Tecnológico de Monterrey

REPORTE FINAL ¨LOS PECES Y EL MERCURIO”

Inteligencia artificial avanzada para la ciencia de datos

TC3007C.502

Módulo 5: Estadística Avanzada para Ciencia de Datos

Carlos David Toapanta Noroña A01657439

03 de diciembre del 2022

Resumen

A través de esta actividad se busca determinar, a través de un análisis estadístico, cuáles son los factores que, obtenidos de análisis en lagos de Florida, permiten describir qué influye en el nivel de contaminación por mercurio. La primera parte para resolver este reto fue un análisis de normalidad de los datos con la ayuda del Test de Mardia, Test de Anderson Darling, Contour Graph, Multivariate QQplot Graph y Mahalanobis distancia. La segunda parte consistió en un análisis de componentes principales que ayudaron a definir las variables que inciden en la contaminación, las cuales resultaron ser: la concentración media de mercurio en el tejido muscular del grupo de peces estudiados en cada lago, la concentración máxima de mercurio en cada grupo de peces y estimación de la concentración de mercurio en el pez de 3 años.

Introducción

La contaminación por mercurio de los peces en agua dulce comestible es una amenaza directa para nuestra salud. Se realizó un estudio reciente en 53 lagos de Florida para examinar los factores que influyen en el nivel de contaminación por mercurio. A partir de este estudio se determinaron 12 variables que describen las condiciones de cada lago y cuyo propósito es determinar las causas de la contaminación para entender cómo erradicar el problema.

La importancia del problema radica en el impacto directo que tiene sobre el ser humano que se alimenta de los peces en estos lagos o incluso su consumo de agua es abastecido por ellos. Es necesario determinar las principales causas de la alta concentración de mercurio en estos recursos naturales para buscar implementar estrategias que reduzcan sus niveles de concentración y así garantizar la salud de la población y la conservación del medio ambiente.

Análisis de resultados.

Mardia's Test.

	Beta-hat	kappa	p-val
Skewness	53.74505	474.747945	0.0000000000
Kurtosis	135.31273	3.597949	0.0003207365

Tomando en consideración la hipótesis:

HO: Las variables siguen una distribución normal multivariable

$H1$: Las variables no siguen una distribución normal multivariable

Y la regla de decisión con un nivel de significancia del 5%:

Se rechaza $H0$ si: el valor p es menor a $\alpha = 0.05$

La conclusión es que ya que el p-value de la curtosis es 0.0003 y del sesgo 0.0000, menores que Alpha; la hipótesis nula se rechaza, por lo no se tiene evidencia para decir que las variables del set de datos siguen una distribución multivariable.

Anderson-Darling Test.

```
data:  D$X4
A = 0.34956, p-value = 0.4611


        Anderson-Darling normality test

data:  D$X5
A = 4.051, p-value = 3.193e-10


        Anderson-Darling normality test

data:  D$X6
A = 5.4286, p-value = 1.4e-13


        Anderson-Darling normality test

data:  D$X7
A = 0.92528, p-value = 0.0174


        Anderson-Darling normality test

data:  D$X8
A = 8.6943, p-value < 2.2e-16


        Anderson-Darling normality test

data:  D$X9
A = 1.977, p-value = 4.161e-05


        Anderson-Darling normality test

data:  D$X10
A = 0.65847, p-value = 0.08099
```

Tomando en consideración la hipótesis:

HO: Los datos siguen una distribución normal

$H1$: Los datos no siguen una distribución normal

Y la regla de decisión con un nivel de significancia del 5%:

Se rechaza $H0$ si: el valor p es menor a $\alpha = 0.05$

La conclusión es que ya que el p-value de X4 es 0.46 y de X10 es 0.08, mayores que Alpha; la hipótesis nula no se rechaza, por lo que ambas variables siguen una distribución normal.

Mardia's test a variables que presentan normalidad.

