Inteligencia Artificial Avanzada para Ciencia de Datos

Instituto Tecnológico de Estudios Superiores de Monterrey

Reporte final “Los peces y el mercurio”

Rodrigo Montelongo Pinales, A00827757

Módulo 5: Estadística avanzada para ciencia de datos, Grupo 502

03 de diciembre del 2022

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 para para poder examinar los factores que influyen en el nivel de contaminación por mercurio.

Para este análisis primero se exploraron los datos y se limpiaron. Se utilizaron herramientas de visualización como histogramas y matriz de correlación. En cuanto a estadística, se realizó un análisis de componentes principales.

Los principales resultados arrojaron que las variables que más influyen son la concentración mínima y máxima de mercurio en el grupo de peces, así como el estimado de un pez de 3 años. No obstante, esto es debido a la relación evidente entre estos datos y la variable de concentración media. Por lo tanto, las variables que más influyen después de éstas son el PH y la alcalinidad.

Introducción

La principal pregunta de investigación que surge de este estudio es ¿cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?, ¿las concentraciones de alcalinidad, clorofila y calcio en el agua del lago influyen en la concentración del mercurio de los peces?

Análisis de los resultados

Lo primero es explorar las variables. Podemos ver que existen 12 variables. Estas se pueden clasificar en cuantitativas y cualitativas.

Cuantitativas:

- X1: Número de identificación
- X3: Alcalinidad
- X4: PH
- X5: Calcio
- X6: Clorofila
- X7: Concentración media de mercurio en el tejido muscular del grupo medio de peces estudiados

- X8: Número de peces estudiados en el lago

- X9: Mínimo concentración de mercurio en peces estudiados

- X10: Máximo concentración de mercurio en peces estudiados

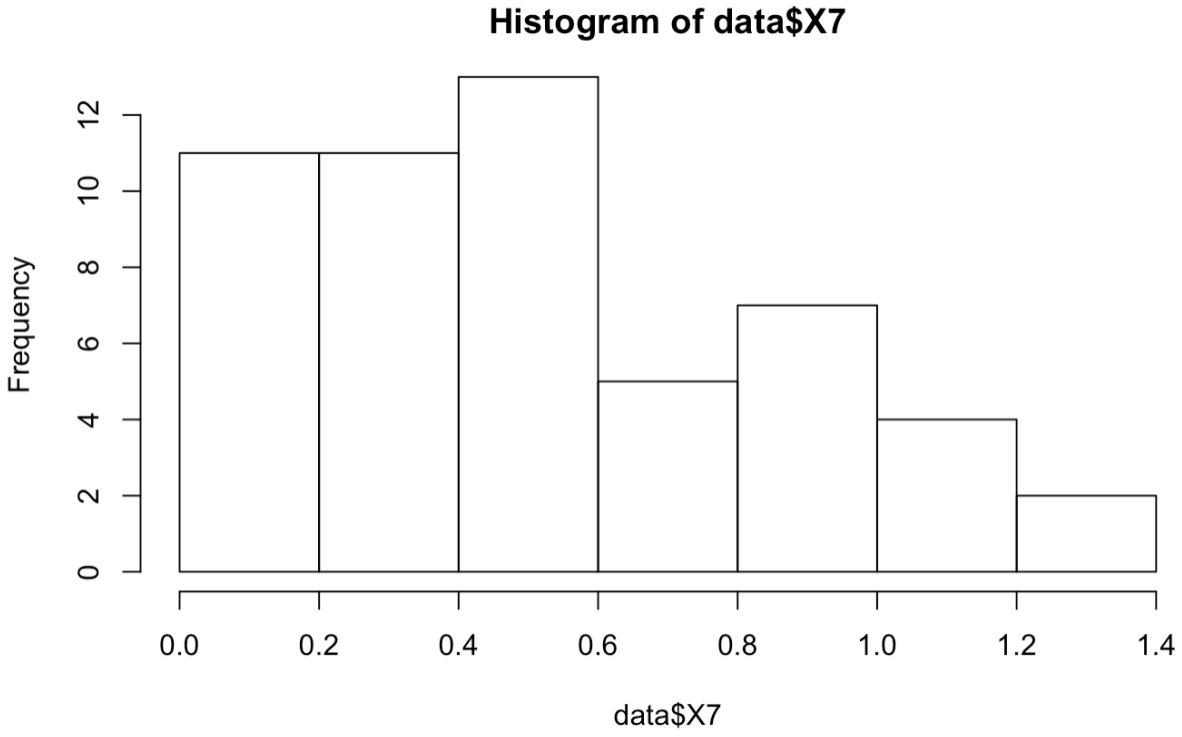
- X11: Estimación de la concentración de mercurio en el pez de 3 años

