Universidad de Granada

Doble Grado en Física y Matemáticas

AED en variables sobre Estados.

Aarón Benitez Barón

Índice

Abstract

En este informe se recoge el estudio estadístico de distintas variables demográficas recogidas sobre países del mundo. Estudiaremos supuestos de normalidad de las variables con el objetivo de poder aplicar técnicas de reducción de la dimensión a estos datos. Por último intentaremos obtener un modelo mediante análisis discriminante si un país esta desarrollado o no.

1. Introducción

La reducción de dimensión es una técnica muy útil en el campo del machine learning ya que este permite reducir el tamaño de memoria que consumen los datos, el tiempo operacional y la mejora de la visualización de las variables. En esta práctica aplicaremos distintos métodos como el Análisis de Componentes Principales o el Análisis Factorial para llevar a cabo esta técnica. Después de ello intentaremos aplicar una técnica muy conocida en el reconocimiento de patrones como es el Análisis de Discriminante para discernir propiedades de la base de datos elegida. En nuestro caso, estudiaremos datos demográficos sobre 34 países distintos que nos darán información sobre la situación de desarrollo del Estado. Dichas variables las expondremos en el siguiente apartado. El objetivo será tratar la base de datos que nos den, tratar con sus ouliers, valores perdidos, etc. y a partir de ahí, estudiar tanto el comportamiento de las variables por separado como sus correlaciones que nos permitan aplicar las técnica del análisis estadístico multivariante.

2. Materiales y métodos

2.1. Materiales

En este apartado, presentaremos la base de datos a estudiar. Estudiaremos 11 variables distintas que serán:

- Densidad de población.
- Población en núcleos urbanos.
- Población trabajando en el sector agrícola.
- Población trabajando en el sector servicio.
- Coeficiente entre población activa y total.
- Esperanza de vida.
- Tasa de mortalidad infantil.
- Tasa entre el número de individuos en ejército de tierra y población total.

- Tasa de médicos por habitante.
- Número de libro publicados.
- Tasa de consumo energéico.

Todas estas variables están normalizadas debido a las distintas unidades que puedan tener. A continuación, presentamos una tabla que expone los diferentes estadísticos más relevantes sobre estas variables:

	Centrimedia	CVc
Densidad población	-0.2293829	-4.7611940
Población urbana	0.1147624	18.6903015
Población agricola	-0.20433276	-11.4243309
Población servicio	0.00064957	12.8296875
Población activa	0.02917055	7.5584589
Esperanza de vida	0.17613181	10.4877506
Mortalidad infantil	-0.2268480	-30.8993711
Tasa de militares	-0.2192510	-1.3081800
Tasa de médicos	-0.149734266	-130.4526316
Libros publicados	-0.2615358	-7.2311230
Consumo energético	-0.26271176	-6.871948

Cuadro 1: Valores de centralidad y coeficientes de variación cuartílica para las distintas variables.
He decidido usar estas medidas de centralidad y dispersión porque son mucho más significativas para estas variables que las usuales media y desviación típica.

2.2. Métodos

En esta sección presentaremos las técnicas que vamos a utilizar para estudiar estas variables:

- En primer lugar, estudiaremos la normalidad univariante de cada variable a través de gráficos qq-plot.
- A continuación, comprobaremos que realmente existen correlaciones entre las variables con un test de Barlett.
- Aplicaremos el Análisis de Componentes principales, con diferentes métodos para decidir cuantas componentes tomar.
- También aplicaremos un Análisis Factorial mediante el modelo varimax para obtener las variables latentes en los datos.
- Por último aplicaremos un Análisis Discriminante lineal y cuadrático con el cuál seremos capaces de predecir si un país dado es desarrollado o no.

3. Resultados

Una vez tratada la base de datos, eliminando los outliers y los valores perdidos pertinentes, procedemos a desarrollar los resultados.

