Nombre: **Roel Adrián De la Rosa Castillo**

Entrega: **Reporte final de "Los peces y el mercurio"**

Clase: **Inteligencia artificial avanzada para la ciencia de datos Grupo 102**

Profesora: **Blanca Rosa Ruiz Hernandez** Fecha: **18 de Septiembre de 2022**

1 Resumen

Durante este reporte se analizará un archivo csv que contiene datos acerca de la contaminación por mercurio en los peces de diferentes lagos. Se realizará un análisis de correlación para observar la relación que se tienen entre las diferentes variables. No se pretende encontrar una relación de causa-efecto, si no asegurar con un nivel de confianza si ciertas variables tienen asociación entre sí. Además de eso se realizo un Análisis de Componentes Principales (PCA) para entender, en cierta medida, que variables son las que tienen una mayor importancia al momento de tomar en cuenta la variabilidad de los datos. A partir de estos dos análisis se llego a la conclusión de que la Alcalinidad, PH, Calcio y Clorofila son los factores que más afectan a la concentración de mercrurio en los peces de estos lagos.

2 Introducción

El problema que se plantea es el siguiente:

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.

El objetivo de este reporte es utilizar diferentes métodos estadísticos para realizar análisis con los datos trabajados en este estudio y encontrar cuales son los pricipales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida. Estos análisis son importantes, pues tras el crecimiento de la industria en las últimas décadas, los niveles de contaminación por mercurio son bastante altos. Esto puede afectar negativamente a la salud de la población que consume estos peces. Se han realziado estudios y a los niños que consumen estos productos, o aquellos que estuvieron expuestos durante el embarazo pueden llegar a tener problemas en sus sistema nervioso. [1]

3 Análisis de los Resultados

3.1 Exploración de los datos

Lo primero que se realizo fue explorar y entender las variables que se tienen en el set de datos. Esto con objetivo de conocer como es que se comportan dichas variables y qué tanta utilidad pueden darnos. A primera vista podemos darnos cuenta que la variable X_1, la cual es el lago en el que se realizo el estudio, tienen valores únicos los cuales no se repiten en otros registros, por lo que podemos simplemente quitarlo del análisis para evitar ruido.

A su vez también realizamos unos gráficos para entender como es que se distribuye la concenctración de mercurio.

Por lo que se puede observar en el histograma y el boxplot, se tienen una media y medianas bastante cercanas a 0.5 mg de Hg/kg, lo cual podría indicar que no cumplen 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). Esto puede indicarnos que realmente es necesario saber cuales son los factores más relevantes sobre el mercurio en los peces.

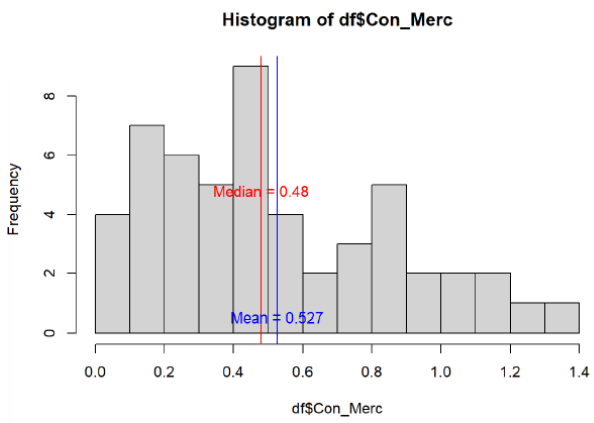

Figure 1: Histograma de la concentración de mercurio con media y mediana

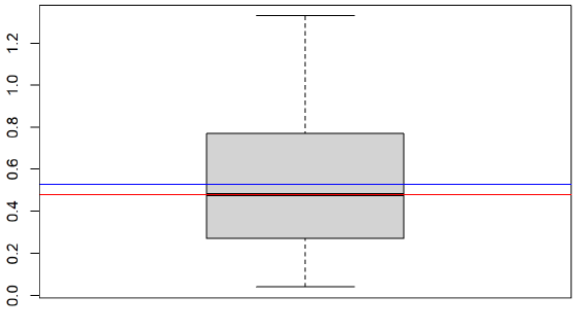

Figure 2: Boxplot de la concentración de mercurio con media y mediana

3.2 Análisis de Correlación

La correlación es una medida que toma en cuenta la covarianza lineal entre dos variables, es decir mide que tanto es que estas variables varian conjuntamente. Esto podemos utilizarlo para entender que variables tienen una relación entre sí, teniendo como objetivo saber cuales son las que pueden afectar o se ven afectadas por el nivel de mercurio en los peces.

