Inteligencia comercial mediante segmentación RFM y modelos predictivos en una empresa B2B del sector eléctrico

María Isabel Cortes Cedula

Asesor

Mireya García García

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

Selco Advance Control es una empresa del sector eléctrico que, después de un año de fuerte crecimiento en 2023, enfrentó en 2024 una caída cercana al 47 % en la demanda. Esta situación llevó a la organización a cuestionar la forma en que toma decisiones comerciales y a preguntarse cómo aprovechar mejor la información que ya tiene para entender a sus clientes y planear mejor su estrategia de mercado.

Para el desarrollo del proyecto, la empresa entregó dos bases de datos: una de clientes y otra de ventas del año 2023. Tras normalizar el NIT e intentar unirlas en Python, solo fue posible encontrar siete registros en común, lo que puso en evidencia problemas de calidad e integración de datos. Ante este escenario, se decidió trabajar de manera exclusiva con la base de ventas 2023, que contiene 6.805 facturas y 545 clientes, y convertirla en insumo central para un modelo de inteligencia comercial.

Siguiendo la metodología CRISP-DM, se construyeron indicadores RFM (Recency, Frequency, Monetary) para cada cliente y se aplicó el algoritmo K-means, validando el número de grupos con el método del codo y el coeficiente de silueta. Así se identificaron cuatro segmentos de clientes con patrones de compra claramente diferenciados. A partir de los segmentos de mayor aporte se definió una variable de “alto valor” y se entrenó en Python un modelo predictivo supervisado, utilizando regresión logística para estimar la probabilidad de que cada cliente pertenezca a los grupos más valiosos.

El modelo obtuvo métricas de desempeño muy altas (exactitud superior al 95 % y AUC cercano a 0,99), lo que permitió generar un score de probabilidad de alto valor para cada cliente. La combinación entre la segmentación RFM y este modelo predictivo configura una herramienta de inteligencia comercial que ayuda a Selco Advance a priorizar sus esfuerzos de ventas, diseñar

acciones diferenciadas según el tipo de cliente y reducir el riesgo de concentración de ingresos en pocos actores.

Palabras clave: Segmentación de mercado, inteligencia comercial, RFM, K-means, regresión logística, business intelligence.

Abstract

Selco Advance Control, a company in the electrical sector, experienced strong growth in 2023 but faced a significant decline in demand of approximately 47% during 2024. This situation prompted the organization to reassess how commercial decisions are made and how existing data could be better used to understand customer behavior and guide market strategy.

For the development of this project, the company provided two datasets: a customer database and a 2023 sales database. After normalizing the NIT identifier and attempting to merge both datasets in Python, only seven matching records were found, revealing significant data quality and integration issues. As a result, the analysis focused exclusively on the 2023 sales dataset, which contains 6,805 invoices and 545 clients, transforming it into the main input for a commercial intelligence model.

Following the CRISP-DM methodology, RFM indicators (Recency, Frequency, Monetary) were constructed for each client, and the K-means algorithm was applied to perform customer segmentation. The optimal number of clusters was validated using the elbow method and the silhouette coefficient, leading to the identification of four customer segments with clearly differentiated purchasing patterns. Based on the segments with the highest contribution, a “high-value” variable was defined, and a supervised predictive model was developed in Python using logistic regression to estimate the probability that each client belongs to the most valuable groups.

The model achieved high performance metrics, with accuracy above 95% and an AUC close to 0.99, allowing the generation of a high-value probability score for each client. The integration of RFM segmentation and the predictive model constitutes a commercial intelligence tool that enables Selco Advance Control to prioritize sales efforts, design differentiated strategies

according to customer type, and reduce the risk of revenue concentration among a limited number of clients.

Keywords: Market segmentation, commercial intelligence, RFM, K-means, logistic regression, business intelligence.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Justificación

En la práctica diaria de Selco Advance, la información comercial ha estado disponible, pero no siempre ha sido aprovechada de manera estratégica. La planeación de ventas se ha apoyado sobre todo en reportes agregados y en la experiencia del equipo comercial, sin una segmentación clara de la cartera ni modelos que permitan anticipar riesgos u oportunidades. En un entorno competitivo, y con una caída reciente tan importante en la demanda, esta forma de trabajo resulta insuficiente.

Desde la literatura en analítica de clientes y CRM se ha mostrado que los modelos de segmentación y los modelos predictivos son piezas claves para transformar datos transaccionales en decisiones concretas de negocio (Ngai & Wu, 2022). El uso de algoritmos de clustering como K-means ha probado ser especialmente útil para agrupar clientes con comportamientos similares y diseñar campañas a la medida en sectores como retail, banca y comercio electrónico (Cortez, Clarke, et al, 2021; Ngai & Wu, 2022; Omol et al, 2024; Nugroho, 2024 y Tabianan, 2022).

El modelo RFM, por su parte, ofrece una forma sencilla pero poderosa de resumir el comportamiento de compra de los clientes y ha sido utilizado ampliamente para segmentar, priorizar y orientar acciones comerciales (Brei, 2020; Chambi, 2023). La regresión logística complementa este enfoque al permitir estimar probabilidades de pertenencia a un segmento, por ejemplo, clientes de alto valor o clientes en riesgo de abandono (Calle, 2023).

En el caso de Selco Advance, la situación es particular: al intentar unir la base de clientes con la base de ventas, solo se encontraron siete coincidencias, lo que revela un problema de gobernanza de datos e integración de fuentes. Aun así, la base de ventas 2023, una vez depurada, contiene información suficiente para construir indicadores RFM y alimentar un modelo de segmentación y predicción. Este Proyecto se justifica porque:

- Visibiliza la necesidad de mejorar la calidad e integración de datos en la empresa;
- Demuestra que, incluso con restricciones de información, es posible generar valor analítico usando datos transaccionales;
- Y entrega a Selco Advance un modelo concreto de inteligencia comercial basado en segmentación y scoring, listo para ser integrado en su gestión cotidiana.

Pregunta Problema

¿Cómo puede la empresa Selco Advance Control, una compañía B2B del sector eléctrico en Colombia, aprovechar sus datos históricos de ventas para segmentar estratégicamente su cartera de clientes y diseñar acciones comerciales diferenciadas y efectivas?

Dado que la integración entre la base de clientes y la base de ventas solo arrojó siete coincidencias, la pregunta se responde empíricamente a partir de la base de ventas 2023, tomando a cada cliente identificado por su NIT como unidad de análisis para la segmentación y el modelado predictivo.

Objetivos

Objetivo General

Diseñar un modelo de inteligencia comercial basado en la segmentación de clientes mediante indicadores RFM y en un modelo predictivo de alto valor, utilizando los datos de ventas 2023 de Selco Advance Control, empresa B2B del sector eléctrico en Colombia.

Objetivos Específicos

Integrar, limpiar y preparar la base de ventas 2023, construyendo indicadores RFM por cliente.

Caracterizar el comportamiento de la cartera mediante análisis exploratorio y correlaciones entre las variables RFM.

Aplicar el algoritmo K-means, apoyado en el método del codo y el coeficiente de silueta, para agrupar clientes en segmentos homogéneos en función de su comportamiento de compra.

Definir una variable objetivo que represente clientes de alto valor y entrenar un modelo predictivo supervisado de regresión logística que estime la probabilidad de pertenecer a dicho segmento.

Describir los segmentos y el modelo predictivo resultante, formulando recomendaciones comerciales diferenciadas y proponiendo un prototipo de dashboard de inteligencia comercial.

Marco Teórico

La segmentación de mercado es una práctica esencial para entender la diversidad de clientes y adaptar las estrategias comerciales a sus características reales. En empresas B2B del sector eléctrico, como Selco Advance, la ausencia de una segmentación formal dificulta distinguir entre clientes estratégicos, clientes con potencial de crecimiento y clientes de bajo aporte, lo que se traduce en uso ineficiente del tiempo comercial y en oportunidades desaprovechadas (Cortez et al, 2021; O'Brien et al, 2020).

En el contexto colombiano, la implementación de estrategias basadas en datos enfrenta retos concretos: muchas pymes no cuentan con procesos sólidos de recolección, depuración e integración de datos, lo que limita el uso de técnicas de aprendizaje automático (Contreras, et al, 2022; Ngai & Wu, 2022). Esta realidad coincide con la situación de Selco Advance, donde las bases entregadas no están alineadas entre sí, como lo demuestra el bajo número de coincidencias entre la base de clientes y la base de ventas.

Diversos autores destacan que la falta de capacidades analíticas internas también es un freno importante. Según Brei (2020) señala que la transición hacia modelos basados en analítica exige no solo herramientas tecnológicas, sino también talento capaz de interpretar algoritmos y traducir resultados en decisiones de negocio. Calle (2023) y Chambi (2023) muestran que, cuando la analítica se integra correctamente en marketing, mejora la precisión de la segmentación y el retorno de las campañas, incluso en contextos complejos como los que dejó la pandemia.

El modelo RFM (Recency, Frequency, Monetary) se ha consolidado como una de las formas más accesibles de describir el comportamiento de compra. Recency indica cuán reciente fue la última compra; Frequency, cuántas veces compra el cliente; y Monetary, cuánto gasta.

Esta triada permite identificar, por ejemplo, clientes recientes y de alto valor, clientes activos, pero de bajo gasto, y clientes que se encuentran prácticamente inactivos (Nugroho, 2024; Omol et al., 2024; Tabianan, 2022).

Sobre esta base se pueden aplicar algoritmos de clustering. K-means es uno de los más utilizados por su equilibrio entre simplicidad e interpretabilidad: agrupa observaciones tratando de minimizar la distancia entre cada cliente y el centro de su grupo. En la práctica, elegir el número adecuado de clusters suele apoyarse en técnicas como el método del codo y el coeficiente de silueta, que ayudan a encontrar un punto de equilibrio entre detalle y claridad (Ngai & Wu, 2022; Gómez y Rodríguez, 2021; Bernal Ospina, 2024; Martínez y Pérez, 2019).

Por otro lado, los modelos predictivos supervisados, como la regresión logística, permiten dar un paso más allá de la descripción y estimar probabilidades asociadas a eventos relevantes para la empresa: la pérdida de un cliente, la compra de una nueva línea de productos o la pertenencia a un segmento de alto valor (Ngai & Wu, 2022; Brei, 2020). La regresión logística es especialmente útil porque sus resultados pueden interpretarse de manera relativamente sencilla: un aumento en la frecuencia de compra o en el valor monetario, por ejemplo, se traduce en un incremento en la probabilidad de que el cliente sea considerado “alto valor”.

Finalmente, la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) ofrece un marco estructurado para organizar este tipo de proyectos. Sus fases, comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue; permiten conectar las decisiones técnicas con los objetivos de negocio y garantizan que el análisis no se limite a correr algoritmos, sino que concluya en propuestas de implementación concretas (Shearer, 2000; Cobos et al., 2010).

Metodología

Se aplicó la metodología CRISP-DM, con las siguientes fases

Comprensión del Negocio

Durante la fase de comprensión del negocio, es requerido la identificación y alimentación del modelo, proseguido de la ejecución de la metodología CRISP-DM, en las figuras 1 a 4 se presentan las capturas del código inicial aplicado. (Figura 1,Figura 2,Figura 3,Figura 4).

