Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 502)

Módulo 5: Estadística Avanzada para ciencia de datos y nombre de la concentración.

Reporte final de "Los peces y el mercurio"

Angel Corrales Sotelo | A01562052

Diciembre 2022

Resumen

El problema que se enfrenta es el de identificar los factores que influyen en la contaminación de mercurio en los peces utilizando los datos de lagos de Florida.

En cuanto a los métodos estadísticos, se realizó una análisis de los datos mediante pruebas de normalidad, sesgo, curtosis, así como distribución normal multivariada y análisis de componentes principales.

Se encontró que el mejor modelo para predecir la concentración media de mercurio en un lago es mediante los atributos alcalinidad y clorofila.

Introducción

En un mundo tan globalizado es complicado saber el origen de cada una de los alimentos que se consumen lo cual puede llegar a ser peligroso si a estos no se les da el cuidado necesario o provienen de lugares contaminados, como en este caso, en el que se encontró que se espera conocer los factores que influyen en el nivel de contaminación por mercurio de peces en los lagos de Florida.

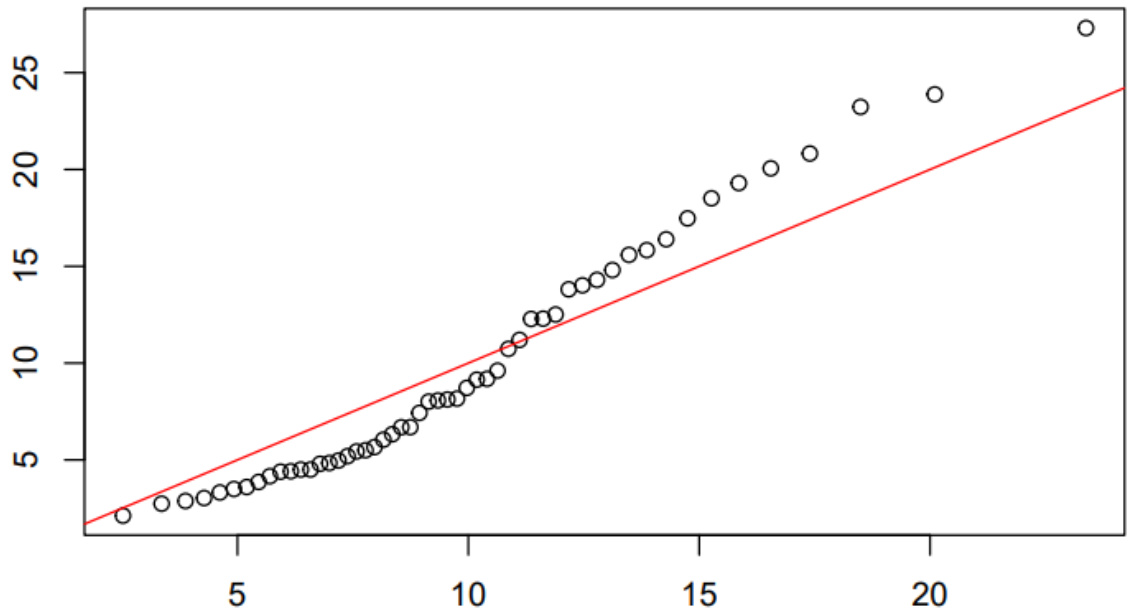
Esto es importante pues dependiendo de cuáles sean los factores que contribuyan a esta contaminación, las medidas a realizar podrían ser totalmente opuestas.

En la etapa anterior se determinaron los factores alcalinidad y clorofila como los más influyentes en la contaminación y se realizó un modelo para predecirlo, no obstante, en esta ocasión se realizaron distintos acercamientos para analizar los datos al agregar el análisis de la normal multivariada y componentes principales, de manera que es posible encontrar distintos y tal vez mejores resultados a los obtenidos anteriormente, a continuación se muestra lo logrado.

