CASO DE ESTUDIO N.º 17 – ANÁLISIS EXPLORATORIO Y SEGMENTACIÓN DE CLIENTES MEDIANTE TÉCNICAS DE CLUSTERING

AUTORES:

- PERCY CONDORY YUCRA
- JULIO CESAR HALLASI AMBROCIO
- JOEL SANTOS GOMEZ ALANOCA

Enlace del DEMO: **https://shre.ink/DEMO-Caso17**

Enlace del repositorio: **https://github.com/ioelgomez2019/clasificacionffvv**

1. Entendiendo el negocio

1.1 Caso de estudio

El presente informe desarrolla el análisis del **'Caso de Estudio N.º 17'** relacionado con fuerza de ventas y segmentación de clientes para asignación de carteras. El objetivo es identificar grupos homogéneos de clientes mediante técnicas de clustering para apoyar toma de decisiones comerciales.

1.2 Planteamiento del problema

Objetivo del negocio: Identificar segmentos de clientes con características socio-demográficas y comportamentales diferenciadas para diseñar estrategias comerciales específicas.

Población objetivo: Registros incluidos en el **dataset procesado (N = 697**), dataset sin procesar es de 1000 registros.

1.3 Presentación del dataset y características principales

Tabla: Resumen de columnas, tipos y estadísticas básicas

Columna	Tipo	Variables / Valores	
0	Edad	Categórica	Mayor, Adulto, Joven
1	Genero	Categórica	Mujer, Hombre

2	CasaPropia	Categórica	Propia, Alquilada
3	EstadoCiv	Categórica	Soltero, Casado
4	Ubicación	Categórica	Lejos, Cerca
5	Salario	Numérica	min=10100, max=168800, mean=56103.90
6	Niños	Numérica	min=0, max=3, mean=0.93
7	Historia	Categórica	Alta, Baja, Media
8	Catalogos	Numérica	min=6, max=24, mean=14.68
9	MontoDinero	Numérica	min=38, max=6217, mean=1216.77

2. Análisis exploratorio

2.1 Variables y tipos de datos

El dataset contiene variables categóricas (Edad, Genero, CasaPropia, EstadoCiv, Ubicación, Historia) y numéricas (Salario, Niños, Catalogos, MontoDinero). Se procedió a inspección inicial de tipos, duplicados y valores nulos.

2.2 Revisión y tratamiento de missings

Reemplazo de valores NaN en 'Historia' (conteo por categoría):

Con el objetivo de manejar los valores ausentes en la columna **'Historia'**, se procedió a reemplazar los valores **NaN** por la palabra **'Nulo'**, utilizando el siguiente comando:

```
df['Historia'] = df['Historia'].fillna('Nulo')
```

De esta forma, todos los registros sin información en dicha variable fueron clasificados como **'Nulo'**.

Posteriormente, se verificó el cambio mediante el conteo de categorías:

Categoría	Frecuencia
Nulo	**303**
Alta	**255**
Baja	**230**

Media	**212**

2.3 Detección y tratamiento de outliers

Resumen IQR y conteo de outliers por variable:

	Variable	Q1	Q3	IQR	Límite inferior	Límite superior	Outliers (#)	% Outliers
0	Salario	29975.00	77025.0	47050.00	-40600.00	147600.00	1	0.1
1	Niños	0.00	2.0	2.00	-3.00	5.00	0	0.0
2	Catalogos	6.00	18.0	12.00	-12.00	36.00	0	0.0
3	MontoDinero	488.25	1688.5	1200.25	-1312.12	3488.88	27	2.7

2.4 Análisis de correlaciones (multicolinealidad)

Se calculó la matriz de correlación de Spearman para evaluar relaciones monotónicas entre variables numéricas y dummies. A continuación se incluye la matriz completa extraída del notebook.