Figura 1

Pre-Carga de Datos – Colab

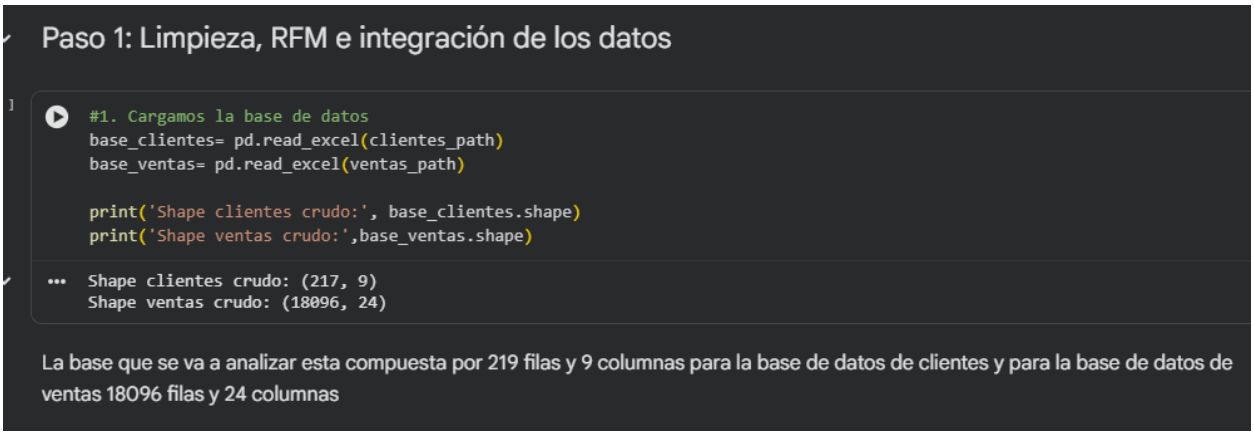


Figura 2

Limpieza de Base

```
#1.1 Realizamos limpieza de la base de clientes

# Recargamos la base_clientes desde el archivo, indicando que los nombres de columna
# están en la fila con índice 3 (la cuarta fila del archivo Excel, 0-indexada).
# Esto automáticamente ignora las filas anteriores y usa la fila 3 como encabezado.
base_clientes = pd.read_excel(clientes_path, header=3)

# La variable 'clientes' ahora se usará para el DataFrame limpio.
# Realizamos una copia para evitar SettingWithCopyWarning en operaciones posteriores.
clientes = base_clientes.copy()

# Eliminamos las columnas que estén completamente vacías.
clientes = clientes.dropna(axis=1, how='all').copy()

# Si hay columnas duplicadas, nos quedamos con la primera aparición.
clientes = clientes.loc[:, ~clientes.columns.duplicated()].copy()

# Renombramos las columnas a minúsculas para consistencia.
# Asegúrate de que los nombres de columna existentes coincidan con las claves del diccionario.
clientes = clientes.rename(columns={
    "EMPRESA": "empresa",
    "NIT": "nit",
    "CONTACTO": "contacto",
    "DIRECCION": "direccion",
    "TELEFONO": "telefono",
    "CORREO": "correo",
    "LUGAR": "ciudad"
}).copy()

# Quitamos las filas que quedaron completamente vacías después de las operaciones anteriores.
clientes = clientes.dropna(how='all').copy()

print(clientes.head())
print(clientes.columns)
```

```
                  empresa           nit                  contacto  \
0  Instrumentos y Controles   860.031068-3            ING JAIRO ARIZA
1                 Seingel  830.013.692-4         ING GUSTAVO GARRIDO
2                   Tecna  860.080.005-9  ING ANGELA PATRICIA BENITEZ
3                 Seringel  800.055.990-5       ING CRISTIAN FERREIRA
4  Grupo Ind. Metalmecanico  830.005.424-3       ING JENNY PAOLA SIERRA

                    direccion telefono                    correo  ciudad  \
0            CALLE 39 NO 24-45  2088080       jariza@iycsa.com.co  Bogota
1               CR 73 NO 48-92  5489221     seingel@seingel.com.co  Bogota
2            CR 32 NO 28 D-60  3684311       compras@tecna.com.co  Bogota
3  Carrera 7 No 180-75 modulo 3  6771569     cferreira@seringel.com  Bogota
4              CR 73A-68B-28  2526888  gim_proyectos3@etb.net.co  Bogota

     Unnamed: 8
0  leutron, GMI
1           NaN
2           NaN
3           NaN
4   leutron, JM
Index(['empresa', 'nit', 'contacto', 'direccion', 'telefono', 'correo',
       'ciudad', 'Unnamed: 8'],
      dtype='object')
```

Figura 3

Resultado de la Limpieza

```
clientes.head()
```

	empresa	nit	contacto	direccion	telefono	correo	ciudad	Unnamed: 8	NIT_norm
0	Instrumentos y Controles	860.031068-3	ING JAIRO ARIZA	CALLE 39 NO 24-45	2088080	jariza@iycsa.com.co	Bogota	leutron, GMI	8600310683
1	Seingel	830.013.892-4	ING GUSTAVO GARRIDO	CR 73 NO 48-92	5489221	seingel@seingel.com.co	Bogota	NaN	8300136924
2	Tecna	860.080.005-9	ING ANGELA PATRICIA BENITEZ	CR 32 NO 28 D-60	3684311	compras@tecna.com.co	Bogota	NaN	8600800059
3	Seringel	800.055.990-5	ING CRISTIAN FERREIRA	Carrera 7 No 180-75 modulo 3	6771569	cferreira@seringel.com	Bogota	NaN	8000559905
4	Grupo Ind. Metalmecanico	830.005.424-3	ING JENNY PAOLA SIERRA	CR 73A-68B-28	2526888	gim_proyectos3@etb.net.co	Bogota	leutron, JM	8300054243

```
#eliminamos la columna unnamed

# Eliminamos cualquier columna cuyo nombre empiece por 'Unnamed'
clientes = clientes.loc[:, ~clientes.columns.str.contains('^Unnamed')]

print(clientes.columns)
clientes.head()
```

```
Index(['empresa', 'nit', 'contacto', 'direccion', 'telefono', 'correo',
       'ciudad', 'NIT_norm'],
      dtype='object')
```

	empresa	nit	contacto	direccion	telefono	correo	ciudad	NIT_norm
0	Instrumentos y Controles	860.031068-3	ING JAIRO ARIZA	CALLE 39 NO 24-45	2088080	jariza@iycsa.com.co	Bogota	8600310683
1	Seingel	830.013.892-4	ING GUSTAVO GARRIDO	CR 73 NO 48-92	5489221	seingel@seingel.com.co	Bogota	8300136924
2	Tecna	860.080.005-9	ING ANGELA PATRICIA BENITEZ	CR 32 NO 28 D-60	3684311	compras@tecna.com.co	Bogota	8600800059
3	Seringel	800.055.990-5	ING CRISTIAN FERREIRA	Carrera 7 No 180-75 modulo 3	6771569	cferreira@seringel.com	Bogota	8000559905
4	Grupo Ind. Metalmecanico	830.005.424-3	ING JENNY PAOLA SIERRA	CR 73A-68B-28	2526888	gim_proyectos3@etb.net.co	Bogota	8300054243

En esta fase se revisó el contexto de Selco Advance Control: caída cercana al 47 % en demanda en 2024, alta dependencia de ciertos clientes de gran tamaño y ausencia de una segmentación formal que oriente la gestión del equipo comercial. Se definió la pregunta problema y se acordó con la empresa que el objetivo principal del proyecto sería:

segmentar a los clientes a partir de su comportamiento de compra;

y construir un modelo que estime la probabilidad de que un cliente pertenezca al segmento de alto valor, para priorizarlo en la planificación comercial.

Comprensión de los Datos

Figura 4

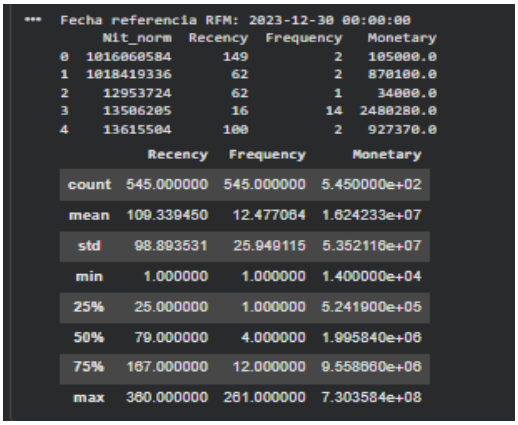
Normalizar y Guardar Datos

```
#1.2limpieza de la base de datos de ventas


ventas = base_ventas.copy()

# Revisar columnas principales
print(base_ventas.columns)

# Parsear fecha de factura
ventas["Fecha Fact"] = pd.to_datetime(ventas["Fecha Fact"], errors="coerce")

# Mantener solo registros con fecha válida
ventas = ventas[ventas["Fecha Fact"].notna()]

# Asegurar que el total sea numérico
ventas["Total"] = pd.to_numeric(ventas["Total"], errors="coerce")
ventas = ventas[ventas["Total"].notna()]

# Normalizar Nit Cliente para cruzar con clientes
ventas["Nit_norm"] = (
    ventas["Nit Cliente"]
    .astype(str)
    .str.replace(".", "", regex=False)
    .str.replace("-", "", regex=False)
    .str.strip()
)

print(ventas[["Fecha Fact", "Nit Cliente", "Nit_norm", "Total"]].head())
```

```
Index(['Prefijo Fact', 'Nro Fact', 'Fecha Fact', 'Nit Cliente',
       'Nombre Cliente', 'Codigo Ref', 'Cod Bodega', 'Nombe Ref',
       'Nombre Linea', 'Total', 'Iva', 'Neto', 'Costo', 'Cod Vendedor',
       'Nombre Vendedor', 'Estado Fact', 'Nombre Marca', 'Costo Producto',
       'Cantidad', 'Existencia Actual', 'MES', 'AÑO', 'Pc. Venta', 'Margen'],
      dtype='object')
  Fecha Fact  Nit Cliente    Nit_norm       Total
0 2023-02-09  811026728-8  8110267288         0.0
1 2023-09-19  811039383-7  8110393837         1.0
2 2023-01-06  800203838-9  8002038389    182000.0
3 2023-02-08  900090874-1  9000908741     52000.0
4 2023-12-08  890937250-6  8909372506  22170600.0
```

Figura 5

Construcción de Variables

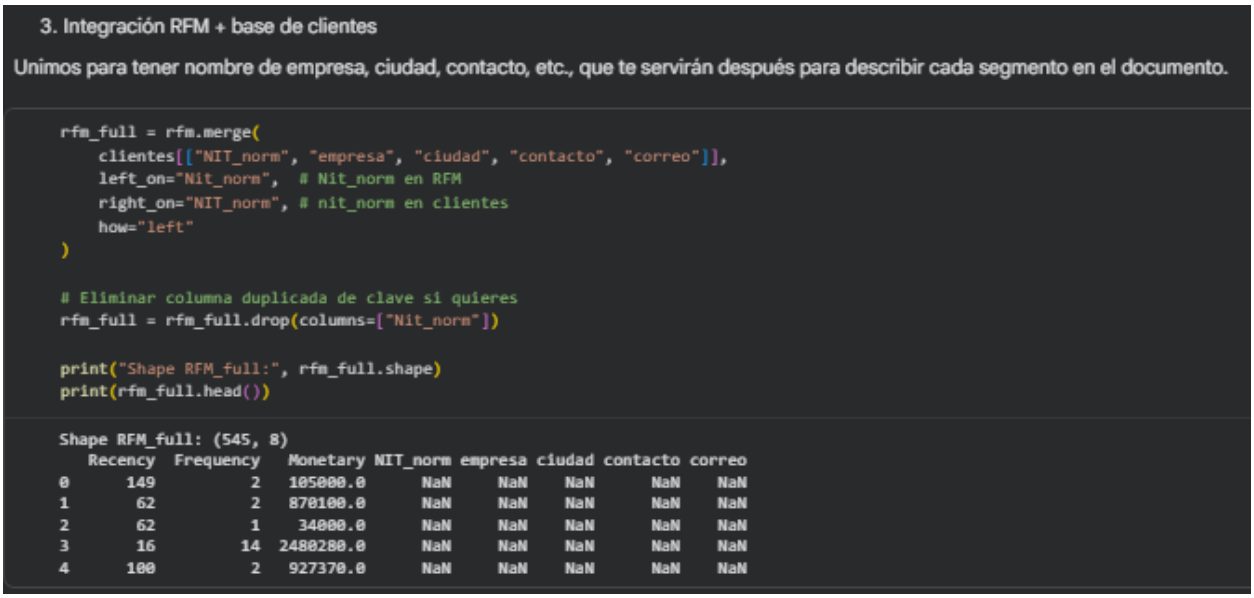


```
***  Fecha referencia RFM: 2023-12-30 00:00:00
        Nit_norm  Recency  Frequency   Monetary
0  1016060584      149          2   105000.0
1  1018419336       62          2   870100.0
2    12953724       62          1    34000.0
3    13506205       16         14  2480280.0
4    13615504      100          2   927370.0
```

