# MODELO DE MACHINE LEARNING PARA LA PREVENCIÓN DE FUGA DE CLIENTES ENFOCADO AL SECTOR DE GASES INDUSTRIALES

**LUISA FERNANDA PAREJA**

**LUISA FERNANDA LOPERA**

**JUAN CAMILO ESCOBAR**

Trabajo de Grado presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor: Ingrid Durley Torres Pardo

Título de Posgrado

Universidad Católica Luis Amigó
Facultad de Ingenierías y
Arquitectura
Especialización en Big Data e Inteligencia de Negocios
Medellín, Colombia

2024

## Dedicatoria

En el camino hacia la culminación de este trabajo de grado en Big Data e inteligencia de negocios, nos detenemos para expresar nuestra más sincera gratitud y reflexión hacia esas personas que hacen de un objetivo, una meta, un sueño. Una realidad palpable para el profesional que se convierte en especialista. Dentro de este contexto agradecemos a nuestros padres, a nuestros docentes y a las empresas que componen nuestro ciclo de vida educativo y productivo.

Este trabajo representa horas de estudio y análisis, es la producción de un año de conocimiento, al igual que un profundo aprendizaje sobre el valor de lo que el esfuerzo representa para el profesional que quiere sobresalir en su campo.

## Agradecimientos

A cada una de las personas que influyó en nuestras carreras, desde nuestros padres, hasta nuestros profesores y asesora de grado, desde nuestros inicios en el conocimiento, hasta el perfeccionamiento que nos brinda nuestra asesora de grado Ingrid Torres Pardo para generar valor a la sociedad y a la empresa Air Products, que nos brindó la oportunidad de generar una intervención completa en su actividad económica para potenciarla.

Por ende a ustedes, Carlos, Sandra, Felipe, Edilma, Nutella y Alma. Les agradecemos porque fueron, son y serán siempre nuestros pilares durante toda la vida.

## Resumen

Este estudio de predicción y Clústerización se realizan predicciones que permiten recomendar productos a clientes corporativos, así como se analizan características de clientes en una empresa, como sector, volumen y valor de compras, fechas de vinculación y compra, contrato de suministro, y frecuencia de compra. El objetivo es determinar qué características están asociadas con diferentes tipos de clientes, identificando fugas de clientes y recomendaciones de productos. Se encontró que ciertos productos tienen demanda geográfica y sectorial específica, lo que puede mejorar la rentabilidad de las plantas con problemas de capacidad instalada ineficiente.

El diseño del estudio involucra análisis cuantitativo y cualitativo de datos de clientes. Se aplicaron métodos de Clústerización como Random Forest, SVC, silhouette, k-means y t-SNE para predecir compras de productos según datos y segmentar clientes según su comportamiento de compra y características demográficas.

Los resultados muestran clústeres de clientes con comportamientos de compra similares y patrones geográficos específicos. Se identificaron segmentos de clientes con alta rentabilidad y potencial de crecimiento. Además, se proponen estrategias para retener clientes, mejorar la oferta de productos y optimizar la capacidad instalada. Al igual que se evidencian contrates de compras de acuerdo a variables primarias que permiten visualizar recomendaciones de compra.

En conclusión, este enfoque analítico de predicción y Clústerización permite una recomendación de productos tangibles y segmentación efectiva de clientes, facilitando la toma de decisiones empresariales basadas en datos.



Palabras clave: Predicción, Clústerización, análisis de datos, Random Forest, silhouette, k-means, t-SNE, Herarquical, rentabilidad, estrategias empresariales.

# Contenido

## Lista de tablas



## Lista de ilustraciones

## Introducción

En el mundo empresarial, la retención de clientes es un factor crucial para el éxito y la sostenibilidad de una compañía. En este contexto, AIRPRODUCTS COLOMBIA, una empresa con una vasta trayectoria en la producción y comercialización de gases para diversas industrias en Colombia, enfrenta un desafío significativo: la pérdida de clientes. Esta problemática no solo afecta su rentabilidad y metas corporativas, sino que también evidencia la necesidad de implementar estrategias proactivas para evitar la fuga de clientes y mantener su posición competitiva en el mercado; por lo tanto, es fundamental el análisis de datos, se debe contemplar que la forma en que estos departamentos adquieren, estructuran y manejan los datos puede representar una barrera o una ventaja competitiva;  el análisis de datos se debe enfocar en identificar similitudes entre clientes, categorizándolos y agrupándolos según características comunes que los distinguen de otros grupos. Estos grupos, que son homogéneos internamente pero heterogéneos entre sí, se generan mediante técnicas de Machine Learning y algoritmos de aprendizaje no supervisado, conocidos como "Clústeres" (Rungruang et al., 2024). Incorporar técnicas de segmentación de clientes mediante Clústerización dentro de la estrategia de fidelización e incentivo de la empresa conlleva diversas ventajas:

- Reconocimiento de patrones de comportamiento, como qué compran, cuánto compran, cómo lo hacen, qué factores influyen en ciertas situaciones, cuándo consumen o utilizan el servicio, entre otras.
- Comprensión del perfil del cliente mediante patrones de consumo homogéneos, como niveles de gasto, CLV (Valor del Tiempo del Cliente), ubicación, hábitos de compra, variables sociodemográficas, edad, entre otras.
- Creación y ejecución de campañas y promociones personalizadas, así como políticas comerciales que reflejen los datos, mejorando su eficiencia.
- Priorización de clientes: concentración de esfuerzos de fidelización en clientes de mayor valor.
- Captación de nuevos clientes (similares): implementación de estrategias de adquisición dirigidas al público objetivo con características similares al Clúster.
- Dirigir el flujo de clientes hacia Clústeres de mayor valor, incluso desarrollando técnicas de venta cruzada.
- Incremento de la retención: la predicción de Clústeres propensos a la deserción permite

implementar acciones para reducir esta tasa.

El presente trabajo tiene como objetivo principal desarrollar un modelo de machine learning, específicamente de Clústerización, que permita anticipar y prevenir la fuga de clientes en AIRPRODUCTS COLOMBIA. Esta iniciativa surge de la falta de herramientas predictivas en la empresa, lo que ha llevado a estrategias de retención y recuperación basadas únicamente en la experiencia comercial, en lugar de en patrones de comportamiento de los clientes.

El método propuesto implica la aplicación de técnicas de Clústerización sobre datos históricos de clientes, con el fin de identificar grupos o "Clústeres" con características similares y patrones de comportamiento que indiquen una mayor probabilidad de fuga. A través de esta segmentación, la empresa podrá tomar decisiones anticipadas y personalizadas para retener a sus clientes, optimizar sus ingresos y costos de adquisición, y mantener su competitividad en el mercado.

Este documento aborda el problema de retención de clientes en la empresa Air Products, se comienza con una introducción al problema, seguido por una argumentación sobre la idoneidad del uso de herramientas de machine learning para abordar este desafío. Se proporciona un análisis de referentes y antecedentes sobre la aplicación de esta técnica en investigaciones previas. Posteriormente, se detallan los objetivos del estudio, seguido por la descripción paso a paso de los métodos aplicados; finalmente, se presentan los resultados obtenidos a partir de estos métodos, los cuales son utilizados para generar conclusiones y recomendaciones específicas para la empresa.

Es importante reconocer que este trabajo presenta ciertas limitaciones; en primer lugar, la eficacia del modelo dependerá en gran medida de la calidad y disponibilidad de los datos históricos de clientes y para nuestro caso, son muy pocos los datos sociodemográficos que se tienen de los clientes, ya que estos son empresas y la información es más limitada que cuando se trata de personas naturales (compras por unidad); se cuenta con la ubicación, tipo de empresa y mercado y el histórico de compras que realizan, pero se carece de otras variables que se referencian en los estudios que son vitales para la retención de clientes como el número de contactos que se le realizan al cliente y las quejas y reclamos; Además, el proceso de Clústerización puede enfrentar desafíos relacionados con la interpretación y validez de los resultados obtenidos. A pesar de estas limitaciones, se espera que este enfoque contribuya significativamente a la capacidad de AIRPRODUCTS COLOMBIA para gestionar y retener a

sus clientes de manera más efectiva y proactiva.

## Planteamiento del Problema

AIRPRODUCTS COLOMBIA es una compañía con 77 años de presencia en el mercado, dedicada a la producción y comercialización de gases para la industria medicinal, industrial, científica y aplicaciones a nivel nacional en Colombia (Cryogas.com)

Actualmente la empresa no cuenta con una herramienta que permita predecir la fuga de clientes, por ende, solo se puede evidenciar la pérdida cuando el cliente notifica la cancelación del contrato o meses después cuando se logra determinar que el cliente no ha tenido movimientos de ventas. Por esta razón, las estrategias retención o recuperación se basan en la experiencia comercial y no en patrones de comportamiento de clientes; además, el proceso de recuperación se realiza de manera reactiva cuando se identifica la pérdida del cliente y no de manera preventiva.

Las consecuencias asociadas a la fuga de clientes identificadas en la compañía impactan significativamente los resultados y metas corporativas. Las áreas más afectadas son las ventas dejando de percibir una facturación de un millón de dólares en el último año y la producción con una disminución en el porcentaje de utilización o carga de las tres plantas con las que cuenta la compañía a nivel Nacional: actualmente la planta 1 (Sibaté) pasó del 86% al 34% de utilización, la planta 2 (Barbosa) pasó del 97% al 86% y la planta 3 (Galapa) pasó del 60% al 40%, generando un incremento de costos que se han trasladado al precio final.

La fuga de clientes no es un problema exclusivo de la empresa AIRPRODUCTS COLOMBIA sino que es una situación que día a día enfrentan las empresas y que puede generar pérdidas económicas representativas. Identificar los elementos que llevan a un cliente a dejar de consumir un bien o servicio es una tarea compleja; sin embargo, mediante su comportamiento es posible estimar una probabilidad de fuga asociada a cada uno de ellos.

Las áreas comerciales y de servicio al cliente tienen el rol fundamental de poder establecer relaciones comerciales para atraer clientes, pero también para retenerlos mediante campañas de marketing y fidelización, debido a que, como se analizó en la revisión de literatura, retener un cliente resulta más rentable para una empresa que atraer uno nuevo puesto que un nuevo cliente implica altos gastos operacionales y gastos asociados a acuerdos comerciales, y solo se transforma en una fuente de beneficios para la empresa cuando comienza a consumir el servicio contratado y otros servicios derivados (Miranda *et al*, 2005).

Se considera fuga de un cliente cuando hay cancelación del servicio prestado y este deja de consumirlo, sin embargo, existen algunos casos donde el usuario deja de consumir el servicio durante un periodo determinado antes de cancelarlo (Pérez, 2014).

En este orden de ideas, este trabajo tiene como finalidad construir un modelo de maching learning (clúster) que permita evitar la fuga de clientes de la compañía AIRPRODUCTS COLOMBIA , de tal manera que la empresa pueda tomar decisiones anticipadas buscando retener sus clientes, mantener o incrementar sus ingresos, optimizar el costo de adquisición de nuevos clientes y finalmente llevar al mercado precios más competitivos sin afectar su rentabilidad ni el cumplimiento de sus objetivos.

## Justificación

El presente trabajo de grado se enfoca en abordar el desafío de la fuga de clientes en la compañía de gases AIRPRODUCTS COLOMBIA. La retención de clientes y la adquisición de nuevos son esenciales para el crecimiento sostenible de la empresa, sin embargo, el beneficio de retener clientes radica en evitar los costos operativos y asociados a acuerdos comerciales que conlleva la adquisición de nuevos clientes. Cuando un cliente ya está establecido y comienza a utilizar los servicios contratados, así como otros servicios adicionales, se convierte en una fuente rentable para la empresa, ya que retener a un cliente resulta más rentable para una empresa que atraer uno nuevo (Miranda *et al*, 2005).

El uso de tecnologías de machine learning, como la Clústerización, brinda a las empresas la capacidad de comprender mejor el comportamiento de compra de los clientes. Al identificar patrones de compra y segmentar a los clientes en grupos homogéneos, (Han, J., Kamber, M., & Pei, J. (2011)

AIR PRODUCTS COLOMBIA podrá diseñar estrategias personalizadas para retener a los clientes existentes y prevenir la fuga de aquellos que muestran signos de posible abandono.

La implementación de técnicas de machine learning proporcionará a AIR PRODUCTS COLOMBIA algunas métricas de alto valor para la toma de decisiones informadas y el diseño oportuno de estrategias innovadoras de retención de clientes. Esto incluye mejorar la experiencia del cliente, estudiar a la competencia, garantizar el abastecimiento de productos y mejorar la experiencia post venta. En última instancia, este enfoque impulsará la lealtad del cliente, aumentará la rentabilidad, haciendo que sus plantas de producción estén a un nivel óptimo, evitando con esto sobre costos en los productos y promoverá el crecimiento continuo de AIR PRODUCTS en el mercado de gases.

## Marco de Referencias

En el mundo empresarial actual, la competencia por las ventas es feroz y constante. Las empresas se enfrentan a un entorno altamente competitivo donde cada decisión cuenta y cada cliente es valioso. En este contexto, el uso efectivo de los datos de compras y las características de los clientes se ha convertido en una estrategia fundamental para mantenerse a la vanguardia y prevenir la fuga de clientes. Por tal motivo, cada día, las empresas se esfuerzan por comprender mejor a sus clientes, anticipar sus necesidades y ofrecerles productos y servicios personalizados que satisfagan sus expectativas. Para lograr esto, es crucial utilizar los datos disponibles de manera inteligente y estratégica, pues el análisis de datos no solo permite a las empresas comprender el comportamiento de compra de sus clientes, sino que también les brinda información valiosa sobre las características y preferencias de cada uno. Con esta información en mano, las empresas pueden crear modelos de machine learning, como la segmentación de clientes, que les permiten agrupar a los clientes en categorías similares en función de sus comportamientos de compra, preferencias y características demográficas. (Anitha & Patil, 2022)

Entre métodos no supervisado más populares de machine learning para Clústerización se encuentran: k-medias, que agrupa los datos en un número predeterminado de clústeres al minimizar la varianza intra-Clúster; el algoritmo de agrupamiento jerárquico, que crea una jerarquía de clústeres utilizando un enfoque aglomerativo o divisivo; el algoritmo DBSCAN (Density-Based Spatial Clústering of Applications with Noise), que agrupa los datos basándose en la densidad local de los puntos; y el algoritmo Mean Shift, que encuentra los centros de los Clústeres moviéndose gradualmente hacia las regiones de alta densidad de puntos. Cada uno de estos algoritmos tiene sus propias ventajas y desventajas, y la elección del algoritmo adecuado depende del tipo de datos y del objetivo del análisis de Clústering. Para nuestro estudio, utilizamos K-medias y el método jerárquico, porque tiene una operación simple y escalabilidad e idoneidad para procesar conjuntos de datos a gran escala, para el caso de K-medias, su principal objetivo es agrupar un conjunto de datos en un número predeterminado de Clústeres, donde cada punto de datos pertenece al clúster con el centroide más cercano; entre sus principales características se encuentran: Versatilidad; el algoritmo K-medias es altamente versátil y puede aplicarse a una amplia variedad de conjuntos de datos y problemas, eficiencia computacional; es computacionalmente eficiente y puede manejar grandes cantidades de datos de manera efectiva, lo que lo hace adecuado para conjuntos de datos de

gran escala, facilidad de implementación; es relativamente sencillo de implementar y entender, lo que lo convierte en una opción popular para muchas aplicaciones de Clústering, interpretabilidad; los Clústeres obtenidos a través del algoritmo K-medias suelen ser fácilmente interpretables, lo que facilita la comprensión y la toma de decisiones basadas en los resultados del Clústering. Por otra parte, el método jerárquico agrupa los datos en una jerarquía de Clústeres, donde los Clústeres más pequeños están contenidos dentro de Clústeres más grandes. Este enfoque puede ser utilizado tanto para análisis exploratorio de datos como para generar insights sobre la estructura subyacente de los datos. Una de las principales ventajas del método jerárquico es su capacidad para revelar la estructura natural de los datos, lo que puede ser útil cuando no se tiene conocimiento previo sobre el número de Clústeres óptimo. Además, el método jerárquico es intuitivo y fácil de interpretar, ya que produce un dendrograma que muestra la relación entre los Clústeres a diferentes niveles de la jerarquía. (Salim et al., 2023)

Es importante resaltar que la evaluación de los Clústeres es un paso crucial en el análisis de Clústering para determinar la calidad y coherencia de los Clústeres obtenidos. Para el caso específico del algoritmo k-medias, una de las técnicas de evaluación más comunes es el método de la silueta (silhouette method), también se utiliza el "método del codo" (elbow method). Este método consiste en trazar un gráfico que muestra la relación entre el número de Clústeres y la variabilidad intra-Clúster (inertia). La variabilidad intra-Clúster mide qué tan compactos y coherentes son los Clústeres, siendo menor cuando los puntos dentro de cada Clúster están más cercanos entre sí. En el gráfico, el punto donde la variabilidad intra-Clúster comienza a disminuir rápidamente es considerado como el codo, y el número de Clústeres en ese punto suele ser seleccionado como el número óptimo de Clústeres.

Para el método jerárquico de Clústering, una técnica de evaluación común es el dendrograma. Un dendrograma es una representación gráfica de la jerarquía de Clústeres, donde los Clústeres se unen gradualmente en función de su similitud. Al observar el dendrograma, los analistas pueden identificar la altura en la que los Clústeres comienzan a fusionarse rápidamente, lo que sugiere la presencia de un número óptimo de Clústeres. Esta técnica proporciona una visión visual de la estructura de los Clústeres y puede ayudar a determinar cuántos Clústeres son adecuados para los datos en cuestión.

(Salim et al., 2023)

## Marco conceptual

**Machine Learning:** Es una de las ramas de la inteligencia artificial. Es la responsable de crear algoritmos capaces de aprender de los datos y los patrones encontrados, de tal manera que no tienen que ser programados en cada ejecución. De esta manera, el algoritmo aprende cómo responder ante todos los posibles escenarios (Sandoval, 2018).

**Algoritmo:** Este concepto hace referencia a un conjunto de pasos para realizar una tarea determinada. Las computadoras aplican los algoritmos para ejecutar procesos de manera más eficiente y ágil (GCF Global, s.f.)

**Aprendizaje automático:** Este concepto es considerado como el arte de enseñar a las computadoras a generar predicciones basadas en datos. Inicialmente se pueden presentar una gran cantidad de errores en la predicción, pero en la medida que se va ejecutando el proceso, se actualiza el algoritmo y va mejorando la precisión de los resultados (Normal, 2019)

**Aprendizaje profundo:** Corresponde a un subcampo del aprendizaje automático. El aprendizaje profundo trabaja con redes neuronales para reconocer relaciones, patrones y predicciones más complejas con grandes cantidades de datos, lo cual implica mayor capacidad de procesamiento. (Rouhiainen, 2018)

**Supervisado (Supervised Learning):** Definición: Un tipo de aprendizaje automático en el que el modelo se entrena en un conjunto de datos etiquetado, es decir, donde las entradas están asociadas con las salidas deseadas. (Alpaydin, 2014)

**No Supervisado (Unsupervised Learning):**: Un tipo de aprendizaje automático en el que el modelo se entrena en un conjunto de datos no etiquetado y debe encontrar patrones o estructuras ocultas en los datos. (Hastie, Tibshirani, & Friedman, 2009).

**Validación de Clústeres:** Es el proceso de evaluar la calidad y coherencia de los clústeres obtenidos por un algoritmo de Clústering, utilizando medidas como la compacidad intra-Clúster y la separabilidad inter-Clúster. (Halkidi et al., 2002)

## Antecedentes

Se han observado diversos estudios prácticos aplicados cuyos objetivos han sido tener un acercamiento con un modelo basado en datos cuantitativos y cualitativos que estime la probabilidad de abandono de los clientes. En las investigaciones analizadas se han identificado las principales técnicas de machine learning que tienen mayor efectividad para el tipo de problema abordado.