Matriz de correlación (Spearman) completa:

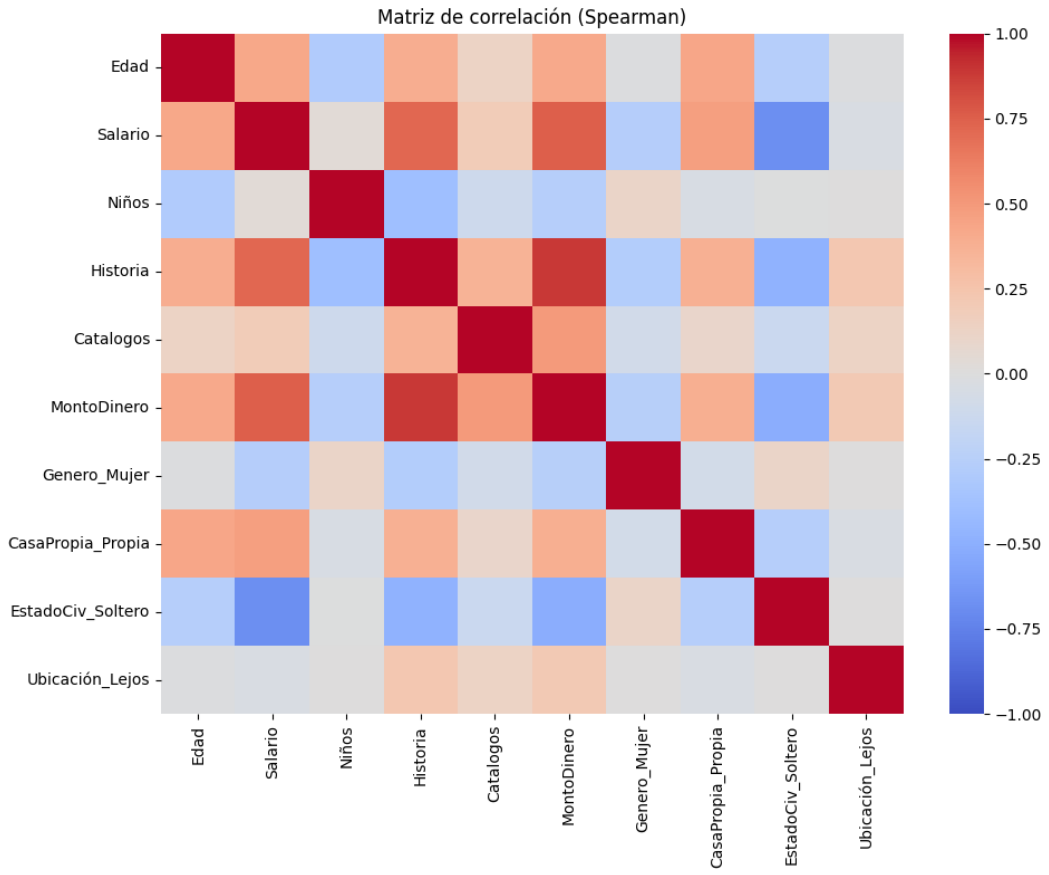


Figura: Heatmap de correlación (Spearman).

3. Procesamiento de datos (clustering)

3.1 Preparación y encoding

Se aplicaron transformaciones: reemplazo de NaN en variables categóricas, codificación (OneHotEncoder/OrdinalEncoder según pipeline) y escalado (StandardScaler) en variables numéricas. Também se aplicó winsorización por IQR para atenuar outliers.

Preprocesamiento: límites IQR y resumen de datos listos para clustering:

Límites IQR usados para winsorizar:

- Salario: [-36850.0, 149950.0]
- Niños: [-3.0, 5.0]
- Catalogos: [3.0, 27.0]

- MontoDinero: [-1373.0, 3547.0]

Datos listos para clustering:

- Registros: 697
- Variables transformadas: 14

3.2 Reducción de dimensionalidad (PCA)

Resultados PCA (varianza explicada por componentes):

PCA 3D - Varianza explicada por componente: [0.4199 0.1624 0.1237]

