Los peces y el mercurio 2

Juan Pablo Yáñez González
A00829598
Módulo 2: Inteligencia artificial avanzada para la ciencia de datos
Grupo 2

Resumen:

La contaminación de mercurio en peces es una amenaza directa contra la salud humana, es por eso que este estudio busca encontrar los principales factores que afectan a esta concentración, de acuerdo con un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influyen en el nivel de contaminación por mercurio. Los resultados mostraron que hay 4 factores importantes en la concentración de mercurio como lo es la alcalinidad en agua, el PH, el calcio y la clorofila.

Introducción:

En este estudio se busca encontrar las variables que afectan la concentración media de mercurio en los lagos de florida, esto es importante ya que considerando las normativas de referencia para evaluar los niveles máximos de Hg (Reglamento 34687-MAG y los reglamentos internacionales CE 1881/2006 y Codex Standard 193-1995) establecen que la concentración promedio de mercurio en productos de la pesca no debe superar los 0.5 mg de Hg/kg. Además de esto en nuestro estudio surgieron otras incógnitas a las cuales se les dio respuesta, estas fueron: ¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana?, ¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?, y ¿Las concentraciones de alcalinidad, clorofila, calcio en el agua del lago influyen en la concentración de mercurio de los peces?.

Análisis de los datos

En nuestros datos existen 12 variables, las cuales vienen siendo las siguiente:

X1 = número de identificación

X2 = nombre del lago

X3 = alcalinidad (mg/l de carbonato de calcio)

X4 = PH

X5 = calcio (mg/l)

X6 = clorofila (mg/l)

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

X8 = número de peces estudiados en el lago

X9 = mínimo de la concentración de mercurio en cada grupo de peces

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

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)

X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Ya al estar familiarizados con los nombres de los valores, continuamos a sacar valores importantes los cuales nos servirán a futuro para nuestro análisis.

```
       X1              X2                  X3               X4        
 Min.   : 1   Length:53          Min.   :  1.20   Min.   :3.600  
 1st Qu.:14   Class :character   1st Qu.:  6.60   1st Qu.:5.800  
 Median :27   Mode  :character   Median : 19.60   Median :6.800  
 Mean   :27                      Mean   : 37.53   Mean   :6.591  
 3rd Qu.:40                      3rd Qu.: 66.50   3rd Qu.:7.400  
 Max.   :53                      Max.   :128.00   Max.   :9.100  
       X5              X6                  X7               X8        
 Min.   : 1.1   Min.   :  0.70   Min.   :0.0400   Min.   : 4.00  
 1st Qu.: 3.3   1st Qu.:  4.60   1st Qu.:0.2700   1st Qu.:10.00  
 Median :12.6   Median : 12.80   Median :0.4800   Median :12.00  
 Mean   :22.2   Mean   : 23.12   Mean   :0.5272   Mean   :13.06  
 3rd Qu.:35.6   3rd Qu.: 24.70   3rd Qu.:0.7700   3rd Qu.:12.00  
 Max.   :90.7   Max.   :152.40   Max.   :1.3300   Max.   :44.00  
       X9              X10              X11              X12        
 Min.   :0.0400   Min.   :0.0600   Min.   :0.0400   Min.   :0.0000  
 1st Qu.:0.0900   1st Qu.:0.4800   1st Qu.:0.2500   1st Qu.:1.0000  
 Median :0.2500   Median :0.8400   Median :0.4500   Median :1.0000  
 Mean   :0.2798   Mean   :0.8745   Mean   :0.5132   Mean   :0.8113  
 3rd Qu.:0.3300   3rd Qu.:1.3300   3rd Qu.:0.7000   3rd Qu.:1.0000  
 Max.   :0.9200   Max.   :2.0400   Max.   :1.5300   Max.   :1.0000
```

En estos resultados podemos ver el mínimo y máximo de cada variable, además de la media, la mediana y la representación de unos cuantos cuartiles. Podemos concluir en que tenemos 7 variables ya que el ID, nombre del lago, el mínimo y máximo de concentración (debido a que son derivadas de lo que deseamos predecir) y la estimación mediante regresión, son variables despreciables haciendo que las eliminemos de nuestro data frame.