Uno de estos trabajos es el realizado por Gattermann y Thonemann (2022) aplicado a una tienda mayorista de estrategia B2B, localizada en Europa. Esta investigación presenta un caso semejante al problema abordado en este trabajo en el sentido que las dos compañías comercializan productos tangibles y alrededor del 56% de los clientes son extracontractuales, es decir, no hay un contrato formal entre la empresa y el cliente y en cualquier momento este puede dejar de adquirir los productos. De acuerdo con los autores, esta situación hace más difícil prever la fuga y para estos casos sugieren aplicar los indicares FRM: Frecuencia: número de compras realizadas por un cliente en un período específico; Reciente: representa cuánto tiempo ha pasado desde la última compra hasta el momento de la recopilación de datos; y Monetario: equivale a la cantidad de dinero que gastó un cliente en tres meses anteriores y 2. Las compras generales disminuyeron un 30% o más en comparación con los tres meses anteriores.

Otras variables que se consideraron importantes son: el tiempo desde el último contacto que se realizó con el cliente para hacerle seguimiento (gestión de relación con los clientes), tiempo transcurrido en la entrega de productos y cantidad de contactos realizados al cliente. Además, de acuerdo con el estudio, se debe identificar el tiempo promedio que hay entre cada compra por cliente para definir cuándo se considera que hay un abandono parcial. Finalmente, una vez identificadas las características o variables a tener presente, se desarrollaron tres modelos de aprendizaje automático que predice la pérdida de clientes: Bosque aleatorio, Regresión logística y SVM lineal.

Por otra parte, en un estudio realizado por Baghla y Gupta (2022) en una empresa de E-Commerce en Brasil de estrategia corporativa B2C, se identificó que las variables utilizadas para la predicción de fuga de clientes son las mismas que las abordadas en el caso anterior: Indicador de frecuencia, indicador de actualidad haciendo referencia a la última fecha de

compra y el tiempo promedio de entrega del producto. Adicionalmente, estos autores consideran fundamental analizar los comentarios de los clientes. Las variables mencionadas fueron evaluadas mediante la técnica NCA que realiza la selección de características para maximizar la precisión de la predicción de los algoritmos de clasificación y regresión y, para este caso, esas características seleccionadas contaron con un peso de 0,99.

Para la selección de características se han utilizado técnicas de análisis de componentes principales y análisis de componentes vecinales. Respecto al algoritmo utilizado, se implementaron varias técnicas de aprendizaje automático: Redes neuronales, Máquinas de vectores de soporte, Naïve Bayes, Bosque aleatorio y la técnica de aprendizaje profundo de Adam. Para efectos de este estudio se definió como abandono a los clientes que no realizan transacciones en la plataforma de comercio electrónico durante 90 días consecutivos.

Finalmente, los autores Fredy Troncoso y Javiera Tapia (2020) realizaron un estudio sobre la predicción de fuga de clientes en la empresa de distribución de gas natural Gar Sur. S.A en las ciudades de Concepción y Los Ángeles en Chile. En este trabajo se utilizó la metodología de Knowledge Discovery in Databases KDD que consiste que consiste en los siguientes 5 pasos: 1. Selección de datos, 2. Preprocesamiento de la base, 3. Transformación y selección de las variables, 4. Minería de datos y 5. Interpretación y evaluación. Adicionalmente, se utilizó una muestra de 36.484 clientes y seleccionaron registros demográficos asociados a contratos, reclamos, servicio al cliente y consumos, comprendidos en los años 2018 y 2020.

Para obtener el patrón que caracteriza a los clientes que se fugan y luego obtener una predicción respecto a este comportamiento, se aplicaron dos técnicas de machine learning: Redes neuronales artificiales y Support vector machine. Además, para la evaluación se consideró la Matriz de Confusión, la curva ROC y el AUC de cada algoritmo de machine learning entrenado, determinando que el mejor algoritmo era Support Vector Machine (mySVM) ajustado con kernel radial y parámetro $\gamma = 0.5$ parámetro c=0 y parámetro $\varepsilon = 0.001$.

Respecto a los costos de clasificación, en la fuga de clientes de Gas Sur S.A. el costo de error tipo I está asociado a la pérdida del consumo por parte de un cliente y el costo de error tipo II hace referencia a aplicar un descuento a un cliente que no se fugará. Estos costos corresponden respectivamente a un 70% y 30% del consumo de un cliente. Mediante la incorporación de estos costos de error de clasificación del modelo, fue posible determinar el valor de umbral que minimiza el costo de error de clasificación en la predicción. Este valor de

umbral es 0.707 e implica que cualquier cliente que tenga un valor de predicción igual o superior será clasificado como un cliente que se fugará.

La utilización del algoritmo Support Vector Machine seleccionado permitió a Gas Sur S.A. identificar aquellos clientes con mayor probabilidad de fuga, priorizar las acciones de retención en aquellos con un mayor consumo y minimizar el costo de error de clasificación, alcanzando a maximizar el beneficio de la acción de retención de los clientes.

un período específico, pues se considera que los clientes no necesariamente dejan de comprar radicalmente los productos, sino que gradualmente van disminuyendo la frecuencia de sus pedidos en dos sentidos: 1. En al menos una categoría las compras disminuyeron un 50% o más en comparación con los Por el ultimo y como propuesta para el desarrollo de la Construcción de un modelo predictivo para la prevención de fuga de clientes enfocado al sector de gases industriales, se tendrán en cuenta factores que no se evidenciaron en las investigaciones realizadas; será enfocado en productos tangibles para el sector de la industrial de gases en Colombia, se aplica una combinación para la estrategia B2B Y B2C, se tendrán en cuenta la aplicación de más de dos técnicas de machine learning y sugerirán estrategias de retención.

| **INVESTIGACIÓN** | **ENFOCADO A B2B & B2C** | **VARIEDAD DE TECNICAS DE MACHINE LEARNING>2** | **VARIABLES ADAPTADAS A TANGIBLES** | **ESTRATEGIAS PARA RETENCIÓN B2B & B2C** |
|---|---|---|---|---|
| Gattermann, T & Thonemann, U. (2022). *Proactive customer retention management in a non-contractual B2B setting based on churn prediction with* | ✗ | ✓ | ✗ | ✗ |

| | | | | |
|---|---|---|---|---|
| *random forests.* Industrial Marketing Management | | | | |
| Baghla, S. & Gupta, G. (2022). *Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce*. Microprocessors and Microsystems | ✗ | ✓ | ✗ | ✗ |
| Troncoso, F. & Tapia, J. (2020). *Predicción de fuga de clientes en una empresa de distribución de gas natural mediante el uso de minería de datos.* Universidad Ciencia y Tecnología | ✗ | ✗ | ✓ | ✗ |
| Construcción de un modelo predictivo para la prevención de fuga de clientes enfocado al sector de gases industriales | ✓ | ✓ | ✓ | ✓ |

*Tabla 1. Comparativa de investigación versus referencias.*

# Objetivos

## Objetivo General

Construir un modelo basado en técnicas de machine learning, que mitigue la fuga de clientes para la empresa AIRPRODUCTS COLOMBIA.

## Objetivos Específicos

- Implementar un proceso de análisis exploratorio de los datos incluyendo extracción, limpieza, transformación, selección y análisis de datos, de los diferentes clientes que posee la empresa AIRPRODUCTS COLOMBIA.
- Utilizar las técnicas de machine learning para la construcción de modelos predictivos a partir de los datos de los clientes.
- Evaluar el desempeño de los modelos construidos.

## Viabilidad

En la construcción del modelo se utilizarían recursos asequibles como son un computador con 8GB de RAM, con SSD de 256 GB de espacio de almacenamiento, sistema operativo Windows, software de código abierto como Python por medio de entornos de desarrollo como Visual Studio Code y Jupyter aplicados dentro del navegador de Anaconda para realizar la correcta implementación y documentación de la investigación, y librerías como Pandas, Matplotlib y Seaborn.

Así mismo, se requiere conexión ethernet de 400 mbps simétricos, considerados suficientes recursos para el manejo de una base de datos de más de 1.000 registros de las ventas mensuales de la empresa AIRPRODUCTS COLOMBIA. Esta base se desglosa cliente a cliente desde el 2019 hasta la fecha (2023) y contiene variables tales como: Tipo de cliente-mercado, fecha de vinculación con la empresa, actividad económica, productos que compra, fechas de las compras, valor de las compras, tipo de contrato, tiempo del contrato, localización del cliente, promedio de ventas y fecha de la última la compra.

El alcance de este trabajo corresponde en realizar una óptima clasificación de los clientes, que facilite la personalización de productos y permita comprender mejor las necesidades y preferencias de los diferentes grupos de clientes, al comprender mejor la demanda de los diferentes segmentos de clientes, la empresa podrá optimizar su cadena de suministro para satisfacer esas demandas de manera más eficiente y rentable. Este propósito implica la creación y desarrollo de un modelo de Clústerización capaz de identificar grupos de clientes con características similares, que permita mejorar la eficacia de sus estrategias de marketing, optimizar sus operaciones y generar un mayor valor para sus clientes y accionistas. Este enfoque permite tomar medidas proactivas para implementar la efectiva retención de cada cliente de acuerdo con las métricas establecidas por la compañía. Es importante aclarar que el proyecto no hace énfasis en la implementación de estrategias de retención o recuperación de clientes, sino que se limitará a la identificación de segmentos de clientes que le permitan a la compañía la generación de alertas tempranas de posible fuga de clientes para intervenir proactivamente con estrategias de retención.

En relación con las implicaciones de la investigación se consideran las repercusiones que puede tener la utilización de datos con información sensible referente a los clientes de

AIRPRODUCTS COLOMBIA Para ello, previamente se generan accesos con consentimientos informados y permisos necesarios debidamente firmados y diligenciados por cada uno de los integrantes del equipo; así, se garantiza la recopilación y utilización de datos personales para el modelo dando anonimato a cada cliente, cumpliendo rigurosamente con todas las regulaciones relacionadas con la privacidad y protección de datos, incluyendo las políticas internas de la empresa y el Habeas Data

Finalmente, respecto a las consecuencias positivas que pretende tener esta producción investigativa, una exitosa implementación de un modelo de Clústerización de clientes puede tener un impacto significativo en la operación de la compañía pero también en el resto del gremio empresarial que la pueda adoptar. Dentro de las principales ganancias se resalta la identificación de características de los clientes, estrategias de retención de clientes integrales y sólidas; pero también se puede lograr una asignación de recursos y personal más eficiente optimizando la carga operativa al enfocarse en esfuerzos que ayuden a retener los clientes con alta probabilidad de abandono. Sumado a lo anterior, la mayor generación de ingresos combinada con reducción de costos de adquisición de clientes genera un crecimiento del EBITDA y maximización de los márgenes financieros de la compañía, así como también puede ayudar a incrementar la participación de la empresa en el mercado colombiano. En síntesis, las ganancias de esta propuesta de trabajo generan beneficios tanto para la empresa como para los clientes y el mercado.

## Metodología

| Objetivo específico | Actividad | Entregable | Fase CRISP-DM |
|---|---|---|---|
| Implementar un proceso de análisis exploratorio de los datos incluyendo extracción, limpieza, transformación, selección y análisis de datos, de los diferentes clientes que posee la empresa AIRPRODUCTS COLOMBIA. | Descripción de la organización, objetivos y procesos del negocio.<br><br>Reconocer la base de datos de clientes, analizar la naturaleza de cada una de sus variables.<br><br>Realizar proceso de limpieza de datos.<br><br>Cargue el dataset en Colab (lenguaje Python) para el proceso mencionado.<br><br>Transformación de variables | Base de datos final con el proceso EDA.<br><br>notebook con la limpieza realizada<br><br>Gráficas explicativas del proceso de EDA. | ➢ FASE I- Entendimiento del negocio del Negocio.<br><br>➢ FASE II- Comprensión de los datos - Objetivo específico 1<br><br>➢ FASE III- Preparación de los datos - Objetivo específico 1 |
| Utilizar las técnicas de machine learning para la construcción de | Generar el modelo predictivo, se evidencian resultados y se concluye que el tipo | Ejecución del modelo predictivo.<br><br>Presentación de | ➢ Fase IV- Modelado |

| | | | |
|---|---|---|---|
| modelos predictivos a partir de los datos de los clientes. | de datos aportado por la empresa no es útil para un modelo predictivo de fuga, para lo cual se procede a generar un modelo acorde a los datos obtenidos.<br><br>Para el cumplimiento del objetivo se recurre a Generar un modelo de Clústerización: Se Corre el modelo, se realiza reducción de dimensionalidad. Aplicar un modelo de Clústering que facilite la segmentación de los clientes a partir de sus diferentes características, realizado por medio de Colab con base en lenguaje python para aplicar al dataset ya estructurado anteriormente.<br><br>Evidenciar la aplicación de diferentes métodos para la identificación de clústeres y | gráficos y resultados.<br><br>Identificación de Overfiting y datos poco relevantes para un modelo predictivo.<br><br>Identificación de un modelo acorde al cumplimiento esperado de prevención de fuga de clientes (Clústerización).<br><br>Presentación de resultados. | |

| | | | |
|---|---|---|---|
| | generación de su respectiva caracterización. | | |
| Fase V- Evaluación<br><br>Evaluar el desempeño de los modelos construidos y Visualizar los resultados del modelo y generar recomendaciones con base en las características de cada clúster.<br><br>Evaluar el cumplimiento de cada uno de los objetivos. | Evaluación inicial de las falencias en la construcción del modelo predictivo, desde los datos aportados por la empresa, hasta la visualización de los resultados.<br><br>Identificar de los resultados obtenidos por la Clústerización.<br><br>Generar estrategias para la empresa Air Products. Por medio de recomendaciones de producto y Clústerización | Identificar del mejor método para la Clústerización de los clientes de la empresa AIRPRODUCTS.<br><br>Caracterizar los clústeres de acuerdo a los resultados<br><br>Sugerencias de campañas para cada clúster, identificación de patrones y acciones a realizar para cada segmento encontrado. | Fase V- Evaluación<br>Fase VI- Despliegue |

*Tabla 2. Metodología utilizada*

Las metodologías más utilizadas en analítica de datos son KDD, CRISP-DM, SEMMA y Catalyst, para esta intervención se utilizará CRISP DM (Cross-Industry Standard Process for

Data Mining), la cual se define como el marco guía de estructura y análisis de datos para una completa comprensión del negocio y de sus datos, una preparación de estos últimos, óptima para su modelado perteneciente al objetivo deseado, contrastado por su evaluación en la cual se reflejan el rendimiento, precisión, sensibilidad, así como también se pueden realizar caracterizaciones de dichos resultados, que permitan a la empresa llevar a la realidad y materializar cada uno de esos objetivos en el despliegue en la organización.

De acuerdo a esto se inicia la metodología CRISP-DM por su primera etapa.

**Desarrollo del Objetivo Específico 1**

Implementar un proceso de análisis exploratorio de los datos incluyendo extracción, limpieza, transformación, selección y análisis de datos, de los diferentes clientes que posee la empresa AIRPRODUCTS COLOMBIA.

El trabajo desarrollado se aplica con la intención de contribuir al área comercial de la empresa AirProducts, en la acertada predicción de los clientes con intención de fuga, generando un plan de acción para el abandono de estos y adicional, la fidelización de quienes aún no poseen dicha intención, encontrando de esta manera que así mismo puede mejorarse la eficiencia de las plantas de producción, ya que al evaluar la información proporcionada por la empresa, es identificada una merma productiva en varias plantas, por lo cual el evitar costos excesivos de consecución de clientes nuevos es indispensable, por lo cual, retener a quienes ya confían en la organización siendo esta una causal directamente proporcional de al incremento en ventas.

En este objetivo se desarrollarán 3 fases de la metodología CRISP - DM

**FASE I: Entendimiento del negocio.**

**Actividad 1: Descripción de la organización, objetivos y procesos del negocio.**

**Contextualización de la empresa**

AIRPRODUCTS COLOMBIA es una compañía con 77 años de presencia en el mercado, dedicada a la producción y comercialización de gases para la industria medicinal, industrial, científica y aplicaciones a nivel nacional en Colombia (Cryogas.com)

**Misión**

Air Products será la empresa de gas industrial más segura, diversa y rentable del mundo y

brindará un excelente servicio a nuestros clientes.

**Visión**

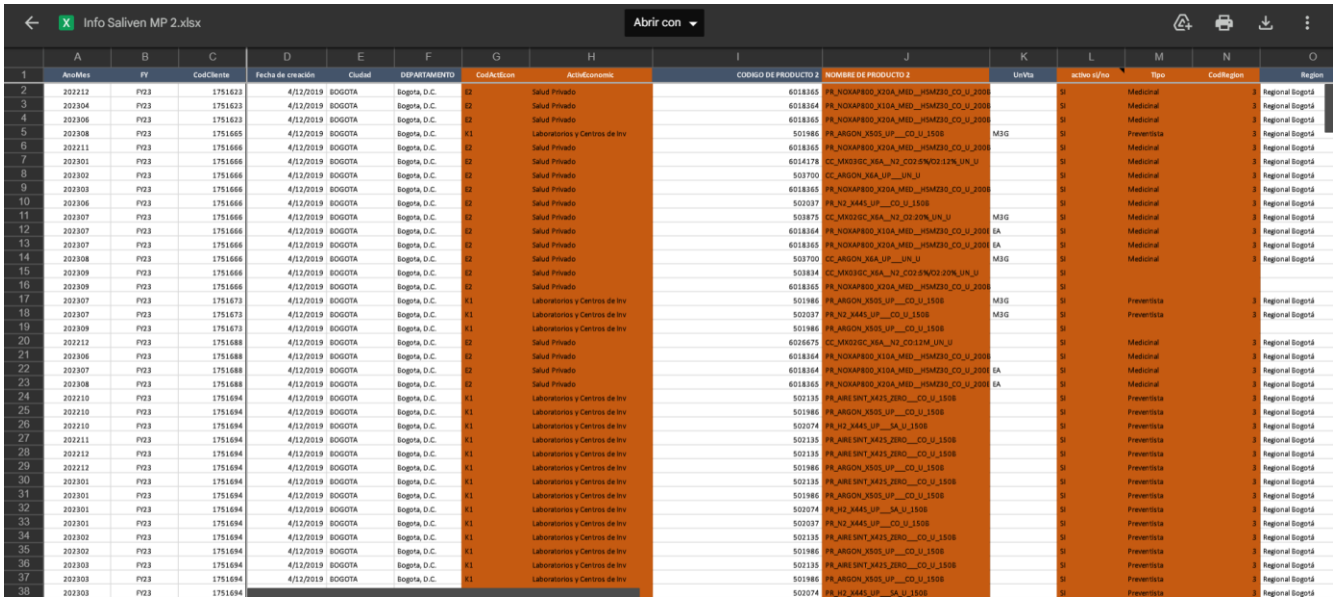
Ser los mejores, siendo los más seguros y rentables en Sudamérica, logrando la preferencia continua de los clientes y el compromiso de nuestros colaboradores

**Área Objetivo**

Área comercial, específicamente, marketing, como complemento a la gestión comercial de la compañía.

**Actividad 2. Reconocimiento de la base de datos de clientes, análisis de la naturaleza de cada una de sus variables.**

En principio, la base de datos cuenta con 9.101 registros y 36 variables.




