Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Optimización estratégica de captación de clientes para la Cámara de Comercio Colombo-británica.

Presentado por:

Ana María Pulido Carvajal, Chavelin Karina Riaño Cuervo, Yuli Natalia Suárez Díaz

Bogotá, D.C. 25 de mayo de 2025

Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Optimización estratégica de captación de clientes para la Cámara de Comercio Colombo-británica.

Presentado por:

Ana María Pulido Carvajal, Chavelin Karina Riaño Cuervo, Yuli Natalia Suárez Díaz

Bajo la dirección de:

Yudy Constanza Castaño Aristizábal

Bogotá, D.C. 25 de mayo de 2025

Contenido

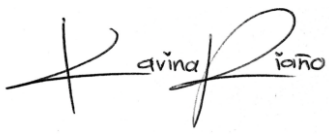
Preliminares

Declaración de Originalidad y Autonomía

Declaro(amos) bajo la gravedad del juramento, que he(mos) escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi(nuestra) propia cuenta y que, por lo tanto, su contenido es original.

Declaro(amos) que he(mos) indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

Ana María Pulido Carvajal



Chavelin Karina Riaño Cuervo

Yuli Natalia Suarez Diaz

Firmado en Bogotá, D.C. el 25 de mayo de 2025

Declaración de Exoneración de Responsabilidad

Declaro(amos) que la responsabilidad intelectual del presente trabajo es exclusivamente de su(s) autor(es). La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

Ana María Pulido Carvajal

Chavelin Karina Riaño Cuervo

Yuli Natalia Suarez Diaz

Firmado en Bogotá, D.C. el 25 de mayo de 2025

Lista de Figuras

Lista de Tablas

Abreviaturas

- **BritCham**: Cámara de Comercio Colombo-británica.
- **CRISP-DM:** Cross-Industry Standard Process for Data Mining, en español Proceso estándar intersectorial para minería de datos.
- **DOFA:** Debilidades, oportunidades, fortalezas y amenazas.
- **NLP**: Natural Language Processing (Procesamiento de leguaje natural).
- **WordCloud**: Nube de palabras.
- **AmCham:** Cámara de Comercio Colombo Americana.

Glosario

1. **Análisis competitivo del mercado:** Ayuda a comprender e identificar a las empresas del mismo sector que compiten por los posibles clientes según la línea de producto o servicio.
2. **Cámaras binacionales:** Son organizaciones privadas, no gubernamentales y sin ánimo de lucro conformadas por expertos o empresarios de diferentes sectores con el fin de facilitar las relaciones comerciales o negociaciones entre dos países, de manera que ambos resulten favorecidos y aprovechen oportunidades de negocio.
3. **Euro Cámaras:** Son asociaciones de cámaras de comercio binacionales europeas en diferentes países.
4. **Metodología CRISP-DM:** Son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos. Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
5. **Dashboard:** Herramienta de gestión de la información que monitoriza, analiza y muestra de manera visual los indicadores clave de desempeño (KPI), métricas y datos fundamentales para hacer un seguimiento del estado de una empresa, un departamento, una campaña o un proceso específico.
6. **Inteligencia empresarial (BI):** Inteligencia de negocios en español, se refiere a un conjunto de herramientas que permiten el análisis de datos para facilitar la toma de decisiones.
7. **Procesamiento de lenguaje natural:** Es una tecnología de inteligencia artificial (IA) que permite a las computadoras entender, interpretar y manipular el texto a analizar.

8. **Nube de palabras:** Es una representación visual de las palabras más repetidas en un texto o de las etiquetas de un sitio web.

9. **Proceso AS IS:** Análisis a la situación actual de la empresa en relación con el proceso que se quiere optimizar.

10. **Proceso TO BE:** Proyección de la situación futura del proceso, más específicamente a donde se quiere llegar del proceso con sus correspondientes mejoras.

11. **Legiscomex:** Es una plataforma digital que ofrece información sobre el comercio exterior en Colombia.

12. **Emis Next:** Es una plataforma integral centrada en ofrecer información confiable de mercados emergentes.

13. **Valor FOB:** Es el valor de una mercancía de exportación cuando se encuentra embarcada en el buque de transporte.

14. **Modelo de segmentación:** Es una técnica utilizada para dividir un grupo grande en subgrupos más pequeños, basados en características compartidas, con el objetivo de comprender mejor sus comportamientos y necesidades.

15. **Modelo de clasificación:** Es un algoritmo que predice a qué categoría o clase pertenece un nuevo dato, basado en datos de entrenamiento previamente categorizados.

16. **Normalización:** es una técnica que transforma los datos a un formato común para que puedan usarse en análisis y algoritmos de aprendizaje automático.

17. **Análisis de componentes principales:** es una técnica para el análisis exploratorio de datos. A menudo se utiliza para reducir la dimensionalidad del dataset para poder identificar entidades y patrones de los datos.

Anexos

1. Actas de reunión con BritCham (Archivo Word).
2. Análisis competitivo y DOFA (Archivo Excel).
3. Información entregada por la Cámara de Comercio Colombo- británica (PDF).
4. Lista de afiliados (PDF).
5. Precios de las Cámaras de Comercio (Archivo Excel).
6. Diccionario (Archivo Excel).
7. Cronograma (Archivo Excel).
8. Proceso AS IS BritCham (Miro https://miro.com/app/board/uXjVLNEysxI=/).
9. Proceso TO BE BritCham (Miro https://miro.com/app/board/uXjVLNEysxI=/).
10. Desarrollo de los modelos de segmentación y clasificación (Archivo ipynb).
11. Visualización por medio de Power BI.

Resumen Ejecutivo

Optimización estratégica de la captación de clientes para la Cámara de Comercio Colombo-británica.

La Cámara de Comercio Colombo-británica (BritCham) enfrenta un reto significativo en su estrategia de crecimiento: actualmente cuenta con 85 empresas afiliadas, mientras que sus principales competidores superan los 300 afiliados. Esta brecha limita su alcance y posicionamiento dentro del comercio binacional.

Con el objetivo de apoyar su expansión, se diseñó una iniciativa estratégica basada en analítica de datos. A través de la metodología CRISP-DM, se analizó información de financiera y de comercio exterior entre Europa y Colombia, correspondiente al periodo enero 2022 – mayo 2024, con el fin de identificar patrones, segmentar empresas y detectar aquellas con alto potencial de afiliación para BritCham.

Para comprender a fondo el negocio y su entorno competitivo, se desarrolló un análisis DOFA, complementado con un ejercicio de *benchmarking* potenciado por técnicas de Procesamiento de Lenguaje Natural (NLP), lo cual permitió una evaluación comparativa frente a sus principales competidores. Posteriormente, se integraron bases de datos del portal Legiscomex y Emis Next con información comercial y financiera de empresas exportadoras e importadoras, obteniendo así una base consolidada, limpia y enriquecida.

A partir de esta información, se aplicaron modelos de segmentación y clasificación mediante técnicas de Machine Learning, lo que permitió identificar empresas con alto potencial de afiliación. Además, se desarrolló un dashboard interactivo en Power BI, que facilita la visualización de sectores estratégicos y mejora la toma de decisiones comerciales basada en datos.

Esta solución transforma la manera en que BritCham podrá identificar y abordar posibles afiliados, dejando la dependencia de la búsqueda reactiva o limitada, y avanza a una estrategia proactiva, enfocada en los segmentos con mayor valor, fortaleciendo su red de afiliados, incrementando el impacto comercial y mejorando su posicionamiento en el ecosistema binacional.

Palabras clave

Analítica de datos, Cámara de Comercio Colombo-británica (BritCham), CRISP-DM, captación de clientes y segmentación de clientes, exportaciones e importaciones.

Abstract

Strategic Optimization of Client Acquisition for the British Colombian Chamber of Commerce.

The Colombian British Chamber of Commerce (BritCham) faces a significant challenge in its growth strategy: it currently has only 85 member companies, while its main competitors exceed 300 affiliates. This gap limits its reach and position within the binational trade landscape.

To support its expansion, a strategic data analytics initiative was designed. Using the CRISP-DM methodology, foreign trade and financial data between Europe and Colombia from January 2022 to May 2024 was analyzed to identify patterns, segment companies, and detect those with high potential for affiliation with BritCham.

To gain a deep understanding of the business and its competitive environment, a SWOT analysis was carried out, complemented by a benchmarking exercise enhanced by Natural Language Processing (NLP) techniques. This enabled a comparative evaluation of BritChams main competitors. Subsequently, databases from the Legiscomex portal were integrated with financial information from import and exporting companies, resulting in a consolidated, clean, and enriched dataset.

Based on this information, segmentation and classification models were applied using Machine Learning techniques, enabling the identification of companies with high affiliation potential. Additionally, an interactive Power BI dashboard was developed to visualize strategic sectors and enhance data-driven commercial decision-making.

This solution transforms the way BritCham identifies and approaches potential affiliate shifting from a reactive or limited search process to a proactive strategy focused on high-value

segments. It strengthens its affiliate network, increases commercial impact, and enhances its positioning within the binational ecosystem.

Keywords

Data analytics, British Colombian Chamber of Commerce (BritCham), CRISP-DM, customer acquisition and customer segmentation, exports and imports.

1. Introducción

La Cámara de Comercio Colombo-británica (BritCham) es una organización binacional que tiene como misión facilitar, promover y fortalecer las relaciones comerciales, de inversión y culturales entre el Reino Unido y Colombia. Entre sus principales funciones destacan la facilitación de negocios, la promoción del comercio bilateral y la prestación de servicios de apoyo a sus afiliados. Actualmente, BritCham cuenta con 85 afiliados, distribuidos en diversos sectores económicos, mayormente acumulados en servicios financieros, legales, educativos y empresariales, quienes pagan una suscripción anual que varía entre $2.300.000 COP y $13.750.000 COP, dependiendo del valor de sus activos (BritCham Colombia, 2025) (Anexo 3).

Con base en lo anterior, BritCham tiene el objetivo de ampliar su base de afiliados, fortalecer las relaciones con éstos y aumentar su reconocimiento en comparación con otras Cámaras binacionales. Por tal motivo, se propone el desarrollo e implementación de una herramienta analítica donde se pueda segmentar, clasificar y visualizar posibles clientes, lo que apalancará la creación de estrategias para la captación de nuevos afiliados por parte de la Cámara.

El comercio exterior entre Colombia y Europa se ha gestado por medio del acuerdo comercial, el cual data de más de 10 años (2013), define reglas claras en materia de comercio de bienes y servicio, lo que permite un mayor crecimiento económico y generación de empleos estables y remunerados, logrando así una relación preferencial y

permanente con actores clave en la economía mundial, representando múltiples ventajas para Colombia como: nuevas y mayores oportunidades de mercado, estableciendo nuevos vínculos en las cadenas de producción y suministro, y teniendo la posibilidad de establecer alianzas productivas y comerciales con diferentes países, ampliando la red de clientes y consumidores, con el fin de brindarle al consumidor colombiano mayores opciones para sus compras con mejores precios (Comercio, Industria y Turismo, 2012).

En cuanto a la relación entre el Reino Unido y Colombia, se firmó en 2019 el instrumento que preservará el marco comercial de relacionamiento que tiene actualmente en el acuerdo con la Unión Europea, garantizando que se mantengan las condiciones de integración luego de la salida de Reino Unido de la Unión Europea. Para Colombia es importante seguir manteniendo y ampliando las relaciones que hoy existen con el Reino Unido para el comercio de mercancías y los servicios, la inversión, las compras públicas y la propiedad intelectual (Comercio, Industria y Transporte, 2019).

Se propone trabajar con dicha información de comercio exterior (importaciones y exportaciones) entre Colombia y Europa, con enfoque en el Reino Unido, siguiendo la metodología CRISP-DM; para lo cual se realizará el análisis a detalle del comportamiento de importaciones y exportaciones por medio de diferentes factores involucrados como: país, departamento, aduana y lugar de ingreso o salida, así como el detalle de lo que compra o vende una empresa en el exterior, los fletes que pagó, los tributos aduaneros que canceló,

entre muchos más datos de la operación comercial (Legiscomex- Estadísticas de comercio exterior, 2024).

La metodología CRISP- DM es ampliamente reconocida por el orden sistemático en el análisis de datos, caracterizada por su fácil entendimiento, flexibilidad y personalización. Según Chapman et al. (2000), la metodología es una guía de desarrollo compuesto por seis fases interrelacionadas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Gracias a su carácter iterativo, garantiza que el abordaje en los problemas de negocio sea efectivo utilizando datos relevantes y modelos precisos.

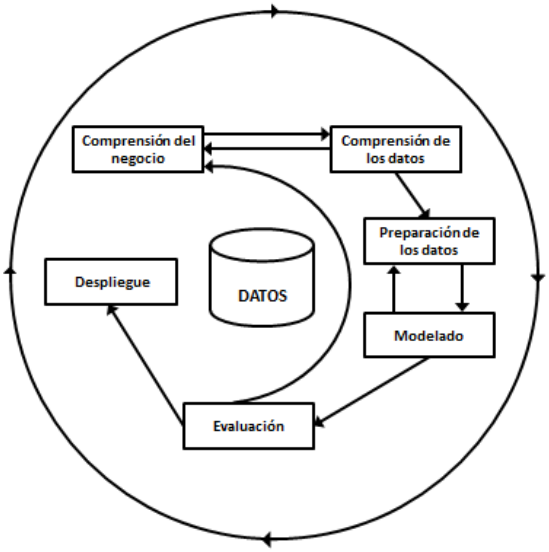


Figura 1 Componentes de la metodología CRISP- DM

1. **Comprensión del negocio:** Recopilación de la situación y procesos internos de la cámara, identificación de necesidades y objetivos de negocio planteados por BritCham mediante el análisis FODA, compilación del comportamiento del mercado binacional y análisis de benchmarking potenciado por Procesamiento de Lenguaje Natural (NLP).

2. **Comprensión de los datos:** Identificación de las fuentes de datos para llevar a cabo el análisis de bases de comercio exterior (importaciones y exportaciones) y estados financieros, extraídos de las plataformas Legiscomex y Emis Next, y de estas realizar el proceso de entendimiento de los datos de manera descriptiva, frecuencias, correlaciones y variabilidad y relaciones entre variables.

3. **Preparación de los datos:** Limpieza, análisis, construcción de nuevos datos e integración en una base final enriquecida con base en la información original de Legiscomex y Emis Next.

4. **Modelado:** Construcción de 2 modelos complementarios

Modelo de segmentación: para identificar patrones comerciales y económicos entre las empresas analizadas y agruparlas según sus variables más fuertes.

Modelo de clasificación: con el objetivo de predecir el grupo o segmento al que pertenece una empresa futura considerando las variables predictoras del modelo.

5. **Evaluación:** Revisión de los diferentes modelos empleados, evaluación de los resultados y determinación del modelo apto para el objetivo de negocio.

6. **Despliegue:** Implementación y validación de las herramientas desarrolladas para asegurar su utilidad en la toma de decisiones estratégicas. Adicionalmente, la creación de un tablero de visualización en Power BI con la información más relevante obtenida luego del análisis realizado.

La completitud satisfactoria de las 6 fases da como resultado el cumplimiento del objetivo por el cual se estableció este proyecto, proporcionando a BritCham herramientas que le permitan tener una ventaja competitiva frente a las otras Cámaras binacionales, incrementando sus afiliados y fortaleciendo su posición en el sector. Este enfoque integrado

contribuirá al crecimiento de la organización y sentará las bases para estrategias de expansión sostenibles en el futuro.

2. Objetivos

2.1. Objetivo General

Desarrollar una solución de analítica de negocio basada en la metodología CRISP-DM que permita identificar patrones comerciales y económicos en los datos de exportación e importación entre Europa y Colombia, con el fin de optimizar la identificación y captación estratégica de clientes potenciales para BritCham, considerando que al menos el 5 % de estos clientes presenta un alto potencial de afiliación.

2.2. Objetivos Específicos

- Analizar el posicionamiento competitivo de BritCham mediante técnicas de procesamiento de lenguaje natural (NLP), identificando oportunidades de crecimiento a través de un análisis comparativo (benchmark) frente a sus principales competidores.
- Implementar modelos de segmentación y clasificación que permitan agrupar empresas con características económicas y comerciales similares, identificar segmentos estratégicos y predecir la pertenencia de nuevas empresas a dichos segmentos.
- Desarrollar un tablero interactivo que consolide los principales indicadores, visualice los resultados de los modelos analíticos y facilite la toma de decisiones estratégicas basadas en los datos procesados para BritCham.

- Diseñar estrategias para cada segmento identificado, con el fin de proporcionar a BritCham lineamientos claros sobre las acciones recomendadas para captar posibles empresas afiliadas.

3. Alcance

El presente proyecto contempla el diseño, desarrollo y despliegue de una solución de analítica de negocio basada en la metodología CRISP-DM, enfocada en identificar patrones comerciales y económicos a partir de datos de exportación e importación entre Europa y Colombia, en el período comprendido entre enero de 2022 y junio de 2024. En este marco se desarrollarán los siguientes componentes:

Benchmarking mediante técnicas de procesamiento de lenguaje natural (NLP), con el fin de comparar el posicionamiento de BritCham frente a otras cámaras y Cámaras binacionales relevantes, identificando oportunidades de crecimiento.

Un modelo de clusterización para segmentar empresas con características económicas y comerciales similares, facilitando la identificación de grupos estratégicos, los cuales tendrán estrategias para cada segmento identificado, proponiendo líneas de acción específicas para facilitar la afiliación de empresas relevantes.

Un modelo de clasificación para estimar el segmento de nuevas empresas identificadas por BritCham como afiliados potenciales, basándose en variables clave del comercio exterior y contexto económico, en conjunto con tableros dinámicos en Power BI que consoliden los datos clave del análisis, incluyendo todo el proceso realizado, los cuales

serán diseñados para facilitar la toma de decisiones con relación a nuevos afiliados a BritCham.

Los componentes mencionados están orientados a mejorar la operación en cuanto a la identificación y captación de clientes potenciales para la Cámara, con soportes cuantitativos y visualizaciones interactivas durante el proceso de toma de decisiones.

3.1. Entregables del Proyecto

Se realizará los siguientes entregables alineados con los objetivos específicos del proyecto, considerando los mínimos criterios de aceptación como: funcionalidad, accesibilidad, compatibilidad y claridad.

Resultados sobre el Procesamiento de Lenguaje Natural, con sus principales recomendaciones estratégicas para las oportunidades observadas en el benchmarking.

Power BI integrado con los modelos de segmentación y clasificación realizados, donde se identifica la arquitectura de los datos y de igual forma, las visualizaciones necesarias por medio de filtros, gráficos e insights del modelo de clasificación.

3.2. Beneficios Esperados

Al finalizar el proyecto, se espera que BritCham obtenga beneficios tangibles en sus operaciones y estrategias de captación de clientes, incluyendo:

1. **Incremento de afiliados**: Gracias al uso del modelo de clasificación, se optimizarán los tiempos y esfuerzos necesarios para captar nuevos clientes potenciales, lo que podría derivar en un aumento futuro de afiliados a la cámara.

2. **Ampliación de servicios**: El análisis estratégico de los tableros permitirá identificar patrones de comportamiento, preferencias y necesidades de los clientes, facilitando el diseño de estrategias efectivas para aumentar las afiliaciones y diversificar los servicios ofrecidos.

Finalmente, el proyecto sentará las bases para la implementación de futuros KPIs, alineados con los objetivos de crecimiento y eficiencia de BritCham, fortaleciendo su capacidad analítica y su enfoque centrado en los datos.

4. Metodología

El proyecto se desarrolló siguiendo la metodología CRISP-DM, siendo una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos, ya que permite abordar problemas de negocio de manera estructurada, con su primera puesta en marcha en 1997 por el Programa de Investigación y Desarrollo en Tecnologías de

Información de la Unión Europea. Hoy en día es principalmente expuesta y recomendada por la compañía de tecnología IBM (Espinosa-Zúñiga, 2020), compuesta por 6 etapas:

4.1. Fase 1 Comprensión del Negocio

Es considerada la etapa más importante, ya que en ella se colabora estrechamente con los interesados del proyecto y stakeholders para definir los objetivos, comprender los requisitos comerciales y los problemas específicos del negocio. El propósito es captar adecuadamente las necesidades del negocio, ya que un mal entendimiento en esta etapa afectaría significativamente las siguientes fases del proyecto.

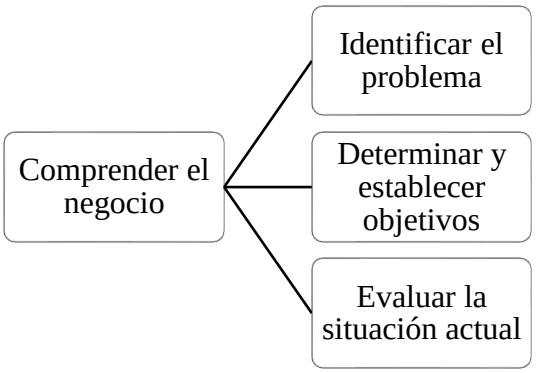


Figura 2 Comprensión del negocio

4.2. Fase 2 Comprensión de los Datos

En esta etapa se exploran y evalúan los datos, analizando su calidad, integridad, relevancia y disponibilidad. Este proceso se apoya en el análisis estadístico exploratorio, que permite observar el comportamiento de los datos, comprender su estructura y conocer las características que los componen. Todo esto contribuye a obtener una visión general de

lo que se puede lograr con la información disponible. Esta fase complementa el trabajo realizado en la etapa anterior, mediante un análisis guiado por el conocimiento del negocio previamente adquirido.

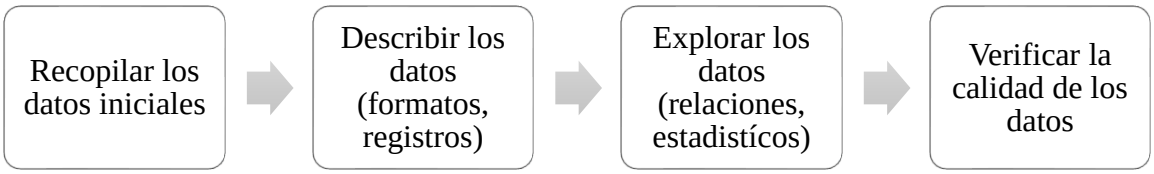


Figura 3 Comprensión de los datos

4.3. Fase 3 Preparación de los Datos

Esta fase, consiste en limpiar, integrar y transformar la información en un formato adecuado para el modelado, considerando retos que se pueden encontrar en los datos tales como: faltantes, duplicados, outliers e inconsistencias, los cuales se pueden manejar por medio de la aplicación de técnicas de preprocesamiento (Astera, 2025):

Limpieza de datos: proceso fundamental del preprocesamiento de datos, ya que permite eliminar errores, imputar valores faltantes y rectificar inconsistencias.

Uniformidad de los datos o normalización: donde las medidas en diferentes escalas se ajustan a una más uniforme, permitiendo las comparaciones sin sesgos o erróneas.

Enriquecimiento de datos: con el fin de mejorar los datos con fuentes adicionales o atributos derivados, que puede proporcionar más profundidad o contexto.

Imputación de datos: los datos faltantes pueden distorsionar el análisis y dar lugar a modelos inexactos, para manejar los valores faltantes se puede aplicar la imputación, lo cual significa completar los valores faltantes con medidas estadísticas como la media o la mediana.

Reducción de dimensionalidad: donde se busca reducir las variables consideradas, simplificando el modelo sin perder información significativa, este método puede mejorar el rendimiento del modelo y reducir la complejidad.

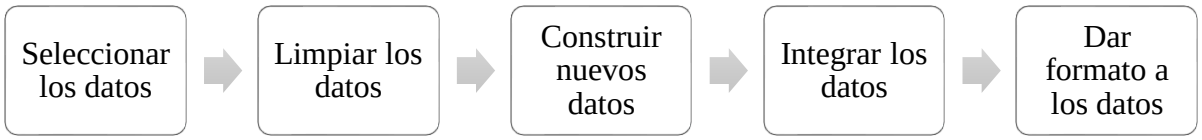


Figura 4 Preparación de los datos

4.4. Fase 4 Modelado

Luego de la preparación de los datos, en esta fase es donde los resultados empiezan a dar una salida sobre el problema planteado en la primera etapa de comprensión del negocio. Se ejecutan varios modelos utilizando los parámetros predeterminados y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias por el modelo.

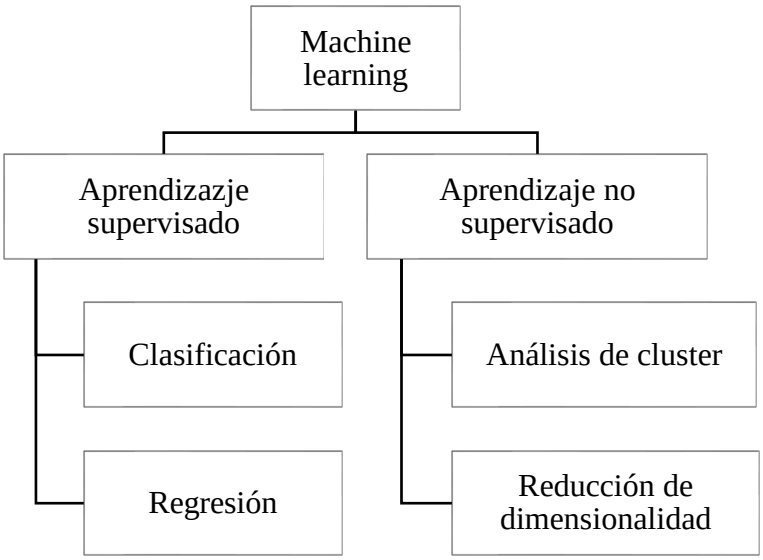


Figura 5 Modelado

Modelos supervisados: son modelos donde cada dato a analizar está etiquetado y asociado a una categoría. Se usan técnicas de clasificación para desarrollar modelos predictivos, donde se hace uso de un conjunto datos de entrenamiento para enseñar a los modelos a producir el resultado deseado, puede dividirse en dos tipos de problemas: clasificación y regresión.

