Procesamiento de datos multivariados

Felipe Gabriel Yépez Villacreses - A01658002

2022-10-26

Módulo 5: Estadística Avanzada para ciencia de datos y nombre de la concentración Grupo 2

Resumen

Se analizó una base de datos para comprender los niveles de concentración de mercurio en peces de ciertos lagos de Florida dado que existe un límite máximo de concentración de este componente para que puedan ser aptos para el consumo humano.

Se realizó un análisis de normalidad para detectar posible normalidad multivariada entre grupos de variables. De igual forma se realizó un análisis de componentes principales con el objetivo de identificar los factores principales que influyen en la contaminación por mercurio de los peces. De esta forma se puede ver la relación que tiene cada variable respecto a la concentración de mercurio y reducir la dimensionalidad de la base de datos a la vez que se mantiene la explicabilidad de la misma al seleccionar aquellos componentes que expliquen en mayor medida la varianza de la concentración de mercurio.

Introducción

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.

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.

Se desea encontrar cuáles son los principales factores que influyen en la concentración de mercurio en los peces de estos lagos.

Explorando la base

Lectura y resumen de los datos:

```
##       num     nombre_lago         alcalinidad          ph        
##  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
##      calcio       clorofila      concentracion_mercurio   num_peces
##  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
##  min_concentracion max_concentracion     est_3          edad_peces
##  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
```

Como se puede observar se cuenta con 12 columnas de datos. En total existen 53 registros de lagos diferentes. Se obtienen los cuartiles de los datos con su media.

Para poder analizar las variables, es necesario eliminar la columna del nombre del Lago, pues al no ser numérica no se puede utilizar para generar modelos o analizarla estadísticamente. No es relevante esta columna dado que cada lago tan solo cuenta con 1 registro que engloba y resume todos los datos recopilados del mismo.

De igual forma, no va a servir el identificador de cada registro para ningún modelo por lo que se lo da de baja del conjunto de datos.

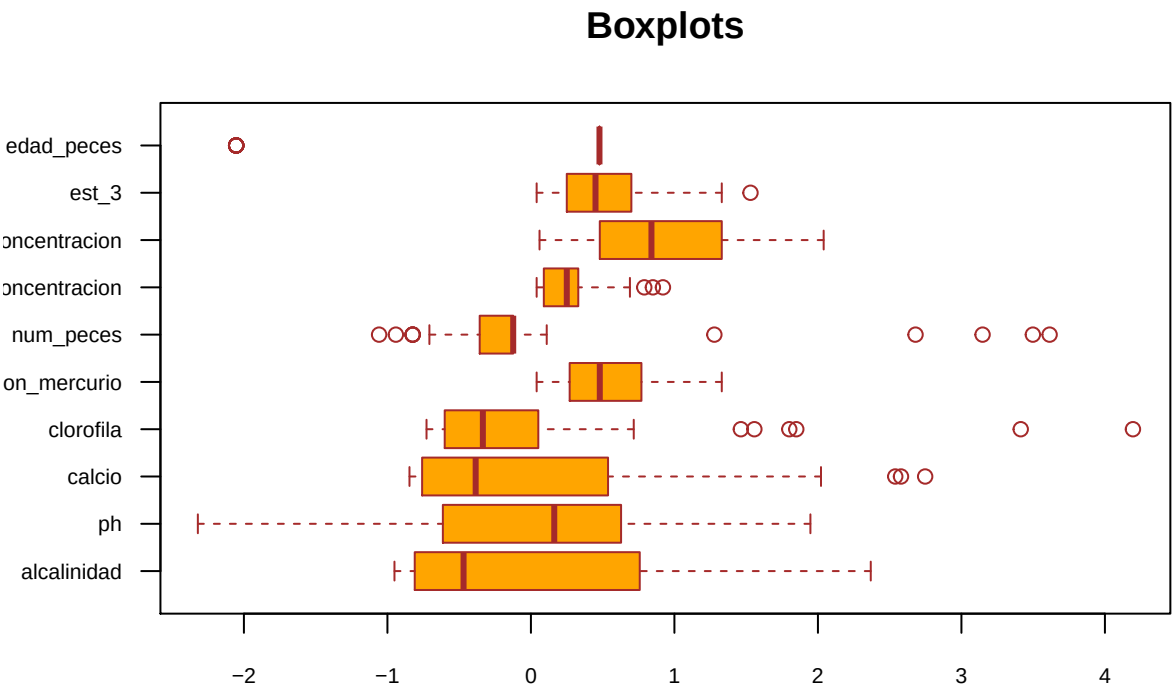
Es necesario verificar que el conjunto de datos no cuente con datos faltantes por lo que se busca valores nulos.

```
##            alcalinidad                    ph                 calcio
##                      0                     0                      0
##              clorofila concentracion_mercurio              num_peces
##                      0                     0                      0
##      min_concentracion      max_concentracion                  est_3
##                      0                     0                      0
##             edad_peces
##                      0
```

Como se puede observar, no existen valores nulos en el conjunto de datos por lo que se puede trabajar con el mismo.

Dado que se va a utilizar el dataset para realizar un análisis multivariado y entender la relación que tienen múltiples predictores con la variable objetivo, es necesario que cada variable contribuya por igual al análisis. Por esta razón es necesario escalar los datos con Z-score Standarization, la cuál se encarga de restar la media y dividir entre la desviación estándar a los datos.