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AnoMes | FY | CodCliente | Fecha de creación | Ciudad | DEPARTAMENTO | CodActEcon | ActivEconomic | CODIGO DE PRODUCTO 2 | NOMBRE DE PRODUCTO 2 | UnVta | activo si/no | Tipo | CodRegion | Region |
| 2 | 202212 | FY23 | 1751623 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | | SI | Medicinal | 3 | Regional Bogotá |
| 3 | 202304 | FY23 | 1751623 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018364 | PR_NOXAP800_X10A_MED__H5MZ30_CO_U_200B | | SI | Medicinal | 3 | Regional Bogotá |
| 4 | 202306 | FY23 | 1751623 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | | SI | Medicinal | 3 | Regional Bogotá |
| 5 | 202308 | FY23 | 1751665 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | M3G | SI | Preventista | 3 | Regional Bogotá |
| 6 | 202211 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | | SI | Medicinal | 3 | Regional Bogotá |
| 7 | 202301 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6014178 | GC_MX03GC_X6A__N2_CO2:5%/O2:12%_UN_U | | SI | Medicinal | 3 | Regional Bogotá |
| 8 | 202302 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 503700 | GC_ARGON_X6A_UP___UN_U | | SI | Medicinal | 3 | Regional Bogotá |
| 9 | 202303 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | | SI | Medicinal | 3 | Regional Bogotá |
| 10 | 202306 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 502037 | PR_N2_X445_UP___CO_U_150B | | SI | Medicinal | 3 | Regional Bogotá |
| 11 | 202307 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 503875 | GC_MX02GC_X6A__N2_O2:20%_UN_U | M3G | SI | Medicinal | 3 | Regional Bogotá |
| 12 | 202307 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018364 | PR_NOXAP800_X10A_MED__H5MZ30_CO_U_200B | EA | SI | Medicinal | 3 | Regional Bogotá |
| 13 | 202307 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | EA | SI | Medicinal | 3 | Regional Bogotá |
| 14 | 202308 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 503700 | GC_ARGON_X6A_UP___UN_U | M3G | SI | Medicinal | 3 | Regional Bogotá |
| 15 | 202309 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 503834 | GC_MX03GC_X6A__N2_CO2:5%/O2:20%_UN_U | | SI | | | |
| 16 | 202309 | FY23 | 1751666 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | | SI | | | |
| 17 | 202307 | FY23 | 1751673 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | M3G | SI | Preventista | 3 | Regional Bogotá |
| 18 | 202307 | FY23 | 1751673 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502037 | PR_N2_X445_UP___CO_U_150B | M3G | SI | Preventista | 3 | Regional Bogotá |
| 19 | 202309 | FY23 | 1751673 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | | SI | | | |
| 20 | 202212 | FY23 | 1751688 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6026675 | GC_MX02GC_X6A__N2_CO:12M_UN_U | | SI | Medicinal | 3 | Regional Bogotá |
| 21 | 202306 | FY23 | 1751688 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018364 | PR_NOXAP800_X10A_MED__H5MZ30_CO_U_200B | | SI | Medicinal | 3 | Regional Bogotá |
| 22 | 202307 | FY23 | 1751688 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018364 | PR_NOXAP800_X10A_MED__H5MZ30_CO_U_200B | EA | SI | Medicinal | 3 | Regional Bogotá |
| 23 | 202308 | FY23 | 1751688 | 4/12/2019 | BOGOTA | Bogota, D.C. | 32 | Salud Privado | 6018365 | PR_NOXAP800_X20A_MED__H5MZ30_CO_U_200B | EA | SI | Medicinal | 3 | Regional Bogotá |
| 24 | 202210 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502135 | PR_AIRE SINT_X42S_ZERO___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 25 | 202210 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 26 | 202210 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502074 | PR_H2_X445_UP___5A_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 27 | 202211 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502135 | PR_AIRE SINT_X42S_ZERO___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 28 | 202212 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502135 | PR_AIRE SINT_X42S_ZERO___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 29 | 202212 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 30 | 202301 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502135 | PR_AIRE SINT_X42S_ZERO___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 31 | 202301 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 32 | 202301 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502074 | PR_H2_X445_UP___5A_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 33 | 202301 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502037 | PR_N2_X445_UP___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 34 | 202302 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502135 | PR_AIRE SINT_X42S_ZERO___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 35 | 202302 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 36 | 202303 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 502135 | PR_AIRE SINT_X42S_ZERO___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 37 | 202303 | FY23 | 1751694 | 4/12/2019 | BOGOTA | Bogota, D.C. | X1 | Laboratorios y Centros de Inv | 501986 | PR_ARGON_X50S_UP___CO_U_150B | | SI | Preventista | 3 | Regional Bogotá |
| 38 | 202303 | FY23 | 1751694 | | | | | | 502074 | PR_H2_X445_UP___5A_U_150B | | SI | Preventista | 3 | Regional Bogotá |

Ilustración 1. Base de datos primaria

**FASE II-Identificación de los datos - Primer objetivo específico**

**Actividad 3. Realizar proceso de limpieza de datos.**

Para efectos del análisis, inicialmente se decide selecciona 13 atributos considerados importantes que incluyen información como:

1. Año de la compra
2. Mes de la compra
3. Año fiscal
4. Código del cliente
5. Fecha de vinculación del cliente
6. Actividad económica del cliente
7. Tipo de cliente
8. Región
9. Mercado
10. Producto
11. Costos netos
12. Costos por unidad
13. Tipo de gases
14. Contrato de suministro

Adicionalmente, se decide trabajar solo con la información de los clientes activos con una suma neto(ventas) mayor a cero y con el año fiscal correspondiente a 2023. Sin embargo no se realiza ninguna modificación de las mencionadas, hasta tanto no esté el dataset cargado.

**FASE III-Preparación de los datos - Objetivo específico 1**

**Actividad 4. Cargue del dataset en Colab para el proceso mencionado.**

Se genera el cargue por medio de Colab, a través del archivo xlsx para su posterior proceso de visualización, transformación y modelamiento.

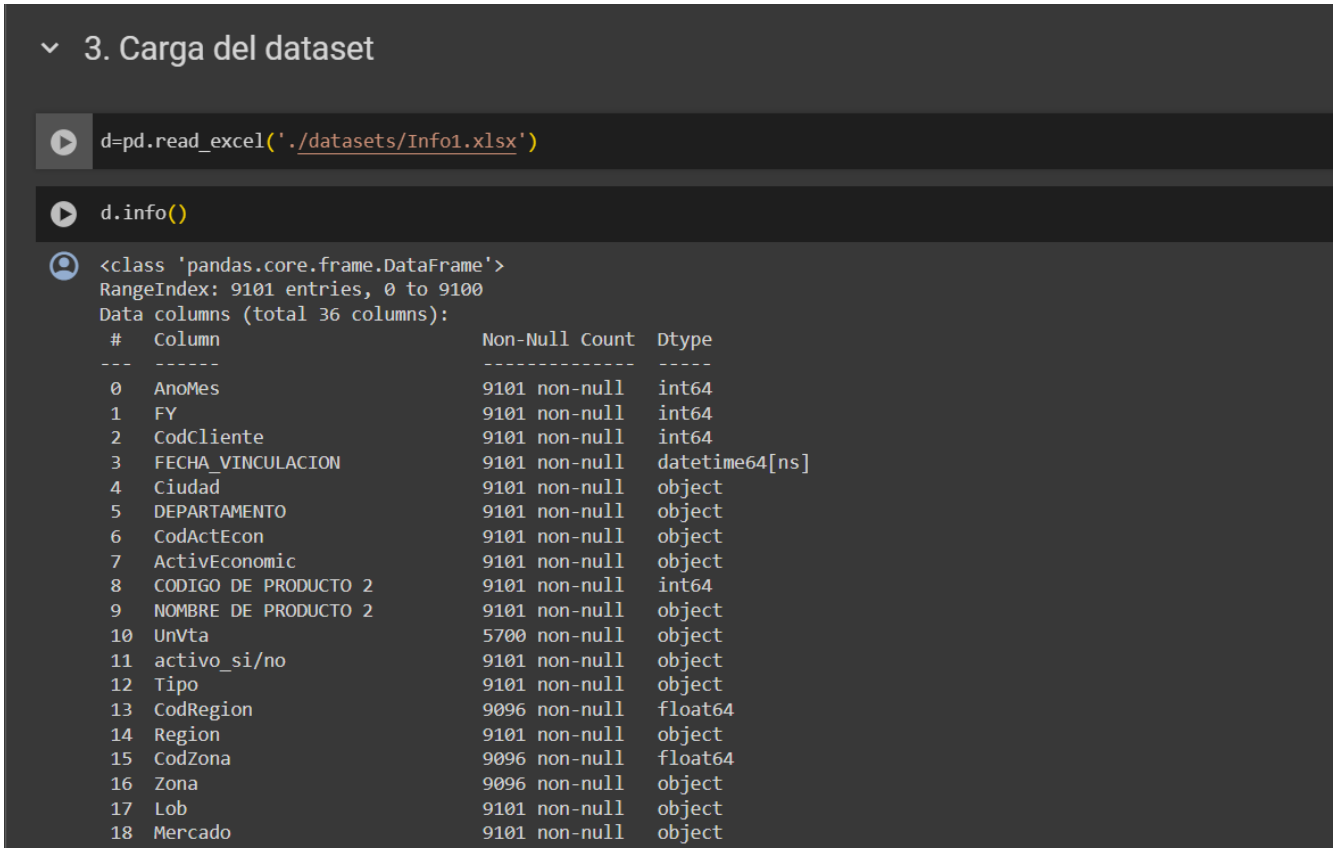



*Ilustración 2.Cargue del dataset a Herramienta Colab*

**Actividad 5. Transformación de variables**

Para lograr hacer el ejercicio de Clústering, es fundamental que se tenga registros únicos por cada cliente, es por ello que se decide agrupar por CodCliente, los atributos considerados. Para ello, por código de cliente se toma la fecha mínima y máxima de compra para tener

registro de cuándo se realizó la primera y última compra del cliente, con relación a los gastos netos y por cantidad se toma la suma total registrada por código cliente, y para el tipo de gas se crean columnas de acuerdo a cada una de las categorías de esta variable y se registra la cantidad adquirida por cada cliente. Para las demás variables estas son fijas para cada registro.

Se realiza un filtro de actividad, para solo realizar foco en los clientes que pueden fugarse, adicional se trabaja con los datos más actuales y completos del año fiscal 2022, ya que se encuentran completos, cuando en 2023 tenemos parciales, por lo cual se decide trabajar con el rango de un año completo. Por último se realizará todo el modelo aplicado a los clientes que posean compras, por lo cual se genera un filtro de valores positivos para evitar outliers innecesarios.

```
d = d[d['activo_si/no'] == 'SI']
d = d[d['FY'] == 2022]
d = d[d['SumaDeNetoProd'] >= 0]
```

Ilustración 3. Filtros aplicados

Posterior al filtro se agrupan las fechas en tiempo de finalización del contrato, para identificar si es a corto, mediano o largo plazo su relacionamiento directo con Air products.

```python
def grupo_fecha(fecha):
    if pd.isnull(fecha) or fecha == 'NO':
        return 'Prioritario'
    elif fecha.year in [2024, 2025]:
        return 'Corto Plazo'
    elif fecha.year in [2026, 2027]:
        return 'Mediano Plazo'
    else:
        return 'Largo Plazo'

d['GRUPO_FECHA'] = d['FINALIZACION_DEL_CONTRATO'].apply(grupo_fecha)

d[['FINALIZACION_DEL_CONTRATO', 'GRUPO_FECHA']]
```

| | FINALIZACION_DEL_CONTRATO | GRUPO_FECHA |
|---|---|---|
| 2616 | NaT | Prioritario |
| 2617 | 2024-12-31 | Corto Plazo |
| 2618 | NaT | Prioritario |
| 2619 | NaT | Prioritario |
| 2620 | NaT | Prioritario |

Ilustración 4. Clasificación y transformación de clientes

Se identifican y eliminan nulos

```python
def grupo_fecha_vinculacion(fecha):
    if pd.isnull(fecha):
        return 'Fecha_erronea'
    elif fecha.year in range(2018, 2021):
        return 'Clientes Antiguos'
    elif fecha.year in range(2021, 2023):
        return 'Clientes Recientes'
    elif fecha.year >= 2023:
        return 'Clientes Nuevos'

d['GRUPO_FECHA_VINCULACION'] = d['FECHA_VINCULACION'].apply(grupo_fecha_vinculacior
d[['FECHA_VINCULACION', 'GRUPO_FECHA_VINCULACION']]
```

| | FECHA_VINCULACION | GRUPO_FECHA_VINCULACION |
|---|---|---|
| 2616 | 2019-04-12 | Clientes Antiguos |
| 2617 | 2019-04-12 | Clientes Antiguos |
| 2618 | 2019-04-12 | Clientes Antiguos |

Ilustración 5. Eliminación de nulos

Concertando de esta manera 10 variables trabajadas con los modelos presentados a continuación:

```
[ ]  d.drop(['DEPARTAMENTO','Ciudad','CodActEcon','ActivEconomic', 'ANegocios','CODIGO

[ ]  d.info()

     <class 'pandas.core.frame.DataFrame'>
     Index: 2762 entries, 2616 to 5470
     Data columns (total 10 columns):
      #   Column                        Non-Null Count  Dtype
     ---  ------                        --------------  -----
      0   AnoMes                        2762 non-null   int64
      1   FY                            2762 non-null   int64
      2   Region                        2762 non-null   int64
      3   Mercado                       2762 non-null   int64
      4   SumaDeNetoProd                2762 non-null   int64
      5   SumaDeCantProd                2762 non-null   float64
      6   TIPO_GASES                    2762 non-null   int64
      7   Contrato de Suministro SI/NO  2762 non-null   int64
      8   GRUPO_FECHA                   2762 non-null   int64
      9   GRUPO_FECHA_VINCULACION       2762 non-null   int64
     dtypes: float64(1), int64(9)
     memory usage: 237.4 KB
```

Ilustración 6.Identificación de variables a trabajar

Adicional a esto en la transformación se realizan los siguientes puntos:

- agrupar la fecha de inicio vinculación OK
- agrupar la fecha de terminación de contrato OK
- Hacer un “if else”, para la categoría si/no antes del drop para que si conserve el registro y después de que todo quede ahí, eliminar la variable que ya no queda aportando nada.
- Eliminar los FY de otros años diferentes a 2022

Visualización de variables dummies y escalado de variables

```
[ ] dDum =pd.get_dummies(dDum, drop_first=1)

dDum.info()

<class 'pandas.core.frame.DataFrame'>
Index: 2762 entries, 2616 to 5470
Data columns (total 35 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   AnoMes                                2762 non-null   int64
 1   FY                                    2762 non-null   int64
 2   SumaDeNetoProd                        2762 non-null   int64
 3   SumaDeCantProd                        2762 non-null   float64
 4   Region_Regional Antioquia             2762 non-null   bool
 5   Region_Regional Bogotá                2762 non-null   bool
 6   Region_Regional Costa                 2762 non-null   bool
 7   Region_Regional Eje Cafetero          2762 non-null   bool
 8   Region_Regional Occidente             2762 non-null   bool
 9   Region_Regional Oriente               2762 non-null   bool
 10  Region_Regional Principal             2762 non-null   bool
 11  Mercado_Procesos Metalmecánicos       2762 non-null   bool
 12  Mercado_Salud                         2762 non-null   bool
 13  TIPO_GASES_AIRE                       2762 non-null   bool
 14  TIPO_GASES_ARGON                      2762 non-null   bool
 15  TIPO_GASES_BUTANO                     2762 non-null   bool
 16  TIPO_GASES_DIOXIDO DE CARBONO         2762 non-null   bool
 17  TIPO_GASES_ETILENO                    2762 non-null   bool
 18  TIPO_GASES_HEXAFLUORURO               2762 non-null   bool
 19  TIPO_GASES_HIDROGENO                  2762 non-null   bool
 20  TIPO_GASES_METANO                     2762 non-null   bool
```

Ilustración 7. Transformación de variables a dummies

```
7. Escalar variables numéricas

[ ] X_scaled = scale(dDum)
```

Ilustración 8. Escalado de variables numéricas

**Desarrollo del Objetivo Específico** 2.

Utilizar las técnicas de machine learning para la construcción de modelos predictivos a partir de los datos de los clientes.

**Fase IV- Modelado**

**Actividad 1. Generar el modelo predictivo y visualizar los resultados, generar conclusiones de la viabilidad del modelo predictivo**

Inicialmente se importan todas las librerías y modelos, dentro de los cuales se utilizan pandas, seaborn y matplotlib, se importan adicional los modelos de árbol de decisión, regresión logística, máquinas de vectores soporte, random forest y xgboost, también se importan métricas como la precisión, la sensibilidad, la exactitud y la división de los datos por medio de train_test_split.

Se crean 4 regionales específicas que permitan identificar ventas por medio de tablas.

```
Bogota = datos[datos['Region'] == 'Regional Bogota']
Periferia =datos[datos['Region'] == 'Periferia Bogota']
Regional_Centro = pd.concat([Bogota, Periferia], axis=0)
Regional_Centro
```

| | AnoMes | ActivEconomic | activo si/no | Tipo | Region | Mercado | SumaDeNetoProd | SumaDeCantProd | GASES PUROS Y ESPECIALES |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 202010 | Laboratorios y Centros de Inv | 1 | Preventista | Regional Bogota | Cientifico | 354735.0 | 6.00000 | ACETILENO |
| 1 | 202010 | Laboratorios y Centros de Inv | 1 | Preventista | Regional Bogota | Cientifico | 825200.0 | 14.00000 | ACETILENO |
| 2 | 202010 | Laboratorios y Centros de Inv | 1 | Preventista | Regional Bogota | Cientifico | 1249098.0 | 21.00000 | ACETILENO |
| 3 | 202010 | Laboratorios y Centros de Inv | 0 | Preventista | Regional Bogota | Cientifico | 2440200.0 | 49.00000 | ACETILENO |

Ilustración 9. Creación de regional centro.

```
Occidente = datos[datos['Region'] == 'Regional Occidente']
Eje =datos[datos['Region'] == 'Regional Eje Cafetero']
#Centro
Regional_Occidente = pd.concat([Occidente, Eje], axis=0)
Regional_Occidente
```

| | AnoMes | ActivEconomic | activo si/no | Tipo | Region | Mercado | SumaDeNetoProd | SumaDeCantProd | GASES PUROS Y ESPECIALES |
|---|---|---|---|---|---|---|---|---|---|
| 24 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Occidente | Cientifico | 413858.0 | 7.000000 | ACETILENO |
| 25 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Occidente | Cientifico | 371947.0 | 7.000000 | ACETILENO |
| 26 | 202010 | Laboratorios y Centros de Inv | 1 | Ejec. Negocios | Regional Occidente | Cientifico | 324105.0 | 7.000000 | ACETILENO |
| 27 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Occidente | Cientifico | 827716.0 | 14.000000 | ACETILENO |

Ilustración 10. Creación de la regional Occidente

```
Antioquia = datos[datos['Region'] == 'Regional Antioquia']
Oriente =datos[datos['Region'] == 'Regional Oriente']
#Centro
Regional_Antioquia = pd.concat([Antioquia, Oriente], axis=0)
Regional_Antioquia
```

| | AnoMes | ActivEconomic | activo si/no | Tipo | Region | Mercado | SumaDeNetoProd | SumaDeCantProd | GASES PUROS Y ESPECIALES |
|---|---|---|---|---|---|---|---|---|---|
| 12 | 202010 | Ind.Metalmecanica (B) | 1 | Ejec. Negocios | Regional Antioquia | Procesos Metalmecánicos | 701356.0 | 14.000000 | ACETILENO |
| 13 | 202010 | Laboratorios y Centros de Inv | 1 | Ejec. Negocios | Regional Antioquia | Cientifico | 709470.0 | 12.000000 | ACETILENO |
| 14 | 202010 | Laboratorios y Centros de Inv | 1 | Ejec. Negocios | Regional Antioquia | Cientifico | 354735.0 | 6.000000 | ACETILENO |
| 15 | 202010 | Laboratorios y Centros de Inv | 1 | Ejec. Negocios | Regional Antioquia | Cientifico | 827716.0 | 14.000000 | ACETILENO |
| 16 | 202010 | Laboratorios y Centros de Inv | 1 | Ejec. Negocios | Regional Antioquia | Cientifico | 295610.0 | 7.000000 | ACETILENO |

Ilustración 11. Creación de la regional Antioquia.

```
Regional_Costa = datos[datos['Region'] == 'Regional Costa']
Regional_Costa
```

| | AnoMes | ActivEconomic | activo si/no | Tipo | Region | Mercado | SumaDeNetoProd | SumaDeCantProd | GASES PUROS Y ESPECIALES |
|---|---|---|---|---|---|---|---|---|---|
| 28 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Costa | Cientifico | 199279.0 | 5.000000 | ACETILENO |
| 29 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Costa | Cientifico | 289702.0 | 12.000000 | ACETILENO |
| 30 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Costa | Cientifico | 365594.0 | 7.000000 | ACETILENO |
| 66 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Costa | Cientifico | 454545.0 | 19.307076 | AIRE |
| 67 | 202010 | Laboratorios y Centros de Inv | 1 | Especialista GE | Regional Costa | Cientifico | 463932.0 | 12.871384 | AIRE |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Ilustración 12. Creación de la regional Costa

Se escalan los datos y se instancian los modelos para inicializar una mejor comprensión de estos al analizarlos, ya que cifras de dinero se hace mucho más agradable e intuitiva una escalabilidad de los resultados.

```
scaler = StandardScaler()


X_train_normalized = scaler.fit_transform(X_train)


X_test_normalized = scaler.transform(X_test)
```

+ Código + Texto

```
# Inicializar los clasificadores
logistic_regression = LogisticRegression()
svm_classifier = SVC()
decision_tree = DecisionTreeClassifier()
random_forest = RandomForestClassifier()

# Entrenar los modelos
logistic_regression.fit(X_train_normalized, y_train)
svm_classifier.fit(X_train_normalized, y_train)
decision_tree.fit(X_train_normalized, y_train)
random_forest.fit(X_train_normalized, y_train)

# Hacer predicciones
y_pred_lr = logistic_regression.predict(X_test_normalized)
y_pred_svm = svm_classifier.predict(X_test_normalized)
y_pred_dt = decision_tree.predict(X_test_normalized)
y_pred_rf = random_forest.predict(X_test_normalized)

# Evaluar los modelos
def evaluate_model(y_true, y_pred):
    accuracy = accuracy_score(y_true, y_pred)
```

Ilustración 13.Escalado de variables y entrenamiento de modelos.

