# Segmentación de Clientes y Predicción de Ofertas Focalizadas


## Resumen

En esta primera entrega del proyecto, presentamos el problema que abordaremos y la propuesta de solución basada en modelos de aprendizaje automático. Realizamos una breve revisión de la literatura sobre el problema de la asignación de ofertas a clientes, enfocándonos en la industria alimentaria. Como insumo técnico, profundizamos en el análisis exploratorio de datos y, finalmente, mostramos la bibliografía consultada hasta el momento.


## Introducción

En Colombia, las empresas dedicadas a la producción y comercialización de alimentos enfrentan grandes retos en muchos aspectos, especialmente en el mercadeo. Con la entrada de las tiendas de descuento (tipo D1), los comerciantes buscan ofrecer productos de alta calidad a un menor costo para sus clientes. Desde el punto de vista estratégico, una práctica antigua y muy importante es la de las ofertas a los clientes, ya que una correcta asignación no solo aumenta las ventas, sino que también garantiza la presencia del producto y la fidelización de los clientes con la marca. Así pues, la asignación óptima de ofertas a los clientes es un desafío crucial en el que participan empresas, tiendas y consumidores. Este proyecto se centra en abordar este problema mediante el uso de modelos de aprendizaje automático. La motivación detrás de esta investigación radica en la necesidad de mejorar la eficiencia y efectividad de las estrategias de marketing en una compañía de alimentos en Colombia. Al implementar técnicas avanzadas de segmentación de clientes y predicción de probabilidades de compra, buscamos desarrollar un sistema que permita asignar ofertas, incrementando la satisfacción del cliente, optimizando procesos internos y garantizando los ingresos de la empresa.

## Revisión de Literatura

La asignación óptima de ofertas a clientes es un tema de gran interés en el ámbito del marketing, especialmente con el auge de las técnicas de aprendizaje automático. Diversos estudios han abordado este problema desde diferentes perspectivas, proporcionando una base sólida para el desarrollo de estrategias efectivas.

Uno de los estudios más relevantes es el de Dynamic Customer Segmentation: Using Machine Learning to Identify and Address Diverse Customer Needs in Real-Time. Este trabajo explora la aplicación de técnicas de aprendizaje automático para lograr una segmentación dinámica de clientes. La segmentación dinámica permite identificar y abordar rápidamente las necesidades cambiantes de los clientes, lo cual es crucial en un entorno de mercado altamente competitivo. Los autores destacan la importancia de utilizar algoritmos de clustering y análisis predictivo para adaptar las ofertas en tiempo real, mejorando así la satisfacción del cliente y la efectividad de las campañas de marketing.

Otro estudio significativo es el de Customer Profiling, Segmentation, and Sales Prediction Using AI in Direct Marketing. Este artículo investiga la segmentación de clientes basada en el análisis de recencia, frecuencia y valor monetario (RFM), utilizando algoritmos de clustering como K-means para identificar segmentos de clientes distintos. La segmentación

basada en RFM permite a las empresas categorizar a sus clientes según su comportamiento de compra, lo que facilita la personalización de las ofertas y la optimización de los recursos de marketing. Los resultados del estudio muestran que la segmentación precisa y la predicción de ventas mediante inteligencia artificial pueden aumentar significativamente la efectividad de las estrategias de marketing directo.

Estos estudios proporcionan una comprensión profunda de cómo las técnicas de aprendizaje automático pueden transformar la segmentación de clientes y la asignación de ofertas, ofreciendo una base sólida para el desarrollo de soluciones innovadoras en la industria alimentaria.

**Descripción y Análisis de los datos**

Para el proyecto, se tomaron los datos de venta de una empresa colombiana que produce alimentos. A continuación, se hace la descripción de cada variable y se comenta sobre su contenido:

1. Mes (Año/Mes):
    - Representa datos desde 202401 (enero de 2024) hasta 202407 (julio de 2024).
2. Cliente (ID de Cliente):
    - El conjunto de datos contiene 1,234,289 registros, indicando un gran número de transacciones.
3. CategoriaMarca (Código de Categoría/Marca):
    - Existen 21 categorías o marcas únicas, siendo "CM26" la más frecuente.
4. SKU:
    - Hay 224 SKUs únicos, con "M1115" como el SKU más común.
5. VtaTon (Volumen de Ventas en Toneladas):
    - El volumen medio de ventas es de 0.0307 toneladas, con un rango significativo desde -2 hasta 86.09 toneladas.
    - La presencia de valores negativos sugiere posibles devoluciones o errores de entrada de datos que necesitan más investigación.
6. VtaValor (Valor de Ventas):
    - El valor de ventas varía ampliamente de -8,957,606 a 262,380,100, con una media de 178,766.30.
    - La existencia de valores negativos indica potenciales reembolsos o ajustes.
7. Departamento (Departamento/Región):
    - Hay 27 departamentos únicos, siendo "BOGOTÁ" el más frecuente.
8. Poblacion (Población/Localidad):
    - 484 localidades únicas, con "BOGOTÁ D.C." apareciendo con más frecuencia.
9. TipologiaCliente (Tipología del Cliente):
    - Existen 8 tipologías de clientes únicas, con "T16" siendo la más frecuente, representando una porción significativa de los registros.
10. DiaAtencion (Día de Atención/Servicio):
    - Hay 6 días de atención únicos, con "MA" siendo el más común.
11. CentroDespacho (Centro de Despacho) y OficinVentas (Oficina de Ventas):

- El centro de despacho y la oficina de ventas tienen pocos códigos únicos, siendo "6038" y "6030" los más comunes, respectivamente.

Ahora, se presentan los detalles más técnicos y estadísticos del dataset:

1. Información del dataset:

```
Información del dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1234289 entries, 0 to 1234288
Data columns (total 12 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   Mes               1234289 non-null  int64
 1   Cliente           1234289 non-null  int64
 2   CategoriaMarca    1234289 non-null  object
 3   SKU               1234289 non-null  object
 4   VtaTon            1234289 non-null  float64
 5   VtaValor          1234289 non-null  float64
 6   Departamento      1234289 non-null  object
 7   Poblacion         1234289 non-null  object
 8   TipologiaCliente  1234289 non-null  object
 9   DiaAtencion       1234289 non-null  object
 10  CentroDespacho    1234289 non-null  int64
 11  OficinVentas      1234289 non-null  int64
dtypes: float64(2), int64(4), object(6)
memory usage: 113.0+ MB
None
```

Se tienen 12 columnas en donde 6 son cualitativas y 6 cuantitativas, las últimas 6 se pueden separar en 4 de tipo discreto y 2 continuas.

2. Estadísticas descriptivas de las variables cuantitativas:

```
Resumen estadístico de las variables numéricas:
                Mes       Cliente        VtaTon      VtaValor  CentroDespacho  \
count  1.234289e+06  1.234289e+06  1.234289e+06  1.234289e+06    1.234289e+06
mean   2.024041e+05  3.028022e+06  3.069266e-02  1.787663e+05    6.035504e+03
std    1.990772e+00  1.538783e+04  3.111323e-01  1.302653e+06    3.005983e+00
min    2.024010e+05  3.001295e+06 -2.000000e+00 -8.957606e+06    6.001000e+03
25%    2.024020e+05  3.013169e+06  1.380000e-03  1.735713e+04    6.033000e+03
50%    2.024040e+05  3.030708e+06  4.500000e-03  4.330296e+04    6.038000e+03
75%    2.024060e+05  3.043733e+06  1.600000e-02  1.173809e+05    6.038000e+03
max    2.024070e+05  3.072785e+06  8.609400e+01  2.623801e+08    6.039000e+03

       OficinVentas
count  1.234289e+06
mean   6.031131e+03
std    1.433605e+00
min    6.030000e+03
25%    6.030000e+03
50%    6.030000e+03
75%    6.032000e+03
max    6.034000e+03
```

A pesar de que tenemos 6 variables numéricas, sólo dos de ellas tienen real sentido dentro del análisis (VtaTon y VtaValor) y este se presenta a continuación:

VtaTon (Ventas en Toneladas):

- Media: 0.0369 toneladas, pero con un mínimo de -2 toneladas, lo que sugiere posibles errores o valores negativos que podrían requerir revisión.

- El 75% de los valores son mayores a 0, lo que implica que la mayoría de las ventas son positivas.

VtaValor (Valor de Ventas):

- Media: 178,766.3, con un rango desde -8,957,660 a 262,380,000, lo cual sugiere que hay ventas con valores negativos, posiblemente devoluciones o errores.