	Recency	Frequency	Monetary
count	545.000000	545.000000	5.450000e+02
mean	109.339450	12.477064	1.624233e+07
std	98.893531	25.949115	5.352116e+07
min	1.000000	1.000000	1.400000e+04
25%	25.000000	1.000000	5.241900e+05
50%	79.000000	4.000000	1.995840e+06
75%	167.000000	12.000000	9.558660e+06
max	380.000000	281.000000	7.303584e+08

Figura 6

Integración de RFM + Base Clientes



3. Integración RFM + base de clientes

Unimos para tener nombre de empresa, ciudad, contacto, etc., que te servirán después para describir cada segmento en el documento.

```
rfm_full = rfm.merge(
    clientes[["NIT_norm", "empresa", "ciudad", "contacto", "correo"]],
    left_on="Nit_norm",  # Nit_norm en RFM
    right_on="NIT_norm", # nit_norm en clientes
    how="left"
)

# Eliminar columna duplicada de clave si quieres
rfm_full = rfm_full.drop(columns=["Nit_norm"])

print("Shape RFM_full:", rfm_full.shape)
print(rfm_full.head())
```

```
Shape RFM_full: (545, 8)
   Recency  Frequency   Monetary NIT_norm empresa ciudad contacto correo
0      149          2   105000.0      NaN     NaN    NaN      NaN    NaN
1       62          2   870100.0      NaN     NaN    NaN      NaN    NaN
2       62          1    34000.0      NaN     NaN    NaN      NaN    NaN
3       16         14  2480280.0      NaN     NaN    NaN      NaN    NaN
4      100          2   927370.0      NaN     NaN    NaN      NaN    NaN
```

Figura 7

Resultado Validación por NIT

```
# 1. ¿Cuántos NIT únicos hay en cada base?
print("NIT únicos en clientes :", clientes["nit_norm"].nunique())
print("NIT únicos en ventas   :", ventas["nit_norm"].nunique())

# 2. Intersección de NIT entre ambas tablas
set_clientes = set(clientes["nit_norm"].dropna())
set_ventas   = set(ventas["nit_norm"].dropna())

interseccion = set_clientes & set_ventas
print("NIT que aparecen en las dos bases:", len(interseccion))

# 3. Ver algunos NIT que sí coinciden
print("Algunos NIT en común:", list(interseccion)[:20])
```

```
NIT únicos en clientes : 30
NIT únicos en ventas   : 545
NIT que aparecen en las dos bases: 7
Algunos NIT en común: ['8001498711', '8040160727', '8600800059', '8600310683', '8301228358', '8300054243', '8301312816']
```

La empresa entregó dos archivos

- Base de clientes 2023
- Base de ventas 2023 (facturas, fechas, NIT, valores, líneas de producto).

En Python se normalizó el NIT en ambas bases (eliminando puntos, guiones y espacios) y se intentó realizar un join por este identificador. El resultado fue solo siete coincidencias, lo que evidenció problemas de gobernanza y calidad de datos en la organización.

Ante esta situación, y dado el alcance del trabajo de grado, se decidió acotar el análisis empírico a la base de ventas 2023, manteniendo la base de clientes como contexto cualitativo, pero no como fuente cuantitativa principal. Esta decisión se documenta como hallazgo clave y como línea prioritaria de mejora para la organización.

Con la base de ventas se elaboró un panorama general: número de facturas, número de clientes, ventas totales, ventas por mes y concentración de ingresos en los principales clientes. Se

observó una alta concentración: un conjunto reducido de clientes representa una proporción muy relevante del ingreso anual.

Preparación de los Datos

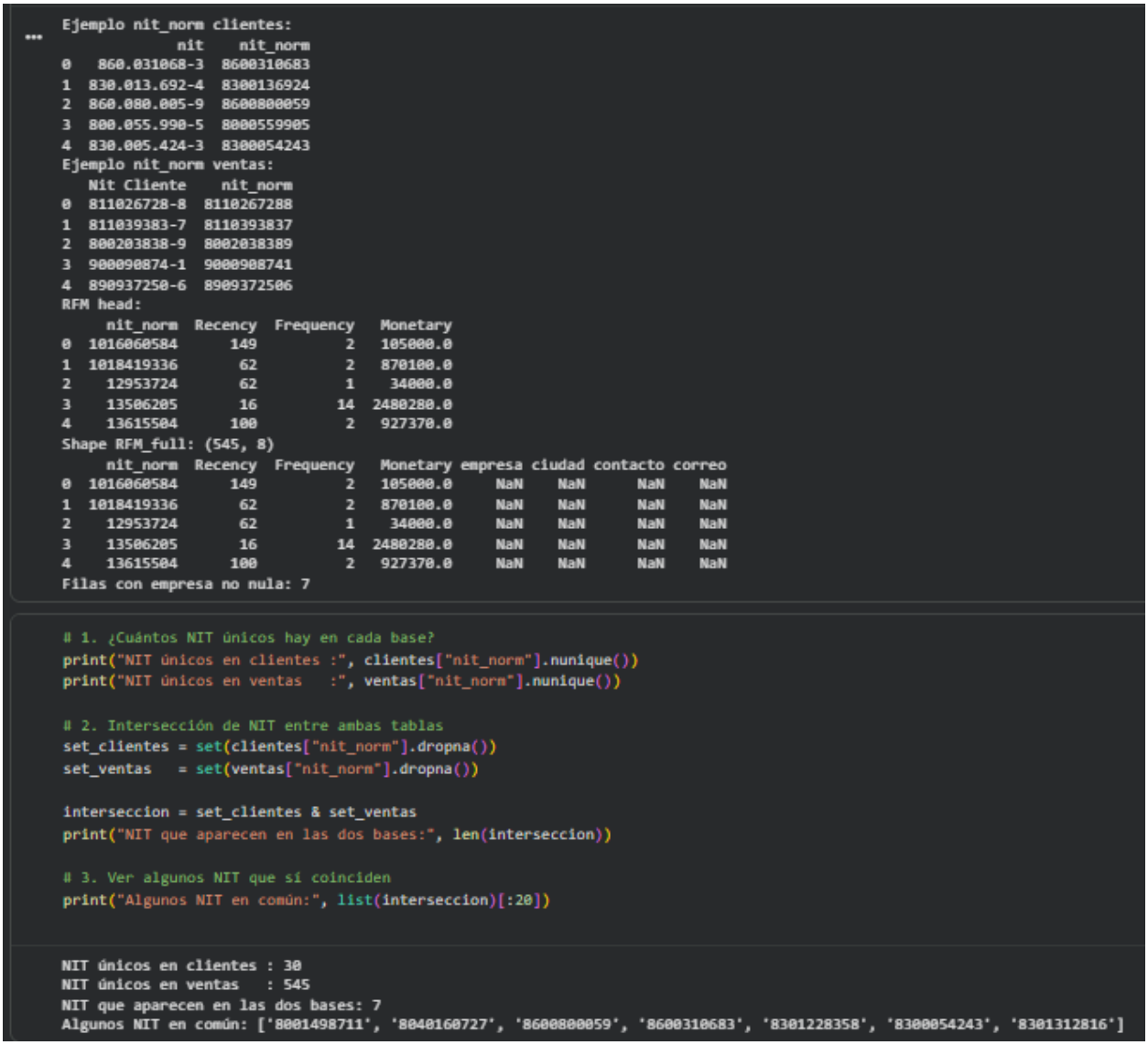
Figura 8

Normalización de los Datos

```
# ============================revisar por que salen datos nulos
# 1. Normalizar NIT en CLIENTES
# ============================
clientes["nit_norm"] = (
    clientes["nit"]
    .astype(str)
    .str.replace(".", "", regex=False)
    .str.replace("-", "", regex=False)
    .str.strip()
)

print("Ejemplo nit_norm clientes:")
print(clientes[["nit", "nit_norm"]].head())

# ============================
# 2. Normalizar NIT en VENTAS
# ============================
ventas["nit_norm"] = (
    ventas["Nit Cliente"]
    .astype(str)
    .str.replace(".", "", regex=False)
    .str.replace("-", "", regex=False)
    .str.strip()
)

print("Ejemplo nit_norm ventas:")
print(ventas[["Nit Cliente", "nit_norm"]].head())

# ============================
# 3. RFM usando nit_norm
# ============================
fecha_ref = ventas["Fecha Fact"].max() + pd.Timedelta(days=1)

rfm = ventas.groupby("nit_norm").agg(
    Recency   = ("Fecha Fact", lambda x: (fecha_ref - x.max()).days),
    Frequency = ("Nro Fact", "nunique"),
    Monetary  = ("Total", "sum")
).reset_index()

print("RFM head:")
print(rfm.head())

# ============================
# 4. MERGE RFM + CLIENTES
# ============================
rfm_full = rfm.merge(
    clientes[["nit_norm", "empresa", "ciudad", "contacto", "correo"]],
    on="nit_norm",    # misma llave en ambos
    how="left"
)

print("Shape RFM_full:", rfm_full.shape)
print(rfm_full.head())

# Comprobar cuántos clientes quedaron con datos de empresa
print("Filas con empresa no nula:", rfm_full["empresa"].notna().sum())
```

Figura 9

Resultado Normalización de Datos



```
Ejemplo nit_norm clientes:
              nit    nit_norm
0   860.031068-3  8600310683
1  830.013.692-4  8300136924
2  860.080.005-9  8600800059
3  800.055.990-5  8000559905
4  830.005.424-3  8300054243
Ejemplo nit_norm ventas:
   Nit Cliente    nit_norm
0  811026728-8  8110267288
1  811039383-7  8110393837
2  800203838-9  8002038389
3  900090874-1  9000908741
4  890937250-6  8909372506
RFM head:
     nit_norm  Recency  Frequency   Monetary
0  1016060584      149          2   105000.0
1  1018419336       62          2   870100.0
2    12953724       62          1    34000.0
3    13506205       16         14  2480280.0
4    13615504      100          2   927370.0
Shape RFM_full: (545, 8)
     nit_norm  Recency  Frequency   Monetary empresa ciudad contacto correo
0  1016060584      149          2   105000.0     NaN    NaN      NaN    NaN
1  1018419336       62          2   870100.0     NaN    NaN      NaN    NaN
2    12953724       62          1    34000.0     NaN    NaN      NaN    NaN
3    13506205       16         14  2480280.0     NaN    NaN      NaN    NaN
4    13615504      100          2   927370.0     NaN    NaN      NaN    NaN
Filas con empresa no nula: 7
```

```
# 1. ¿Cuántos NIT únicos hay en cada base?
print("NIT únicos en clientes :", clientes["nit_norm"].nunique())
print("NIT únicos en ventas   :", ventas["nit_norm"].nunique())

# 2. Intersección de NIT entre ambas tablas
set_clientes = set(clientes["nit_norm"].dropna())
set_ventas   = set(ventas["nit_norm"].dropna())

interseccion = set_clientes & set_ventas
print("NIT que aparecen en las dos bases:", len(interseccion))

# 3. Ver algunos NIT que sí coinciden
print("Algunos NIT en común:", list(interseccion)[:20])
```

```
NIT únicos en clientes : 30
NIT únicos en ventas   : 545
NIT que aparecen en las dos bases: 7
Algunos NIT en común: ['8001498711', '8040160727', '8600800059', '8600310683', '8301228358', '8300054243', '8301312816']
```

En Python (Google Colab) se ejecutó la limpieza y transformación de la base de ventas

1. Conversión del campo Fecha Fact a tipo datetime, con eliminación de registros con fechas inválidas.

2. Conversión del campo Total a numérico, filtrando entradas vacías o no numéricas.

3. Construcción de un identificador normalizado nit_norm a partir del NIT del cliente.

4. Definición de una fecha de referencia (un día después de la última compra registrada).

5. Cálculo, por cada nit_norm, de los indicadores RFM:

Recency: días desde la última compra hasta la fecha de referencia.