Se realizan predicciones de los modelos en los cuales se evidencia un overfiting en los resultados hallados, ya que los datos generados para dichas predicciones tienden a ser obvios

```python
# Hacer predicciones
y_pred_lr = logistic_regression.predict(X_test_normali
y_pred_svm = svm_classifier.predict(X_test_normalized)
y_pred_dt = decision_tree.predict(X_test_normalized)
y_pred_rf = random_forest.predict(X_test_normalized)

# Evaluar los modelos
def evaluate_model(y_true, y_pred):
    accuracy = accuracy_score(y_true, y_pred)
    precision = precision_score(y_true, y_pred)
    recall = recall_score(y_true, y_pred)
    f1 = f1_score(y_true, y_pred)
    roc_auc = roc_auc_score(y_true, y_pred)
    return accuracy, precision, recall, f1, roc_auc

print("Logistic Regression:")
print(evaluate_model(y_test, y_pred_lr))
print("SVM:")
print(evaluate_model(y_test, y_pred_svm))
print("Decision Tree:")
print(evaluate_model(y_test, y_pred_dt))
print("Random Forest:")
print(evaluate_model(y_test, y_pred_rf))
```

Ilustración 14. Evaluación de modelos.

```
Logistic Regression:
(1.0, 1.0, 1.0, 1.0, 1.0)
SVM:
(0.9983507421660253, 0.9983202687569989, 1.0, 0.9991594284113197, 0.9583333333333333)
Decision Tree:
(1.0, 1.0, 1.0, 1.0, 1.0)
Random Forest:
(1.0, 1.0, 1.0, 1.0, 1.0)
```

Ilustración 15. Predicciones

Se realiza un ajuste en hiper parámetros en regresión lineal, para generar predicciones acertadas evitando el overfiting, después de también haber realizado una normalización de los datos

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


ridge_model = Ridge()


alphas = [0.001, 0.01, 0.1, 1, 10, 100]


scorer = make_scorer(mean_squared_error, greater_is_better=False)


param_grid = {'alpha': alphas}
grid_search = GridSearchCV(ridge_model, param_grid, scoring=scorer, cv=5)
grid_search.fit(X_train_scaled, y_train)


best_model = grid_search.best_estimator_
best_alpha = best_model.alpha


print(f"Mejor valor de alfa: {best_alpha}")


y_pred = best_model.predict(X_test_scaled)
```

Ilustración 16. Creación de hiperparámetros (Grid search)

```
Mejor valor de alfa: 10
Error Cuadrático Medio en el conjunto de prueba: 21275355846388.617
```

Ilustración 17.Mejor valor

Sin embargo se evidencia que al momento de generar predicciones aleatorias no genera unos buenos resultados, independientemente de su escalado como se evidencia a continuación.

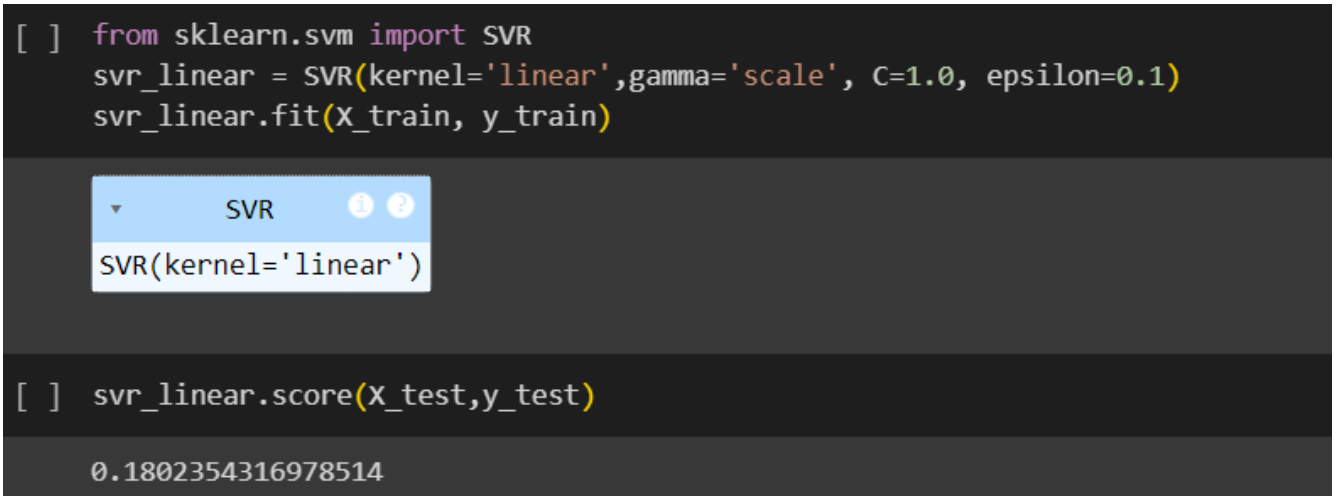



Ilustración 18. Score de SVR.

Como se evidencia con este valor, la predicción en cuanto a máquinas de vectores de soporte, el resultado es demasiado bajo, tendiendo al azar.

Al no generar la predicción suficientemente acertada, se procede a concluir que los datos aportados por la empresa no son materia suficiente para un estudio predictivo de fuga y se necesitaría información más detallada en variables del cliente para generar una correcta clasificación de dicha fuga posible, en este caso, datos sensibles de los clientes que imposibilitan su divulgación y que podrían enriquecer el modelo, pero a la cual no se puede acceder.

Sin embargo se trata de aportar valor a Air Products y al lector generando un modelo de machine learning diseñado para la predicción de productos consumidos por el cliente, con una clasificación multivariable.

```
[ ] ##clasificacion multivariable de productos

# Seleccionar las columnas categóricas
cat_columns_datos = [ 'ActivEconomic','Tipo','Mercado','Region']
# Obtener todas las categorías únicas de todas las columnas categóricas en los sub-dataframes
all_categories = set()
for col in cat_columns_datos:
    all_categories.update(datos[col].unique())


# Crear variables dummy para cada columna categórica en el DataFrame concatenado
df_final_datos_multiple = pd.get_dummies(datos, columns=cat_columns_datos)


df_final_datos_multiple
```

| | AnoMes | activo si/no | SumaDeNetoProd | SumaDeCantProd | GASES PUROS Y ESPECIALES | Contrato de Suministro SI/NO | ActivEconomic_ASTILLEROS / Naval | ActivEconomic_Acuicultura | ActivEconomic_Agroindustria | ActivEconomi No r |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 202010 | 1 | 354735.0 | 6.000000 | ACETILENO | 1 | False | False | False | |
| 1 | 202010 | 1 | 825200.0 | 14.000000 | ACETILENO | 1 | False | False | False | |
| 2 | 202010 | 1 | 1249098.0 | 21.000000 | ACETILENO | 1 | False | False | False | |
| 3 | 202010 | 0 | 2440200.0 | 49.000000 | ACETILENO | 0 | False | False | False | |
| 4 | 202010 | 1 | 295613.0 | 5.000000 | ACETILENO | 1 | False | False | False | |

Ilustración 19. Creación de dataframe final dumificada.

Se realiza un Label encoder para generar nuevas variables correspondientes a los productos de gases puros y especiales consumidos por los clientes y así poder predecir con un 20% de los datos, si un cliente con ciertas características puede comprar “X” producto.

```
[ ] y = df_final_datos_multiple["GASES PUROS Y ESPECIALES"]


    X = df_final_datos_multiple.drop(["GASES PUROS Y ESPECIALES"],axis=1)

from sklearn.preprocessing import LabelEncoder

# Crear un objeto LabelEncoder
label_encoder = LabelEncoder()

# Aplicar el etiquetado a la variable objetivo
y_etiquetada = label_encoder.fit_transform(y)

y_etiquetada

array([ 0,  0,  0, ..., 15, 15, 17])

[ ] ##train and test

[ ]

    X_train, X_test, y_train, y_test = train_test_split(X, y_etiquetada, test_size=0.2, random_state=42)
```

Ilustración 20. Aplicación de Label encoder.

Se escalan los valores y se instancian los modelos para su posterior análisis con base en el accuracy obtenido de máquinas vectores de soporte y Random Forest

```
# Entrenar un modelo de SVM
svm_classifier = SVC(kernel='linear', random_state=42)
svm_classifier.fit(X_train_normalized, y_train)

# Predecir con el modelo de SVM
y_pred_svm = svm_classifier.predict(X_test_normalized)
y_test_svm=y_test
# Evaluar el rendimiento del modelo de SVM
accuracy_svm = accuracy_score(y_test_svm, y_pred_svm)
print("Accuracy SVM:", accuracy_svm)



# Entrenar un modelo de Random Forest
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
rf_classifier.fit(X_train_normalized, y_train)

# Predecir con el modelo de Random Forest
y_pred_rf = rf_classifier.predict(X_test_normalized)
y_test_rf=y_test
# Evaluar el rendimiento del modelo de Random Forest
accuracy_rf = accuracy_score(y_test_rf, y_pred_rf)
print("\nAccuracy Random Forest:", accuracy_rf)
```

```
Accuracy SVM: 0.4370533260032985

Accuracy Random Forest: 0.8103353490929082
```

Ilustración 21. Mayores accuracy de los modelos.

Evidenciando una predicción correcta por medio del Random Forest, por lo cual se procede a imprimir una comparativa del dato real vs la predicción para el modelo con mejor accuracy.

PREDICCION RANDOM FOREST

```
y_pred_ori_rf = label_encoder.inverse_transform(y_pred_rf)
y_test_ori_rf = label_encoder.inverse_transform(y_test_rf)
comparison_df = pd.DataFrame({'Real':y_test_ori_rf , 'Predicted': y_pred_ori_rf})


print(comparison_df)
```

```
                          Real                 Predicted
0                    NITROGENO                      AIRE
1                      OXIGENO                   OXIGENO
2                    NITROGENO                 NITROGENO
3                         AIRE                      AIRE
4                         AIRE                      AIRE
...                        ...                       ...
1814                   OXIGENO                   OXIGENO
1815  MEZCLA ARGON - METANO P-10  MEZCLA ARGON - METANO P-10
1816                 NITROGENO                    ARGON
1817                      AIRE                      AIRE
1818                     ARGON                 ACETILENO

[1819 rows x 2 columns]
```

Ilustración 22. Predicción aplicada de resultados versus datos reales Random Forest.

Para este resultado visualizado se evidencia que el modelo es ajustado en la predicción de productos recomendados que puede comprar el cliente, por lo cual se puede utilizar para generar campañas de retención a clientes que consumen algún producto y que tal vez no tengan una oferta proactiva por parte de la empresa, adicional funciona para generar campañas de acuerdo al producto recomendado para actividades económicas.

Posteriormente a esto, se busca solucionar por medio de alternativas, la problemática de la empresa, de esta manera se halla la forma de visualizar clientes a través de sus características, por medio de agrupaciones que abarcan rangos específicos divididos según el modelo relacionado a continuación, denominado “Clústering”. Con él, se pretenden identificar características que definen agrupaciones de clientes pertenecientes a un título que se le da para su diferenciación.

**Actividad 2.** Aplicar un modelo de Clústering que facilite la segmentación de los clientes a partir de sus diferentes características, realizado por medio de Colab con base en lenguaje python para aplicar al dataset ya estructurado anteriormente.


Ilustración 23.Importación de Librerías y funciones

```python
def plot_multiples_graficas(df, cols, num_cols, num_rows, tipo, targetVar, figsize=(16,8)):
    plt.rcParams['figure.figsize'] = figsize

    #num_plots = len(cols)
    #num_cols = math.ceil(np.sqrt(num_plots))
    #num_rows = math.ceil(num_plots/num_cols)

    fig, axs = plt.subplots(num_rows, num_cols)

    for ind, col in enumerate(cols):
        i = math.floor(ind/num_cols)
        j = ind - i*num_cols

        if num_rows == 1:
            if num_cols == 1:
                if tipo == 'c':
                    sns.countplot(y=df[col], ax=axs, dodge = False, color= '#89A2BE')
                    plt.tick_params(axis='y', labelsize=8)
                if tipo == 'b':
                    sns.boxplot(x=df[col], y=df[targetVar], ax=axs)
                if tipo == 's':
                    sns.scatterplot(x=df[col], y=df[targetVar], ax=axs)
            else:
                if tipo == 'c':
                    sns.countplot(y=df[col], ax=axs[j], dodge = False, color= '#89A2BE')
                    plt.tick_params(axis='y', labelsize=8)
                if tipo == 'b':
                    sns.boxplot(x=df[col], y=df[targetVar], ax=axs[j])
                if tipo == 's':
                    sns.scatterplot(x=df[col], y=df[targetVar], ax=axs[j])
```

Ilustración 24. Configuración  inicial de parámetros.

Se realiza la importación de la base de datos nuevamente y se realiza nuevamente la limpieza y la de

```python
import os
from google.colab import drive
drive.mount('/content/drive')

os.chdir("/content/drive/MyDrive/1. Clases dadas/4. Luisa Lopera")
df = pd.read_excel("Info1.xlsx")
df.shape
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True)
(9101, 36)
```

En principio, la base de datos cuenta con 9.101 registros y 36 variables. Para efectos del análisis, inicialmente se decide selecciona 13 atributos considerados importantes que incluyen información como el año y mes de la compra, el año fiscal, el código del cliente, la fecha de vinculación del cliente, su actividad económica, tipo de cliente, región, mercado, producto, costos netos y por unidad, tipo de gases y contrato de suministro.

Adicionalmente, se decide trabajar solo con la información de los clientes activos con una suma neto mayor a cero y con el año fiscal correspondiente a 2023. Para lograr hacer el ejercicio de Clústering, es fundamental que se tenga registros únicos por cada cliente, es por ello que se decide agrupar por CodCliente, los atributos considerados. Para ello, por código de cliente se toma la fecha mínima y máxima de compra para tener registro de cuándo se realizó la primera y última compra del cliente, con relación a los gastos netos y por cantidad se toma la suma total registrada por código cliente, y para el tipo de gas se crea columnas de acuerdo a cada una de las categorías de esta variable y se registra la cantidad de cada cliente. Para las demás variables estas son fijas para cada cliente.

```python
df = df[df["FY"] == 2023]
df = df[df['activo_si/no'] == 'SI']
df = df[df['SumaDeNetoProd'] >= 0]

columnas = ["AnoMes","FY", "CodCliente", "FECHA_VINCULACION", "Tipo", "Region", "Mercado", "

df = df[columnas]
df.reset_index(drop=True, inplace=True)

tipos_gases = list(set(df["TIPO_GASES"]))
# Agrupar por CodCliente
grouped = df.groupby('CodCliente').agg({
    'AnoMes': ['min', 'max'],
    'FECHA_VINCULACION': 'first',
    'Tipo': 'first',
    'Region': 'first',
    'Mercado': 'first',
    'Producto_Mercado': 'first',
    'SumaDeNetoProd': 'sum',
    'SumaDeCantProd': 'sum',
    'Contrato de Suministro SI/NO': 'first'
})

# Aplanar las columnas
grouped.columns = ['AnoMes_Min', 'AnoMes_Max', 'FECHA_VINCULACION', 'Tipo', 'Region', 'Mercad

# Contar los tipos de gases
tipo_gases_count = df.pivot_table(index='CodCliente', columns='TIPO_GASES', aggfunc='size',

# Unir los resultados
final_df = grouped.join(tipo_gases_count)
final_df.reset_index(inplace=True)
```

Ilustración 25. Filtros y transformaciones.