A partir de lo anterior se puedeobservar en el primer heatmap como es que las variables tienen correlación entre sí. Para poder entender mejor la relación entre las variablesse han calulado los p-values entre las correlaciones de todas las variables en el segundo heatmap. Después se han usado pruebas dehipótesis para ver si las variables tienen independencia o asociación entre sí.

Figure 3: Matriz de correlación entre todas las variables

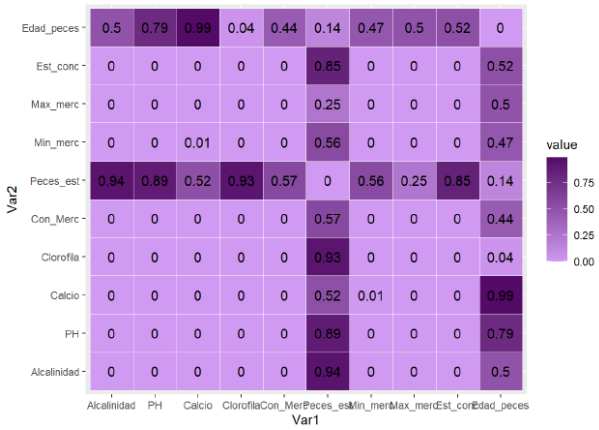

Figure 4: Matriz con el p-value de la correlación entre todas las variables

$$\begin{aligned} H_0 &: \rho = 0 \text{ El caso de independencia entre las variables} \\ H_1 &: \rho \neq 0 \text{ El caso de asociación entre variables} \end{aligned} \tag{1}$$

$$\textbf{Regla de decisión: Rechazar} H_0 \textbf{ si el p-value} < 0.05 \tag{2}$$

Dado que queremos ver que variables afectan a la concentración de mercurio, podemos observar que Alcalinidad, PH, Calcio, Clorofila, el mínimo de mercurio, el máximo de mercurio y la estimación de mercurio tienen un p-value menor a 0.05, por lo que se rechaza H_0 y se llega a la conclusión que estas variables son las que pueden afectar a la concentración de mercurio. [2]

3.3 Análisis de Componentes Principales (PCA)

Antes de hacer el PCA debemos de saber si es necesario escalar los datos. Para poder saber esto primero vamos a ver las distribuciones delas variables.

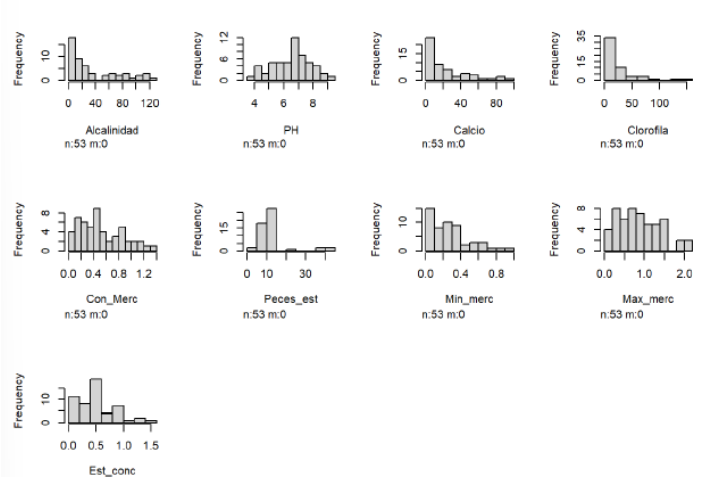

Figure 5: Distribución de las variables

Podemos observar que algunas variables que tienen valores pequeños, mientras que algunas otras tienen valores relativamente mucho másgrandes, por ello vamos a buscar normalizar los datos para que cuando se realize el PCA no se tenga algún sesgo hacia las variables con valores más altos.

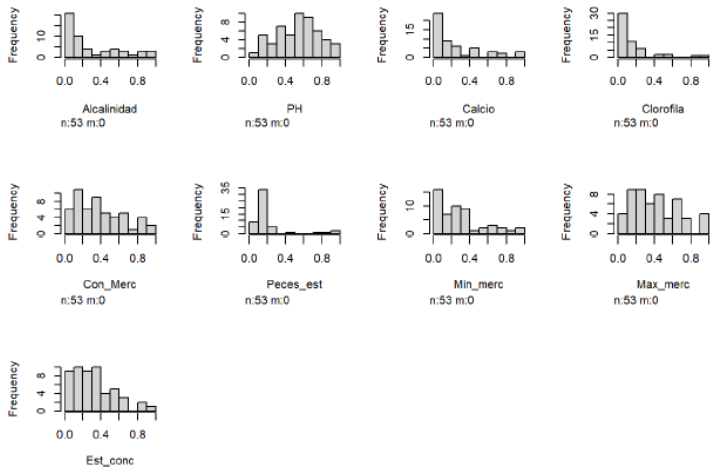

Figure 6: Distribución de las variables normalizadas

Ya que tenemos los datos normalizados, podemos aplicar el PCA para ver cuantos

componentes principales representan la mayoría de lavariabilidad de los datos y que variables son las que tienen mayor relevancia en esos componentes.