Varianza explicada acumulada (3D): 0.7061

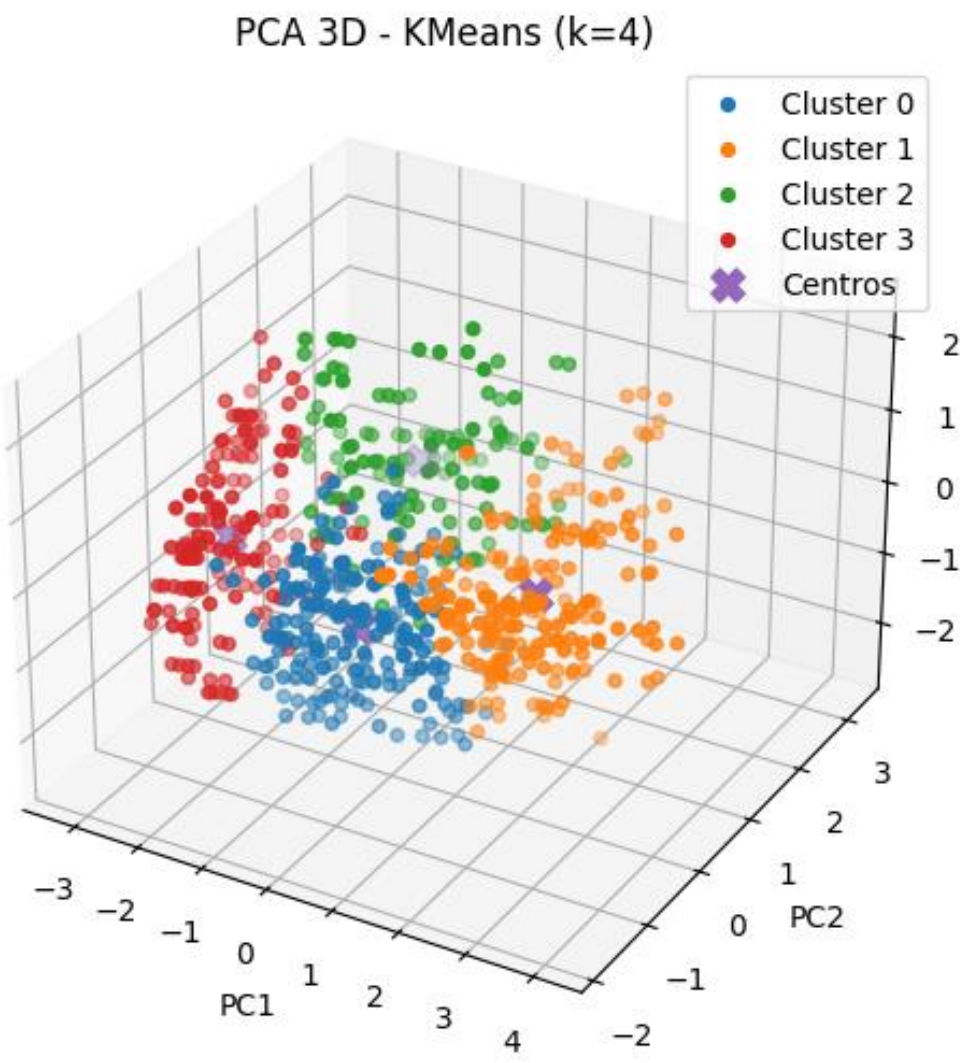


Figura: PCA 3D - proyección de los registros y separación por clústeres (si aplica).

3.3 Selección de k y algoritmo

El análisis exploratorio y criterios visuales llevaron a seleccionar k = 4 para aplicar KMeans. A continuación se presentan las métricas y el perfilamiento de clusters resultante.

Resultados de clustering: conteos por cluster, promedios numéricos y modas categóricas:

Conteo por cluster:

Cluster	Cantidad
0	203
1	200
2	126
3	168

Promedios numéricos por cluster:

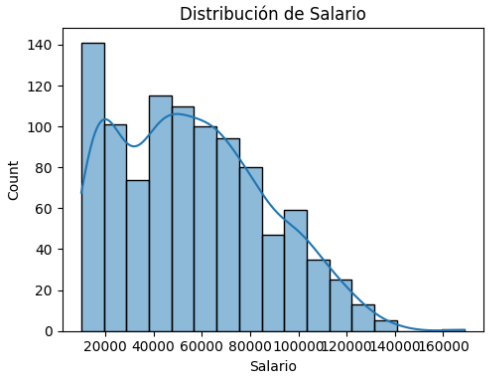
Cluster	Salario	Niños	Catálogos	Monto Dinero
0	52,306.90	0.22	14.63	1,062.36
1	88,903.75	0.57	19.56	2,329.78
2	66,852.38	2.47	13.33	611.14
3	21,564.29	0.94	11.71	370.04

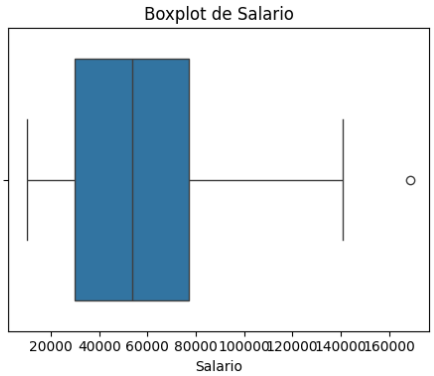
Modas de variables categóricas (ordinales):

