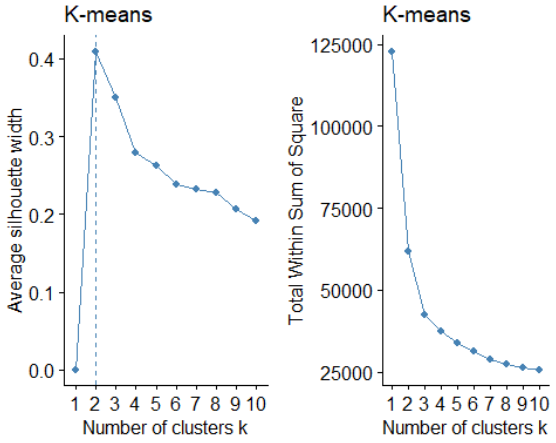
Observamos que la observación 2331 es un outlier moderado, es decir, que no es explicada por el modelo. Esta observación corresponde al año 2010 en Haití, donde hubo un terremoto que supuso una gran tragedia humana en el país. Cómo no ha habido sucesos similares en los últimos 30 años esta observación no es explicada adecuadamente por el modelo. Al no ser un outlier severo no es necesario eliminarla.

Cómo se puede comprobar en el Anexo 2, el modelo PCA generado sin el valor anómalo es idéntico al anterior.

4. Análisis Clustering

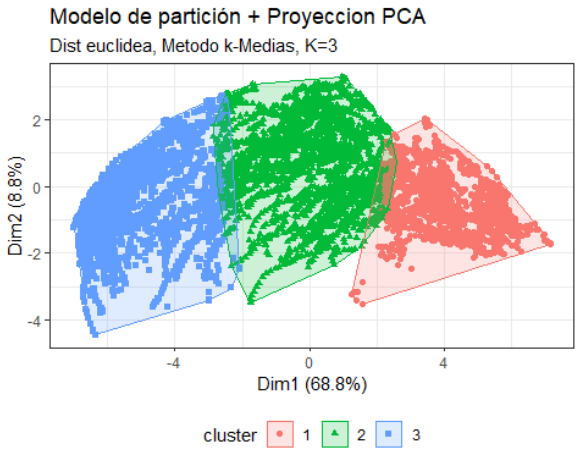
Para realizar el clustering hemos realizado diferentes modelos [Anexo 2], pero en la memoria únicamente mostraremos el obtenido con el método de k-medias. Para llegar hasta aquí, primero hemos tenido que calcular la matriz de distancias mediante la distancia euclídea, ya que nos interesa qué países están más cercanos en cuanto a su valor, porque para nuestro proyecto es más relevante agrupar observaciones con valores similares de DALY (Disability-Adjusted Life Years), es decir, con una incidencia de las causas de mortalidad parecida, en lugar de países que sigan la misma variación entre causas.

Gráfico de Silhouette y de suma de cuadrados intra-cluster



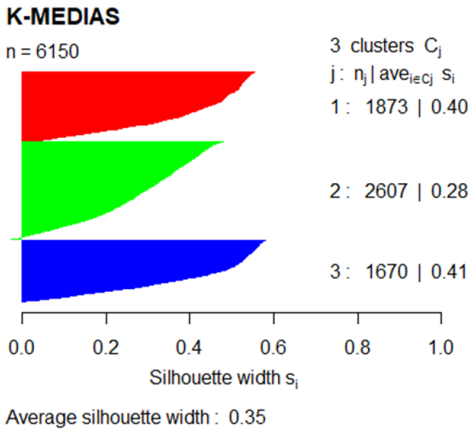
Tras observar Silhouette, vemos que el valor medio más alto se obtiene con 2 clusters, pero tras observar la suma de cuadrados intra-cluster vemos que este presenta un valor todavía muy alto, por lo que consideramos adecuado descartar esta cantidad. Decidimos escoger 3, puesto que aunque tenga un valor de Silhouette menor, es bastante similar al valor máximo, y además su suma de cuadrados intra-cluster es muy inferior y más adecuada.