Medidas de posición



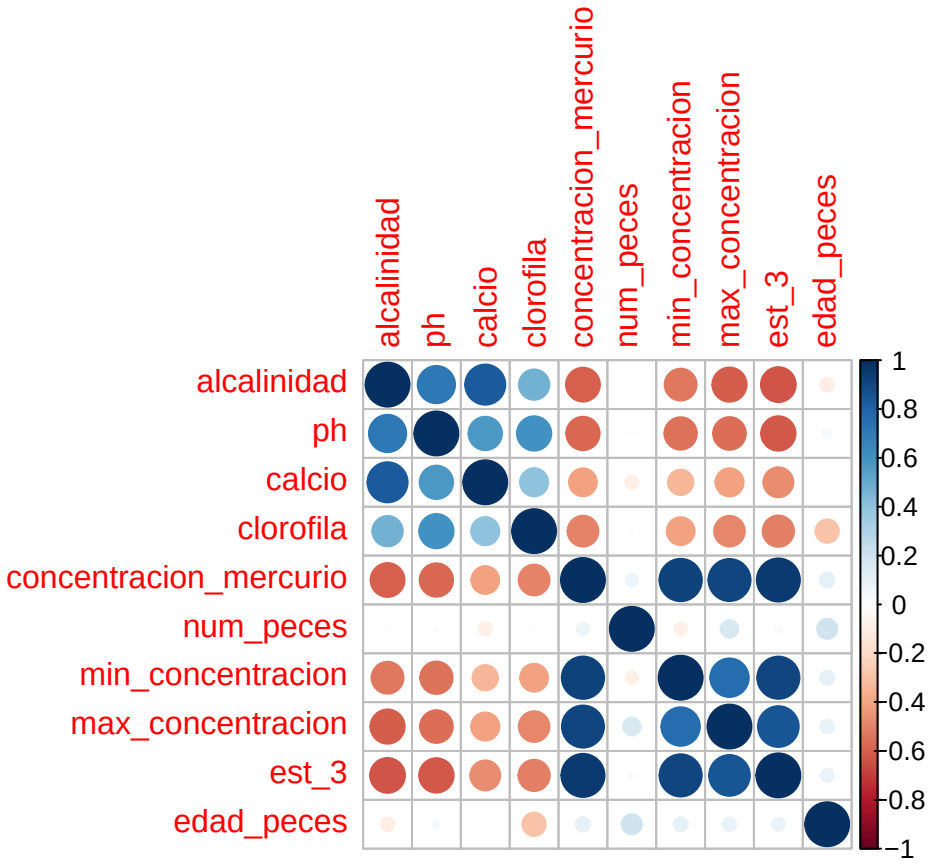
Se puede observar la distribución de cada variable. Al estar escalados se los puede observar juntos bajo una escala muy similar. En varias de las columnas existen valores outliers, pues sus valores mínimos y máximos exceden su respectivo cuartil más cercano con la diferencia de 1.5 veces el rango intercuartil.

Muchos de los datos no muestran distribución simétrica, pues unos cuartiles están más grandes que el otro y la extensión de sus barras son unas más grandes que otras contando con valores outliers algunas columnas.

Correlación

Se obtiene los valores de correlación de los datos.

```
## corrplot 0.92 loaded
```



De color azulado se pueden observar las correlaciones positivas y de color rojizo las negativas.De igual forma el diametro de cada círculo muestra el nivel de correlación entre variables.

Se puede notar que existe correlación alta entre el mínimo, promedio, maxímo y estimación de concentración de mercurio. Esto no es beneficioso para generar modelos dado que puede generarse ruido entre estas variables al utilizarlas en un modelo.

De las variables de concentración, la mínima no resulta útil para realizar un modelo ya que se estaría sesgando los valores obtenidos de cada lago a tan solo los más bajos por lo que podría resultar peligroso consumir peces de los lagos. De igual forma el valor máximo es un gran indicativo que dice que existen muy pocos lagos cuyos registros de concentración de mercurio puede llegar a ser seguro para el consumo de estos peces. Sin embargo, están los valores sesgados a tan solo los más altos. Podría ser interesante utilizar los valores máximos para asegurar que el consumo no llegue a ser peligroso pero se dejarían de lado todo el resto de registros que pudieron haber sido seguros.

Por esta razón se decide utilizar la media de concentración dado que representa de mejor forma al Lago en términos generales referente a concentración de mercurio de cada pez y de esta forma se puede encontrar los componentes que tengan mayor efecto en la concentración de mercurio en términos generales de todos los registros para cada lago.

Análisis de normalidad

Mediante el siguiente análisis de normalidad se busca identificar aquellas variables que sean normales dado que si los datos provienen de una distribución normal se podrán utilizar para realizar algún otro análisis más avanzado. Por ejemplo se los podría usar para reducir la dimensionalidad de los datos mediante PCA y con esto utiliar los componentes principales para poder generar modelos de manera más acertada a pesar de PCA no requiere que .

Normalidad de Mardia y prueba de Anderson Darling Mediante la prueba de Mardia se puede probar si existe normalidad multivariada en un conjunto de datos al revisar si su sesgo y kurtosis son consistentes con un conjunto de datos multivariados normalmente distribuido.

H_0 La muestra proviene de una distribución normal multivariante

Se utilizará un nivel de significancia de $alpha = 0.05$ al igual que como se sugiere en la academia.

```
## Warning: package 'MVN' was built under R version 4.0.5
```

```
##              Test        Statistic              p value Result
## 1 Mardia Skewness 474.747945136975 8.64265750182786e-21     NO
## 2 Mardia Kurtosis 3.59794900484948 0.000320736483631068     NO
## 3             MVN             <NA>                  <NA>     NO
```

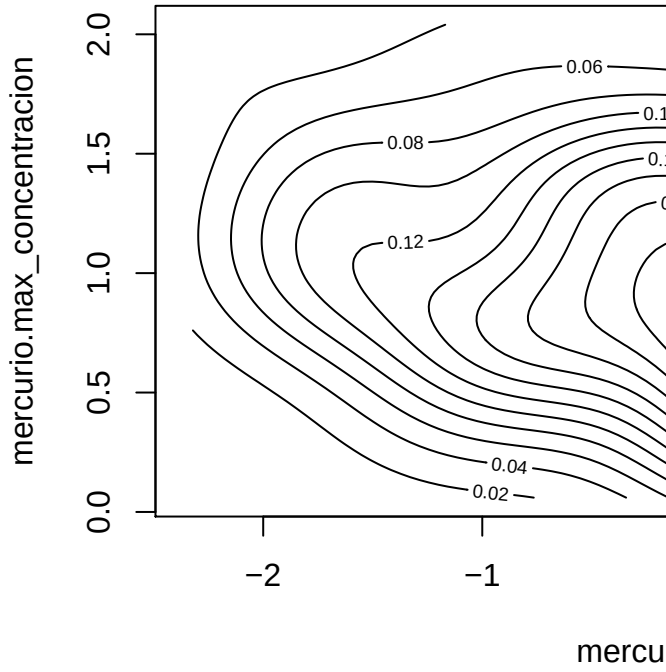
Con Mardia Skewness se obtiene un p value muy cercano a 0 y con Mardia Kurtosis se obtiene un p value de 0.00032, es decir se puede rechazar H_0 para ambas pruebas de mardia que dicen que los datos se distribuyen normalmente. De esta manera se identifica que las variables no están distribuidas normalmente dado que el valor p obtenido es menor a nuestro nivel de significancia de 0.05 en ambos casos.

