Tecnológico de Monterrey

Reporte “Los peces y mercurio”

Materia

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 501)
Módulo 5: Estadística

Autor

Facundo Vecchi - A01283666

27 de noviembre del 2022, Monterrey N.L.

Resumen

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Ya que se cuentan con una gran cantidad de variables, antes de determinar cuales son las que mas influyen sobre el nivel de mercurio en los peces, se realizó un análisis de componentes principales para determinar si se puede reducir la dimensionalidad de los datos. Se encontró que los datos se pueden reducir a 6 componentes explicando el 96% de la variabilidad de los datos.

Introducción

A través del análisis realizo se intenta contestar la siguiente pregunta:

¿Se puede reducir la dimensionalidad de los datos utilizando componentes principales?

Antes de poder determinar cuales son los factores que mas influyen en la concentración de mercurio de los peces de los lagos, tenemos que encontrar si se puede reducir la dimensionalidad de los datos. Esto con el propósito de facilitar la exploración de estos mismos y la elaboración del análisis para encontrar que factores son los más importantes.

Análisis de los resultados

A continuación se demostrará el proceso que se llevó a cabo para llegar a los resultados previamente mencionados.

Análisis de normalidad

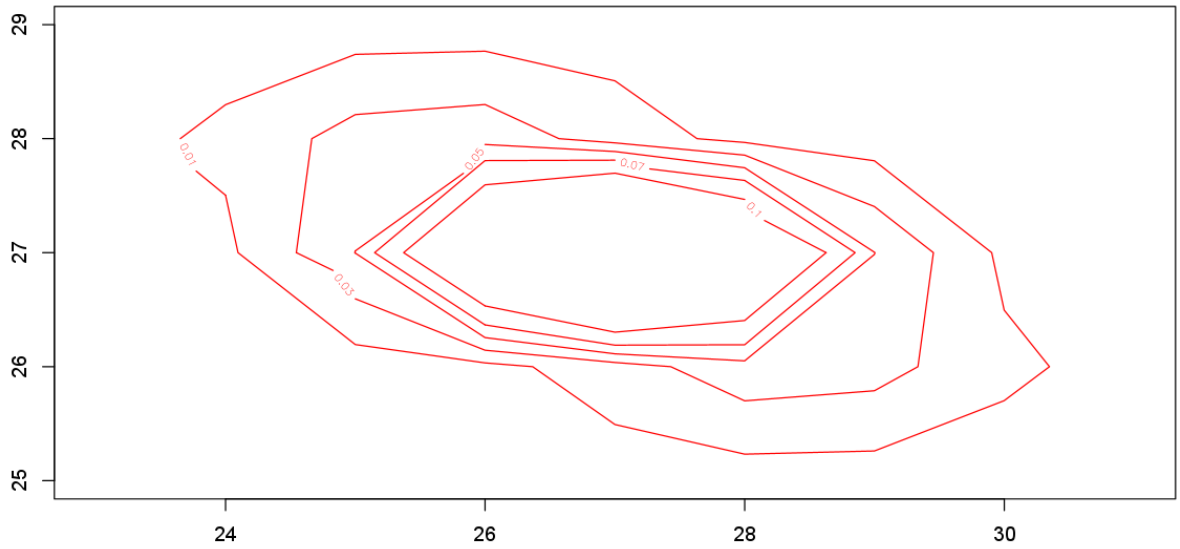
Antes de realizar un análisis de componentes principales, se realizó un análisis de normalidad. Se realizaron pruebas de Anderson y de Mardia y se obtuvieron los siguientes resultados.

```
$multivariateNormality
             Test        Statistic              p value Result
1 Mardia Skewness 474.747945136975 8.64265750182764e-21     NO
2 Mardia Kurtosis 3.59794900484948 0.000320736483631068     NO
3             MVN             <NA>                  <NA>     NO

$univariateNormality
               Test  Variable Statistic   p value Normality
1  Anderson-Darling    X3        3.6725  <0.001      NO
2  Anderson-Darling    X4        0.3496  0.4611      YES
3  Anderson-Darling    X5        4.0510  <0.001      NO
4  Anderson-Darling    X6        5.4286  <0.001      NO
5  Anderson-Darling    X7        0.9253  0.0174      NO
6  Anderson-Darling    X8        8.6943  <0.001      NO
7  Anderson-Darling    X9        1.9770  <0.001      NO
8  Anderson-Darling    X10       0.6585   0.081      YES
9  Anderson-Darling    X11       1.0469  0.0086      NO
10 Anderson-Darling    X12      14.3350  <0.001      NO
```

Como se puede observar en los resultados, solo las variables X4 y X10 se distribuyen como una normal, así que se realizaron las mismas pruebas pero solo utilizando esas dos variables.

```
$multivariateNormality
             Test         Statistic           p value Result
1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
3             MVN              <NA>              <NA>    YES

$univariateNormality
               Test  Variable Statistic   p value Normality
1 Anderson-Darling    X4        0.3496    0.4611    YES
2 Anderson-Darling    X10       0.6585    0.0810    YES
```

Como se puede observar en la imagen anterior las dos pruebas indican que hay normalidad multivariada. Para observar esto, se elaboro una grafica de contornos que se muestra a continuación.