La clasificación utiliza un algoritmo para asignar con precisión los datos de prueba en categorías específicas, algunos ejemplos de estos modelos son clasificadores lineales, las máquinas de vectores de soporte (SVM), los árboles de decisión, el método K de los vecinos más próximos y el random forest.

La regresión se utiliza para comprender la relación entre variables dependientes e independientes y comúnmente para hacer proyecciones. La regresión lineal, la regresión logística y la regresión polinomial son de algoritmos más conocidos.

Estos modelos enfrentan algunos retos en cuanto a su construcción, ya que requieren de cierto nivel de experiencia y su entrenamiento puede tomar mucho más tiempo, los datos pueden tener una mayor probabilidad de error humano, generando algoritmos con un aprendizaje incorrecto y no pueden agrupar ni clasificar los datos por si solos, en comparación con los modelos no supervisados.

Modelos no supervisados: en este caso los datos de estudio no son etiquetados y el objetivo es estudiar muchos datos complejos para hacerlos más simples, se rigen por 3 tareas principales: agrupamiento, asociación y reducción de dimensionalidad, su enfoque está relacionado con descubrir patrones ocultos o agrupaciones de datos.

La agrupación es una técnica que agrupa datos sin etiquetar en función de sus similitudes o diferencias, estos algoritmos de agrupación en clústeres se pueden clasificar en 3:

1. Agrupación exclusiva y superpuesta: estipula que un punto de datos solo puede existir en un clúster. Esto también se puede denominar agrupamiento "duro", por ejemplo, el algoritmo K-means realiza agrupación exclusiva. Los clústeres superpuestos difieren de los clústeres exclusivos en que permiten que los puntos de datos pertenezcan a varios clústeres con grados de membresía distintos.

Agrupación jerárquica: conocida como análisis de agrupamiento jerárquico, que puede categorizar de dos formas: pueden ser aglomerados o divisivos:

La **agrupación aglomerativa:** es cuando los puntos de datos se aíslan inicialmente como agrupaciones separadas y luego se fusionan de forma iterativa según la similitud hasta que se logra crear un grupo.

La **agrupación divisiva** es lo opuesto a la aglomerativa, un solo clúster de datos se divide en función de las diferencias entre los puntos de datos, sin embargo, no se utiliza comúnmente.

Agrupación probabilística: ayuda a resolver problemas de estimación de densidad o de agrupamiento, los puntos de datos se agrupan en función de la probabilidad de que pertenezcan a una distribución particular. El modelo de mezcla gaussiana (GMM) es uno de los métodos de agrupación probabilística más utilizados.

2. Reglas de asociación: es un método basado en reglas para encontrar relaciones entre variables en un conjunto de datos determinado.

3. Reducción de dimensionalidad: es una técnica que se utiliza cuando el número de características o dimensiones de un conjunto de datos determinado es demasiado alto. Reduce la cantidad de entradas de datos a un tamaño manejable y al mismo tiempo preserva la integridad del conjunto de datos tanto como sea posible. Se usa comúnmente en la etapa de preprocesamiento de datos, y se pueden usar algunos métodos diferentes de reducción de

dimensionalidad, como: análisis de componentes principales, descomposición en valores singulares y codificadores automáticos.

4.5. Fase 5 Evaluación del Modelo

La fase de Evaluación implica una revisión crítica del modelo desarrollado y la interpretación de los resultados obtenidos, se determina también la calidad del modelo con base en el análisis de ciertas métricas estadísticas del mismo, contrastando esto con las opiniones de los stakeholders o expertos en el negocio.

De acuerdo con los resultados de esta etapa se determina seguir con la última fase de la metodología, regresar a alguna de las etapas anteriores o partir de cero con un nuevo.

4.6. Fase 6 Despliegue o Implementación

Esta fase involucra la implementación de soluciones basadas en datos en entornos operativos del negocio, garantizando la implementación exitosa y el monitoreo continuo del sistema desplegado.

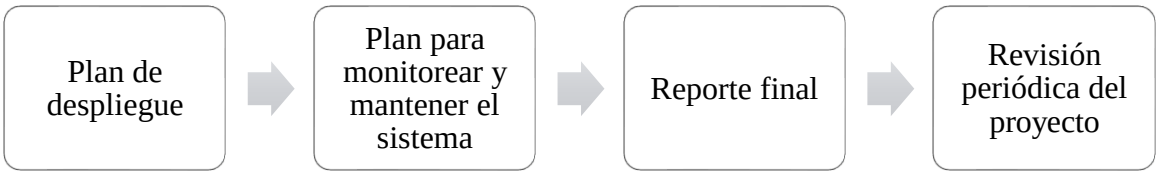


Figura 6 Despliegue

5. Cronograma

Se dio inicio al proyecto con reuniones de entendimiento de negocio desde el mes de junio de 2024, teniendo en cuenta las fases mencionadas de la metodología CRISP-DM y establecimiento de hitos para el desarrollo de cada una. La finalización del proyecto está planificada para mediados de mayo 2025.

Planificación y comprensión del negocio: Establecer y entender las necesidades del usuario por medio de la definición clara de objetivos y la aceptación de estos durante esta fase, donde se entregue el primer documento de avance.

Comprensión de datos: Identificar las fuentes internas y externas para hacer una revisión preliminar de las bases y evaluar la integridad de los datos y su relevancia dentro del proyecto.

Preparación de datos: Lograr una base enriquecida de datos que permita el análisis integral de la información.

Modelado: Evaluar y seleccionar una técnica de modelado de segmentación y clasificación.

Evaluación y ajuste: Comparar los modelos propuestos, escogiendo el que mejor se destaque en su ejecución con base a los criterios de evaluación de cada uno.

Despliegue: Implementar el Power BI con la información final de las fases anteriores, considerando los modelos de segmentación y clasificación dentro de la visualización.

Tabla 1 Cronograma del proyecto

Tarea	Mes							
	1	2	3	5	6	7	8	9
Reunión inicial para definición con el equipo	X							
Establecer metas claras para el sistema de análisis de datos	X							
Análisis de flujos de trabajo actuales	X	X						
Identificación de cuellos de botella y áreas de mejora	X	X						
Identificar y clasificar las fuentes de datos internas			X	X				
Identificar fuentes de datos externas necesarias			X	X				
Revisión preliminar para evaluar la calidad de datos					X			
Verificar la relevancia de los datos con respecto a los objetivos del proyecto					X			
Eliminación de duplicados, corrección de errores.					X			
Normalización de formatos de datos para análisis consistente					X			
Integrar datos de múltiples fuentes en un único repositorio						X		
Selección de técnicas de modelado						X		
Construcción de modelos de analítica						X		
Definición de variables clave y parámetros del modelo						X		
Pruebas de los modelos							X	
Optimización de modelos basada en resultados							X	
Evaluación de modelos en condiciones de test							X	
Ajustes basados en feedback y pruebas adicionales								X
Despliegue final y configuración del entorno de producción								X

6. Descripción de las Fuentes de Datos

6.1. Fuentes de Datos Externas

Una de las fuentes de datos utilizadas en proyecto proviene de Legiscomex, una plataforma reconocida por su especialización en información de comercio internacional que

facilita la gestión y el análisis del comercio exterior, proporcionando información clave para optimizar la toma de decisiones estratégicas en los negocios internacionales. (Legiscomex,2024.).

Se proporcionan bases de datos que incluyen detalles de exportaciones e importaciones entre Colombia y Europa, desde enero de 2022 hasta mayo de 2024.

La base de datos de Importaciones cuenta con 31 variables que detallan las transacciones de productos importados a Colombia desde Europa, cubriendo información clave sobre el origen y destino de las mercancías, como año, mes, aduanas de entrada y destino, y productos importados (clasificación arancelaria y descripción).

Asimismo, proporciona datos económicos como el valor FOB (Free on board) y cantidad de productos importados, también, detalles sobre los actores involucrados, como el importador.

Tabla 2 Extracto diccionario de datos tabla importaciones

Variable	Tipo de dato	Descripción
Razón Social del Importador	Texto	Nombre legal de la empresa importadora.
Cantidad(es)	Numérico	Número de unidades o volumen total de la mercancía que se importa.
País de compra	Texto	País donde se realizó la transacción comercial para la compra de la mercancía.
Departamento Destino	Texto	Departamento de Colombia al que está destinada la mercancía importada.
Municipio	Texto	Localidad específica dentro del departamento de destino de la mercancía.

Valor FOB (USD)	Numérico	Valor total de la mercancía en términos FOB (Libre a Bordo) expresado en dólares estadounidenses.

La base de datos de exportaciones incluye 68 variables que detallan las transacciones de productos exportados desde Colombia hacia Europa. Las variables cubren aspectos clave como la información temporal (año, mes, día) y geográfica (municipio de origen, aduana de salida).

También incluye detalles sobre los productos exportados, tales como el capítulo del arancel y la descripción del producto. En cuanto a los aspectos económicos, se registran datos como el valor FOB (valor de la mercancía sin incluir el costo de transporte o seguro) y la cantidad (peso bruto y neto). Adicionalmente, proporciona información sobre los actores involucrados, como el exportador, así como el modo de transporte utilizado para la exportación y el país de destino.

Tabla 3 Extracto diccionario de datos tabla exportaciones

Variable	Tipo de dato	Descripción
NIT del exportador	Identificador	Número de Identificación Tributaria del exportador, que lo identifica ante las autoridades fiscales y aduaneras.
Razón social actual Exportador	Texto	Nombre legal de la empresa importadora o exportadora.
Moneda de negociación	Texto	Moneda en la que se realizó la transacción comercial.
Valor FOB (USD)	Numérico	Valor total de la mercancía en términos FOB expresado en dólares estadounidenses.

Valor FOB (COP)	Numérico	Valor total de la mercancía en términos FOB expresado en pesos colombianos.
Continente Destino	Texto	Continente al que se dirige la mercancía exportada.

Por otro lado, también se extrajeron datos de información financiera relevante de empresas obtenida por medio de la plataforma Emis Next, la cuál se encarga de ofrecer información confiable sobre el mercado de cada país, permitiendo proporcionar la claridad necesaria sobre la toma de decisiones, inversiones y riesgos de manera informada.

Los estados financieros obtenidos contienen información de activos, patrimonio y pasivos totales con corte al año fiscal auditado, y variables que permiten conocer a la empresa como NIT, país, ciudad, contacto, número de empleados y fecha de incorporación. De tal forma que no solo se establece como relevante el comportamiento comercial, sino su desempeño económico según su estado operacional.

Tabla 4 Extracto diccionario de datos tabla estados financieros

Variable	Tipo de dato	Descripción
NIT	Identificador	Número de Identificación Tributaria de la empresa
País	Texto	País donde se encuentra la empresa
Sector	Texto	Sector económico al que pertenece la empresa
Activos totales	Numérico	Total de activos por empresa considerando el último año fiscal registrado
Patrimonio total	Numérico	Total de patrimonio por empresa considerando el último año fiscal registrado
Pasivos totales	Numérico	Total de pasivos por empresa considerando el último año fiscal registrado

6.2. Fuentes de Datos Internas

Así mismo, BritCham ofrece información clave, incluyendo una descripción detallada de su modelo de negocio con un listado de los 85 afiliados considerando información del sector económico al que pertenece, NIT y para funcionalidad del proyecto si se realiza alguna de las actividades de comercio exterior analizadas.

Tabla 5 Extracto diccionario de datos tabla de afiliados BritCham

Variable	Tipo de dato	Descripción
NIT	Identificador	Número de Identificación Tributaria de la empresa
Nombre de la empresa	Texto	Nombre de identificación de la empresa
Sector económico	Texto	Sector económico al que pertenece la empresa
Exportaciones	Texto	Países a los que exporta
Importaciones	Texto	Países de donde importa

7. Descripción de la Situación Organizacional

La Cámara de Comercio Colombo-británica (BritCham) tiene como propósito fortalecer los lazos comerciales, gubernamentales y culturales entre Colombia y el Reino Unido. Cuenta con una experiencia especializada y un enfoque innovador ofreciendi una red de contactos extensa, respaldada por el apoyo institucional de entidades como la Embajada Británica en Colombia y el Departamento de Comercio Internacional, lo que facilita la creación de oportunidades de negocios y la expansión de operaciones entre ambos países. (BritCham, 2024).

Los miembros de BritCham acceden a servicios personalizados, que incluyen estudios de mercado, presentaciones de alto nivel y estrategias de marketing para fortalecer las marcas británicas en la región. Además, tienen la oportunidad de participar en eventos exclusivos de networking, seminarios y otros espacios de discusión empresarial. (BritCham, 2024).

Gracias a la apertura a nuevos negocios en ambos países y la estrecha relación comercial y diplomática, BritCham se posiciona como un enlace confiable para empresas que buscan aprovechar las oportunidades de los mercados colombiano y británico, proporcionando orientación y acompañamiento estratégico (BritCham Colombia, 2025).

BritCham opera en un entorno competitivo, donde entidades similares, como Eurocamaras y AmCham Colombia, compiten por atraer y retener empresas afiliadas mediante servicios de networking, promoción comercial y desarrollo empresarial. Además, enfrenta una competencia indirecta de instituciones como ProColombia, Enterprise Europe Network, La Embajada Británica en Bogotá y Cámara de Comercio Colombo-británica en Londres que ofrecen servicios de apoyo empresarial y promoción de exportaciones (BritCham Colombia, 2025).

En cuanto a los servicios que ofrece, BritCham proporciona una amplia gama de servicios para sus afiliados (BritCham Colombia, 2025), entre los que destacan:

- **Networking empresarial:** Creación de conexiones estratégicas entre empresas afiliadas y socios potenciales.

- **Promoción comercial:** Organización de eventos, ferias y misiones comerciales que facilitan el acceso a nuevos mercados.
- **Capacitación empresarial:** Programas como BritCham Academy, diseñados para mejorar las competencias empresariales de los afiliados.
- **Eventos emblemáticos:** Actividades como el Torneo de Golf BritCham y los premios Lazos a la Sostenibilidad, que fortalecen la relación con los afiliados y promueven la visibilidad de sus logros.

BritCham cuenta con 85 empresas afiliadas, divididas en múltiples sectores económicos, en las cuales los sectores que tienen un amplio porcentaje se encuentran Banca, servicios jurídicos, servicios empresariales y bebidas y alimentos, el sector que presenta menos porcentaje corresponde a Turismo, con membresías anuales. Dichos afiliados pagan una membresía por el uso de los servicios que BritCham ofrece, varían entre $2.410.000 y $14.500.000 COP dependiendo del tamaño de la empresa en relación con el valor de los activos (BritCham Colombia, 2025).

BritCham opera en el sector del comercio internacional, específicamente facilitando conexiones comerciales y empresariales bilaterales. Este sector es altamente dinámico y depende de la capacidad de las organizaciones para adaptarse a cambios en las políticas comerciales, macroeconomía global y necesidades específicas de los sectores. La misión de BritCham en este contexto es potenciar el comercio bilateral mediante el apoyo a sus afiliados y la generación de oportunidades de negocios con valor agregado.

Según el Departamento Administrativo Nacional de Estadística (2024), en octubre de 2024 las exportaciones del país fueron US$4.311,8 millones FOB (Free On Board) y presentaron un crecimiento de 3,8% en relación con octubre de 2023; este resultado se debió principalmente al crecimiento de 34,8% en las ventas externas del grupo de Agropecuarios, alimentos y bebidas (DANE, 2024), y en relación a las importaciones fueron US$5.162,8 millones CIF (Cost, Insurance, and Freight) y presentaron un crecimiento de 4,4% con relación al mismo mes de 2023 (DANE, 2024). Este comportamiento obedeció principalmente al aumento de 5,0% en el grupo de Manufactura.

Con un enfoque en la optimización de la afiliación y retención de clientes, BritCham busca implementar soluciones basadas en análisis de datos y estrategias personalizadas que maximicen su impacto en el comercio bilateral. Su compromiso es garantizar el crecimiento sostenible de la red empresarial, posicionándose como un actor clave en el comercio internacional entre Colombia y Reino Unido.

El desarrollo del proyecto se fundamenta en principios de inteligencia de negocios (BI) y análisis competitivo, a través del uso de herramientas analíticas, donde se busca identificar sectores clave y patrones de comportamiento empresarial que permitan dar una solución insumo para que la empresa pueda diseñar estrategias efectivas para la afiliación y retención de clientes.

7.1. Análisis del Proceso "AS IS" y "TO BE" en BritCham:

Este análisis muestra la optimización del proceso una vez el proyecto se encuentre funcionando dentro de la entidad, generando mayor eficiencia en la captación de clientes para BritCham (Angeli, 2018).

7.1.1. Análisis del Proceso "AS IS" de BritCham

1. **Objetivo:** Mejorar la adquisición de nuevos clientes potenciales, basado en procesos manuales y herramientas limitadas.
2. **Proceso:** Los flujos actuales dependen de herramientas manuales y lineamientos básicos, lo que genera mayor tiempo y esfuerzo operativo para lograr la captación de nuevos clientes.
3. **Aplicaciones:** Se emplean herramientas ofimáticas y One Drive, lo cual limita la integración de datos y procesos más avanzados.
4. **Datos:** Se manejan datos básicos de los afiliados, como el nombre del representante, correo electrónico, dirección y tarifa de afiliación, lo que dificulta la personalización de estrategias y segmentación eficiente.

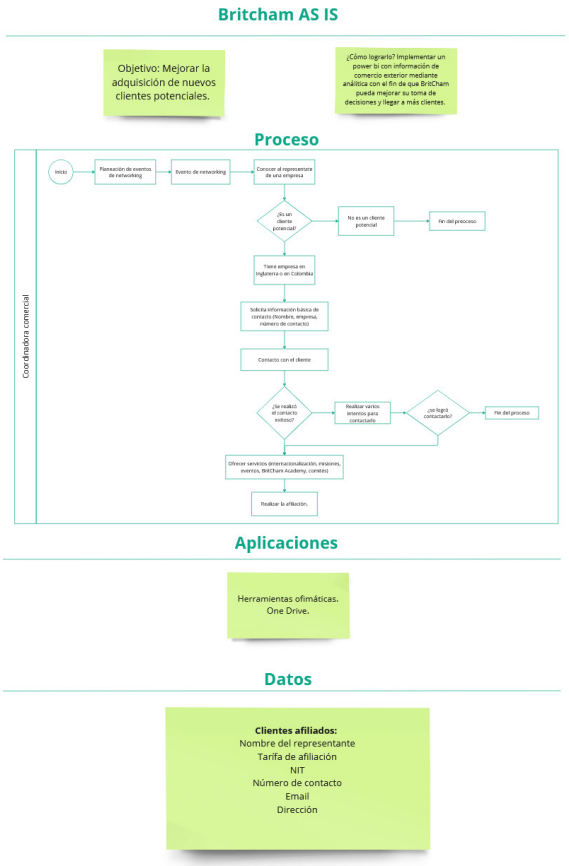


Figura 7 Proceso AS IS BritCham (Visualización más detallada Anexo 8)

7.1.2. Análisis del Proceso TO BE de BritCham

1. Objetivo: Optimizar la identificación y captación de nuevos clientes potenciales mediante la implementación de tecnología avanzada y el aprovechamiento de bases de datos robustas, garantizando estrategias más eficientes y personalizadas.

2. Proceso: Implementación de flujos de trabajo optimizados con actividades claras y procesos más eficientes, lo que reduce la carga operativa y mejora los resultados.

3. Aplicaciones: Incorporación de herramientas avanzadas como Power BI, Python y bases de datos robustas, permitiendo un manejo más eficiente de información y análisis.

4. Datos: Uso de bases de datos sobre históricos de exportaciones e importaciones para permitir a la entidad realizar estrategias más personalizadas y segmentadas.

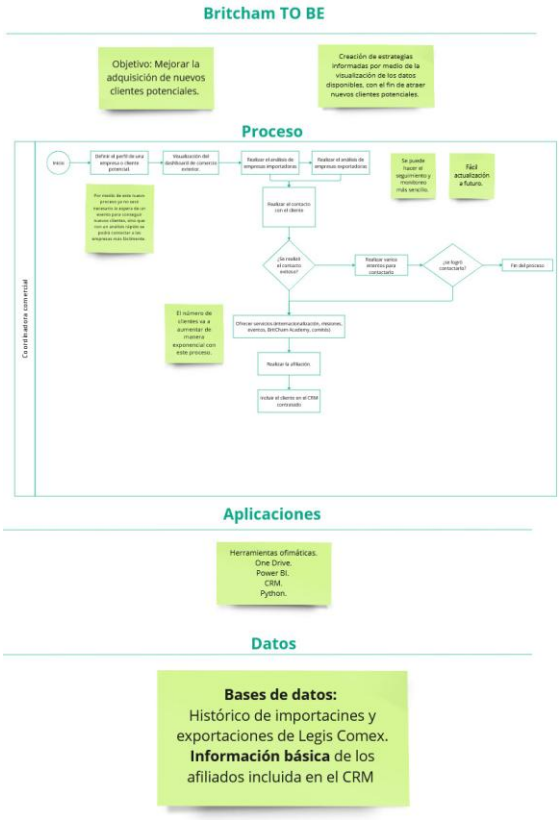


Figura 8 Proceso TO BE BritCham (Visualización más detallada Anexo 9)

En conclusión, se espera que BritCham presente un crecimiento en relación con la modernización y la eficacia, potenciando la capacidad de captar nuevos clientes y fortalecer su red comercial. Este cambio no solo responde a las necesidades actuales, sino que también

prepara a la organización para enfrentar desafíos futuros con una base tecnológica sólida y datos estratégicos.

8. Descripción de la Situación Estudio de Caso y/o Problemática Empresarial

BritCham ofrece servicios a sus empresas afiliadas, sin embargo, enfrenta un desafío crítico relacionado con su posicionamiento frente a las principales euro cámaras. A pesar de contar con experiencia y una oferta de servicios de calidad, la cámara presenta un número de afiliados considerablemente inferior al de sus de benchmarking, que muestra las siguientes cifras de afiliados en algunas de las principales euro cámaras:

Tabla 6 Afiliados de la competencia

Euro Cámara	Número de Afiliados
Suiza	50
Reino Unido	85
España	150
Alemania	278
Francia	292
Holanda	300
Italia	338
Estados Unidos	950

Eurocámaras como Francia (292 afiliados), Alemania (278 afiliados) y Holanda (300 afiliados) tienen una ventaja considerable en términos de captación de afiliados. Esto indica que hay una fuerte competencia por captar empresas y miembros dentro del mismo segmento de mercado.

Las Cámaras con mayor número de afiliados, como Estados Unidos, Italia, Francia y Alemania, pueden aprovechar economías de escala, tienen más recursos para ofrecer una gama más amplia de servicios, lo que les permite atraer y retener un mayor número de afiliados, según el análisis de mercado realizado.

En mercados altamente competitivos, como las Cámaras de Comercio, la visibilidad es clave, ya que las cámaras más grandes tienen mayor presencia mediática, lo que refuerza su posición y visibilidad ante los empresarios.

Por lo anterior se pone en evidencia la necesidad urgente de que BritCham desarrolle una estrategia de posicionamiento más robusta, que le permita diferenciarse y captar nuevos clientes de manera más efectiva. La falta de un proceso de toma de decisiones estratégico e informado está limitando el crecimiento de su portafolio de empresas afiliadas y dificultando la mejora de sus vínculos, lo que impacta negativamente en su operación general.

Si este problema no se aborda de manera inmediata, podrían surgir consecuencias adicionales, como una mayor pérdida de competitividad frente a otro euro cámaras, la estancación en la innovación de los servicios ofrecidos, una disminución en los ingresos por afiliaciones y la posible pérdida de fidelidad de los clientes actuales. Estos problemas no solo tendrían repercusiones operacionales, sino que también afectarían la reputación y estabilidad financiera de la cámara a largo plazo, creando desafíos más complejos y difíciles de resolver en el futuro.

Para enfrentar este desafío, la solución propuesta en este proyecto consiste en implementar tableros de inteligencia de negocios y un modelo de clasificación de empresas potenciales para afiliarse. Estas herramientas permitirán una toma de decisiones informada y estratégica, optimizando la captación de afiliados y mejorando la personalización de los servicios ofrecidos, lo que facilitará un posicionamiento fuerte y sostenible en el mercado.

9. Descripción de las Acciones que se Toman en el Análisis de la Solución a la Problemática

9.1. Entendimiento del Negocio

Para abordar los desafíos identificados en BritCham, se implementó la metodología CRISP-DM para el desarrollo del proyecto. Para el entendimiento del negocio, se tuvo reuniones para conocer la operación de la Cámara y se realizó un análisis DOFA con los interesados del proyecto, considerando dentro de sus debilidades la limitación en filtros de afiliación y recursos financieros, mientras que en oportunidades se destacaron por el reconocimiento, la llegada del bicentenario del Reino Unido y la expansión de marca. En cuanto a las fortalezas se encontró la relación cercana con la Embajada Británica y la sólida reputación, dejando por último como amenaza la competencia con cámaras mejor financiadas.