Análisis de normalidad

Después de esto continuamos por realizar un análisis de normalidad en nuestras variables ya que el tamaño de la muestra es muy pequeño, así sabremos el patrón de distribución que estas siguen. Para esto planteamos una hipótesis H0: p = 0 muestra distribución normal, H1: p!0 muestra distribución anormal. Para esto realizamos las pruebas de Anderson y Mardia.

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	264.072513156727	1.74090404676373e-20	NO
Mardia Kurtosis	3.97661926862428	6.99019470364881e-05	NO
MVN	NA	NA	NO

Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
X3	3.6725	<0.001	NO
X4	0.3496	0.4611	YES
X5	4.0510	<0.001	NO
X6	5.4286	<0.001	NO
X7	0.9253	0.0174	NO
X8	8.6943	<0.001	NO
X12	14.3350	<0.001	NO

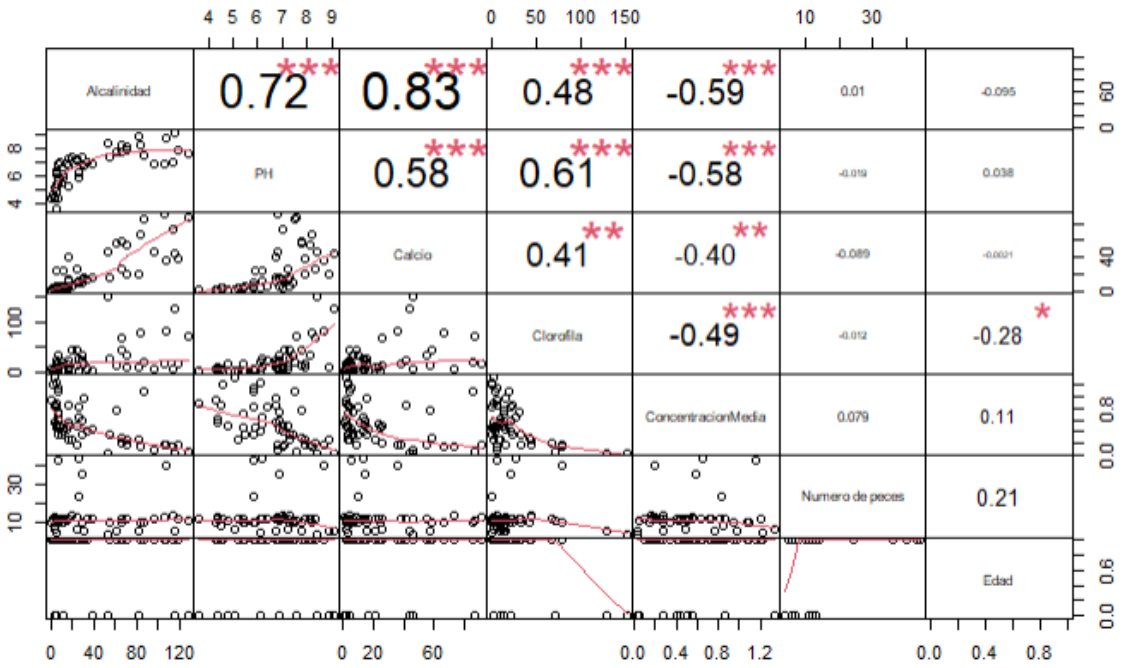
Como podemos observar solo una variable salió con un p value cercano al 0 este fue el PH, para corregir esto hicimos uso de la función boxcox para para estimar el parámetro de transformación por estimación de máxima verosimilitud. Al transformar los datos volvimos a realizar las pruebas y nos dieron los siguientes resultados.

Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
newDf.X4	0.3496	0.4611	YES
nd_X3	0.8704	0.0239	NO
nd_X5	0.7818	0.0398	NO
nd_X6	0.1744	0.9213	YES

Estos resultados demuestran que la clorofila y el ph siguen una distribución normal, cumpliendo al igual con la normalidad del sesgo y la kurtosis.

Después de esto se realizó un análisis de componentes principales para determinar cuáles son las variables que presentan una importancia en la concentración media de mercurio. Utilizaremos este análisis ya que se trata de un conjunto de datos multidimensionales, permitiéndonos obtener la variabilidad de los datos y determinar cuales son los de mayor importancia.

Comenzaremos por realizar una matriz de correlación para determinar si nuestras variables presentan alguna relación entre ellas.

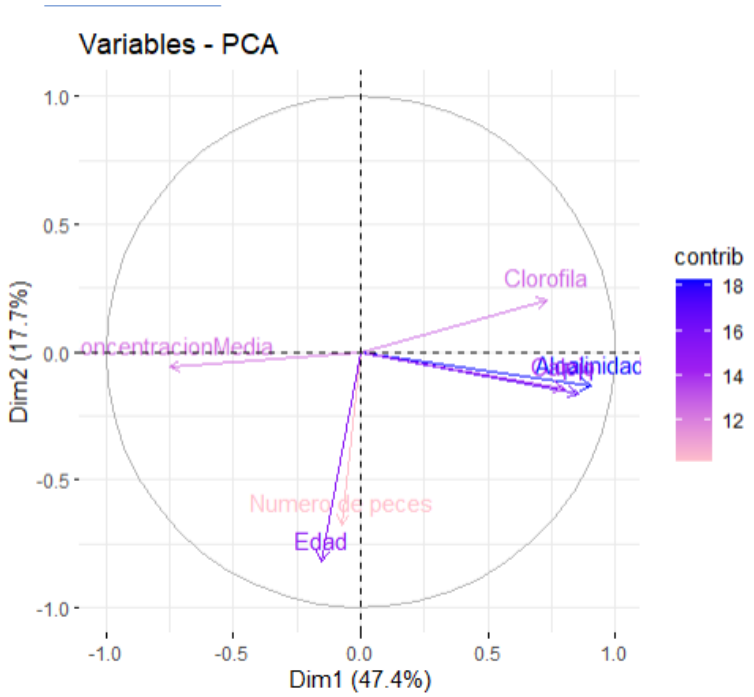


Como podemos observar nuestra gráfica demuestra que todas nuestras variables presentan una relación entre ellas a excepción del número de peces y la edad de estos mismos, contestando una pregunta de nuestra incógnita al saber que la edad no altera la concentración media porque no presenta relación entre los componentes.

Al momento de realizar la prueba de PCA se nos arrojan los siguientes datos.