Análisis de los resultados

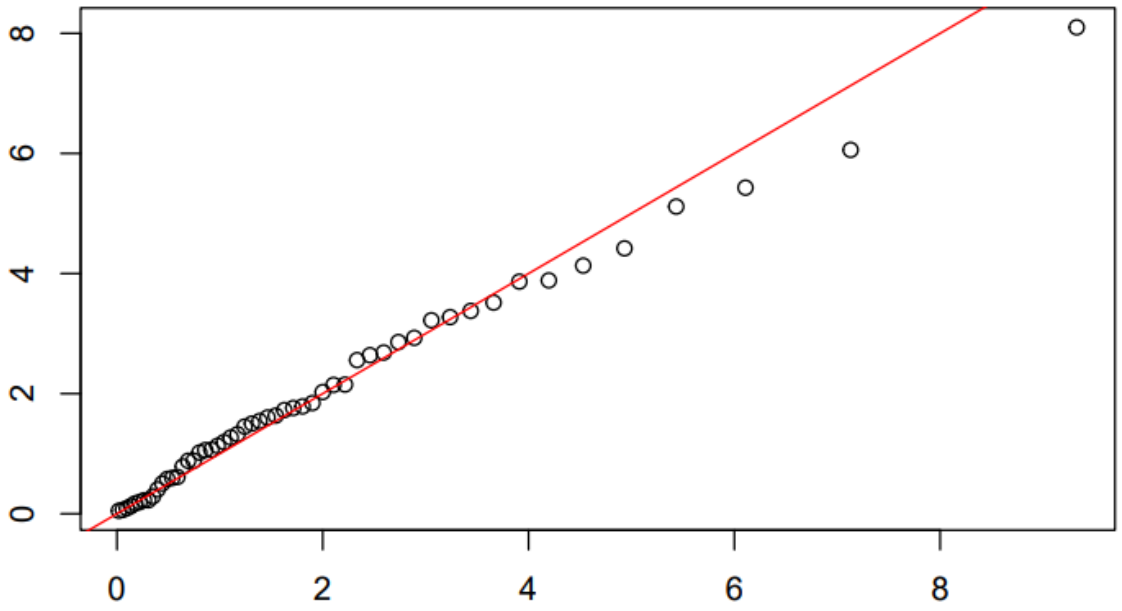
Prueba de normalidad

Mediante la prueba de normalidad de Mardia y la prueba de Anderson Darling se determinó que únicamente las variables X4 (PH) y X10 (máximo de la concentración de mercurio en cada grupo de peces) se comportan de manera normal y los datos mostraban curtosis y sesgo como puede observarse en la siguiente gráfica.




```
$multivariateNormality
             Test         Statistic               p value Result
1 Mardia Skewness 474.747945136974 8.64265750182826e-21     NO
2 Mardia Kurtosis 3.59794900484947 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
```

Sin embargo, al utilizar solo las variables anteriormente mencionadas que sí se comportan de forma normal, las pruebas indican que ya no hay curtosis ni sesgo, lo cual se puede observar en siguiente gráfica.




```
$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
```

Datos atípicos

Para encontrar los datos atípicos en la normal multivariada con las dos variables anteriores se encontraron las siguientes distancias de Mahalanobis

```
 [1] 1.19340732 3.86623312 4.13223576 0.06401297 2.55854258 0.59679099
 [7] 1.78765285 1.50256004 1.01802464 1.63445643 1.26957081 0.22408796
[13] 0.61327299 5.43003539 2.14558317 2.15128332 2.68426029 5.11541618
[19] 0.78631278 1.84332204 3.88561874 0.04988963 0.88374242 8.09988683
[25] 2.02671417 0.58001802 0.50566324 2.64029211 2.92897033 1.06510339
[31] 0.88216107 0.40467236 6.05908470 3.51592769 1.45168260 1.32485980
[37] 1.13164718 2.85893264 0.22445752 3.37787450 1.54397421 1.72672287
[43] 0.28960341 1.76103444 1.60277372 0.09262808 0.19257450 3.22197239
[49] 3.27393627 0.17321589 0.12786732 4.41942776 1.06000856
```

Teniendo que (X - miu)' Eˆ-1 (X - miu) <= Xpˆ2(alpha). Todas aquellas distancias de mahalanobis que sean menor a Xpˆ2(alpha) caen dentro del contorno de probabilidad estimado del 99.73% de una distribución normal bivariada

Dado que X_2ˆ2(0.9973) = 11.82901 Todas las observaciones caen dentro del contorno de probabilidad estimado, el cual fue elegido para determinar los valores atípicos como los

valores alejados a 3 o más desviaciones estándar de la media. Por lo tanto podemos concluir que no hay valores atípicos así como no hay valores atípicos influyentes.

