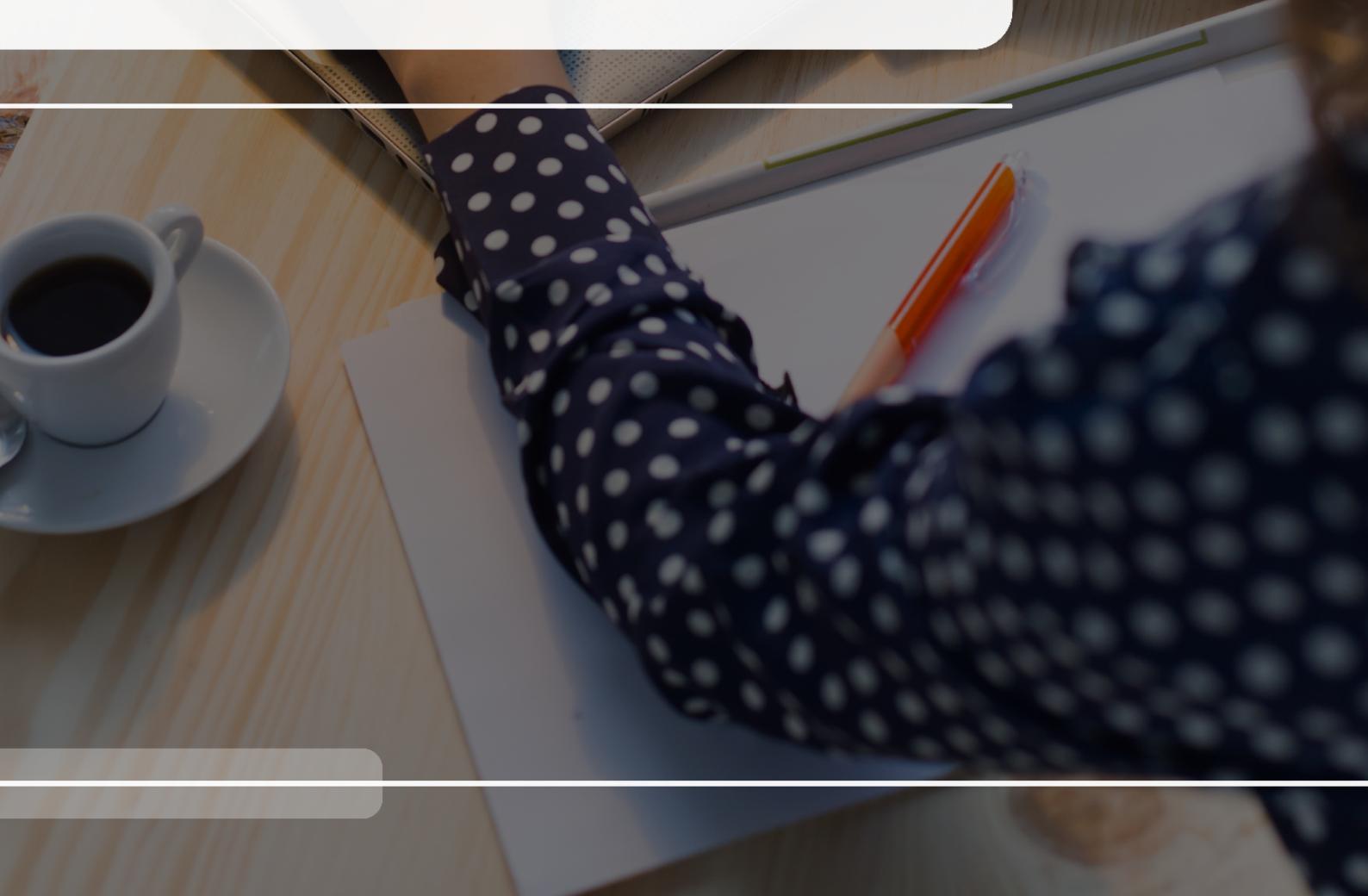




# ABPro

**Módulo:** Aprendizaje de máquina no supervisado



## EVALUACIÓN MÓDULO

### Ciencia de Datos

Módulo: Aprendizaje de máquina no supervisado

### Nivel de Dificultad

Medio

### Nombre del proyecto

Evaluación Módulo 6

### Tema

Aprendizaje de máquina no supervisado

### Objetivo del proyecto (Competencias del módulo)

Elaborar un modelo predictivo a partir de un set de datos utilizando técnicas de aprendizaje de máquina no supervisado implementados en lenguaje Python para resolver un problema.

### Ejecución

Grupal

# Descripción del ejercicio

## Contexto

### Parte I:

1. La principal diferencia entre los métodos supervisados (I) y no supervisados (II) es que:

- a) (I) requieren que el usuario especifique algunos hiperparámetros.
- b) (II) no tienen restricciones y/o supuestos.
- c) (I) usan la variable respuesta para entrenar el modelo.
- d) (II) se aplican a problemas autónomos.

2. Considere las siguientes afirmaciones:

- (i) PCA es un método no supervisado.
- (ii) Todos los componentes principales de un PCA son ortogonales entre si.
- (iii) PCA busca las direcciones en las que los datos tienen la mayor varianza.
- (iv) El número máximo de componentes principales es menor o igual al número de variables.

Elija la opción con el mayor número de ítems correctos:

- a) (i) e (iii).
- b) (ii) e (iii).
- c) (i), (ii) e (iii).
- d) (i), (ii), (iii) e (iv).

