* **Título del proyecto.**
* **Resumen**. Una síntesis de lo hecho hasta el momento. Ésta es una parte critica de un trabajo, de hecho, es a veces lo único que se lee. Es por ello, que es una parte integral del proyecto y en esta primera entrega se espera que ustedes tengan una primera aproximación que irán mejorando con el tiempo
* **Introducción**. Donde se establece el problema o la pregunta bien definida con una motivación clara para resolverla.
* **Revisión preliminar de antecedentes en la literatura**. Revisión de literatura nacional e internacional, para enriquecer el proyecto con lo que ya se haya escrito sobre el tema que se va a tratar.
* **Descripción detallada de los datos**. Descripción que incorpore estadísticas descriptivas y/o visualizaciones.
* **Propuesta metodológica**. En esta sección debe mostrarse cómo el proyecto incorpora aprendizaje no supervisado y qué algoritmo planean utilizar. (Este algoritmo puede ser preliminar y pueden modificarlo para la entrega final, si ven que a lo largo del curso incorporan nuevas herramientas que pueden resultar más apropiadas)
* **Bibliografía**. Incorpora los artículos mencionados en el texto.

# Segmentación de mercados usando Google Analytics: Una vista de segmentación de mercadeo B2C en una compañía cosmética

## Resumen

El proyecto está enfocado en desarrollar un análisis de clusters para una compañía que produce y distribuye la toxina U.S. botulinum, un grupo de exotoxinas que interfieren la transmisión neural causando parálisis muscular (Nigam & Nigam, 2010). El mercado de Estados Unidos para este producto esta valorado en US$4.6 Billones, y se espera que crezca año a año al 5% (Fortune Business Insights, 2023), lo que crea una oportunidad para esta compañía de capturar mayores porciones de mercado. Con base en una sabana de datos de los usuarios recogida por medio de la plataforma de Google Analytics, se espera desarrollar la segmentación de estos usuarios con el fin de maximizar el impacto de las campañas de mercadeo que la compañía lanzará por medio de otras plataformas digitales.

## Introducción

Una compañía del sector cosmético cuyo mercado se enfoca a Estados Unidos (USA) produce fármacos inyectables para el rostro, para todos los consumidores que deseen preservar el estado de su piel (prevenir tener más arrugas) o estirar la piel del rostro para mantener un aspecto joven. De acuerdo con (Fortune Business Insights, 2023) en 2022 el mercado de la toxina U.S. botulinum (un componente principal para la fabricación de estos inyectables) estuvo valorado en US$4.6 billones, y para el 2030 se espera que crezca a US$6.7 billones, lo que se traduce en oportunidades de mercado para este tipo de productores.

Las compañías de este sector desarrollan distintas estrategias de mercadeo en dos frentes (B2B y B2C) para capturar porciones de este mercado, dado que, estos inyectables son distribuidos y aplicados a compañías del sector salud con capacidades para realizar procedimientos de belleza (e.g. Hospitales, centros dermatológicos, centros de inyectores, entre otros). Es decir, el consumidor final (los usuarios que se inyectan) no pueden comprar estas toxinas y aplicárselas a si mismos, sino que se acercan a un centro de salud, donde les inyectan los productos que ellos demandan. En este sentido, las compañías que ofertan dichas toxinas deben desarrollar estrategias de mercadeo para corporaciones (lo que se llama mercadeo B2B), y para consumidores finales (o más conocido como B2C). El presente trabajo estará enfocado en el segmento B2C (sin dejar de lado que el análisis de clusters también puede desarrollarse para el segmento B2B). Las bases de datos suministradas son capturadas desde sitios web enfocados en el segmento B2C.

La compañía cosmética ha recogido información de los usuarios que visitan su página web mediante Google Analytics (“una plataforma que recoge datos de [los] sitios web … para crear informes que proporcionan estadísticas para el provecho de [la] empresa” (Google, 2023) ). Dicha información está compuesta por datos de rango de edad, genero, intereses de los usuarios, ente otros. Es importante resaltar que la plataforma no distribuye información personal identificable (PII por sus siglas en ingles), por lo que no se usa datos de carácter sensible en este análisis. Con los datos proveídos, la compañía espera que segmentar a los usuarios que visitan su página web en un número reducido de clusters.

