REPORTE DE SOLUCIÓN DEL RETO

Los Peces y el Mercurio

Armando de Jesús Cerda de la Rosa A01570376

10 de noviembre del 2022

Inteligencia artificial avanzada para la ciencia de datos

Blanca Ruiz

Resumen

El problema de este reporte consistió en entender la data proporcionada de manera estadística en cuanto a normalidad y variables representativas gracias a componentes principales. La data muestra la contaminación de mercurio en 53 lagos en Florida. Los dos análisis hechos fueron sobre normalidad y componentes. El primer análisis comenzó con encontrar qué variables tenían normalidad y en conjunto cuales tenían normalidad multivariada, para luego encontrar datos atípicos de la normal multivariada. El segundo análisis comenzó con una matriz de correlaciones, para luego ver cuales eran los componentes que mejor resume las variables de la data, y luego que tanto influenciaron cada variable a los componentes principales.

Introducción

- Normalidad
  - ¿Qué variables son normales?
  - ¿Qué grupo de variables tienen normalidad multivariada?
  - ¿Qué se interpreta de sesgo y curtosis?
- Componentes principales
  - ¿Porqué es adecuado el uso de componentes principales?
  - ¿Cuántos componentes principales son adecuados?
  - ¿Cuáles son las variables que impactan a PC1 y PC2?
- Análisis
  - ¿Cuáles son los factores principales?
  - ¿Cómo ayuda la normalidad con los outliers?

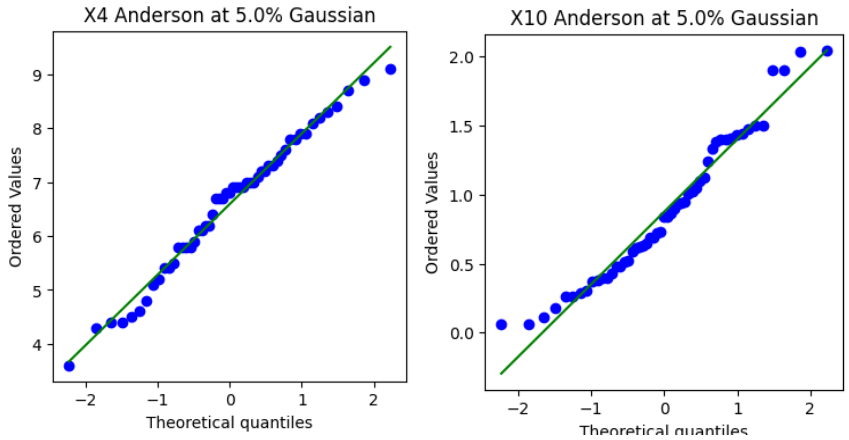
Análisis

Código

[Link]

Normalidad

¿Qué variables son normales?



De todas las variables que se tenían en el set de datos solo estas cumplían con la prueba de Anderson al ser menor que .05 y se refleja bastante bien en cómo el plot está super cerca de la línea.

¿Qué grupo de variables tienen normalidad multivariada?

```
Multivariate Skewness 0.7402136949443168
Statistic 6.538554305341465
P-Value 0.16237730235450726
Normality yes

Multivariate Kurtosis 7.022738670293162
Statistic 0.7908922569787501
P-Value 0.3738304629001151
Normality yes
```

Si solo ponemos las variables que se encontraron como normales individualmente y las analizamos en conjunto para evaluar normalidad multivariada encontramos que la prueba es positiva.

¿Qué se interpreta de sesgo y curtosis?