Frequency: número de facturas únicas en 2023.

Monetary: suma del valor total de las compras en el año.

El resultado fue un DataFrame rfm con 545 clientes y sus respectivos indicadores RFM, listo para el análisis exploratorio y el modelado. Estas operaciones se realizaron en Python por su capacidad para manejar volúmenes de datos, reproducir pasos de limpieza y facilitar la aplicación de algoritmos de machine learning de forma transparente y replicable.

Modelado

El modelado es clave porque traduce los datos históricos de ventas en un modelo predictivo de inteligencia comercial accionable para Selco Advance. Hasta este punto se había logrado comprender el negocio, revisar la calidad de la información y construir indicadores RFM por cliente; sin embargo, sin un modelo formal que relacione estos indicadores con el valor de los clientes, la empresa seguiría dependiendo de informes agregados y de la intuición del equipo comercial. El propósito del análisis que se desarrolla en este punto es, precisamente, transformar los datos transaccionales en reglas y probabilidades que permitan priorizar clientes, anticipar oportunidades y reducir el riesgo de concentración de ventas, en línea con lo que la literatura plantea sobre el uso de analítica y machine learning en marketing y CRM.

El modelado se organizó siguiendo la metodología CRISP-DM, que propone avanzar de forma iterativa desde la comprensión del negocio hasta el despliegue de soluciones, garantizando que cada decisión técnica responda a objetivos comerciales concretos. En primer lugar, el análisis exploratorio y la matriz de correlaciones permiten verificar que las variables RFM resumen de manera adecuada el comportamiento de compra y están relacionadas entre sí de forma coherente (por ejemplo, clientes más frecuentes tienden a generar mayor valor monetario, mientras que una alta recencia suele asociarse con menor frecuencia y menor aporte). Esta etapa es indispensable para justificar el uso posterior de RFM como base del modelo y para asegurar que no se está construyendo el modelo sobre relaciones espurias.

En segundo lugar, se aplica K-means sobre las variables RFM, apoyado en el método del codo y el coeficiente de silueta, con el fin de descubrir segmentos de clientes con patrones de compra homogéneos. Esta decisión responde a dos razones: por un lado, la evidencia empírica y la literatura muestran que la combinación RFM + clustering es una práctica robusta y ampliamente utilizada para segmentar clientes en distintos sectores; por otro lado, la empresa necesitaba una segmentación fácilmente explicable al equipo comercial (clientes VIP, de alto valor, regulares, inactivos), algo que K-means y los perfiles promedio por cluster permiten comunicar de manera sencilla.

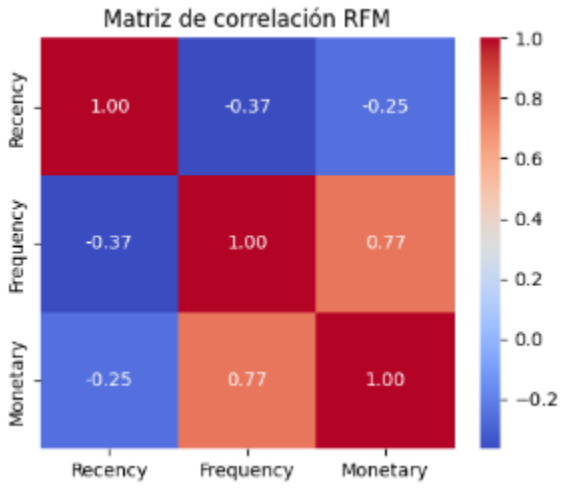
Finalmente, a partir de los segmentos identificados se define una variable objetivo de “alto valor” y se entrena un modelo supervisado de regresión logística que estima la probabilidad de que cada cliente pertenezca a dicho grupo. Esta elección metodológica prioriza la interpretabilidad y la facilidad de implementación: la regresión logística ofrece buenas métricas de desempeño con el volumen de datos disponible, permite entender el efecto de cada componente RFM en la probabilidad de alto valor y puede integrarse de forma sencilla en un

dashboard de inteligencia comercial. De esta manera, el análisis desarrollado en el punto 4 no es solo un ejercicio estadístico, sino la pieza central del modelo predictivo de inteligencia comercial, que conecta la segmentación de clientes con la toma de decisiones sobre a quién contactar primero, qué tipo de campaña ofrecer y dónde concentrar los recursos de la fuerza de ventas.

Análisis Exploratorio y Correlaciones

Figura 10

Visualización de la Correlación



Se calculo la matriz de correlaciones de Pearson entre las variables del modelo RFM(Recency, Frecuency y Monetary). Los resultados muestran una correlación positiva alta entre Frequency y Monetary (r aproximado 0.77), lo que indica que los clientes que compran con mayor frecuencia tienden también a generar un mayor de ventas anual.

Entre Recency y Frecuency se observa una correlación negativa moderada (r aprox. 0.37): A mayor número de días desde la última compra, menor es la frecuencia de compra del

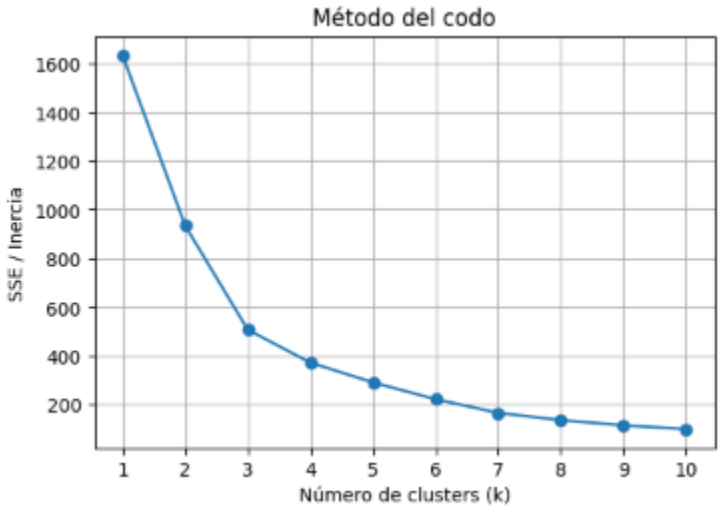
cliente. Es decir, quienes llevan más tiempo sin comprar son, en general, clientes poco recurrentes.

Por su parte, la correlación entre Recency y Monetary es negativa débil (r aproximado -0.25). Esto sugiere que los clientes con mayor nivel de gasto tienden a ser algo más recientes que el promedio, aunque la relación no es tan fuerte como en el caso de Frequency- Monetary

En resumen, los resultados son acordes con el comportamiento esperado en un esquema RFM: Las compras que aporta un cliente (Frequency y Monetary) está estrechamente asociada, mientras que la recencia de la última compra se relaciona de forma inversa con su frecuencia y valor, reflejando el riesgo de inactividad cuando aumenta el tiempo desde la última transacción.

Figura 11

Gráfico Método del Codo



Para determinar el número adecuado de clusters se aplicó el método del codo sobre las variables estandarizadas del modelo RFM.

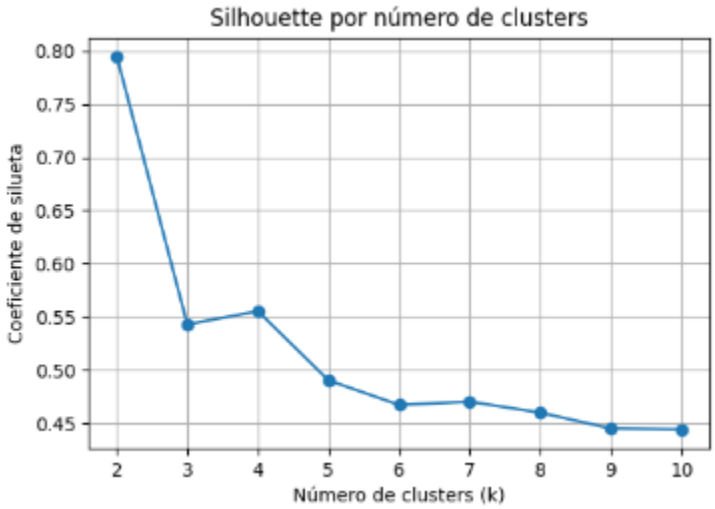
En la gráfica se observa que la suma de cuadrados intra-cluster (SSE/Inercia) disminuye fuertemente entre K=1 (aproximado 1600) k=2, y vuelve a caer de manera importante al pasar a k=3 (aproximado 400).

A partir del k > 4 la reducción de la inercia es cada vez menor y la curva se vuelve más plana, lo que indica rendimientos decrecientes al seguir aumentando el número de grupos.

Es así y gracias a este comportamiento que sugiere el método de codo que la curva se encuentra alrededor de k=3, por lo que se considera tres clusters como un número adecuado de segmentos, nos ayuda a mejorar sustancialmente la compactación de los grupos frente a 1 o 2 clusters, sin introducir una complejidad excesiva en el modelo. Este valor será contrastado posteriormente con el coeficiente de silueta y con criterios de interpretación del negocio.

Figura 12

Modelo de Silueta (Silhouette)



Para complementar el método de codo, se calculó el coeficiente de silueta para diferentes valores de K (entre 2 y 10 clusters). Los resultados muestran que el mayor valor de silueta se

obtiene con K=2 (S aproximado de 0.80), lo que indica una excelente separación entre dos grupos de clientes.

Sin embargo, desde el punto de vista de negocio, una segmentación en solo dos grupos resulta un poco sesgada o demasiado general para diferenciar estrategias comerciales; por ello se analizaron valores de k superiores, Entre los modelos con tres o más clusters, el mejor desempeño se alcanza con k=4 que presenta un coeficiente de silueta aproximado de 0.55, ligeramente superior al de k=3 (s = 0.54). A partir de K> 5 el coeficiente desciende y se estabiliza alrededor de 0.45, lo que indica que aumentar el número de clusters no mejora la calidad de la partición y solo fragmenta innecesariamente los grupos.

Por lo tanto, combinando la evidencia del método de codo (que sugiere un punto de inflexión en torno a 3-4 grupos) y los resultados de coeficiente de silueta, se selecciona K=4 clusters como solución final de segmentación, por ofrecer un buen equilibrio entre calidad estadística de los grupos y nivel de detalle útil para la toma de decisiones comerciales.

En conclusión, se calcularon estadísticas descriptivas de RFM y la matriz de correlaciones de Pearson. Los resultados mostraron: Correlación positiva alta entre Frequency y Monetary (clientes que compran más veces tienden a comprar mayor valor); correlación negativa moderada entre Recency y las otras dos variables (clientes que llevan mucho tiempo sin comprar suelen ser menos frecuentes y de menor aporte).

Estos hallazgos son consistentes con la teoría RFM y confirman que estas variables son adecuadas para segmentar la cartera.

Clustering RFM con K-Means

Para la segmentación, se siguió el siguiente procedimiento:

1. Estandarización de las variables RFM mediante StandardScaler para que tuvieran media 0 y desviación estándar 1, evitando que la escala de Monetary dominara el cálculo de distancias.

2. Aplicación del método del codo: se entrenó K-means para k entre 1 y 10 y se analizó la evolución de la suma de cuadrados intra-cluster (SSE). Se observó una disminución pronunciada hasta k=3–4 y mejoras marginales a partir de ese punto.