Como se puede observar en la imagen, parece que ambas de las variables se distribuyen como una normal. Por ultimo se realizo una última prueba para detectar los valores atípicos y se obtuvo el siguiente resultado, indicando que no se encontraron valores atípicos.

```
$multivariateOutliers
NULL
```

Análisis de componentes principales

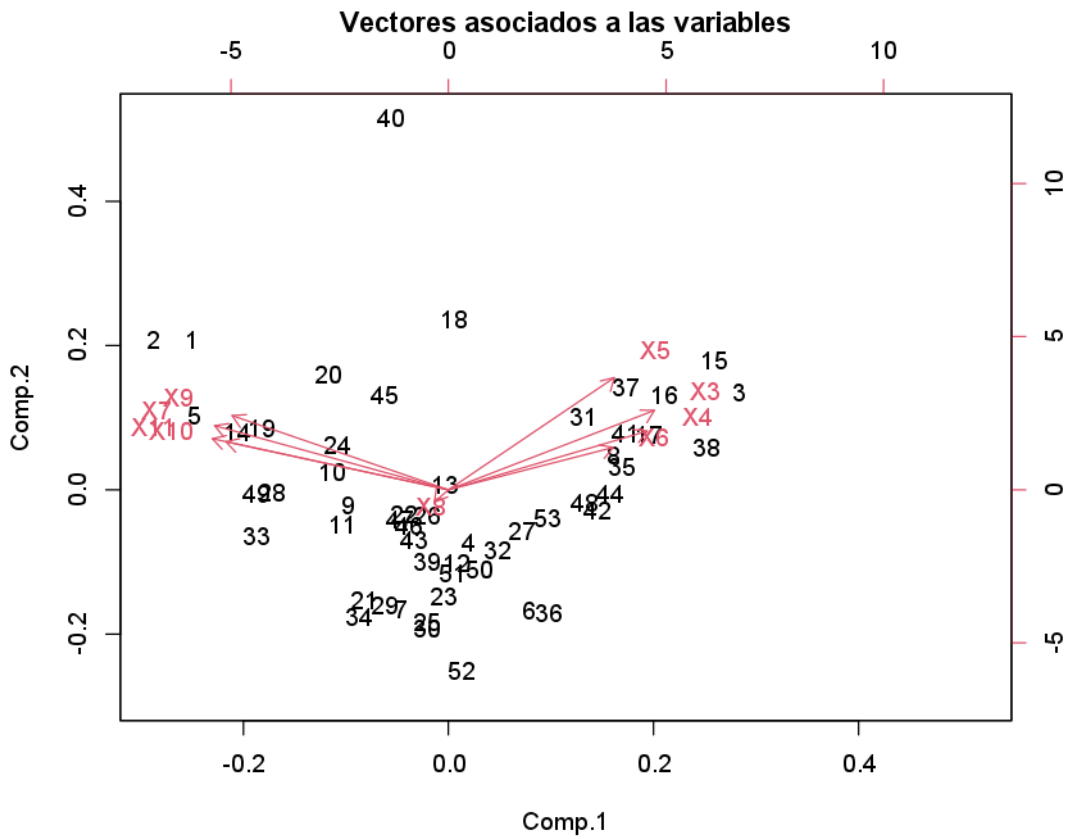
Para poder contestar la pregunta establecida en la introducción, se requiere realizar un análisis de componentes principales. Primero que todo se debe de justificar por qué este análisis es apropiado para este conjunto de datos.

Este análisis de componentes principales es apropiado para este conjunto de datos ya que el objetivo final es obtener un modelo de regresión para poder predecir la contaminación del mercurio en el agua de los lagos. Trabajar con mas de 10 variables diferentes vuelve complicada la selección de estas para dicho modelo. Es por esto que a través de componentes principales, es apropiado buscar reducir la dimensionalidad de los datos para reducir la complejidad.