Posteriormente, se llevó a cabo el mismo análisis para los competidores directos e indirectos, identificando fortalezas como el reconocimiento internacional, el respaldo gubernamental y alianzas estratégicas sólidas. En las oportunidades destacaron las áreas de

innovación tecnológica, expansión digital y el aprovechamiento de tratados comerciales. Sin embargo, se encuentran como amenazas la inestabilidad económica, la competencia global y una dependencia financiera pública, que se relaciona con debilidades como la adaptación limitada a cambios regulatorios y recursos insuficientes.

Dado que el análisis DOFA se realizó inicialmente con información proporcionada por el equipo de BritCham, se decidió adicionar al estudio el uso de una metodología analítica basada en datos. Para ello, se empleó un modelo de procesamiento de lenguaje natural (NLP) desarrollado en Python, utilizando información pública disponible en las páginas web de los competidores, específicamente en las secciones relacionadas con sus objetivos y servicios ofrecidos. En el proceso, se eliminaron preposiciones y palabras irrelevantes, como "contacto", "dirección", "menú" y "más", para garantizar que el análisis se enfocara exclusivamente en contenido relevante.

El resultado de este modelo fue la generación de una nube de palabras para cada página analizada, permitiendo identificar tanto el enfoque como los servicios destacados de los competidores. Este análisis proporcionará a BritCham una herramienta estratégica para diseñar iniciativas de crecimiento competitivo. Los resultados completos se encuentran en el Anexo técnico 2, sin embargo, en las siguientes figuras se encuentran algunos ejemplos realizados a partir de los siguientes links:

Tabla 7 Links usado modelo NLP

Titulo	enlace
BritCham - Objetivos	https://britcham.com.co/la-camara/
BritCham - Servicios	https://britcham.com.co/servicios/
HollandHouse - Servicios	https://hollandhouse-colombia.com/aliados/membresias

Figura 9 Nube de palabras construcción propia a partir de los objetivos BritCham

Figura 10 Nube de palabras construcción propia a partir de los servicios BritCham

Figura 11 Nube de palabras construcción propia a partir de los servicios HollandHouse

El análisis permitió identificar que BritCham tiene grandes oportunidades para consolidar su posición en el mercado mediante la diversificación de su portafolio de servicios, integrando asesorías legales, comerciales y de inversión para apoyar a empresas en su expansión entre Colombia y el Reino Unido. También puede optimizar su estructura de membresía con niveles de beneficios personalizados que permitirán atraer nuevos afiliados y atender mejor a empresas en diferentes etapas de la afiliación.

La inversión en estrategias digitales y de marketing, incluyendo la automatización y una mayor presencia en redes sociales, maximizará su alcance y visibilidad. Por otro lado, fortalecer alianzas estratégicas, especialmente con la Embajada Británica y otras cámaras internacionales, ampliará su red de contactos y generará nuevas oportunidades para sus miembros.

La adopción de tecnologías avanzadas y plataformas de inteligencia empresarial añadirá valor mediante análisis personalizados, mientras que la promoción de valores empresariales británicos como sostenibilidad, innovación y ética reforzará su identidad en

el mercado, haciendo que estas iniciativas posicionen a BritCham como un actor clave en el comercio bilateral, asegurando su relevancia y un crecimiento sostenido en el entorno binacional.

9.2. Comprensión de los Datos

En cuanto al entendimiento de los datos de las bases de exportaciones e importaciones, se realizó por medio Google Colab.

9.2.1. Exportaciones

La base de datos de exportaciones tiene 242.873 registros y 68 columnas en el periodo comprendido entre enero de 2022 y marzo de 2024, 10 de las 68 columnas contaban con valores nulos:

Tabla 8 Columnas con valores nulos exportaciones

Nombre de columna	# registros	% de nulos
Número de la declaración definitiva	242.843	0,01%
Fecha de Declaración de Exportación Anterior	175.303	27,82%
Número de declaración de exportación anterior	72.392	70,19%
Fecha De Declaración De Importación Anterior	144.923	40,33%
Número De Declaración De Importación Anterior	619	99,75%
Dirección agente aduanero	241.810	0,44%
Razón social del importador	242.831	0,02%
Dirección del importador	242.827	0,02%
Descripción de la mercancía	117.694	51,54%
Número de Autorización de Embarque	242.515	0,15%

Los tipos de datos encontrados en la base están distribuidos de la siguiente forma:

Tabla 9 Tipos de datos exportaciones

Objeto	35	51,47%
Float	21	30,88%
Integer	12	17,64%

35 Strings : Capitulo Del Arancel, Tipo de declaración, Modalidad de importación, Tipo De Datos, Exportación en Tránsito, Aduana, Aduana De Embarque, Oficina Min Comercio, Agente aduanero(s), Usuario, Razón social actual Exportador, Municipio, Dirección agente aduanero, Clase de Exportación, Razón social del importador, Dirección del Importador, Descripción de la partida arancelaria, Descripción de la Mercancía, Unidad comercial, País de Destino, Departamento Origen, Departamento De Procedencia, Lugar de salida, Código de embarque, Vía de transporte, Nacionalidad del medio de transporte, régimen exportación, Modalidad de exportación, Certificado de Origen, Sistemas Especiales, Moneda de negociación, Forma de pago, Continente Destino.

18 Float: Número de declaración de exportación anterior, Número De Declaración De Importación Anterior, NIT del exportador, Dirección agente aduanero, Dirección del Importador, Cantidad(es), Peso en kilos netos, Peso en kilos brutos, Valor FOB (USD), Valor FOB (COP), Valor Agregado Nacional (VAN), Valor Flete, Valor seguro, Valor otros, Precio Unitario FOB (COP) Peso Neto, Precio Unitario FOB (COP) Peso Bruto, Precio Unitario FOB (USD) Peso Neto, Precio Unitario FOB (USD) Peso Bruto, Precio Unitario FOB (USD) Cantidad, Precio Unitario FOB (COP) Cantidad.

12 Integer: Fila, Año, Mes, Dia, Año de la Declaración Definitiva, Mes de la declaración definitiva, Día de la Declaración Definitiva, Fecha de Declaración de Exportación Definitiva, Número de la declaración definitiva, Fecha de Declaración de Exportación Anterior, Fecha De Declaración De Importación Anterior, Código Agente aduanero, Código De Usuario, Código Partida, Número de artículos, Fecha de Embarque, Número de Autorización de Embarque.

En el análisis de frecuencias se observó que variables como " Descripción de la Mercancía " presentan una dispersión considerable en la información, lo que sugiere que no aportarían valor significativo al análisis. Por otro lado, la variable "Fecha de Declaración de Exportación Anterior" resulta útil, ya que permite realizar un análisis de recencia, el cual analiza el tiempo transcurrido entre una exportación actual y una anterior, brindando insights sobre patrones temporales de exportación.

En cuanto a la variable "Tipo de Datos", se concluye que no contribuye de manera significativa al análisis, mientras que la "Razón Social del Exportador" y la "Razón Social del Importador" serán tratadas como identificadores únicos (IDs) debido a su naturaleza. En relación con la columna "Municipio", se identifica el potencial de realizar un análisis sectorial enfocado en localidades específicas dentro del departamento de origen o destino de la mercancía, lo que podría ofrecer información valiosa a nivel regional.

Respecto a la "Clase de Exportación", la variable no tiene relevancia en este contexto, dado que el 99.94% de los registros corresponden al tipo "privado", con solo un 0.06% en la

categoría "público". En cuanto a la "Unidad Comercial", debido a la diversidad de unidades (unidades, kilogramos, litros, metros cúbicos, etc.), se ha decidido utilizar la variable "Peso en kilos netos", ya que ofrece una medida más general y uniforme para segmentar la información. Adicional la variable "Departamento De Procedencia" se considera irrelevante para el análisis, ya que el origen de la mercancía no tiene un impacto significativo en los objetivos del estudio.

Finalmente, la variable "País de Destino" es crucial para evaluar la frecuencia de las exportaciones hacia diversos países, al igual que "Departamento de Origen", lo que permitirá realizar un análisis detallado de los destinos más frecuentes. En cuanto a "Forma de Pago", se ha determinado que no proporciona información relevante, ya que solo clasifica las transacciones en categorías de pago con "ítem completo" o "sin ítem completo".

Tabla 10 Frecuencias País de destino (TOP 10).

País	F	F.R	F.A	F.R.A
España (UE)	2478	25.1%	2478	25.1%
Países bajos (UE)	1958	19.8%	4436	44.9%
Reino unido (UE)	1183	12.0%	5619	56.9%
Bélgica (UE)	889	9.0%	6508	65.9%
Alemania (UE)	693	7.0%	7201	72.9%
Italia (UE)	517	5.2%	7718	78.1%
Francia (UE)	515	5.2%	8233	83.3%
Polonia (UE)	275	2.8%	8508	86.1%
Rusia	209	2.1%	8717	88.2%
Suiza	182	1.8%	8899	90.0%

Los 10 primeros países de destino concentran el 90% de las exportaciones, siendo España el principal receptor con el 25.1%. Rusia y Suiza, aunque no tan representativos, siguen siendo destinos relevantes, con Rusia alcanzando un 2.1% y Suiza un 1.8%. Este patrón muestra una fuerte concentración de las exportaciones hacia un grupo reducido de países, lo que puede ser útil para orientar estrategias comerciales y de diversificación de mercados.

En relación con los estadísticos, se observa que la variable "Cantidad" no es adecuada para un análisis claro debido a que se encuentra representada en unidades no uniformes. En cambio, una de las variables más relevantes es "Peso en Kilos Netos", ya que refleja el peso real de la mercancía excluyendo el embalaje. Además, el "Valor FOB (USD)" es clave, ya que representa el valor total de la mercancía en términos FOB en dólares estadounidenses, permitiendo realizar análisis más significativos.

Variables como "Número de artículos", "Valor Agregado Nacional (VAN)", "Valor Flete", "Valor Seguro" y "Valor otros" no se consideran relevantes, dado que están mayormente en cero o uno (75% de los registros), lo que indica que no se reportan o no se aplican en muchas exportaciones. Esto resalta que dichas variables no son útiles para el análisis global, ya que en gran parte de los casos no aportan información significativa. Las demás variables, como "Precio Unitario FOB (USD)" y "Valor FOB (USD)", son relevantes para el análisis, ya que ofrecen información útil sobre el valor de la mercancía. Se ha decidido

tomar variables únicamente en moneda a USD para facilitar la interpretación, omitiendo el valor en COP, esto simplifica el análisis y asegura la coherencia en los datos.

Tabla 11 Estadísticos Valor FOB y Valor Unitario FOB (USD) Exportaciones

	Valor FOB (USD)	Unitario FOB (USD)
Cuenta	$ 9.883	$ 9.883
Media	$ 14 mil millones	$ 846
Desviación estándar	$ 342.694	$ 69.626
Mínimo	$ 0	$ 0
25%	$ 587	$ 1.85
50%	$ 4.615	$ 5.03
75%	$ 31.008	$ 10
Máximo	$ 13 mil millones	$ 6.916.000

El análisis exploratorio de los datos de exportaciones revela patrones relevantes en cuanto a destinos, productos, regiones y actores económicos involucrados en el comercio exterior durante el periodo observado. A continuación, otras observaciones identificadas:

Productos más exportados:

- Prendas y complementos de vestir (103.126 unidades),
- Plantas vivas y productos de la floricultura (77.717),
- Frutas y cortezas de cítricos (61.140),
- Café, té, yerba mate y especias (13.701), y
- Plásticos y sus manufacturas (4.791)

En cuanto al comercio específicamente dirigido al Reino Unido, se identificaron las siguientes empresas como principales exportadoras según el valor FOB:

- LCC (La Guajira): $96.992.330 – productos: combustibles minerales.
- Comercializadora Internacional Bananeros Unidos de Santa Marta (Magdalena): $83.740.147 – frutas cítricas.
- Unión de Bananeros de Urabá Uniban (Antioquia): $68.661.432 – frutas cítricas.
- CITECNICAS Baltime de Colombia (Magdalena): $51.839.125 – frutas cítricas.
- Flores El Capiro (Antioquia): $49.473.923 – productos de floricultura.

9.2.2. Importaciones

La base cuenta con 1.797.673 registros y 31 columnas para el mismo periodo de tiempo de la base de exportaciones, teniendo 9 columnas con vacíos, pero no con una relevancia considerable respecto al número total de registros, ya que el porcentaje promedio de nulos es de 0.28% en toda la base. Los tipos de datos se encuentra distribuidos de la siguiente forma:

Tabla 12 Tipos de datos importaciones

Objeto	18	58,06%
Float	9	29,03%
Integer	4	12,90%

18 Strings: Capítulo de arancel, tipo de importación, razón social del importador, actividad económica del importador, departamento del importador, dirección del importador, teléfono del importador, descripción de la partida arancelaria, unidad comercial, país de origen, País de procedencia, país de compra, departamento de destino, municipio, proveedor, país del exportador, ciudad del proveedor, dirección del proveedor.

9 Float: Código agente aduanero, NIT del importador, código importador, código de Partida, cantidad (es), subpartidas, peso en kilos netos, valor FOB (USD), pecio unitario FOB (USD), cantidad unidad comercial, valor CIF (COP).

4 Integer: Fila, año, mes.

Tras analizar el contenido de las variables, se decidió excluir del modelo las siguientes: fila, código agente aduanero, código importador (10 Dig), código partida, proveedor, ciudad del proveedor y dirección del exportador. Estas variables fueron descartadas porque algunas corresponden a códigos generados por la oficina de Aduanas, que son altamente variables y carecen de patrones consistentes, mientras que otras contienen datos textuales diversificados y sin categorización, lo que dificulta su integración en el modelo.

Adicionalmente, se realizó un análisis de frecuencias para las variables capítulo del arancel, tipo de importación, actividad económica del importador, departamento del importador, descripción de la partida arancelaria, unidad comercial, subpartidas, país de origen, país de procedencia, país de compra, departamento destino, país del exportador y

municipio. Con este análisis, se identificó el grado de repetición de cada categoría dentro de la base de datos. Los resultados indicaron que las variables capítulo del arancel, tipo de arancel, actividad económica del importador, descripción de la partida arancelaria y subpartidas no aportan valor significativo al modelo debido a la gran cantidad de categorías (más de 100) y a la dispersión de los datos, lo que dificulta identificar comportamientos consistentes.

Por último, las variables restantes, que requieren un análisis más detallado, podrían ser categorizadas o agrupadas en rangos para facilitar su inclusión y análisis en el modelo.

Tabla 13 Frecuencias por país de origen (TOP 10)

País	F	F.R	F.A	F.R.A
Alemania	16821	26,3%	16821	26,3%
Italia	8420	13,1%	25241	39,4%
España	6674	10,4%	31915	49,8%
Francia	5940	9,3%	37855	59,1%
Turquía	3888	6,1%	41743	65,2%
Reino unido	3165	4,9%	44908	70,1%
Rumania	2510	3,9%	47418	74,0%
Suiza	2400	3,7%	49818	77,8%
Polonia	1659	2,6%	51477	80,4%
Suecia	1579	2,5%	53056	82,9%

El análisis de frecuencias muestra que Alemania (UE) es el principal socio comercial con 16,821 registros, representando el 26.27% del total, seguido por Italia (UE) (13.15%) y España (UE) (10.42%). Estos tres países concentran casi el 50% de las relaciones comerciales analizadas, lo que indica una fuerte dependencia hacia estos mercados.

Por otro lado, los países con menor frecuencia, como Bielorrusia (0.0016%) y Chipre (UE) (0.0031%), tienen una participación insignificante, lo que sugiere áreas con bajo nivel de comercio o potencialmente descuidadas.

En términos acumulados, los primeros 10 países concentran más del 80% del total, lo que refleja una distribución altamente concentrada en pocos actores clave. Sin embargo, hay otras variables que pueden demostrar el comportamiento de las importaciones con base en las frecuencias obtenidas, como departamento destino, país de procedencia, departamento del importador y municipio.

En cuanto a los estadísticos obtenidos, se decide solamente trabajar con las variables peso en Kilos netos, valor FOB (USD) y precio unitario FOB (USD), sin el uso de cantidades importadas ya que pueden sesgar los análisis a realizar, los datos se comportan de la siguiente manera considerando las variables de más interés.

Tabla 14 Estadísticos Valor FOB y Valor Unitario FOB (USD)

	Valor FOB (USD)	Unitario FOB (USD)
Cuenta	$ 64.028.000.000	$ 64.028.000.000
Media	$ 14.073.680.000	$ 2.268.377.000
Desviación estándar	$ 207.761.300.000	$ 30.840.070.000
Mínimo	$ -	$ -
25%	$ 152.907.500	$ 9.541.621
50%	$ 794.065.000	$ 44.357.180
75%	$ 4.088.045.000	$ 238.444.800
Máximo	$ 16.854.950.000.000	$ 2.853.280.000.000

9.2.3. Estados Financieros

La base presenta valores financieros de los últimos 2 años registrados por las empresas e información de contacto. Tiene 13 columnas y 111.790 registros, 5 columnas presentan vacíos en sus registros, sin embargo, no son significativos, ya que la columna "página web" con el porcentaje más alto de vacíos es 94% sin mucha relevancia para el análisis, y con un promedio total de 21% de vacíos en la base.

Tabla 15 Columnas con valores nulos Estados Financieros

Nombre de columna	% de nulos
Número de empleados	10%
Fecha de Incorporación	0,3%
Correo Electrónico	0,01%
Teléfono	0,15%
Página Web	94%

Los tipos de datos encontrados son Objetos (8) y Float (5), lo que significa que la mayoría de los datos en esta base son cualitativos como Compañía, ciudad, estatus operacional, correo, teléfono y página web.

En cuanto a los datos estadísticos de esta base se recalca que:

Tabla 16 Valores estadísticos Estados Financieros

Estadístico	Activos Totales	Pasivos Totales	Total de patrimonio
Media	$ 1.82 billones	$ 9.02 billones	$ 9.22 billones
Des. estdr	$ 2.69 billones	$ 1.25 billones	$ 1.63 billones
Mínimo	$ 0	-$739	-$ 333.136
25%	$ 795	$ 2.512.125	$ 3.297.975

50%	$ 1.740.245	$ 724	$ 825.625
75%	$ 5.090.745	$ 2.312	$ 2.472
Máximo	$ 42.844.950	$ 16.480.904	$ 26.364.046

Los activos totales presentan un valor promedio muy alto (≈ $1.82 billones), pero también una desviación estándar considerable ($2.69 billones), lo que sugiere una gran dispersión en los valores. Lo mismo ocurre con pasivos y patrimonio totales, ambos con medias altas (≈ $9.02 billones y $9.22 billones respectivamente), acompañadas de desviaciones estándar también elevadas, lo que indica una fuerte variabilidad entre las observaciones.

Las diferencias entre los cuartiles y los valores máximos sugieren una distribución altamente asimétrica y posiblemente con valores atípicos, aunque el valor máximo de los activos totales es $42.844.950, el 75% de los datos se encuentran por debajo de $5.090.745.

La diferencia entre la media y la mediana (50%) en las tres variables sugiere que los datos no siguen una distribución normal, en los activos totales, la media es mucho mayor que la mediana ($1.8 billones vs. $1.740.245), lo cual indica que existen valores extremos que están sesgando la media hacia la derecha.

9.3. Preparación de los Datos

Para garantizar la calidad, integridad y coherencia de los datos para el modelado, se llevó a cabo un la limpieza y transformación sobre las bases de datos relacionadas con exportaciones, importaciones, estados financieros y afiliados a BritCham. Este proceso

incluyó varios pasos secuenciales orientados a la depuración, validación y enriquecimiento de la información.

Inicialmente, se eliminaron los registros con valores nulos en campos clave como el NIT, ya que este identificador era esencial para el cruce entre fuentes, y también los que no cumplían con el formato estándar de 9 dígitos requerido por la DIAN.

Se eliminaron columnas categóricas con alta dispersión o varianza que dificultaban el análisis por su baja representatividad, así como aquellas sin relevancia para los objetivos del estudio como “Aduana”, “Aduana De Embarque”, “Oficina Min Comercio”, “Agente aduanero(s)”, “Usuario” o, por el contrario, variables que no tuvieran registros suficientes como “Número De Declaración De Importación Anterior” en la base de exportaciones. De igual forma, se descartaron columnas numéricas susceptibles de introducir sesgos en el modelado posterior como “Cantidades”, “Peso en kilos netos” y “Peso en kilos brutos”.

Con base en la comprensión de los datos, se seleccionaron variables clave que agregaban valor al análisis, como el departamento de origen, país de destino, capítulo del arancel, valor FOB, entre otras y se realizó la agrupación por NIT y periodo (Año-Mes) para obtener un único registro por empresa tanto en exportación como en importación.

A partir de esta agregación, se generaron nuevas variables de resumen como el valor FOB máximo y total, precios unitarios promedios y máximos, la frecuencia de exportaciones, creando también variables de frecuencia ajustadas a ventanas móviles de 3 y 6 meses y el número de países distintos a los que exportó cada empresa.

Una vez estructurada la base de comercio exterior, esta fue integrada con los estados financieros de las compañías, permitiendo combinar información económica y operativa. En este proceso, también se eliminaron variables poco significativas de los estados financieros y se imputaron valores faltantes en variables como activos, pasivos, patrimonio y antigüedad utilizando la mediana como medida robusta, ya que proporciona una buena estimación de los valores que faltan.

Finalmente, se incorporó una variable binaria que indica si la empresa está afiliada a BritCham, a partir de la base de datos de afiliados. El conjunto resultante consolidó 130.034 registros (114.817 importaciones y 19.794 exportaciones) junto con 34 columnas.

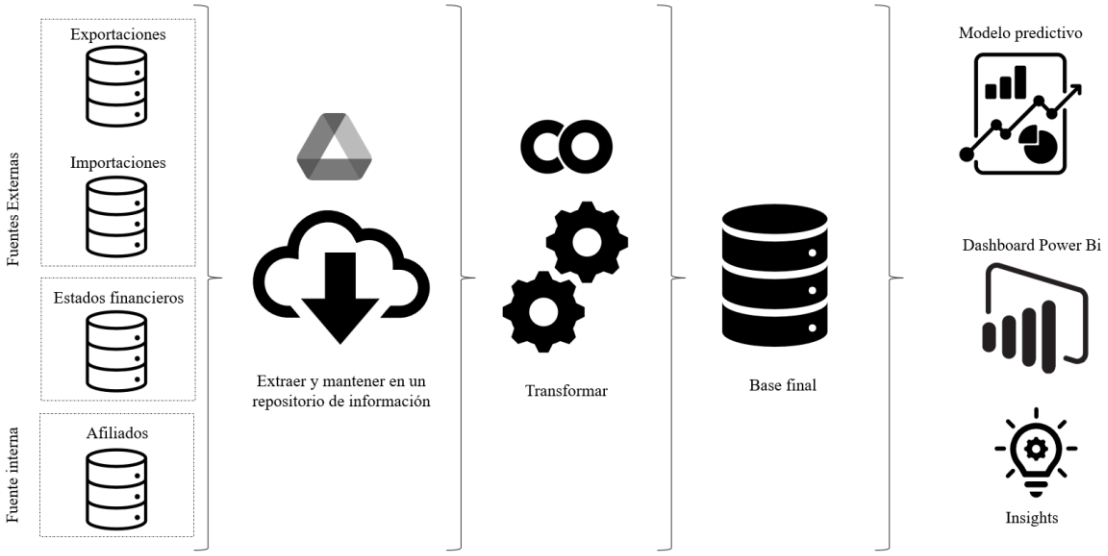


Figura 12 ETL del proyecto

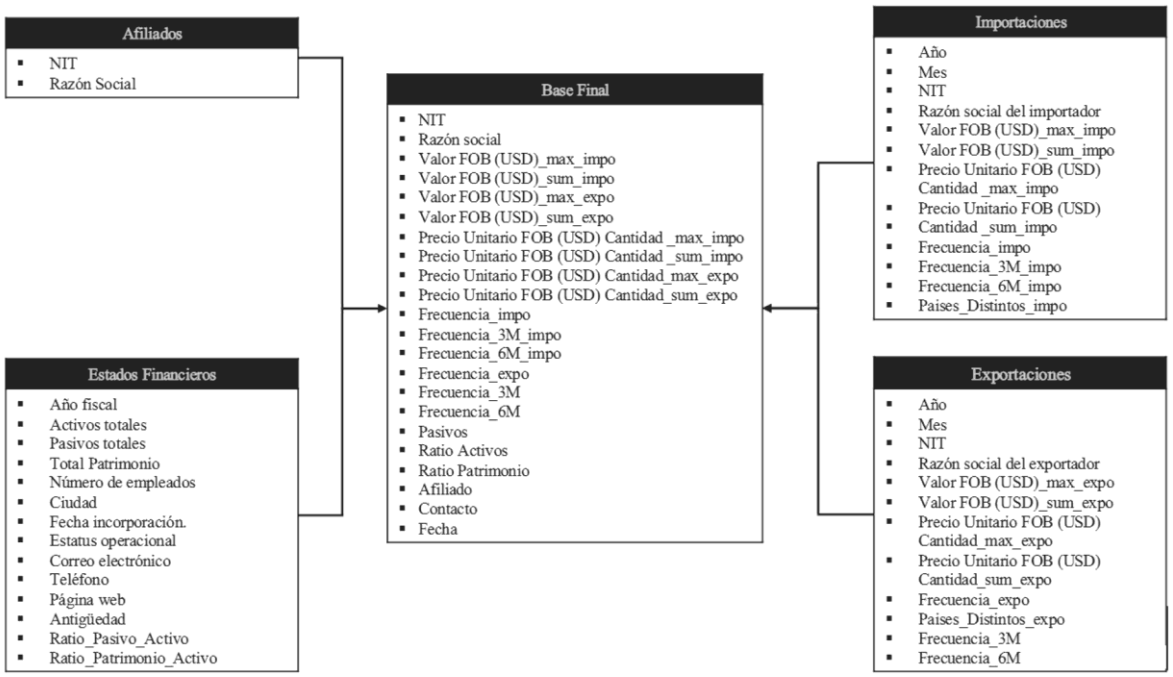


Figura 13 Arquitectura del modelo de datos

Considerando la permanencia en el tiempo, se suguiere que la actualización de los datos del tablero, se realice teniendo en cuenta la actualización de estados financieros de las empresas anualmente, brindados por la plataforma Emis Next, de tal forma que BritCham pueda establecer un objetivo y porcentaje anual de captación de clientes con base en la información suministra de más de 2 millones de registros incluidos en las bases de datos de comercio exterior por la fuente de información, Legiscomex.