Ahora describiremos, un **gráfico de scores** para observar cómo se distribuyen cada uno de los clusters:



Tras analizar el gráfico, vemos que el solapamiento es mucho menor en comparación con los métodos anteriores[Anexo 3] y, por tanto, la explicación de los clusters es mejor, puesto que podemos diferenciarlos claramente. En esta visualización, se han identificado tres clusters utilizando el método de k-medias (k-means) con K=3, lo que ha permitido una mejor separación de los datos en el espacio de las componentes principales. Los puntos verdes (cluster 1), los triángulos azules (cluster 2) y los puntos rojos (cluster 3) muestran una mayor separación entre los clusters, lo que indica que este método proporciona una agrupación más clara y precisa. Además, la primera componente principal (Dim1) sigue siendo la que más contribuye a la variabilidad de los datos, explicando el 68.8% de la misma, mientras que la segunda componente principal (Dim2) explica un 8.8%.

Selección y validación del modelo

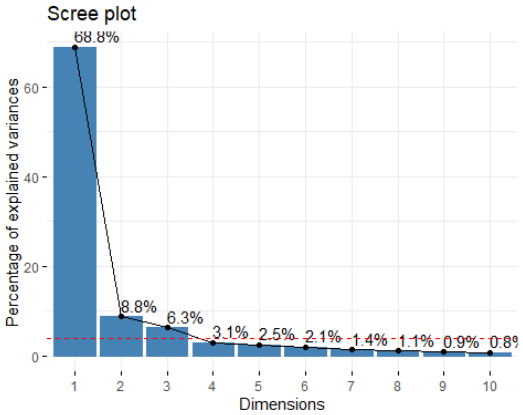


Descartamos el método de Ward y el de la media por el elevado número de observaciones con valores negativos en el estadístico de Silhouette, es decir, observaciones clasificadas en cluster erróneos. Entre los métodos de partición de k-medias y k-medoides las diferencias son prácticamente inexistentes. Nos decantamos por el método de k-medias porque tiene un valor medio del estadístico de Silhoutte ligeramente superior.[Anexo 3]

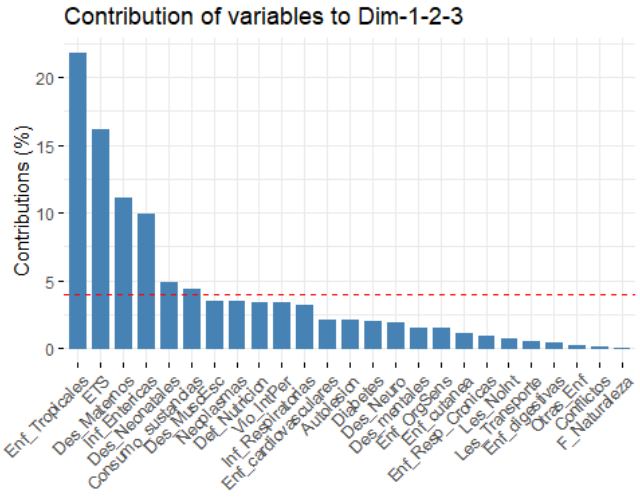
Interpretación de los resultados obtenidos mediante PCA

Vamos a realizar un PCA para ver qué variables han contribuido más a la determinación de clusters con el algoritmo de k-medias.

Realizamos el PCA y obtenemos que la primera componente explica casi el 70% de los datos, por lo que tendrá una gran importancia:

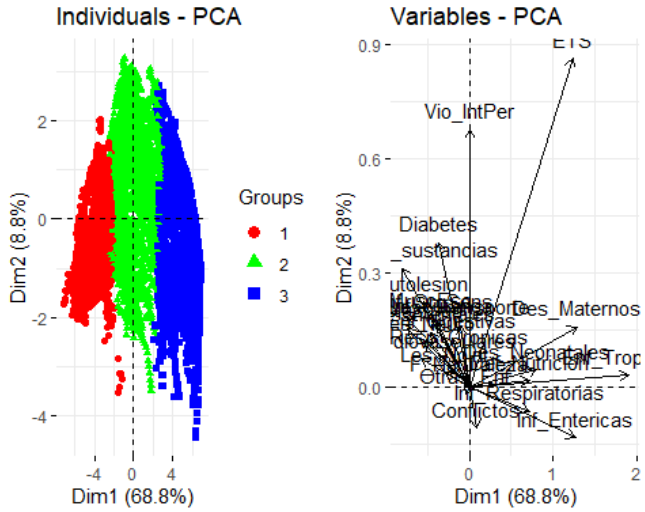


Graficamos las **contribuciones de las variables** para las diferentes dimensiones