Dentro de las ventajas que se tienen por incluir el análisis de clusters en las estrategias de mercadeo, se encuentran: (1) ayudar a segmentar a los consumidores basados en sus características, preferencias, comportamientos, o necesidades heterogéneas (dado que estas últimas están cambiando constantemente); permitiendo incrementar el entendimiento de la base de consumidores para adaptar los esfuerzos de mercadeo y maximizar el impacto de recursos limitados (LinkedIn, 2023; Dibb, Market segmentation: strategies for success, 1998), (2) ofrecer recomendaciones personalizadas, venta cruzada, cupones, descuentos, o promociones a los usuarios de manera enfocada (LinkedIn, 2023), (3) incrementar la eficiencia las estrategias de focalización (targeting), al añadir claridad al proceso de desarrollo de estrategias de mercadeo (Fonseca, 2011; Dibb, Market segmentation: strategies for success, 1998), (4) identificar los segmentos no explorados para crear una ventaja competitiva, desarrollando productos y servicios llamativos únicos para dichas audiencias (Fonseca, 2011; Dibb, Market segmentation: strategies for success, 1998), (5) soportar los lanzamientos de nuevos productos (Saunders, 1980), (6) mejorar el rendimiento organizacional (Dibb, Stern, & Robin, 2002), entre otras.

El análisis de clusters para mercadeo comprende tres fases: segmentación, focalización y posicionamiento (llamado STP por sus siglas en ingles) (Dibb, Market segmentation: strategies for success, 1998).

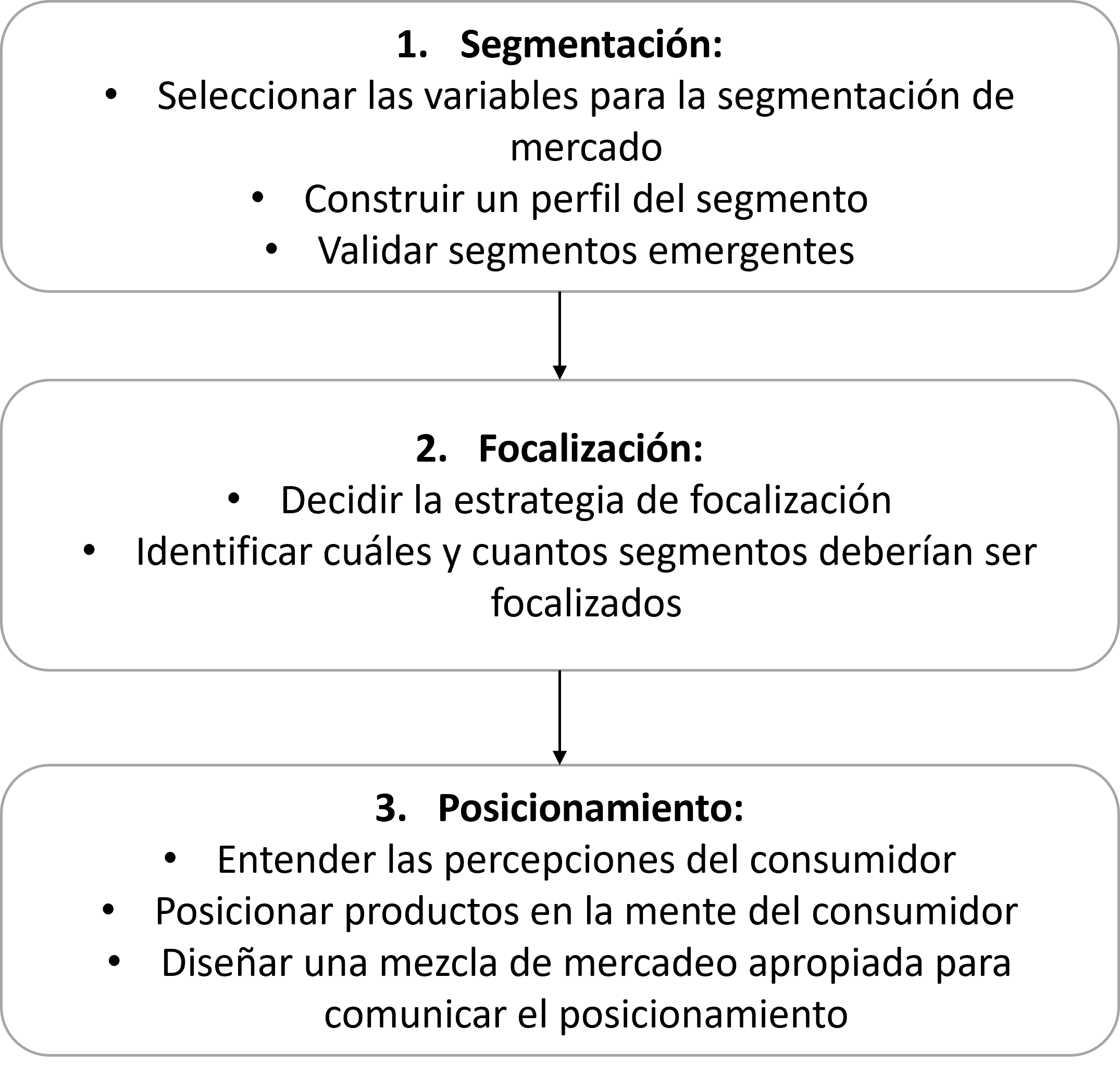


Ilustración 1 Adaptado de (Dibb, Market segmentation: strategies for success, 1998)

El presente trabajo estará destinado en cubrir la primera fase de este análisis de clusters, segmentación.