```
[ ]  def grupo_fecha_vinculacion(fecha):
         if pd.isnull(fecha):
             return 'Fecha_erronea'
         elif fecha.year in range(2018, 2021):
             return 'Clientes Antiguos'
         elif fecha.year in range(2021, 2023):
             return 'Clientes Recientes'
         elif fecha.year >= 2023:
             return 'Clientes Nuevos'

     final_df['GRUPO_FECHA_VINCULACION'] = final_df['FECHA_VINCULACION'].apply(grupo_fecha_vincula
     final_df.drop(["FECHA_VINCULACION"], axis=1, inplace=True)
```

Ilustración 26. Filtrado de antigüedad.

Después de realizado el filtro, al igual que la unificación de código único de registro, se evidencia un consolidado total de 544 clientes únicos en el año fiscal correspondiente a 2023

```
catCols=final_df.select_dtypes(include = ["object"]).columns.tolist()
plot_multiples_graficas(final_df,catCols, 3, 2, 'c', None, (34,12))
```

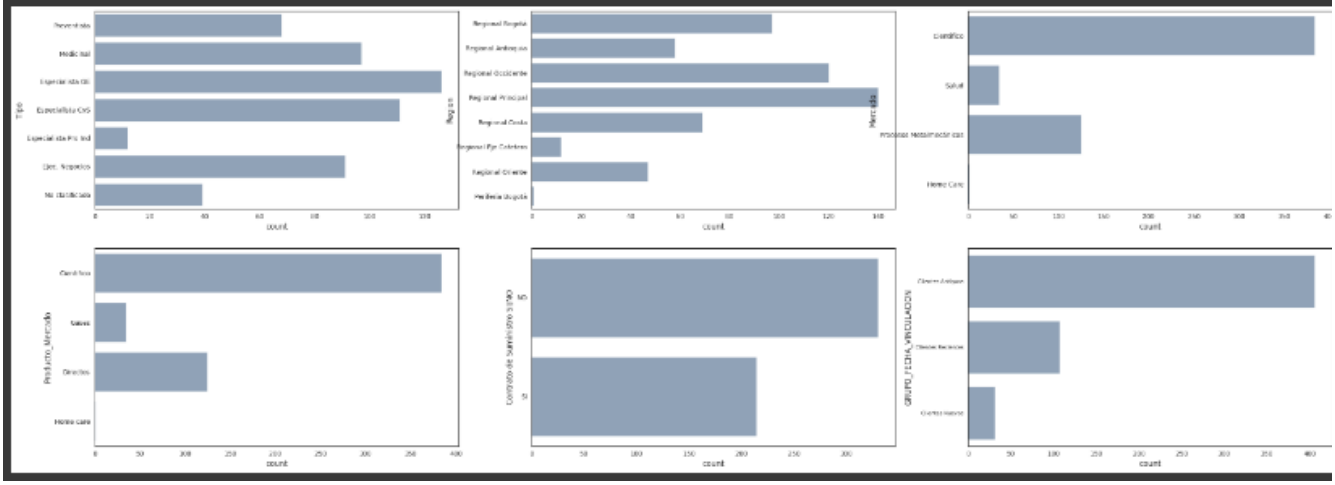


Ilustración 27. Plot múltiple.

La gran mayoría de los clientes analizados son antiguos, y los especialistas en GE y CyS son los más comunes entre ellos. Principalmente, se concentran en la Región Principal, seguida de la región occidental. Nuestro mercado se inclina hacia lo científico, y la mayoría de nuestros clientes no tienen contrato de suministro.

```
[ ]  plt.rcParams['figure.figsize'] = (15,4)
     fig, axs = plt.subplots(2, 1)

     sns.boxplot(ax=axs[0], x=final_df.SumaDeCantProd,flierprops=dict(marker='*'))
     sns.boxplot(ax=axs[1], x=final_df.SumaDeNetoProd,flierprops=dict(marker='*'))
     plt.ticklabel_format(style='plain', axis='x')
```

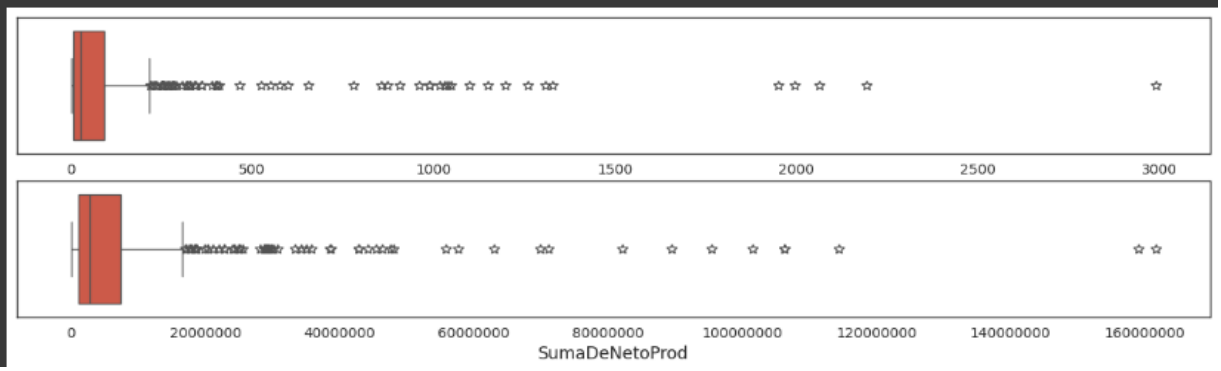



Ilustración 28. Identificación de Outliers.

Se identifican outliers y posteriormente se procede a su normalización.

```
[ ]  plt.rcParams['figure.figsize'] = (15,4)
     fig, axs = plt.subplots(2, 1)

     sns.boxplot(ax=axs[0], x=final_df.SumaDeCantProd,flierprops=dict(marker='*'))
     sns.boxplot(ax=axs[1], x=final_df.SumaDeNetoProd,flierprops=dict(marker='*'))
     plt.ticklabel_format(style='plain', axis='x')
```

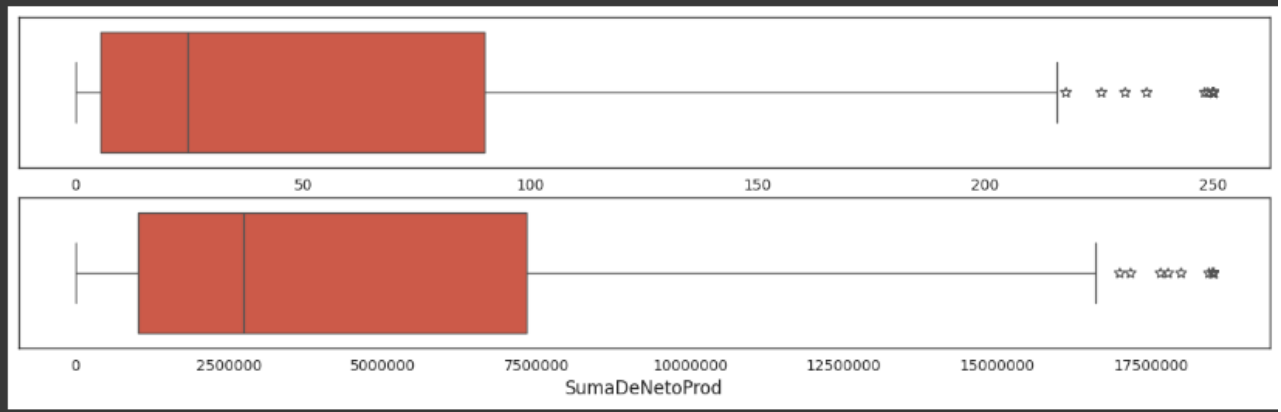



Ilustración 29. Identificación de Outliers posterior.

.

Cerca del 10% de los clientes registran una Suma Neta mayor a los 18 millones y medio aproximadamente, siendo esta una suma elevada, por su parte hay un 2.5% de los clientes

que registran esta suma por encima de los 51 millones aproximadamente llegando incluso a máximos de casi 162 millones, sumas muy atípicas.

```
final_df[["SumaDeNetoProd", "SumaDeCantProd"]].describe([.01,.1,.2,.3,.4,.5,.6,.7,.8,.9,.975])
```

| | SumaDeNetoProd | SumaDeCantProd |
|---|---|---|
| count | 544.00 | 544.00 |
| mean | 8268953.17 | 112.71 |
| std | 17247337.70 | 287.42 |
| min | 0.00 | 0.00 |
| 1% | 178218.27 | 0.72 |
| 10% | 527386.20 | 1.09 |
| 20% | 834280.80 | 2.92 |
| 30% | 1370070.00 | 6.44 |
| 40% | 1805456.40 | 13.36 |
| 50% | 2728279.00 | 24.74 |
| 60% | 3949490.00 | 40.91 |
| 70% | 5953328.20 | 64.00 |
| 80% | 10555595.80 | 120.06 |
| 90% | 18559204.60 | 244.34 |
| 97.5% | 51367707.85 | 1022.80 |
| max | 161633350.00 | 2990.60 |

Ilustración 30. Visualización de cuartiles para las variables monto y cantidad.

```
tab_tipogas=pd.DataFrame(final_df[["ACETILENO", "AIRE", "ARGON", "BUTANO", "DIOXIDO DE CARBONC
          "MEZCLA FUNCION PULMONAR- Pletismografía", "MEZCLAS EMISIONES VEHICULARES", "MEZCLAS
            columns=["Cantidad"]).reset_index().rename(columns={"index":"Variable"}).sort_val
tab_tipogas
```

| | Variable | Cantidad |
|---|---|---|
| 15 | NITROGENO | 617 |
| 1 | AIRE | 489 |
| 2 | ARGON | 407 |
| 13 | MEZCLAS ESPECIALES | 395 |
| 0 | ACETILENO | 353 |
| 17 | OXIGENO | 190 |
| 7 | HIDROGENO | 174 |
| 4 | DIOXIDO DE CARBONO | 56 |
| 12 | MEZCLAS EMISIONES VEHICULARES | 29 |
| 9 | MEZCLA ARGON - METANO P-10 | 19 |
| 8 | METANO | 18 |
| 6 | HEXAFLUORURO | 4 |
| 3 | BUTANO | 4 |
| 11 | MEZCLA FUNCION PULMONAR- Pletismografía | 3 |
| 18 | PROPANO | 3 |
| 14 | MONOXIDO | 2 |
| 16 | OXIDO NITROSO | 2 |

Ilustración 31. Agrupación de gases por cantidad.

La mayoría de clientes han comprado como tipo de gas el nitrógeno con una cantidad de compras registradas en 617, seguido del aire y el argón con cantidades de 489 y 407 respectivamente. Los tipos de gases menos demandados son la mezcla de argón con metano P-5, óxido nitroso, helio UAP y monóxido respectivamente. Luego del proceso anteriormente realizado se convierten las variables a booleanas y se escalan utilizando PCA

```
ianzas_explicativas = pca.explained_variance_ratio_
nt("Las varianzas explicativas de los componentes son: {}".format(varianzas_explicativas))
```

```
Las varianzas explicativas de los componentes son: [0.12653405 0.08701027 0.06720248 0.05367047
 0.04089993 0.03882348 0.03654604 0.03560692 0.03536396 0.03471117]
```

Ilustración 32. Varianzas explicativas.

**Actividad 3. Evidenciar la aplicación de diferentes métodos para la identificación de clústeres y generación de su respectiva caracterización.**

Luego se utiliza el método del codo, para realizar una correcta asignación de clústeres de acuerdo a su cambio con respecto a su variación en la gráfica evidenciada a continuación:

```
modelo = KMeans(n_clusters=4, random_state=1234, n_init=10)
codo = KElbowVisualizer(modelo, k=10, timings=False)
codo.fit(PCA_ds)
codo.show()
```

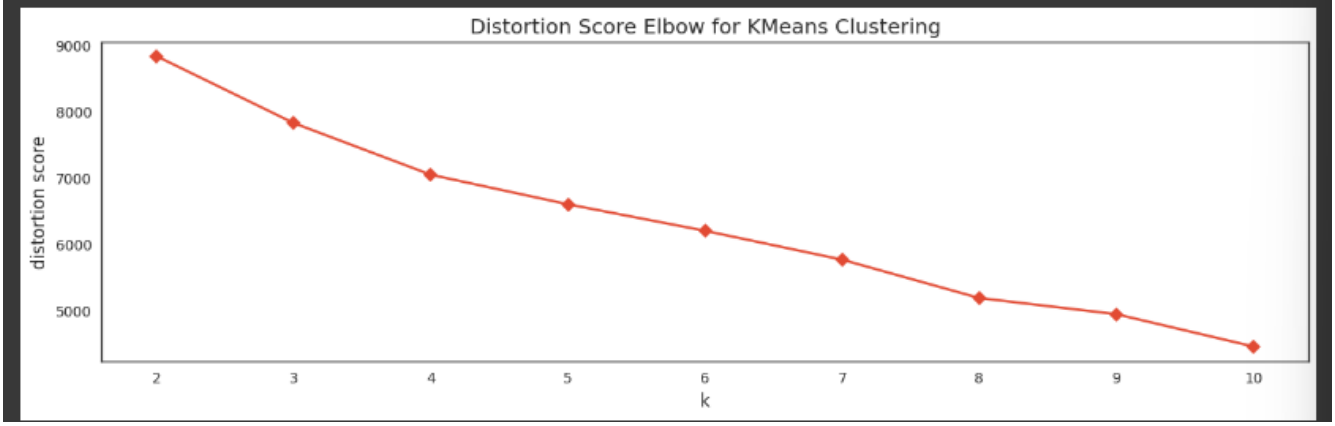



Ilustración 33. Método del Codo (KMeans)

Algo muy similar cuando se trabaja con el modelo silhouette para la asignación completa de clústeres

```
valores_medios_silhouette = []
range_n_clusters = range(2,15)
for n_clusters in range_n_clusters:
    modelo_kmeans = KMeans(
                        n_clusters   = n_clusters,
                        #n_init       = 20,
                        random_state = 10
                    )
    cluster_labels = modelo_kmeans.fit_predict(X)
    silhouette_avg = silhouette_score(X, cluster_labels)
    valores_medios_silhouette.append(silhouette_avg)

fig, ax = plt.subplots(1, 1, figsize=(15, 5))
ax.plot(range_n_clusters, valores_medios_silhouette, marker='o')
ax.set_title("Evolución de media de los índices silhouette")
ax.set_xlabel('Número clusters')
ax.set_ylabel('Media índices silhouette');
```

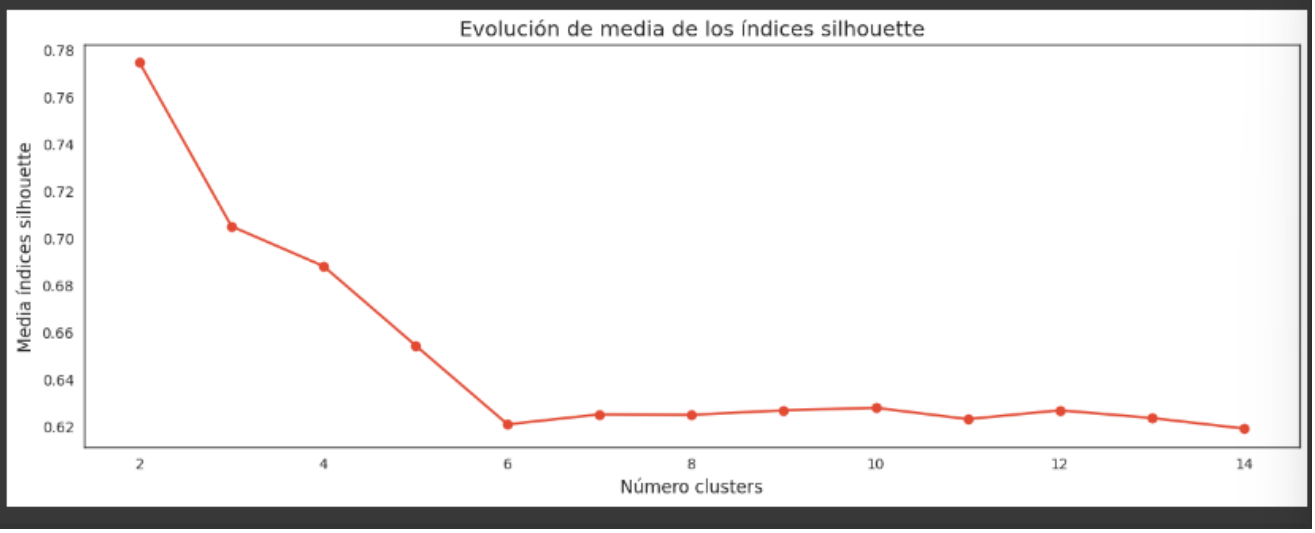



Ilustración 34. Método Silhouette.

De acuerdo al primer gráfico sobre la Evolución de la varianza intra-clúster total basándose en el método de codo (Número de clúster óptimo donde la varianza intra-clúster disminuya considerablemente) y además en base al análisis de la silueta, se decide dividir a los clientes en 4 grupos.

Se instancia el modelo y adicionalmente se procede a visualizar los datos. generados, utilizando PCA como base para dicha segmentación diferenciando cada una por colores, representando sus características.

PCA

```
[ ]  from sklearn.decomposition import PCA

     pca = PCA(n_components=2)
     pca_d = pca.fit_transform(X)
     pca_d_df = pd.DataFrame(data=pca_d, columns = ['componente_1', 'componente_2'])
     pca_tiposClientes = pd.concat([pca_d_df, dClus[['cluster']]], axis =1 )


     fig = plt.figure(figsize = (20,6))
     ax = fig.add_subplot(1,1,1)
     ax.set_xlabel('componente 1')
     ax.set_ylabel('componente 2')
     ax.set_title('Componentes Principales')

     ax.scatter(x=pca_tiposClientes.componente_1, y= pca_tiposClientes.componente_2, c=dClus['cluster'], cmap='v:
     plt.show()
```

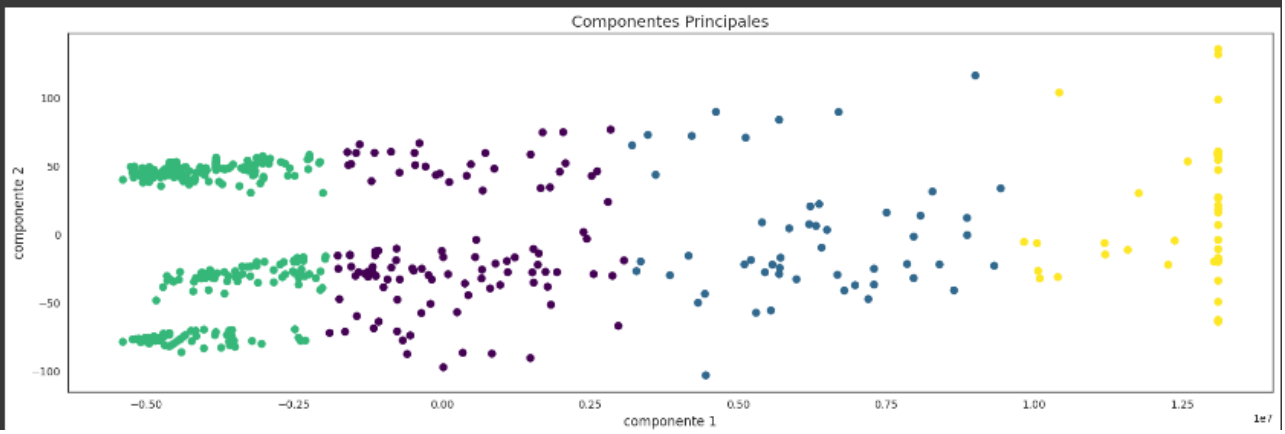



Ilustración 35. Visualización de aglomeraciones PCA.

t-SNE

```
from sklearn.manifold import TSNE

tsne = TSNE(n_components=2, perplexity=100, random_state=42)
tsne_d = tsne.fit_transform(X)
tsne_d_df = pd.DataFrame(data=tsne_d, columns = ['componente_1', 'componente_2'])
tsne_tiposClientes = pd.concat([tsne_d_df, dClus[['cluster']]], axis =1 )

fig = plt.figure(figsize = (20,6))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('componente 1')
ax.set_ylabel('componente 2')
ax.set_title('Componentes t-SNE')


#colors = np.array(list(range(0, 101, 5)))

ax.scatter(x=tsne_tiposClientes.componente_1, y= tsne_tiposClientes.componente_2, c=dClus['cluster'], cmap=
plt.show()
```

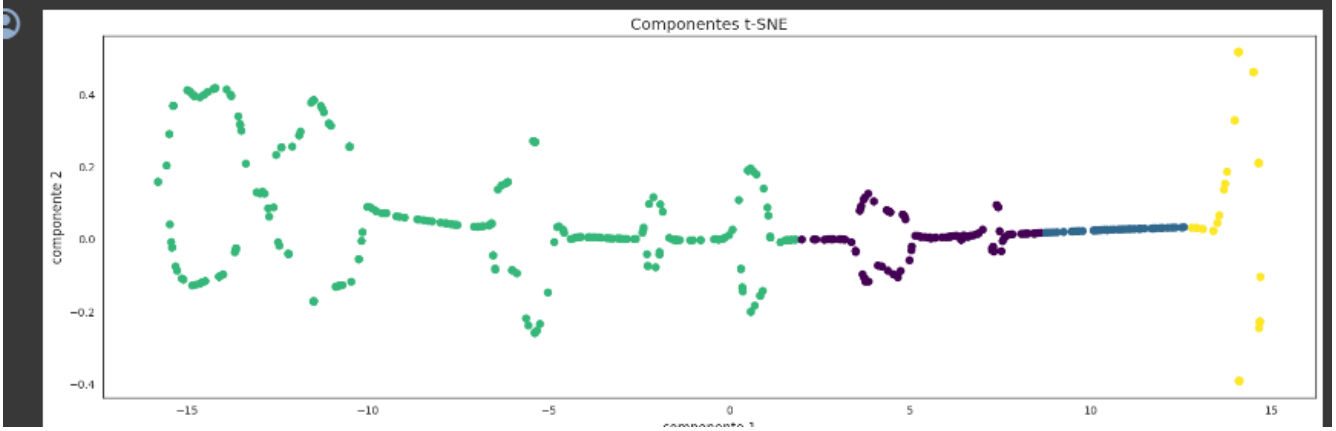



Ilustración 36. Visualización de aglomeraciones T-SNE.

Para confirmación adicional también se busca una homogeneidad en datos, generando por medio de K-means una varianza mínima  y una distancia óptima. Así como, por medio de T-SNE, se conserva la estructura  y las distancias entre los datos vecinos.

## 8. Hierarchical Clustering

```
[ ]  def plot_dendrogram(model, num_clusters, **kwargs):
        '''
        Esta función extrae la información de un modelo AgglomerativeClustering
        y representa su dendograma con la función dendogram de scipy.cluster.hierarchy
        '''

        counts = np.zeros(model.children_.shape[0])
        n_samples = len(model.labels_)
        for i, merge in enumerate(model.children_):
            current_count = 0
            for child_idx in merge:
                if child_idx < n_samples:
                    current_count += 1  # leaf node
                else:
                    current_count += counts[child_idx - n_samples]
            counts[i] = current_count

        linkage_matrix = np.column_stack([model.children_, model.distances_,
                                          counts]).astype(float)

        # Calcula el color_threshold
        distances = linkage_matrix[:, 2]
        idx = np.argsort(distances)
        sorted_distances = distances[idx]
        color_threshold = sorted_distances[-num_clusters+1]

        # Plot
        dendrogram(linkage_matrix, color_threshold=color_threshold, **kwargs)
```

Ilustración 37. Instanciación de Herarquical Clústering.