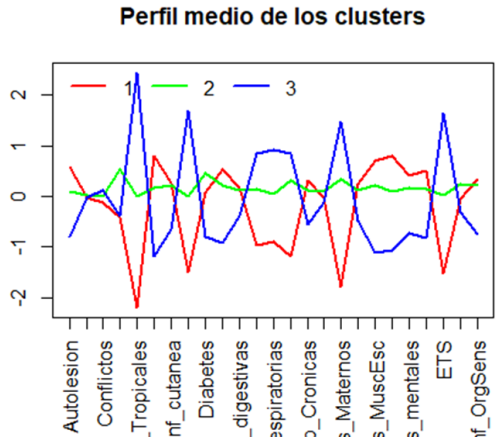
El gráfico muestra la contribución de diversas variables a las dimensiones 1, 2 y 3 en un análisis de componentes principales (PCA). Las barras representan el porcentaje de contribución de cada variable, con las "Enf_Tropicales" (enfermedades tropicales) siendo la que más contribuye, con más del 20%. La línea roja punteada indica el umbral de contribución promedio esperada (alrededor del 5%), destacando las variables que superan esta contribución promedio. Las variables más significativas después de "Enf_Tropicales" son "Des_Nutricio" (desnutrición) y "Enf_ETS" (enfermedades de transmisión sexual), entre otras. Esto sugiere que estas variables son las más influyentes en las primeras tres dimensiones del análisis.

El hecho de que estas enfermedades o estos problemas sean tan importantes para la creación de las dimensiones, se puede deber a su gran variabilidad y a la diferencia de incidencia entre países, estas suelen tener mayor importancia en países subdesarrollados, produciendo graves problemas sobre estos y una gran cantidad de muertes, y por otra parte en los países más desarrollados se cuenta con protección ante estos, produciendo esa variabilidad.



Tras observar el gráfico de scores, vemos cómo es posible diferenciar los clusters únicamente con la primera dimesión, ya que aunque se produzcan algunos solapamientos la diferenciación entre ellos es muy buena. Tras obervar el gráfico de las variables, observamos como la primera dimensión está explicada por 'Enf_Tropicales', 'Inf_Entericas','Des_Maternos' y 'Des_Neonatales' entre otras, por lo que la primera dimensión representa enfermedades relacionadas con países subdesarrollados, ya que son producto de infecciones y de falta de alimentación , y por tanto son variables que tienen una gran variabildad entre los diferentes pasies.. Por otra parte, en el caso de la segunda dimensión está explicada por 'ETS' y 'Vio_IntPer'.

Por último realizaremos un gráfico de perfiles medios para observar qué aspecto tiene cada cluster:



El gráfico muestra los perfiles medios de tres clusters identificados en un análisis de clustering. Cada línea representa un cluster diferente: el cluster 1 en rojo, el cluster 2 en verde y el cluster 3 en azul. El eje horizontal representa diferentes variables o factores (como “Autolesion,” “Conflitos,” “Enf_Tropicales,” etc.), mientras que el eje vertical representa la media de las puntuaciones normalizadas de estas variables dentro de cada cluster.

Interpretación:

El primer grupo toma valores bajos en la mayoría de variables, exceptuando algunas como las autolesiones y el consumo de sustancias. El segundo grupo toma valores cercanos a la medida en todas las variables. El tercer grupo, toma valores elevados en casi todas las variables, destacando las enfermedades tropicales, las infecciones entéricas y las ETS entre otras. Podemos suponer que los grupos están asociados respectivamente a los países desarrollados, en vías de desarrollo y subdesarrollados.