3. Cálculo del coeficiente de silueta para k entre 2 y 10. Los valores más altos y estables se presentaron para k=4, indicando una buena separación entre grupos y una estructura útil para fines comerciales.

4. Entrenamiento definitivo de K-means con k=4, asignando a cada cliente un cluster entre 0 y 3.

5. Cálculo de perfiles promedio por cluster (Recency, Frequency, Monetary, número de clientes, ventas aproximadas asociadas) y construcción de un mapa de calor de medias normalizadas para facilitar la interpretación.

Definición de Clusters y Variable Objetivo

Figura 13

Técnicas de Clustering (K-Means)

```
#Nuestro valor de k optimo es de 4

k_optimo = 4

kmeans_final = KMeans(n_clusters=k_optimo, random_state=42, n_init=10)
rfm_features_scaled = X_scaled.copy()
cluster_labels = kmeans_final.fit_predict(rfm_features_scaled)


# Añadimos el cluster al DataFrame original RFM
rfm["cluster"] = cluster_labels
print(rfm.head())
rfm["cluster"].value_counts()
```

```
      nit_norm  Recency  Frequency   Monetary  cluster
0  1016060584      149          2   105000.0        0
1  1018419336       62          2   870100.0        0
2    12953724       62          1    34000.0        0
3    13506205       16         14  2480280.0        0
4    13615504      100          2   927370.0        0
```

cluster	count
0	352
1	147
3	36
2	10

dtype: int64

```
#Perfil de cada cluster

cluster_profile = rfm.groupby("cluster").agg(
    n_clientes = ("nit_norm", "count"),
    Recency_mean   = ("Recency", "mean"),
    Frequency_mean = ("Frequency", "mean"),
    Monetary_mean  = ("Monetary", "mean")
).reset_index()

print(cluster_profile)
```

```
   cluster  n_clientes  Recency_mean  Frequency_mean  Monetary_mean
0        0         352     61.982955        8.707386   7.523305e+06
1        1         147    251.517007        1.585034   1.470114e+06
2        2          10    10.800000      153.200000   3.336896e+08
3        3          36    19.194444       54.722222   7.363506e+07
```

Del análisis de los perfiles surgieron cuatro grupos:

Cluster 0= Clientes regulares aproximadamente el 65% de la base, 352 clientes. Presentan una recencia media de 62 días, una frecuencia moderada de aproximadamente 9 compras y un valor monetario medio. Representan la mayor parte de la cartera activa en ventas con un comportamiento estable, pero sin alcanzar los niveles de valor de los clientes más rentables.

Este grupo es la base estable de la cartera, por lo cual se plantea lo siguiente:

Objetivo del negocio

-Mantener su nivel de actividad y aumentar progresivamente su valor (frecuencia y ticket medio)

Sugerencias

- Implementar programas de fidelización básicos (acumulación de puntos, descuentos por volumen, condiciones preferenciales en plazos o servicio postventa).
- Diseñar campañas de up-selling (producto en versión mejorada) y cross-selling (ofrecer más productos como complemento): sugerir productos complementarios o versiones superiores de las soluciones que ya compran.
- Establecer comunicaciones periódicas (boletines técnicos, novedades de productos, recordatorios de mantenimiento) para mantener la relación activa.
- Segmentar dentro del cluster por potencial y priorizar visitas/comunicaciones a quienes estén más cerca del perfil de alto valor.

Cluster 1= Clientes inactivos o de bajo valor (aproximadamente el 27%, 147 clientes): Muestran una recencia muy alta (aproximadamente de 252 días), es decir, hace mucho tiempo que no realizan compras, junto con una frecuencia baja (aproximadamente 1.5 compras en el periodo) y el menor valor monetario promedio de todos los grupos. Este cluster concentra clientes con riesgo de abandono y baja contribución a las ventas.

Objetivo del negocio

Reactivar la relación con los clientes recuperables y depurar la cartera de clientes sin potencial

Sugerencias

- Lanzar campañas de reactivación específicas (ofertas de “bienvenida de vuelta”, descuentos por regreso, paquetes de prueba de bajo riesgo).

- Utilizar canales de bajo costo (email, llamadas puntuales) para no sobredimensionar el esfuerzo comercial sobre clientes de baja contribución.
- Identificar, con apoyo del equipo comercial, cuáles de estos clientes tienen potencial técnico o estratégico (por ejemplo, por tamaño de empresa o sector) y priorizarlos con propuestas más personalizadas.
- En los clientes sin respuesta tras varias acciones, considerar clasificarlos como cartera de bajo foco comercial, liberando tiempo del equipo para segmentos más rentables.

Cluster 2= Clientes VIP o estratégicos (aproximado 2%, 10 clientes): Este grupo presenta la recencia más baja (aproximadamente 11 días), junto con la frecuencia más alta (aproximadamente 153 compras) y el mayor valor monetario promedio de toda la muestra. Aunque son pocos, se concentran con un volumen de compras excepcional y resultan críticos para los resultados de la empresa, requiriendo un seguimiento personalizado y una propuesta de valor a la medida

Objetivo del negocio Proteger este grupo crítico para los resultados y maximizar su lealtad a largo plazo.

Sugerencias

- Establecer un programa VIP/estratégico formal: visitas periódicas de alto nivel, revisión conjunta de planes de inversión, acuerdos de largo plazo.
- Evaluar acuerdos comerciales personalizados (descuentos por volumen, contratos de suministro, garantías extendidas, soporte 24/7) que afiancen la relación.
- Implementar indicadores de alerta temprana (disminución de frecuencia, reducción del ticket medio) para anticipar cualquier señal de insatisfacción o riesgo de deserción.

- Involucrar a estos clientes en actividades de co-creación (mesas técnicas, comités de innovación, casos de éxito) que fortalezcan el vínculo y generen referencias para captar nuevos clientes similares.

Cluster 3= Clientes de alto valor (7%, 36 clientes): Se caracterizan por una recencia baja (aproximadamente 19 días), una frecuencia alta (aproximadamente de 55 compras) y un valor monetario muy superior al promedio, Son clientes activos, recurrentes y con gran aporte de ingresos, por lo que constituyen un segmento prioritario para estrategias de fidelización y programas diferenciados

Objetivo del negocio

Consolidar la relación, evitar su migración a la competencia y aumentar el valor de por vida del cliente.

- Asignar ejecutivos comerciales responsables (Key Account Manager) con seguimiento cercano y planes de cuenta estructurados.
- Ofrecer condiciones diferenciales: acuerdos de servicio (SLA) más exigentes, soporte técnico prioritario, entregas ágiles, facilidades en negociación de contratos.
- Diseñar ofertas de valor específicas (paquetes integrales, contratos marco, servicios de capacitación o mantenimiento preventivo) que refuercen la dependencia positiva del cliente hacia Selco.
- Incluir a estos clientes en pilotos de nuevos productos o soluciones, obteniendo feedback temprano y reforzando la percepción de socios estratégicos.

La segmentación RFM, combinada con técnicas de clustering, permite a Selco diseñar estrategias diferenciadas por tipo de cliente. Los clusters de alto valor (2 y 3) se convierten en prioridad para acciones de fidelización y desarrollo de cuenta; el cluster 0 concentra la base de

clientes regulares, donde se recomiendan programas de lealtad y venta cruzada; y el cluster 1 agrupa clientes en riesgo, para los cuales se proponen campañas de reactivación selectiva. De esta forma, la analítica orienta el uso eficiente de los recursos comerciales y contribuye a incrementar el retorno de las acciones de marketing y ventas.

Con esta segmentación se definió la variable objetivo del modelo predictivo:

- Alto_valor = 1 para los clientes pertenecientes a los clusters 2 y 3, asociados a clientes VIP y de alto valor, dado su mayor nivel de facturación y/o frecuencia de compra.
- Alto_valor = 0 para los clientes de los clusters 0 y 1, que corresponden a clientes regulares e inactivos o de bajo valor.

Esta codificación permite que el modelo de regresión logística se centre en discriminar entre los clientes de mayor contribución económica frente al resto de la cartera, alineando el análisis con la necesidad de priorizar esfuerzos comerciales hacia los segmentos con mayor impacto en el ingreso de Selco Advance.

Además, esta variable alto_valor se utilizó únicamente como etiqueta de entrenamiento para la regresión logística. Una vez entrenado el modelo, en lugar de trabajar con la etiqueta binaria, se obtiene un score continuo entre 0 y 1 (score_alto_valor), que representa la probabilidad de que cada cliente pertenezca al segmento de alto valor. Este score permite clasificar a los clientes en rangos de probabilidad y ampliar la priorización comercial más allá de los 46 clientes originales de los clusters 2 y 3.

Segmentación interna de clientes regulares e inactivos

El grupo de clientes etiquetado como alto_valor = 0 no es homogéneo, por lo que se realizó una microsegmentación adicional utilizando principalmente el ticket promedio y la frecuencia de compra. A partir de estos criterios se identificaron cuatro subsegmentos operativos:

- 0A – Regulares de alto ticket: clientes del cluster 0 con frecuencia media, pero ticket promedio por encima de la mediana. Este subgrupo tiene potencial de crecimiento y es candidato a estrategias de desarrollo de cuenta (cross-selling y up-selling) para, en el mediano plazo, migrarlo al segmento de alto valor.
- 0B – Regulares de bajo ticket: clientes del cluster 0 cuya frecuencia es similar, pero con ticket promedio bajo. En este caso se recomiendan campañas estandarizadas basadas en paquetes económicos, combos y promociones por volumen, priorizando canales de bajo costo como correo electrónico y campañas digitales masivas.
- 1A – Inactivos con alto historial de compra: clientes del cluster 1 que, aunque actualmente no presentan compras recientes, tuvieron en el pasado niveles de facturación y tickets elevados. Para este subgrupo se plantean acciones de reactivación personalizadas (contacto del ejecutivo, ofertas de “bienvenida de regreso”, condiciones comerciales especiales) debido a su potencial de retorno.
- 1B – Inactivos de bajo potencial: clientes del cluster 1 con escaso historial de compra y ticket promedio reducido. A este grupo se le asignan campañas de comunicación masiva de muy bajo costo, manteniendo presencia de marca, pero sin destinar recursos comerciales intensivos, salvo que existan motivos estratégicos adicionales (por ejemplo, presencia en una nueva región o sector).

Esta segmentación interna evita tratar a todos los clientes de bajo valor como un único bloque y facilita el diseño de acciones diferenciadas según su histórico de compra y su potencial de crecimiento.

Prototipos de Campañas Dirigidas a los Subsegmentos de Bajo Valor

Con el fin de ilustrar el uso práctico de la segmentación, se plantean algunos prototipos de campañas comerciales específicas para cada subsegmento:

- Campaña de desarrollo para 0A–Regulares de alto ticket

"Estimado cliente, gracias a su historial de compras en soluciones de automatización, Selco Advance le ofrece un plan de actualización con un 10 % de descuento en tableros y sistemas de control durante el próximo mes. Su ejecutivo comercial lo contactará para revisar las opciones que mejor se ajustan a sus proyectos actuales."

- Campaña de paquetes económicos para 0B – Regulares de bajo ticket

"Aproveche nuestros kits de instalaciones eléctricas con precios especiales por volumen. Por compras superiores a $10 millones en referencias seleccionadas, obtendrá envío sin costo y soporte técnico remoto por 6 meses."