3. Como parte de un análisis de datos se analizaron 11 indicadores económicos y sociales de 96 países. Las variables observadas son:

- X1: Tasa anual de crecimiento de la población,  
 X2: Tasa de mortalidad infantil por cada 1000 nacidos vivos,  
 X3: Porcentaje de mujeres en la población activa,  
 X4: PNB en 2005 (en millones de dólares),  
 X5: Producción de electricidad (en millones kW/h),  
 X6: Líneas telefónicas por cada 1000 habitantes,  
 X7: Consumo de agua per cápita, X8: Proporción de la superficie del país cubierta por bosques,  
 X9: Proporción de deforestación anual,  
 X10: Consumo de energía per cápita,  
 X11: Emisión de CO<sub>2</sub> per cápita.

Dada la gran cantidad de variables se aplicó un análisis de componentes principales, utilizando la matriz de correlación, donde los vectores de carga de las dos primeras componentes son:

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$
$Y_1$	-0.314	-0.392	0.116	0.295	0.259	0.446	0.092	0.006	-0.244	0.415	0.375
$Y_2$	0.348	-0.041	-0.583	-0.177	-0.174	-0.027	0.321	-0.457	-0.154	0.233	0.292

A partir d la información anterior, se puede concluir que:

- a) El porcentaje de variabilidad explicado por las dos primeras componentes es 63.45%.
- b) Asumiendo las condiciones necesarias sobre las componentes no reportadas, entonces es posible que las variables X2, X6, X10 y X11 son las que m'as contribuyen en la primera componente principal.
- c) Asumiendo las condiciones necesarias sobre las componentes no reportadas, entonces es posible que las variables X2, X6, X10 y X11 son las que m'as contribuyen en la segunda componente principal.

- d) No es posible interpretar los resultados anteriores debido a que es un error haber utilizado la matriz de correlación y en su lugar se debería haber utilizado la matriz de covarianzas.

4. Considere las siguientes observaciones:

$i$	1	2	3	4	5	6	7	8
$x_i$	10	8	34	9	46	68	80	50
$y_i$	4	99	44	50	77	30	25	35
$z_i$	50	31	86	57	75	14	40	60

Sin escalar las variables, describa tres iteraciones del algoritmo K-means para k = 2. Use los centroides  $C1 = (47.5, 37.5, 21.8)$  y  $C2 = (53.2, 22.4, 75.3)$ .

**Observación:** El objetivo del ejercicio es saber si comprenden como funcionan internamente el algoritmo, así que no basta con sólo la respuesta en Python u otro lenguaje. Aunque sí se pueden apoyar en algún software para los cálculos de cada paso.

5. Enuncie al menos tres diferencias entre el análisis factorial y el método de componentes principales.

6. ¿Qué significa que el método de clusterización sea jerárquico?

## Parte II

Suponga que tenemos una empresa, cuya área de marketing, desea generar una clusterización para poder tener mayor éxito con sus campañas. Para esto nos entregó una base de datos (llamada evaluación\_mkt\_campaign.csv) que contiene los siguientes campos:

1. ID: Identificador único de cliente
2. Age: Edad del cliente
3. Seniority: Días desde el enrolamiento del cliente
4. Children: Número hijos
5. Ingreso: Ingreso anual del cliente

6. Recency: Días desde la última compra del cliente
7. MntWines: Gasto en vino en los últimos 2 años
8. MntFruits: Gasto en fruta en los últimos 2 años
9. MntMeatProducts: Gasto en carne en los últimos 2 años
10. MntFishProducts: Gasto en pescado en los últimos 2 años
11. MntSweetProducts: Gasto en dulces en los últimos 2 años
12. MntGoldProds: Gasto en oro en los últimos 2 años
13. NumDealsPurchases: Número de compras realizada bajo descuento
14. NumWebPurchases: Número de compras realizadas por la web
15. NumCatalogPurchases: Número de compras realizadas por catálogo
16. NumStorePurchases: Número de compras realizadas en la tienda
17. NumWebVisitsMonth: Número de visitas a la página web

En base a estos datos, realice lo siguiente:

**Búsqueda de Anomalías:** Usando las columnas de comportamiento de compra, encuentre registros anómalos correspondientes al 5% de la muestra, y descártelos.

**Feature Extraction:** Usando los datos filtrados reduzca el número de variables. Busque una representación que le permita explicar un 90% de la varianza original de los datos.

### **Clustering:**

- a) Usando los resultados anteriores clusterice el comportamiento de sus clientes. El área de marketing solicitó que el número de clusters sea no mayor a 8. Encuentre el número óptimo de clusters considerando esa restricción
- b) Reporte los centroides de cada cluster en términos de las variables originales. Interprete.

# Requerimientos de los participantes

## Conocimientos previos:

- Inferencia estadística

## Actitudes para el trabajo:

- Cumplimiento de plazos
- Buenas prácticas de codificación
- Diseño y Estructura
- Trabajo en equipo
- Optimización del tiempo

## Valores:

- Tiempo de resolución.
- Enfoque al requerimiento.
- Estructura de Solución.

## Duración del proyecto:

1 jornada de clases

## Tips o listados de preguntas guía

- ¿Cómo clasificar?

## Productos para obtener durante la realización del proyecto

Documento de conclusión con análisis de la información.

## Especificaciones de desempeño

Deberá realizar la actividad según requerimientos técnicos y en un plazo máximo de 1 fin de semana; el resultado deberá ser un documento de conclusión con análisis que incluya las respuestas teóricas y un archivo con las respuestas a la pregunta aplicada en py.