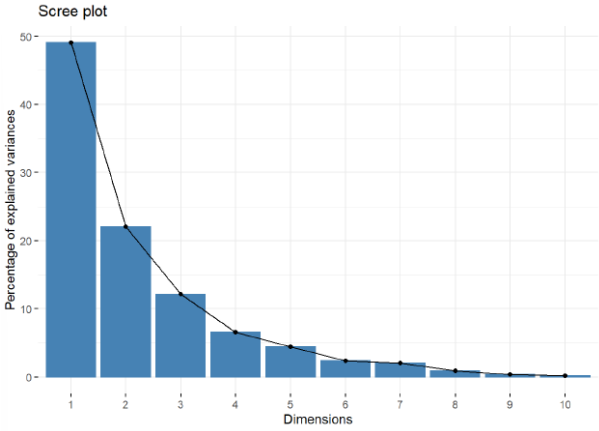

Figure 7: Variabilidad de los datos por número de componente principal

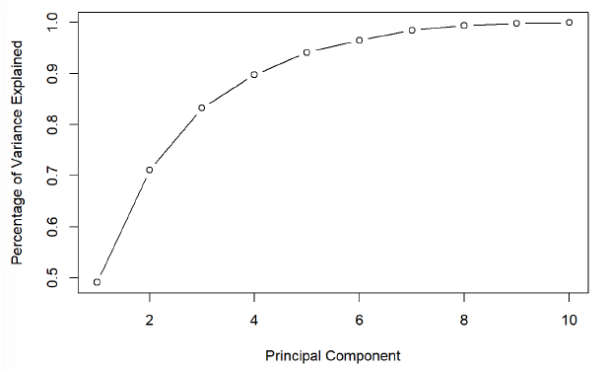

Figure 8: Variabilidad de los datos acumulada por número de componente principal

Cómo se puede observar con 4 componentes principales se tiene cerca del 90% de la variabilidad de los datos. Ahora solo falta observar esos componentes principales.

Cómo se puede observar en la tabla anterior:

- El primer componente principal se ve afectado en su mayoria por Alcalinidad, PH, Calcio, Clorofila, la concentración de mercurio, el minimo demercurio, el máximo de mercurio y la estimación de la concentración.
- El segundo componente principal se ve afectado principalmente por la variable que dice si un pez es joven o adulto

```
## Standard deviations (1, .., p=10):
##  [1] 0.59581904 0.39917379 0.29659562 0.21724577 0.17846501 0.13020889
##  [7] 0.12054968 0.08034046 0.05387504 0.03552301
##
## Rotation (n x k) = (10 x 10):
##                     PC1         PC2         PC3          PC4         PC5
## Alcalinidad  0.42599010  0.07110301 -0.45380835 -0.092834387  0.22948361
## PH           0.30061298  0.11482867 -0.17870490 -0.069014050 -0.55950798
## Calcio       0.31893493  0.11188130 -0.59403512  0.099361373  0.28519334
## Clorofila    0.21582602 -0.09664014 -0.08869493 -0.142529907 -0.70038086
## Con_Merc    -0.40348016 -0.02205194 -0.34761399 -0.065785424 -0.07823426
## Peces_est   -0.02601474  0.14242070  0.04076418 -0.928309977  0.12533319
## Min_merc    -0.36228310 -0.02807425 -0.37522624  0.161153945 -0.16959234
## Max_merc    -0.38487657 -0.02403028 -0.27289223 -0.221215975 -0.01713438
## Est_conc    -0.34987337 -0.03625495 -0.24119831  0.008891634 -0.06986001
## Edad_peces  -0.12083256  0.96770451  0.07375163  0.123927733 -0.08157767
##                     PC6         PC7         PC8          PC9         PC10
## Alcalinidad -0.17189310  0.43307746  0.57319435 -0.01376188  0.02023250
## PH          -0.63015912  0.03565135 -0.38101324 -0.02411459 -0.05073958
## Calcio       0.29582901 -0.42882768 -0.40366493  0.08098410 -0.02767055
## Clorofila    0.54136410 -0.15098977  0.31542473  0.08653852  0.05010209
## Con_Merc    -0.06531699  0.04565994 -0.03511790  0.01657562  0.83520926
## Peces_est    0.14595043  0.14393496 -0.23038905 -0.03708317 -0.05862557
## Min_merc     0.24594093  0.43568870 -0.18643890 -0.52315245 -0.33715926
## Max_merc    -0.31539964 -0.58865459  0.39981950 -0.19248498 -0.29102414
## Est_conc    -0.01896973  0.21235311 -0.03264383  0.82000169 -0.30694627
## Edad_peces   0.08289019 -0.01571581  0.11766871  0.01975141  0.01188972
```