9.4. Modelado

Luego de la limpieza de los datos se escogieron las variables más relevantes dentro de la base integrada de exportaciones e importaciones, con el fin de poder dar

cumplimiento al segundo objetivo en relación con los modelos de segmentación y clasificación, considerando 3 factores importantes: las empresas que realizan importaciones, las empresas que realizan exportaciones y las empresas que realizan ambas actividades.

El diseño de los modelos se llevó a cabo en Google Colab, servicio en la nube de Jupyter Notebook, que permite escribir y ejecutar código Python, permitiendo el desarrollo de las dos etapas anteriores: comprensión de los datos y preparación de los datos, así como también el modelado; de esta manera, las siguientes variables fueron seleccionadas como representativas para cada segmento en el que se dividió la base:

Tabla 17 Variables relevantes dentro del modelo de segmentación

Exportador	Importador	Importador y Exportador
Valor FOB (USD)_sum_expo	Valor FOB (USD)_sum_impo	Valor FOB (USD)_sum_impo
Precio Unitario FOB	Precio Unitario FOB (USD)	Valor FOB (USD)_sum_expo
Cantidad_sum_expo	Cantidad _sum_impo	Precio Unitario FOB (USD)
Frecuencia expo	Frecuencia_impo	Cantidad _sum_impo
Frecuencia_3M_expo	Frecuencia_3M_impo	Precio Unitario FOB (USD)
Frecuencia_6M_expo	Frecuencia_6M_impo	Cantidad_sum_expo
Antigüedad	Activos Totales	Frecuencia_impo
Número de empleados	Pasivos Totales	Frecuencia_expo
Activos totales	Total de patrimonio	Frecuencia_3M_impo
Ratio_Pasivo_Activo	Número de empleados	Frecuencia_3M_expo
Ratio_Patrimonio_Activo		Frecuencia_6M_impo
		Frecuencia_6M_expo
		Año Fiscal
		Antiguedad
		Número de empleados
		Activos Totales
		Ratio_Pasivo_Activo
		Ratio_Patrimonio_Activo

En los 3 factores mencionados anteriormente, las variables fueron escaladas con StandardScaler, con el fin de normalizar las variables númericas en un conjunto de datos, transformando los datos con media 0 y desviación 1, logrando que todas las variables tuvieran el mismo peso, mejorando el desempeño y estabilidad del modelo (Scikit Learn, 2025). Así mismo, el proceso para llevar a cabo el modelado, en las 3 situaciones siguió el mismo flujo de trabajo:

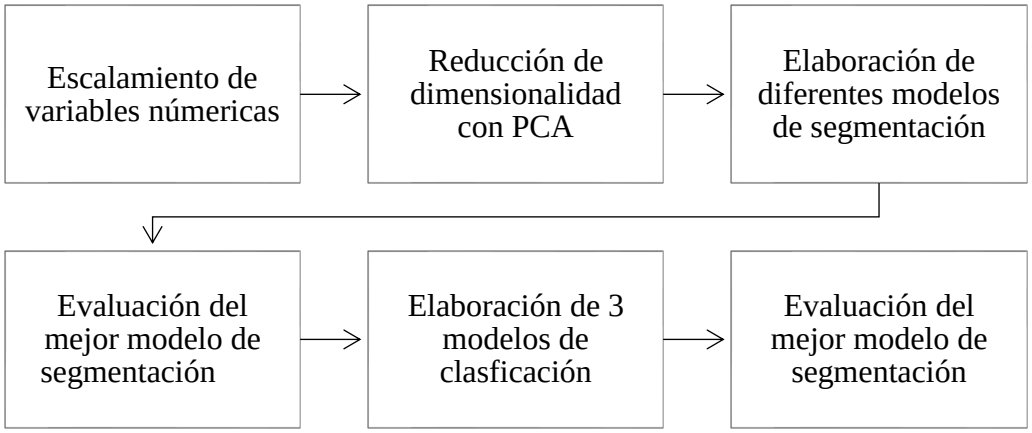


Figura 14 Proceso de Modelado

9.4.1. Modelado de Segmentación en Exportaciones

Luego de escalar las variables, se aplicó el Análisis de Componentes Principales (PCA), una técnica que permite reducir la dimensionalidad del conjunto de datos transformando las variables originales —potencialmente correlacionadas— en un nuevo conjunto de variables llamadas componentes principales. Estas componentes capturan la

mayor parte de la variabilidad del conjunto original, facilitando el análisis sin perder información relevante (IBM, 2025).

Para determinar cuántas componentes conservar, se utilizó el análisis de varianza acumulada, con el cual se identificó que con 6 componentes principales se explica más del 90% de la varianza total, lo que garantiza una representación fiel de los datos manteniendo la simplicidad del modelo.

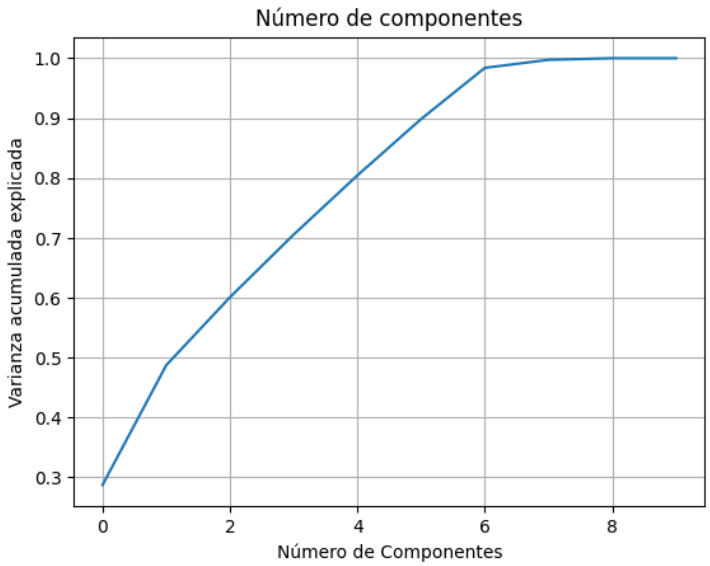


Figura 15 Número de componentes principales PCA

Cada componente principal es una combinación lineal de las variables originales, diseñada para capturar patrones distintos y no redundantes. El primer componente (PC1), por ejemplo, representa la dimensión más significativa del comportamiento exportador de las empresas, distribuida de la siguiente manera:

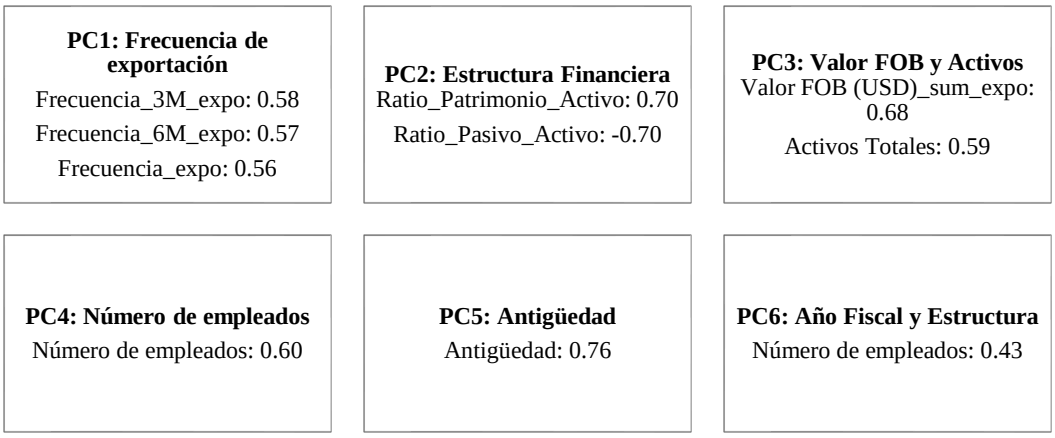


Figura 16 Componentes principales PCA exportaciones

Una vez realizada este análisis, se procedió a aplicar los 3 algoritmos de segmentación no supervisada: K- Means, Gaussian Mixture Models y BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), con el objetivo de segmentar a las empresas exportadoras según patrones similares.

9.4.1.1. K- Means

K-Means es un algoritmo de aprendizaje no supervisado que se utiliza para agrupar datos sin etiquetas en diferentes clústeres, basándose en su similitud. Funciona dividiendo los datos en k grupos, donde cada grupo se forma alrededor de un centroide, que representa el punto medio de los datos dentro del clúster, el algoritmo asigna cada punto de datos al clúster más cercano según una medida de distancia (generalmente la distancia euclidiana) y ajusta los centroides de manera iterativa para minimizar la distancia total entre los puntos y sus respectivos centros. Este proceso se realiza hasta que los grupos dejan de cambiar significativamente (IBM, 2025).

Es una técnica utilizada por su simplicidad y eficiencia, el número de clústeres (k) puede ajustarse según el nivel de detalle que se desea.

Para determinar el número óptimo de clústeres, se utilizaron dos herramientas complementarias: el método del codo, que evalúa la inercia dentro de los grupos (suma de distancias a los centroides), y el Silhouette Score, que mide la coherencia interna de los clústeres formados.

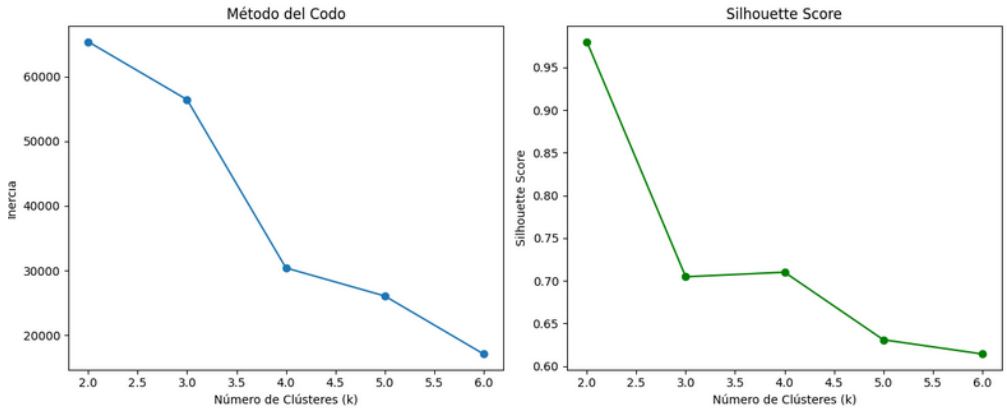


Figura 17 Método de codo y Silhouette score K-Means

Ambos métodos señalaron que cuatro clústeres eran una elección adecuada, al mostrar una reducción significativa en la inercia y un valor alto del índice de Silhouette. Posteriormente, se entrenó el modelo definitivo con k=4. La agrupación permitió identificar perfiles diferenciados entre los exportadores.

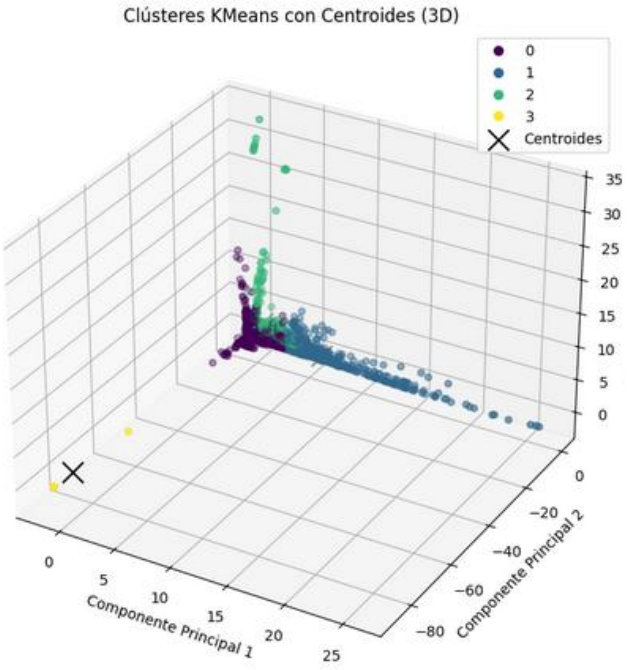


Figura 18 Vista en 3 dimensiones de la distribución de los datos con k=4 K-Means

9.4.1.2. Gaussian Mixture Models (GMM)

El Gaussian Mixture Models (GMM) es un modelo probabilístico de aprendizaje no supervisado, en el que se busca encontrar clúster de puntos en el conjunto de datos, los cuales comparten algunas características en común, y se asume que los datos provienen de una mezcla de distribuciones gaussianas (Carrasco, 2024).

Para determinar el número óptimo de grupos, se evaluaron distintos modelos con entre 1 y 10 componentes, utilizando los siguientes criterios:

Bayesian Information Criterion (BIC) es una medida estadística que permite seleccionar el modelo que mejor representa un conjunto de datos, considerando tanto su ajuste como su complejidad. Se basa en la función de verosimilitud, pero incluye una

penalización por el número de parámetros para evitar el sobreajuste (Geeks For Geeks, 2024).

En modelos de segmentación como los Modelos de Mezcla Gaussiana (GMM), el BIC se utiliza para determinar el número óptimo de clústeres. Evalúa qué tan bien se ajusta el modelo a los datos y castiga los modelos innecesariamente complejos. El modelo más adecuado será aquel con el BIC más bajo, ya que logra el mejor equilibrio entre precisión y simplicidad.

Akaike Information Criterion (AIC) es una medida utilizada para comparar modelos estadísticos y evaluar cuál se ajusta mejor a un conjunto de datos. Cuanto menor sea el valor de AIC, mejor es la calidad del modelo en términos de equilibrio entre ajuste y simplicidad. Este criterio penaliza la complejidad excesiva, evitando el sobreajuste (Zajic, 2025).

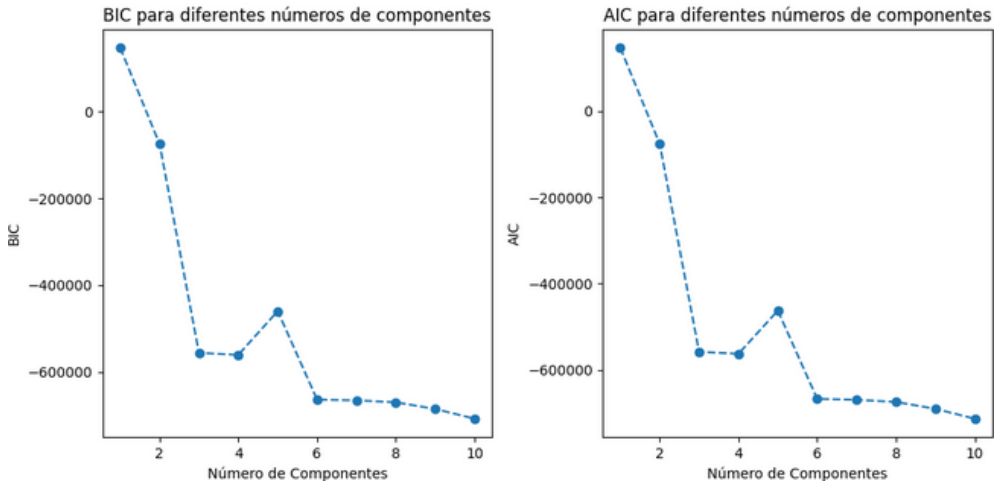


Figura 19 Gráfica de BIC y AIC para el número óptimo de componentes GMM

En modelos como el GMM, donde se debe decidir el número de clústeres a utilizar, el AIC es clave: si se eligen muy pocos clústeres, el modelo puede ser demasiado simple y no capturar la estructura real de los datos; si se eligen demasiados, puede ajustarse demasiado y perder generalización, permitiendo encontrar el punto de equilibrio óptimo, seleccionando el número de componentes que mejor representa los datos sin añadir complejidad innecesaria.

En otras palabras, ambos indicadores miden el ajuste del modelo penalizando la complejidad, y el valor más bajo indica el mejor equilibrio entre precisión y simplicidad. En este caso, el modelo con 10 componentes presentó el menor valor de BIC, por lo que se seleccionó como el número óptimo de clústeres.

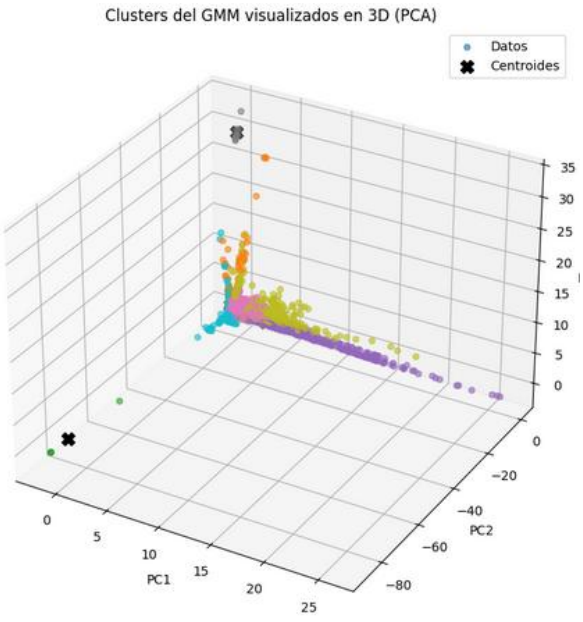


Figura 20 Vista en 3 dimensiones de la distribución de los datos con 10 clúster GMM

Sin embargo, al compararlo con otras métricas de rendimiento de modelos de segmentación, su desempeño con 10 clústeres no era el apropiado, por lo que se realizó una segunda versión del modelo, evaluado en 2 clústeres, respondiendo a la necesidad de comparar una segmentación más general frente al modelo más detallado, los resultados obtenidos, indicaron que, aunque el modelo con dos clústeres presentaba una menor granularidad, logró una separación razonable entre dos grandes grupos de empresas exportadoras y un mejor desempeño en las métricas internas.

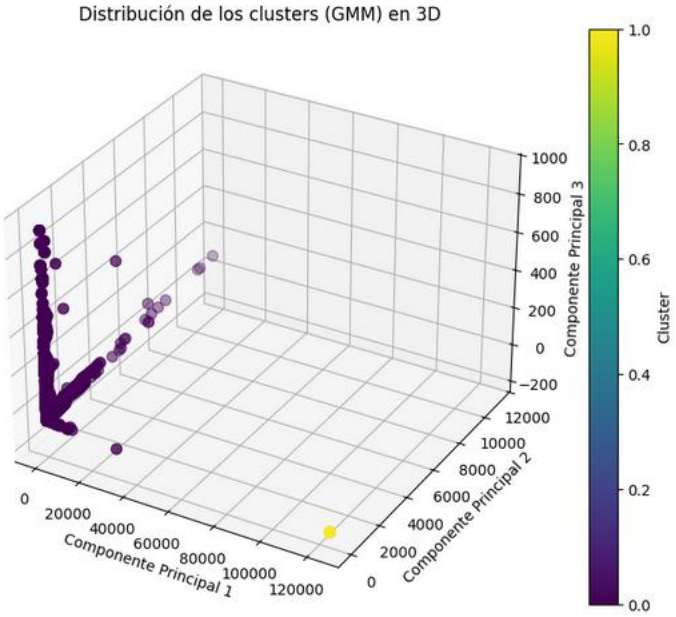


Figura 21 Vista en 3 dimensiones de la distribución de los datos con 2 clúster GMM

9.4.1.3. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Birch es un tipo de clustering usado para grandes conjuntos de datos, siendo útil cuando la escalabilidad (número de datos) de los datos aumenta, tiene como objetivo reducir el número de datos que se ingresan como inputs, generando datos que resumen el

data set original (Otavalo, 2020). El modelo organiza los datos en una estructura similar a un árbol, diseñada para identificar grupos de empresas con características similares y se agrupan de forma incremental y eficiente a medida que se corre cada nodo del árbol.

Cada rama del árbol contiene pequeños grupos de datos (subclústeres), que almacenan información básica como: cuántas empresas hay en ese grupo, cuánto suman sus valores y cuál es su dispersión.

De tal forma que, cuando una nueva empresa se analiza, el modelo buscará el grupo más cercano en el árbol, luego actualiza los valores del grupo con la nueva empresa y repite el proceso hasta llegar a una “hoja” del árbol, donde se finaliza la asignación, permitiendo organizar mejor los datos y mantener la calidad de la segmentación.

Durante el proceso de llevar a cabo el modelo, se probó éste con distintos valores de clúster de 2 a 15, evaluando su desempeño, lo que permitió seleccionar el número óptimo de clúster como 4.

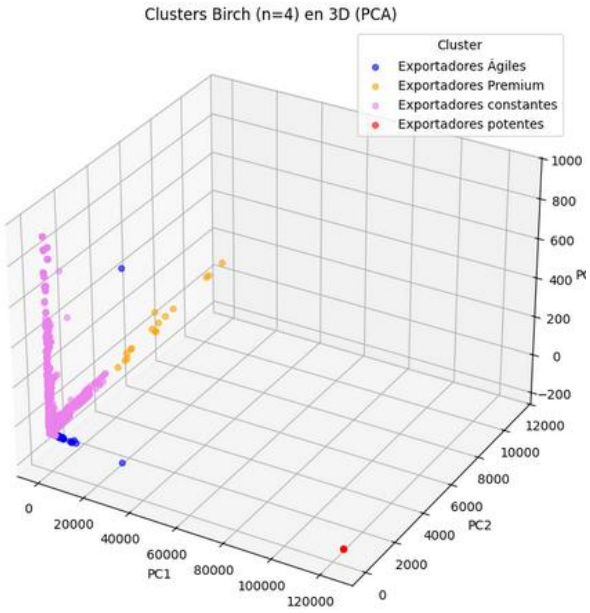


Figura 22 Vista en 3 dimensiones de la distribución de los datos con 4 clúster BIRCH

En conclusión, la aplicación de los modelos de segmentación permitió identificar grupos diferenciados de empresas exportadoras en función de variables clave como el volumen exportado, la frecuencia, la antigüedad, el tamaño y la estructura financiera.

En los 3 modelos desarrollados se evidenció un clúster mayoritario, lo que refleja un patrón de comportamiento homogéneo compartido por la mayoría de las empresas, y los clústeres más pequeños agrupan a aquellas empresas cuyo comportamiento difiere del patrón general, y por tanto, pueden representar nichos estratégicos o casos de alto interés para BritCham.

9.4.2. Modelado de Clasificación en Exportaciones

Con el fin de realizar un modelo mucho más robusto, se desarrollaron 3 modelos de clasificación, donde se buscó dividir los puntos de datos en grupos denominados clases, o

en este caso segmentos, de tal forma que se pueda predecir por medio de las características de cada segmento a partir de los datos de entrada y así asignar posibles segmentos a los nuevos datos que se ingresen considerando las características aprendidas.

Estos algoritmos de clasificación se utilizan en la ciencia de datos para predecir patrones y resultados, y tienen una gran variedad de casos de uso actualmente, siendo principalmente tareas de clasificación binaria o multiclases. Para efectos de este trabajo, se trabajará con una clasificación multiclases, obteniendo la clasificación de los datos en 2 o más segmentos.

9.4.2.1. Random Forest

El random forest o bosque aleatorio es un algoritmo de aprendizaje automático, que combina el resultado de múltiples árboles de decisión para llegar a un resultado único (IBM, 2025).

Según IBM, empresa multinacional de tecnología y consultoría de Estados Unidos: “el algoritmo de bosque aleatorio se compone de una colección de árboles de decisión y cada árbol está compuesto por una muestra de datos extraída de un conjunto de entrenamiento con reemplazo, llamado bootstrapping. De esa muestra de entrenamiento, un tercio de ella se reserva como datos de prueba, lo que se conoce como la muestra fuera de bolsa (obb), y luego, se inyecta otra instancia de aleatoriedad mediante el embolsado de características, lo que agrega más diversidad al conjunto de datos y reduce la correlación entre los árboles de decisión” (IBM, 2025).