```
fig, axs = plt.subplots(3, 1, figsize=(20, 12))

plot_dendrogram(modelo_hclust_average, 4, ax=axs[0])
axs[0].set_title("Distancia euclídea, Linkage average")

plot_dendrogram(modelo_hclust_complete, 4,ax=axs[1])
axs[1].set_title("Distancia euclídea, Linkage complete")

plot_dendrogram(modelo_hclust_ward,  4,ax=axs[2])
axs[2].set_title("Distancia euclídea, Linkage ward")

plt.tight_layout();
```

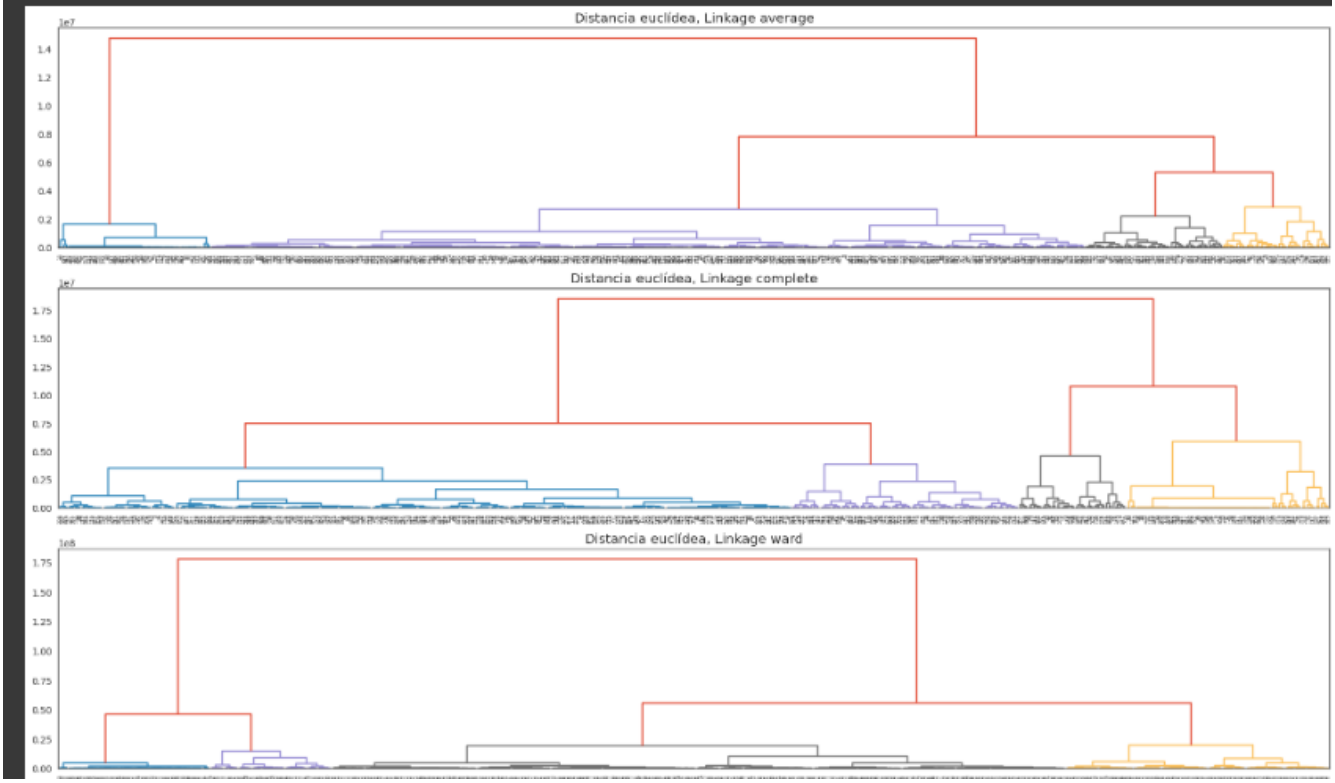



Ilustración 38. Visualización de Herarquical Clústering.

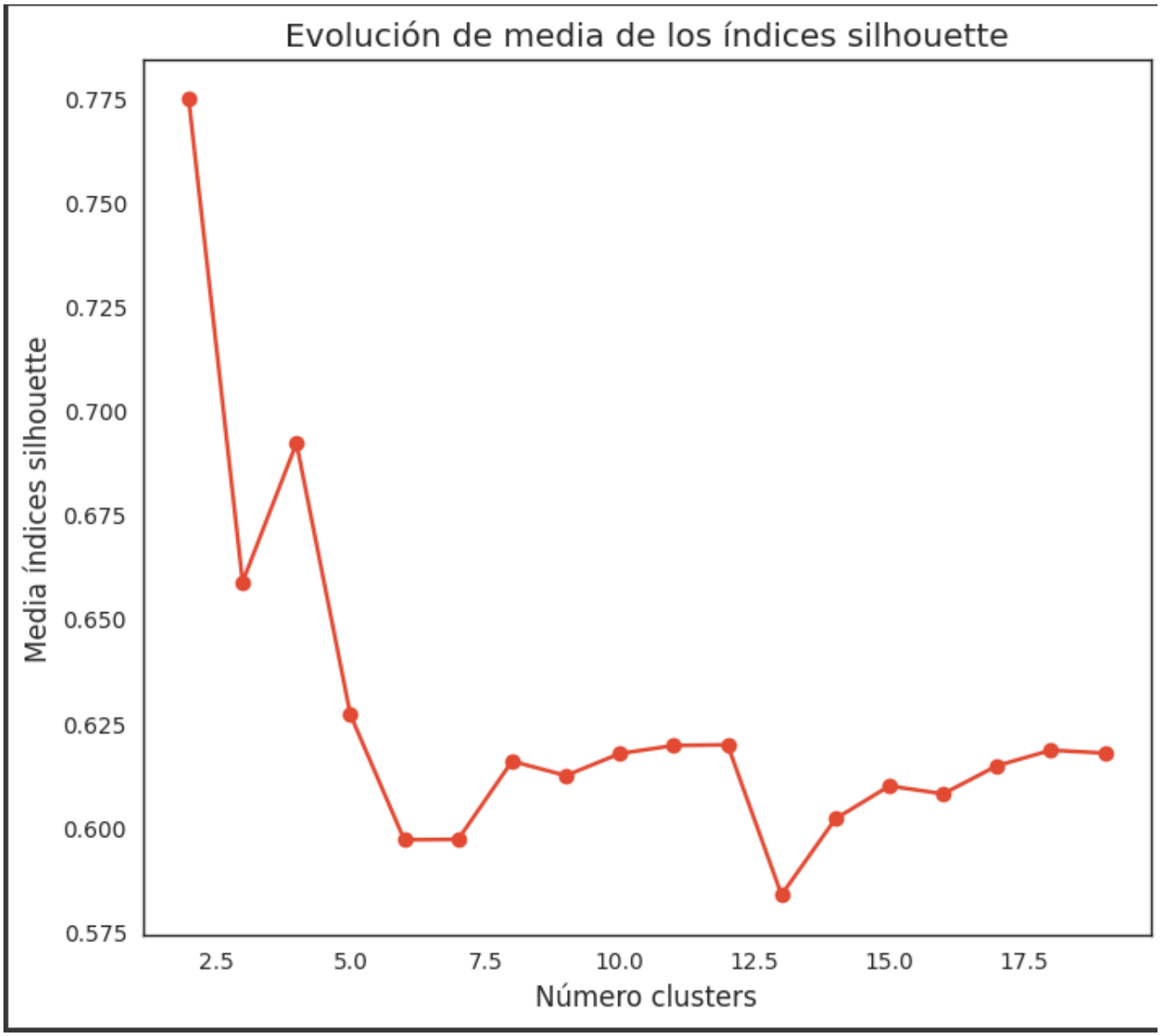



Ilustración 39. Media Índice Silhouette.

Se decide trabajar con 4 clústeres dado que con esta cantidad se obtiene un muy buen índice de silueta cercano a 0.7. Además del método de enlace Ward por presentar en el dendrograma una mejor discriminación entre grupo.

8.2 Visualizacion reducción dimensionalidad

PCA

```
[ ]  from sklearn.decomposition import PCA

     pca = PCA(n_components=2)
     pca_d = pca.fit_transform(X)
     pca_d_df = pd.DataFrame(data=pca_d, columns = ['componente_1', 'componente_2'])
     pca_tiposClientes = pd.concat([pca_d_df, dClus[['cluster2']]], axis =1 )


     fig = plt.figure(figsize = (20,6))
     ax = fig.add_subplot(1,1,1)
     ax.set_xlabel('componente 1')
     ax.set_ylabel('componente 2')
     ax.set_title('Componentes Principales')

     ax.scatter(x=pca_tiposClientes.componente_1, y= pca_tiposClientes.componente_2, c=dClus['cluster2'], cmap='
     plt.show()
```

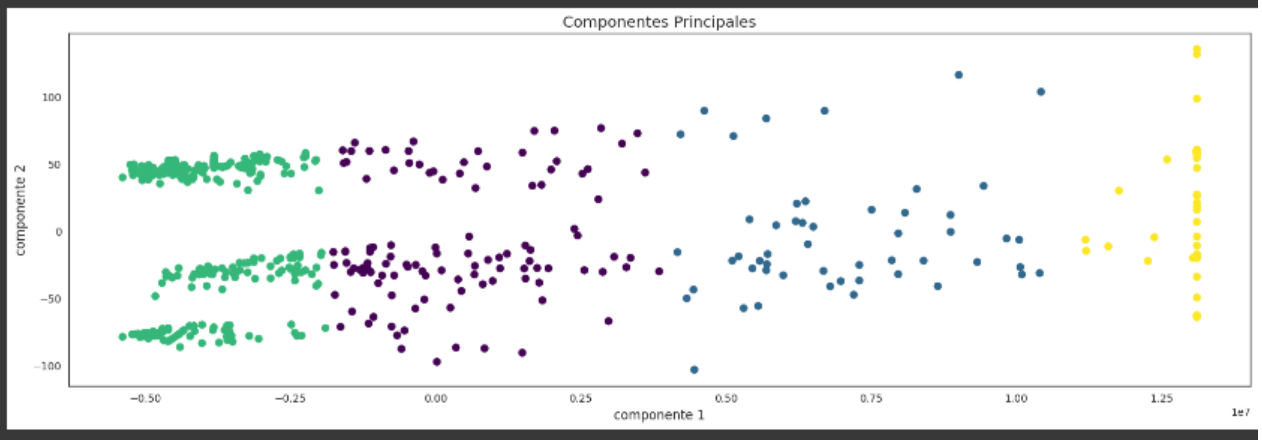



Ilustración 40. Aplicación de PCA.

**Fase V- Evaluación**

**Actividad 1. Evaluar el desempeño de los modelos construidos y Visualizar los resultados del modelo y generar recomendaciones con base en las características de cada clúster**

Decisión del mejor algoritmo de Clústering

Para evaluar y decidir el mejor método de Clústering se tendrá en cuenta además de la visualización de la segmentación generada, el puntaje de la silueta la cual mide se encuentra

entre -1 y 1 y entre más cercana a 1 mejor segmentación, debido a que 1 significa que los Clústeres son muy densos y bien separados. La puntuación de 0 significa que los clústeres se superponen. La puntuación inferior a 0 significa que los datos pertenecientes a los conglomerados pueden ser erróneos/incorrectos.

```
[ ]  cluster1=kmeans.labels_
     print('Silueta K-Medias: '+str(silhouette_score(X, cluster1, metric='euclidean')))

     Silueta K-Medias: 0.6886806988862654

[ ]  cluster2= modelo_hclust_ward.labels_
     print('Silueta Cluster jerárquico Aglomerativo: '+str(silhouette_score(X, cluster2, metric='euclidean')))

     Silueta Cluster jerárquico Aglomerativo: 0.692259043680856
```

Ilustración 41. Ilustración del mejor modelo de Clústering.

Dado lo anterior se decide continuar el análisis con el algoritmo Jerárquico dado un puntaje de silueta mayor e igual a 0.6922 el más alto de los 2 obtenidos.

## 10. Caracterización de los grupos

```
[ ]  final_df["Cluster"]= cluster2
     tab_cluster=pd.DataFrame(final_df["Cluster"].value_counts()).sort_values("Cluster").reset_index().rename(co
     tab_cluster["Porcentaje"] = round(tab_cluster["Cantidad"]/final_df.shape[0]*100,2)
     tab_cluster
```

| | Cluster | Cantidad | Porcentaje |
|---|---|---|---|
| 0 | 0 | 113 | 20.77 |
| 1 | 1 | 52 | 9.56 |
| 2 | 2 | 314 | 57.72 |
| 3 | 3 | 65 | 11.95 |

Ilustración 42. Distribución por clúster.

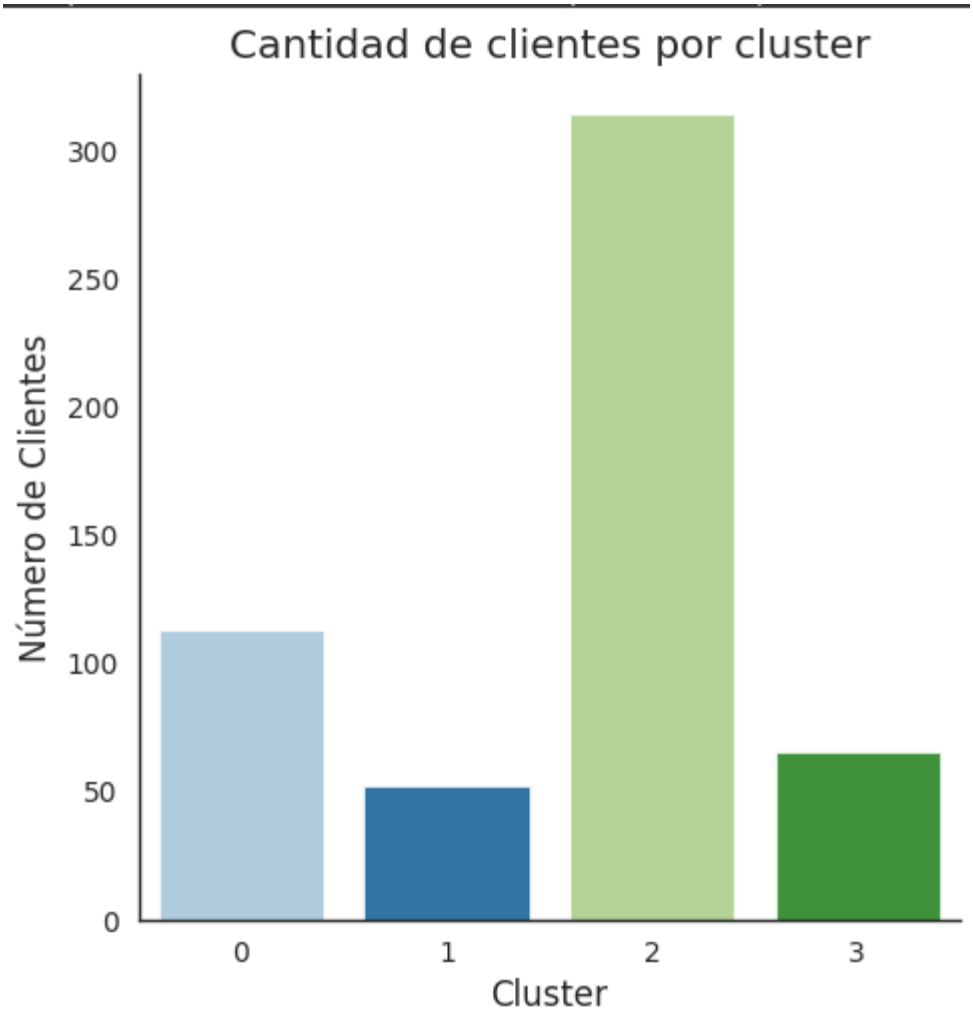



Ilustración 43. Cantidad gráfica por clúster.

La gran mayoría de clientes, precisamente el 57.72%, pertenecen al grupo 2, seguido del grupo 0 que contiene el 20.77% del total de clientes, el grupo 3 se compone del 11.95% de clientes, y el grupo 1 es el que menos clientes tiene, tan solo el 9.56% que son solo 52 en total.

```python
statistics_SumaDeNetoProd = final_df[["Cluster", "SumaDeNetoProd"]].groupby("Cluster").agg({
    "SumaDeNetoProd": [np.mean, np.median, np.std, np.min, np.max, lambda x: np.percentile(x, 25), lambda x
})
# Renombrar las columnas para mayor claridad
statistics_SumaDeNetoProd.columns = ['Promedio', 'Mediana', 'Std Dev', 'Min', 'Max', 'Q1', 'Q3']
statistics_SumaDeNetoProd
```

| Cluster | Promedio | Mediana | Std Dev | Min | Max | Q1 | Q3 |
|---|---|---|---|---|---|---|---|
| 0 | 5729007.95 | 5411775.00 | 1528061.38 | 3630685 | 9240787 | 4321650.00 | 6939055.00 |
| 1 | 12347055.33 | 11982105.50 | 1816208.27 | 9555318 | 15816142 | 10964956.50 | 13704910.00 |
| 2 | 1432805.81 | 1315647.00 | 900059.67 | 0 | 3492765 | 708898.75 | 2101939.25 |
| 3 | 18363871.26 | 18500000.00 | 428470.40 | 16582369 | 18500000 | 18500000.00 | 18500000.00 |

Ilustración 44. Percentiles pertenecientes a cada clúster con relación en ventas.