5. Análisis Discriminante (método opcional)

Vamos a realizar un análisis discriminante de nuestros datos. Para ello, primero llevaremos a cabo un pequeño tratamiento de la base de datos (BBDD) para adecuarla y poder analizarla correctamente. Dado que nuestra base de datos contiene datos continuos, vamos a hacer uso de una nueva columna llamada "cluster" para poder realizar el análisis discriminante. En esta columna, clasificamos los diferentes individuos en tres grupos distintos, como lo hicimos en el análisis de clustering anterior. Ahora, utilizaremos esta clasificación para intentar hacer predicciones utilizando el resto de las variables.

Creamos una variable que contenga los nombres de columnas que nosotros queremos para poder entender adecuadamente los resultados que obtendremos posteriormente.

Tablas de frecuencias

Como vamos a hacer uso de la variable “Cluster” como variable respuesta, queremos conocer cómo se distribuye esta en frecuencia y porcentajes.

Var1	Freq

Var1	Freq

1	1873
2	2606
3	1673

1	30.46024
2	42.38087
3	27.15889

Como podemos observar, las clases no están perfectamente equilibradas, ya que para ello cada una tendría que poseer el 33.33% de las observaciones totales de la base, pero el desequilibrio es muy ligero por lo que podemos hacer uso de esta perfectamente.

Modelos

Ahora, dividiremos la base de datos en dos partes: la primera se llamará 'train_daly' y contendrá el 80% de los datos, mientras que la segunda se llamará 'test_daly' y contendrá el 20% restante. Utilizaremos el primer data frame para entrenar el modelo que desarrollaremos a continuación y el segundo para probarlo. Aunque la cantidad de datos de cada uno será diferente, ya que como hemos dicho antes tienen distinto tamaño, la distribución es la misma, es decir contendrán el mismo porcentaje de observaciones de cada cluster.

Ahora separamos lo que hemos explicado anteriormente, es decir el dataframe de 'train' y el de 'test'

Para comprobar que se ha realizado correctamente la separación, volveremos a crear las tablas de frecuencias. Tras su creación, podemos observar que está creada correctamente y que esta tiene la misma distribución que la originales.

Para poder ejecutar correctamente el modelo, tenemos que eliminar aquellas variables que sean de tipo texto, es decir, en este caso eliminaremos "Pais" ,"Codigo" y "Año", ya que no nos son útiles para la clasificación:

A continuación, crearemos una nueva variable asociada a "train_daly", que llamaremos "train_dalyEsc". Esta contendrá los datos escalados, excepto los de la variable "Cluster", ya que no tendría sentido escalar la variable que se utilizará para realizar la clasificación. En el mismo cuadro de código generamos también el modelo lineal discriminante sobre los datos de entrenamiento y evaluamos su bondad de clasificación sobre estos datos y sobre los datos test. En esta ocasión, utilizaremos la función lda directamente para generar el modelo, en lugar de la librería caret, y no realizaremos validación cruzada sobre los datos de entrenamiento.

```
## Coefficients of linear discriminants
##                              LD1       LD2
## Autolesion          -0.083780906  0.054721191
## F_Naturaleza         0.071296362 -0.038402431
## Conflictos          -0.023403870 -0.103770061
## Vio_IntPer          -0.091537789 -0.327257100
## Enf_Tropicales       0.388737111  0.268317680
## Consumo_sustancias   0.093322728 -0.019443772
## Enf_cutanea          0.262013545 -0.667767249
## Inf_Entericas        0.199731172  0.276700406
## Diabetes             0.187990875 -0.824253581
## Enf_cardiovasculares -0.319134936 -0.565912974
## Enf_digestivas       -0.149011982 -0.104298367
## Def_Nutricion        -0.060595036 -0.165310408
## Inf_Respiratorias    0.338172279 -0.404034616
## Des_Neonatales       0.003419465 -0.952211608
## Enf_Resp_Cronicas    0.118820079  0.207559965
## Otras_Enf            0.202407376 -0.000430931
```

```
## Des_Maternos          0.952747903  0.007973042
## Les_NoInt            -0.201500647 -0.130971564
## Des_MuscEsc          -1.218953081  0.082101795
## Neoplasmas           -0.198265341  0.570629542
## Des_mentales         -0.306697680  0.371536589
## Des_Neuro             0.275306623 -0.261752968
## ETS                   0.481298539 -0.090868065
## Les_Transporte        0.026914698 -0.196363068
## Enf_OrgSens           0.205500514 -0.269989185
##
## Proportion of trace:
##     LD1    LD2
## 0.8259 0.1741
```

