Escuela de Ingeniería y Ciencias

Maestría en Inteligencia Artificial Aplicada (MNA-V)

Semana 4. Actividad Semanal. Proceso de PCA

TC4029.10 Ciencia y Analítica de Datos

Profesor: Jobish Vallikavungal Devassia

EQUIPO 5:

- **Laura Irán González Ojeda**
  - **Matrícula: A01794099**
- **Marcela Alejandra Rosales Jiménez**
  - **Matrícula: A01032022**

|MARTES, 11 DE OCTUBRE DE 2022.

1 Parte 2.

1. **¿Cuál es el número de componentes mínimo y por qué?**

Después de hacer el análisis PCA y ver la Proporción Acumulada, hemos encontrado que los componentes mínimos para lograr varianza acumulada de más del 95% son en total 9. Lo cual nos permite reducir de 15 columnas a 9, sin perder demasiada información.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Standard deviation	2.8305	1.0302	0.9613	0.9178	0.9001	0.8661	0.8617	0.8327	0.8033	0.5397	0.3877	0.1871	0.1360	0.1200	0.1009
Proportion of variance	0.5341	0.0708	0.0616	0.0562	0.0540	0.0500	0.0495	0.0462	0.0430	0.0194	0.0100	0.0023	0.0012	0.0010	0.0007
Cumulative proportion	0.5341	0.6048	0.6664	0.7226	0.7766	0.8266	0.8761	0.9223	0.9654	0.9848	0.9948	0.9971	0.9984	0.9993	1.0000

2. **¿Cuál es la variación de los datos que representan esos componentes?**

La Varianza acumulada del Componente Principal 9 es de 96.53596276679052.

```
varianza_acumulada_pc9 = ((pcsSummary_df.iloc[2][8])*100)
print(f"Varianza acumulada Componente Principal 9 = {varianza_acumulada_pc9}")
```

```
Varianza acumulada Componente Principal 9 = 96.53596276679052
```

3. **¿Cuál es la pérdida de información después de realizar PCA?**

La pérdida de información sería de un 3.4640372332094.

100 - 96.53596276679052 = 3.46403723320948

4. **De las variables originales, ¿Cuál tiene mayor y cuál tiene menor importancia en los componentes principales?**

Concluimos que los componentes PC1, PC2 y PC3 contienen los coeficientes más representativos en el análisis de datos porque tienen mayor proporción de varianza, lo contrario a los componentes PC4 a PC8.

El análisis individual de los 3 componentes más importantes representa diferentes correlaciones de las variables originales:

- **PC1**. Correlación de las columnas X1, X12, X13, X14, X15, X16, X17. (Correlación entre los meses de deuda en estados de cuenta y el monto otorgado a cada cliente).

```
pcsComponents_df.PC1.nlargest(7)
```

```
X15    0.341287
X14    0.341176
X16    0.338614
X13    0.338563
X12    0.334578
X17    0.333334
X1     0.295097
Name: PC1, dtype: float64
```

  - Análisis obtenido:
    - Los meses de abril a septiembre del 2005 los clientes tienen una mayor cantidad de deuda en sus estados de cuenta.
    - Podemos ver que están relacionados entre ellos. Esto tiene sentido ya que esas 6 categorías pertenecen al mismo grupo (valor del estado de cuenta), pero en distinto mes.
    - También nos muestra que las categorías X12-X17 se relacionan con X1, que es el monto de crédito otorgado a cada cliente.
    - Por lo que podríamos decir que, a mayor valor en el estado de cuenta, mayor crédito es otorgado a un cliente.
- **PC2.** Correlación de las columnas Y y X5. (Correlación entre la probabilidad de otorgar un crédito y la edad del cliente).

```
print(pcsComponents_df.PC2.nsmallest(2))
```

```
Y    -0.869760
X5   -0.296909
Name: PC2, dtype: float64
```

  - Análisis obtenido:
    - Podemos visualizar que la variable Y (probabilidad de otorgar crédito) y X5 (edad) están relacionadas ya que obteniendo los números negativos interpretamos que a menos edad es menos la probabilidad de que se otorgue un crédito.

- **PC3.** Correlación de las columnas Y, X18, X19, X20, X21, X22, X23. (Correlación entre la probabilidad de otorgar un crédito y el historial mensual de pagos).

```
print(pcsComponents_df.PC3.nlargest(7))
```

```
X23    0.528168
Y      0.373907
X21    0.343306
X18    0.296673
X22    0.290993
X19    0.254781
X20    0.212447
Name: PC3, dtype: float64
```

  - Análisis obtenido:
    - Interpretamos que a mejor historial de pagos (X18-X23) es mayor la probabilidad de otorgar o mantener un crédito (Y).

5. ¿Cuándo se recomienda realizar un PCA y qué beneficios ofrece para Machine Learning?

El Análisis de Componentes Principales o PCA es aplicado para reducir la complejidad de los datos en un dataset grande que pareciera no estar correlacionado e identificar las características más importantes.

PCA permite identificar las variables que aportan más información de un dataset, y descartar las menos relevantes. El proceso de descarte de las variables menos relevantes se le conoce como la *reducción de dimensionalidad*, que de hecho es una de las aplicaciones principales del PCA.

Otra de las aplicaciones de PCA es la *detección de anomalías,* ya que PCA analiza las variables que definen lo que corresponde a un comportamiento normal, para después aplicar distintas métricas de distancia (varianza) que identifiquen los casos que se alejan de este comportamiento.

Las dos aplicaciones anteriormente descritas (*reducción de dimensionalidad y detección de anomalías)* son de suma importancia cuando se aplican algoritmos de Machine Learning, sobre todo en aquellos algoritmos que pertenecen al Aprendizaje no supervisado (unsupervised learning), pues los métodos de *unsupervised learning* tienen el objetivo de predecir una variable a partir de una serie de predictores. El principal problema al que se enfrentan los métodos de *unsupervised learning* es la dificultad para validar los resultados dado que no se dispone de una variable respuesta que permita contrastarlos. Por lo que *PCA* permite "condensar" la información aportada por múltiples variables en pocos componentes. (Amat, 2017)

Referencias

Amat, J. (2017, Junio). *Análisis de Componentes Principales*. Retrieved Octubre 9, 2022, from cienciadedatos.net: https://www.cienciadedatos.net/documentos/35_principal_component_analysis#Introducci%C3%B3n

Kane, F. (2017). *Hands-On Data Science and Python Machine Learning.* Birmingham: Packt Publishing Ltd.

Na8. (2018, Octubre 8). *Comprende Principal Component Analysis*. Retrieved Octubre 8, 2022, from Aprende Machine Learning: https://www.aprendemachinelearning.com/comprende-principal-component-analysis/

Recuero, P. (2018, Junio 6). *Python para todos : ¿Qué es el análisis de Componentes Principales o PCA?* Retrieved Octubre 10, 2022, from Think Big: https://empresas.blogthinkbig.com/python-para-todos-que-es-el-pca/