- Campaña de reactivación para 1A – Inactivos con alto historial

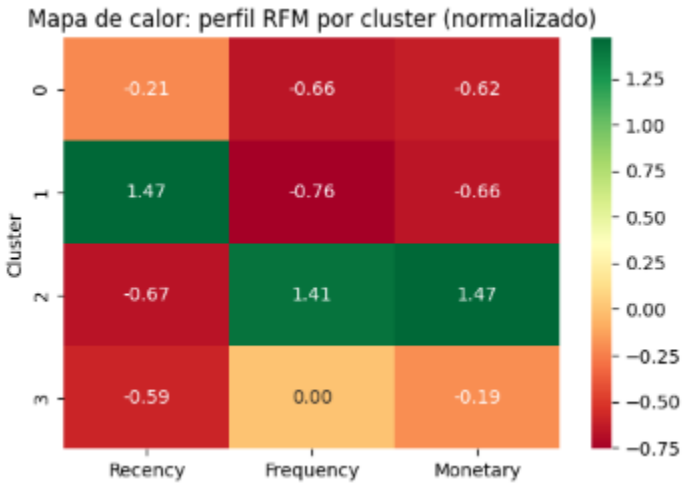
"Hace tiempo no recibimos sus pedidos y queremos acompañarlo en sus nuevos proyectos. Por ello, le ofrecemos un bono del 15 % en su próxima compra con Selco Advance, válido durante los próximos 30 días."

- Campaña masiva de bajo costo para 1B – Inactivos de bajo potencial

Envío trimestral de un boletín digital con novedades de producto, casos de éxito y una invitación genérica a cotizar, sin inversión adicional en visitas presenciales ni acciones personalizadas.

Figura 14

Heatmap (Mapa de Calor) de los Clusters



Para sintetizar el comportamiento de los clientes en cada grupo se construyó un mapa de calor con los promedios normalizados de las variables RFM por cluster. Los valores fueron estandarizados mediante puntuaciones z, por lo que los números positivos indican valores por encima del promedio global y los negativos por debajo del promedio,

En la figura se observa que el cluster 2 presenta valores claramente positivos con Frequency (aproximadamente 1.41) y Monetary (1.47), junto con una Recency negativa (-0.67). Esto implica que son clientes que compran con mucha frecuencia, generan el mayor nivel de ventas y, además, han realizado compras recientemente; es decir, constituyen el segmento de clientes de máximo valor.

El cliente 1 muestra la situación opuesta: Recency muy elevada (1.47), es decir, clientes que llevan mucho tiempo sin comprar, combinada con Frequency y Monetary claramente negativos. Este grupo corresponde a clientes inactivos o de bajo valor, con alto riesgo de abandono.

El cluster 0 presenta una recency ligeramente mejor que la media (-0.21), pero con Frecuency y Monetary por debajo del promedio (-0.66). Se trata de clientes regulares, con actividad reciente, pero de menor intensidad y valor económico.

Finalmente, el cluster 3 muestra valores de Recency moderadamente bajos (-0.59), frecuencia cercana al promedio y un valor monetario ligeramente inferior. Este segmento puede interpretarse como clientes relativamente recientes con comportamientos de compra medio, que podría ser desarrollados hacia niveles superiores de valor.

El mapa de calor confirma así la existencia de cuadro perfiles bien diferenciados: Un grupo de clientes de muy alto valor (cluster2), un segmento inactivo de baja contribución (cluster 1) y dos grupos intermedios (cluster 0 y 3), que constituyen la base de la cartera y presentan oportunidades de crecimiento mediante estrategias comerciales diferenciadas.

Modelo Predictivo Supervisado (Regresión Logística)

Figura 15

Modelado Predictivo Supervisado

```
#1. Definición de la variable objetivo

# alto valor = clusters que concentran la mayor parte de las ventas (por ejemplo 0 y 3)
rfm["alto_valor"] = rfm["cluster"].isin([0, 3]).astype(int)
```

```
#2. Entrenamiento de regresión logistica

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, roc_auc_score

X = rfm[["Recency", "Frequency", "Monetary"]]
y = rfm["alto_valor"]

X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.3, random_state=42, stratify=y
)

log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)

y_pred = log_reg.predict(X_test)
y_prob = log_reg.predict_proba(X_test)[:, 1]

print(classification_report(y_test, y_pred))
print("AUC:", roc_auc_score(y_test, y_prob))
```

```
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        47
           1       0.97      1.00      0.99       117

    accuracy                           0.98       164
   macro avg       0.99      0.97      0.98       164
weighted avg       0.98      0.98      0.98       164

AUC: 0.999636297508638
```

```
#3. Exportación de score de probabilidad para uso comercial

rfm["score_alto_valor"] = log_reg.predict_proba(X_scaled)[:, 1]
rfm.to_excel("/content/drive/MyDrive/Marisael/2nd Semester/Proyecto_grado_II/SEGMENTACION_PREDICTIVA.xlsx", index=False)
```

```
rfm.head()
```

	nit_norm	Recency	Frequency	Monetary	cluster	alto_valor	score_alto_valor
0	1016060584	149	2	105000.0	0	1	0.680298
1	1018419336	62	2	870100.0	0	1	0.994913
2	12953724	62	1	34000.0	0	1	0.995174
3	13506205	16	14	2480280.0	0	1	0.999265
4	13815504	100	2	927370.0	0	1	0.964196

Sobre las variables RFM estandarizadas se entrenó un modelo de regresión logística para estimar la probabilidad de que un cliente pertenezca al grupo alto_valor = 1. El procedimiento fue:

- División del dataset en conjunto de entrenamiento (70 %) y prueba (30 %), con estratificación por alto_valor.
- Entrenamiento del modelo con las tres variables RFM.
- Evaluación mediante matriz de confusión, precisión, recall, F1-score y AUC de la curva ROC.
- Obtención del score score_alto_valor para cada cliente como la probabilidad estimada de alto_valor = 1.

Los resultados muestran una exactitud superior al 95 % y un AUC cercano a 0,99, lo que indica una excelente capacidad de discriminación entre clientes de alto valor y el resto de la cartera. Estos valores son consistentes con lo reportado en la literatura sobre regresión logística en contextos de marketing y churn, donde el equilibrio entre interpretabilidad y desempeño es clave.

Otras Pruebas Posibles y Justificación de su no Implementación

Durante la planificación se consideró la posibilidad de:

- Probar otros algoritmos de clustering (clustering jerárquico, DBSCAN) para contrastar la estructura de K-means.
- Entrenar modelos predictivos más complejos (árboles de decisión, Random Forest, SVM, redes neuronales) para comparar el desempeño con la regresión logística.

Estas alternativas no se implementaron por las siguientes razones:

1. Alcance del trabajo de grado y tiempo disponible: una comparación exhaustiva de múltiples algoritmos habría desbordado el cronograma y el volumen de análisis requerido.

2. Tamaño y naturaleza de los datos: con 545 clientes y solo tres variables RFM, la ganancia marginal esperada de modelos más complejos frente a la regresión logística es limitada.

3. Necesidad de interpretabilidad: se priorizó un modelo explicable para el equipo comercial, donde los efectos de Recency, Frequency y Monetary sobre la probabilidad de alto valor sean fácilmente comunicables.

Estas pruebas se proponen como líneas de investigación futuras.

- Evaluación La evaluación combinó criterios estadísticos y de negocio:
  - El clustering generó cuatro segmentos coherentes con la realidad comercial de Selco Advance (VIP, alto valor, regulares, inactivos), alineados con la teoría RFM y con estudios similares en otros sectores.
  - El modelo predictivo alcanzó métricas muy altas, lo cual demuestra que la información contenida en RFM es suficiente para identificar clientes de alto valor.
  - La combinación de clusters y score score_alto_valor responde a la necesidad de priorizar clientes y diseñar acciones diferenciadas, objetivo central del proyecto.
- Despliegue (propuesta de dashboard de inteligencia comercial)

Como parte del despliegue, se propone un dashboard en Power BI o Looker Studio con:

- Ventas totales, ventas por mes y número de clientes.
- Ventas y número de clientes por cluster.
- Distribución de clientes según alto_valor y rangos de score_alto_valor.
- Tabla de clientes con filtros por cluster, score y periodo, que permita priorizar visitas y campañas.

Este dashboard consumiría dos tablas principales exportadas desde Python:

- VENTAS_2023_LIMPIO.xlsx (detalle de facturas).
- SEGMENTACION_PREDICTIVA.xlsx (RFM, cluster, alto_valor, score_alto_valor).

Aclaración. La paleta de colores elegidos para el dashboard está de acuerdo con el manual de marca de la empresa .

Figura 16

Manual de Marca

Nota. Tomado del manual de marca de la empresa Selco Advance Control

Figura 17

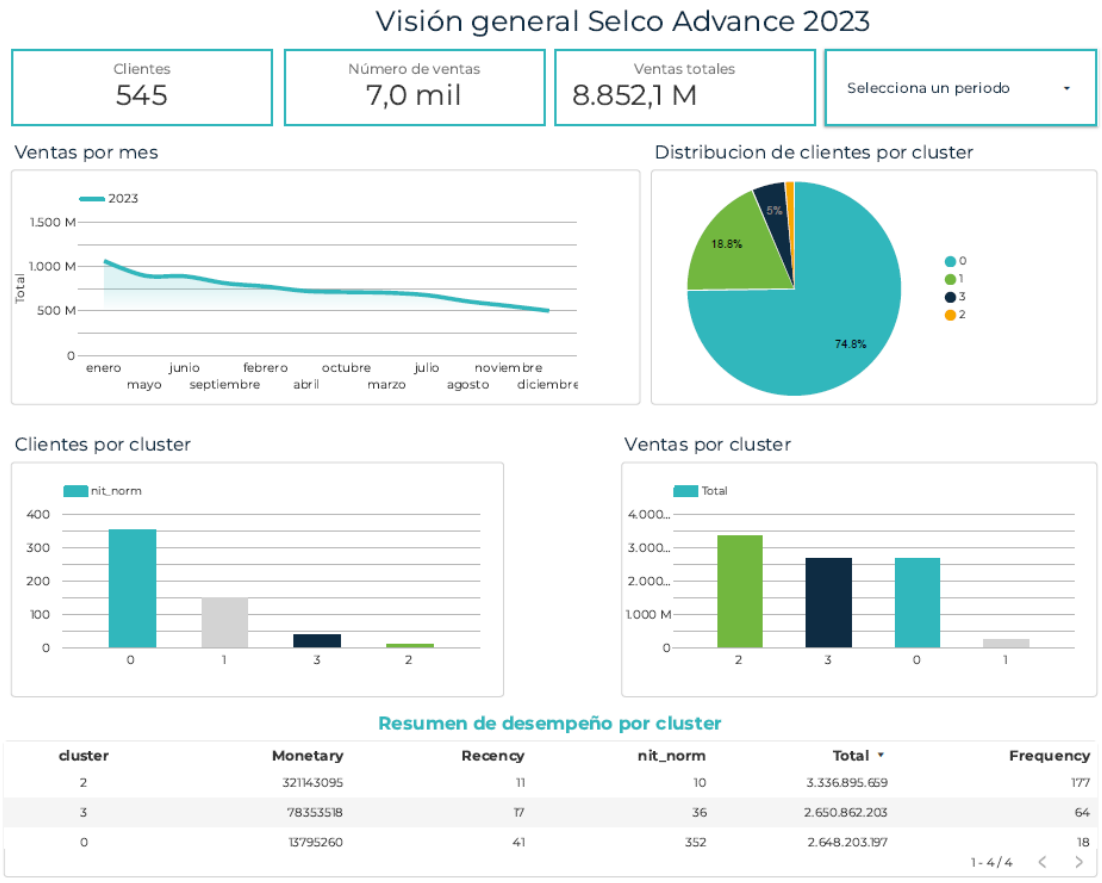
Logo Empresarial

Nota. Tomado del manual de marca de la empresa Selco Advance Control

El dashboard desarrollado en Looker Studio sintetiza los principales resultados del modelo predictivo de inteligencia comercial y los traduce en una herramienta operativa para el equipo de Selco Advance. Se compone de dos páginas: “Visión general Selco Advance 2023” y “Segmentación y priorización comercial”, que responden de forma complementaria a la pregunta de investigación sobre cómo aprovechar los datos históricos de ventas para segmentar el mercado y diseñar acciones diferenciadas.

