Reporte “Los peces y el mercurio”

Inteligencia artificial avanzada para ciencia de datos II

Modulo 5: Estadística Avanzada para ciencia de datos

Grupo 502

Luis Gabriel Martínez Rentería

A01651812

Fecha de entrega: 04 / 12 / 22

Resumen

Uno de los principales problemas con la contaminación de los lagos y los ríos es la afectación que esto lleva a nuestra salud, ya que los animales que viven en estos lagos también se ven afectados por esta contaminación y la ingesta de estos animales puede llevar a intoxicaciones. En este reporte se analizarán los datos de un estudio hecho por investigadores de Florida, en el cual se midieron los niveles de mercurio de 53 lagos de la zona para examinar las características del agua. Mediante análisis de normalidad como de componentes principales se identificarán los factores principales en la contaminación de estos lagos y de las afectaciones que estos tienen en los peces.

Introducción

Este problema es uno de los más importantes a atender ya que el pescado es una de las principales fuentes de alimento de Florida, y al estar expuestos a la contaminación, las afecciones en la salud que estos pueden tener pueden llegar a ser muy graves. Actualmente existen diversos estudios relacionados a los riesgos asociados a la ingesta de los contaminantes del pescado, además de distintas regulaciones por parte de los países que dictaminan el límite máximo permitido de concentración de mercurio en pescados y mariscos, variando desde 0,5 hasta 1,5 ppm. Tomando en cuenta lo mencionado, usaremos herramientas de estadística para identificar las variables principales que intervienen en la contaminación por mercurio de los peces para los diferentes lagos de Florida.

Análisis de los resultados

Análisis de normalidad

Se realizará un análisis de normalidad, con el objetivo de analizar cuánto es que difiere la distribución de los datos observados respecto a lo esperado.

H0: Los datos siguen una distribución normal

H1: Los datos no siguen una distribución normal

Con un nivel de significancia de a = 0.05

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	523.558480961006	3.65025368733176e-16	NO
Mardia Kurtosis	2.73381287599958	0.00626056123016561	NO
MVN	*NA*	*NA*	NO

Podemos observar que ninguna de las pruebas indica normalidad multivariante, por lo que los datos no siguen una distribución normal.

Se hace una segunda prueba de normalidad univariante, donde

H0: los datos siguen una distribución normal

H1: los datos no siguen una distribución normal

Esta se hace con una prueba de Anderson-Darling

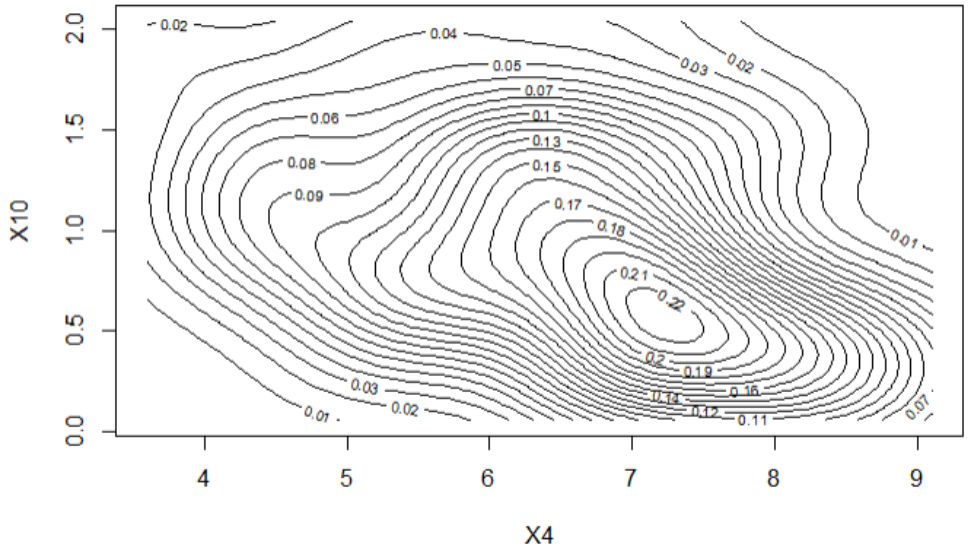
```
        Anderson-Darling normality test

data:  M$X3
A = 3.6725, p-value = 2.706e-09


        Anderson-Darling normality test

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


        Anderson-Darling normality test

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


        Anderson-Darling normality test

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


        Anderson-Darling normality test

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