## Revisión preliminar de antecedentes en la literatura

La clusterización jerárquica es una técnica de análisis de datos ampliamente utilizada que se encuentra en la intersección de la estadística, la minería de datos y la inteligencia artificial. Surgió en la década de 1960 y desde entonces ha sido objeto de investigación continúa debido a su capacidad para revelar estructuras y patrones ocultos en conjuntos de datos multidimensionales sin requerir etiquetas previas.

Uno de los primeros enfoques en la clusterización jerárquica se introdujo en el trabajo pionero de Sibson (1973) con la presentación del algoritmo SLINK, que se destacó por su eficiencia computacional en el cálculo de enlaces jerárquicos. Esta contribución fue fundamental para comprender los fundamentos de la clusterización jerárquica y establecer la base para futuras investigaciones en este campo.

La aplicación de la clusterización jerárquica se ha extendido a diversas áreas, como la biología y la genómica. Eisen, Spellman, Brown y Botstein (1998) utilizaron esta técnica para analizar patrones de expresión génica y descubrir subgrupos funcionales de genes. Este estudio subraya la capacidad de la clusterización jerárquica para revelar relaciones biológicas ocultas en datos genómicos.

En el ámbito del marketing y la segmentación de clientes, Jain y Dubes (1988) proporcionaron una visión general completa de los algoritmos de clusterización, incluida la clusterización jerárquica. Destacaron cómo esta técnica puede utilizarse para identificar grupos de clientes con características similares, lo que es fundamental para la toma de decisiones estratégicas en marketing.

Además, la clusterización jerárquica ha encontrado aplicaciones en el análisis de imágenes y la visión por computadora. Jain y Dubes (1981) exploraron su uso en la segmentación de objetos en imágenes y la detección de patrones visuales, subrayando su versatilidad en campos interdisciplinarios.

Además de los algoritmos y aplicaciones específicos, la clusterización jerárquica se ha beneficiado de investigaciones sobre herramientas y visualización de resultados. Murtagh y Legendre (2014) profundizaron en el método de Ward, uno de los enfoques más utilizados en clusterización jerárquica. Presentaron el algoritmo aglomerativo y la función de criterio de Ward, proporcionando una base sólida para su implementación en software de análisis de datos.

Por último, Inselberg y Dimsdale (1990) introdujeron el concepto de coordenadas paralelas como una herramienta efectiva para visualizar resultados de clusterización jerárquica y explorar estructuras de datos multidimensionales de manera intuitiva.

En un artículo de 1980 (Saunders, 1980) considera que los métodos de clusters más apropiados para desarrollar segmentos de mercado son Segmentación de beneficios y agrupamiento jerárquico. Para este último, se proponen alternativas como el método Q-factor, particiones (optimización), búsqueda de densidad, Aglomeración (clumping en inglés) y agrupamiento restringido.

De acuerdo con (Dibb, Stern, & Robin, 2002) el análisis de cluster es la herramienta más efectiva para incrementar el entendimiento del consumidor, e impactar en el rendimiento organizacional.

Cuando se hace un análisis de clúster para segmentar mercados, (Lin, 2002) propone que no sé considere únicamente variables demográficas, sino también variables psicográficas, es decir, variables que permitan entender el estilo de vida y la personalidad de los consumidores, explorar modelos de consumo e identificar características relevantes de la marca (Lin, 2002).

En resumen, los antecedentes de la clusterización jerárquica reflejan su importancia en diversas disciplinas y su capacidad para revelar estructuras ocultas en datos complejos y multidimensionales. Estas investigaciones y aplicaciones han sentado las bases para la utilización de esta técnica en una amplia gama de campos, desde la biología hasta el marketing y la visión por computadora.