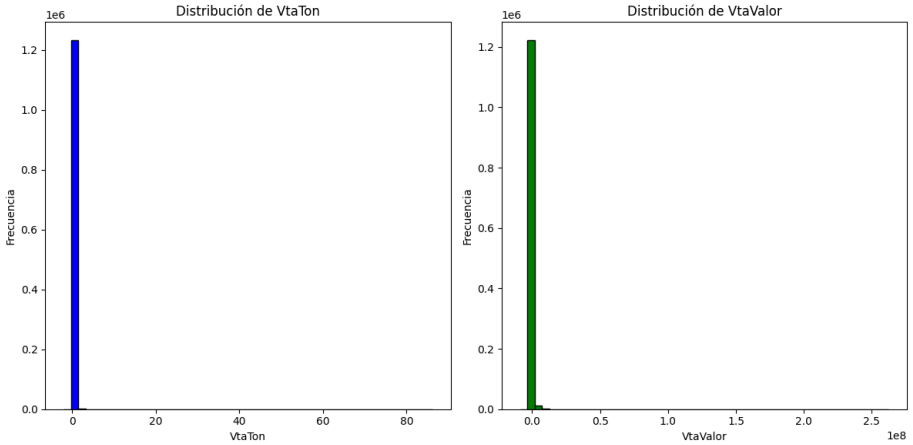
3. Valores faltantes:

Dentro de los datos no se observan datos faltantes como se muestra a continaución:

```
Valores faltantes por columna:
Mes                 0
Cliente             0
CategoriaMarca      0
SKU                 0
VtaTon              0
VtaValor            0
Departamento        0
Poblacion           0
TipologiaCliente    0
DiaAtencion         0
CentroDespacho      0
OficinVentas        0
dtype: int64
```

4. Distribución y relación de las variables numéricas:

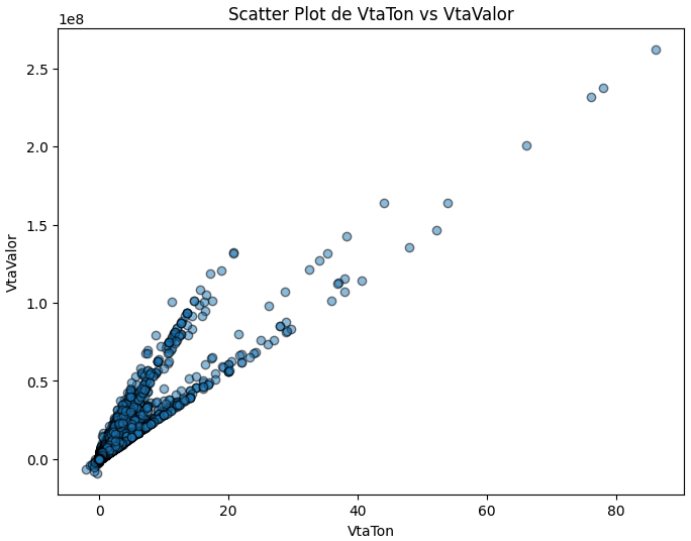
Siguiendo con el análisis de las variables, ahora se presentan la distribución de las variables VtaTon y VtaValor:




Las distribuciones de las variables VtaTon y VtaValor muestran que la mayoría de los valores están concentrados cerca de cero, con algunos valores extremos mucho más altos:

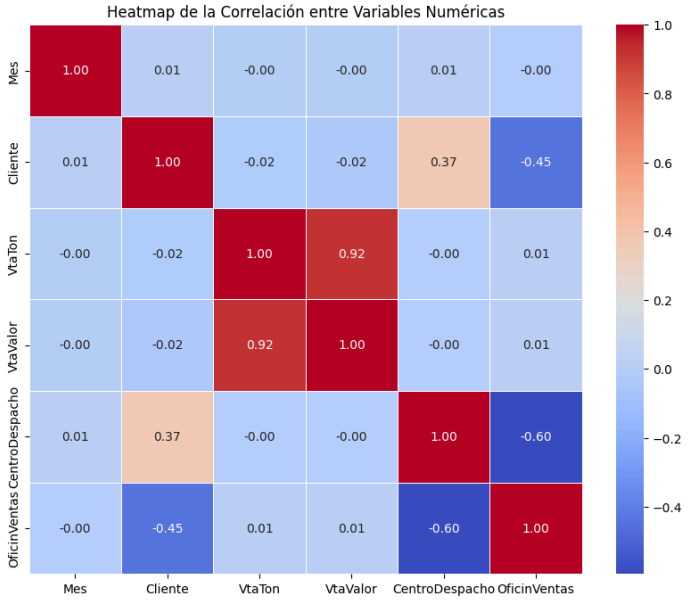
- VtaTon (Ventas en toneladas): La mayoría de las ventas en toneladas son muy pequeñas, y hay pocos casos con valores muy altos.
- VtaValor (Ventas en pesos): Similarmente, la mayoría de los valores de ventas están cerca de cero, con algunos puntos que alcanzan valores significativamente mayores.