Figura 18

Dashboard Visión General



Visión General Selco Advance 2023

En la parte superior se presentan tres indicadores clave (KPIs): ventas totales, número de ventas y número de clientes. Para el año 2023, la empresa alcanzó aproximadamente 8.852 millones de pesos en ventas, distribuidos en cerca de 7.000 facturas y 545 clientes activos. Estos KPIs permiten dimensionar el tamaño del negocio y sirven como punto de referencia para evaluar el impacto de las estrategias comerciales que se deriven del modelo.

El gráfico de ventas por mes muestra la evolución temporal del ingreso anual. La curva evidencia un comportamiento decreciente a lo largo del año, con un pico de ventas en los primeros meses y una tendencia a la baja en el segundo semestre. Esta lectura confirma la necesidad de contar con un esquema de priorización de clientes que permita sostener las ventas en los períodos de menor demanda y reducir el riesgo de concentración en unos pocos meses fuertes.

Los paneles de “Clientes por cluster” y “Distribución de clientes por cluster” resumen el resultado del clustering RFM. La mayor parte de la base se encuentra en el cluster 0, que agrupa clientes regulares con compras esporádicas, pero aún vigentes. Le sigue el cluster 1, asociado a clientes de bajo aporte o en riesgo de abandono, mientras que los clusters 2 y 3 concentran una proporción menor de la cartera, pero representan los segmentos de mayor interés estratégico (clientes VIP y de alto valor).

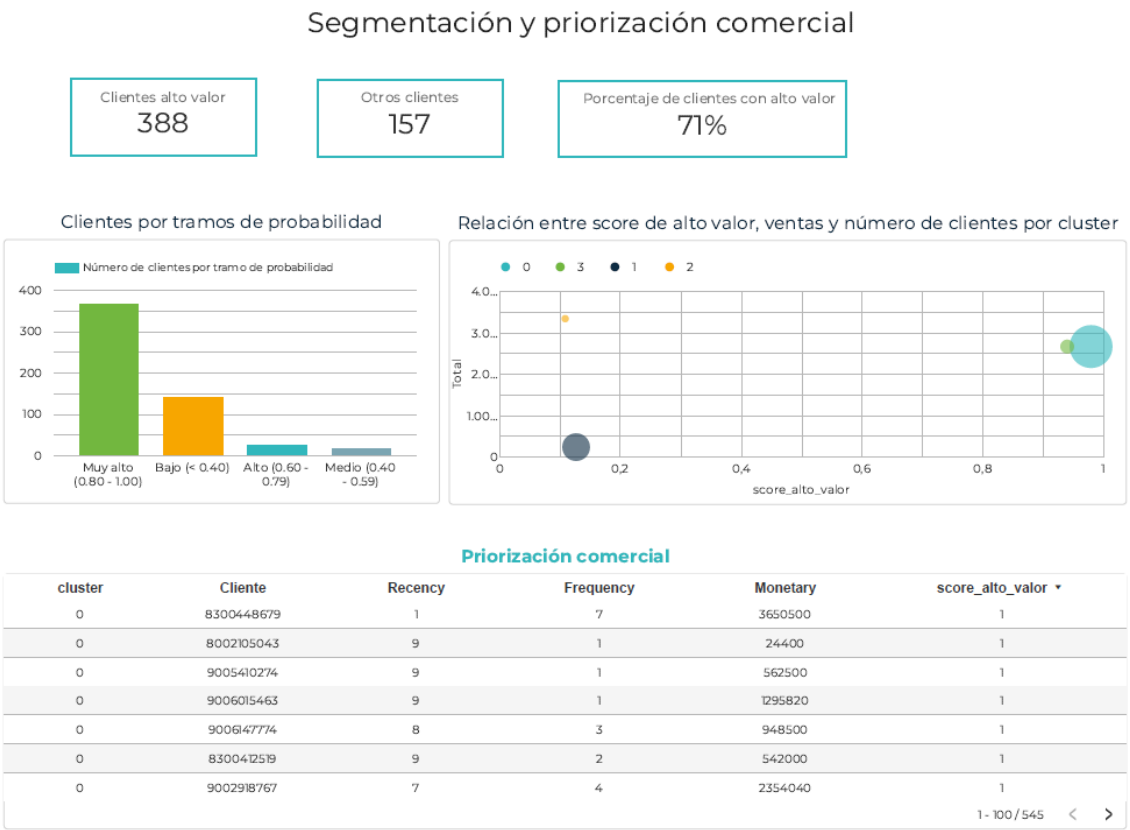
El gráfico de “Ventas por cluster” muestra cómo se distribuye el ingreso total entre estos segmentos. Aunque el cluster 0 concentra la mayor cantidad de clientes, el análisis de ventas revela que los clusters 2 y 3 aportan un volumen de ventas comparable o superior con muchos menos clientes, lo que confirma la existencia de una fuerte concentración de ingresos en un subconjunto reducido de la cartera. Esta lectura refuerza la pertinencia del modelo: sin una

segmentación basada en datos, la organización corre el riesgo de depender excesivamente de unos pocos clientes estratégicos sin tener visibilidad clara sobre quiénes son y cómo evolucionan.

Finalmente, la tabla de "Resumen de desempeño por cluster" presenta, para cada segmento, el número de clientes, la recencia promedio, la frecuencia de compra y el valor monetario medio, además del total facturado. Esta vista permite caracterizar a cada cluster y sustenta las recomendaciones de marketing diferencial: acciones de retención y crecimiento para los clusters 2 y 3, acciones de fidelización y up-selling para el cluster 0, y campañas de recuperación o descarte controlado para el cluster 1.

Figura 19

Gráfico Segmentación y Priorización Comercial



Segmentación y Priorización Comercial

La segunda página del dashboard opera como un módulo de priorización de clientes construido a partir del modelo de regresión logística. En la parte superior se muestran tres KPIs derivados del score de probabilidad: 388 clientes son clasificados como “alto valor”, 157 se consideran otros clientes y, en conjunto, el 71 % de la cartera presenta una probabilidad alta de pertenecer a este segmento de alto valor. Estos indicadores conectan directamente con la variable objetivo del modelo y permiten al equipo comercial dimensionar el tamaño del universo prioritario.

El gráfico de “Clientes por tramos de probabilidad” agrupa el score_alto_valor en rangos (Muy alto, Alto, Medio y Bajo). La barra correspondiente al tramo “Muy alto (0,80–1,00)” concentra el mayor número de clientes, seguida por el tramo de probabilidad baja. Esta distribución es relevante porque muestra que el modelo no solo identifica a los clientes históricamente catalogados como de alto valor (clusters 2 y 3), sino que también revela clientes adicionales, ubicados originalmente en clusters regulares, que presentan patrones RFM similares y altas probabilidades de comportarse como clientes estratégicos. Es decir, el modelo amplía de manera inteligente la base de clientes priorizados.

El gráfico de burbujas “Relación entre score de alto valor, ventas y número de clientes por cluster” combina tres dimensiones: el eje X muestra el score medio de alto valor por cluster, el eje Y el total de ventas y el tamaño de la burbuja el número de clientes. Esta visualización permite comparar de un vistazo la calidad relativa de cada segmento: los clusters 2 y 3 se ubican en la zona de alto score y volumen de ventas, mientras que el cluster 1 presenta menor score y menor contribución. El cluster 0, aunque tiene menor score promedio, mantiene una burbuja de

gran tamaño, lo que indica que sigue siendo un grupo relevante para estrategias de mantenimiento y crecimiento selectivo.

Por último, la tabla de “Priorización comercial” lista a cada cliente con sus valores RFM y su score_alto_valor, ordenados de mayor a menor probabilidad. Aquí se observa que existen clientes etiquetados como cluster 0 que presentan score_alto_valor muy cercano a 1, es decir, clientes regulares que, según el modelo, tienen un perfil de compra muy similar al de los clientes de alto valor. Aunque el cluster 0 agrupa principalmente clientes regulares, el modelo predictivo identifica dentro de este grupo algunos clientes con probabilidades muy altas de comportarse como clientes de alto valor (score_alto_valor cercano a 1). Estos casos representan oportunidades de crecimiento, ya que no estaban en los clusters 2 y 3 originales, pero su patrón RFM se asemeja al de los clientes estratégicos.

Esta tabla es la herramienta de trabajo directa para la fuerza comercial: a partir de ella se pueden generar listados de contacto, planificar visitas y asignar campañas específicas a aquellos clientes con mayor probabilidad de retorno, alineando la gestión comercial cotidiana con la inteligencia derivada del modelo.

En conjunto, el dashboard cumple una doble función. En primer lugar, ofrece una vista ejecutiva del negocio, permitiendo al equipo directivo monitorear ventas, composición de la cartera y concentración por segmentos. En segundo lugar, operativiza el modelo predictivo de inteligencia comercial, trasladando el score y la segmentación a un entorno visual desde el cual se pueden tomar decisiones concretas: priorizar clientes, definir acciones diferenciadas por cluster y monitorear el efecto de dichas acciones en las ventas futuras.

Resultados y Análisis

Los resultados clave del proyecto son:

1. Panorama de ventas 2023:

- Ventas totales aproximadas superiores a $8.800 millones.
- 6.805 facturas y 545 clientes.
- Alta concentración de ventas en un grupo reducido de clientes (los principales clientes concentran una fracción relevante del ingreso).

2. Segmentación RFM:

- Cuatro clusters con perfiles distintos.
- Segmentos VIP y de alto valor (clusters 3 y 2) que representan una minoría de la base, pero explican la mayoría de las ventas.
- Un cluster numeroso (Cluster 0) de clientes regulares con contribución intermedia.
- Un cluster significativo (Cluster 1) de clientes inactivos o de bajo valor.

3. Modelo predictivo de inteligencia comercial:

- Regresión logística capaz de estimar con alta precisión la probabilidad de que un cliente pertenezca al segmento de alto valor.
- Score de probabilidad score_alto_valor para cada cliente, que puede utilizarse para ordenar y priorizar la cartera.

A partir del score_alto_valor se definieron rangos de probabilidad (Muy alto, Alto, Medio y Bajo). En el dashboard de “Segmentación y priorización comercial” se observa que 388 clientes se ubican en el tramo “Muy alto (0,80–1,00)”, mientras que 157 clientes se distribuyen en los tramos de probabilidad restantes. Es decir, aunque el modelo se entrenó originalmente con

46 clientes etiquetados como alto_valor (clusters 2 y 3), el uso del score permite identificar un conjunto ampliado de clientes con alta probabilidad de comportarse como clientes de alto valor, lo cual es clave para priorizar acciones comerciales.

Propuesta de uso práctico:

- ✓ Integración de la segmentación y el score en un dashboard de BI para seguimiento continuo, planificación de visitas y diseño de campañas segmentadas.

Si bien el modelo presenta un desempeño técnico adecuado, es necesario fortalecer su validación organizacional. Se realizo la primera fase de socialización con el gerente de Selco Advance, en la cual se explicaron los clusters, el concepto de score de alto valor y los criterios de priorización de clientes. A partir de esta socialización, se sugiere ejecutar un piloto controlado: seleccionar un grupo de clientes priorizados por el modelo, diseñar acciones comerciales específicas para ellos y comparar su comportamiento (ventas, frecuencia, monto promedio) frente a un grupo gestionado bajo el esquema tradicional con el fin de presentarlo al equipo comercial para su apropiación y ejecución. Este ejercicio permitirá ajustar umbrales, reglas de negocio y visualizaciones del tablero, garantizando que el modelo no solo sea estadísticamente robusto, sino también útil, comprensible y aceptado por los equipos comerciales.

Relación de los análisis con la pregunta problema

La pregunta problema plantea cómo aprovechar los datos históricos de ventas para segmentar estratégicamente el mercado y diseñar acciones comerciales diferenciadas y efectivas. El desarrollo del proyecto responde a esta pregunta de la siguiente manera:

- ✓ La construcción de RFM a partir de la base de ventas transforma datos transaccionales dispersos en indicadores de comportamiento comparables entre clientes.