	Beta-hat	kappa	p-val
Skewness	0.6991004	6.175387	0.1864276
Kurtosis	6.7602297	-1.128208	0.2592321

Tomando en consideración la hipótesis:

HO: Las variables siguen una distribución normal multivariable

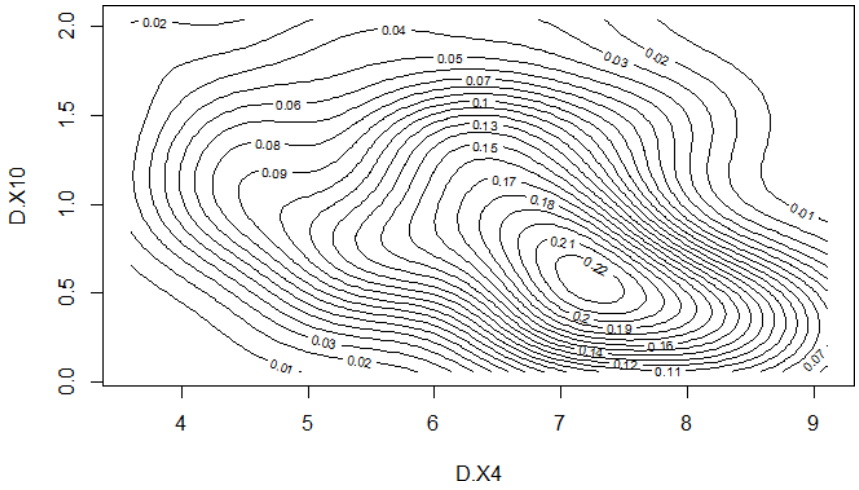
$H1$: Las variables no siguen una distribución normal multivariable

Y la regla de decisión con un nivel de significancia del 5%:

Se rechaza $H0$ si: el valor p es menor a $\alpha = 0.05$

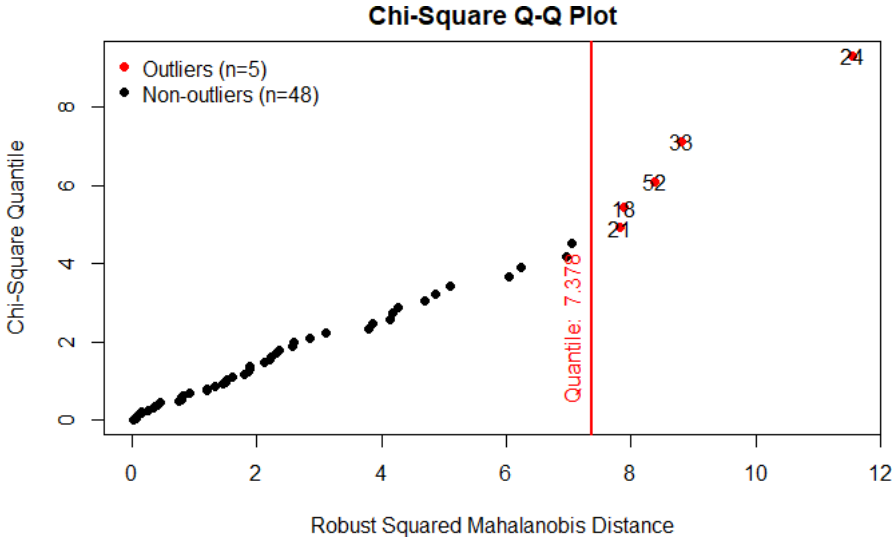
La conclusión es que ya que el p-value de la curtosis es 0.25 y del sesgo 0.18, mayores que Alpha; la hipótesis nula no se rechaza, por lo que las variables X4 y X5 siguen una distribución normal multivariable.

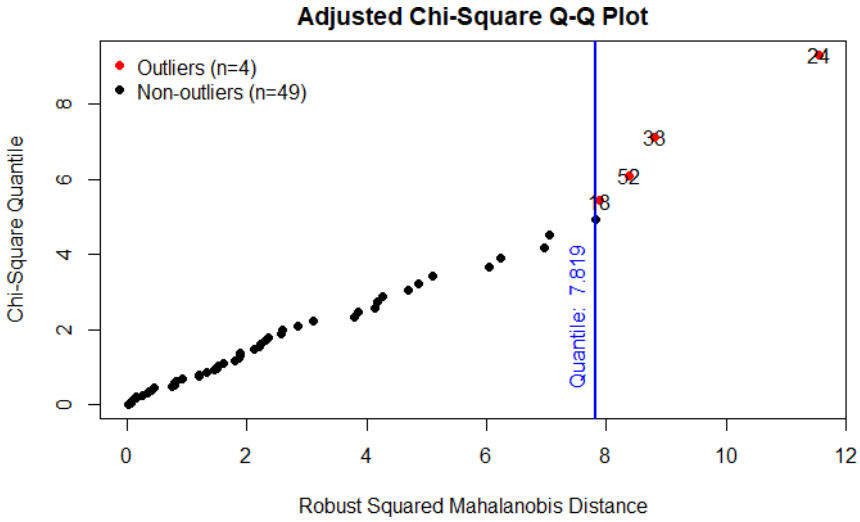
Gráfica de contorno.