5. Gráfico de dispersión:




El gráfico de dispersión muestra la relación entre las variables VtaTon (ventas en toneladas) y VtaValor (valor de ventas). La mayoría de los puntos están agrupados cerca del origen, lo que sugiere que hay muchas observaciones con ventas bajas tanto en toneladas como en valor. Sin embargo, también se observan algunos puntos dispersos que representan ventas mucho mayores

6. Gráfico de correlación:




- VtaTon y VtaValor tienen una alta correlación positiva (0.92), lo cual es esperado, ya que ambas métricas están relacionadas con las ventas.
- Cliente y CentroDespacho tienen una correlación moderada (0.37), sugiriendo cierta relación entre el cliente y el centro de despacho.
- CentroDespacho y OficinVentas presentan una correlación negativa significativa (-0.60), indicando que, a medida que una variable aumenta, la otra tiende a disminuir.
- No hay otras correlaciones fuertes entre las demás variables.

**Metodología**

La metodología que se presenta a continuación, tiene 2 partes: segmentación de clientes y predicción de la probabilidad de compra.

1. Metodología para la Segmentación de Clientes

Paso 1: Preparación de los Datos

- Limpieza de Datos: Se debe garantizar que los datos estén limpios, eliminando duplicados, manejando valores faltantes y corrigiendo posibles errores en los datos (como valores negativos en ventas que no correspondan a devoluciones).
- Transformación de Variables: Crear variables adicionales que puedan ser relevantes para la segmentación, como:
    - Recencia: Tiempo desde la última compra del cliente.
    - Frecuencia: Número de compras realizadas por el cliente en un período determinado.
    - Valor Monetario: Valor total gastado por el cliente en el mismo período.
    - Categoría Preferida: Categoría o marca con mayor frecuencia de compra.
    - Día Preferido de Atención: Día de la semana en el que el cliente suele realizar más compras.

Paso 2: Selección del Algoritmo de Segmentación

- K-means es útil cuando se tiene un gran conjunto de datos y se necesita una segmentación rápida y eficiente. Es fácil de interpretar y se puede aplicar iterativamente para ajustar el número de clusters óptimo.

Paso 3: Evaluación y Selección del Número de Clusters

- Método del Codo: Método para determinar el número óptimo de clusters en K-means. Analiza la varianza intra-cluster y busca el punto donde agregar más clusters no mejora significativamente la varianza explicada.

Paso 4: Interpretación y Perfilado de los Segmentos

- Analizar cada cluster resultante para entender su perfil:
- Asignar nombres y descripciones a cada segmento para facilitar su uso en el modelo de predicción.

2. Metodología para la Predicción de la Probabilidad de Compra

Paso 1: Preparación del Conjunto de Datos de Entrenamiento

- Feature Engineering: Utilizar los segmentos de clientes resultantes como una característica (feature) clave. Además, agregar otras variables relevantes como:
    - Histórico de compras (frecuencia, cantidad, categorías preferidas).
    - Actividad reciente (recencia de la última compra).

    - Respuesta a ofertas anteriores (si se tiene esta información).
- Variable Objetivo: Definir la variable objetivo como una variable binaria que indica si un cliente compró una oferta específica (1 = compra, 0 = no compra).

Paso 2: Selección del Algoritmo de Predicción

Random Forest: Este modelo de ensamble basado en múltiples árboles de decisión es robusto y eficaz para conjuntos de datos con múltiples variables. Maneja bien la complejidad y la no linealidad de las características del cliente.

Paso 3: Evaluación del Modelo

- Uso de métricas para evaluar el desempeño del modelo.
- Realizar validación cruzada para garantizar que el modelo generalice bien a nuevos datos.

**Bibliografía**


1. Dynamic Customer Segmentation: Using Machine Learning to Identify and Address Diverse Customer Needs in Real-Time. Este estudio explora la aplicación de técnicas de aprendizaje automático para lograr una segmentación dinámica de clientes, permitiendo identificar y abordar rápidamente las necesidades cambiantes de los clientes.

2. Customer Profiling, Segmentation, and Sales Prediction Using AI in Direct Marketing. Este artículo investiga la segmentación de clientes basada en el análisis de recencia, frecuencia y valor monetario (RFM), utilizando algoritmos de clustering como K-means para identificar segmentos de clientes distintos.

3. Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov. Este libro ofrece una introducción sólida a machine learning en poco más de 100 páginas, combinando teoría y práctica con ilustraciones, modelos y algoritmos escritos en Python.

4. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media. Este libro es una guía práctica para aprender machine learning utilizando bibliotecas populares como Scikit-Learn, Keras y TensorFlow.

5. Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media. Un libro excelente para principiantes con experiencia en Python, que cubre los conceptos básicos y avanzados de machine learning.