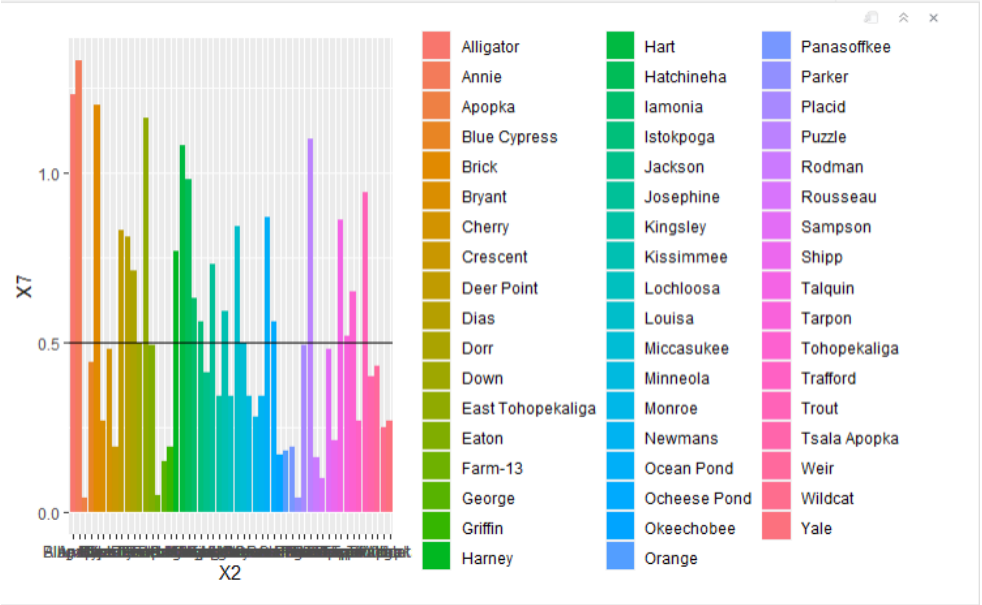
```
See two factoextra-related books at https://goo.gl/ve3wBa
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.8206 1.1141 0.9492 0.80962 0.70773 0.53413 0.31841
Proportion of Variance 0.4735 0.1773 0.1287 0.09364 0.07155 0.04076 0.01448
Cumulative Proportion  0.4735 0.6509 0.7796 0.87321 0.94476 0.98552 1.00000
```

En la proporción de varianza observamos que nuestros componentes tienen un valor menor al 0 lo que nos indica que esos son los datos relevantes, es decir que son los datos que alteran la concentración media de mercurio en el lago.



Al ver esto gráficamente en la gráfica vectorial, observamos mejor como las variables relacionadas se dirigen hacia un mismo lado mientras que las no relacionadas van en un sentido opuesto, además la variable la cual logran encontrar la relación va en dirección opuesta, en este caso sería la concentración media. Con esto podemos asegurar que la Clorofila, el Ph, el calcio y la alcalinidad son factores que sí influyen en la concentración media de mercurio en los lagos de Florida.

Por último realizaremos un histograma para poder comprobar la concentración media de mercurio en cada lago y obtener el promedio de esta concentración, demostrando si se podría decir que esta es dañina.



En el gráfico superior podemos observar que hay varios lagos con una concentración superior a la media de 0.5 dañina para el consumo humano. Además de que el promedio del CMM resulta en 0.48, si bien aún no llega al punto en el que se considere normativamente dañino, esto sigue siendo alarmante.

Conclusión

En conclusión podríamos decir que los factores importantes influyentes en la concentración media de mercurio en los lagos de florida, son el calcio, el mercurio, el ph y la alcalinidad, dependiendo de las concentraciones de estos minerales en el agua se podría sacar un buen aproximado del porcentaje de mercurio que se encuentra en ella. Además podemos decir que si hay evidencia suficiente para indicar que la cantidad de mercurio en ciertos lagos de florida es dañina y que se deberían de tomar acciones para corregirlo ya que hay evidencia suficiente como para decir que el promedio de la concentración de mercurio, no está normativamente por encima del 0.5 dañino para los humanos pero está

extremadamente cerca y muy probable en unos años llegue a estar por encima de ese nivel.

Referencias:

What Is Principal Component Analysis (PCA) and How It Is Used? (2020). Sartorius. https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186#:~:text=Principal%20component%20analysis%2C%20or%20PCA,more%20easily%20visualized%20and%20analyzed.

Wikipedia Contributors. (2022, November 10). Principal component analysis. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Principal_component_analysis

Anderson-Darling Test in R (Quick Normality Check) | R-bloggers. (2021, November 9). R-Bloggers. https://www.r-bloggers.com/2021/11/anderson-darling-test-in-r-quick-normality-check/

Zach. (2019, April 22). How to Conduct an Anderson-Darling Test in R - Statology. Statology. https://www.statology.org/anderson-darling-test-r/

Anexos:

https://drive.google.com/drive/folders/1TPSuifP10dXi_pWhyH31JBM9z0bfiO9f?usp=sharing