Reporte final “Los peces y el mercurio”

Ileana del Carmen Parra Enríquez

A00827284

28 de noviembre de 2022

Módulo 1: Estadística para la ciencia de datos.

TC3006C – Inteligencia Artificial Avanzada para la Ciencia de Datos. Gpo. 102

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. Como continuación del análisis anterior, se realizó un análisis de normalidad de las variables continuas, encontrando normalidad en las variables X4 y X10. También se realizó un análisis de componentes principales, encontrando que las variables que más influyen en el componente uno son las variables X11 y X7.

Introducción

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. Se busca responder la principal pregunta del estudio: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? Además, en esta continuación del trabajo, se busca facilitar el análisis al utilizar componentes principales y buscar normalidad en las variables.

La importancia de este problema reside en que una alta concentración de mercurio en lo que consumimos, afecta directamente nuestra salud. Es por esto que encontrar la raíz del problema y observarlo directamente es necesario. Con este reporte se pretende dar a conocer algunos de los factores que influyen en la concentración de mercurio para poder vigilar estos factores y así reducir la concentración de mercurio en los lagos y peces.

Análisis de los resultados

Prueba de normalidad

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	410.214790601478	7.04198777815398e-23	NO
Mardia Kurtosis	4.59612555772731	4.30419392238868e-06	NO
MVN	*NA*	*NA*	NO

La prueba de Mardia indica un sesgo de 410.214 lo que significa que los datos presentan un sesgo a la derecha, también nos indica que contamos con una distribución platicúrtica ya que la curtosis es mayor a 3.

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	Column1	3.6725	<0.001	NO
2	Anderson-Darling	Column2	0.3496	0.4611	YES
3	Anderson-Darling	Column3	4.0510	<0.001	NO
4	Anderson-Darling	Column4	5.4286	<0.001	NO
5	Anderson-Darling	Column5	0.9253	0.0174	NO
6	Anderson-Darling	Column6	8.6943	<0.001	NO
7	Anderson-Darling	Column7	1.9770	<0.001	NO
8	Anderson-Darling	Column8	0.6585	0.081	YES
9	Anderson-Darling	Column9	1.0469	0.0086	NO

La prueba de Anderson-Darling nos indica que no tenemos normalidad multivariada ya que solo dos variables son normales.

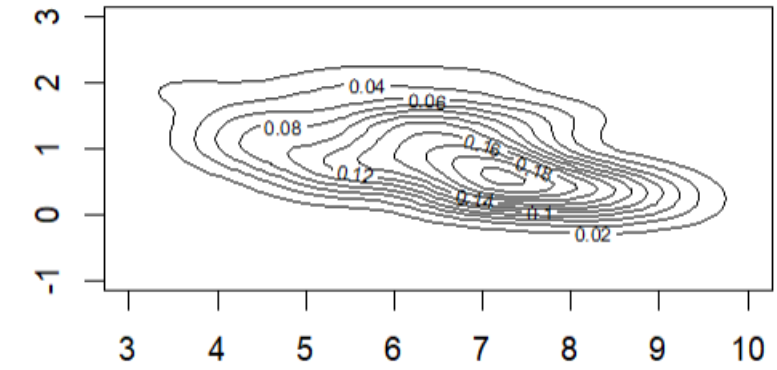
Volvemos a realizar las pruebas de Mardia y Anderson Darling en las variables que sí tuvieron normalidad, en este caso las variables X2 y X8.

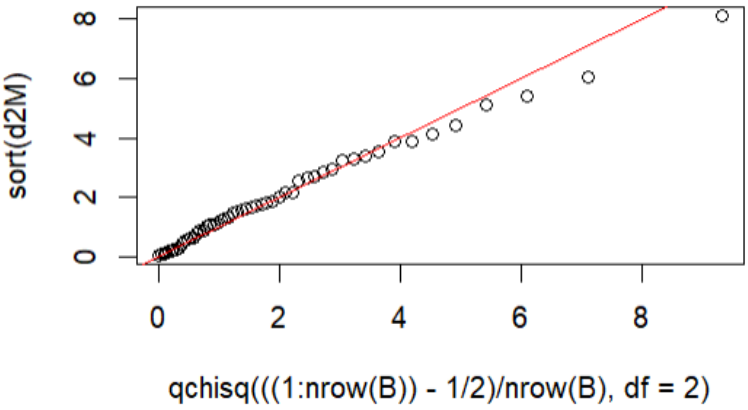
Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	6.17538668676458	0.186427564928852	YES
Mardia Kurtosis	-1.12820795824432	0.25923210375991	YES
MVN	*NA*	*NA*	YES

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	Column1	0.3496	0.4611	YES
2	Anderson-Darling	Column2	0.6585	0.0810	YES

- La prueba de Mardia indica un sesgo a la derecha y una distribución leptocúrtica.
- La prueba de Anderson-Darling nos indica que sí hay normalidad multivariada ya que ambas variables son normales.