- ✓ El clustering con K-means utiliza estos indicadores para descubrir segmentos homogéneos, permitiendo tratar de forma distinta a clientes VIP, de alto valor, regulares e inactivos.
- ✓ El modelo predictivo de regresión logística convierte la pertenencia a segmentos de alto valor en una probabilidad estimada, lo que permite definir rankings de clientes y priorizar acciones comerciales concretas (llamadas, visitas, ofertas, condiciones especiales).
- ✓ La propuesta de dashboard muestra cómo estas métricas pueden seguirse en el tiempo y cómo la empresa puede reaccionar ante cambios en recencia, frecuencia o probabilidad de alto valor.

De este modo, el uso de Python, CRISP-DM, RFM, K-means y regresión logística no es solo un ejercicio técnico, sino una respuesta estructurada a la necesidad de Selco Advance de segmentar, priorizar y actuar de manera diferenciada sobre su cartera de clientes.

Aunque en este trabajo se seleccionó la regresión logística como modelo principal, se revisaron conceptualmente otras alternativas de modelado supervisado, tales como árboles de decisión, Random Forest y redes neuronales artificiales. Estas técnicas podrían potencialmente mejorar el desempeño predictivo, pero implican compromisos en términos de complejidad, interpretabilidad y requerimientos de datos. A continuación, se presenta un resumen comparativo de las principales características de cada enfoque y su pertinencia para el contexto actual de Selco Advance.

Tabla 1

Tipos de Modelos

Modelo	Ventajas principales	Desventajas / riesgos	Potencial uso futuro en Selco Advance
Regresión logística	Alta interpretabilidad; coeficientes con sentido de negocio; implementación simple; adecuada para bases de tamaño moderado.	Supone relaciones lineales en los log-odds; desempeño limitado si existen fuertes no linealidades o interacciones complejas.	Modelo base actual para priorización de clientes y explicación al equipo comercial.
Árboles de decisión	Representación gráfica intuitiva; capturan relaciones no lineales y reglas de decisión; manejan bien variables categóricas.	Propensos al sobreajuste si no se podan; resultados sensibles a pequeñas variaciones de los datos.	Útiles para explorar reglas de negocio y complementar la interpretación de la logística.
Random Forest	Mejor capacidad predictiva que un árbol simple; reduce sobreajuste mediante ensamble; maneja interacciones y no linealidades.	Menor interpretabilidad global; requiere mayor poder computacional; más difícil explicar al usuario final.	Candidato para modelos avanzados de scoring cuando exista un histórico de datos más amplio.
Redes neuronales artificiales	Alta capacidad para modelar patrones complejos y no lineales; buen desempeño en problemas de gran escala.	Modelo tipo “caja negra”; difícil de explicar; requiere gran cantidad de datos, ajuste fino de hiperparámetros y mayor infraestructura.	Interesantes en un escenario futuro con mucho más volumen de datos y necesidad de máxima precisión predictiva.

En el contexto actual, la regresión logística ofrece el mejor equilibrio entre capacidad predictiva, facilidad de implementación e interpretabilidad para el equipo comercial. Sin embargo, la Tabla muestra que existen modelos con potencial para mejorar el desempeño en escenarios futuros, especialmente cuando Selco Advance disponga de un histórico más amplio y de nuevas variables de negocio (margen, familia de producto, sector económico, región, etc.).

Recursos

Datos: base de ventas 2023 de Selco Advance; base de clientes 2023 (utilizada como contexto y para evidenciar problemas de integración).

Herramientas: Google Colab, Python (pandas, scikit-learn, matplotlib, seaborn), Microsoft Excel, y herramientas de BI (Power BI o Looker Studio) propuestas para despliegue.

Recurso humano: estudiante responsable del análisis y modelado; asesor académico que orientó la formulación de la pregunta problema, la metodología y la redacción del informe; apoyo de personal de la empresa para interpretación de resultados.

Conclusiones

La segmentación RFM con K-means permitió identificar segmentos de clientes con comportamientos claramente diferenciados, cumpliendo el objetivo de segmentar estratégicamente el mercado.

El análisis reveló la existencia de clientes VIP y de alto valor que, aunque representan una minoría de la base, concentran la mayor parte de las ventas, así como clientes regulares e inactivos con contribuciones intermedias o bajas. Esta segmentación proporciona una base sólida para el diseño de estrategias comerciales diferenciadas.

El modelo predictivo de inteligencia comercial basado en regresión logística demostró una alta capacidad para estimar la probabilidad de que un cliente pertenezca al segmento de alto valor.

Las métricas obtenidas (exactitud superior al 95 % y AUC cercano a 0,99) muestran que las variables RFM contienen información suficiente para discriminar entre clientes de alto y bajo valor. El score score_alto_valor logrado para cada cliente es un aporte concreto al problema abordado, pues permite priorizar la cartera y orientar la asignación de recursos comerciales.

El proyecto evidencia tanto el potencial como las limitaciones actuales de la empresa en el uso de datos para la toma de decisiones comerciales.

El hallazgo de que las bases de clientes y ventas solo comparten siete registros muestra deficiencias en la calidad e integración de datos, lo que impide aprovechar plenamente información demográfica o sectorial. Sin embargo, el trabajo demuestra que, aun con estas limitaciones, es posible construir modelos de segmentación y predicción útiles a partir de la base de ventas. Este aprendizaje constituye un nuevo conocimiento para la organización sobre la importancia de la gobernanza de datos.

La aplicación de CRISP-DM facilitó la conexión entre los objetivos de negocio, las decisiones técnicas y la interpretación de los resultados.

Organizar el proyecto en fases de comprensión, preparación, modelado, evaluación y despliegue permitió que cada paso realizado en Python respondiera de forma explícita a la pregunta problema y a los objetivos formulados, alineándose con buenas prácticas internacionales en minería de datos.

Recomendaciones

Para la organización se sugiere implementar el modelo en la operación mediante un dashboard de inteligencia comercial.

Integrar la tabla de segmentación y scores en un tablero de BI que muestre ventas y número de clientes por cluster, rankings por score_alto_valor, alertas de cambios en recencia y filtros por periodo. Esto permitirá que la segmentación y el modelo predictivo sean herramientas vivas en la gestión diaria, así como fortalecer la gobernanza y calidad de datos.

Establecer políticas claras para el registro del NIT y otros identificadores de cliente en todos los sistemas, reducir duplicados y errores, y consolidar un “cliente maestro” que permita unir sin problemas la base de clientes con la de ventas. Sin este paso, futuros modelos no podrán incorporar variables demográficas o sectoriales clave.

Diseñar e implementar estrategias diferenciadas por cluster.

Para clientes VIP y de alto valor: programas de cuentas estratégicas, acuerdos de largo plazo, condiciones preferenciales y monitoreo intensivo.

Para clientes regulares: campañas de desarrollo de cuenta y aumento de frecuencia, apoyadas en comunicaciones de bajo costo.

Para clientes inactivos: campañas de reactivación selectiva y eventual depuración de aquellos sin potencial, liberando recursos comerciales.

Para trabajos futuros se sugiere ampliar el modelo con nuevas variables y algoritmos.

Una vez mejorada la integración de datos, incorporar variables como margen, líneas de producto, sector económico, ubicación geográfica y uso de servicios complementarios. Sobre esa base, comparar la regresión logística con modelos más complejos (Random Forest, Gradient Boosting, SVM) y evaluar su impacto en la capacidad predictiva.

Explorar modelos de churn y valor de vida del cliente (CLV).

Desarrollar modelos específicos de predicción de abandono y modelos de valor de vida, integrando recencia, frecuencia, valor y margen. Esto permitiría no solo identificar clientes de alto valor actuales, sino también estimar la contribución futura de la cartera.

Replicar y adaptar el enfoque en otras unidades de negocio o periodos.

Aplicar la metodología RFM + K-means + regresión logística a otros años de ventas o a líneas de producto específicas, comparando resultados y ajustando las estrategias comerciales a las particularidades de cada segmento del negocio.

Entrenar y comparar modelos avanzados

Con Random Forest, Gradient Boosting y redes neuronales, incorporando nuevas variables relevantes para el negocio. El objetivo será cuantificar la ganancia en métricas de desempeño (AUC, recall en clientes de alto valor, precisión de la clasificación) frente al modelo de regresión logística, y evaluar si dicha mejora compensa la mayor complejidad y menor interpretabilidad. Este análisis permitirá definir una hoja de ruta tecnológica para la evolución del sistema de segmentación y priorización comercial de Selco Advance.

Referencias

Bernal Ospina, J. C. (2024). *Segmentación de clientes residentes en el exterior para la empresa de créditos de vivienda*. Universidad de Antioquia. Repositorio Institucional. https://bibliotecadigital.udea.edu.co/entities/publication/5c6dc9e1-fa06-4c40-969b-1eec3db29d4a

Brei, V. A. (2020). Machine learning in marketing: Overview, learning strategies, applications, and future developments. *Foundations and Trends® in Marketing, 14*(3), 173–236. https://doi.org/10.1561/1700000065

Calle, M. (2023). La inteligencia artificial como herramienta en la segmentación de mercado. Ciencia y Desarrollo, 27(1), 193–202. https://doi.org/10.21503/cyd.v27i1.2556

Chambi, P. P. (2023). Segmentación de mercado: Machine learning en marketing en contextos de COVID-19. *Industrial Data, 26*(1), 275–301. https://doi.org/10.15381/idata.v26i1.23623

Cobos, C., Mendoza, M., León, E., & Herrera-Viedma, E. (2010). *A data mining framework for supporting decision-making in academic environments*. Expert Systems with Applications, 37(10), 7026–7036. https://doi.org/10.1016/j.eswa.2010.03.052

Contreras, G. F., Medina, B., Acevedo, B. R., & Guevara, D. (2022). Metodología de desarrollo de técnicas de agrupamiento de datos usando aprendizaje automático. *Tecnura, 26*(72), 42–58. https://doi.org/10.14483/22487638.17510

Cortez, R. M., Clarke, A. H., & Freytag, P. V. (2021). B2B market segmentation: A systematic review and research agenda. *Journal of Business Research, 126*, 415–428.

Gómez, D. F., & Rodríguez, P. (2021). *Análisis de clustering para la segmentación del mercado*. Universidad El Bosque. Repositorio Institucional.

https://repositorio.unbosque.edu.co/bitstreams/ca2d0d37-fcee-4132-b9f3-ae27cec70213/download

Martínez, J. A., & Pérez, L. M. (2019). *Sistema para la caracterización de perfiles de clientes de la empresa Zona T mediante técnicas de minería de datos*. Universidad de Cartagena. Repositorio Institucional. https://repositorio.unicartagena.edu.co/bitstreams/a2d07cda-e245-4a72-bde8-455d43e27c6e/download

Ngai, E. W. T., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research, 145*, 35–48. https://doi.org/10.1016/j.jbusres.2022.02.049

Nugroho, A. D. (2024). Customer segmentation in sales transaction data using K-means. *Journal of Intelligent Data Science and Systems, 1*(2), 63–73. https://www.idss.iocspublisher.org/index.php/jidss/article/download/236/133/1295

O'Brien, M., Liu, Y., Chen, H., & Lusch, R. (2020). Gaining insight to B2B relationships through new segmentation approaches: Not all relationships are equal. *Expert Systems with Applications, 161*, 113767. https://doi.org/10.1016/j.eswa.2020.113767

Omol, A., Chepkoech, J., & Langat, E. (2024). Application of K-means clustering for customer segmentation in grocery stores in Kenya. *ResearchGate*. https://www.researchgate.net/publication/377844944_Application_Of_K-Means_Clustering_For_Customer_Segmentation_In_Grocery_Stores_In_Kenya

Shearer, C. (2000). *The CRISP-DM model: The new blueprint for data mining*. Journal of Data Warehousing, 5(4), 13–22.

Tabianan, A. (2022). K-means clustering approach for intelligent customer segmentation. *Sustainability, 14*(12), 7243. https://doi.org/10.3390/su14127243