Instituto Tecnológico de Estudios Superiores de Monterrey
Campus Monterrey

“Apegándome a la Integridad Académica de los Estudiantes del Tecnológico de Monterrey, me comprometo a que mi actuación en esta actividad esté regida por la integridad académica. En congruencia con el compromiso adquirido, realizaré este trabajo de forma honesta y personal, para reflejar, a través de él, mi conocimiento y aceptar, posteriormente, la evaluación obtenida”

Inteligencia artificial avanzada para la ciencia de datos II
Gpo 502

Módulo 5: Estadística
Procesamiento de datos multivariados

Alumno:
A01750164 | Paul Martín García Morfín

Profesor:
Blanca Rosa Ruiz Hernandez

Fecha: 30/11/2022

Los peces y el mercurio

Resumen

Se realiza el análisis de un conjunto de datos que contiene información sobre la contaminación por mercurio en los peces comestibles de agua dulce, en el que se incluye información como el nombre del lago, la alcalinidad, el PH, el calcio, la clorofila, el número de peces y su madurez, así como medidas de la concentración de mercurio. El presente trabajo se interesa por identificar aquellos factores que influyen en el nivel de contaminación por mercurio. Para ello, se utilizaron herramientas estadísticas como el análisis de normalidad y multinormalidad, gráficos para entender y visualizar el comportamiento de los datos, además de un análisis de componentes principales. Se llegó a la conclusión de que los principales factores son la estimación de concentración de mercurio, la concentración media, el número de peces y la edad de estos. Además, por la variabilidad que explican, es posible decir que el calcio, clorofila y PH tamhién influyen en la concentración de mercurio de los peces.

Introducción

La contaminación por mercurio de peces comestibles en el agua dulce es una amenaza directa contra la salud de las personas. El consumo de pescado forma parte de una dieta balanceada por su importancia como fuente de proteína y para mantener una buena salud cardiovascular, no obstante, la contaminación que llega a los cuerpos de agua tales como ríos, mares y lagos da lugar a que especies de peces acumulen cantidades de mercurio en su organismo, así como otros contaminantes, que en cantidades significativas resultan ser dañinos para nuestro organismo después de ingerirlos.

En el caso del mercurio, este se trata de un metal pesado que es sumamente tóxico y puede causar afecciones en el desarrollo del sistema nervioso central, en los riñones, el hígado y en los órganos reproductivos. La contaminación de animales y plantas por mercurio se puede dar por procesos naturales a través del ambiente, sin embargo, la actividad humana ha causado un aumento en las cantidades de mercurio que se transfiere a los organismos. Este metal se transmite de forma acumulativa gracias a la cadena alimenticia, siendo el pescado la mayor fuente de exposición a este metal para los humanos.

En ese sentido, el conjunto de datos con el que se trabaja se creó a partir del estudio de 53 lagos de Florida con el propósito de examinar los factores que influyen en el nivel de contaminación por mercurio. De forma más detallada, el archivo de datos contiene la siguiente información:

- X1 = número de identificación del lago
- X2 = nombre del lago analizado

- X3 = alcalinidad (mg/l de carbonato de calcio)
- X4 = PH
- X5 = calcio (mg/l)
- X6 = clorofila (mg/l)
- X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
- X8 = número de peces estudiados en el lago
- X9 = mínimo de la concentración de mercurio en cada grupo de peces
- X10 = máximo de la concentración de mercurio en cada grupo de peces
- X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Este trabajo es la continuación de un análisis anterior en el que se utilizó un modelo de regresión múltiple, por lo que uno de los objetivos es contrastar los resultados relacionados con los principales factores que influyen en la contaminación por mercurio. Además, se busca responder a las siguientes preguntas:

- ¿En qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas?
- ¿Cómo ayudan los componentes principales a abordar el problema?

Procedimiento

1. Exploración de la base de datos

A. Acceso a la base de datos y exploración de variables

Se consiguió cargar el conjunto de datos.