```python
# Crear la figura y los subplots
fig, axes = plt.subplots(1, 2, figsize=(12, 5))

# Graficar el primer boxplot
sns.boxplot(ax=axes[0], x="Cluster", y="SumaDeNetoProd", data=final_df, flierprops=dict(marker='*'), palett
axes[0].set_title("Suma Neta Producto")

# Graficar el segundo boxplot
sns.boxplot(ax=axes[1], x="Cluster", y="SumaDeCantProd", data=final_df, flierprops=dict(marker='*'), palett
axes[1].set_title("Suma por Cantidad Producto")

# Mostrar los gráficos
plt.tight_layout()
plt.show()
```

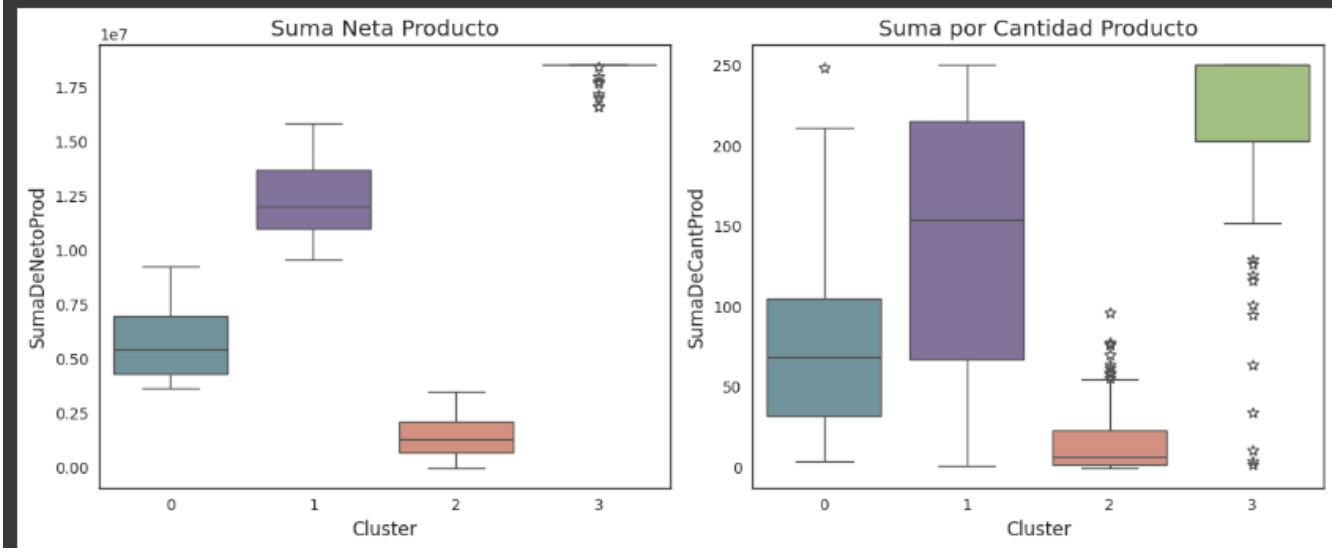



Ilustración 45. Gráficos de bigotes.

El grupo 2 es quien menos ha invertido en las compras de los productos, al presentar en general las menores sumas netas, en este grupo el 75% de los clientes gastan menos de 2 millones 100 mil aproximadamente. Mientras que por su parte el grupo 3 es quien gasta más, donde la mitad de los clientes que lo componen pueden llegar a gastar más de 18 millones y medio. Por su parte, el grupo 0 es el segundo en menos en gastar en los productos, donde el 75% de los clientes gastan menos de 7 millones aproximadamente, y el grupo 1 suele gastar más, pues la mitad de sus clientes pueden llegar a gastar más de 12 millones aproximadamente.

```
statistics_AnoMes_Min = final_df[["Cluster", "AnoMes_Min"]].groupby("Cluster").agg({
    "AnoMes_Min": [np.mean, np.median, np.std, np.min, np.max, lambda x: np.percentile(x, 25), lambda x: np.
})
# Renombrar las columnas para mayor claridad
statistics_AnoMes_Min.columns = ['Promedio', 'Mediana', 'Std Dev', 'Min', 'Max', 'Q1', 'Q3']
statistics_AnoMes_Min
```

| Cluster | Promedio | Mediana | Std Dev | Min | Max | Q1 | Q3 |
|---|---|---|---|---|---|---|---|
| 0 | 202239.46 | 202211.00 | 43.16 | 202210 | 202308 | 202210.00 | 202301.00 |
| 1 | 202226.62 | 202210.00 | 35.51 | 202210 | 202307 | 202210.00 | 202212.00 |
| 2 | 202262.57 | 202301.00 | 46.40 | 202210 | 202308 | 202211.00 | 202304.00 |
| 3 | 202229.03 | 202210.00 | 37.71 | 202210 | 202307 | 202210.00 | 202212.00 |

Ilustración 46. Percentiles pertenecientes a cada clúster con relación a año de ingreso.

```
statistics_AnoMes_Max = final_df[["Cluster", "AnoMes_Max"]].groupby("Cluster").agg({
    "AnoMes_Max": [np.mean, np.median, np.std, np.min, np.max, lambda x: np.percentile(x, 25), lambda x:
})
# Renombrar las columnas para mayor claridad
statistics_AnoMes_Max.columns = ['Promedio', 'Mediana', 'Std Dev', 'Min', 'Max', 'Q1', 'Q3']
statistics_AnoMes_Max
```

| Cluster | Promedio | Mediana | Std Dev | Min | Max | Q1 | Q3 |
|---|---|---|---|---|---|---|---|
| 0 | 202295.29 | 202307.00 | 30.48 | 202210 | 202308 | 202304.00 | 202308.00 |
| 1 | 202304.65 | 202307.00 | 13.25 | 202212 | 202308 | 202305.00 | 202308.00 |
| 2 | 202284.75 | 202304.00 | 38.79 | 202210 | 202308 | 202301.00 | 202306.00 |
| 3 | 202299.66 | 202308.00 | 25.72 | 202211 | 202308 | 202306.00 | 202308.00 |

Ilustración 47.Percentiles pertenecientes a cada clúster con relación a año de ingreso

En cuanto a la antigüedad de los clientes, los grupos 1 y 3 exhiben un patrón similar, con la mitad de sus clientes realizando su primera compra antes de octubre de 2022, la fecha más antigua registrada para el año fiscal seleccionado. Por otro lado, el grupo 0 muestra que la mitad de sus clientes se vincularon antes de noviembre de 2022. En contraste, el grupo 2 parece conformarse por clientes con vínculos más recientes, ya que la mitad de ellos realizaron su primera compra después de enero de 2023.

En lo que respecta a la fecha de la última compra registrada de los clientes, destaca que el grupo 3 parece tener las compras más recientes, con la mitad de sus clientes realizando su última compra después de agosto de 2023. Le siguen los grupos 0 y 1, donde la mitad de los clientes efectuaron su última compra después de julio de 2023. Mientras tanto, el grupo 2 está compuesto por clientes cuya última compra fue hace más tiempo, ya que la mitad de ellos compraron por última vez después de abril de 2023.

Estos hallazgos sugieren que el grupo 2 está integrado por clientes que tardaron más en vincularse, pero que también cesaron sus compras más rápidamente, evidenciado por la antigüedad de su última compra. Además, el grupo 3 se caracteriza por tener clientes con mayor antigüedad, pero también con compras más actuales, lo que sugiere una mayor fidelidad en este segmento de clientes. En cuanto a los grupos 0 y 1, presentan una antigüedad y recencia similares, encontrándose en un equilibrio donde son unos de los más antiguos pero también han realizado compras recientemente.

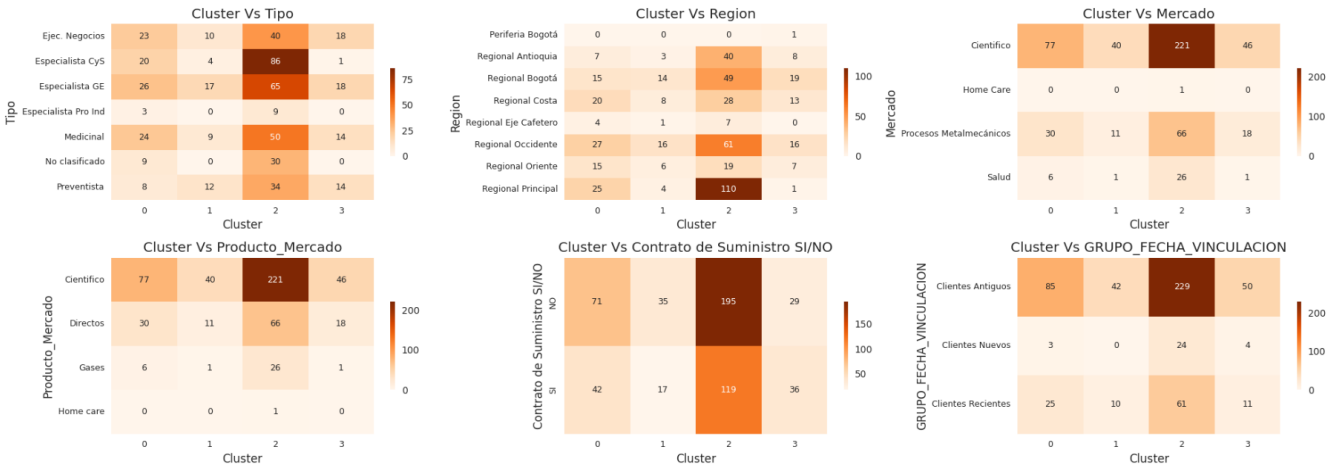



Ilustración 48. Visualización varia de Clústeres

Según el tipo de clientes, se observa que el grupo 1 está mayormente compuesto por

Especialistas GE, representando el 32.7% del total, seguido de los Preventistas. En cambio, en el grupo 2, la mayoría (27.4%) son Especialistas CyS. El grupo 0 abarca una variedad de tipos de clientes, desde Especialistas GE hasta Ejecutivos de Negocios y Médicos. Mientras tanto, el grupo 3 consiste exclusivamente en Ejecutivos de Negocios, Especialistas GE, Médicos y Preventistas.

En cuanto a la región, el grupo 0 se concentra mayormente en las regiones Occidental y Principal. En contraste, el 35% del grupo 2 se encuentra en la región Principal. Los clientes del grupo 1 están principalmente distribuidos en las regiones Occidental y de Bogotá. Por último, el grupo 3 tiene una representación significativa en varias regiones, con predominio en la Occidental, Bogotá y la Costa.

Respecto al mercado, prevalece una tendencia uniforme en los cuatro grupos, donde los productos científicos dominan con una representación superior al 68% del total de clientes.

Respecto al contrato de suministros, se destaca que en el grupo 3, la mayoría (55.4%) sí tienen este contrato, a diferencia de los otros tres grupos donde más del 60% de los clientes no lo poseen.

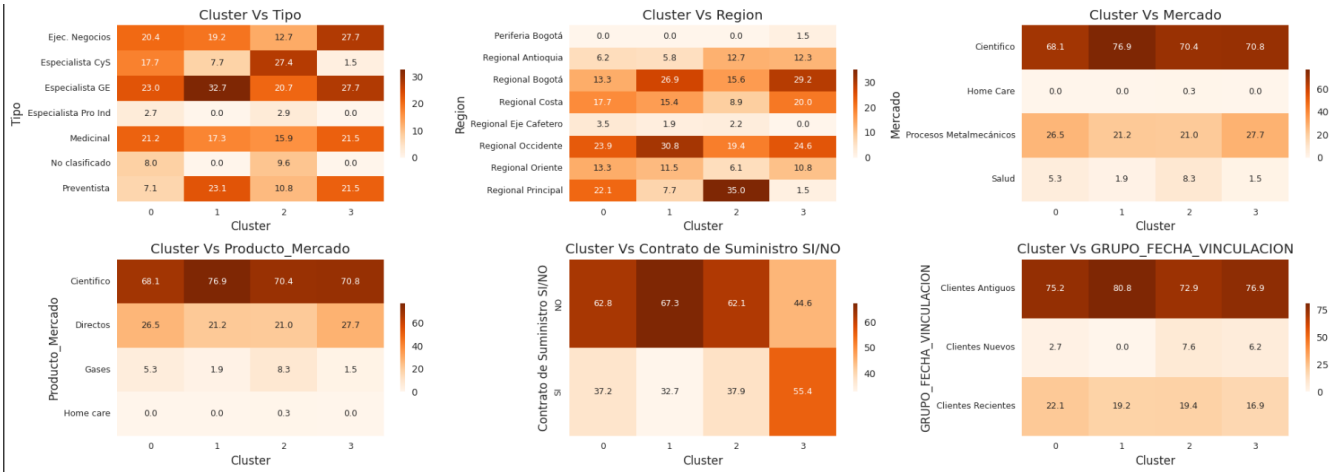



Ilustración 49. Visualizaciones varias por clústeres.

```
suma_gases=final_df[["Cluster","ACETILENO", "AIRE", "ARGON", "BUTANO", "DIOXIDO DE CARBONO", "HELIO UAP", "HEXAFLUORURO", "HIDROGENO", "METANO", "MEZCLA ARGON - METANO P-10",
          "MEZCLA ARGON - METANO P-5","MEZCLA FUNCION PULMONAR- Pletismografía", "MEZCLAS EMISIONES VEHICULARES", "MEZCLAS ESPECIALES", "MONOXIDO", "NITROGENO", "OXIDO NITROSO",
          "OXIGENO", "PROPANO"]].groupby("Cluster").sum()