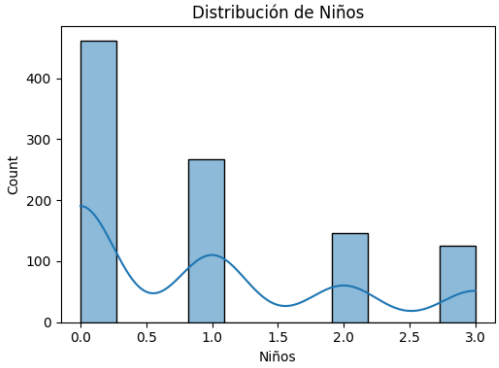
Cluster	Edad	Historia
0	Adulto	Media
1	Adulto	Alta
2	Adulto	Baja
3	Joven	Baja

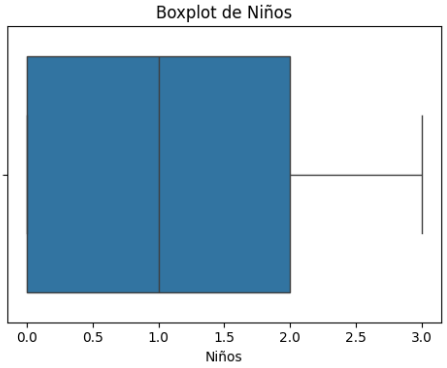
Modas de variables categóricas (nominales):

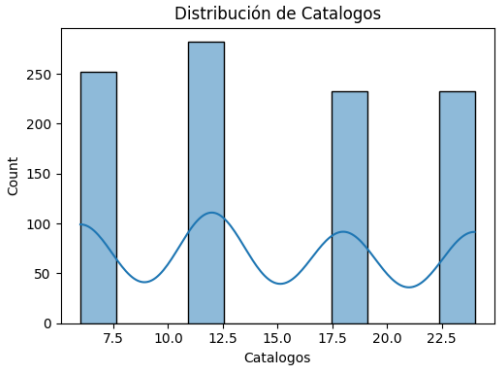
Cluster	Género	Casa Propia	Estado Civ	Ubicación
0	Hombre	Propia	Soltero	Cerca
1	Hombre	Propia	Casado	Cerca
2	Mujer	Propia	Casado	Cerca
3	Mujer	Alquilada	Soltero	Cerca

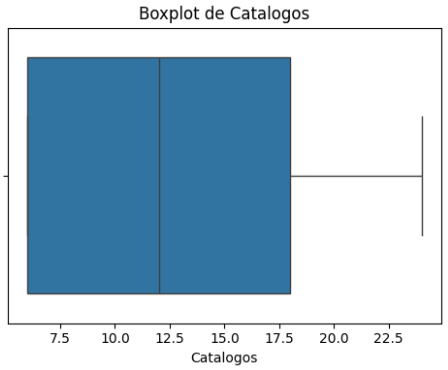

Distribución de Salario
Count
Salario
0
20
40
60
80
100
120
140
20000
40000
60000
80000
100000
120000
140000
160000