Alcalinidad <dbl>	PH <dbl>	Calcio <dbl>	Clorofila <dbl>	ConMedia_Mercurio <dbl>	No_Peces <int>	Min_Con <dbl>	Max_Con <dbl>	ConEst_Mercurio <dbl>	Edad_Peces <int>
5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53	1
3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33	0
116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04	0
39.4	6.9	16.4	3.5	0.44	12	0.13	0.84	0.44	0
2.5	4.6	2.9	1.8	1.20	12	0.69	1.50	1.33	1
19.6	7.3	4.5	44.1	0.27	14	0.04	0.48	0.25	1
5.2	5.4	2.8	3.4	0.48	10	0.30	0.72	0.45	1
71.4	8.1	55.2	33.7	0.19	12	0.08	0.38	0.16	1
26.4	5.8	9.2	1.6	0.83	24	0.26	1.40	0.72	1
4.8	6.4	4.6	22.5	0.81	12	0.41	1.47	0.81	1

Se tienen 12 variables con 53 observaciones cada una, de las cuales hay:

- 9 variables numéricas: Alcalinidad, PH, Calcio, Clorofila, Media, Peces, Mínimo, Máximo, Estimación

- 3 variables categóricas: ID, Lago, Edad

2. Análisis de normalidad

El objetivo de este análisis es identificar variables normales.

Hipótesis
- H0: La muestra proviene de una distribución normal.
- H1: La muestra no proviene de una distribución normal.

Nivel de significancia
- El nivel de significancia con el que se trabajará es de 0.05. $\alpha = 0.05$

Criterio de decisión
- Si $P < \alpha$ Se rechaza H0
- Si $P >= \alpha$ No se rechaza H0

Variables continuas a analizar: Alcalinidad, PH, Calcio, Clorofila, Media de mercurio, Mínimo de mercurio, Máximo de mercurio, Estimación de mercurio.

A. Pruebas de normalidad para identificar variables normales y detectar posible normalidad multivariada

a. Prueba de Anderson-Darling

	Test <S3: Asis>	Variable <S3: Asis>	Statistic <S3: Asis>	p value <S3: Asis>	Normality <S3: Asis>
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	Media	0.9253	0.0174	NO
6	Anderson-Darling	Mínimo	1.9770	<0.001	NO
7	Anderson-Darling	Máximo	0.6585	0.081	YES
8	Anderson-Darling	Estimación	1.0469	0.0086	NO

Como se puede observar, las únicas variables con distribución normal según este test son el PH y la Concentración máxima de mercurio.

b. Prueba de Mardia

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	306.147576400225	3.13807898771458e-18	NO
Mardia Kurtosis	4.59976628417483	4.22965210922222e-06	NO
MVN	*NA*	*NA*	NO

Se comprobaron los test de normalidad en los que se utilizó el método de Anderson-Darling, llegando al mismo resultado que anteriormente. Por otro lado, con este grupo de variables no se detectó multinormalidad, según el test de Mardia, los métodos de sesgo y kurtosis.

A continuación, se harán las mismas pruebas pero sólo con aquellas variables que sí presentaron normalidad.

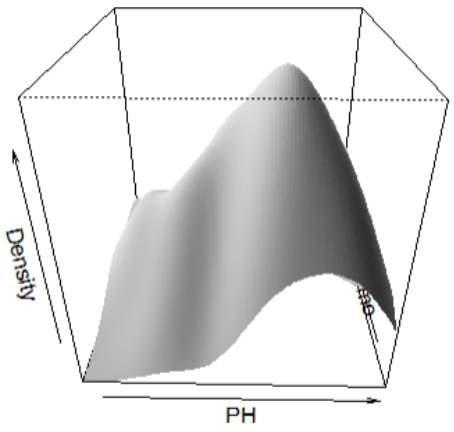
B. Pruebas de normalidad en variables con distribución normal, interpretación de resultados, análisis de sesgo y curtosis