Tras observar los resultados obtenidos en las anteriores salidas, vemos que la traza de la primera función discriminante es la más significativa, puesto que es la que mayor valor tienen de las dos (LD1=0.8237 > LD2=0.1763), por lo que seguramente será muy importante a la hora de separar los grupos en la clasificación, a continuación realizaremos un mapa que evidencia esto. Por otra parte cabe destacar que Desorden Muscular-Esquelético, la variable que tiene más influencia para clasificar una observación en un cluster, ya que su valor es el más alto, pero este es negativo(-1.39), cosa que indica que la variable es crucial para la discriminación entre los grupos y que existe una relación inversa entre la variable y la función discriminante.

Tras realizar la función discriminante para los datos de “train_dalyEsc” y observar sus resultados, ahora crearemos matrices de confusiones para observar cuáles son los resultados de los índices y cómo de bien clasifican los datos estos, para ello primero realizaremos el tratamiento que hemos hecho antes con “train_daly”, pero para “test_daly”

Creación de la matriz de confusión para los datos obtenidos con el primer modelo, es decir, los datos de “train_daly”:

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    1 2          3
##         1 1414   25         0
##         2   85 2040   58
##         3          0   20 1278
##
## Overall Statistics
##
##               Accuracy : 0.9618
##                 95% CI : (0.956, 0.967)
##         No Information Rate : 0.4238
##         P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 0.9413
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity                0.9433   0.9784   0.9566
## Specificity                0.9927   0.9496   0.9944
## Pos Pred Value             0.9826   0.9345   0.9846
## Neg Pred Value             0.9756   0.9836   0.9840
## Prevalence                 0.3047   0.4238   0.2715
```

```
## Detection Rate          0.2874   0.4146   0.2598
## Detection Prevalence    0.2925   0.4437   0.2638
## Balanced Accuracy       0.9680   0.9640   0.9755
```

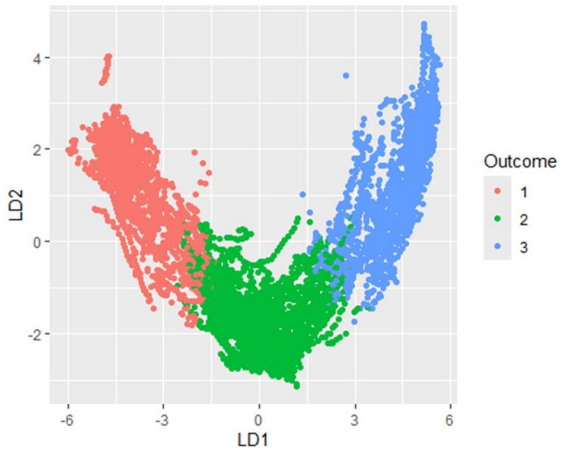
Vemos que los valores obtenidos en los índices para las tres clases son muy altos, por encima del 0.95 en sensibilidad y especifidad, además la aproximación es del 0.9624 por lo que sí que podemos probarlo con los datos de tes, aunque hayan algunas clasificaciones erróneas.

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction   1   2   3
##         1 352   1   0
##         2  22 516  11
##         3   0   4 323
##
## Overall Statistics
##
##               Accuracy : 0.9691
##                 95% CI : (0.9578, 0.978)
##         No Information Rate : 0.4239
##         P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 0.9525
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity             0.9412   0.9904   0.9671
## Specificity             0.9988   0.9534   0.9955
## Pos Pred Value          0.9972   0.9399   0.9878
## Neg Pred Value          0.9749   0.9926   0.9878
## Prevalence              0.3043   0.4239   0.2718
## Detection Rate          0.2864   0.4199   0.2628
## Detection Prevalence    0.2872   0.4467   0.2661
## Balanced Accuracy       0.9700   0.9719   0.9813
```

Tras realizarlo en los datos de test los resultados son muy similares a los de train, por lo que la aproximación es bastante buena, con una valor de accuracy del 0.9569, lo que significa que las malas clasificaciones son muy bajas.