Análisis de componentes principales

Debido a que la correlación entre las variables es considerablemente alta (la cual se puede observar en la tabla de abajo) sería conveniente reducir la dimensionalidad encontrando los componentes principales.

	X3	X4	X5	X6	X7	X8
X3	1.00000000	0.71916568	0.832604192	0.47753085	-0.59389671	0.01029074
X4	0.71916568	1.00000000	0.577132721	0.60848276	-0.57540012	-0.01860607
X5	0.83260419	0.57713272	1.000000000	0.40991385	-0.40067958	-0.08937901
X6	0.47753085	0.60848276	0.409913846	1.00000000	-0.49137481	-0.01182027
X7	-0.59389671	-0.57540012	-0.400679584	-0.49137481	1.00000000	0.07903426
X8	0.01029074	-0.01860607	-0.089379013	-0.01182027	0.07903426	1.00000000
X9	-0.52535654	-0.54196524	-0.332476229	-0.40045856	0.92720506	-0.08165278
X10	-0.60479558	-0.55181523	-0.407916635	-0.48497215	0.91586397	0.16109174
X11	-0.62795845	-0.61284905	-0.464409465	-0.50644193	0.95921481	0.02580046
X12	-0.09493882	0.03800021	-0.002111124	-0.28300234	0.10873896	0.20795617

	X9	X10	X11	X12
X3	-0.52535654	-0.60479558	-0.62795845	-0.094938825
X4	-0.54196524	-0.55181523	-0.61284905	0.038000214
X5	-0.33247623	-0.40791663	-0.46440947	-0.002111124
X6	-0.40045856	-0.48497215	-0.50644193	-0.283002338
X7	0.92720506	0.91586397	0.95921481	0.108738958
X8	-0.08165278	0.16109174	0.02580046	0.207956171
X9	1.00000000	0.76535319	0.91908939	0.100661967
X10	0.76535319	1.00000000	0.85975810	0.093752072
X11	0.91908939	0.85975810	1.00000000	0.089411267
X12	0.10066197	0.09375207	0.08941127	1.000000000

Es importante mencionar que este análisis se realizó con la matriz de correlación en lugar de la matriz de varianza-covarianza debido a que los datos tienen rangos diferentes de valores y necesitan ser escalados.

Para este análisis es necesario obtener los valores propios y vectores propios, los cuales se muestran a continuación.