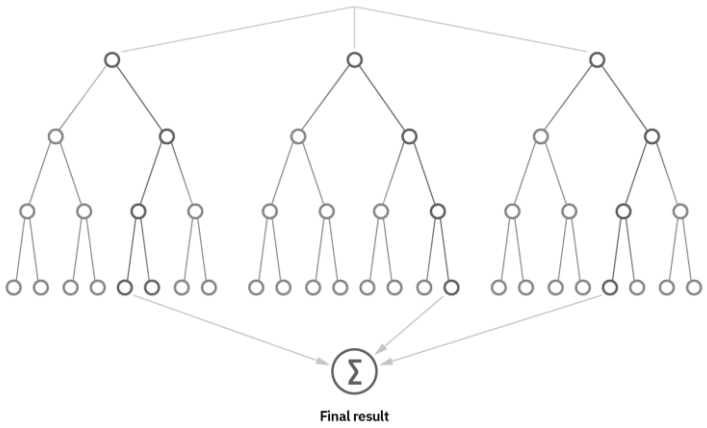


Figura 23 Representación de un bosque aleatorio por IBM

Para este proyecto, en aras de predecir el segmento se utilizó (según el modelo de segmentación Birch) el que pertenece a cada empresa exportadora a partir de variables clave como volumen exportado, frecuencia de operación, antigüedad, número de empleados y ratio financiero. Se dividieron los datos en conjuntos de entrenamiento (70) y prueba (30) y su desempeño se evaluó mediante la matriz de confusión, accuracy, precision, recall, F1-score y las curvas ROC por segmento o clase, lo que permitió medir su capacidad de distinguir correctamente entre los distintos clústers, obteniendo los siguientes resultados:

Matriz de confusión: muestra cómo se comporta el modelo al clasificar cada empresa en su segmento real frente al segmento predicho, es decir, como se clasifican en su respectivo clúster. Cada fila de la matriz representa las verdaderas clases (segmentos Birch), y cada columna representa las predicciones del modelo, concluyendo que, el clúster mayoritario (2) fue clasificado correctamente en la mayoría de los casos, lo que es coherente con su predominancia en los datos, lo que podría generar un “sesgo” donde el modelo tiende a predecir más fácilmente ese grupo; sin embargo, a pesar de estas

limitaciones, el modelo logra distinguir adecuadamente entre varios segmentos. En cuanto a los demás criterios:

- Accuracy: el modelo alcanzó un buen nivel de accuracy, impulsado por su buen desempeño en el clúster mayoritario.
- Precision: se evidencia una alta presición para el clúster 2, es decir que cuando el modelo precide “Cluster 2” es cierto.
- Recall: mide cuántos casos reales de una clase el modelo fue capaz de identificar correctamente, en este caso el recall para el clúster 0 es el más bajo, implicando que varias empresas realmente pertenecientes a ese grupo fueron mal clasificadas.
- F1 Score: indica que el modelo es robusto para todo el conjunto de exportadores y no solo para el clúster 2 (grupo mayoritario)
- Curva ROC: se observan curvas ROC para cada clase cercanas a 1, indicando un excelente desempeño y poder de discriminación.

9.4.2.2. Gradient Boosting

El gradient boosting es un algoritmo de aprendizaje conjunto que produce predicciones precisas combinando múltiples árboles de decisión en un solo modelo, utiliza modelos base para construir sobre sus fortalezas, corregir errores y mejorar las capacidades predictivas. Al capturar patrones complejos en los datos, el aumento de gradiente sobresale en diversas tareas de modelado predictivo.

Es una técnica de conjunto que entrena iterativamente modelos para corregir errores anteriores. Da más peso a los casos mal clasificados en los modelos posteriores, permitiéndoles centrarse en puntos de datos desafiantes y, en última instancia, mejorando el rendimiento general (IBM, 2025).

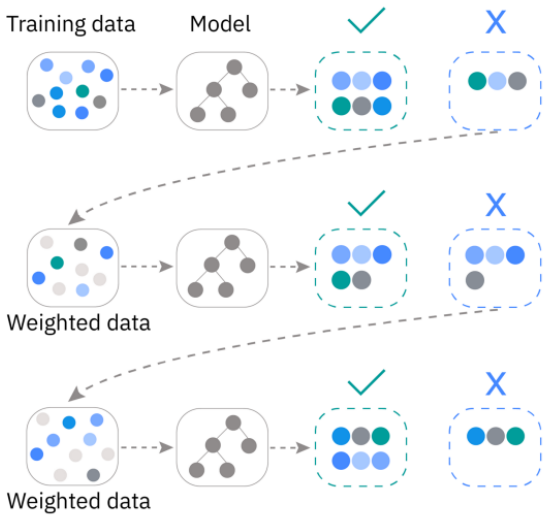


Figura 24 Representación de un Gradient Boosting por IBM

Para efectos del proyecto, el objetivo de la aplicación de este modelo tuvo como objetivo predecir el segmento de cada empresa exportadora a partir de variables operativas y financieras. Al igual que se realizó para el Random Forest, se entrenó y evaluó el modelo utilizando un conjunto de entrenamiento y uno de prueba (70/30) y las mismas variables predictoras utilizadas en el modelo Random forest: valor FOB exportado, la frecuencia de exportación, el número de empleados.

Se evaluaron los mismos criterios que en el modelo mencionado anteriormente, matriz de confusión, accuracy, precision, recall, F1-score y las curvas ROC por segmento o clase:

- Matriz de confusión: se evidenció un comportamiento similar al del modelo Random Forest, el clúster mayoritario con alta precisión.
- Accuracy: El modelo mostró una alta exactitud general, especialmente por su buen desempeño en el clúster dominante.
- Precision, recall y F1 Score: El modelo logró un buen equilibrio entre precisión y recall en las clases más frecuentes, y un desempeño moderado en los clústeres menos representados. El F1-score mostró una ligera mejora respecto al modelo RF en algunas clases, lo que indica que el GB puede capturar patrones más sutiles en los datos.
- Curva ROC: El modelo alcanzó valores AUC elevados para las clases frecuentes, mientras que las clases minoritarias presentaron curvas menos definidas, lo que es consistente con el comportamiento observado en otras métricas.

9.4.2.3.Red neuronal MLP Classifier

La red neuronal MLP (Multi Layer Perception) Classifier es un algoritmo de aprendizaje supervisado, conocido por ser una red supervisada de perceptrones multicapa,

capaz de capturar patrones complejos no lineales en los datos, basado en una arquitectura de capas interconectadas (Scikt learn , 2025).

Con el mismo objetivo que se desarrollaron los 2 modelos anteriores, se llevó a cabo la ejecución del MLP Classifier para predecir el segmento Birch de cada empresa exportadora, con las mismas variables predictoras. El modelo se entrenó sobre datos normalizados utilizando un pipeline que incluía *StandardScaler()* [1]que permitiera garantizar una coincidencia adecuada.

Luego de segmentar las empresas exportadoras, se desarrollaron 3 modelos de clasificación (Random Forest, Gradient Boosting y MLPClassifier) con el propósito de predecir automáticamente el segmento de pertenencia de nuevas empresas. En general, todos los modelos presentaron un buen desempeño, sobretodo en el clúster mayoritario (Cluster 2), cada modelo aportó ventajas específicas:

- **Random Forest**: robustez, buena interpretabilidad y estabilidad.
- **Gradient Boosting:** logró una mayor capacidad de ajuste, capturando relaciones más complejas.
- **MLPClassifier:** mostró un buen comportamiento general pero mayor sensibilidad a la normalización de los datos.

[1] Estandarización de las características eliminando la media y escalada a la varianza.

Los 3 modelos permiten automatizar la asignación de nuevos exportadores a los segmentos identificados, facilitando así el análisis prospectivo, con el fin de poder brindar e e implementar estrategias diferenciadas según el perfil de cada empresa.

9.4.3. Modelado de Segmentación en Importaciones

Al igual que para las empresas exportadoras, para esta clase de importadores se evaluaron múltiples modelos, de tal forma que se pudieran medir criterios de desempeño entre sí.

Es importante resaltar que, para este conjunto de datos con información de importaciones, se aplicó el análisis de componentes principales al igual que los datos de exportaciones y las variables: Valor FOB (USD)_sum_impo, Precio Unitario FOB (USD) Cantidad _sum_impo, Frecuencia_impo, Frecuencia_3M_impo, Frecuencia_6M_impo, Activos Totales, Pasivos Totales, Total de patrimonio, Número de empleados, las cuales fueron escaladas con StandardScaler ().

Del análisis de componentes principales se obtuvo que: los primeros 3 componentes (PC1 + PC2 + PC3) explican más o menos 84% de la varianza total. Para reducir la dimensionalidad y mantener buena parte de la información.

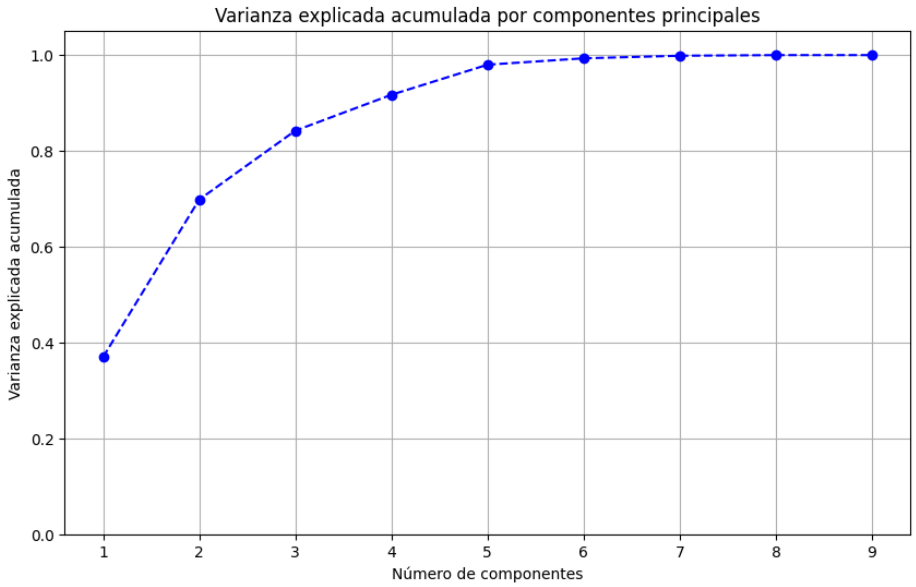


Figura 25 Varianza explicada PC3

Cada componente se define con las siguientes variables:

PC1: Comercio exterior	**PC2: Estabilidad financiera**	**PC3: Dinamismo reciente**
Valor FOB total	Activos totales	Frecuencia 3M y 6M
Precio unitario promedio	Pasivos totales	Precio unitario
Frecuencia general de importación	Patrimonio total	

Figura 26 Componentes principales PCA importaciones

9.4.3.1. K- Means

Considerando la explicación anteriormente descrita para el modelo K-Means en el apartado de exportaciones, se realizó el mismo proceso con base en los datos de importaciones, evaluando el número de clúster en k= 3, k=4 y k=5, según el método de codo.

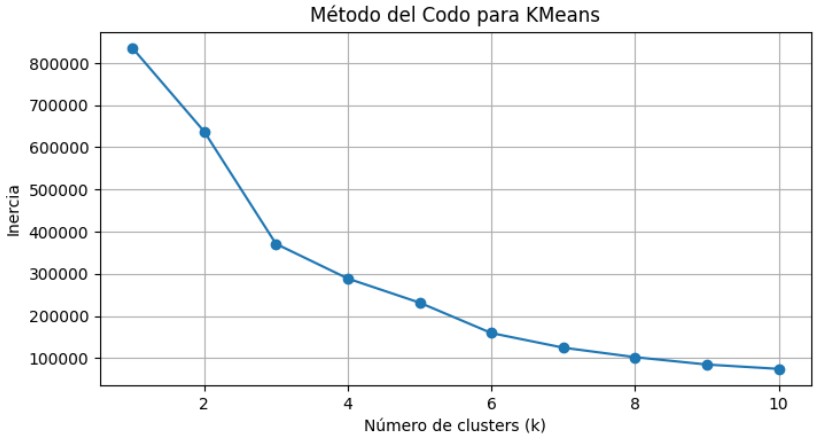


Figura 27 Número de clúster K-Means

Donde se identificó que,

k = 3: tiene una segmentación más general, capturando los grupos más amplios y diferenciados.

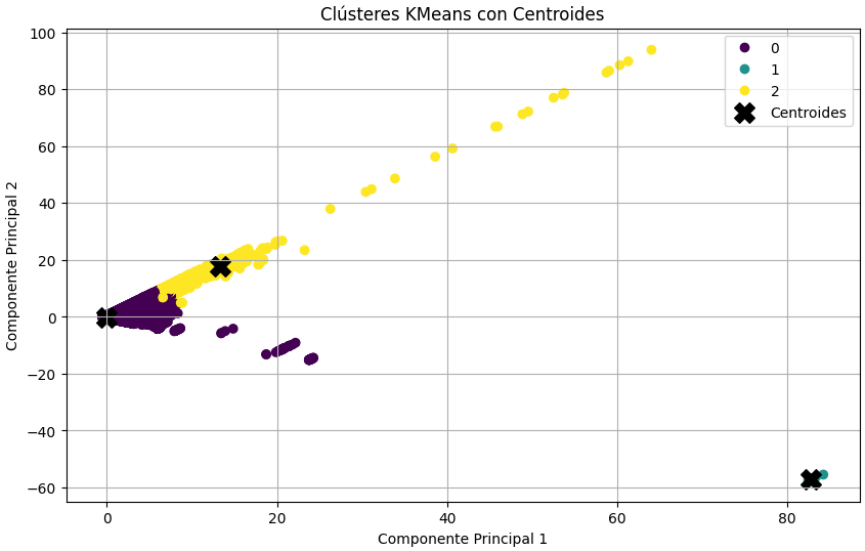


Figura 28 Segmentación con k= 3 K- Means

k = 4 y **k = 5**: cuentan con mayor granularidad, permitiendo observar subgrupos dentro de segmentos más grandes.

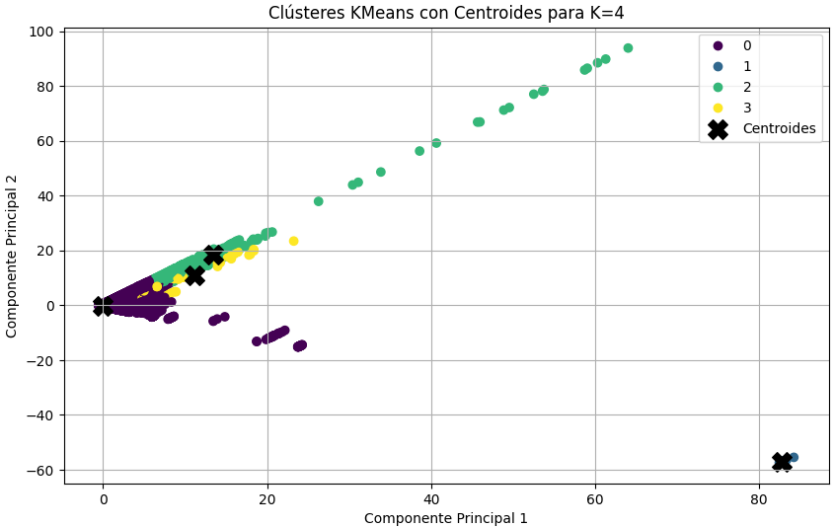


Figura 29 Segmentación k=4 K- Means

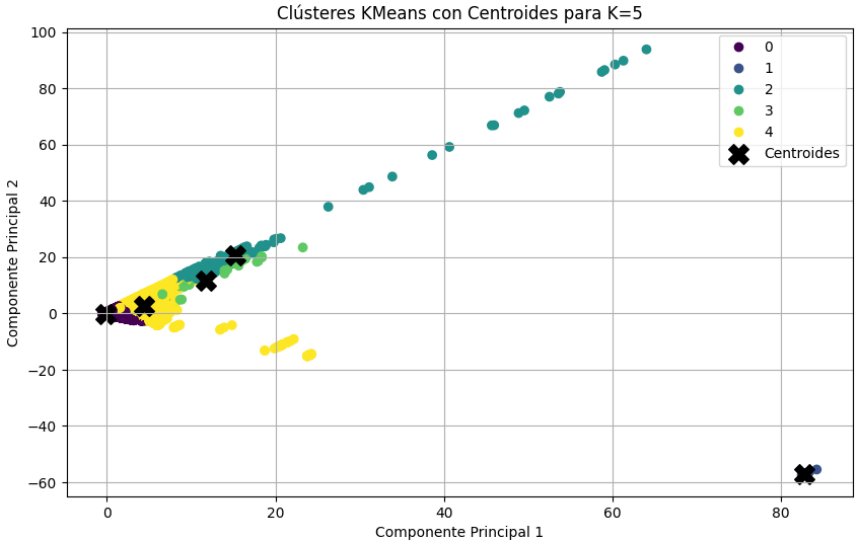


Figura 30 Segmentación k=5 K-Means

9.4.3.2. DBSCAN

El modelo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de agrupamiento basado en densidad. Su funcionamiento se basa en identificar regiones con alta concentración de puntos (densidad) para formar clústeres, y clasificar como ruido aquellos puntos que no pertenecen a ninguna región densa (Geeks For Geeks, 2025). Este modelo clasifica los datos en tres categorías:

Puntos centrales: tienen un número suficiente de vecinos dentro de un radio definido.

Puntos fronterizos: están cerca de puntos centrales, pero no cumplen por sí mismos con el mínimo de vecinos requerido.

Puntos de ruido: no pertenecen a ningún grupo, ya que están aislados o en regiones poco densas.

A diferencia del algoritmo K-Means, DBSCAN no requiere predefinir el número de clústeres, lo cual lo hace especialmente útil para explorar estructuras no lineales o no claramente delimitadas dentro del conjunto de datos.

Para su implementación en este proyecto, se utilizó previamente el Análisis de Componentes Principales (PCA) para la reducción de dimensiones, y se tomó una muestra aleatoria del 20% de la base de datos original. Esta decisión se basó en consideraciones computacionales, ya que el tamaño completo de la base excede la capacidad de

procesamiento de Google Colab en su versión gratuita. Los parámetros definidos para DBSCAN fueron los siguientes:

eps = 1.5: define el radio del vecindario alrededor de un punto de datos.

min_samples = 10: indica el número mínimo de puntos dentro del radio eps necesarios para considerar una región como densa.

Estos dos parámetros interactúan para definir qué se considera una región suficientemente densa para formar un clúster. A partir de estos puntos densos se construyen los grupos, mientras que los puntos que no cumplen estas condiciones se etiquetan como ruido.

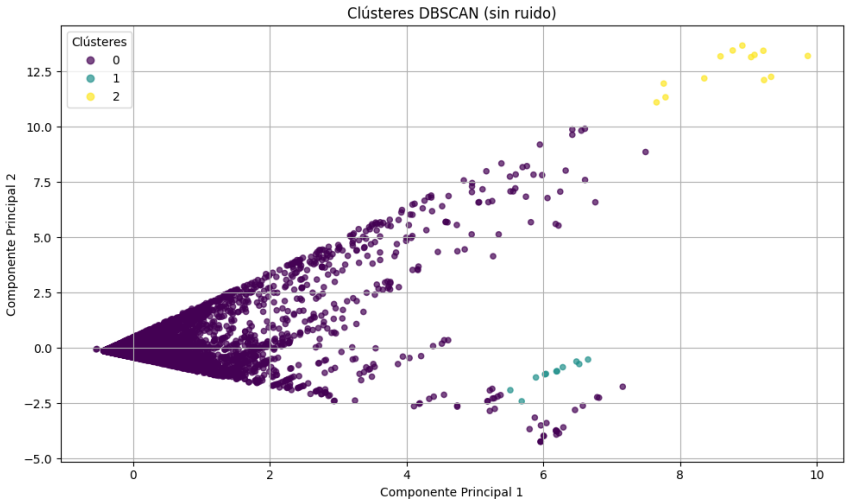


Figura 31 Segmentación BDSCAN 3 clúster

En este caso, BDSCAN ha generado clúster menos definidos, lo que sugiere que la estructura de los datos no es ideal para clustering basado en densidad, a comparación del K-Means que generó clúster más compactos y separados.

9.4.3.3. Spectral Clustering

Este modelo es un algoritmo de partición de datos basado en la teoría de grafos espectrales y álgebra lineal (Data Scientest, 2023), donde la idea es segmentar un gráfico en varios grupos pequeños con caracteristicas similares, considerando los siguientes pasos:

1. Construcción de un grafo a través de la matriz de afinidad (o matriz de similitud, matriz adyacente).
2. Segmentación de puntos de datos en espacios dimensionales más pequeños.
3. Uso de valores y vectores propios para definir subgrafos.

Permitiendo analizar relaciones entre los puntos en forma de un grafo, detectando clústers con formas arbitrarias o no convexas.

Al igual que con el BDSCAN, en este modelo se tomó una muestra representativa del 20%. Los parametros fueron los siguientes:

- n_clusters = 3: número de clústeres a identificar.
- affinity = 'nearest_neighbors': define la afinidad a partir de los vecinos más cercanos.
- n_neighbors = 10: número de vecinos considerados para construir el grafo.
- assign_labels = 'kmeans': usa K-Means para asignar etiquetas finales tras el análisis espectral.

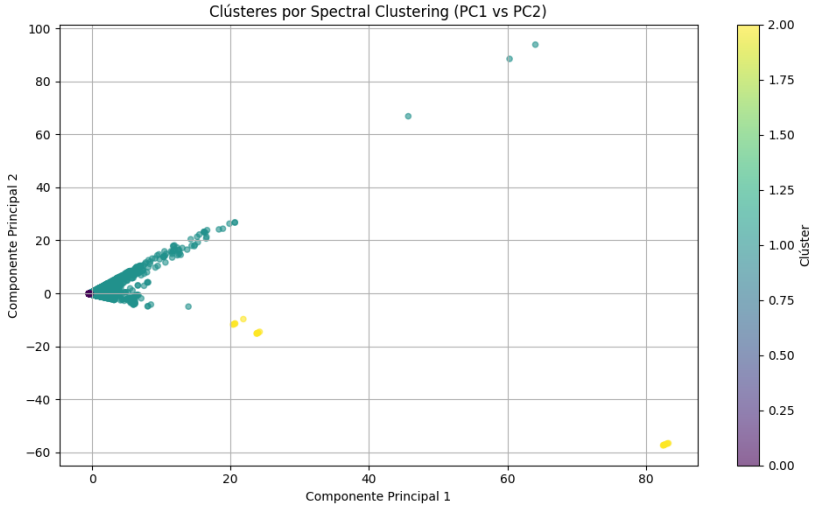


Figura 32 Segmentación Spectral Clustering

9.4.3.4. Clustering Jerárquico

IBM define el clustering jerárquico como: “un algoritmo de machine learning no supervisado que agrupa los datos en un árbol de clústeres anidados. Los principales tipos incluyen aglomerantes y divisivos. El análisis de clústeres jerárquicos ayuda a encontrar patrones y conexiones en conjuntos de datos. Los resultados se presentan en un diagrama de dendrograma que muestra las relaciones de distancia entre los clústeres” (IMB, 2025).

Enfoque aglomerativo o ascendente que fusiona repetidamente los clústeres en otros más grandes hasta que surge un solo clúster.

Enfoque divisivo o descendente que comienza con todos los datos de un solo clúster y continúa dividiendo los clústeres sucesivos hasta que todos los clústeres sean únicos.

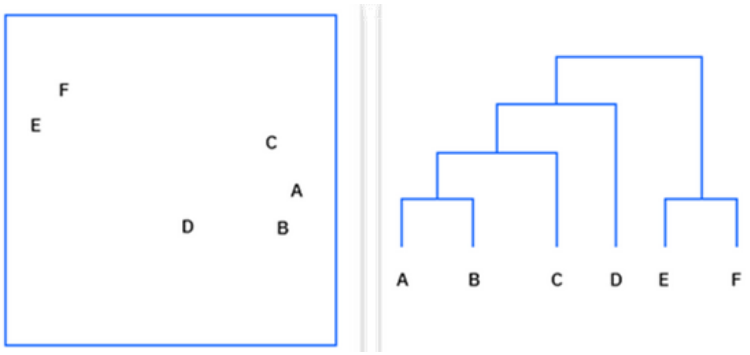


Figura 33 Ejemplo de dendrograma clustering jerárquico

Se suele utilizar un diagrama en forma de árbol llamado dendrograma[2] para visualizar la jerarquía de los clústeres. Muestra el orden en el que se han fusionado o dividido los clústeres y muestra la similitud o distancia entre los puntos de datos. Los dendogramas también pueden entenderse como una lista anidada de listas con atributos (IMB, 2025).

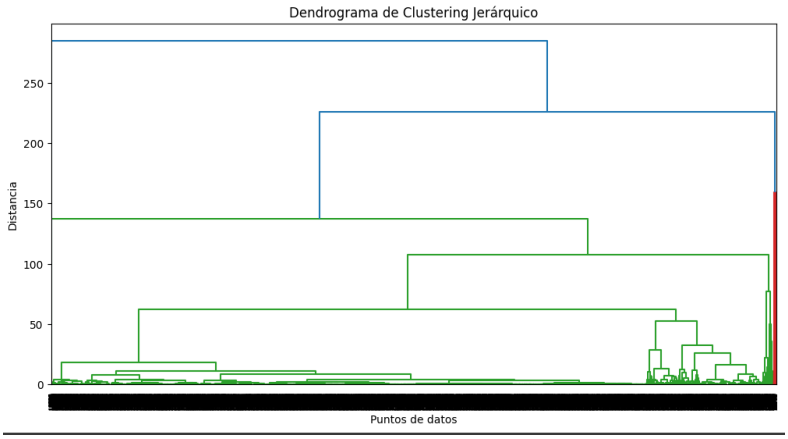


Figura 34 Dendograma con 3 clúster

[2] Un dendograma es un diagrama en forma de árbol que visualiza las relaciones jerárquicas entre objetos o grupos

El dendrograma permitió confirmar visualmente que tres agrupaciones representaban una división razonable, ya que el corte natural del árbol se encontraba en ese punto.