Representación gráfica

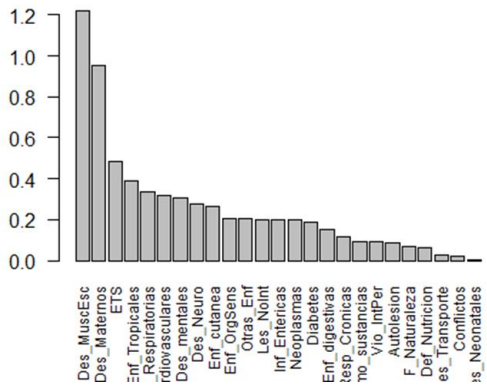
Ahora representaremos gráficamente las puntuaciones discriminantes (ahora en un gráfico de dos dimensiones) para todos los datos, para poder observar los resultados de una formas más visual:



Observamos cómo la primera función discriminante separa muy bien los tres clusters, no de manera perfecta, pero sí con una alta precisión tal y como hemos visto anterirormente. Por otra parte LD2 no se considera necesaria, ya que en ella se solapan los tres cluseters y la representación no es muy buena. Así pues, en este ejemplo, nos quedaríamos únicamente con la función LD1.

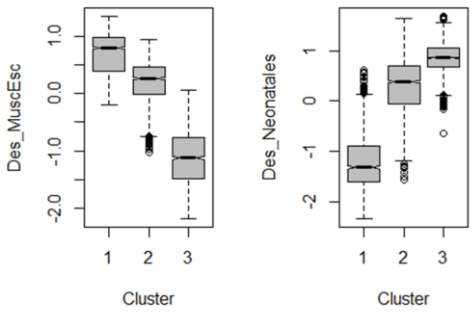
Contribuciones

Variables que más han contribuido a clasificar:



En el gráfico superior, hemos representado la contribución de las variables en la función discriminante 1. En este podemos observar que hay dos variables muy superiores al resto en cuanto a la aportación sobre la primera función discriminante: Des_MuscEsc y Des_Maternos. El hecho de que estas variables tengan un valor tan superior al del resto de variables en cuanto a la separación de los grupos en la primera función discriminante nos podría indicar que son dos variables a tener muy en cuenta. Estas variables son cruciales para la diferenciación de los grupos, sugiriendo que las diferencias en desórdenes musculoesqueléticos y desórdenes maternos son las más determinantes para separar los grupos en este análisis. Por lo tanto, deben ser consideradas prioritarias tanto en la interpretación de los resultados como en la toma de decisiones estratégicas en cuanto a tomar acción en estos ámbitos.

Realizamos boxplot de las variables que mejor y peor clasifican, para observar su valor en cada uno de los clusters.



Tras observar los gráficos vemos que los box-plots están bien diferenciados en cada uno de los clusters, por lo que ambas variables(`Des_MuscEsc',‘Des_Neonatales') son de utilidad para diferenciar los grupos. Pese a que ambas son buenas en cuanto a la separación de grupos, la que más importancia y más aporta en este aspecto es la primera tal y como hemos podido ver antes.

6. Análisis PLS

El objetivo de este modelo PLS es predecir el nivel de desarrollo según el HDI a partir de los valores de DALY de distintas causas de mortalidad o enfermedades. Puesto que el objetivo es discriminar qué observaciones tienen mayores probabilidades de estar en un determinado nivel, realizamos un modelo PLS-DA o discriminante.

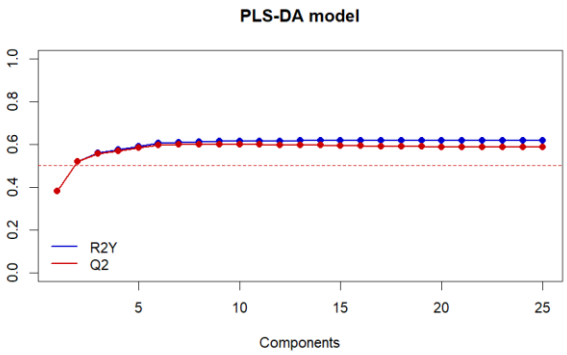
Dado que solo tenemos una variable respuesta, esta no puede tener valores nulos, por lo que eliminamos todas las observaciones con valores faltantes en la variable respuesta.