Gráfica de contornos y QQplot





De acuerdo con la gráfica, los datos no se comportan como una normal, ya que presentan un sesgo. Los puntos que se alejan de la gráfica se identifican como outliers.

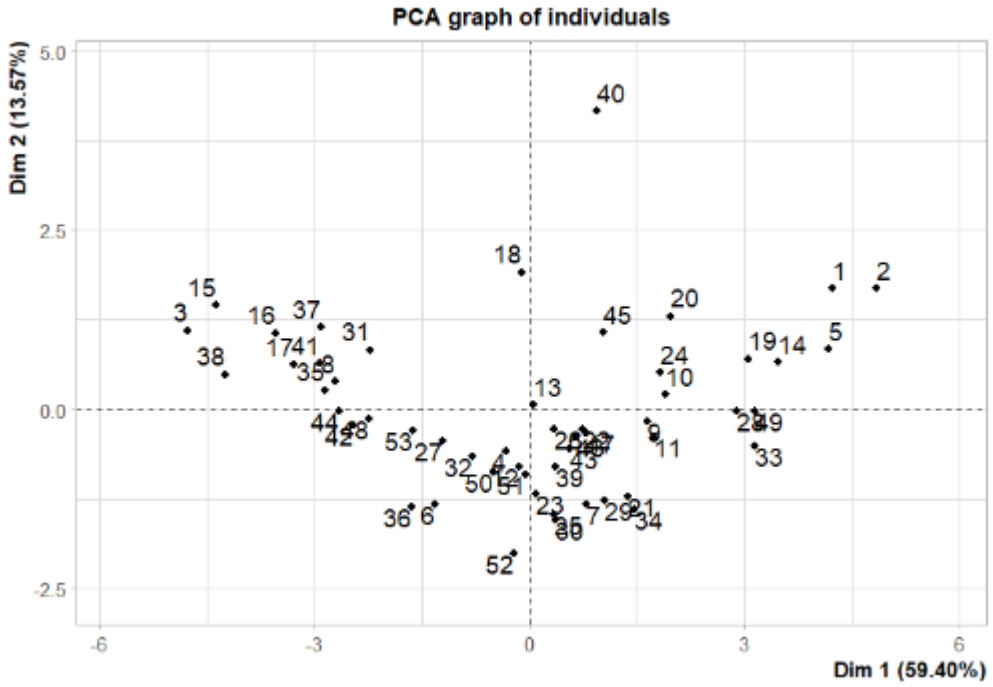
Análisis de componentes principales

Como se puede ver en la matriz de correlaciones, hay una fuerte correlación entre la mayoría de las variables, por lo que el uso de componentes principales es adecuado.