3.1. Normalidad univariante

Comprobemos en primer lugar si las variables por separado siguen aproximadamente un comportamiento normal. Para ello utilizaremos un qq-plot:

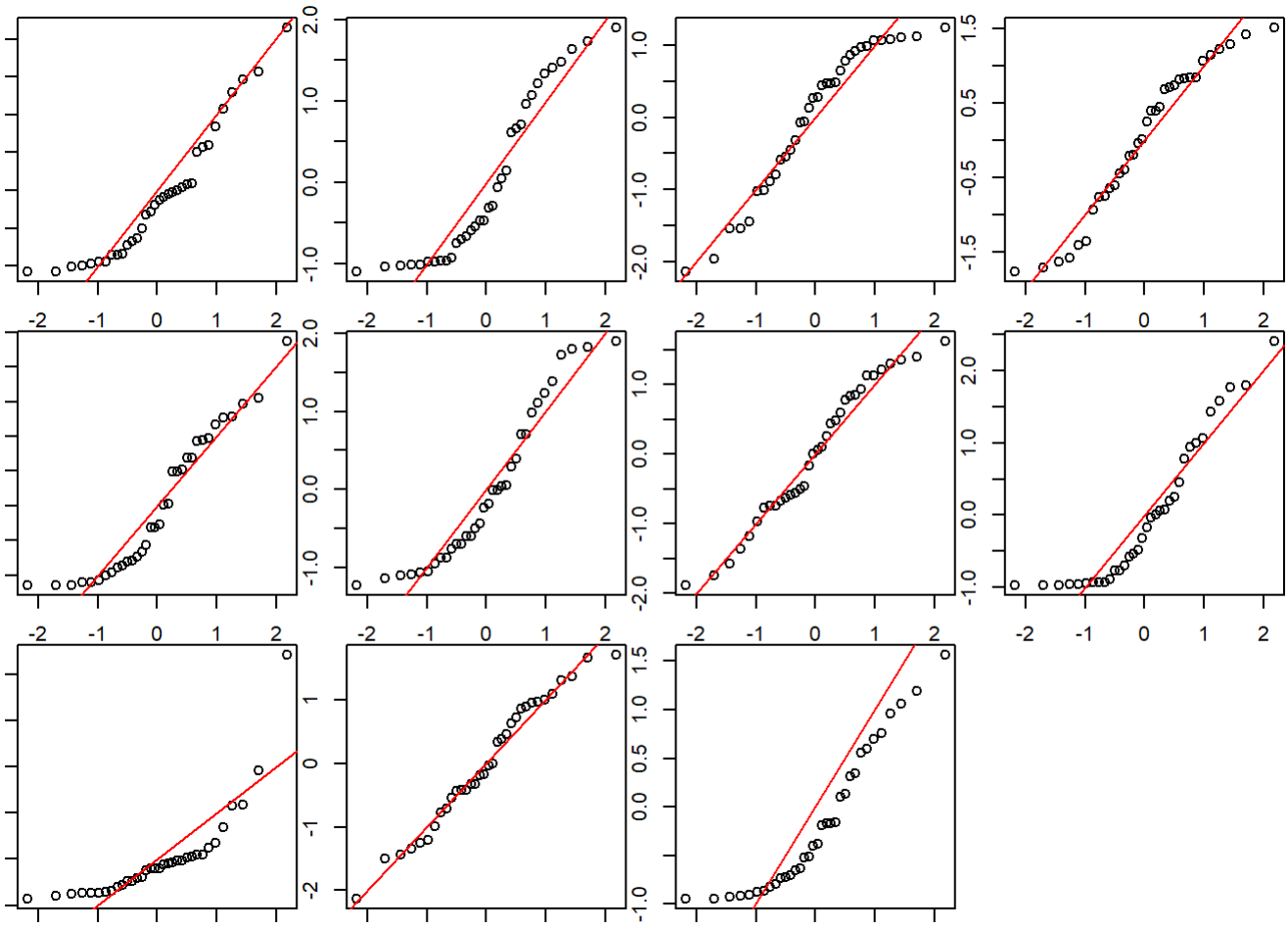

Figura 1: Estudio de la normalidad de las variables.

Observamos que en general, nuestras variables se desvían del comportamiento normal. Sin embargo, como no es una desviación muy extrema (salvo en la variable 9, correspondiente a la tasa de militares) asumiremos normalidad en dichas variables.