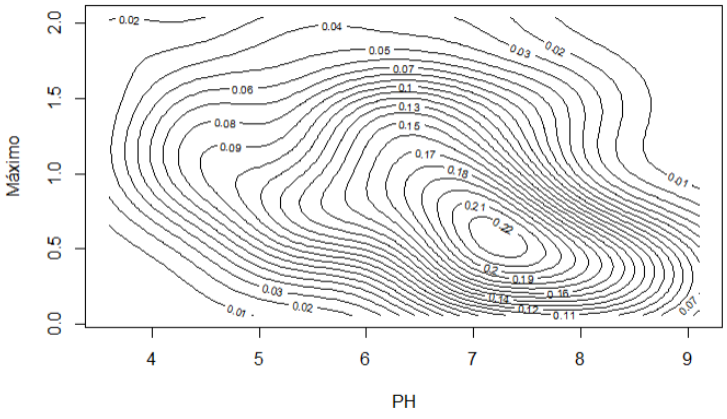
Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	6.17538668676458	0.186427564928852	YES
Mardia Kurtosis	-1.12820795824432	0.25923210375991	YES
MVN	*NA*	*NA*	YES

Se puede notar que en este caso sí existe multinormalidad para estas dos variables (PH y Concentración máxima de mercurio) según la prueba de Mardia así como con los métodos de sesgo y kurtosis.

Los datos de PH son moderadamente simétricos, ya que presentan un ligero sesgo a la izquierda, con baja kurtosis, teniendo una forma platicúrtica.
La variable de concentración máxima de mercurio tiene una distribución moderadamente sesgada a la derecha, con baja kurtosis y, por tanto, una forma platicúrtica.
Parece que no hay presencia de datos atípicos, pero esto se analizará a detalle más adelante.

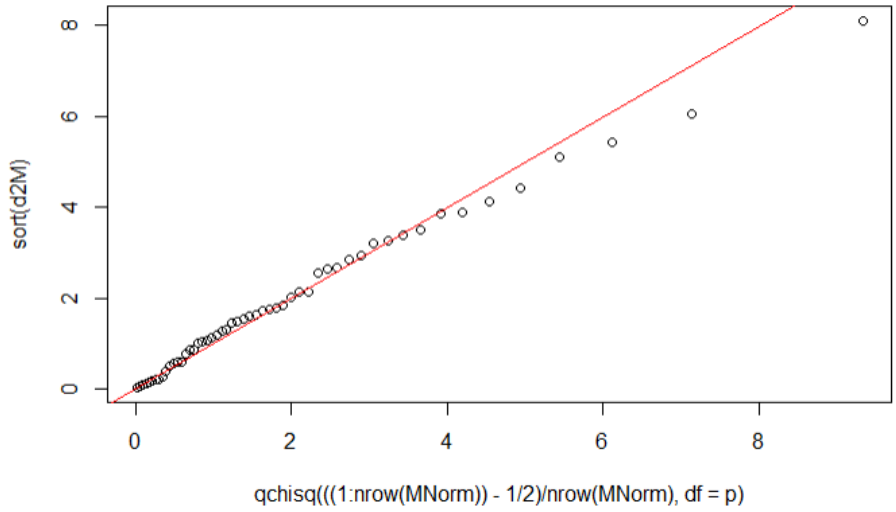
C. Gráficas de normal multivariada

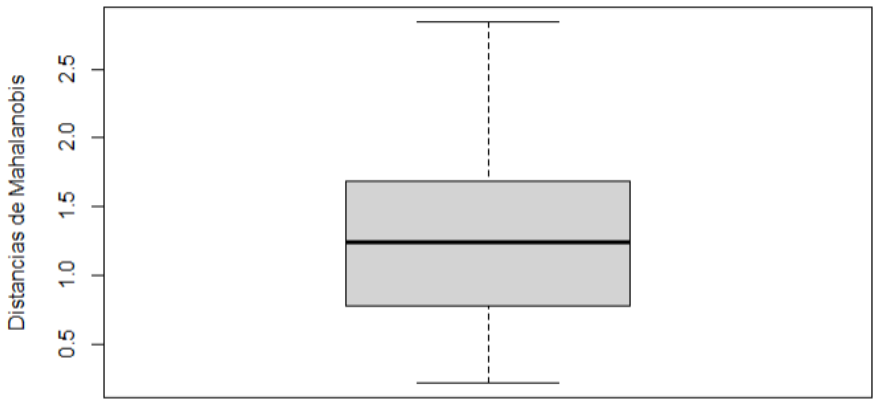




El diagrama de perspectiva nos proporciona información sobre cómo se correlacionan dos variables entre sí. Cuando los datos están distribuidos normalmente, se espera obtener un gráfico tridimensional en forma de campana, lo cual sí ocurre en este caso. El gráfico de contorno es una proyección bidimensional de la gráfica de perspectiva, así que se debe observar un patrón similar, en este caso, con contornos con forma elíptica.