Antes de realizar el análisis, definimos los niveles de desarrollo que vamos a tener en consideración.

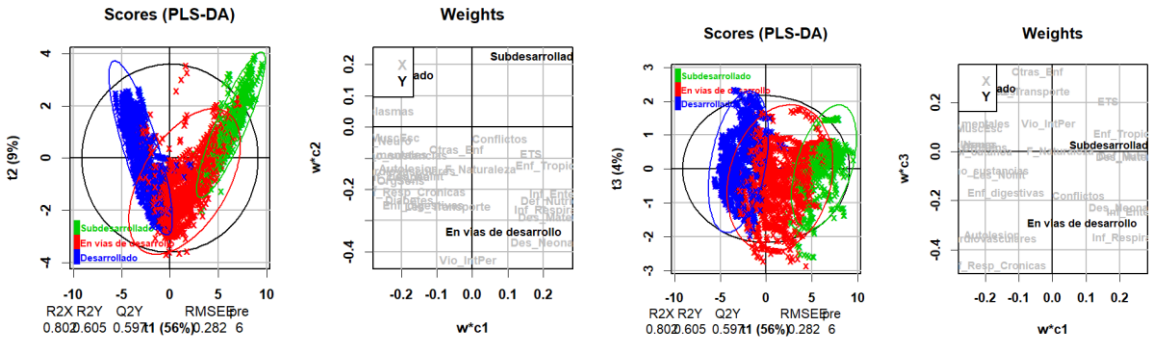
- 0-0.45: Subdesarrollado
- 0.45-0.7:En vías de desarrollo
- 0.7-1: Desarrollado

Realizamos un modelo PLS-DA con validación cruzada 100-fold con los siguientes resultados:

```
## PLS-DA
## 3796 samples x 25 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort   pR2Y    pQ2
## Total    0.802    0.605   0.597 0.282   6   0 0.0333 0.0333
```



Apreciamos que con 6 componentes el modelo se ajusta bien al conjunto de variables predictoras, pero no consigue valores completamente satisfactorios de bondad de ajuste y predicción para la variable respuesta. Como podemos ver, modificando el número de componentes no obtenemos cambios significativos.



Estudiando los gráficos de scores y weights, apreciamos que el modelo es capaz de discriminar los grupos para la mayoría de observaciones, pero las zonas comunes compartidas entre países en vías de desarrollo con países desarrollados o subdesarrollados son demasiado amplias.

Esto provoca que si una observación cae en esas zonas comunes, el modelo tenga problemas para predecir a que grupo pertenece.

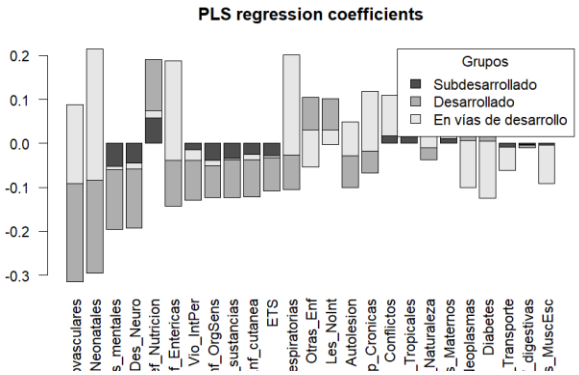
Teniendo en cuenta que un país puede cambiar su categoría modificando una décima su Índice de Desarrollo Humano, comprendemos que el modelo tenga estos problemas en los límites entre grupos puesto que el desarrollo no es una variable discreta sino continua.

Del mismo modo, el modelo consigue diferenciar perfectamente las observaciones de países subdesarrollados de las de países desarrollados.

Analizando el gráfico de weights vemos que los países desarrollados están correlacionados positivamente con un grupo de afecciones y correlacionados negativamente con el grupo de afecciones asociado a los países subdesarrollados.