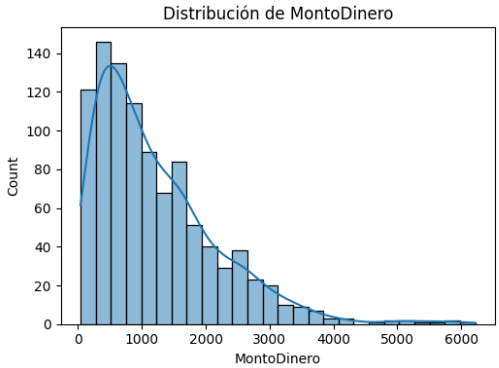

Boxplot de Salario
Salario
20000
40000
60000
80000
100000
120000
140000
160000

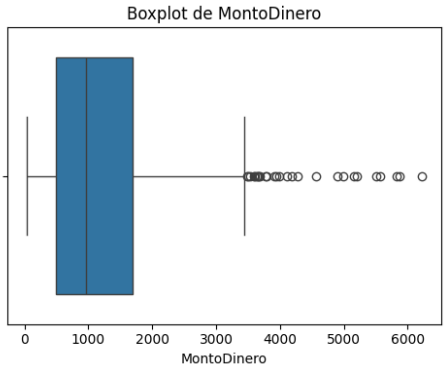

Distribución de Niños
Count
Niños
0
100
200
300
400
0.0
0.5
1.0
1.5
2.0
2.5
3.0

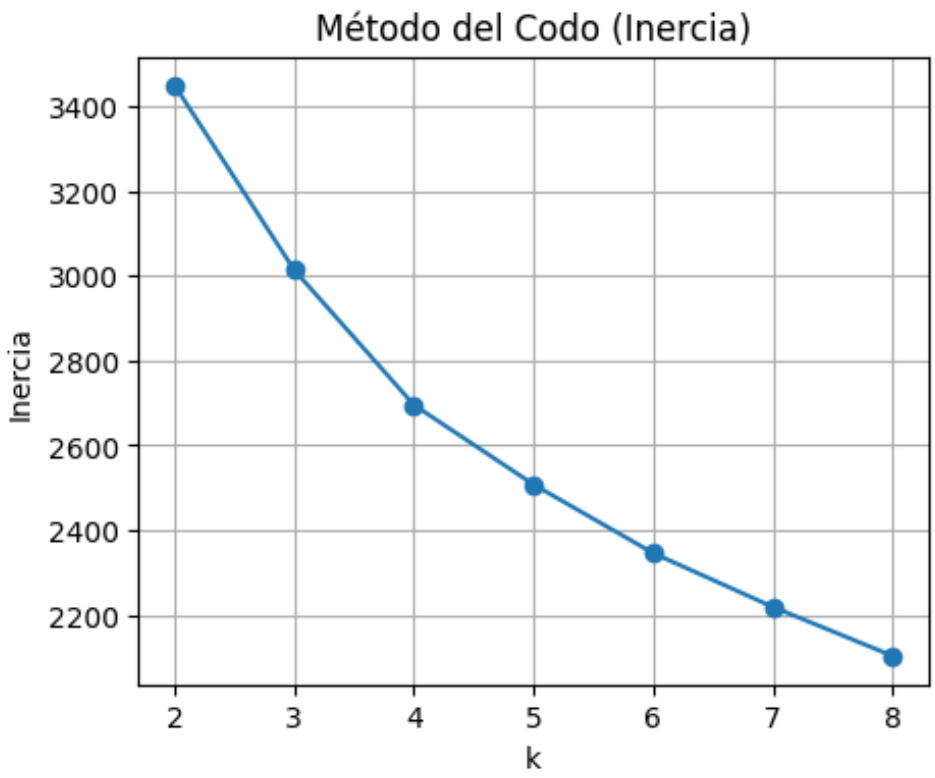

Boxplot de Niños
Niños
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Distribución de Catalogos
Count
Catalogos
0
50
100
150
200
250
7.5
10.0
12.5
15.0
17.5
20.0
22.5


Boxplot de Catalogos
Catalogos
7.5
10.0
12.5
15.0
17.5
20.0
22.5


Distribución de MontoDinero
Count
MontoDinero
0
20
40
60
80
100
120
140
0
1000
2000
3000
4000
5000
6000


Boxplot de MontoDinero
MontoDinero
0
1000
2000
3000
4000
5000
6000



4. Interpretación de resultados y conclusiones

4.1 Análisis detallado de cada cluster

A continuación, se presenta un análisis detallado de los grupos identificados en el proceso de segmentación. Los resultados permiten observar diferencias claras entre los *clusters* en términos de características demográficas, económicas y de comportamiento de consumo.

Distribución de observaciones por *cluster*

Cluster	Conteo
0	203
1	200
2	126
3	168

Interpretación:

El *cluster* con mayor cantidad de individuos es el **Cluster 0** (203 observaciones), seguido del **Cluster 1** (200). El **Cluster 2** es el más pequeño, con 126

observaciones, lo que sugiere un grupo más específico o de nicho dentro del conjunto total.