Las regiones con valores más altos se pueden apreciar a partir de los niveles de contorno, mismos que revelan un pico centrado en aproximadamente 7 para el valor de X4 y en 0.5 para el valor de X10, las puntuaciones en esta región pico son superiores a 0.2

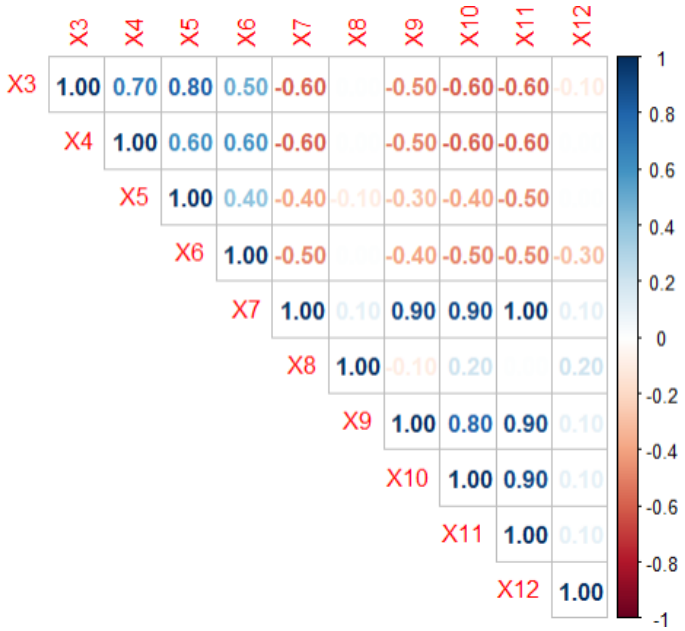
Datos atípicos o influyentes en la normal multivariada





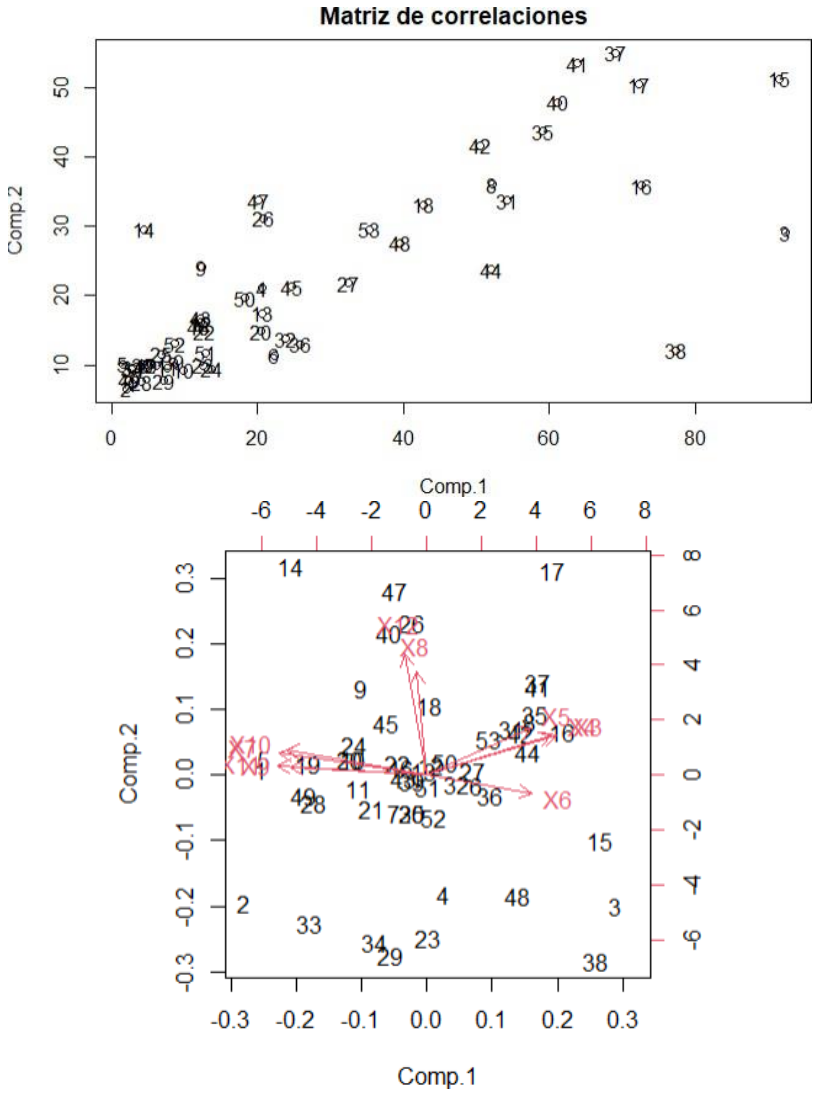
De las gráficas obtenidas, la distancia de Mahalanobis declara 5 observaciones como valor atípico multivariado, mientras que la distancia Mahalanobis ajustada declara 4