Para observar el comportamiento de cada variable se realiza la prueba de Anderson Darling para probar cuáles variables se ajustan a una distribución normal.

```
##                Test                Variable Statistic   p value Normality
## 1  Anderson-Darling      alcalinidad          3.6725  <0.001      NO
## 2  Anderson-Darling           ph              0.3496  0.4611      YES
## 3  Anderson-Darling         calcio            4.0510  <0.001      NO
## 4  Anderson-Darling       clorofila           5.4286  <0.001      NO
## 5  Anderson-Darling concentracion_mercurio    0.9253  0.0174      NO
## 6  Anderson-Darling       num_peces           8.6943  <0.001      NO
## 7  Anderson-Darling   min_concentracion       1.9770  <0.001      NO
## 8  Anderson-Darling   max_concentracion       0.6585   0.081      YES
## 9  Anderson-Darling         est_3             1.0469  0.0086      NO
## 10 Anderson-Darling       edad_peces         14.3350  <0.001      NO
```

Mediante Anderson-Darling se obtiene mediante su p value bajo el mismo nivel de significancia que las variables ph y máxima concentración de mercurio si son normales mientras que el resto no lo es.

Normalidad de Mardia y prueba de Anderson Darling para variables normales Se realizará la



prueba de normalidad con ph y máxima concentración de mercurio.

```
## $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
##
## $univariateNormality
##               Test                  Variable Statistic   p value Normality
## 1 Anderson-Darling        mercurio.ph            0.3496    0.4611    YES
## 2 Anderson-Darling mercurio.max_concentracion    0.6585    0.0810    YES
##
## $Descriptives
##                               n         Mean   Std.Dev    Median       Min
## mercurio.ph                  53 4.003161e-17 1.0000000 0.1625473 -2.321058
## mercurio.max_concentracion 53 8.745283e-01 0.5220469 0.8400000  0.060000
##                                    Max       25th      75th       Skew   Kurtosis
## mercurio.ph                   1.947639 -0.6135795 0.6282234 -0.2458771 -0.6239638
## mercurio.max_concentracion 2.040000  0.4800000 1.3300000  0.4645925 -0.6692490
```

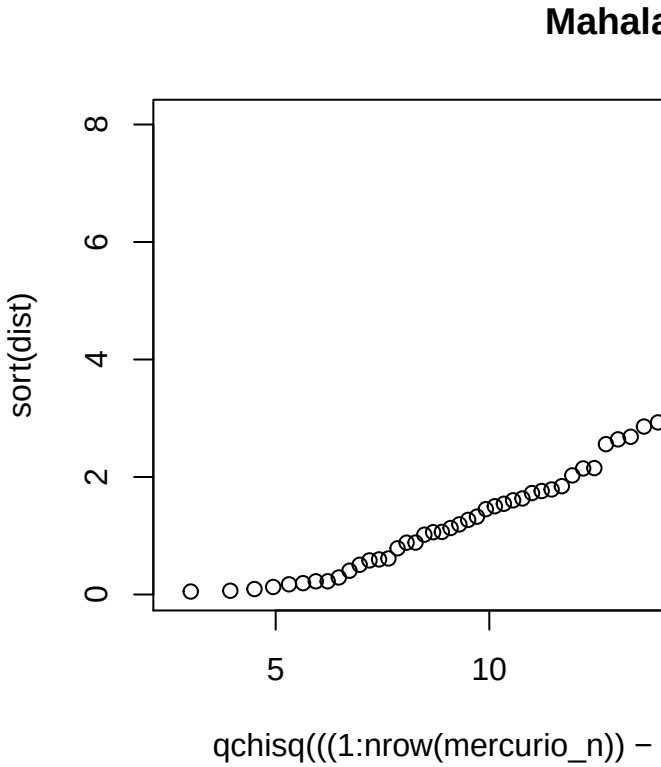
Al realizar las mismas pruebas con las variables cuya prueba de normalidad resultó positiva anteriormente con Anderson Darling, se obtiene con Mardia Skewness y Mardia Kurtosis que si provienen de una distribución normal multivariante y con Anderson-Darling también se determina que ambas variables son normales a partir del mismo nivel de significancia anteriormente elegido de $alpha = 0.5$.

La variable ph tiene -0.2458 de sesgo y -0.6239 de curtosis, un valor negativo para ambos casos. El sesgo

negativo significa que existen observaciones con valores bajos de ph en comparación con la mayor parte de las observaciones del conjunto de datos. La curtosis negativa significa que la mayor parte de las observaciones de ph se encuentran cercanas a la media del conjunto de datos, es decir menos datos están cerca a las colas.

Por otro lado, la concentración máxima de mercurio es de 0.4645 con curtosis de -0.6692. El valor de curtosis es parecido, por lo que se puede asumir lo mismo que la mayoría de los datos se encuentran cercanos a la media. Al tener sesgo positivo se puede entender que el valor de la media es mayor al de la mediana, contrario a las observaciones de ph.

En el gráfico de contorno, se observa que el máximo no es alcanzado en el origen (0,0), es decir, existe correlación entre ambas variables. De igual forma, al no ser completamente circulares los contornos se puede decir que existe correlación entre las variables.



Datos atípicos o influyentes con distancia de Mahalanobis

Mediante la distancia de Mahalanobis se puede calcular la distancia de un punto a una distribución, y en este caso se puede observar que tan solo el valor más alejado en el gráfico en la esquina superior derecha podría ser considerado un Outlier.

Análisis de Componentes Principales

Se realizará un análisis de Componentes Principales con la finalidad de conocer los factores principales que afectan la concentración de mercurio en los Lagos de Estados Unidos.