9.4.4. Modelado Clasificación en Importaciones

Para realizar el modelado de la clasificación, se tuvieron en cuenta 3 modelos: Radom Forest, XG Boost y LightGBM, con el objetivo de robustecer el análisis y que se pueda clasificar el segmento de importación al que pertenece.

Escogiendo como variables predictoras las siguientes:

- Valor FOB (USD)_sum_impo
- Precio Unitario FOB (USD) Cantidad _sum_impo
- Frecuencia_6M_impo
- Número de empleados
- Antigüedad
- Ratio_Pasivo_Activo

9.4.4.1. Random Forest

Para el análisis propuesto de los importadores, se implementó Random Forest con el objetivo de predecir automáticamente el segmento al que pertenece cada empresa, tomando como variable objetivo las etiquetas generadas por el modelo de segmentación escogida (K-Means).

Al igual que los modelos de clasificación para exportación, los 3 modelos mencionados usados para importaciones se entrenaron y probaron con la siguiente proporción 70/30. Su desempeño se evaluó mediante la precisión global (accuracy), la matriz de confusión, y un reporte de clasificación detallado por clase (precisión, recall, F1-score).

- **Accuracy:** se alcanzó una precisión global significativa, respaldando la capacidad del modelo para capturar patrones relevantes en los datos.
- **Matriz de confusión:** se observó un alto grado de acierto en el clúster dominante, y un rendimiento aceptable en los otros clústers.

En resumen, el modelo demostró ser sólido para predecir la pertenencia de nuevas empresas importadoras a los segmentos identificados. Aunque su rendimiento fue mayor en el grupo más representado, también mostró un comportamiento aceptable en los clústeres minoritarios.

9.4.4.2. XGBoost

El XGBoost utiliza árboles de decisión potenciados por gradiente, un algoritmo de boosting del aprendizaje supervisado que hace uso del descenso por gradiente, es una versión optimizada del Gradient Boosting, conocido por su velocidad, eficacia y capacidad para escalar bien con grandes conjuntos de datos.

En términos generales el XGBoost se comportó como un clasificador altamente competitivo, superando en algunos aspectos al Random Forest, especialmente en su

capacidad para manejar relaciones no lineales entre variables y ajustar mejor a las características del conjunto de datos, respaldado principalmente por su desempeño en el F1. Score, ya que el XGBoost tiene mejor desempeño en clústeres minoritarios, se está ajustando mejor a la estructura de los datos.

En cuanto al Accuracy y la Matriz de confusión; alcanzó un buen nivel de exactitud en el conjunto de prueba y demostró un desempeño sólido en el clúster predominante y una clasificación aceptable en los clústeres minoritario, lo cual lo hace una alternativa óptima para el despliegue operativo.

9.4.4.3. LightGBM

LightGBM es un algoritmo de refuerzo de gradientes basado en modelos de árboles de decisión. Puede ser utilizado para la categorización, clasificación y demás tareas de aprendizaje automático, en las que es necesario maximizar o minimizar una función objetivo, consiste en combinar clasificadores sencillos, como árboles de decisión de profundidad limitada; se destaca por ser veloz en su entrenamiento, por su mayor precisión y soporte de aprendizaje paralelo y soporte para GPUs (Universitat Oberta de Catalunya, 2025).

Al igual que en todos los modelos utilizados, las variables predictoras son indicadores operativos y financieros seleccionados y la división de datos es la misma: 70% para entrenamiento y 30% para prueba, con estratificación por clúster.

En cuanto a su desempeño, alcanzó una precisión destacada al clasificar correctamente una alta proporción de las empresas en sus respectivos clústers; de igual forma, la matriz de confusión demostró una buena ejecución, con errores reducidos y distribuidos, lo que indica una clasificación más balanceada. Para el reporte de clasificación, se generaron reportes por cada clase, donde se observa un rendimiento competitivo, con capacidad para clasificar correctamente tanto los clústers más representados como los menos frecuentes.

En conclusión, cada modelo fue entrenado con las mismas variables númericas predictoras y evaluado bajo el mismo estandar: precisión global (accuracy), matriz de confusión, precision, recall y F1-score por clase. De cada modelo se puede destacar:

- R**andom Forest**: robustez, estabilidad y facilidad de interpretación, ofreciendo una clasificación efectiva en el clúster mayoritario.
- **XGBoost:** mayor capacidad para ajustarse a patrones complejos y no lineales, lo cual se reflejó en un mejor desempeño en los clústers menos representados.
- **LightGBM**: combinó ambos enfoques, sobresaliendo por su eficiencia y precisión.

9.4.5. Modelado de Segmentación de Importaciones y Exportaciones

Para llevar a cabo el análisis de las empresas que realizan ambas actividades alrededor de 4577 registros en esta base, las variables numéricas fueron escaladas por el StandardScaler() con el fin de efectuar el desarrollo de los modelos de segmentación.

La identificación de la estructura optima de agrupamiento, en esta base se realizó a través de una evaluación comparativa entre múltiples algoritmos, demostrando otro método para agrupar y evaluar los modelos propuestos, considerando un análisis de componentes principales con 4 componentes principales que explican más del 75% de la información original.

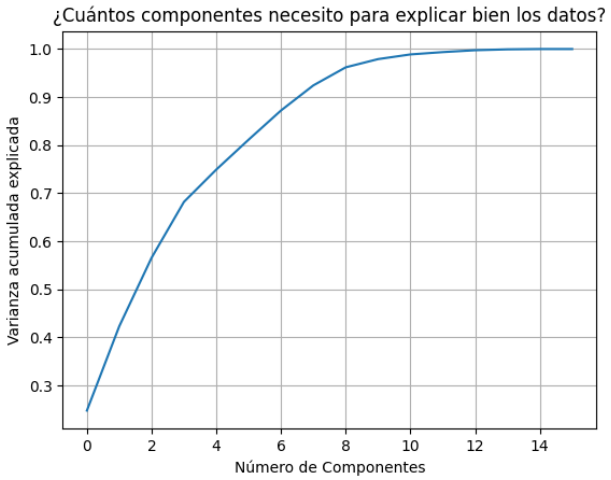


Figura 35 Número de componentes principales

Se consideraron métodos clásicos y como métodos avanzados, por ejemplo:

- **K-Means** con 3, 4 y 5 clústeres.
- **Gaussian Mixture Models (GMM)** con 3, 4 y 5 componentes.
- **Fuzzy C-Means**, que permite asignación parcial de pertenencia a múltiples clústeres.
- **DBSCAN**, que agrupa según densidad y detecta automáticamente outliers.

- **Mean Shift**, que identifica clústeres sin necesidad de definir su número previamente.
- **Spectral Clustering**, basado en teoría de grafos, con 3, 4 y 5 grupos.
- **Deep Embedded Clustering (DEC)**: una combinación de autoencoders con GMM sobre el espacio latente, evaluado también con 3, 4 y 5 clústeres.

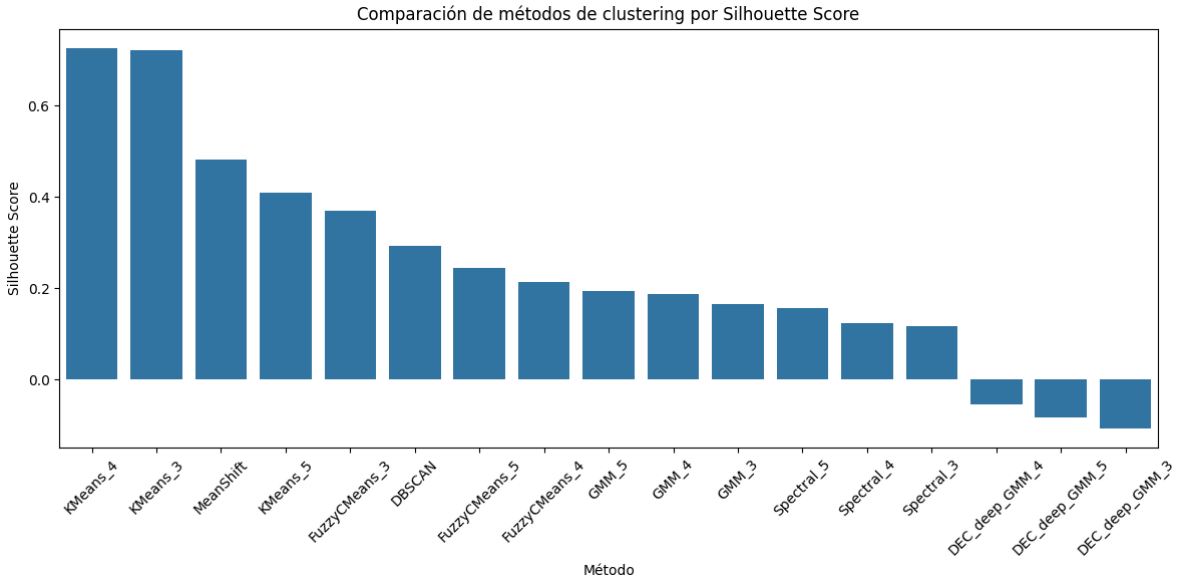


Figura 36 Modelos de segmentación importación y exportación

Esta comparación permitió identificar qué modelos ofrecen un mejor ajuste a la estructura real de los datos, facilitando la toma de decisiones sobre qué técnica utilizar como base para posteriores análisis de clasificación y caracterización de empresas.

9.4.6. Modelado de Clasificación de Importaciones y Exportaciones

Luego de la revisión de múltiplos modelos de segmentación, se evaluaron 3 modelos de clasificación supervisada, con el mismo objetivo de las otras 2 bases

mencionadas (importaciones o exportaciones). Con base en las etiquetas generadas por el modelo de segmentación seleccionado (K- Means), se entrenaron modelos de clasificación para automatizar la asignación de nuevas empresas a los segmentos definidos. Los algoritmos utilizados fueron Random Forest, XGBoost y K- Nearest Neighbors (KNN).

Las variables predictoras para todos los modelos son:

- Valor FOB (USD)_sum_impo, Valor FOB (USD)_sum_expo
- Precio Unitario FOB (USD) Cantidad _sum_impo
- Precio Unitario FOB (USD) Cantidad_sum_expo
- Frecuencia_impo
- Frecuencia_expo
- Año Fiscal
- Antigüedad
- Número de empleados
- Ratio_Pasivo_Activo

9.4.6.1. Random Forest

Al igual que en otros casos, el modelo se entrenó y testeo, 70/30 respectivamente, sin necesidad de escalamiento previo gracias a la naturaleza del algoritmo. En cuanto a los criterios de desempeño ya mencionados en anteriores modelos, el accuracy en este modelo se alcanzó un rendimiento sólido, destacándose especialmente en el clúster mayoritario. En cuanto a la matriz de confusión: permitió observar el comportamiento por clase, mostrando

alta precisión en la clase dominante y niveles aceptables de precisión en las clases menos representadas

En conclusión, este algoritmo demostró ser una herramienta robusta y efectiva para la clasificación multiclase en este contexto.

9.4.6.2. XGBoost

Este modelo demostró ser un modelo muy competitivo, ya que tiene la capacidad de ajustarse a relaciones complejas entre las variables y así mismo mantener el rendimiento balanceado, en cuanto al accuracy, demostró una alta precisión, superior al Random Forest, la matriz de confusión confirma una clasificación acertada en las clases con menor error relativo en los clústeres menos frecuentes. La curva ROC por clase: muestran un AUC elevados en las tres clases, lo cual indica una fuerte capacidad de discriminación entre segmentos.

9.4.6.3. KNN (K- Nearest Neighbors)

El algoritmo de k-nearest neighbors (KNN) es un clasificador de aprendizaje supervisado, que emplea la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual, partiendo del supuesto de que se pueden encontrar puntos similares cerca uno del otro (IBM, 2025).

Figura 37 Modelo KNN

Para la ejecución del modelo, se escalaron los datos debido a su sensibilidad en términos de distancia de las variables y la escala de estas, sin embargo, fue entrenado 70/30 como los modelos anteriores.

En términos generales, el modelo KNN ofreció un desempeño razonable en tareas de clasificación multiclase, destacándose por su simplicidad y fácil interpretación. Sin embargo, su sensibilidad al escalado de variables y al desbalance entre clases lo hicieron menos robusto en comparación con Random Forest y XGBoost.

En el accuracy se logró una precisión global aceptable, pero menor a los otros dos modelos con los que se comparó. En la matriz de confusión se tuvo una mayor proporción de errores en los clústeres menos representativos, debido a que depende de la densidad local. Para el reporte de clasificación se evidenció precisión y recall moderadas, pero un F1 Score menor al XGBoost y Random Forest. La curva ROC se identificó discriminación aceptable, aunque más limitada que los modelos de árboles.

En conclusión, cada modelo se destacó de alguna forma con las siguientes características:

- **Random Forest** robustez, buena generalización y facilidad de interpretación.
- **XGBoost** mostró el mejor equilibrio entre precisión y cobertura, particularmente en los clústeres minoritarios, considerándolo como el modelo más competitivo del conjunto.
- **KNN**, aunque más simple, ofreció una clasificación aceptable pero más sensible al desbalance y a la escala de los datos, lo que limitó su rendimiento frente a los otros dos modelos.

9.5. Evaluación del Modelo

Una vez ajustados los modelos, su desempeño fue evaluado a través de métricas estándar para cada uno de los modelos de segmentación y clasificación

9.5.1. Evaluación Modelos de Segmentación

Las métricas de evaluación para los modelos de segmentación se basaron principalmente en métricas internas, especialmente en el Silhouette Score, sin embargo, también se consideraron 2 más: Davies-Bouldin Index y Calinski-Harabasz Score. Estas métricas permiten evaluar qué tan bien estructurados están los clústers generados, sin necesidad de contar con etiquetas reales.

Silhouette Score: evalúa la calidad de la agrupación en la informática. Mide la coherencia de los clústeres, con un coeficiente más elevado que indica grupos más coherentes. El coeficiente oscila entre -1 y 1, con valores cercanos a 1 que indican que una

muestra está lejos de los cúmulos vecinos, y valores negativos que sugieren que las muestras pueden haber sido asignadas al grupo equivocado (ScienceDirect, 2021).

El coeficiente se calcula sobre la base de la cohesión de los grupos y la separación de grupos, que representan las distancias medias entre instancias y puntos de datos dentro y entre clústeres, respectivamente (ScienceDirect, 2021).

Davies-Bouldin Index: es una métrica para evaluar la calidad de los clústeres, donde el índice es que una agrupación de calidad debería producir clústeres separados y compactos. Se basa en relacionar la dispersión dentro de los clústeres y la separación entre clústeres (Rodriguez, 2023).

Por un lado, la dispersión intra-clúster mide la separación de los puntos dentro de cada clúster, un resultado bajo indica que los puntos dentro de un grupo están muy cercanos entre sí, algo que es deseable en un buen clustering (Rodriguez, 2023).

Y, la dispersión inter-clúster mide la separación entre los grupos, si esta medida es alta significa que los grupos están muy alejados entre sí, lo que también es deseable en un buen clustering (Rodriguez, 2023).

En conclusión, el índice de Davies-Bouldin se construye como el cociente de ambos valores. Por lo que cuando los clústeres están separados y son compactos el valor de este índice se minimiza.

Calinski-Harabasz Score: es una medida de evaluación de algoritmos de agrupamiento. Se utiliza comúnmente para evaluar la bondad de la división mediante un algoritmo de agrupamiento de K- Means para un número determinado de conglomerados, se calcula como la relación de la suma de la dispersión entre clúster y la suma de la dispersión dentro de los conglomerados para todos los clústeres, un valor alto indica que los clústeres están bien definidos: compactos internamente y separados entre sí (Towards Data Science, 2022).

En resumen, la interpretación es la siguiente:

Tabla 18 Interpretación métricas de evaluación modelos de segmentación

Silhouette Score	Alto para un mejor agrupamiento global
Davies-Bouldin Index	Bajo para un menor solapamiento entre los clústeres.
Calinski-Harabasz Score	Alto para una buena separación entre los grupos.

9.5.1.1. Segmentación exportaciones

Para el caso de empresas exclusivamente exportadoras, el modelo Gaussian Mixture Model presentó un mejor desempeño, con una segmentación de 2 clúster, sin embargo, para efectos del análisis no es relevante contar con 2 clúster, por tal motivo se selecciona el siguiente mejor, con desempeño destacable dentro de la evaluación de modelos, logrando un excelente balance entre cohesión y separación.

Tabla 19 Desempeño modelos segmentación exportaciones

Modelo	Silhouette	Davies Bouldin	Calinski Harbasz	Número clúster
K- Means	0.75	0.52	481787.35	4
GMM	0.99	0.00	290358.87	2
Birch	0.95	0.28	2111685.33	4

Los clústeres obtenidos fueron los siguientes:

Tabla 20 Clusters Birch

Cluster	Nombre	Cuenta	Características	Resumen
0	**Exportadores ágiles**	79	Alta frecuencia, buen volumen, pequeñas pero muy activas.	Empresas consolidadas con alto volumen y frecuencia exportadora, pero de tamaño pequeño.
1	**Exportadores premium**	16	Bajo volumen, pero precios unitarios altísimos. Envíos selectos y exclusivos.	Exportaciones esporádicas de alto valor unitario. Posibles productos premium o maquinaria.
2	**Exportadores constantes**	15115	Gran mayoría, actividad regular, desempeño medio y estable.	Empresas medianas con actividad exportadora frecuente pero moderada.
3	**Exportadores potentes**	7	Altísimos activos y patrimonio, aunque exportan poco. Empresas con gran capacidad.	Empresas grandes que exportan poco pero con gran capacidad financiera. Posible perfil institucional.

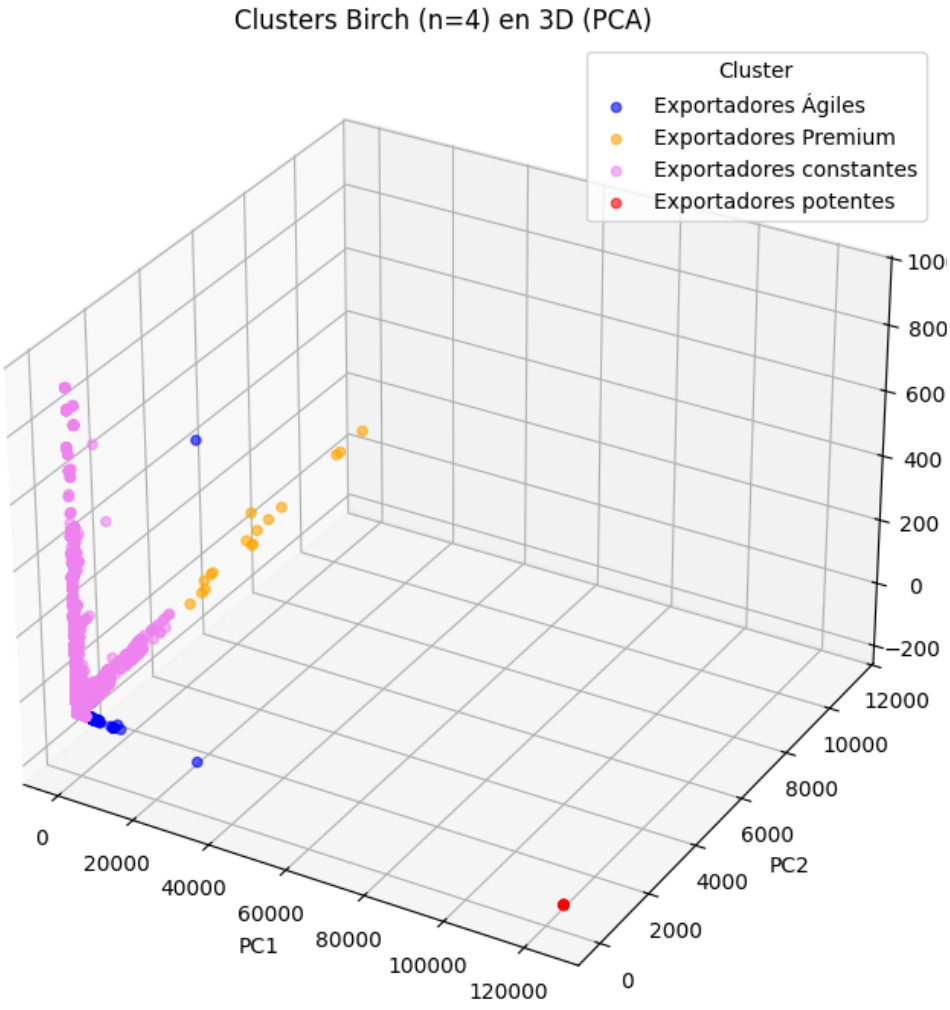


Figura 38 Segmentación 3D modelo Birch

Ventajas del modelo escogido (Birch):

- Diseñado para trabajar con grandes volúmenes de datos.
- Realiza una segmentación progresiva, es decir, permite construir subclusters de manera incremental.

- Tiene un mejor rendimiento cuando los clústeres son de manera irregular o de tamaño desigual.

En general, los 4 clúster que se identificaron reflejan perfiles de empresas exportadoras diferenciados, cada clúster destaca por una característica en específico, permitiendo crear estrategias más centradas y personalizadas para cada tipo de clúster.

9.5.1.2.Segmentación importaciones

Para el siguiente segmento de empresas, relacionado a las empresas que realizan importaciones, se evaluaron variantes del K- Means, DBScan, Spectral Clustering y Clustering Jerarquico, teniendo como resultado de la comparación que el modelo K-Means con 3 clúster es el de mejor rendimiento, con resultados consistentes y estables, por el contrario, Spectral Clustering y DBSCAN mostraron resultados más bajos.

Tabla 21 Desempeño modelos segmentación importaciones

Modelo	Silhouette	Número clúster
K- Means	0.967	3
K- Means	0.964	4
K- Means	0.951	5
DBScan	0.897	2
Spectral	0.435	3
Jerárquico	0.966	3

Los clúster obtenidos fueron los siguientes:

Tabla 22 Clusters K= 3 importaciones

Cluster	Nombre	Cuenta	Características	Resumen
0	**Importador MiPymes**	109853	Bajo valor FOB, baja frecuencia, pocos empleados, bajos activos y patrimonio.	Empresas con baja frecuencia y valor de importación. Principalmente micro y pequeñas empresas
1	**Importador constante**	359	Alta frecuencia, valor FOB medio, activos y empleos moderados.	Empresas con alta frecuencia de importación y operación sostenida. Importadores regulares.
2	**Importador Premium**	28	Valor FOB elevado, grandes activos, alta carga laboral, pero frecuencia menor que constante.	Empresas grandes con alta capacidad financiera y operativa. Importadores estratégicos.

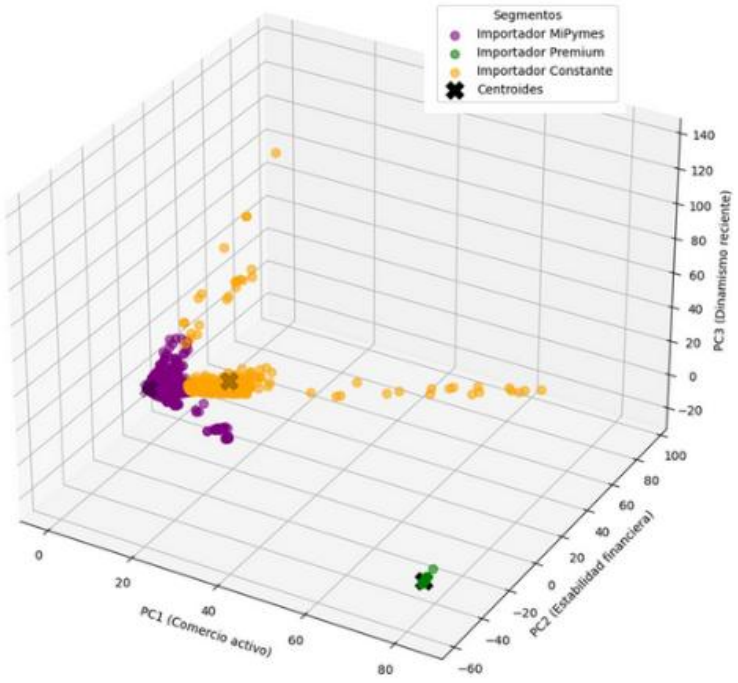


Figura 39 Segmentación 3D modelo K- Means K=3 importaciones

Al igual que en la culturización de exportaciones, se reflejan 3 perfiles de empresas importadoras diferenciados, el grupo más numeroso es el Importadores MiPymes, el grupo de constantes importadores agrupa empresas con importaciones sostenidas y el clúster premium demuestran importadores estratégicos con alta carga en importaciones y recursos significativos. De tal forma que se facilita la creación de estrategias diferenciadas para cada uno.

9.5.1.3.Segmentación importaciones y exportaciones

En el caso de empresas con operaciones mixtas, demostró que el K-Means con 3 clúster tiene el mejor desempeño, aunque con métricas generales más bajas que en los casos anteriores. Esto sugiere que la estructura interna de estos datos es más compleja. Otros modelos evaluados como el GMM, Spectral y DBSCAN presentaron resultados más bajos, especialmente en: Silhouette Score y Calinski-Harabasz.