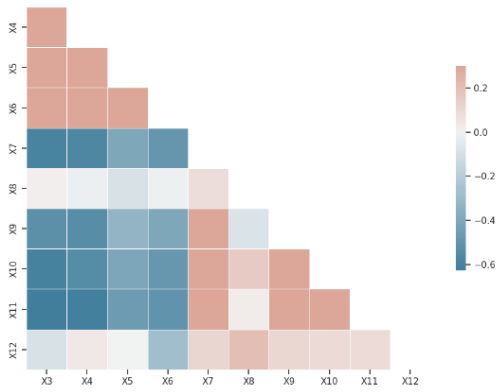
```
Multivariate Skewness 56.90573699461316
Statistic 502.6673434524163
P-Value 3.627769397753002e-24
Normality No

Multivariate Kurtosis 140.56710967185447
Statistic 23.35345626402786
P-Value 1.3480107591351132e-06
Normality No
```

Con la prueba de mardia encontramos que en conjunto todas las variables del dataset no presentan normalidad. Lo que también se puede ver visualmente en la distribución de las variables al tener la mayoría sesgos a la izquierda.

Componentes principales

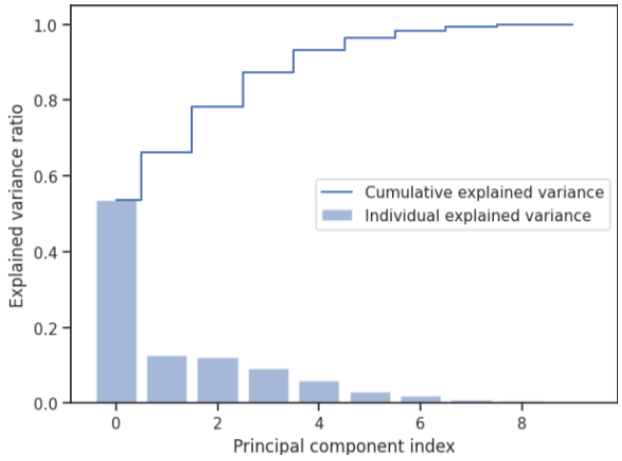
¿Cuándo es adecuado el uso de componentes principales?



Si las variables en un dataset están altamente correlacionadas como se muestra en la matriz de correlación para el dataset de lagos, un análisis de componentes principales puede ayudar a reducir la redundancia de varianza. Menos variables con mayor representación

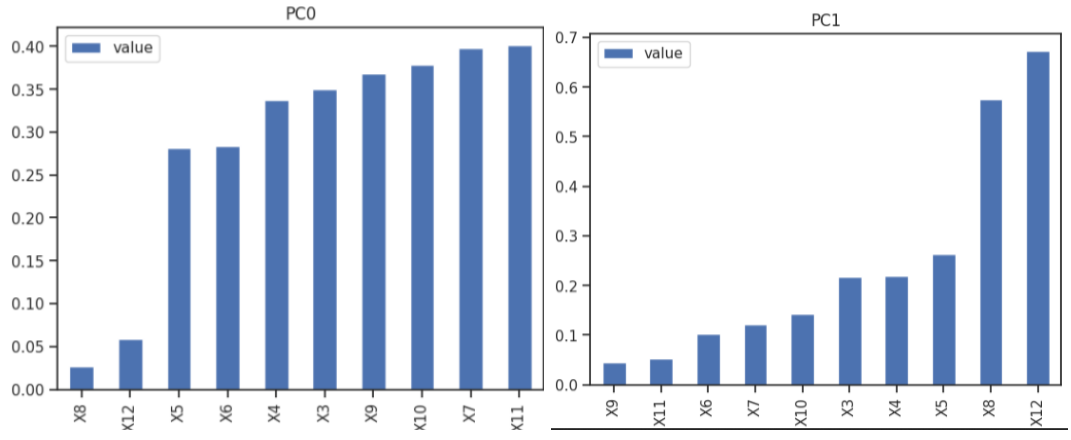
de la varianza ayuda a que con menos datos se pueda hacer manos, evitando la Curse of Dimensionality.

¿Cuántos componentes principales son adecuados?