- X12: Indicador de la edad de los peces

Cualitativas

- X2: Nombre del lago

Se realizó un histograma de la concentración de mercurio para ver si en realidad se están infringiendo los límites permitidos.



Como se puede observar podemos ver que existen varios lagos que están por el límite de lo permitido de concentración de mercurio en los peces, la cual es 0.5. Es necesario encontrar las variables que pueden tener mayor relación con este fenómeno.

Al analizar los datos podemos ver que sólo uno no es numérico, X2 (el lago). Por lo tanto, debemos explorarlo para ver si es necesario incluirlo en el modelo. Esto ya que la técnica de análisis de componentes principales que utilizaremos no permite que haya variables no numéricas.

```
'data.frame':   53 obs. of  12 variables:
 $ X1 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ X2 : Factor w/ 53 levels "Alligator","Annie",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ X3 : num  5.9 3.5 116 39.4 2.5 19.6 5.2 71.4 26.4 4.8 ...
 $ X4 : num  6.1 5.1 9.1 6.9 4.6 7.3 5.4 8.1 5.8 6.4 ...
 $ X5 : num  3 1.9 44.1 16.4 2.9 4.5 2.8 55.2 9.2 4.6 ...
 $ X6 : num  0.7 3.2 128.3 3.5 1.8 ...
 $ X7 : num  1.23 1.33 0.04 0.44 1.2 0.27 0.48 0.19 0.83 0.81 ...
 $ X8 : int  5 7 6 12 12 14 10 12 24 12 ...
 $ X9 : num  0.85 0.92 0.04 0.13 0.69 0.04 0.3 0.08 0.26 0.41 ...
 $ X10: num  1.43 1.9 0.06 0.84 1.5 0.48 0.72 0.38 1.4 1.47 ...
 $ X11: num  1.53 1.33 0.04 0.44 1.33 0.25 0.45 0.16 0.72 0.81 ...
 $ X12: int  1 0 0 0 1 1 1 1 1 1 ...
```

En los registros podemos ver que cada lago (X2) es diferente, por lo que se puede descartar para el análisis.

De igual forma eliminamos la variable X1 ya que es el número de identificación y no aporta nada al modelo.

Procedemos a crear la matriz de correlación. Esto nos ayudará para saber cuáles son las variables que más afectan en el nivel de contaminación de los peces. En este caso nos enfocaremos en X7, la cual es la concentración media de mercurio en los peces de cada lago. Podemos observar que existe mayor correlación con X3, X4, X9, X10 y X11. Éstas serán las variables que utilizaremos para nuestro modelo.