Por otro lado, los países en vías de desarrollo están correlacionados con un abanico de afecciones que comprende algunas propias de los países desarrollados y otras propias de los subdesarrollados; por lo que se entiende que comparte zonas comunes con estos grupos.

Observamos que la primera y la segunda componente son las relevantes a la hora de discriminar los distintos grupos.



Estos son los coeficientes de regresión PLS para clasificar cada grupo propuesto. Comparando este gráfico con los gráficos de weights, podemos discernir qué variables y de qué manera están relacionados con cada grupo de la variable respuesta.

```
## Confusion Matrix and Statistics
##
                          Reference

  Prediction              Subdesarrollado En vías de desarrollo Desarrollado
```

```
  Subdesarrollado                360                110            0
  En vías de desarrollo          144                1179           179
  Desarrollado                     0                200            1624

##
               Accuracy : 0.8332
                  95% CI : (0.821, 0.845)
    No Information Rate : 0.475
    P-Value [Acc > NIR] : < 2.2e-16


                  Kappa : 0.7221
```

La matriz de confusión muestra que el modelo clasifica con bastante precisión las observaciones pertenecientes a los tres grupos, pero suele cometer errores con los grupos limítrofes.

Los valores de Accuracy y Kappa no son insatisfactorios una vez hemos comprendido la naturaleza de la base de datos y del modelo.

En conclusión; el modelo de PLS-DA planteado sirve para diferenciar a los países según su nivel de desarrollo, pero hay que tener en cuenta que puede fallar si el país está cerca de cambiar su clasificación.

7. Conclusiones

Gracias a los resultados obtenidos mediante el Análisis de Componente Principales, hemos visto que existe un cierto grupo de enfermedades que modelan el espacio latente, es decir, que son las que más variabilidad explican del modelo. Por ende, estas variables toman valores diferentes dependiendo del país y el año de la observación. En consecuencia, los países se pueden agrupar según los valores que tomen en dichas componentes, mostrando que las variables con más peso en las componentes tienen relación con el tipo de país. Esta información nos da pie para continuar con las siguientes técnicas estadísticas.

Los resultados del clustering nos muestran que hay tres grupos diferenciados de países. Analizando los perfiles de los clusters hemos llegado a las siguientes conclusiones. El primer grupo toma valores bajos en la mayoría de variables, exceptuando algunas como las autolesiones y el consumo de sustancias. El segundo grupo toma valores cercanos a la medida en todas las variables. El tercer grupo, toma valores elevados en casi todas las variables, destacando las enfermedades tropicales, las infecciones entéricas y las ETS entre otras. Podemos suponer que los grupos están asociados respectivamente a los países desarrollados, en vías de desarrollo y subdesarrollados.

Para comprobar la calidad de este análisis hemos efectuado un análisis discriminante tomando los clusters como variable respuesta. El modelo conseguía discriminarlos correctamente con un error muy pequeño, por lo que queda demostrado las diferencias entre los tres grupos definidos.

Por último, comprobamos si la hipótesis que hemos planteado sobre el desarrollo es correcta. Para ello introducimos el IDH en nuestro análisis, y según los valores del índice clasificamos a cada país. Posteriormente realizamos un modelo PLS con la variable del nivel de desarrollo como variable respuesta. El modelo resultante consigue discriminar correctamente los tres grupos, mostrando para cada uno coeficientes de regresión similares a los perfiles obtenidos en el clustering. En cuánto a la predicción, el modelo suele acertar la mayoría de veces, únicamente teniendo problemas cuando se trata de países que están cerca de cambiar de nivel de desarrollo.

8. ANEXOS

 # Anexos Rmarkdown

1. Anexo 1: [codigo_trabajo_sin_pca.Rmd: análisis exploratorio de la base de datos]

2. Anexo 2: [mdp_pca_comentado.Rmd: análisis PCA]

3. Anexo 3: [mdp_clust_def.Rmd: aplicación de clustering a la base de datos]

4. Anexo 4: [mdp_disc_def.Rmd: aplicación del análisis discriminante a la base de datos]

5. Anexo 5: [mdp_plsda.Rmd: aplicación del pls a la base de datos]

6. Anexo 6: [https://www.kaggle.com/datasets/shivkumarganesh/disease-burden-by-cause]