3.2. Test de Barlett

Si representamos la matriz de correlaciones de forma gráfica obtenemos el siguiente resultado:

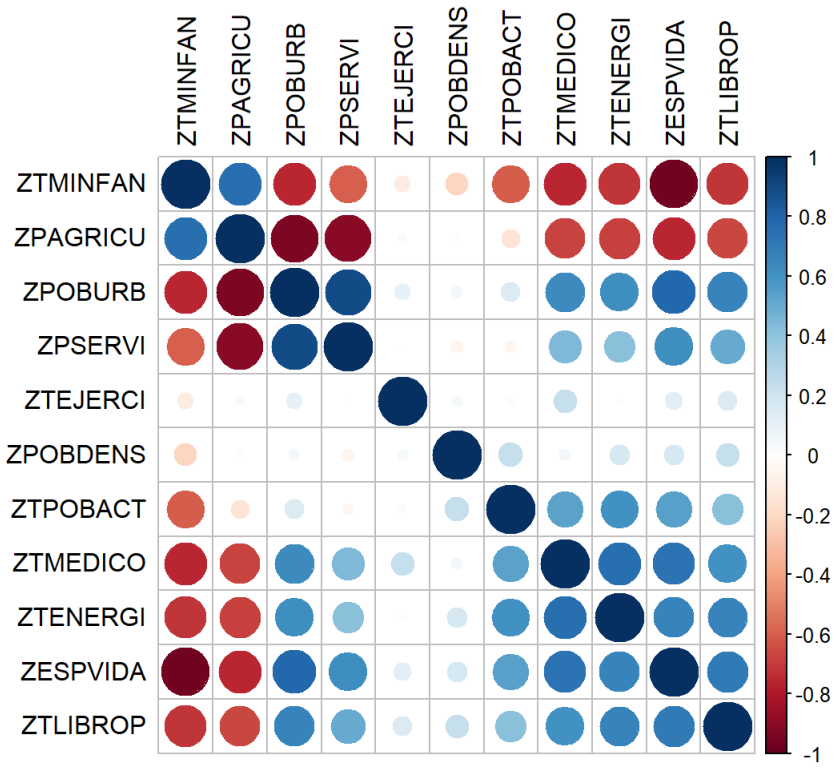

Figura 2: Estudio de la normalidad de las variables.

Observamos que claramente existen correlaciones entre las variables, sin embargo realizamos el test de Barlett para asegurarnos de ello. En el obtuvimos un p-valor muy cercano a cero, por lo que tenemos que rechazar la hipótesis nula H_0 y aseguramos que la matriz de correlaciones no es trivial.

3.3. Análisis de Componentes principales.

A continuación aplicaremos el método de las componentes principales. Para obtener el número de componentes que debemos considerar utilizaremos varios métodos. En primer lugar, presentamos el método del codo:

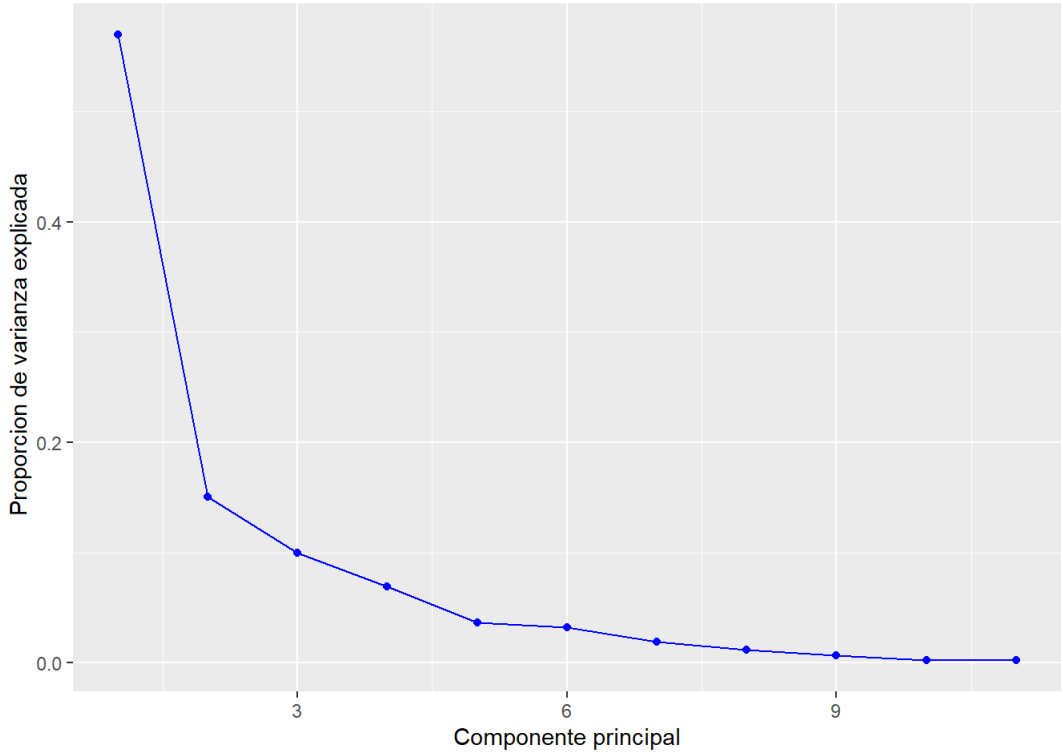

Figura 3: Gráfico de la varianza explicada de cada CP.