Figure 9: Componentes principales y como es que las variables lo afectan.

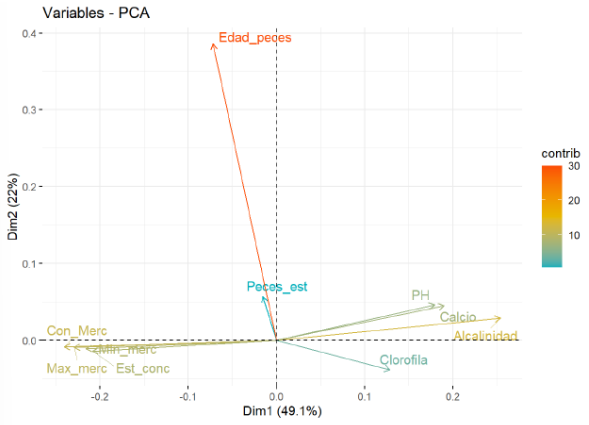

Figure 10: Loading Plot del PCA.

- El tercer componente principal tiene mayor relación con Alcalinidad, Calcio, Concentración de mercurio, el minimo de mercurio, el máximo demercurio y la estimación de la concentración.

- El cuarto componente principal se ve afectado en su mayoría por la cantidad de peces estudiados. Esto es algo que se puede ver gráficamente en el loading plot anterior. Se puede ver como es que esas variables se van agrupando.

Como se puede observar en la siguiente gráfica , parece que el pca si se ve muy sesgado por las variables que tienen mayores rangos, por lo que la normalizaión que realizamos fue un acierto.

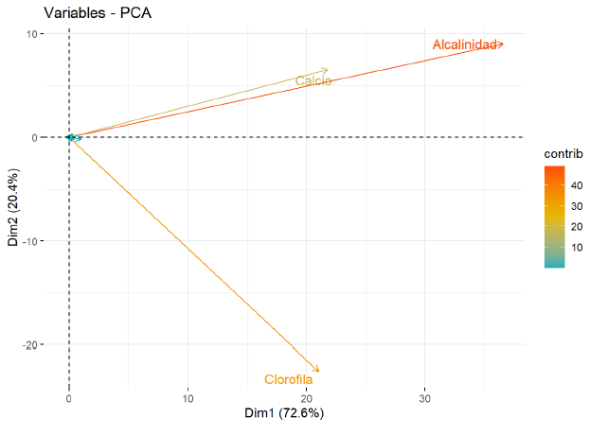

Figure 11: Loading Plot del PCA con datos no normalizados.

4 Conclusión

A partir del análisis de las correlaciones entre las variables y su significancia, además del análisis de componentes principales con los datos normalizados, se llega a la conclusión que los principales factores que influyen en el nivel de contaminación por mercurio en los peces delagos de Florida son:

Alcalinidad, PH, Calcio, Clorofila, el mínimo de mercurio, el máximo de mercurio y la estimación de mercurio. Algo que se debe de tomar en consideración es que el minimo de mercurio, el máximo de mercurio y la estimación de mercurio, por la forma en la que fueron obtenidas y calculadas, tienen bastante relación con la concentración de mercurio, por lo que considero que la Alcalinidad, el PH, el Calcio y la Clorofila son los factores más relevantes.

5 Referencias

References

[1] Mercurio en pescado Ecologistas en Acción. (2022). Retrieved 18 September 2022, from https://www.ecologistasenaccion.org/4975/mercurio-en-pescado/

[2] Correlation and P value. (2022). Retrieved 18 September 2022, from https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/

[3] What Is Principal Component Analysis (PCA) and How It Is Used?. (2022). Retrieved 18 September 2022, from https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

6 Anexos

Liga con el archivo Markdown utilizado para el análisis:
https://drive.google.com/drive/folders/1mxoJPfQE3mi-VQglX-bxEmDbivV37XTQ?usp=sharing