D. Datos atípicos e influyentes en la normal multivariada (distancia de Mahalanobis y gráfico QQplot multivariado)





Gráficamente, se puede observar que el cuadrado de la distancia de Mahalanobis se aproxima a una distribución Chi2, esto da un indicio de que la muestra pertenece a una distribución normal multivariada, sin embargo, gracias a los tests hechos anteriormente esto ya fue comprobado.

La distancia de Mahalanobis se refiere a la distancia entre cada punto de datos y el centro de masa. Por tanto, es cero cuando el dato se encuentra en el centro y es mayor a cero cuando se encuentra alejado. Los puntos que se encuentran muy lejos son datos atípicos. En este caso, se observa que no hay presencia de datos atípicos según el boxplot.

3. Análisis de de componentes principales

Para este análisis se hace uso de la base de datos completa para identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

A. Justificación del uso de componentes principales

Hacer un análisis de componentes principales es una técnica útil para el análisis exploratorio de datos, que permite visualizar de mejor forma la variación presente en un conjunto de datos con muchas variables así como hallar aquellas con un mayor peso. A continuación se muestra la matriz de correlación.

	Alcalinidad	PH	Calcio	Clorofila	Media	Peces	Mínimo	Máximo	Estimación	Edad
Alcalinidad	1.00000000	0.71916568	0.832604192	0.47753085	-0.59389671	0.01029074	-0.52535654	-0.60479558	-0.62795845	-0.094938825
PH	0.71916568	1.00000000	0.577132721	0.60848276	-0.57540012	-0.01860607	-0.54196524	-0.55181523	-0.61284905	0.038000214
Calcio	0.83260419	0.57713272	1.000000000	0.40991385	-0.40067958	-0.08937901	-0.33247623	-0.40791663	-0.46440947	-0.002111124
Clorofila	0.47753085	0.60848276	0.409913846	1.00000000	-0.49137481	-0.01182027	-0.40045856	-0.48497215	-0.50644193	-0.283002338
Media	-0.59389671	-0.57540012	-0.400679584	-0.49137481	1.00000000	0.07903426	0.92720506	0.91586397	0.95921481	0.108738958
Peces	0.01029074	-0.01860607	-0.089379013	-0.01182027	0.07903426	1.00000000	-0.08165278	0.16109174	0.02580046	0.207956171
Mínimo	-0.52535654	-0.54196524	-0.332476229	-0.40045856	0.92720506	-0.08165278	1.00000000	0.76535319	0.91908939	0.100661967
Máximo	-0.60479558	-0.55181523	-0.407916635	-0.48497215	0.91586397	0.16109174	0.76535319	1.00000000	0.85975810	0.093752072
Estimación	-0.62795845	-0.61284905	-0.464409465	-0.50644193	0.95921481	0.02580046	0.91908939	0.85975810	1.00000000	0.089411267
Edad	-0.09493882	0.03800021	-0.002111124	-0.28300234	0.10873896	0.20795617	0.10066197	0.09375207	0.08941127	1.000000000