Tabla 23 Desempeño modelos de segmentación importación y exportación

Modelo	Silhouette	Davies Bouldin	Calinski Harbasz	Número clúster
K- Means	0.63	1.04	1004.39	3
GMM	0.13	2.49	477.10	4
Spectral	0.01	1.52	56.20	4
DBScan	-0.10	1.79	18.77	77

Los clústeres obtenidos fueron los siguientes:

Tabla 24 Cluster K= 3 importaciones y exportaciones

Cluster	Nombre	Cuenta	Características	Resumen
0	**Exportadores activos**	130	Empresas con participación	Empresas exportadoras recientes con crecimiento.

			exportadora marcada, estables económicamente.	
1	**Empresas mixtas pequeñas**	4382	Valores FOB bajos en promedio, pero alta varianza	Empresas de tamaño pequeño a mediano, algunas con comportamientos extremos.
2	**Grandes importadores**	65	Precio unitario promedio, con frecuencias de importación alta y bien establecidas.	Empresas grandes y maduras con alto volumen, pero con estructura financiera riesgosa

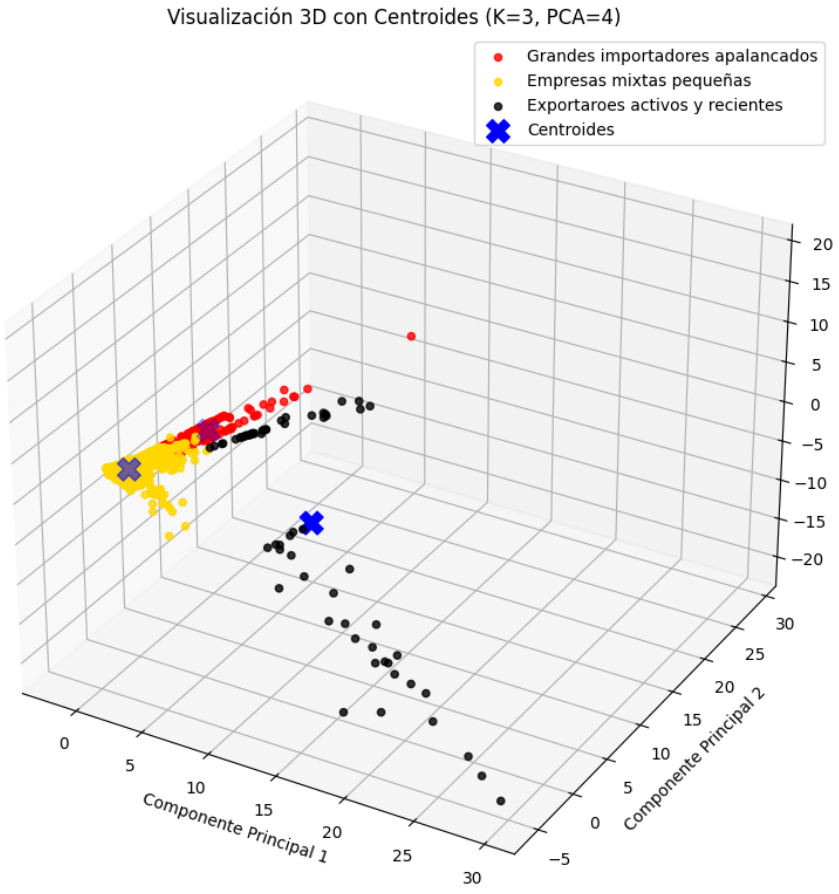


Figura 40 Segmentación 3D modelo K- Means K=3 importaciones y exportaciones

La segmentación para este grupo permite diferenciar tres perfiles de empresas con operaciones mixtas. Las empresas exportadoras activas se destacan por su estabilidad económica y dinamismo, las empresas mixtas pequeñas presentan gran diversidad en su comportamiento, y los grandes importadores concentran operaciones de alto volumen.

Ventajas del modelo K- Means aplicado y escogido en las empresas que solo realizan importaciones y empresas con la actividad mixta:

- Simplicidad, rapidez y de fácil implementación.
- Adecuado para grandes volúmenes de datos.
- Genera clúster con fronteras definidas y facilitando la interpretación visual.

9.5.2. Evaluación Modelos de Clasificación

Las métricas para evaluar un modelo de clasificación se usan para explicar su rendimiento, donde se comparan los valores de pruebas del conjunto de datos seleccionados como pruebas, con el fin de calcular su precisión y así poder mejorar el modelo.

9.5.2.1. Métricas de Desempeño de Modelos de Clasificación

Matriz de Confusión: herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado, permitiendo ver los aciertos y errores que está teniendo el modelo para aprender sobre los datos (IBM, 2025).

Tabla 25 Ejemplo matriz de confusión binomial

Valores predicción		
	Verdaderos positivos (VP)	Falsos negativos
	Falsos positivos	Verdaderos negativos

Valores reales

- **Verdaderos positivos:** número de predicciones correctas para la clase positiva.
- **Falsos positivos:** casos negativos identificados como positivos incorrectamente.
- **Falsos negativos:** casos positivos reales pero predichos como negativos.
- **Verdaderos negativos:** número de predicciones negativas con precisión negativa.

Con relación a los resultados de los modelos para cada uno de los conjuntos de datos, se obtuvo una matriz de confusión, donde cada fila representa un clúster [0, 4] para exportaciones y [0,3] para importaciones, la diagonal son los valores correctos y los valores fuera de esta son los errores que tuvo el modelo.

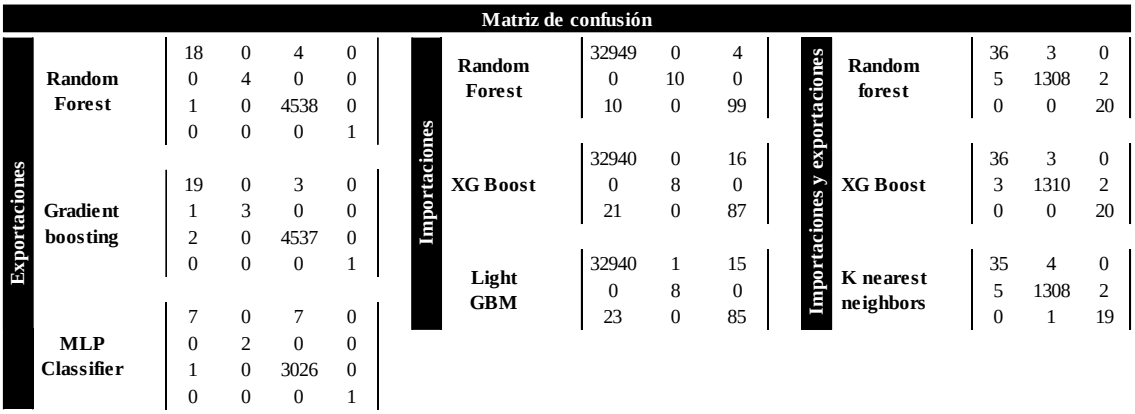


Figura 41 Matriz de confusión para cada modelo de clasificación

Para los modelos del conjunto de datos de empresas exportadoras, el Random Forest y Gradient Boosting predicen muy bien, tienen un buen rendimiento, sin embrago el desempeño del MLP Classifier tiene confusión entre clases, teniendo el peor desempeño de los 3, en otras palabras, el Random Forest y Gradient Boosting son precisos; MLP falla al distinguir clases.

En los modelos de importaciones, el Random Forest tiene un desempeño casi perfecto con 2 errores confundidos como clúster 0, es decir el segmento importador MiPymes. El XG Boost tiene 21 errores clasificados como clúster 0 y el Light GBM tiene buen desempeño, pero tiene 23 errores al clasificar en el clúster 2 (importador premium). En conclusión, el Random Forest tiene el mejor desempeño; los otros dos son competitivos, pero cometen más errores con la clase menos representada.

Por último, los modelos para el conjunto de datos que contiene la información de empresas que hacen ambas actividades: el Random Forest es bueno, sin embargo, falla al

predecir el clúster 2 conocido como grandes importadores (20 predichos como clúster 1), el XG Boost tiene mejor detección en el clúster 2 y el K- nearest neighbors tiene el peor desempeño relativo y falla con el clúster 0, exportadores activos. Es decir, que el XGBoost tiene un mejor desempeño en detección de todas las clases y el KNN es menos preciso.

Con relación a las demás métricas relevantes para evaluar los modelos de clasificación (Alce, 2019), se consideraron las siguientes:

Precisión: se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión.

Recall (sensibilidad): la capacidad del modelo para discriminar los casos positivos, de los negativos.

F1 Score: resume la precisión y sensibilidad en una sola métrica. Por ello es de gran utilidad cuando la distribución de las clases es desigual.

Exportaciones												
Modelos	**Random Forest**				**Gradient boosting**				**MLP Classifier**			
Clusters	0	1	2	3	0	1	2	3	0	1	2	3
Precisión	0,95	1	1	1	0,86	1	1	1	0,88	1	1	1
Recall	0,82	1	1	1	0,86	0,75	1	1	0,5	1	1	1
F1 Score	0,88	1	1	1	0,86	0,86	1	1	0,64	1	1	1

Importaciones									
Modelos	**Random Forest**			**XG Boost**			**Light GBM**		
Clusters	0	1	2	0	1	2	0	1	2
Precisión	1	1	0,96	1	1	0,84	1	1	0,85
Recall	1	1	0,91	1	1	0,81	1	1	0,79
F1 Score	1	1	0,93	1	1	0,82	1	0,94	0,82

Importaciones y exportaciones									
Modelos	**Random Forest**			**XG Boost**			**KNN**		
Clusters	0	1	2	0	1	2	0	1	2
Precisión	0,87	0,99	0,9	0,92	0,99	0,9	0,87	0,99	0,9
Recall	0,92	0,99	1	0,92	0,99	1	0,89	0,99	0,95
F1 Score	0,9	0,99	0,95	0,92	0,99	0,95	0,88	0,99	0,92

Figura 42 Valores de precisión, recall y F1 Score para cada modelo

Para exportaciones, se pueden obtener las siguientes conclusiones, considerando al Random Forest es el modelo más robusto para el conjunto de exportadores.

- Random Forest presentó el mejor desempeño global, con un F1-score promedio de 0.88, y una alta precisión en el clúster dominante (95%).
- Gradient Boosting obtuvo un desempeño muy cercano (F1 = 0.86), con menor recall en el clúster 1 (75%), indicando cierta dificultad para detectar correctamente ese grupo.
- MLP Classifier mostró el rendimiento más bajo (F1 = 0.64), con valores simétricos entre clases, pero sin capacidad para distinguir adecuadamente los clústeres.

Para importaciones, todos los modelos demostraron un buen desempeño, el LightGBM fue el modelo con mejor equilibrio en F1 Score en todos los clústeres un promedio de 0.94, por otro lado, el Random Forest destacó en la métrica de recall para los clúster 1 y 2, importador constante e importador premium, respectivamente. Por último, el XGBoost tuvo un rendimiento inferior, afectado por un recall más bajo en el clúster 1 (importador constante).

En el conjunto de datos de importaciones y exportaciones (empresas que realizan ambas actividades), todos los modelos obtuvieron rendimientos notables, con F1 Score promedio superiores al 0.92. El modelo KNN obtuvo resultados competitivos con un F1 = 0.92, sin embargo, su desempeño fue más bajo en clúster 2. El Random Forest y XGBoost empataron con un desempeño excelente, pero el XGBoots tiene un promedio mejor en la métrica F1 Score con 0.95.

Por otro lado, también se evaluó el Accuracy de cada modelo, donde las empresas dedicadas a exportaciones o importaciones (solo una de las dos actividades) tuvieron una exactitud perfecta, reflejando que los clúster generados mediante la segmentación están diferenciados y consistentes, facilitando la predicción de los modelos de clasificación, por otra parte, los valores para los modelos de empresas que realizan las dos actividades, descendiendo a 0.99, indicando que los datos en este conjunto son más complejos, con algunos solapamientos entre clases o variabilidad en el comportamiento, sin embargo y en resumen, los 9 modelos evaluados mantienen un buen rendimiento.

Tabla 26 Comparación Accuracy de todos los modelos

Modelo	Accuracy
Exportaciones	
Random Forest	1
Gradient Boosting	1
MLP Classifier	1
Importaciones	
Random Forest	1
XG Boost	1
Light GBM	1
Importaciones y exportaciones	
Random Forest	0.99
XG Boost	0.99
KNN	0.99

Los modelos tienen una gran capacidad para generalizar y predecir correctamente los clústeres definidos, ya que están claramente separados y definidos (Google for developers, 2025).

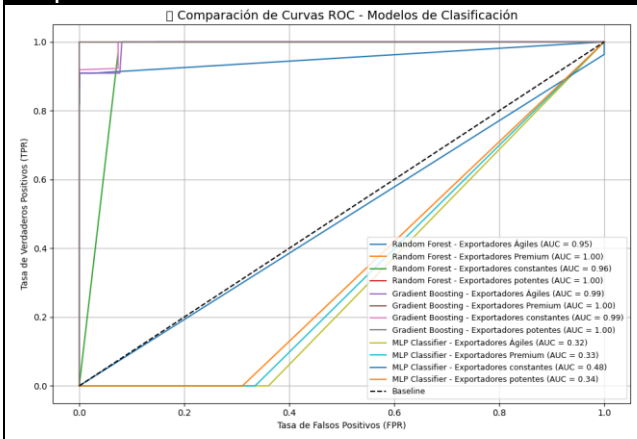
9.5.2.2. ROC y AUC

El área under the line (AUC por sus siglas en inglés) o el Receiver Operating Characteristic Curve (ROC por sus siglas en inglés) es una métrica robusta que evalúa la bondad del modelo en todo el espectro de umbrales, es una métrica de acceso para evaluar el rendimiento de los modelos de clasificación binaria (Data Camp, 2024).

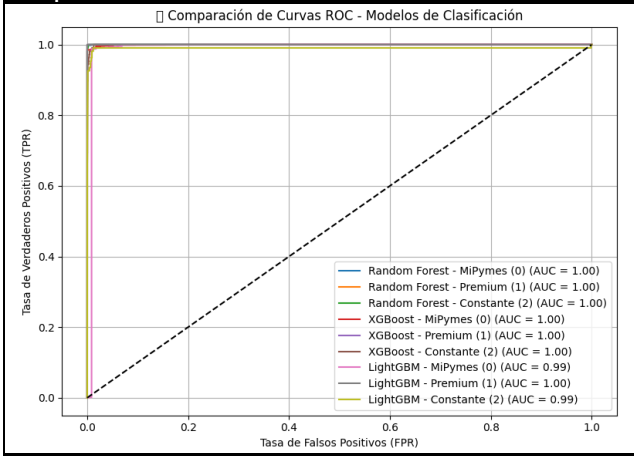
La curva ROC ofrece una representación visual de las compensaciones entre los verdaderos positivos y los falsos negativos en varios umbrales, proporcionando información sobre lo bien que el modelo puede equilibrarlas (Data Camp, 2024).

Tabla 27 ROC para cada conjunto de datos

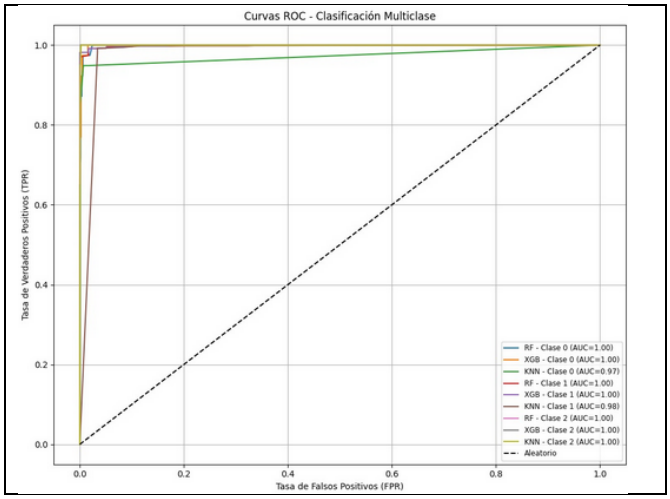
Exportación



Importación



Importaciones y exportaciones



En el caso de empresas únicamente exportadoras o importadoras, los valores de AUC alcanzaron valores como 1, lo que indica una excelente capacidad de discriminación entre los diferentes grupos, confirmando que los clústeres obtenidos en el proceso de segmentación son altamente separables, y que los modelos son capaces de identificar con precisión las categorías obtenidas en estos dos grandes segmentos (7 clúster en total). Para el caso de empresas que realizan las dos actividades, la complejidad aumenta debido a la mayor diversificación, sin embargo, las curvas ROC muestran resultados sólidos en todos los modelos, con valores AUC entre 0.99 y 1.00 para las tres clases, conservando un alto nivel de rendimiento diferenciar entre exportadores activos, empresas mixtas pequeñas y grandes importadores.

9.5.2.3. Validación Cruzada

La validación cruzada es un método para garantizar que los modelos no solo funcionen bien con datos de entrenamiento, sino también con nuevos datos (Scikit Learn, 2025).

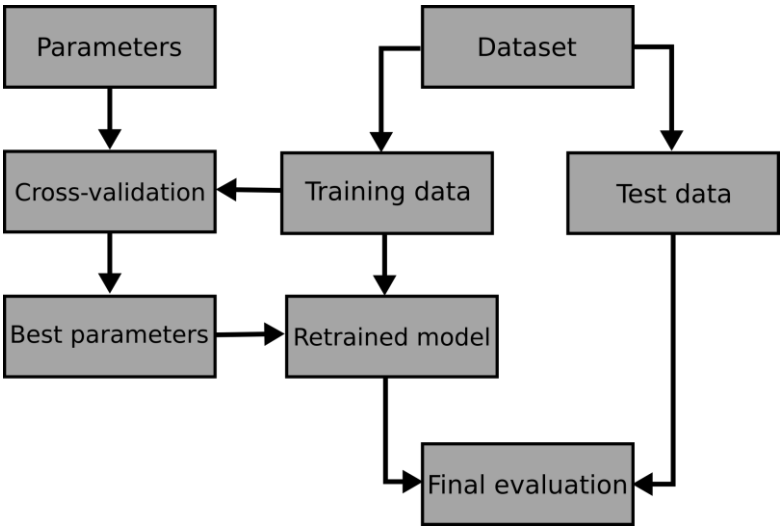


Figura 43 Flujo de trabajo de validación cruzada

Donde se divide el conjunto de datos en varios subconjuntos y se entra el modelo varias veces. Este método se llevó a cabo en cada uno de los conjuntos de datos, concluyendo que:

En el caso de exportaciones, el random forest se destaca como el modelo con mejor equilibrio entre precisión, sensibilidad y F1-score, sin embargo, el modelo MLP presenta una alta accuracy, sus métricas macro indican que no logró predecir correctamente todas las clases. Para las importaciones, todos los modelos alcanzaron niveles muy buenos de

precisión, por medio de la desviación, el Random Forest indica que es el mejor, el XGBoost también es muy competitivo y LightGBM, tiene un rendimiento inferior ya que durante el entrenamiento algunos árboles no fueron posibles de mejorar la predicción. Para ambas actividades económicas, todos los modelos obtuvieron resultados sobresalientes, sin embargo, el KNN presentó un mejor rendimiento promedio y la menor variabilidad, pero el escogido finalmente por su rendimiento general fue el XG Boost.

En conclusión, los modelos escogidos como clasificadores son:

Tabla 28 Modelos clasificadores escogidos

Exportaciones	Random Forest	Robusto y consistente, adaptandose bien estructuras complejas.
Importaciones		
Importaciones y exportaciones	XG Boost	Rendimiento altamente competitivo.

9.6. Despliegue e Implementación

Luego del cumplimiento satisfactorio de las 5 fases anteriores al despliegue, se procedió a implementar los modelos de tal forma que su funcionamiento pudiera aplicar para nuevos datos.

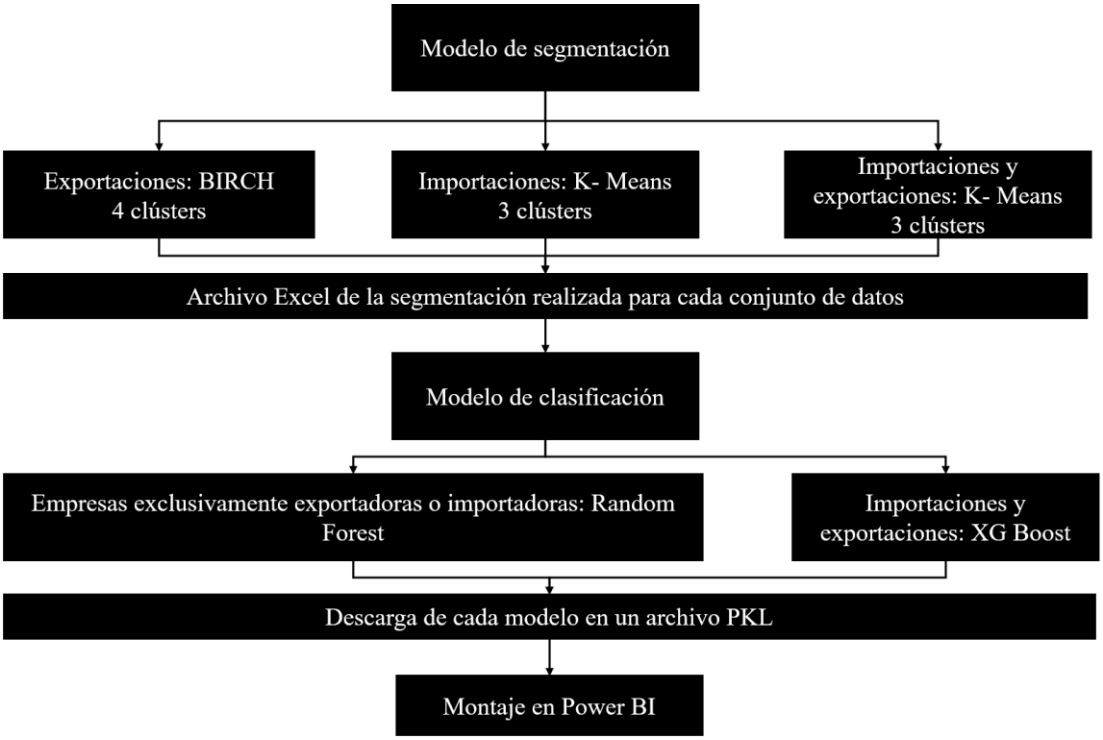


Figura 44 Flujo para el despliegue de los modelos

9.6.1. Desarrollo de Power BI

Con la base final obtenida luego de la preparación de los datos, los archivos con la indicación de cada segmento y los modelos guardados en PKL, se pasó a desarrollar la visualización en Power BI, de tal forma que se tuviera información tanto descriptiva como predictiva.

Se presenta una página inicial con una visión general del alcance del análisis empresarial para BritCham:

Afiliados actuales: Se identifican 75 empresas afiliadas activas a la Cámara.

Empresas analizadas: Se trabajó con una base de datos robusta que incluye más de mil empresas, cubriendo tanto exportadores, importadores y empresas mixtas.

Clientes segmentados: Se logró construir los segmentos empresariales con características diferenciadas, resultado del modelado de segmentación aplicado sobre las empresas activas en comercio exterior.

Figura 45 Panel 1 Power BI

En el segundo panel, se indica información relevante de los 75 afiliados activos a BritCham:

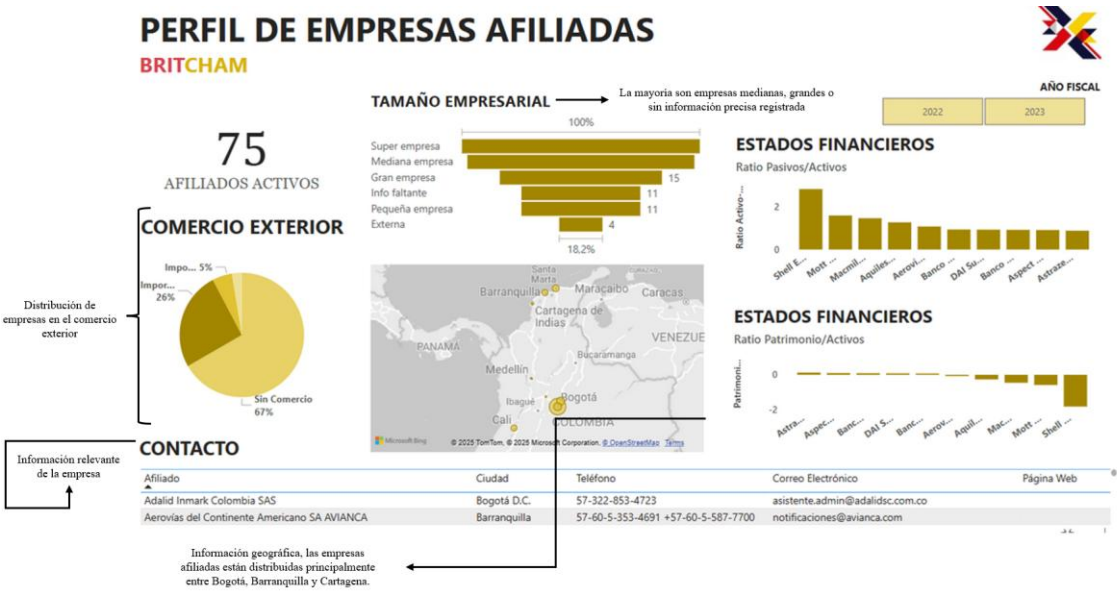


Figura 46 Panel 2 Power BI

En el tercer panel, se representa la fase de analítica descriptiva del tablero, integrando datos históricos, junto con resultados del modelado de empresas segmentadas por cada conjunto.