Al generar componentes principales se evaluará las variables que los conformen y en qué medida para poder determinar lo que influye en la concentración de mercurio.

De esta forma mediante un análisis de componentes principales se puede rápidamente detectar de todas las variables, cuáles son las más influyentes al seleccionar los componentes principales que logren explicar la varianza de los datos.

Matriz de correlaciones

Para realizar el análisis de componentes principales es necesario obtener la matriz de correlaciones con la finalidad de que la escala de los datos no influya en el análisis.

Componentes Principales

Se obtienen los eigen values y eigen vectors a partir de la matriz de correlación para poder realizar PCA.

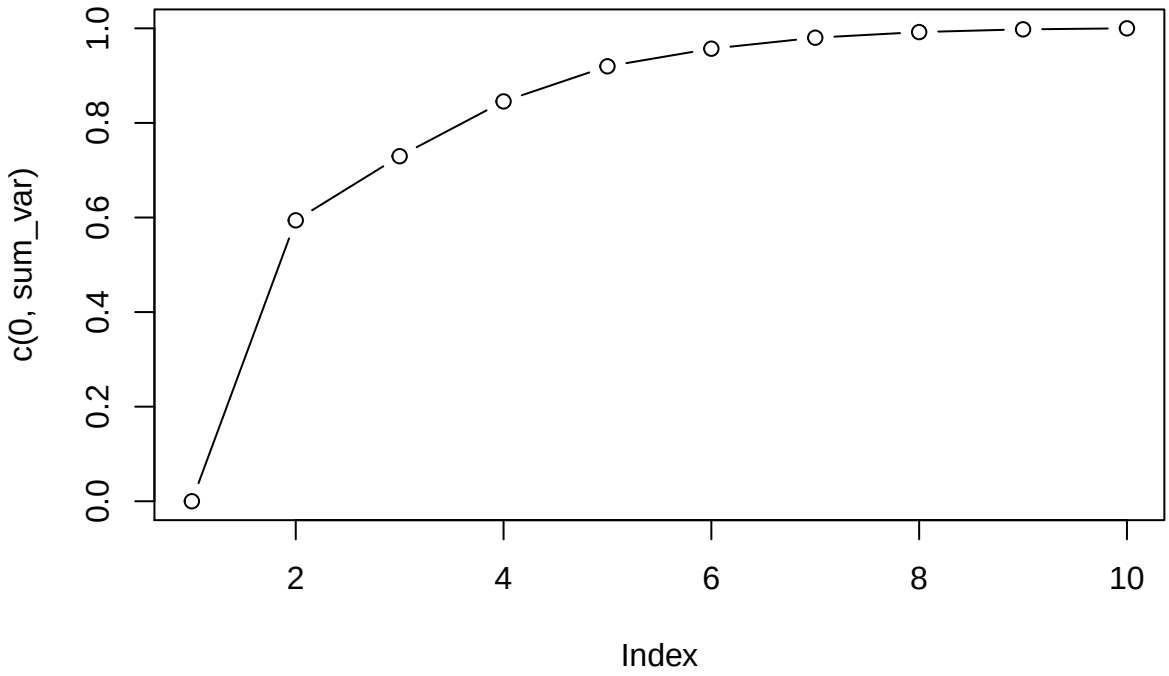
```
## eigen() decomposition
## $values
## [1] 5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403
## [8] 0.05203127 0.01885332
##
## $vectors
##              [,1]        [,2]        [,3]        [,4]        [,5]         [,6]
##  [1,] -0.35136146 -0.40301855 -0.07586402  0.30359419  0.03194121  0.284360283
##  [2,] -0.33907420 -0.29786166 -0.07470140 -0.23236707 -0.82623084  0.054271109
##  [3,] -0.28306469 -0.56943030  0.02991336  0.37427137  0.32816132 -0.298278080
##  [4,] -0.28126962 -0.21524882 -0.06147214 -0.83056128  0.39488490 -0.099142969
##  [5,]  0.39890941 -0.32518645 -0.05648045 -0.04980219 -0.06539303  0.004765464
##  [6,]  0.02398876  0.06261499 -0.96994179  0.05149024  0.09004998  0.149954574
##  [7,]  0.36905050 -0.37647100  0.11743644 -0.11401063  0.10565624  0.489107573
##  [8,]  0.37957032 -0.24428857 -0.16175615 -0.02767633 -0.16523448 -0.711214479
##  [9,]  0.40293860 -0.25922456  0.00756517 -0.07091614 -0.04298253  0.223233955
##              [,7]         [,8]         [,9]
##  [1,]  0.72620919 -0.082971700  0.007161703
##  [2,] -0.22348526  0.009782475 -0.032988603
##  [3,] -0.48766992  0.140957430 -0.017292418
##  [4,]  0.11144724  0.043959526  0.028777382
##  [5,]  0.01398475 -0.053416125  0.849768758
##  [6,] -0.14013431 -0.011952152 -0.041106334
##  [7,] -0.22360542 -0.528271290 -0.340326567
##  [8,]  0.30736177 -0.211913074 -0.311145559
##  [9,]  0.09015694  0.802648566 -0.247594211
```

Variables que más contribuyen

Se puede evidenciar en este caso que el Componente Principal 1 está conformado gran parte por las variables 5 y 9 mientras que el segundo componente principal está conformado mayormente por las variables 6 y 10.