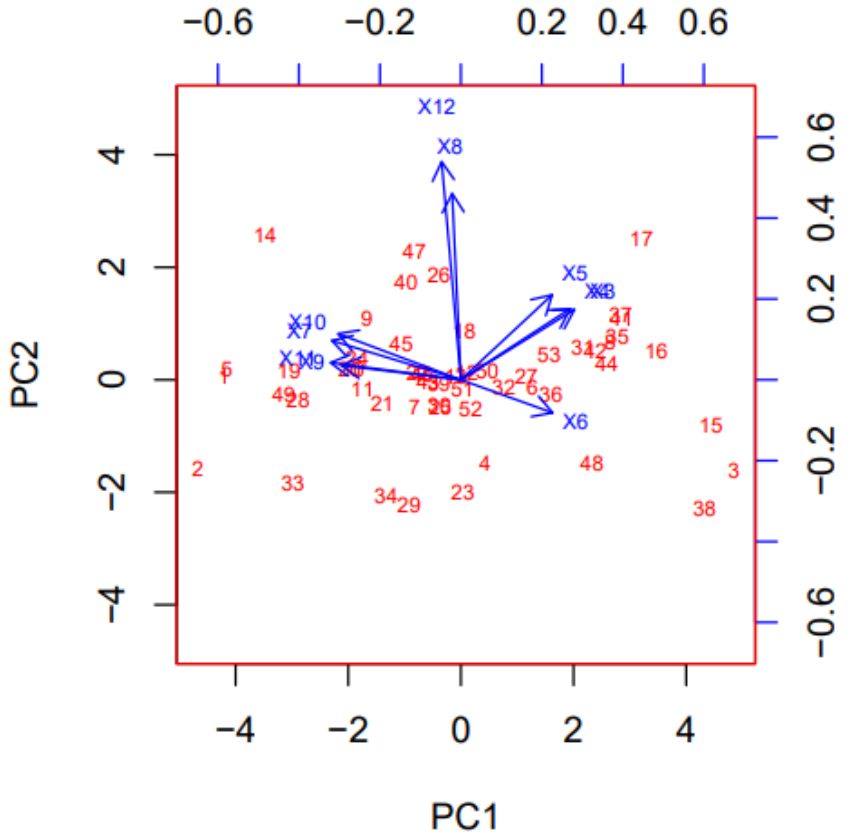
```
$values
 [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741
 [7] 0.20673634 0.08682133 0.05163902 0.01863699

$vectors
             [,1]        [,2]       [,3]         [,4]        [,5]        [,6]
 [1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534 -0.07547497


 [2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038 -0.73121012
 [3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937  0.31342329
 [4,] -0.28334182  0.10195058 -0.2639612 -0.432676049 -0.63093376  0.44112169
 [5,]  0.39830786 -0.12104244 -0.2996635 -0.080630070 -0.03046869 -0.07436922
 [6,]  0.02667579 -0.57556151  0.3050633 -0.692854505  0.19646415  0.05926732
 [7,]  0.36839224 -0.04432459 -0.3876861  0.044658983 -0.13236038  0.19602465
 [8,]  0.37893835 -0.14237181 -0.2024901 -0.167921215  0.02678086 -0.26671839
 [9,]  0.40206100 -0.05279514 -0.2562319 -0.042242268 -0.05607416 -0.03863899
[10,]  0.05931430 -0.67421026  0.2294446  0.521815581 -0.37253140  0.21612970
             [,7]        [,8]        [,9]       [,10]
 [1,] -0.33823501  0.68622998  0.04284021  0.02239801
 [2,] -0.08629646 -0.28769221  0.01363551 -0.04445261
 [3,]  0.34312185 -0.45568753 -0.11508339 -0.02634676
 [4,]  0.13435159  0.19006976 -0.06333133  0.03982419
 [5,] -0.01377825 -0.01674789  0.06243320  0.84827636
 [6,] -0.14693148 -0.16809481  0.02532023 -0.04805976
 [7,] -0.45674057 -0.18260535  0.53803577 -0.35020485
 [8,]  0.67376588  0.33602914  0.18844932 -0.30445219
 [9,] -0.23387764  0.02613406 -0.80648296 -0.24018040
[10,]  0.05759514  0.16451240 -0.02782678  0.01839703
```

Tras calcular las proporciones acumuladas, podemos concluir que la cantidad apropiada de componentes principales a utilizar es 5, la cual corresponde a una proporción de variable explicada de 0.933 como se muestra en la tabla siguiente.

	values
1	0.5361226
2	0.6615488
3	0.7832169
4	0.8741602
5	0.9333019
6	0.9636166
7	0.9842903
8	0.9929724
9	0.9981363
10	1.0000000

Resultados

El hecho de que se hayan tenido que utilizar hasta 5 componentes principales y al observar los valores correspondientes a cada variable muestra que las variables no muestran una importancia muy dominante ante las demás, siendo todas más o menos relevantes de manera intercambiable entre dichos componentes. Esto puede observarse en la gráfica siguiente en la cual no se logra ver una diferencia considerable entre las variables.



Conclusión

En cuanto a lo que esto significa para el problema en cuestión, debido a que los componentes principales no mostraron significancias destacables entre los variables, no es información significativa para determinar mejor los factores que influyen en la contaminación por mercurio de peces de los lagos de Florida. Por lo cual los factores importantes tomando en cuenta la etapa anterior siguen siendo la clorofila y la alcalinidad aun cuando en principio no se comporten de manera normal.

Anexos

https://github.com/AngelCorso/Reporte-final-de-Los-peces-y-el-mercurio