Matriz de correlaciones



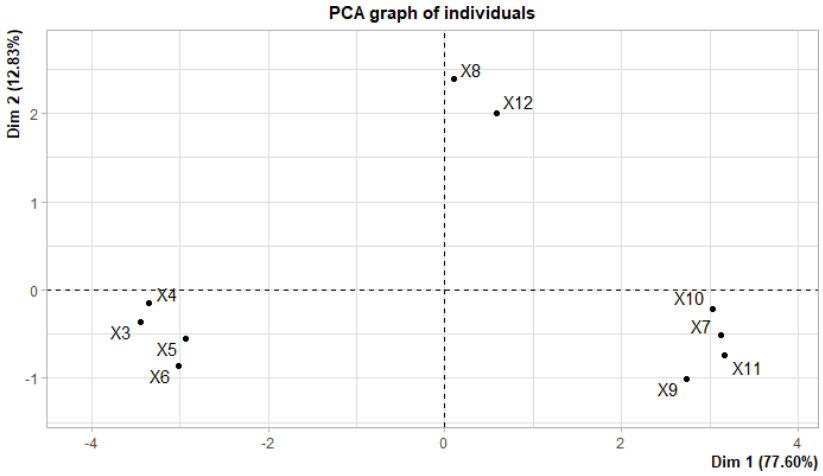
Es adecuado el uso de componentes principales para analizar la base debido a que esto permite identificar aquellas variables que tienen un mayor peso en la contaminación por mercurio de peces en el agua dulce comestible.