Varianza explicada acumulada de los componentes principales

```
## [1] 0.5939898 0.7296462 0.8454831 0.9196901 0.9569915 0.9802068 0.9921239
## [8] 0.9979052 1.0000000
```



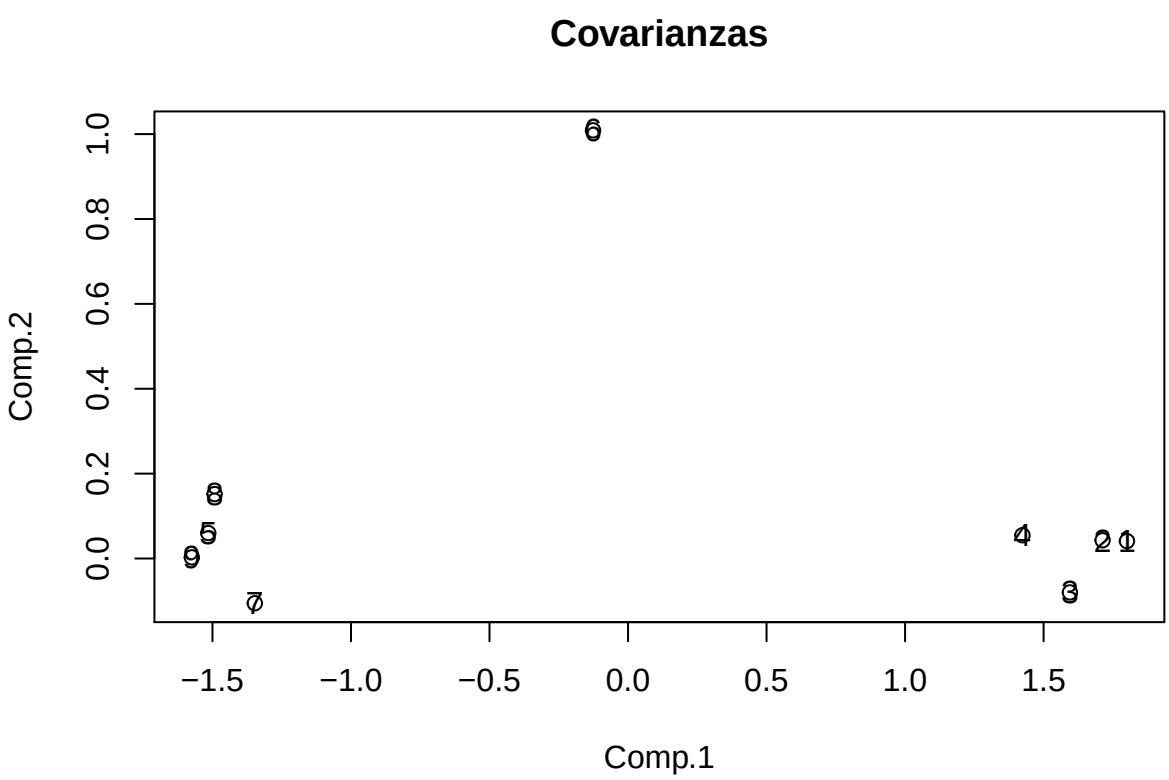
Para lograr explicar al menor el 90% de la varianza se deberían utilizar al menos los 4 primeros componentes principales por lo que se puede evidenciar que si se quisiera usar los componentes principales para realizar algún análisis posterior, no sería de gran utilidad ya que no se logró reducir en gran medida la dimensionalidad de los datos. Al reducir la dimensionalidad del conjunto de datos se pierde cierto nivel de interpretabilidad e información del conjunto inicial de datos.

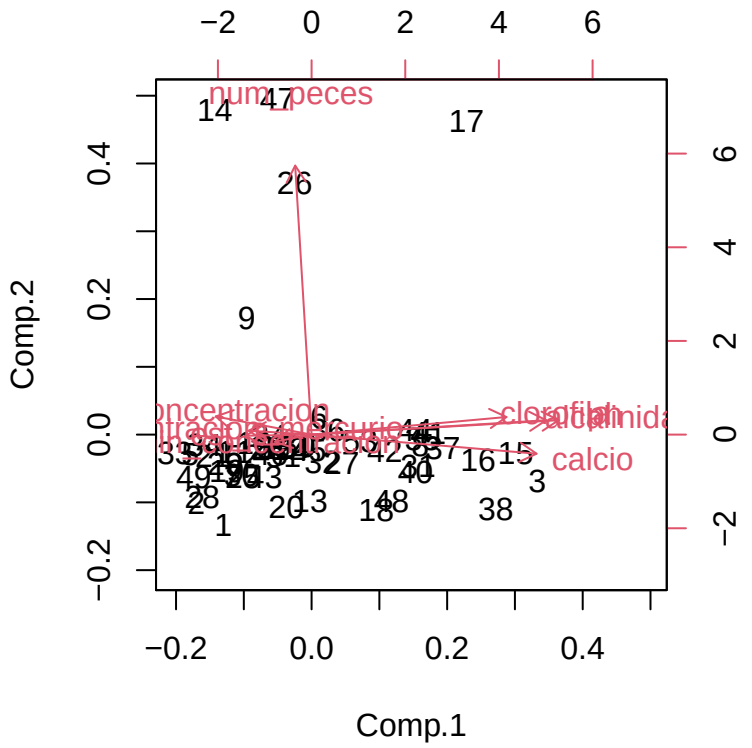
```
## Warning: package 'factoextra' was built under R version 4.0.5
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
## Warning: package 'FactoMineR' was built under R version 4.0.5
```