Promedios de variables numéricas por *cluster*

Cluster	Salario	Niños	Catálogos	Monto de Dinero
0	**52,306.90**	**0.22**	**14.63**	**1,062.36**
1	**88,903.75**	**0.57**	**19.56**	**2,329.78**
2	**66,852.38**	**2.47**	**13.33**	**611.14**
3	**21,564.29**	**0.94**	**11.71**	**370.04**

Interpretación:

- El **Cluster 1** presenta el **mayor nivel de ingresos** (≈ 88,900) y el **mayor gasto promedio** (≈ 2,330), lo que lo posiciona como un segmento de alto poder adquisitivo.
- El **Cluster 0** muestra ingresos medios y un gasto moderado.
- El **Cluster 2** se caracteriza por un número elevado de hijos (≈ 2.5) y un gasto bajo, lo que sugiere familias con responsabilidades económicas mayores.
- El **Cluster 3** agrupa a los individuos con **menor salario** y **menor gasto promedio**, posiblemente jóvenes en etapa inicial de vida laboral.

Modas de variables categóricas (ordinales)

Cluster	Edad	Historia
0	Adulto	Media
1	Adulto	Alta
2	Adulto	Baja
3	Joven	Baja

Interpretación:

Los *clusters* 0, 1 y 2 están formados principalmente por **adultos**, aunque difieren en su **historial crediticio**.

- El **Cluster 1** posee la **mejor historia crediticia (Alta)**, coherente con su nivel de ingresos.
- El **Cluster 2**, en cambio, muestra una **historia baja**, lo que puede limitar su acceso a productos financieros.
- El **Cluster 3** está compuesto mayoritariamente por **jóvenes con historial bajo**, un grupo emergente con potencial de crecimiento futuro.

Modas de variables categóricas (nominales)

Cluster	Género	Casa Propia	Estado Civil	Ubicación
0	Hombre	Propia	Soltero	Cerca
1	Hombre	Propia	Casado	Cerca
2	Mujer	Propia	Casado	Cerca
3	Mujer	Alquilada	Soltero	Cerca

Interpretación:

- Los **Clusters 0 y 1** están compuestos mayoritariamente por **hombres con vivienda propia**, aunque difieren en su estado civil y nivel económico.
- El **Cluster 2** reúne principalmente a **mujeres casadas con vivienda propia**, lo que puede asociarse a estabilidad familiar.
- El **Cluster 3** agrupa a **mujeres jóvenes, solteras y en vivienda alquilada**, representando posiblemente un perfil de independencia reciente o menor estabilidad económica.
- En todos los casos, la ubicación predominante es **"Cerca"**, indicando proximidad geográfica o preferencia por zonas urbanas o de fácil acceso.

4.2 Interpretación técnica

Las diferencias observadas entre los *clusters* se explican principalmente por un conjunto de variables clave: **Salario**, **MontoDinero**, **Niños** y **Historia**, complementadas por factores **sociodemográficos** como **Edad**, **Género**, **CasaPropia** y **EstadoCivil**.

Estas variables determinan patrones distintivos de capacidad económica, estructura familiar y comportamiento financiero. En conjunto, los perfiles obtenidos permiten **diseñar estrategias comerciales diferenciadas**, **ajustar la oferta de productos y servicios** según el segmento, y **optimizar la asignación de recursos** hacia los grupos de mayor potencial o rentabilidad.

4.3 Conclusiones generales

El análisis confirma que el modelo con **k = 4** genera *clusters* con **diferencias estadísticamente significativas** en variables relacionadas con la **capacidad de gasto** y la **composición familiar**.

Asimismo, el análisis de componentes principales (**PCA**) indica que **tres componentes explican aproximadamente el 70.6 % de la varianza total**, lo que garantiza una **representación adecuada de la información** y facilita la **visualización multidimensional** de los grupos.

5. Recomendaciones e implicancias prácticas

A partir de los resultados del análisis de segmentación, se proponen las siguientes **acciones operativas y estratégicas**:

1. **Diseñar campañas comerciales personalizadas** para cada *cluster*, alineadas con los perfiles definidos en el análisis.

2. **Priorizar el Cluster 1** en la oferta de productos **premium** y en estrategias de **cross-selling**, dada su mayor capacidad adquisitiva.

3. **Implementar pruebas A/B por cluster**, evaluando indicadores clave de desempeño (**KPIs**) como tasa de conversión, ticket promedio y retorno de inversión (**ROI**).

4. **Actualizar el modelo de segmentación de forma periódica** (idealmente cada seis meses), a fin de monitorear la evolución de los perfiles y adaptar las estrategias.

5. En el desarrollo de **modelos predictivos**, considerar **reducción de dimensionalidad o técnicas de regularización**, con el objetivo de **mitigar la colinealidad** entre las variables **MontoDinero**, **Salario** e **Historia**, preservando la estabilidad y precisión de los modelos.