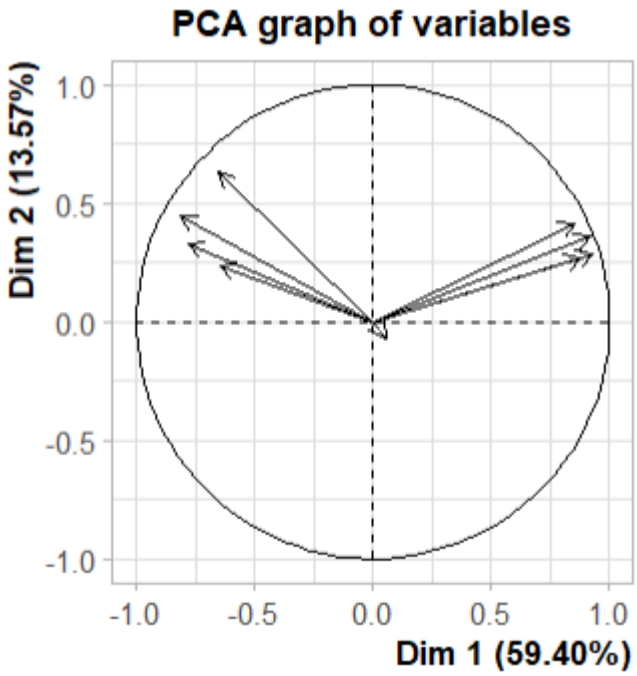
```
             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]        [,7]       [,8]        [,9]
[1,]  1.00000000  0.71916568  0.83260419  0.47753085 -0.59389671  0.01029074 -0.52535654 -0.6047956 -0.62795845
[2,]  0.71916568  1.00000000  0.57713272  0.60848276 -0.57540012 -0.01860607 -0.54196524 -0.5518152 -0.61284905
[3,]  0.83260419  0.57713272  1.00000000  0.40991385 -0.40067958 -0.08937901 -0.33247623 -0.4079166 -0.46440947
[4,]  0.47753085  0.60848276  0.40991385  1.00000000 -0.49137481 -0.01182027 -0.40045856 -0.4849721 -0.50644193
[5,] -0.59389671 -0.57540012 -0.40067958 -0.49137481  1.00000000  0.07903426  0.92720506  0.9158640  0.95921481
[6,]  0.01029074 -0.01860607 -0.08937901 -0.01182027  0.07903426  1.00000000 -0.08165278  0.1610917  0.02580046
[7,] -0.52535654 -0.54196524 -0.33247623 -0.40045856  0.92720506 -0.08165278  1.00000000  0.7653532  0.91908939
[8,] -0.60479558 -0.55181523 -0.40791663 -0.48497215  0.91586397  0.16109174  0.76535319  1.0000000  0.85975810
[9,] -0.62795845 -0.61284905 -0.46440947 -0.50644193  0.95921481  0.02580046  0.91908939  0.8597581  1.00000000
```

```
"La varianza total es:"
3106.289
"La suma acumulada es:"
0.7264164 0.9300767 0.9775266 0.9996963 0.9999067 0.9999882 0.9999979 0.9999994 1.0000000
```

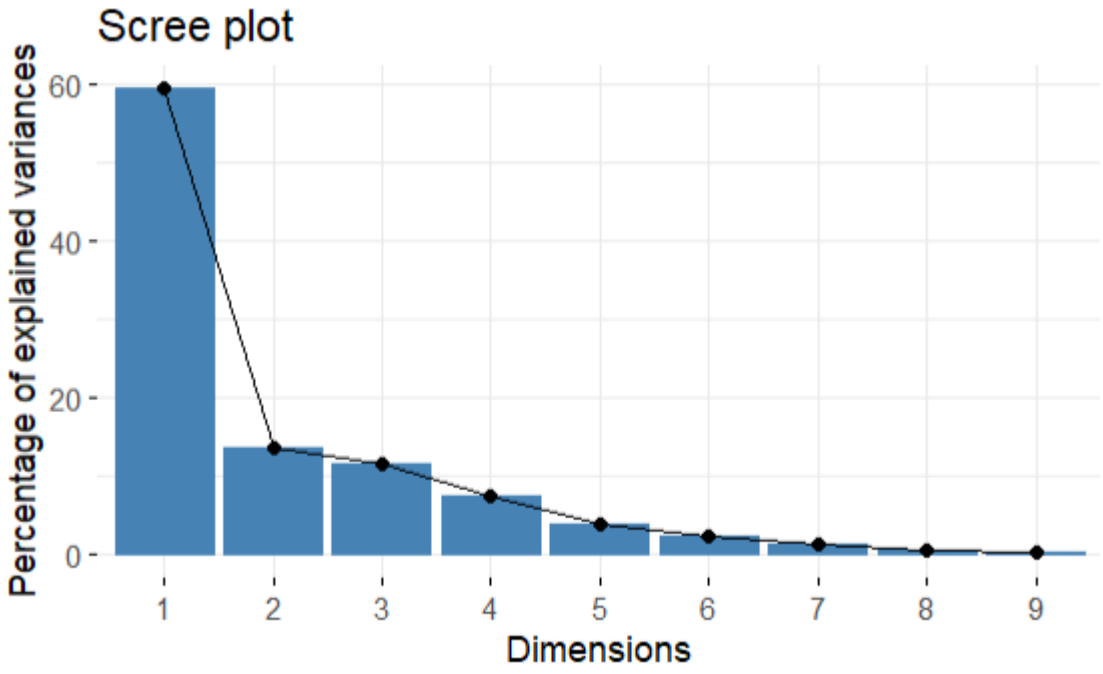
El componente uno explica el 72.64% de la varianza. De acuerdo con el porcentaje de varianza explicada, el número ideal de componentes es 3, ya que con estos se logra explicar el 97.753%.