Observamos que el codo nos dice que tomemos 2 componentes principales. Sin embargo, los métodos de la varianza acumulada y la Regla de Abdi nos .ªconsejan"que tomemos 3. La decisión final será tomar tres componentes para cubrir más del 80 % de la varianza explicada. Las componentes principales quedan así:

```
##                PC1         PC2         PC3
## ZPOBDENS  0.05570327  0.28133608  0.05364209
## ZTMINFAN -0.38613899 -0.15136518 -0.06347649
## ZESPVIDA  0.38379701  0.10210875  0.02945470
## ZPOBURB   0.37212146 -0.29140632 -0.03828412
## ZTMEDICO  0.34118876  0.17867758 -0.11869379
## ZPAGRICU -0.37407604  0.30485020 -0.02794410
## ZPSERVI   0.31138182 -0.48515645  0.02486393
## ZTLIBROP  0.32856071  0.09495763 -0.04290409
## ZTEJERCI  0.05738214  0.14151643 -0.95932898
## ZTPOBACT  0.20051927  0.62673107  0.20272273
## ZTENERGI  0.23930145  0.14325835  0.10965106
```

Figura 4: Resultado de las componentes principales.

Si representamos las contribuciones de cada variable a las componentes obtenemos los

siguientes gráficos:

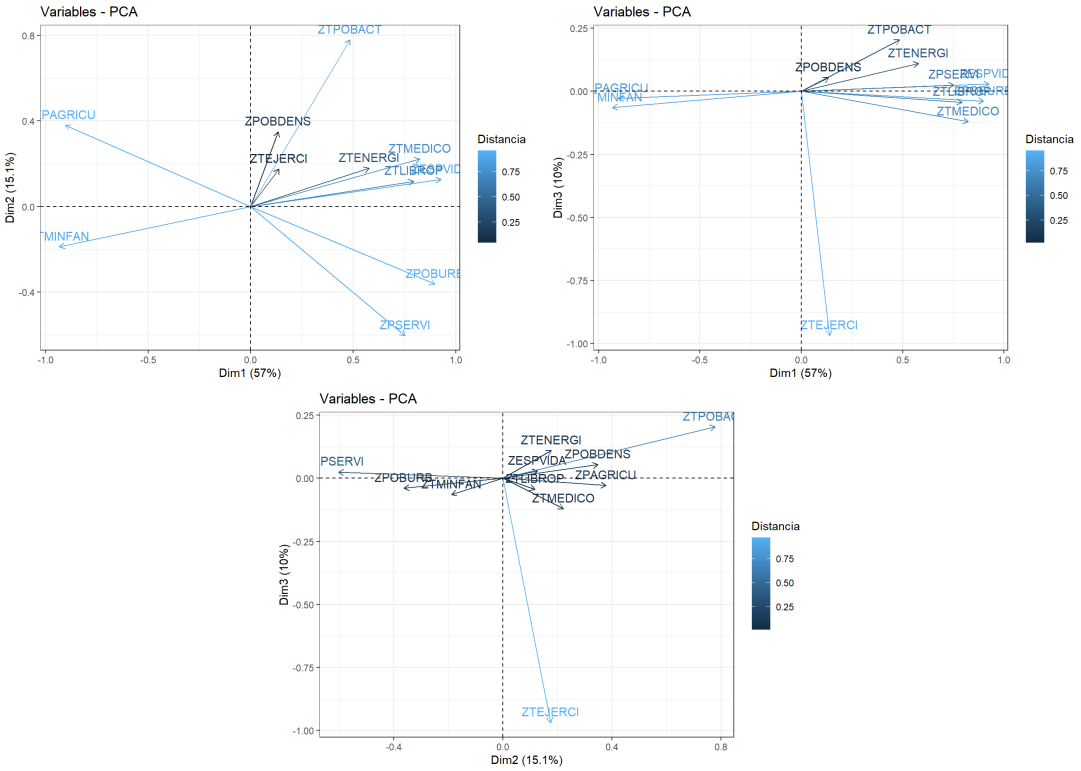

Figura 5: Representación de las contribuciones para cada compoenente principal.