```
        Anderson-Darling normality test

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


        Anderson-Darling normality test

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


        Anderson-Darling normality test

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


        Anderson-Darling normality test

data:  M$X11
A = 1.0469, p-value = 0.008637


        Anderson-Darling normality test

data:  M$X12
A = 14.335, p-value < 2.2e-16
```

Prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores. Interpreta los resultados obtenidos con base en ambas pruebas y en la interpretación del sesgo y la curtosis de cada una de ellas.

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

Gráfica de contorno de la normal multivariada obtenida en el inciso B



Como podemos observar, hay una relación entre el pH y la concentración máxima de mercurio en los peces, ya que estas dos son las únicas variables en la base de datos que se comportan de manera normal y en la gráfica superior se puede ver que las líneas de contorno se encuentran atravesando la diagonal.

Análisis de componentes principales

Justifique por qué es adecuado el uso de componentes principales para analizar la base (haz uso de la matriz de correlaciones)

El análisis de componentes principales nos ayudará a reducir la cantidad de variables, como para también ver cuales son las variables que influyen más en la varianza de los datos.

Matriz de correlación:

	X1	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	1.000	0.104	0.041	0.105	-0.082	-0.277	0.030	-0.269	-0.212	-0.304	0.132
X3	0.104	1.000	0.719	0.833	0.478	-0.594	0.010	-0.525	-0.605	-0.628	-0.095
X4	0.041	0.719	1.000	0.577	0.608	-0.575	-0.019	-0.542	-0.552	-0.613	0.038
X5	0.105	0.833	0.577	1.000	0.410	-0.401	-0.089	-0.332	-0.408	-0.464	-0.002
X6	-0.082	0.478	0.608	0.410	1.000	-0.491	-0.012	-0.400	-0.485	-0.506	-0.283
X7	-0.277	-0.594	-0.575	-0.401	-0.491	1.000	0.079	0.927	0.916	0.959	0.109
X8	0.030	0.010	-0.019	-0.089	-0.012	0.079	1.000	-0.082	0.161	0.026	0.208
X9	-0.269	-0.525	-0.542	-0.332	-0.400	0.927	-0.082	1.000	0.765	0.919	0.101
X10	-0.212	-0.605	-0.552	-0.408	-0.485	0.916	0.161	0.765	1.000	0.860	0.094
X11	-0.304	-0.628	-0.613	-0.464	-0.506	0.959	0.026	0.919	0.860	1.000	0.089
X12	0.132	-0.095	0.038	-0.002	-0.283	0.109	0.208	0.101	0.094	0.089	1.000

Realiza el análisis de componentes principales y justifica el número de componentes principales apropiados para reducir la dimensión de la base.

```
Standard deviations (1, .., p=11):
 [1] 2.3258356 1.1911085 1.1183794 0.9925154 0.8615791 0.7303252 0.5493805 0.4511099 0.2943002 0.2229690
[11] 0.1361971

Rotation (n x k) = (11 x 11):
            PC1         PC2           PC3          PC4         PC5         PC6          PC7         PC8
X1   0.10440468  0.5616080  0.0401873641  0.379627909 -0.58580410 -0.41715723  0.05116561  0.07265883
X3   0.34728626 -0.2041947 -0.3157647705  0.150134568 -0.20443469  0.29448933  0.06072771  0.32908620
X4   0.33254276 -0.1825284 -0.2884607573 -0.004577198  0.16265589 -0.32592230  0.73835183  0.11281481
X5   0.27922859 -0.2613924 -0.4062374898  0.356322981 -0.24719058  0.24437343 -0.30851923 -0.35025274
X6   0.27680608 -0.3662938 -0.0009503956 -0.292675297 -0.04175176 -0.68585978 -0.41574064 -0.11977427
X7  -0.39839259 -0.1843927 -0.2101943425  0.063924756 -0.13557055 -0.09254376  0.08074668  0.02339454
X8  -0.02534764  0.2867635 -0.4592764521 -0.728702504 -0.32025000  0.12839882 -0.06387412  0.14504041
X9  -0.36886370 -0.2537374 -0.1635819424  0.204639230 -0.06616171 -0.18951729 -0.19161497  0.47575294
X10 -0.37790292 -0.1017464 -0.1994670456 -0.029068627 -0.20980793 -0.07781712  0.28902758 -0.65678086
X11 -0.40256718 -0.1805635 -0.1323874530  0.065665030 -0.06330799 -0.07128056  0.03417493  0.22870674
X12 -0.05501346  0.4283867 -0.5578970001  0.196388915  0.59362665 -0.17551267 -0.22019381 -0.07127531
            PC9         PC10         PC11
X1   0.01648750 -0.049935164  0.01070449
X3   0.68817149  0.028339402  0.02372318
X4  -0.28348568  0.007897542 -0.04322898
X5  -0.46086548 -0.099426546 -0.02729800
X6   0.19170668 -0.074977159  0.04291286
X7  -0.01506893  0.067297896  0.84867397
X8  -0.16711684  0.026934290 -0.04863678
X9  -0.16731167  0.523164116 -0.35478927
X10  0.33795828  0.181810786 -0.30653074
X11  0.01536130 -0.817673655 -0.22827278
X12  0.16135528 -0.023092223  0.01758282
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7    PC8     PC9    PC10    PC11
Standard deviation     2.3258 1.1911 1.1184 0.99252 0.86158 0.73033 0.54938 0.4511 0.29430 0.22297 0.13620
Proportion of Variance 0.4918 0.1290 0.1137 0.08955 0.06748 0.04849 0.02744 0.0185 0.00787 0.00452 0.00169
Cumulative Proportion  0.4918 0.6208 0.7345 0.82401 0.89149 0.93998 0.96742 0.9859 0.99379 0.99831 1.00000
```

En la tabla de arriba podemos ver que con los primeros 9 componentes principales ya estamos describiendo el 99% de los datos, pero si solo se quisiera tener arriba del 90% de la varianza acumulada, con 6 componentes principales nos bastaría para describir los datos de manera correcta.

Conclusiones

Una vez hecho el análisis se tiene que los principales factores que influyen en la contaminación de los rios es el pH y el nivel de alcalinidad de los lagos, por lo que si en un futuro se quisiera hacer un modelo de predicción los principales factores que se tendrían que tomar en cuenta son esos. Los análisis de normalidad de las variables nos mostraron que la mayoría tienen un comportamiento no normal.

Anexo:

R markdown:

https://drive.google.com/file/d/1-aNqLMukAhGaomsytFR7BBMUnHpPRayA/view?usp=share_link

Database:

https://drive.google.com/file/d/17M1PM_zrZFGupsPWGxPEnYOh-KBsxuUy/view?usp=share_link