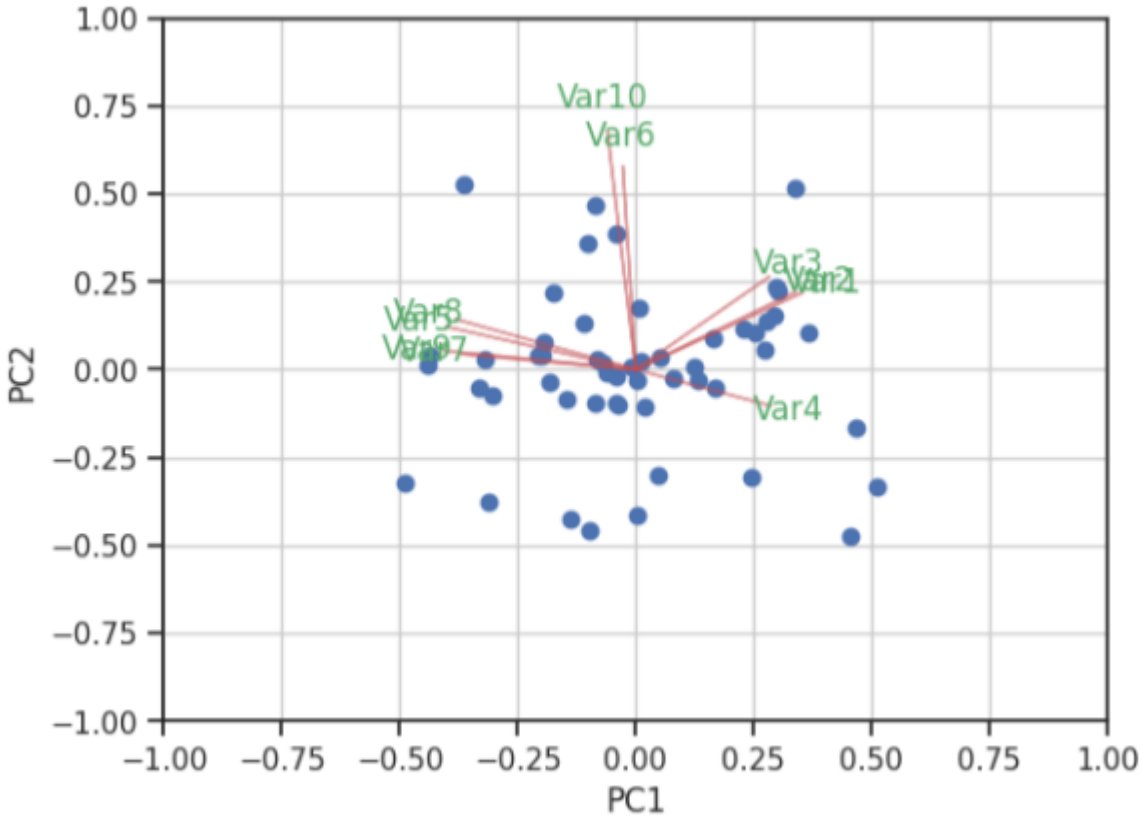
De acuerdo a la siguiente gráfica la mayor cantidad de varianza acumulada estaría entre 3-5 componentes principales siendo el componente 0 el que más explica individualmente. La varianza.

¿Cuáles son las variables que impactan a PC1 y PC2?



El componente 0 tiene como su mayor componente a la varianza de x11, x7, y x10. El componente 1 tiene como su mayor componente a

x12 y a x8 y x5 que curiosamente son los menos contribuyentes de PC0.



Análisis

¿Cómo ayuda la normalidad con los outliers?

Sabiendo que un conjunto de datos tiene normalidad multivariada se puede identificar filas de outliers viendo su significancia.

	X4	X10	mahalanobis	p
13	5.8	2.03	5.430035	0.066204
17	7.8	1.50	5.115416	0.077482
23	6.9	2.04	8.099887	0.017423
32	3.6	1.90	6.059085	0.048338

Conclusión

Gracias a estos análisis, si comenzara algún proceso de implementar modelos de predicción lo que haría en mi preprocessing sería eliminar outliers considerando la normalidad multivariada. Además de eliminar la redundancia aprovechando 3 componentes principales.