3.4. Análisis Factorial

Antes de aplicar el modelo correspondiente para obtener los factores latentes, debemos discernir cuantos factores vamos a considerar. Para ello utilizaremos el análisis paralelo:

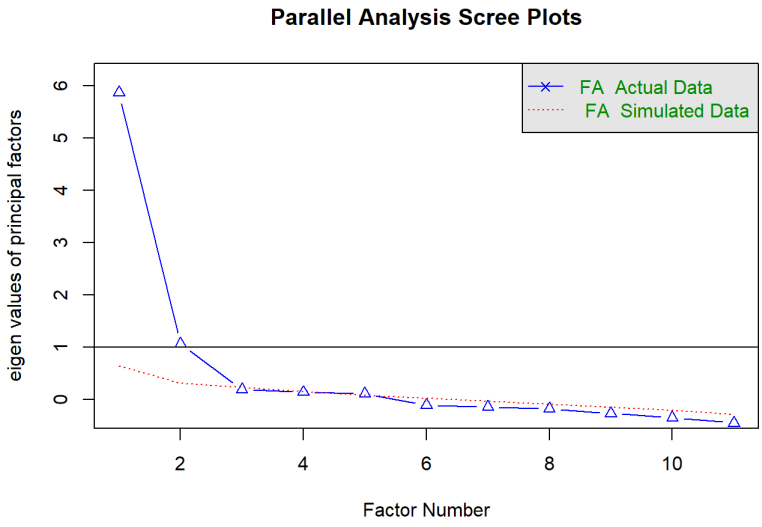

Figura 6: Gráfico del análisis paralelo.

En vistas de este, tomaremos 2 factores latentes. Aplicando el modelo varimax, obtenemos los siguientes factores:

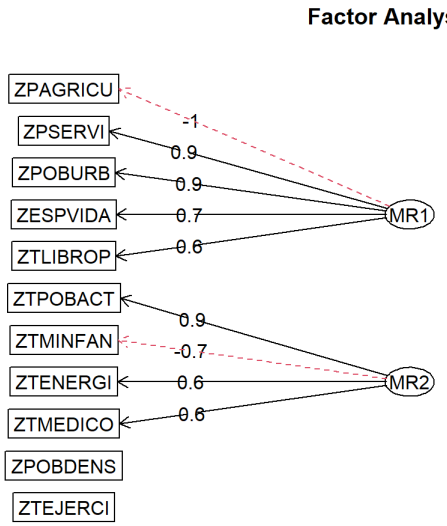

Figura 7: Diagrama de factores latentes.

3.5. Análisis Discriminante

Antes de aplicar este método, estudiaremos el supuesto de normal multivariante a través de distintos test, como el de Royston y el de Henze-Zirkler. Si realizamos, el test con todas las variables, obtenemos que la negación de este hecho. Sin embargo, al eliminar

la variable que recogía la tasa de militares (ya que esta se alejaba demasiado del comportamiento normal) si obtenemos normalidad con el test de Henze-Zirkler. Con esto nos bastará para aplicar el método. El objetivo de este análisis será ser capaces de dilucidar si un país es desarrollado o subdesarrollado según los datos dados. El análisis discriminante lineal nos da casi un 15 % de error al contrastarlo con la matriz de confusión, sin embargo, el cuadrático nos da una tasa de error del 0 % por lo que será este el que realmente tendremos en cuenta. Los resultados de dicho análisis fueron:

```
Prior probabilities of groups:
      des       sub
0.3529412 0.6470588

Group means:
      ZPOBDENS   ZTMINFAN   ZESPVIDA    ZPOBURB   ZTMEDICO   ZPAGRICU
des  0.2356446 -0.9316218  0.9469584  0.8644555  0.6402707 -0.8540436
sub -0.2625487  0.5081573 -0.5165228 -0.4715212 -0.3492386  0.4658420
       ZPSERVI   ZTLIBROP   ZTPOBACT   ZTENERGI
des  0.7781426  0.9582738  0.4761974  0.3566198
sub -0.4244414 -0.5226948 -0.2597440 -0.4557681
```

Figura 8: Resultado del análisis discriminante cuadrático.