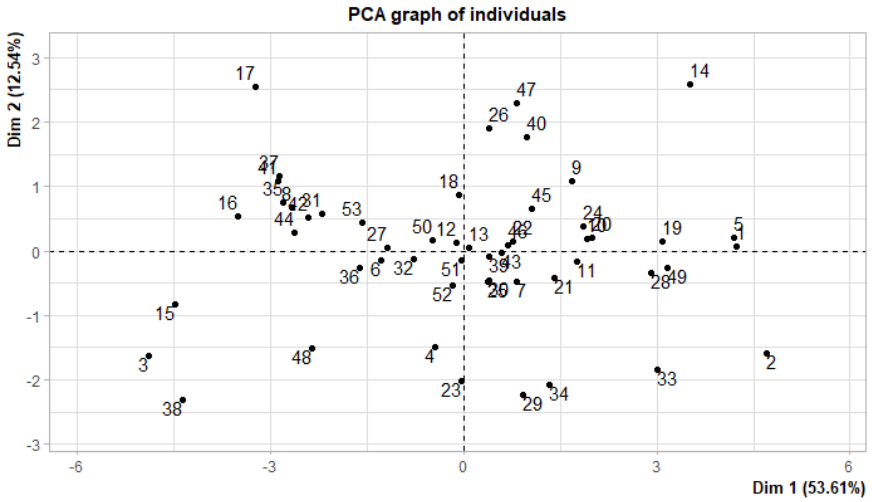
B. Análisis de componentes principales y justificación del número de componentes principales apropiados

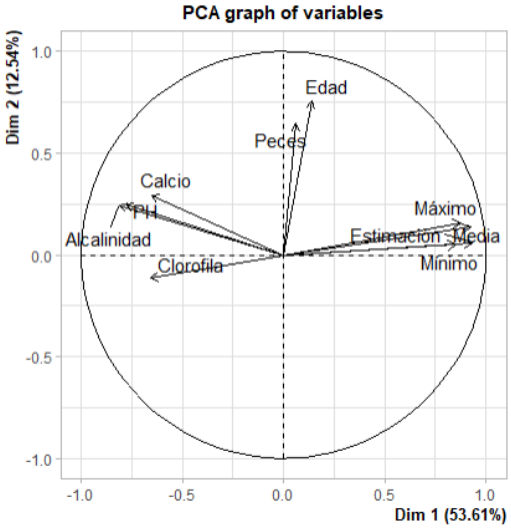
```
eigen() decomposition
$values
 [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741 0.20673634 0.08682133 0.05163902 0.01863699

$vectors
             [,1]        [,2]       [,3]         [,4]        [,5]        [,6]        [,7]        [,8]        [,9]       [,10]
 [1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534  0.07547497 -0.33823501  0.68622998  0.04284021 -0.02239801
 [2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038  0.73121012 -0.08629646 -0.28769221  0.01363551  0.04445261
 [3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937 -0.31342329  0.34312185 -0.45568753 -0.11508339  0.02634676
 [4,] -0.28334182  0.10195058 -0.2639612 -0.432676049 -0.63093376 -0.44112169  0.13435159  0.19006976 -0.06333133 -0.03982419
 [5,]  0.39830786 -0.12104244 -0.2996635 -0.080630070 -0.03046869  0.07436922 -0.01377825 -0.01674789  0.06243320 -0.84827636
 [6,]  0.02667579 -0.57556151  0.3050633 -0.692854505  0.19646415 -0.05926732 -0.14693148 -0.16809481  0.02532023  0.04805976
 [7,]  0.36839224 -0.04432459 -0.3876861  0.044658983 -0.13236038 -0.19602465 -0.45674057 -0.18260535  0.53803577  0.35020485
 [8,]  0.37893835 -0.14237181 -0.2024901 -0.167921215  0.02678086  0.26671839  0.67376588  0.33602914  0.18844932  0.30445219
 [9,]  0.40206100 -0.05279514 -0.2562319 -0.042242268 -0.05607416  0.03863899 -0.23387764  0.02613406 -0.80648296  0.24018040
[10,]  0.05931430 -0.67421026  0.2294446  0.521815581 -0.37253140 -0.21612970  0.05759514  0.16451240 -0.02782678 -0.01839703

[1] 10
[1] 10
 [1] 0.536122641 0.125426109 0.121668138 0.090943267 0.059141736 0.030314741 0.020673634 0.008682133 0.005163902 0.001863699
 [1] 0.5361226 0.6615488 0.7832169 0.8741602 0.9333019 0.9636166 0.9842903 0.9929724 0.9981363 1.0000000
```

La primera componente explica el 53.61% de la varianza observada en los datos, la segunda el 12.54%, la tercera el 12.17%, la cuarta 9.09%, etc. Utilizando las primeras siete componentes

se podría explicar el 98.42% de la varianza observada. Es importante mencionar que al usar la matriz de correlación, las variables son escaladas para tener una varianza unitaria.