Ya que se justificó el análisis, a continuación se aplica la instrucción de *princomp* de R para realizar el análisis de componentes principales y se obtuvieron los siguientes resultados.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.3121220 1.1049470 1.0210443 0.81722905 0.57940716
Proportion of Variance 0.5939898 0.1356564 0.1158368 0.07420704 0.03730141
Cumulative Proportion  0.5939898 0.7296462 0.8454831 0.91969010 0.95699151
                           Comp.6     Comp.7      Comp.8      Comp.9
Standard deviation     0.45709713 0.32749661 0.228103640 0.137307403
Proportion of Variance 0.02321531 0.01191711 0.005781252 0.002094814
Cumulative Proportion  0.98020682 0.99212393 0.997905186 1.000000000
```

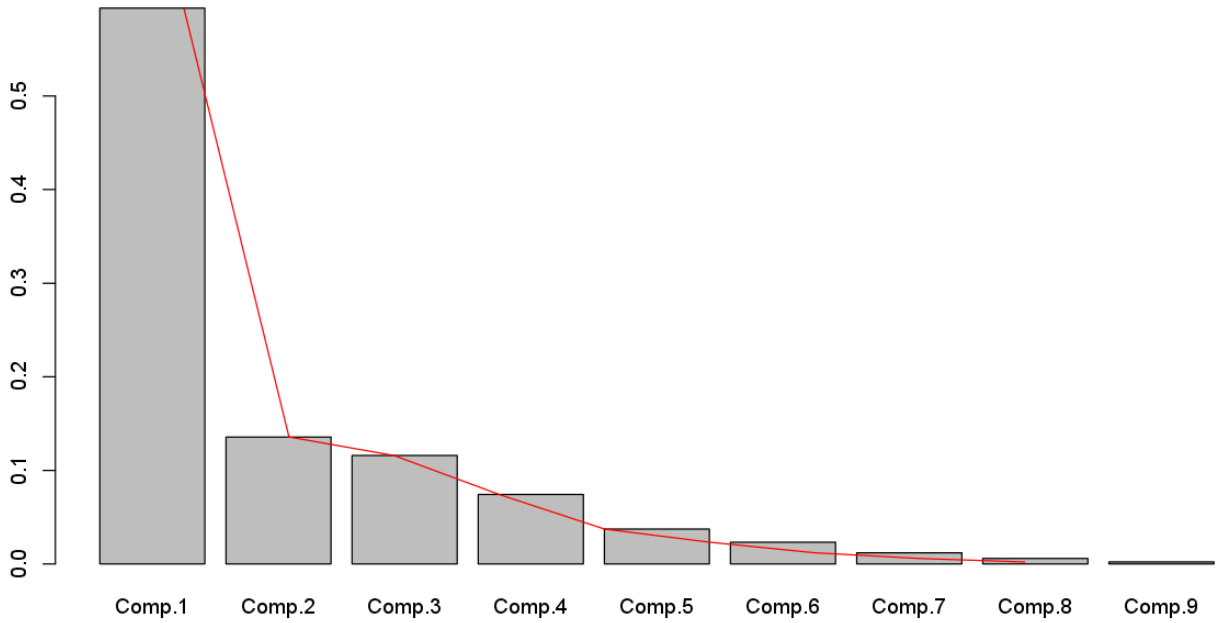
Observando las proporciones de la varianza, se puede notar como del componente 1 al 2 hay un gran salto, pero a partir de ahí empiezan a haber saltos mas pequeños en la varianza explicada de cada componente. Esto indica que los componentes 3 a 9, aunque la varianza explicada acumulada de estos es cerca del 30%, individualmente no aportan mucho. Ya que el propósito de este análisis es reducir la dimensionalidad los mas que se pueda, se utilizaran solo los componentes 1 y 2.



Al graficar los vectores de los componentes se puede observar que que se crearon un total de 3 grupos de variables. El de la izquierda que esta compuesto por las variables X7, X9, X10, X11, el del centro compuesto por X8 y el de la derecha compuesto por X3, X4, X5, X6. Este ultimo grupo esta compuesto por las variables de Alcalinidad, PH, Calcio y Clorofila, que son las variables que se buscan utilizar para un modelo de regresión.

Conclusiones

Con este análisis de componentes principales se concluyo que se puede reducir la dimensionalidad de los datos a dos, utilizando los componente 1 y 2 de este análisis.



En esta grafica se puede notar que del componente 2 en adelante hay una gran diferencia en la variabilidad explicada por cada componente. Indicando que es viable solo utilizar los componentes 1 y 2.

Anexo

- Link al repositorio de GitHub: https://github.com/facund015/ai_avanzada_personal/tree/main/Periodo_3/Portafolios/Implementacion/Mercurio
- Link a Google Drive: https://drive.google.com/drive/folders/10hZ8KOz8bo1MsBDCllxG5dJYjHxmos0u?usp=sharing