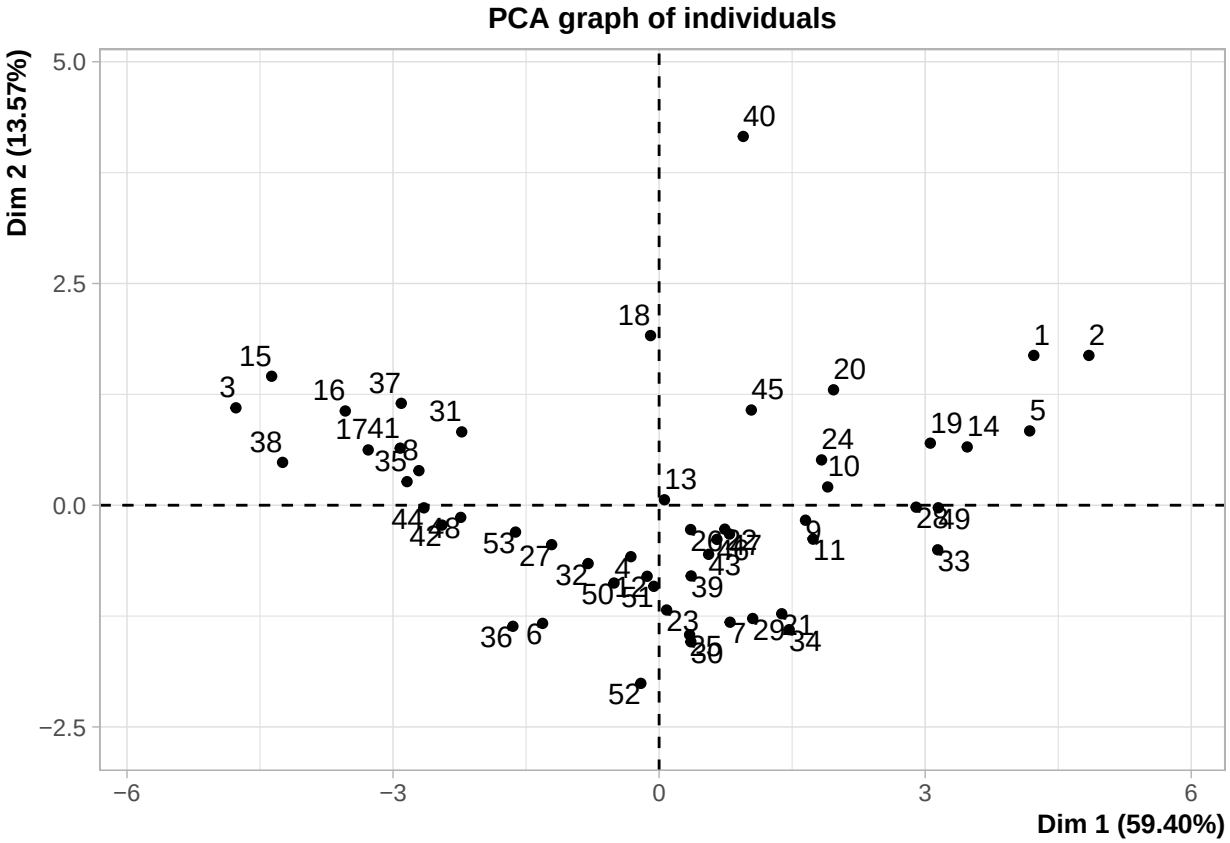

Covarianzas
Comp.2
Comp.1
0.0
0.2
0.4
0.6
0.8
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