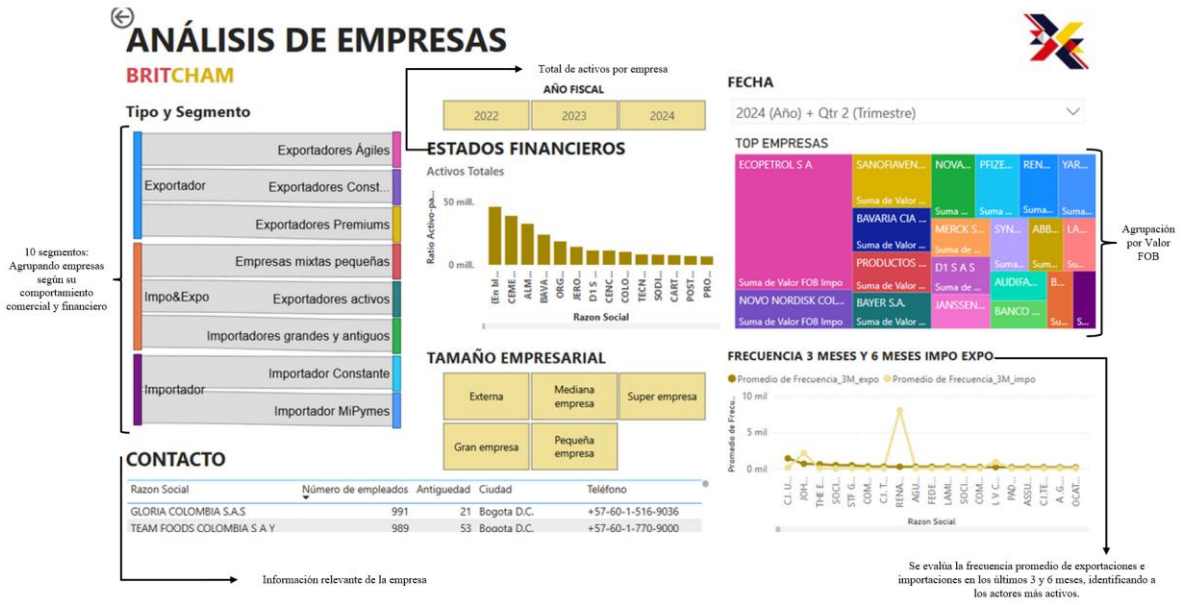


Figura 47 Panel 3 Power BI

Por último, en el panel número 4 se encuentra la información de nuevas empresas, de tal forma que BritCham no solo tiene la opción de ver los clúster asignados a las empresas ya tratadas dentro de la información de importaciones y exportaciones, sino que también se podrá identificar nuevas empresas, donde teniendo información de las variables del modelo presenta en el panel el clúster asignado según sus características y las recomendaciones según el clúster, junto con la posible suscripción basándose en los valores públicados por BritCham dependiendo de la categoría de afiliación.

Figura 48 Panel 4 Power BI

Con el fin de mantener el modelo, a continuación, se indica el flujo de información completo de la propuesta analítica desarrollada:

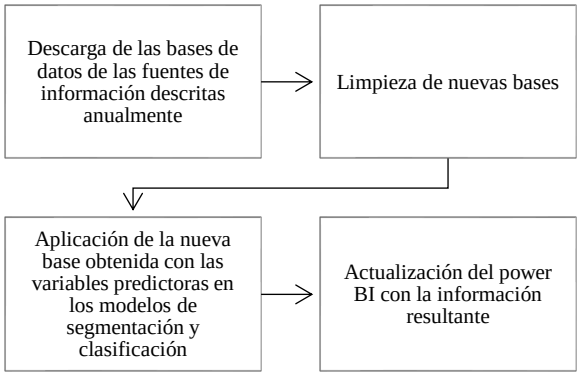


Figura 49 Flujo de actualización

9.6.2. Estrategias para cada uno de los Clústeres Generados

Considerando los diez clústeres identificados a partir de los tres grandes conjuntos de datos analizados, el equipo de analítica propone una serie de estrategias que la Cámara de Comercio Colombo-británica podría implementar para fortalecer su proceso de afiliación y oferta de servicios empresariales.

Tabla 29 Tipos de segmento para cada conjunto de datos

Exportaciones	Exportadores ágiles	79
	Exportadores premium	16
	Exportadores constantes	15.115
	Exportadores potentes	7
Importaciones	Importador MiPymes	109.853
	Importador constante	359
	Importador Premium	28
Importaciones y exportaciones	Exportadores activos	130
	Empresas mixtas pequeñas	4.382
	Grandes importadores	65

Dado que la mayoría de las empresas se concentran en los clústeres de exportadores constantes, importadores MiPymes y empresas mixtas pequeñas, se recomienda enfocar los esfuerzos en captar empresas de los segmentos menos representados. Estos grupos, aunque minoritarios, presentan características estratégicas que podrían beneficiar significativamente los objetivos de la Cámara si se vinculan adecuadamente.

La identificación de perfiles diferenciados permite no solo atraer nuevos afiliados, sino también diseñar servicios más ajustados a las necesidades de cada tipo de empresa. A

partir de las estrategias propuestas, se busca potenciar su internacionalización, fortalecer la conexión con el Reino Unido y contribuir a una operación más sostenible y competitiva.

Para el caso de los exportadores, se reconocen cuatro perfiles clave:

Exportadores Ágiles que se caracterizan por su alta frecuencia de envíos y un volumen significativo pese a su tamaño reducido.

Estrategia: ofrecer un portafolio de servicios centrado en programas de escalamiento, buscando llevar a un nuevo nivel la competitividad y volumen de exportación por medio algunos aspectos clave de diagnóstico de capacidad exportadora, talleres de mercado y formación en inglés comercial y técnico, brindando asesoría y consultoría en este campo, junto con el acceso a redes comerciales y espacios de networking con compradores británicos.

Exportadores Premium con envíos de bajo volumen, pero alto valor unitario.

Estrategia: realizar una mayor visibilidad de marca, acceso a clientes institucionales o de nicho y asistencia en procesos de certificación de calidad o normativas específicas del Reino Unido.

Exportadores Constantes que representan el grupo más numeroso y con actividad exportadora moderada pero sostenida

Estrategia: dar apoyo en formación continua, actualización normativa, permitiendo mantener la competitividad ante cambios en tratados comerciales, requisitos y/ nuevas

exigencias del mercado británico y programas de fidelización institucional, de tal forma que se fortalezca el vínculo entre la cámara y la empresa.

Exportadores Potentes a pesar de su bajo número, son empresas con gran capacidad financiera y activos elevados.

Estrategia: ofrecer membresías corporativas exclusivas, participación en foros de inversión bilateral y proyectos de responsabilidad empresarial, junto con la asesoría especializada para diversificación de mercados o eficiencia logística.

En cuanto a los importadores, se identificaron tres grupos con necesidades diferenciadas:

Empresas MiPymes importadoras, que representan el grueso del universo empresarial importador, con bajo volumen y frecuencia.

Estrategia: ofrecer servicios asequibles como cursos de formación en comercio exterior, apoyo en procesos aduaneros y acceso a catálogos traducidos o ferias virtuales de proveedores británicos o colombianos.

Los Importadores Constantes presentan operaciones sostenidas y una capacidad de importación moderada.

Estrategia: dar la oportunidad de incorporación a programas de mejora logística, como capacitaciones especializadas en optimización de la cadena de suministro, nuevas tecnologías y digitalización de procesos aduaneros logísticos, adopción de estándares

internacionales cumpliendo con las exigencias regulatorias, por último, un networking de comunidades de importadores regulares, de tal forma que se puedan compartir buenas prácticas y alianzas estratégicas.

Importadores Premium, empresas de gran tamaño y capacidad operativa.

Estrategia: fomentar su participación en misiones económicas permitiendo ampliar las oportunidades de negocio para las empresas junto con la consolidación de relaciones entre Colombia y Reino Unido, por otro lado, realizar espacios de co-creación de políticas bilaterales y posicionarlos como aliados estratégicos de BritCham mediante membresías de alto nivel con beneficios diferenciados reconociendo el liderazgo de dichas empresas.

Respecto a las empresas con operación mixta (exportadora e importadora), también se evidencian tres perfiles:

Los Exportadores Activos, en general empresas recientes, pero con crecimiento sostenido.

Estrategia: acompañar de manera técnica a las empresas para consolidar su operación internacional y espacios para visibilizar sus logros, tales como publicaciones de boletines, vitrinas de casos de éxito o certificaciones de calidad alcanzadas.

Las Empresas Mixtas Pequeñas, con comportamientos más heterogéneos y diversificados.

Estrategia: ofrecer diagnósticos personalizados, formación en planeación financiera y fortalecimiento de capacidades de gestión del riesgo, especialmente comerciales, logísticos, operativos y regulatorios por medio de talleres o consultorías.

Los Grandes Importadores, aunque con alto volumen y madurez operativa, muestran señales de riesgo financiero.

Estrategia: ofrecer asesorías o ayudas de valor centrada en la gobernanza empresarial, talleres de gestión financiera y acompañamiento en programas de sostenibilidad o transición verde, a tecnologías ecológicas y amigables con el medio ambiente para la disminución de costos por políticas medioambientales.

9.6.3. Indicadores de Seguimiento

9.6.3.1. Recalibración del Modelo

Con el fin de mantener la vigencia y precisión del modelo de segmentación y clasificación implementado, se propone establecer un indicador de recalibración basado en la tasa de afiliación acumulada:

"*El modelo deberá ser recalibrado automáticamente cuando el 60 % de las empresas clasificadas como potenciales afiliados hayan sido efectivamente vinculadas a BritCham, o anualmente, lo que ocurra primero*".

Con el propósito de que:

- Se pueda asegura que el modelo se ajuste a los cambios estructurales del ecosistema empresarial y las nuevas dinámicas de mercado.
- Se realice una revisión anticipada si el modelo demuestra ser altamente efectivo en corto tiempo, permitiendo refinar criterios de segmentación.

La recalibración debe considerar: la actualización de la base de datos, reentrenamiento de los modelos con las nuevas observaciones, validación de rendimiento y ajuste de estrategias de captación.

9.6.3.2. Tasa de nuevos afiliados

Como parte del diseño estratégico orientado a incrementar la afiliación empresarial a BritCham, se propone la incorporación del siguiente indicador cuantitativo para dar seguimiento: Tasa de nuevos afiliados, de tal forma que se pueda medir el crecimiento relativo de los afiliados durante un periodo y determinar el rendimiento de las estrategias para cada clúster (Mailchimp, 2024).

Fórmula del indicador:

$$Tasa\ de\ nuevos\ afiliados = \frac{(Nuevos\ afiliados\ en\ el\ periodo)}{(Total\ de\ afiliados\ al\ inicio\ del\ periodo} x\ 100$$

Con el uso de este indicador, se podrá validar la efectividad de las estrategias propuestas o modificadas por BritCham según su necesidad y capacidad, permitiendo identificar las acciones que tienen mayor impacto según el perfil empresarial, clúster o comportamiento económico de las empresas.

Por otro lado, y con el apoyo del tablero en Power BI, se realiza la caracterización de nuevos afiliados por medio de los clústeres modelados con Machine Learning y explorar así sus comportamientos.

A nivel institucional, permite a BritCham contribuir en la toma de decisiones comerciales fundamentadas en los datos analizados y en el comportamiento orgánico del indicador.

9.6.3.3. Beneficios de los indicadores

- Poder evaluar el impacto real de las estrategias de afiliación.
- Facilitar la toma de decisiones basadas en datos, con metas trimestrales claras.
- Ayudar a identificar períodos con mayor o menor captación de nuevos miembros y a partir de esto poder crear planes de acción para mantener un comportamiento constante en la periodicidad evaluada a futuro.
- Mejorar los procesos de segmentación y perfilamiento, al permitir analizar la afiliación según sectores, tamaños empresariales o clústeres de comportamiento.
- Identificar el momento adecuado para recalibrar el modelo sin depender exclusivamente de ciclos fijos, garantizando que siempre refleje la realidad actual del entorno empresarial.
- Reducir la necesidad de revisión manual constante al establecer una regla clara que puede integrarse como alerta en el sistema de monitoreo.

En conjunto, tanto la visualización proporcionada, como las estrategias propuestas y el indicador de seguimiento, fortalecen la capacidad analítica de la Cámara, promoviendo la gestión basada en resultados y decisiones analizadas con impactos relevantes en el crecimiento de BritCham, junto con el cumplimiento de los objetivos de negocio de expansión y consolidación del ecosistema empresarial de Cámaras binacionales.

10. Conclusiones

El desarrollo del proyecto permitió implementar con éxito una solución integral de analítica de negocio bajo la metodología CRISP-DM, enfocada en identificar patrones comerciales y económicos en los flujos de exportación e importación entre Colombia y Europa, específicamente con el Reino Unido. Por medio del uso de técnicas de procesamiento, modelado y visualización, se logró estructurar una herramienta capaz de apoyar a BritCham en la identificación y captación estratégica de nuevos afiliados.

BritCham se encuentra en un sector altamente competitivo y dinámico, donde el posicionamiento y reconocimiento son clave para su crecimiento. Mediante la técnica de procesamiento de lenguaje natural (NLP) y la comparación con cámaras binacionales, se identificaron oportunidades de diferenciación y se generaron recomendaciones estratégicas para fortalecer su presencia en este entorno. Como resultado del benchmarking realizado, se propusieron acciones orientadas a la diversificación y adaptación de servicios, la optimización de la estructura de membresía, el fortalecimiento de estrategias digitales y de marketing, la ampliación de alianzas estratégicas. Todo ello, con el propósito de consolidar

una propuesta de valor más robusta y alineada con el fortalecimiento de la cultura empresarial británica en Colombia.

Los modelos de segmentación y clasificación implementados fueron capaces de agrupar a las empresas en clústeres con características económicas y comerciales similares, revelando perfiles estratégicos como “exportadores ágiles”, “importadores premium” o “empresas mixtas consolidadas”. A partir de estos grupos, también se desarrollaron modelos predictivos que permiten clasificar nuevas empresas dentro de los segmentos existentes, ampliando así la utilidad del sistema más allá del análisis descriptivo.

El tablero de inteligencia de negocio diseñado en Power BI consolidó los hallazgos del proyecto en una herramienta visual, interactiva y accesible, fácil de usar, permitiendo filtrar información por tipo de empresa, actividad comercial, tamaño, ubicación y desempeño financiero, facilitando así la toma de decisiones estratégicas basadas en datos actualizados y confiables.

Finalmente, se diseñaron y propusieron estrategias específicas para cada clúster identificado, brindando a BritCham recomendaciones claras para mejorar su proceso de afiliación y promesa de servicios, de tal forma que sean personalizados y permitan fortalecer su presencia en el ecosistema empresarial colombo-británico, y así mismo a nivel de mercado binacional.

En conjunto, estas acciones consolidan un sistema analítico sólido, escalable y alineado con los objetivos estratégicos y de negocio de BritCham, alineado con su necesidad de crecimiento, contribuyendo a su posicionamiento como una entidad proactiva, guiada por los datos y orientada a resultados.

11. Plan y Recomendaciones de Implementación y Aplicación

Con base en el desarrollo del proyecto, basado en metodologías y técnicas de inteligencia de negocios, el plan y recomendaciones de implementación están dadas según el objetivo del negocio y la necesidad específica de la empresa en cuanto a la participación de nuevos afiliados, permitiendo así, que este proyecto aporte en la segmentación, clasificación y entendimiento de sus afiliados antiguos y potenciales, con el fin de optimizar la toma de decisiones comerciales, fortalecer el proceso de afiliación y priorizar los servicios según el perfil empresarial.

Fase 1: Infraestructura y Capacidad

Adquisición de licencias Power BI Pro para garantizar la compartición y actualización colaborativa del tablero por parte de la Cámara de Comercio.

Disponibilidad de servidores o espacio en la nube (Microsoft OneDrive o SharePoint) para alojar bases de datos.

Instalación de herramientas analíticas necesarias como Power BI Desktop, entornos Python/R para mantenimiento de los modelos y Excel avanzado.

Fase 2: Despliegue de Tableros y Visualizaciones

Comprensión del dinamismo del tablero, filtros, información relevante y método de actualización de los bases contemplados en los apartados *9.3 Preparación de los datos* y *9.6. Despliegue*.

Fase 3: Integración

Entendimiento de la interpretación de los clústeres mencionados en la sección de evaluación de los modelos, como resultado del modelado y las estrategias creadas para cada uno en el apartado *9.6 Despliegue*, subnumeral *9.6.2. Estrategias para cada uno de los clústeres generados*, de tal forma que se pueda hacer el uso estratégico del tablero de visualización y así mismo acomodar, modificar y personalizar las acciones propuestas según el tipo de empresa y las decisiones de BritCham.

Fase 4: Indicadores de Éxito

El éxito se verá reflejado luego de su implementación en la Cámara, donde se espera que al menos el 5% de las empresas segmentadas de un alto potencial de afiliación para BritCham, incentivando así el uso y análisis del tablero teniendo en cuenta las necesidades de la Cámara, demostrando su utilidad y apropiación por los usuarios.

Junto con el indicador de seguimiento propuesto también reflejará la medición del impacto de las estrategias de afiliación con metas claras, optimizando la toma de decisiones basada en datos. Además, facilita la identificación de tendencias en la captación de miembros y mejora la segmentación según características clave como sector, tamaño o comportamiento.

Fase 5: Gestión de Riesgos y Plan de Acción

Tabla 30 Gestión de riesgos

Riesgo	Plan de acción	Mitigación
Baja calidad de datos	Automatizar procesos de validación y limpieza de datos antes del modelado.	Se creó un código en Google Colab para llevar a cabo la limpieza de las bases de datos.
Retrasos en la actualización de datos	Definir cronograma de carga con responsables y validación cruzada con fuentes.	El responsable será designado por la Cámara y el tiempo de actualización se propone que sea cada año.
Resistencia al cambio-BritCham	Validación con la Cámara de lo que espera de los tableros	Se validó con la empresa el uso y entendimiento de la información que contiene la visualización.

Como conclusión del plan de recomendación e implementación, se resalta que el proyecto brinda a BritCham una herramienta robusta y estratégica para mejorar la afiliación empresarial mediante el uso de inteligencia de negocio que garantiza no solo la sostenibilidad del tablero, sino también su utilidad práctica para tomar decisiones basadas en datos, segmentar clientes con precisión y diseñar estrategias personalizadas, junto con la implementación adecuada permitirá a la Cámara fortalecer su proceso de afiliación,

optimizar recursos y generar valor a partir del análisis continuo de sus empresas actuales y potenciales.

Referencias Bibliográficas

Alce, I. B. (26 de 07 de 2019). *La matriz de confusión y sus métricas* . Obtenido de https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Angeli, J. (29 de 06 de 2018). *¿Qué es el mapeo de procesos AS IS/TO BE?* Obtenido de https://www.neomind.com.br/es/blog/que-es-el-mapeo-de-procesos-as-is-to-be/?lang=es

Astera. (13 de Marzo de 2025). *¿Qué es el preprocesamiento de datos? Definición, conceptos, importancia, herramientas*. Obtenido de Astera: https://www.astera.com/es/type/blog/data-preprocessing/

BritCham Colombia. (2025). *CATEGORÍAS DE AFILIACIÓN*. Obtenido de https://britcham.com.co/porque-afiliarse-a-britcham/

BritCham Colombia. (2025). *La Cámara*. Obtenido de https://britcham.com.co/la-camara/

BritCham Colombia. (2025). *Servicios*. Obtenido de https://britcham.com.co/servicios/

Carrasco, O. C. (30 de 09 de 2024). *Gaussian Mixture Model Explained*. Obtenido de Built In: https://builtin.com/articles/gaussian-mixture-model

Comercio, Industria y Transporte. (2019). *Acuerdo de continuidad comercial entre Colombia y el Reino Unido*. Obtenido de https://www.tlc.gov.co/acuerdos/vigente/reino-unido

Comercio, Industria y Turismo. (2012). *Abecé del acuerdo comercial con la Unión Europea*. Obtenido de https://www.tlc.gov.co/acuerdos/vigente/union-europea/1-antecedentes/abece-del-acuerdo-comercial-con-la-union-europea

DANE. (2024). *Exportaciones*. Obtenido de https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-internacional/exportaciones

DANE. (2024). *Importaciones*. Obtenido de https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-internacional/importaciones

Data Camp. (10 de 09 de 2024). *AUC and the ROC Curve in Machine Learning*. Obtenido de AUC and the ROC Curve in Machine Learning

Data Scientest. (15 de 10 de 2023). *Spectral Clustering: definition, operation, use*. Obtenido de https://datascientest-com.translate.goog/en/spectral-clustering-definition-operation-use?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Entropia. (2024). *Metodología Lego Serious Play*. Obtenido de https://entropiacreatividad.com/metodologia-lego-serious-play/

Espinosa-Zúñiga, J. J. (30 de Agosto de 2020). *Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública*. Obtenido de Scielo: https://www.scielo.org.mx/scielo.php?pid=S1405-77432020000100008&script=sci_arttext#B6

Geeks For Geeks. (29 de 07 de 2024). *Bayesian Information Criterion (BIC)*. Obtenido de https://www-geeksforgeeks-org.translate.goog/bayesian-information-criterion-bic/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=wa

Geeks For Geeks. (29 de 01 de 2025). *Agrupación en clústeres DBSCAN en aprendizaje automático . Agrupación basada en la consideración*. Obtenido de https://www-geeksforgeeks-org.translate.goog/dbscan-clustering-in-ml-density-based-clustering/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Google for developers. (2025). *Clasificación: exactitud , recuperación , precisión y métricas relacionadas* . Obtenido de https://developers-google-com.translate.goog/machine-learning/crash-course/classification/accuracy-precision-recall?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

IBM. (2025). *¿Qué es el algoritmo de k vecinos más cercanos (KNN)?* . Obtenido de https://www.ibm.com/mx-es/think/topics/knn

IBM. (2025). *¿Qué es el análisis de componentes principales (PCA)?* . Obtenido de https://www.ibm.com/es-es/think/topics/principal-component-analysis

IBM. (2025). *¿Qué es el bosque aleatorio?* . Obtenido de https://www.ibm.com/mx-es/think/topics/random-forest

IBM. (2025). *¿Qué es Gradient Boosting?* . Obtenido de https://www.ibm.com/topics/gradient-boosting

IBM. (2025). *¿Qué es la agrupación en clústeres k-means?* . Obtenido de https://www.ibm.com/mx-es/think/topics/k-means-clustering

IBM. (2025). *¿Qué es una matriz de confusión?* . Obtenido de https://www.ibm.com/es-es/think/topics/confusion-matrix

IBM CORPORATION. (17 de Agosto de 2021). *Conceptos básicos de ayuda de CRISP-DM*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

IMB. (2025). *¿Qué es el clustering jerárquico?* Obtenido de https://www.ibm.com/es-es/think/topics/hierarchical-clustering

Legiscomex- Estadísticas de comercio exterior. (2024). Obtenido de Legiscomex: https://www.legiscomex.com/informacion-estadisticas-de-comercio-exterior

Mailchimp. (2024). *Cómo calcular y mejorar tu tasa de adquisición* . Obtenido de https://mailchimp.com/es/resources/acquisition-rate/#:~:text=La%20tasa%20de%20adquisici%C3%B3n%20mide%20el%20%C3%A9xito,invertir%20recursos%20para%20obtener%20los%20mejores%20resultados.

Otavalo, J. (05 de 12 de 2020). *Birch Clustering que es y como funciona?* Obtenido de Medium: https://juanotavalo.medium.com/birch-clustering-que-es-y-como-funciona-9f94b13246f4

Rodriguez, D. (30 de 06 de 2023). *El índice de Davies-Bouldinen para estimar los clústeres en k-means e implementación en Python*. Obtenido de https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/&ved=2ahUKEwj-w5PFq6mNAxXCRjABHU-qIJoQFnoECDkQAQ&usg=

ScienceDirect. (2021). *Coeficiente Silhouette*. Obtenido de https://www.sciencedirect.com/topics/computer-science/silhouette-coefficient

Scikit Learn. (2025). Obtenido de StandardScaler: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Scikit Learn. (2025). *3.1. Validación cruzada: evaluación del rendimiento del estimador* . Obtenido de https://scikit--learn-org.translate.goog/stable/modules/cross_validation.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Scikt learn . (2025). *Neural network models (supervised)*. Obtenido de https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Towards Data Science. (2022). *Índice de Calinski-Harabasz para la evaluación de agrupamiento de K-medias con Python*. Obtenido de https://towardsdatascience-com.translate.goog/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python-4fefeeb2988e/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Universitat Oberta de Catalunya. (2025). *Espacio de recursos de ciencia de datos-LightGBM*. Obtenido de https://datascience.recursos.uoc.edu/es/lightgbm/

Zajic, A. (17 de 04 de 2025). *What Is Akaike Information Criterion (AIC)?* Obtenido de Built IN: https://builtin.com/data-science/what-is-aic

Zendesk. (14 de Septiembre de 2024). *Cuál es el significado de cluster y cuáles son las ventajas de implementarlo en tu empresa*. Obtenido de https://www.zendesk.com.mx/blog/significado-cluster/#

Anexos Técnicos

1. Nube de PalabrasK (archivo ipynb)
2. Limpieza de las bases (archivo ipynb)
3. Exploración y estádisticos de cada una de las bases (archivo ipynb)
4. Base final consolidada (archivo ipynb)
5. Modelos implementados por cada una de las bases (archivo ipynb)
6. Archivos PKL con los modelos guardados (archivo ipynb).