Observamos que a priori, la probabilidad de que un pais sea desarrollado es de solamente un 35 %. Si representamos algunas variables que relaciona nuestro análisis, vemos como este consigue clasificar los países como se pretendía;

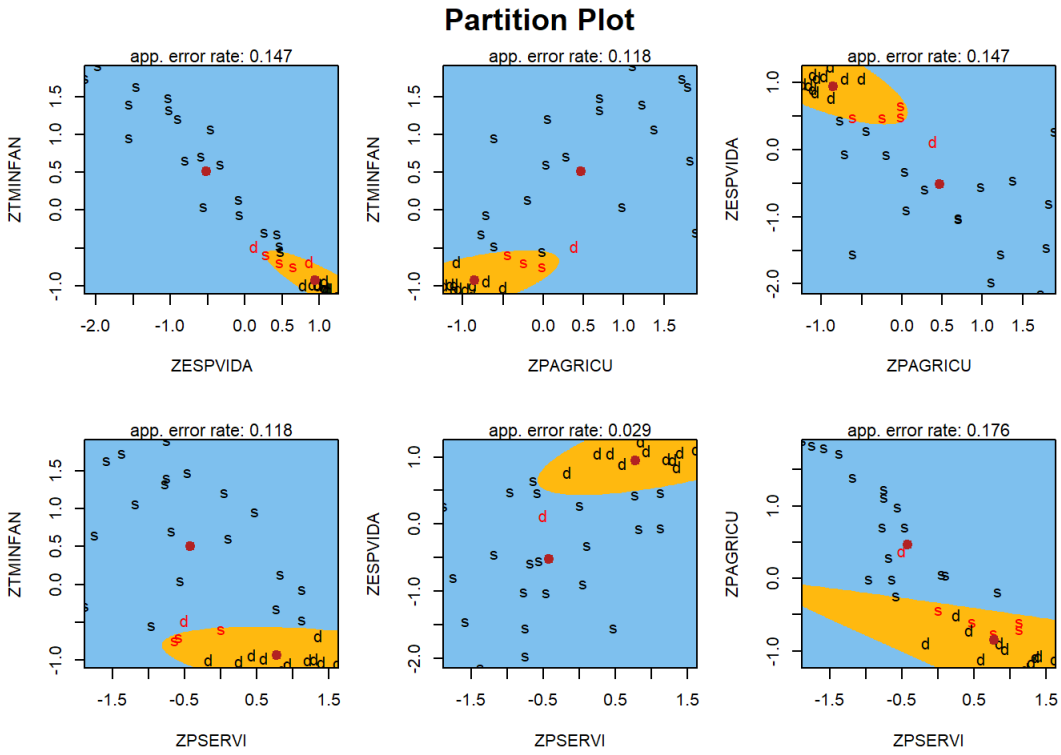

Figura 9: Gráficos de clasificación para el análisis discriminante.

4. Discusión

Vamos a interpretar un poco los resultados obtenidos. Cabe destacar que los supuestos de normalidad, tanto univariante como multivariante no son para nada robustos, por lo que no podemos asegurar con autoridad de que los resultados obtenidos sean completamente correctos. Obviando esto hemos conseguido reducir la dimensión de la muestra como se pretendía con el ACP y el AF. Cabe destacar la interpretación que le podemos dar a los factores latentes obtenidos:

- El MF1 podría darnos un indicador de como de urbanizado esta un país, teniendo en cuenta positivamente las variables de personas trabajando en sector servicio y la población residente en ciudades, además de valorar negativamente la población agrícola. Con menos fuerza también valora positivamente la esperanza de vida y la cantidad de libros publicados.

- El MF2 nos da idea de un desarrollo más general del país, valorando si se dan ciertas necesidades básicas. En este factor se recoge positivamente si hay pocos parados, la tasa de médicos y la energía que se consume. Por último, se valora negativamente la mortalidad infantil en el país. Este indicador podría ser útil para compara países en vías de desarrollo.

Por último, en el análisis Discriminante se han conseguido resultados fructíferos a la hora de hacer una clasificación entre los distintos países, por lo menos si aplicábamos el modelo cuadrático, dando a entender que las relaciones son algo más complejas como para tratarlas únicamente con un modelo lineal.

5. Conclusión

Pese a la limitación de conocimiento por mi parte, creo que se han conseguido resultados interesantes sobre las variables expuestas. Para mejorar el trabajo tendríamos que realmente comprobar si estos métodos son aplicables con los supuestos de normalidad que se han dado. También podríamos completar este informe estudiando muchas más variables demográficas que nos de información más rica para el estudio. En conclusión, los objetivos que se establecían se han cumplido en gran medida por lo que podemos dar por cerrado este trabajo.