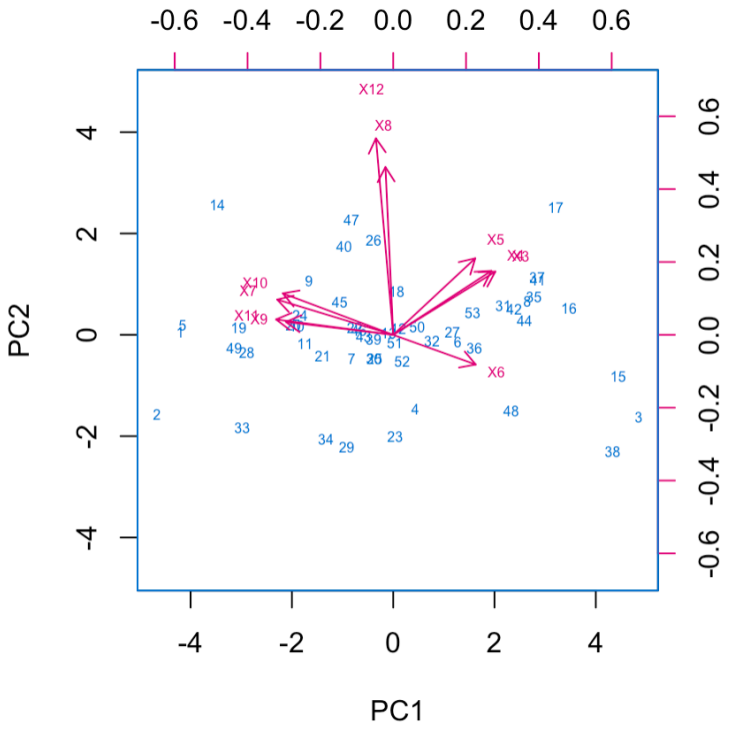
```
             X1          X3          X4           X5          X6          X7
X1   1.00000000  0.10381134  0.04059144  0.105022732 -0.08198056 -0.27746561
X3   0.10381134  1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671
X4   0.04059144  0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012 -
X5   0.10502273  0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958 -
X6  -0.08198056  0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481 -
X7  -0.27746561 -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000
X8   0.03011027  0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426
X9  -0.26878640 -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506 -
X10 -0.21212470 -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397
X11 -0.30402207 -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481
X12  0.13240584 -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896
```

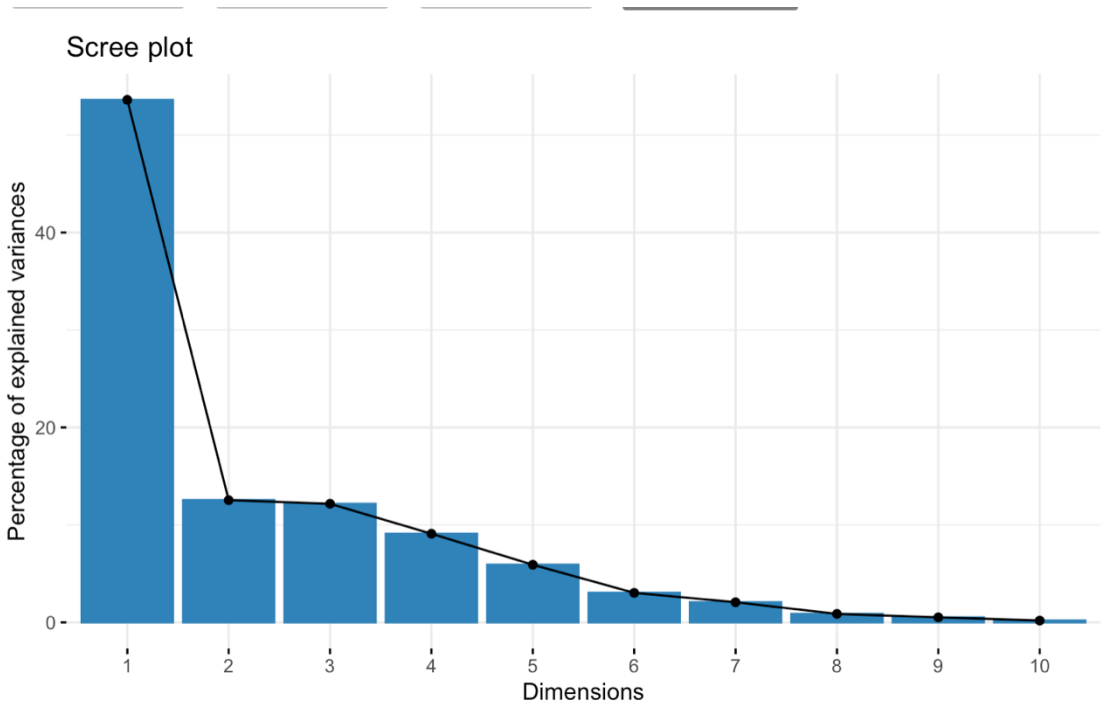
En este caso vemos que las que tienen mayor correlación con la concentración de mercurio son:

- X9 (mínimo de concentración)
- X10(máximo de concentración)
- X3(alcalinidad)
- X4(PH)
- X6(clorofila)

Procedemos a hacer el análisis de componentes principales. Agregamos scale = True para normalizar los datos ya que hay variables con valores muy altos mientras que otras tienen valores muy pequeños.

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.3154 1.1199 1.1030 0.95364 0.76904 0.55059 0.45468 0.29465 0.22724
Proportion of Variance 0.5361 0.1254 0.1217 0.09094 0.05914 0.03031 0.02067 0.00868 0.00516
Cumulative Proportion  0.5361 0.6615 0.7832 0.87416 0.93330 0.96362 0.98429 0.99297 0.99814
                          PC10
Standard deviation     0.13652
Proportion of Variance 0.00186
Cumulative Proportion  1.00000
```





En el análisis de componentes podemos ver que el 87.4% de la varianza se ve explicado en 4 componentes. Vemos que el componente 2 tiene una relación positiva con las variables, mientras que el componente tiene alrededor de la mitad del lado positivo y negativo.

```
Rotation (n x k) = (10 x 10):
             PC1         PC2        PC3          PC4
X3    0.35065869  0.21691594 -0.3472906  0.009131194
X4    0.33700381  0.21940887 -0.2360975 -0.017242162
X5    0.28168286  0.26250672 -0.5113780  0.146950070
X6    0.28334182 -0.10195058 -0.2639612 -0.432676049
X7   -0.39830786  0.12104244 -0.2996635 -0.080630070
X8   -0.02667579  0.57556151  0.3050633 -0.692854505
X9   -0.36839224  0.04432459 -0.3876861  0.044658983
X10  -0.37893835  0.14237181 -0.2024901 -0.167921215
X11  -0.40206100  0.05279514 -0.2562319 -0.042242268
X12  -0.05931430  0.67421026  0.2294446  0.521815581
```

Las variables que más se ven presentes en estos componentes son:

- PC1:
  - X3: Alcalinidad
  - X7: Concentración de mercurio
  - X9: Mínimo concentración de mercurio
  - X10: Máxima concentración de mercurio
  - X11: Estimación concentración de mercurio
- PC2:
  - X3: Alcalinidad
  - X4: PH
  - X5: Calcio
  - X8: Número de peces estudiados
  - X12: Edad del pescado
- PC3:
  - X3: Alcalinidad
  - X5: Calcio
  - X8: Numero de peces estudiados
  - X9: Mínimo de concentración de mercurio
- PC4:
  - X5: Calcio
  - X6: Clorofila
  - X8: Número de peces estudiados
  - X12: Edad del pescado

Conclusión

Las variables que parecen afectar más la concentración de mercurio que vemos en los peces son la concentración mínima, la concentración máxima y la estimación de mercurio. No obstante, estas están muy relacionadas con la variable que queremos predecir por lo que el análisis está muy sesgado. Las variables que mayor relación tienen sin contar las antes mencionadas son el calcio, la alcalinidad y el PH.

Link a carpeta de drive:

https://drive.google.com/drive/folders/18YboF81-MMl1feH0YOaTYkiiHvsZ-RyQ?usp=share_link