C. Gráfico de los vectores asociados y puntuaciones de las observaciones





D. Interpretación de resultados

La primera gráfica nos permite resumir y visualizar la información de las observaciones descritas por las variables cuantitativas interrelacionadas del conjunto de datos, de las dos primeras componentes.

En la segunda gráfica se muestran los vectores y cómo están fijados en el origen de las principales componentes. Sus valores de proyección indican el peso que tienen en esa

componente. En este caso, las tres variables con más peso en la primera componente son la Estimación, la Media y el Máximo; y para la segunda componente son la Edad y el Número de peces.

La primera componente se podría etiquetar como las concentraciones de mercurio, ya que las variables con mayor peso son aquellas relacionadas a medidas como media, máximo, mínimo y estimación de la concentración de mercurio. La segunda componente tiene que ver más con la cantidad de peces analizados y la madurez de estos. Por último, si se observan las demás componentes, la mayoría tiene más relación con variables como el PH, el calcio o la clorofila.

Conclusión

Cuando se trabaja con una gran cantidad de variables se dificulta el proceso de hallar las relaciones que existen entre estas, cosa que puede ser un problema si se presenta el caso de que no haya independencia entre dichas variables, ya que es posible que estén midiendo lo mismo pero bajo distintos puntos de vista. Por ello, resulta importante reducir el número de variables y tomar aquellas que nos brinden más información, es decir, en las que haya una mayor variabilidad.

Existen diferentes métodos para estudiar estas relaciones, siendo uno de ellos el análisis de componentes principales, el cual ayuda a visualizar el orden de importancia de cada variable en cuanto a la variabilidad que explican. Dicho de otra forma, el análisis de componentes principales permite visualizar de mejor forma la variación presente en un conjunto de datos con muchas variables y de esta forma hallar aquellas con un mayor peso.

Por otro lado, en el análisis de componentes principales no se requiere de la suposición de multinormalidad de los datos, no obstante, si se presenta este caso ayuda a dar una mejor interpretación de los componentes.

Se llegó a la conclusión de que los principales factores son la estimación de la concentración de mercurio, la concentración media, el número de peces estudiados y la edad de estos. Además, por la variabilidad que explican, también se puede decir que las concentraciones de calcio, clorofila y PH influyen en la concentración de mercurio de los peces.

Referencias

García, P. (2022, 18 de septiembre). Construcción de un modelo estadístico base: Los peces y el mercurio. ITESM.

Ecologistas en acción. (2020, 07 de julio). Mercurio en pescado. https://www.ecologistasenaccion.org/4975/mercurio-en-pescado/

Carpio, E. (2019, 20 de diciembre). Normalidad univariante y multivariante en R. UNA-PUNO. http://rstudio-pubs-static.s3.amazonaws.com/562854_fb3b0b61a64346dd94544e8641cc4a5f.html

Salazar, W; Zea A. (2019, 14 de abril). Análisis Descriptivo Multivariante. RPubs. https://rpubs.com/azeav/491093

Muñoz, J; Amón I. (2013). Técnicas para detección de outliers multivariantes. UPB. https://repository.upb.edu.co/bitstream/handle/20.500.11912/6582/T%C3%A9cnicas%20para%20detecci%C3%B3n%20de%20outliers%20multivariantes.pdf?sequence=1&isAllowed=y#:~:text=3.1%20La%20distancia%20de%20mahalanobis&text=%C3%89sta%20describe%20la%20distancia%20entre,distancia%20es%20mayor%20a%20cero.

Universidad Autónoma de Madrid. (s.f.). Análisis de componentes principales. https://www.estadistica.net/Master-Econometria/Componentes_Principales.pdf

Anexos

Liga a los documentos de análisis:
https://drive.google.com/drive/folders/1FrhDej6-UmzN_YvVlbT-7nlkSORvnMmw?usp=sharing