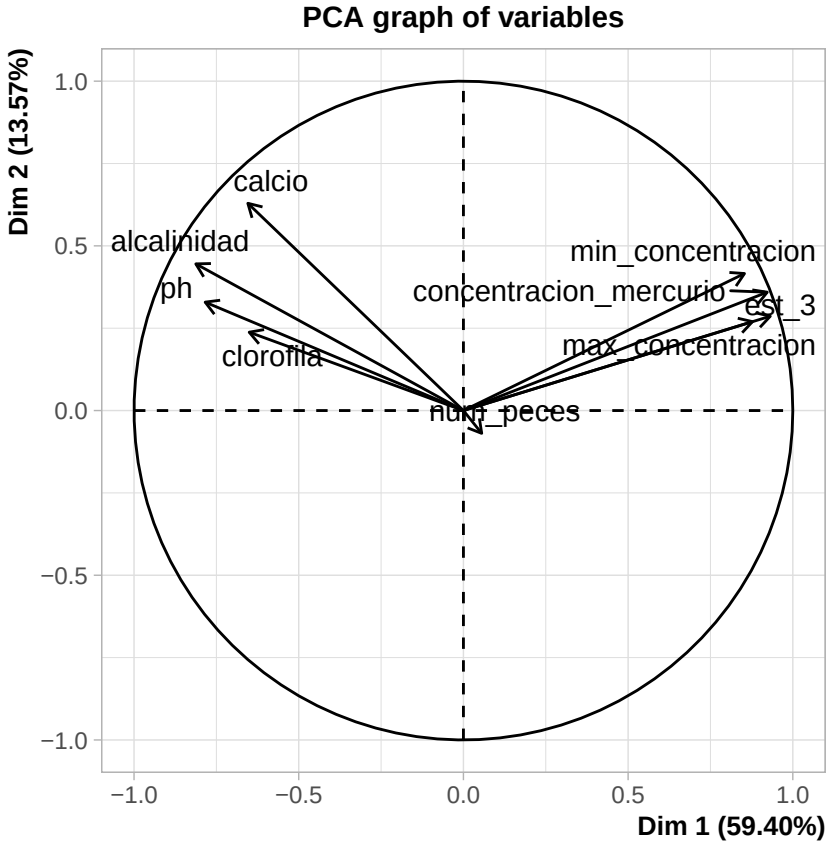


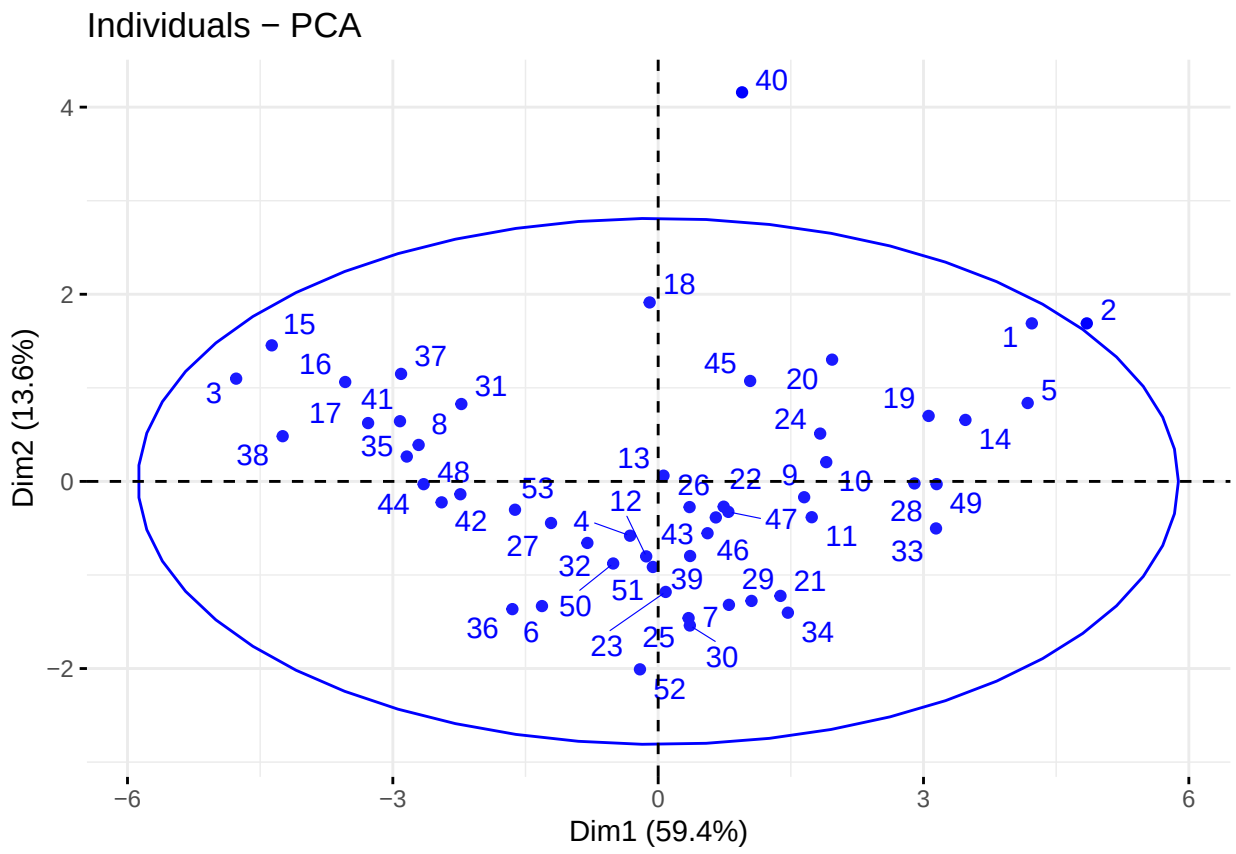
En el primer gráfico se puede evidenciar que los datos no se agrupan mucho dado que se forman varios subgrupos, razón por la cuál es complicado generar componentes principales para este conjunto de datos. Los datos son dispersos.

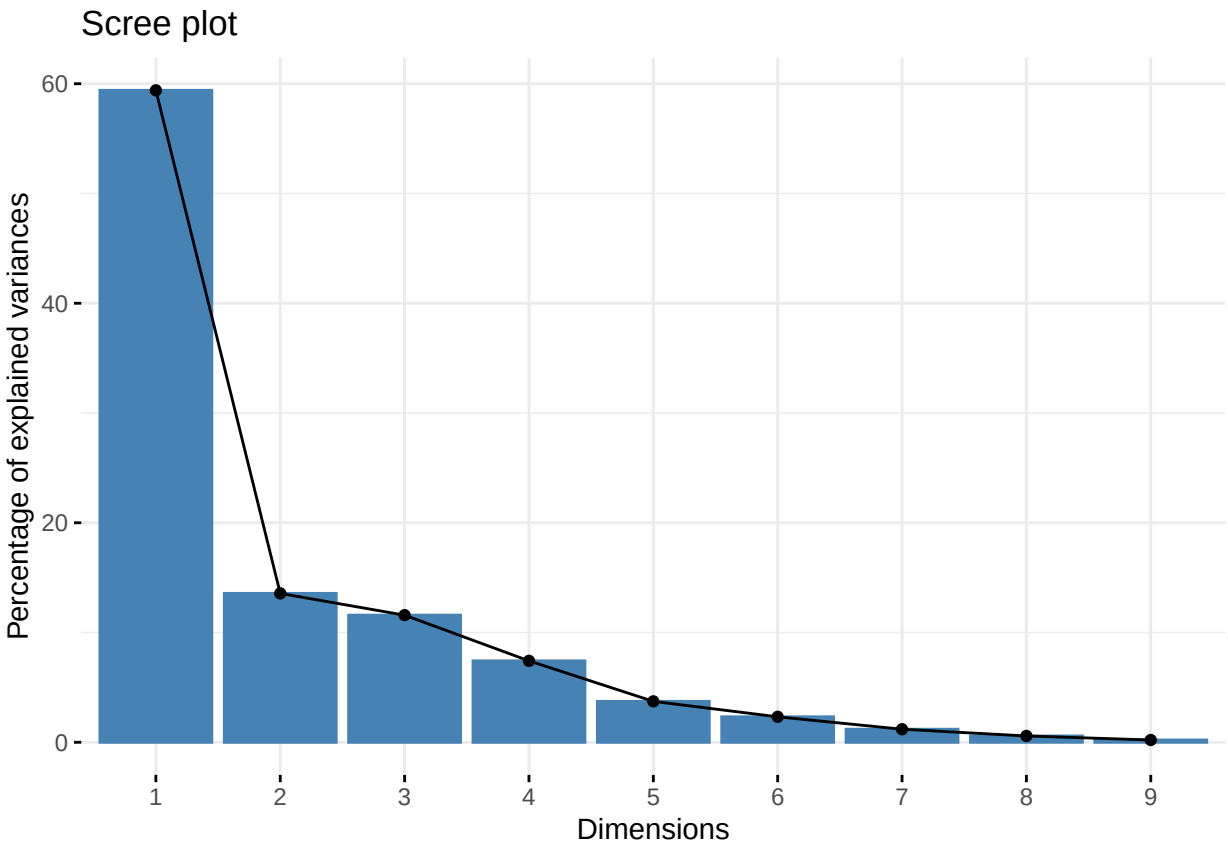
En el segundo gráfico se puede evidenciar aquellas variables más influyentes, en este caso, clorofila, alcalinidad y calcio. A travez de PCA se pueden obtener las variables más influyentes.