suma_por_cluster = suma_gases.sum(axis=1)
prop_gases=suma_gases.div(suma_por_cluster, axis=0).transpose()*100
```

Ilustración 50. Agrupación de gases por cantidad.

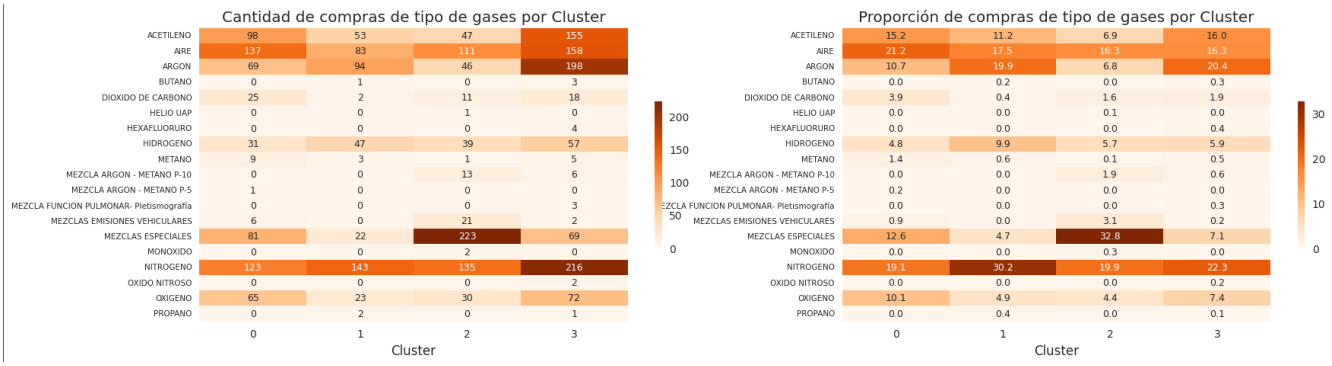



Ilustración 51. Agrupación de gases por clúster.

Finalmente, de acuerdo al tipo de gas, se observa que el grupo 0 en su mayoría compran los tipos gases correspondientes a Aire (21.2%) y Nitrógeno (19.1%). El grupo 1 por su parte, la gran mayoría, el 30.2% compra Nitrogeno, seguido de Argon con el 20%. Por su parte, el grupo 2 en su gran mayoría el 32.8% de clientes se concentran en la compra de mezclas especiales, seguidos del 20% que compra Nitrogeno. Y por su parte, el grupo 3 se compone de clientes que compra mayor variedad de tipos de gases, predominando el nitrogeno (22.3%), el argon (20.4%), el aire (16.3%) y acetileno (16%).

## Actividad 2. Identificar de los resultados obtenidos por la Clústerización.

### Resumen de las caracterizaciones de los grupos

**El grupo 0** se caracteriza por ser uno de los segmentos con menor inversión en productos químicos, con aproximadamente el 75% de sus clientes gastando menos de 7 millones. Sus clientes tienden a tener una antigüedad moderada en términos de la fecha de su primera compra, con la mitad de ellos vinculados antes de noviembre de 2022. Sin embargo, también muestran una recencia significativa en sus compras, con la mitad de ellos realizando su última compra después de julio de 2023. En cuanto al tipo de cliente, abarca una variedad que incluye desde Especialistas GE hasta Ejecutivos de Negocios y Médicos. Se concentra mayormente

en las regiones Occidental y Principal, y muestra una preferencia por los gases como Aire y Nitrógeno, con una proporción considerable de contratos de suministro. En resumen, el grupo 0 se destaca por su diversidad en términos de clientes y preferencias de compra, con una inversión moderada pero con una presencia sólida en diferentes aspectos del mercado químico. Este grupo puede llamarse como **"Clientes Potenciales"** y se caracterizan por tener una frecuencia de compra moderada, un valor monetario moderado y por haber pasado un tiempo desde su última transacción. Aunque han estado inactivos recientemente, estos clientes aún tienen potencial para aumentar su valor si se toman medidas para reactivar su participación.

**El grupo 1** se caracteriza por ser uno de los segmentos de clientes que muestran un mayor gasto en productos químicos, con la mitad de sus clientes capaces de gastar más de 12 millones aproximadamente. Además, este grupo exhibe una antigüedad significativa, ya que la mitad de sus clientes realizaron su primera compra antes de octubre de 2022. A pesar de esta longevidad, también demuestran actividad reciente en sus compras, con la mitad de ellos realizando su última compra después de julio de 2023. En términos de tipos de clientes, este grupo está principalmente compuesto por Especialistas GE, seguidos de Preventistas. En cuanto a la distribución regional, se encuentran principalmente en las regiones Occidental y de Bogotá. A nivel de mercado, muestran una preferencia por los productos científicos, que representan la mayoría de sus compras. En cuanto al tipo de gas, tienen una preferencia por el nitrógeno, seguido del argón. En resumen, el grupo 1 se destaca por su alto gasto, antigüedad en el servicio, actividad reciente en compras, predominio de Especialistas GE, presencia en regiones específicas y preferencia por ciertos tipos de gases, como el nitrógeno y el argón. Este grupo puede llegar a llamarse como **"Clientes Leales"** que se caracterizan por incluir clientes que tienen una alta frecuencia de compra, han realizado compras recientes y poseen un valor monetario moderado. Aunque no gastan tanto como los clientes de alto valor, estos clientes son leales y consistentes en sus compras, lo que los hace valiosos para la empresa a largo plazo.

**El grupo 2** se distingue por ser el que menos ha invertido en compras de productos químicos, con una predominancia de clientes cuyos gastos se sitúan por debajo de los 2 millones 100

mil aproximadamente. Además, este grupo está conformado en su mayoría por clientes con vínculos más recientes, evidenciado por el hecho de que la mitad de ellos realizaron su primera compra después de enero de 2023. Respecto a la fecha de la última compra registrada, el grupo 2 está compuesto por clientes cuya última adquisición fue hace más tiempo, ya que la mitad de ellos compraron por última vez después de abril de 2023. Esto sugiere un patrón de vinculación tardío y una menor frecuencia de compras en comparación con los otros grupos. En cuanto al tipo de clientes, el grupo 2 está principalmente compuesto por Especialistas CyS, representando alrededor del 27.4% del total. Respecto al mercado, al igual que otros grupos, el mercado de productos científicos es dominante entre sus compras, con una representación superior al 68% del total de clientes. En cuanto al tipo de gas, destaca por una alta proporción de clientes que compran mezclas especiales, seguido de Nitrogeno como el segundo tipo de gas más adquirido por este grupo. En resumen, el grupo 2 se compone de clientes que no compran con alta frecuencia, pues sus últimas compras fueron hace mucho y además son compras de bajos montos, posiblemente por realizar compras en su mayoría de mezclas especiales. Este grupo de clientes se puede nombrar como **"Clientes en Riesgo"**, debido a que han estado inactivos durante algun tiempo y tienen una baja frecuencia de compra, bajos valores monetarios y ha pasado tiempo desde su última transacción. Pueden ser clientes que están en riesgo de perderse si no se toman medidas para reactivar su participación, como por ejemplo mayor oferta y promociones en mezclas especiales.

**El grupo 3** se destaca por ser el segmento que realiza las mayores inversiones en productos químicos, con la mitad de sus clientes capaces de gastar más de 18 millones y medio, lo que lo posiciona como el grupo más proclive a realizar compras significativas. Además, este grupo exhibe una combinación peculiar de antigüedad y recencia en sus compras: a pesar de que la mitad de sus clientes están vinculados desde hace bastante tiempo, también presentan las compras más recientes, lo que sugiere una alta fidelidad y actividad continua en términos de adquisiciones de productos químicos. Por otro lado, en términos de tipo de cliente, el grupo 3 se compone exclusivamente de Ejecutivos de Negocios, Especialistas GE, Médicos y Preventistas, lo que podría indicar una preferencia por productos de alto valor y una demanda especializada. Este grupo se puede nombrar como **"Clientes de Alto Valor"** pues se caracterizan por tener una alta frecuencia de compra, realizar compras recientes y tener un alto valor monetario. Representan el grupo más valioso para la empresa, contribuyendo significativamente a los ingresos y mostrando una alta probabilidad de lealtad.

+ El grupo 0 se caracteriza por ser uno de los segmentos con menor inversión en productos químicos, con aproximadamente el 75% de sus clientes gastando menos de 7 millones. Sus clientes tienden a tener una antigüedad moderada en términos de la fecha de su primera compra, con la mitad de ellos vinculados antes de noviembre de 2022. Sin embargo, también muestran una recencia significativa en sus compras, con la mitad de ellos realizando su última compra después de julio de 2023. En cuanto al tipo de cliente, abarca una variedad que incluye desde Especialistas GE hasta Ejecutivos de Negocios y Médicos. Se concentra mayormente en las regiones Occidental y Principal, y muestra una preferencia por los gases como Aire y Nitrógeno, con una proporción considerable de contratos de suministro. En resumen, el grupo 0 se destaca por su diversidad en términos de clientes y preferencias de compra, con una inversión moderada pero con una presencia sólida en diferentes aspectos del mercado químico. Este grupo puede llamarse como "Clientes Potenciales" y se caracterizan por tener una frecuencia de compra moderada, un valor monetario moderado y por haber pasado un tiempo desde su última transacción. Aunque han estado inactivos recientemente, estos clientes aún tienen potencial para aumentar su valor si se toman medidas para reactivar su participación.

+ El grupo 1 se caracteriza por ser uno de los segmentos de clientes que muestran un mayor gasto en productos químicos, con la mitad de sus clientes capaces de gastar más de 12 millones aproximadamente. Además, este grupo exhibe una antigüedad significativa, ya que la mitad de sus clientes realizaron su primera compra antes de octubre de 2022. A pesar de esta longevidad, también demuestran actividad reciente en sus compras, con la mitad de ellos realizando su última compra después de julio de 2023. En términos de tipos de clientes, este grupo está principalmente compuesto por Especialistas GE, seguidos de Preventistas. En cuanto a la distribución regional, se encuentran principalmente en las regiones Occidental y de Bogotá. A nivel de mercado, muestran una preferencia por los productos científicos, que representan la mayoría de sus compras. En cuanto al tipo de gas, tienen una preferencia por el nitrógeno, seguido del argón. En resumen, el grupo 1 se destaca por su alto gasto, antigüedad en el servicio, actividad reciente en compras, predominio de Especialistas GE, presencia en regiones específicas y preferencia por ciertos tipos de gases, como el nitrógeno y el argón. Este grupo puede llegar a llamarse como "Clientes Leales" que se caracterizan por

incluir clientes que tienen una alta frecuencia de compra, han realizado compras recientes y poseen un valor monetario moderado. Aunque no gastan tanto como los clientes de alto valor, estos clientes son leales y consistentes en sus compras, lo que los hace valiosos para la empresa a largo plazo.

+ El grupo 2 se distingue por ser el que menos ha invertido en compras de productos químicos, con una predominancia de clientes cuyos gastos se sitúan por debajo de los 2 millones 100 mil aproximadamente. Además, este grupo está conformado en su mayoría por clientes con vínculos más recientes, evidenciado por el hecho de que la mitad de ellos realizaron su primera compra después de enero de 2023. Respecto a la fecha de la última compra registrada, el grupo 2 está compuesto por clientes cuya última adquisición fue hace más tiempo, ya que la mitad de ellos compraron por última vez después de abril de 2023. Esto sugiere un patrón de vinculación tardío y una menor frecuencia de compras en comparación con los otros grupos. En cuanto al tipo de clientes, el grupo 2 está principalmente compuesto por Especialistas CyS, representando alrededor del 27.4% del total. Respecto al mercado, al igual que otros grupos, el mercado de productos científicos es dominante entre sus compras, con una representación superior al 68% del total de clientes. En cuanto al tipo de gas, destaca por una alta proporción de clientes que compran mezclas especiales, seguido de Nitrogeno como el segundo tipo de gas más adquirido por este grupo. En resumen, el grupo 2 se compone de clientes que no compran con alta frecuencia, pues sus últimas compras fueron hace mucho y además son compras de bajos montos, posiblemente por realizar compras en su mayoría de mezclas especiales. Este grupo de clientes se puede nombrar como "Clientes en Riesgo", debido a que han estado inactivos durante algun tiempo y tienen una baja frecuencia de compra, bajos valores monetarios y ha pasado tiempo desde su última transacción. Pueden ser clientes que están en riesgo de perderse si no se toman medidas para reactivar su participación, como por ejemplo mayor oferta y promociones en mezclas especiales.

**Fase V- Despliegue**

**Actividad 1. Generar estrategias para la empresa Air Products. Por medio de recomendaciones de producto y Clústerización.**

Se realiza una socialización inicial con la empresa Air Products, con los estudiantes de especialización, por medio de presentación de resultados de cada objetivo y se genere una reunión foco al área comercial para implementación de estrategias que complementen el análisis realizado. Se genera una capacitación a dos personas pertenecientes al área para utilizar la herramienta Colab, quienes en adelante, en conjunto con la estudiante Luisa Fernanda Lopera, realizarán la documentación de procesos y resultados obtenidos durante el piloto y la puesta en producción del modelo para la empresa. Con esto, se entregan tanto el modelo de recomendación de productos (Predictivo), como los modelos de Clústerización (K-means, Silhouette, t-SNE y Hierarchical) que beneficiará tanto a la empresa en general en términos de mejor EBITDA por incremento de ventas, reteniendo a clientes importantes y permitiendo enfocarse en clientes nuevos y ganar cuota de mercado. Así como a nivel micro a su área de Marketing, ya que no poseen un área de analítica consolidada y será un proceso de establecimiento a mediano plazo.

Posterior a la socialización, se realiza el despliegue de los modelos generando análisis de los resultados de manera mensual, con foco en los clientes sin contrato de suministro para su monitoreo constante con el modelo de clúster y con foco en los clientes de contrato para recomendación por medio del modelo predictivo. Por ende se debe tener un esfuerzo continuado por la correcta implementación del área de marketing en alimentar correctamente los datos de los clientes que no solo permitan realizar ajustes y mantenimiento constantes a los modelos ya proporcionados a la empresa, sino también para concretar la creación de los modelos predictivos esperados inicialmente, que por falta de información y el no suministro para fines académicos no resultó en una predicción acertada, por lo cual para manejo interno, permitirá a la empresa establecer una cantidad ilimitada de clasificaciones que ayuden a su objetivo por medio de esta materia prima. Por ende, es de vital importancia que se realice una evaluación y retroalimentación de las estrategias proporcionadas a continuación.

**Estrategias**

**Enfocadas al Grupo 0 (Clientes Potenciales):**

1. **Programas de Reactivación:** Realizar una planificación estratégica de reactivación para este grupo por medio de incentivos de retención, apoyado en descuentos para compras futuras, campañas exclusivas para el segmento y recomendación de productos.

2. **Personalización:** Utilizar estrategias de marketing personalizado para ofrecer productos o servicios, como puede ser planificación de pagos de este acreedor a su medida, ya que diferir el pago en días a su capacidad de rotar cartera, por lo que deben requerir adaptabilidad a las preferencias individuales de cada cliente en este grupo.

3. **Seguimiento Activo:** Mantener un seguimiento activo de la actividad de estos clientes, siendo intensivos, sin entorpecer la relación, enviando recordatorios o comunicaciones periódicas para mantener su interés y participación. Así como para identificar necesidades que se traduzcan en mayores ventas.

**Enfocado en Grupo 1 (Clientes Leales):**

1. **Programas de Fidelización:** Implementar programas de fidelización que recompensen la lealtad de estos clientes con beneficios exclusivos, como acceso a eventos VIP, socialización de resultados del grupo al que pertenecen y capacitaciones, adicional a esto, ofrecer descuentos especiales o servicios premium.

2. **Ofertas Personalizadas:** Ofrecer ofertas y promociones personalizadas basadas en el historial de compras y preferencias de productos de cada cliente en este grupo, enfocándose en los recomendados por el modelo para cada actividad económica en especial.

3. **Feedback Activo:** Solicitar retroalimentación regularmente a estos clientes para entender sus necesidades y expectativas, y ajustar las estrategias en consecuencia.

**Enfocado en Grupo 2 (Clientes en Riesgo):**

1. **Reactivación Proactiva:** Implementar campañas proactivas de reactivación, ofreciendo incentivos atractivos y destacando el valor agregado de los productos o servicios de la empresa.

2. **Segmentación Específica:** Segmentar este grupo en subgrupos según sus características y comportamientos de compra para diseñar estrategias más enfocadas y efectivas como ofertas por producto consumido por este grupo, tal y como son los gases con fines científicos.

3. **Comunicación Persuasiva:** Utilizar una comunicación persuasiva que destaque los beneficios únicos que la empresa ofrece a estos clientes, con énfasis en las mezclas especiales y otros productos de interés para ellos.

**Enfocado en Grupo 3 (Clientes de Alto Valor):**

1. **Gestión de Cuentas Clave:** Designar equipos especializados para la gestión de cuentas clave en este grupo, en pocas palabras, generar una cartera para un Key Account Manager brindando un servicio personalizado y atención prioritaria al servicio y venta B2B a la empresa.

2. **Programas VIP:** Implementar programas VIP exclusivos para estos clientes, personalizando pagos de acuerdo a necesidades, creación de ecosistema de aliados con referencias de y relacionamiento de empresas que atiende Air Products para crear conglomerados que compren mayoreo a mejor precio y establezcan precios competitivos, ofreciendo beneficios premium como servicios personalizados a estas empresas durante el primer año, por medio de la compra de cierto volumen de gases y sosteniendo esa promesa de servicio gratuito a través de objetivos anuales de compra. Prioridad en contratos de suministro en caso de emergencias a nivel país y global como la pandemia, en la cual existió escasez, generando contratos y relaciones a largo plazo.

3. **Cross-selling y Up-selling:** Identificar oportunidades de cross-selling y up-selling para aumentar el valor de cada transacción y fortalecer la relación con estos clientes de alto valor.

Estas estrategias pueden adaptarse y ajustarse según las necesidades y características específicas de cada grupo de clientes, con el objetivo de retener su lealtad y maximizar su valor para la empresa.

## Resultados

**Objetivo General:** Construir un modelo basado en técnicas de machine learning, que mitigue la fuga de clientes para la empresa AIRPRODUCTS COLOMBIA.

**Resultado**

Se generan varios modelos de machine learning que permiten realizar una correcta retención de los clientes que poseen contrato de suministro y que compran retail, concluyendo en sugerencias que incluso pueden mejorar el relacionamiento a largo plazo con estos últimos mencionados, todo esto, por medio de la consecución completa de los objetivos específicos y que puede ayudar a reducir de forma significativa la merma productiva actual encontrada durante la realización del objetivo específico número uno.

**Objetivos Específicos**

- Implementar un proceso de análisis exploratorio de los datos incluyendo extracción, limpieza, transformación, selección y análisis de datos, de los diferentes clientes que posee la empresa AIRPRODUCTS COLOMBIA.

  Se realiza un EDA por medio de Colab que permite poseer a la empresa un dataset completamente limpio y preparado para su posterior modelamiento, esto por medio del cumplimiento al 100% del objetivo, esto logrado con el correcto entendimiento del negocio, de sus datos, una minuciosa limpieza de cada una de las variables, caracterizado por una preparación y transformación acorde al objetivo planteado.

- Utilizar las técnicas de machine learning para la construcción de modelos predictivos a partir de los datos de los clientes.

  Se crean un total de 5 modelos enfocados en la prevención de fuga de clientes, los cuales son K-means, T-SNE, Herarquical y clasificación multivariable por medio de SVC y Random Forest, los cuales aportan conocimiento a la organización sobre la intención de abandono y como puede ser solucionada. Siendo los mejores, la

clasificación multivariable y el jerárquico, ya que poseen un buen porcentaje de afinidad al resultado esperado. Específicamente una exactitud del 70% para el mejor de los modelos de clúster y el 81% de los predictivos (Random Forest), con lo cual se cumple el objetivo previsto.

- Evaluar el desempeño de los modelos construidos

  Se realiza un contraste tanto de las recomendaciones realizadas versus las reales contrastando un nivel de efectividad decente para su implementación en los procesos de marketing de la empresa, así como una caracterización de cada clúster creado, describiéndolos y titulando cada uno de los cuatro segmentos creados, ideando estrategias para su posterior despliegue e implementación en los procesos de AirProducts. Exponiendo cada una de sus fases a las personas clave en la organización y de esta manera generar valor.

.

## Conclusiones

En el contexto de Air Products, una empresa enfocada en la venta B2B, es crucial comprender el comportamiento y las necesidades de sus clientes corporativos. Este conocimiento permite a Air Products desarrollar estrategias de marketing y servicios que fortalezcan las relaciones comerciales a largo plazo.

A pesar de la disponibilidad de varias herramientas de Machine Learning en el mercado, el modelo propuesto en este trabajo se adapta de manera óptima a las necesidades específicas de Air Products en la actualidad. Esta capacidad de adaptación permite analizar de manera efectiva los datos relacionados con los clientes empresariales para identificar oportunidades de mejora y fortalecer las operaciones por medio de Clústerización, identificando características.

La retención y satisfacción de los clientes B2B son fundamentales para el éxito continuo de Air Products. Aunque la empresa no trabaja con un ciclo de vida del cliente tradicional ni un índice de retención como tal, se enfoca en comprender y atender las necesidades cambiantes de los clientes a lo largo de la relación comercial.

La evaluación técnica interna de Air Products ha identificado los atributos clave para implementar el modelo de análisis de datos. Estos atributos son fundamentales para comprender la dinámica de las relaciones comerciales y tomar decisiones informadas que beneficien tanto a Air Products como a sus clientes B2B.

Basándose en los resultados obtenidos y en el análisis de los perfiles de los clientes empresariales, Air Products dirige sus esfuerzos hacia la mejora continua de productos y servicios. Al mantener una comunicación abierta y una colaboración estrecha con los clientes, la empresa puede anticiparse a sus necesidades y ofrecer soluciones que agreguen valor a sus operaciones.

## Recomendaciones

Se presentan como una serie de aspectos que se podrían realizar en un futuro para emprender investigaciones similares o fortalecer la investigación realizada.

Se debe hacer una documentación exhaustiva de las causas de fuga de quienes abandonaron la compañía.

Se debe realizar un registro y seguimiento periódico a los contactos que se realizan al cliente y a la recepción de quejas y reclamos, ya que según estudios realizados, son de las principales causantes de la fuga.

Contar con información más completa de los clientes que permita asociar sus comportamientos de compra, con las características de este, ya que, entre más rasgos se conozcan, mejores modelos aplicados pueden tener Air Products para la prevención y análisis de este comportamiento negativo para la empresa.

## Referencias

La bibliografía es la relación de las fuentes documentales consultadas por el investigador para sustentar sus trabajos. Usar la última versión de normas APA. Mínimo 30 referencias bibliográficas


Baghla, S. & Gupta, G. (2022). *Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce*. Microprocessors and Microsystems. Volumen 94, 104680, ISSN 0141-9331. DOI: https://doi.org/10.1016/j.micpro.2022.104680. Disponible en https://www.sciencedirect.com/science/article/abs/pii/S0141933122002101

Caigny, A; Coussement, K; Verbeke, W; Idbenjra, K & Phan, M. (2021). *Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach.* Industrial Marketing Management. Volumen 99, Pages 28-39, ISSN 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2021.10.001. Disponible en https://www.sciencedirect.com/science/article/abs/pii/S0019850121001930

Cryogas (s.f.). Cryogas.com. Disponible en https://www.cryogas.com.co/web/co

Gattermann, T & Thonemann, U. (2022). *Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests.* Industrial Marketing Management. Volumen 107, Pages 134-147, ISSN 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2022.09.023. Disponible en https://www.sciencedirect.com/science/article/abs/pii/S0019850122002255

Miranda, J; Rey, P & Weber, R. (2005). *Predicción de Fugas de Clientes para una Institución Financiera Mediante Support Vector Machines.* Revista Ingeniería de Sistemas. Volumen XIX, pp. 49-68. Disponible en https://www.dii.uchile.cl/~ris/RISXIX/RISXIXpaper4.pdf

Pérez, P. A. (2014). *Modelo de predicción de fuga de clientes de telefonía móvil post pago.* Universidad de Chile, Santiago, Chile. Disponible en

https://repositorio.uchile.cl/bitstream/handle/2250/115942/cf-perez_pv.pdf?sequence=1

Troncoso, F. & Tapia, J. (2020). *Predicción de fuga de clientes en una empresa de distribución de gas natural mediante el uso de minería de datos.* Universidad Ciencia y Tecnología. Volumen 24. DOI: 10.47460/uct.v24i106.399. Disponible en https://www.researchgate.net/publication/346994600_PREDICCION_DE_FUGA_DE_CLIENTES_EN_UNA_EMPRESA_DE_DISTRIBUCION_DE_GAS_NATURAL_MEDIANTE_EL_USO_DE_MINERIA_DE_DATOS

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/https://doi.org/10.1016/j.jksuci.2019.12.011

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121449

Salim, A., Juliandry, Raymond, L., & Moniaga, J. V. (2023a). General pattern recognition using machine learning in the cloud. *Procedia Computer Science*, *216*, 565–570. https://doi.org/https://doi.org/10.1016/j.procs.2022.12.170

Salim, A., Juliandry, Raymond, L., & Moniaga, J. V. (2023b). General pattern recognition using machine learning in the cloud. *Procedia Computer Science*, *216*, 565–570. https://doi.org/https://doi.org/10.1016/j.procs.2022.12.170

Alpaydin, E. (2014). Introduction to Machine Learning. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). On Clústering Validation Techniques.