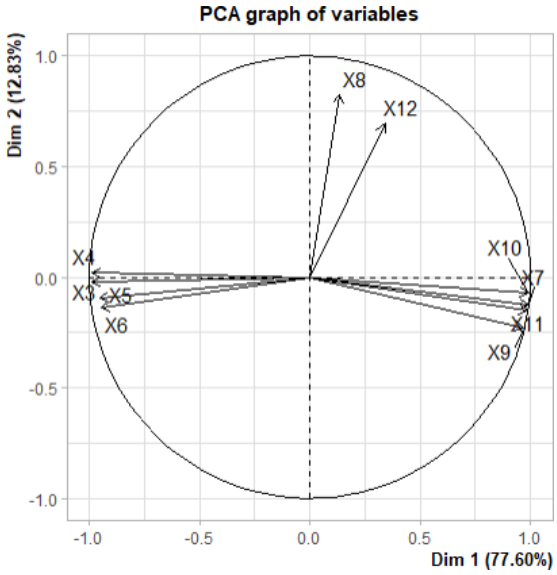
Análisis de componentes principales

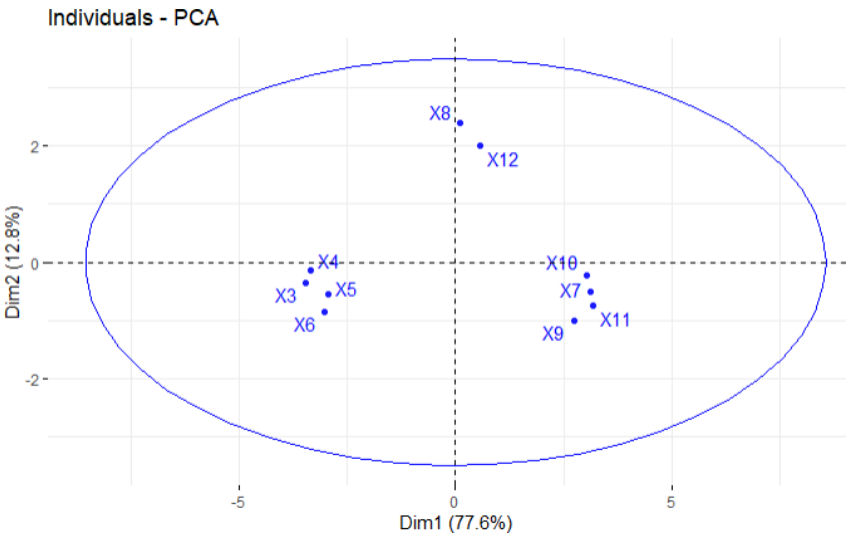


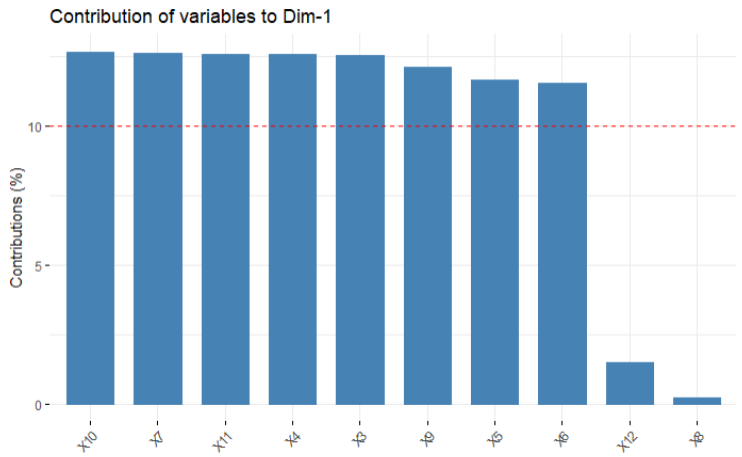
La matriz de correlaciones permite identificar dos componentes que agrupan las variables: la base 1, donde el porcentaje de proporción de varianza explicada es de 77.60% y la base 2 donde el porcentaje es 12.83%%. Este número de componentes principales explica poco más del 90% la exactitud, por lo que resulta ideal recurrir al uso de uno de ellos para reducir la dimensión de la base.

Vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes


PCA graph of individuals
Dim 2 (12.83%)
Dim 1 (77.60%)
X8
X12
X4
X3
X5
X6
X10
X7
X11
X9


PCA graph of variables
Dim 2 (12.83%)
Dim 1 (77.60%)
X8
X12
X4
X3
X5
X6
X10
X7
X11
X9


Individuals - PCA
Dim2 (12.8%)
Dim1 (77.6%)
X8
X12
X4
X3
X5
X6
X10
X7
X11
X9



En el primer gráfico se puede notar el porcentaje de proporción de varianza explicada que tiene cada uno de los dos primeros componentes; se puede notar que para el primero es de 77.60% y para el segundo 12.83%. También, permite definir las variables que tienen una mayor influencia en cada una de ellas. El segundo y tercer gráfico permite entender el mismo resultado descrito en el punto anterior, con una visualización distinta (agrupada por cuadrantes). El penúltimo gráfico permite visualizar la proporción de varianza explicada en cada componente, donde se puede demostrar que el primero es el que mayormente explica los datos. El último gráfico permite visualizar el porcentaje de contribución (peso) de cada variable en el primero componente.

Conclusión

En conclusión, los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida son: la concentración media de mercurio (partes por millón) en el tejido muscular del grupo de peces estudiado en cada lago, la concentración máxima de mercurio en cada grupo de peces y estimación (por regresión) de la concentración de mercurio en el pez de 3 años. La normalidad encontrada en un grupo de variables detectadas ayuda dentro de este estudio a realizar los análisis estadísticos de componentes principales con mayor precisión. La distribución normal se utiliza para conocer la probabilidad de encontrar un valor de la variable que sea igual o menor a cierto valor, conociendo la media, la desviación estándar y la varianza de un conjunto de datos al sustituirlos en la función que describe el modelo. Los componentes principales se basan en la proporción de la varianza explicada que, como se mencionó anteriormente, se ve afectada positivamente cuando se dispone de datos normalizados. Asimismo, los componentes principales permiten determinar aquellas variables que mejor explican, para este caso específico, la contaminación por mercurio de peces en agua dulce comestible.

Anexos

Código en R:

https://drive.google.com/drive/folders/1dnXchXBpEhKbnm6AO7jTo2oNejHn2kF3?usp=sharing