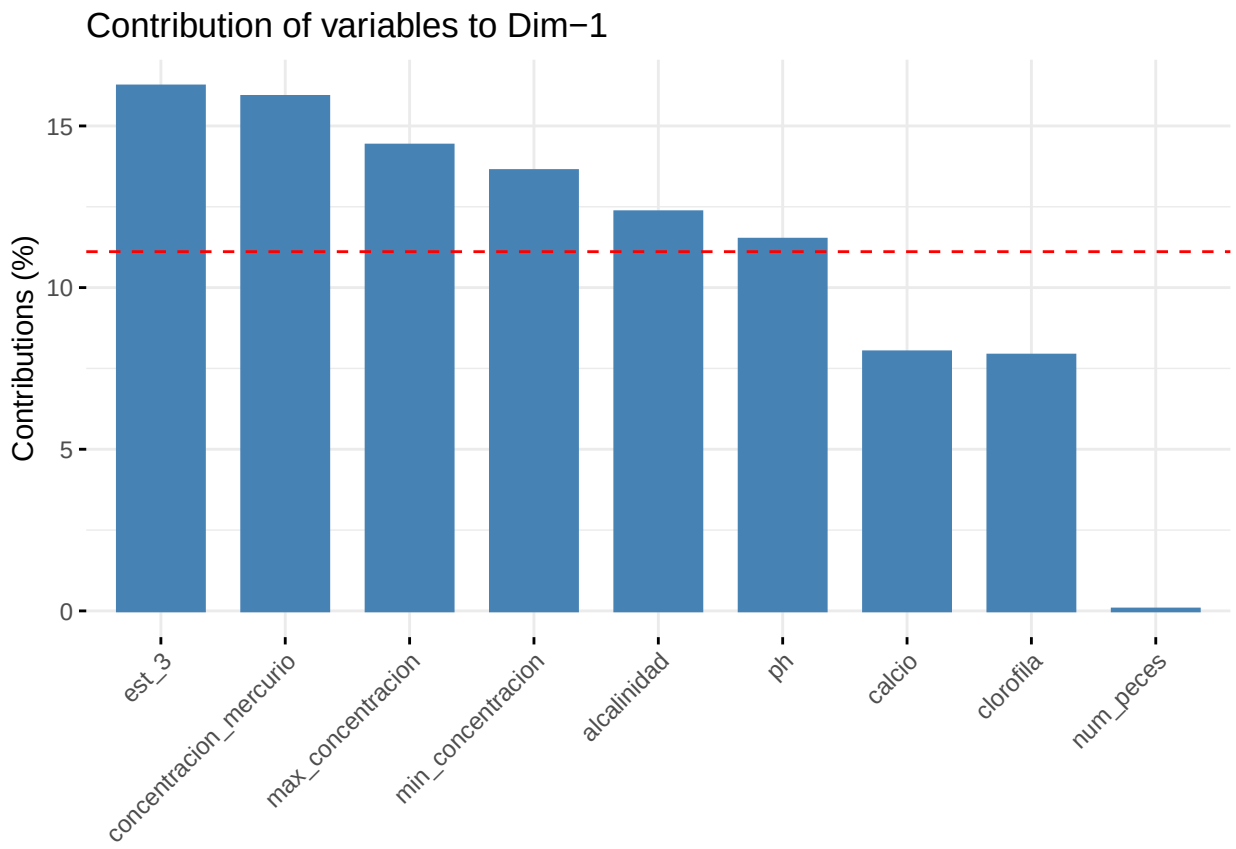
Gráficas para entender PCA






Individuals − PCA
Dim2 (13.6%)
Dim1 (59.4%)
4
2
0
−2
−6
−3
0
3
6
40
18
15
16
37
3
41
31
17
8
38
35
48
44
42
53
12
4
27
32
51
50
36
6
23
25
52
30
7
39
43
46
47
26
22
9
10
11
29
21
34
13
45
20
24
19
14
1
2
5
28
49
33


Scree plot
Percentage of explained variances
60
40
20
0
1
2
3
4
5
6
7
8
9
Dimensions



El la Gráfica 1 se puede observar la disperción de los datos como ya se había visto antes que es complicado agruparlos.

En la Gráfica 2 se puede evidenciar las variables más influyentes, en este caso, la gran parte de las variables tiene similar longitud de vector en cuanto a contribución de los 2 primeros componentes principales.

En la Gráfica 3 se evidencia la dispersión de los datos referente a los 2 primeros componentes principales. No logran agruparse e incluso hay algunos que se salen fuera de los límites definidos por la elipse.

En la gráfica 4 se muestra la varianza explicada por cada componente principal. A medida que aumenta el número de componente principal, menor explicabilidad a la varianza de los datos.

En la última gráfica, se puede observar aquellas variables más influyentes para el primer componente principal, es decir, aquellas que tienen mayor explicabilidad a la varianza de los datos. Con esto se puede llegar a saber que las principales variables son todas las que tienen que ver inmediatamente con concentración de mercurio. Las siguientes variables más influyentes son alcalinidad, ph y calcio.

Conclusión

Se realizó un análisis de normalidad y se logró determinar que los datos no provienen de una distribución. Las únicas variables que provienen de una distribución normal bivariante son ph y máxima concentración de mercurio.

Con ambas variables, es complicado poder utilizarlas separadas del resto del conjunto de datos para realizar un análisis más avanzado. De esta forma sabemos que los datos al no venir de una distribución normal, será complicado poder utilizarlos para realizar modelos que requieran que lo sean. De igual forma, si se realiza un análisis de componentes principales no será adecuado utilizarlos para un análisis posterior ya que al no provenir de una distribución normal no es recomendable.

Posteriormente, se realizó el análisis de componentes principales con el conjunto de datos. No se obtuvieron resultados tan prometedores, pues no se logró reducir en gran medida la dimensionalidad del conjunto de datos sin perder interpretabilidad e información. Fue necesario utilizar la varianza explicada de los 4 primeros componentes principales con la finalidad de tener al menos el 90% de la explicabilidad del conjunto de datos.

Se logró identificar que los datos son muy dispersos y es dificil agruparlos. De igual forma a través de PCA se logró identificar las variables más influyentes en el conjunto de datos.

Las principales variables halladas fueron alcalinidad, ph y calcio.

De esta forma se puede utilizar PCA para saber cuáles son las variables más influyentes del conjunto de datos a pesar de no lograr reducir la dimensionalidad del mismo.

Anexos

Documento de análisis y archivo de la base de datos. https://github.com/FelipeYepez/Inteligencia-Artificial-Avanzada2/tree/main/Estad%C3%ADstica%20Avanzada