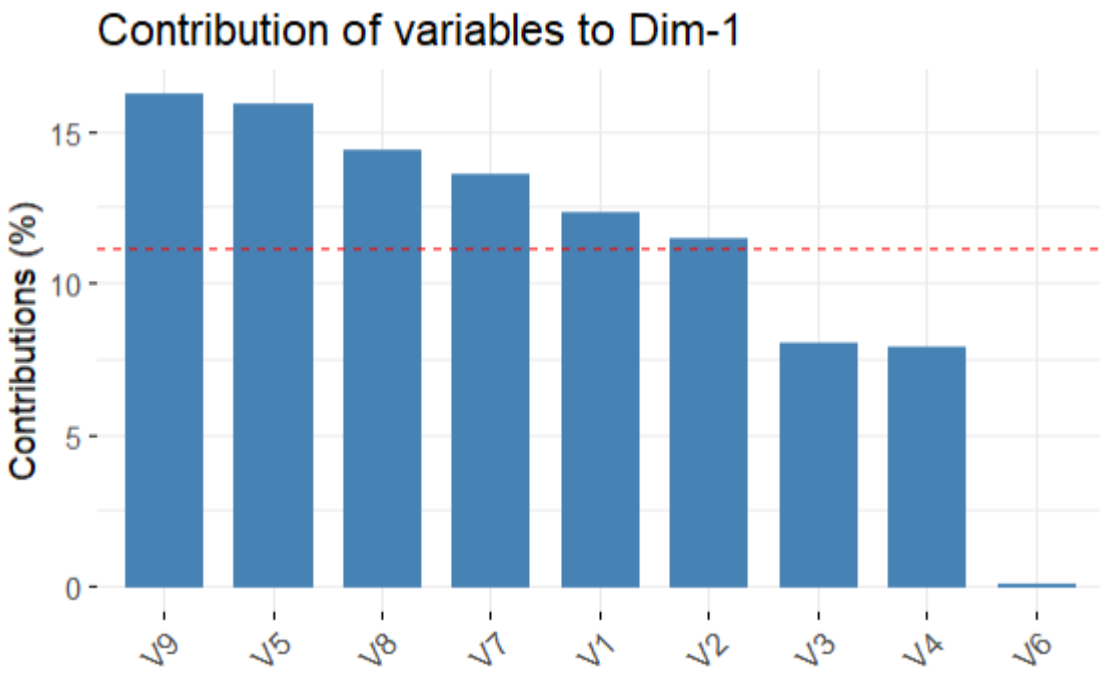
En la gráfica se pueden observar algunos outliers como el 40 y un clúster de variables en el segundo, tercer y cuarto cuadrante.



La gráfica muestra las variables que tienen mayor influencia en los componentes. Se puede ver que el componente dos casi no tiene variables con correlación negativa, mientras que el componente uno la mitad de las variables tienen correlación positiva y la otra mitad negativa. De acuerdo con el gráfico, la variable 9 tiene mayor influencia en el componente uno y la variable 3 en el componente 2.



El gráfico de sedimentación muestra que el primer componente explica la mayor parte de la varianza. Podemos usar este gráfico para justificar el número ideal de componentes el cual estaría entre 3 y 4, ya que después de estos componentes el porcentaje de varianza explicada es mínimo.



El último gráfico nos muestra que las variables 9 y 5 son las que más contribuyen al componente uno.

Conclusiones

- Las variables que más influyen en el componente uno es:

X11: estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)

X7 = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago

- El análisis por componentes principales nos permite reducir la dimensionalidad del conjunto de datos, es decir, nos facilita el trabajo, ya que en este caso reducimos el conjunto de datos de 9 variables a 3.
- Debido a que se encontró normalidad en 2 variables facilita hacer cálculos o análisis donde se necesita normalidad para obtener una respuesta. La normalidad se encontró en las siguientes variables:

X4 = PH

X10 = máximo de la concentración de mercurio en cada grupo de peces

Anexos

https://drive.google.com/file/d/1eaSP3Ud_bdyQLaoZjjF-ra0HSEsRy169/view?usp=share_link