## Descripción detallada de los datos

1. **userType**: Esta variable representa el tipo de usuario o cliente en un conjunto de datos. Puede tomar diferentes valores para categorizar a los usuarios en grupos específicos. Por ejemplo, los valores podrían incluir "New visitor”, “Older” entre otros.
2. **userAgeBracket**: Esta variable se utiliza para agrupar a los usuarios en categorías o brackets de edad. Los brackets de edad son rangos específicos que abarcan un conjunto de edades similares. Por ejemplo, los brackets podrían ser "Menores de 18 años", "De 18 a 25 años", "De 26 a 35 años" y así sucesivamente. Esta variable ayuda a comprender la distribución de edades de los usuarios en el conjunto de datos.
3. **userGender**: Esta variable indica el género o sexo de los usuarios. Puede tomar valores como "Masculino", "Femenino", "No binario" u otras categorías que representen la identidad de género de los usuarios. El propósito de esta variable es analizar la composición de género de la audiencia o realizar análisis específicos relacionados con el género.
4. **interestOtherCategory**: Esta variable representa la categoría o área de interés de los usuarios que no se encuentra en otras variables específicas. Puede ser una categoría general que abarque intereses diversos y no cubiertos por variables más detalladas. Por ejemplo, podría incluir " Arts & Entertainment/Celebrities & Entertainment News" o " Arts & Entertainment/Music & Audio/Radio". Esta variable proporciona flexibilidad para capturar intereses variados de los usuarios.
5. **channelGrouping**: Esta variable se refiere a la forma en que los usuarios llegaron o accedieron a un canal o plataforma específica. Puede incluir categorías como "Email", "Direct", “Display" o " Organic Search". Ayuda a analizar cómo los usuarios interactúan con el canal o plataforma y cómo llegaron a ella.
6. **country**: Esta variable indica el país de origen o ubicación geográfica de los usuarios. Cada valor corresponde a un país específico, como "Estados Unidos", "Canadá", "Reino Unido", etc. El país de origen es útil para realizar análisis geográficos, como la segmentación de usuarios por ubicación geográfica.
7. **date**: Esta variable representa la fecha en que se registraron o realizaron las interacciones de los usuarios en el conjunto de datos. Puede ser una fecha específica en formato YYYY-MM-DD o incluir información de hora si es necesario. Esta variable es esencial para realizar análisis temporales y comprender cómo cambian los patrones con el tiempo.
8. **users**: Esta variable representa el número de usuarios o clientes que cumplen con las características específicas descritas en las variables anteriores. Es una variable cuantitativa que indica la cantidad de usuarios en una categoría o grupo particular. Esta variable es fundamental para realizar análisis estadísticos y cuantitativos.

## Propuesta metodológica

Objetivos:

* Aplicar la clusterización jerárquica como técnica de aprendizaje no supervisado para agrupar datos en clústeres.
* Identificar patrones y estructuras ocultas en el conjunto de datos sin requerir etiquetas previas.
* Generar conocimientos valiosos a partir de la interpretación de los resultados de la clusterización.

Metodología Propuesta:

1. Recopilación y Preprocesamiento de Datos: Obtener el conjunto de datos de origen y realizar una exploración inicial. Realizar tareas de limpieza de datos, tratamiento de valores faltantes y normalización de ser necesario.

2. Selección de Variables (Opcional): Si el conjunto de datos es grande o contiene muchas características, considerar la selección de variables relevantes para simplificar el análisis y mejorar la eficiencia computacional.

3. Configuración de Parámetros: Determinar el método de enlace (single, complete, average, etc.) y la métrica de distancia (euclidiana, Manhattan, etc.) más apropiados para la clusterización jerárquica.

4. Aplicación de Clusterización Jerárquica: Utilizar el algoritmo de clusterización jerárquica para agrupar las muestras en clústeres de manera jerárquica.

Construir el dendrograma para visualizar la estructura jerárquica de los clústeres.

5. Evaluación del Número Óptimo de Clústeres: Emplear métodos como el codo (elbow method), índice de silueta u otros métodos de selección de clústeres para determinar el número adecuado de grupos.

6. Interpretación de Resultados: Evaluar y describir la coherencia de los clústeres identificados. Explorar y documentar las características y patrones comunes dentro de cada clúster.

7. Validación de Resultados (Opcional): Si se dispone de etiquetas verdaderas, se pueden utilizar medidas de validación como el índice de Rand o la pureza para evaluar la calidad de la clusterización.

8. Documentación y Comunicación de Resultados: Generar un informe detallado que incluya hallazgos, conclusiones y visualizaciones que respalden la interpretación de los resultados.

Conclusiones:

Esta metodología propuesta permite aplicar el aprendizaje no supervisado a través del algoritmo de clusterización jerárquica para explorar y descubrir estructuras ocultas en un conjunto de datos. Los resultados obtenidos proporcionarán una mejor comprensión de la naturaleza de los datos y podrán ser utilizados en diversos campos, desde la segmentación de clientes hasta la clasificación de elementos en taxonomías. La estructura detallada de la metodología facilita un análisis sistemático y una interpretación sólida de los resultados.

## Bibliografía

Dibb, S. (1998). Market segmentation: strategies for success. *Marketing intelligence & planning*, 394-406.

Dibb, S., Stern, P., & Robin, W. (2002). Marketing knowledge and the value of segmentation. *Marketing Intelligence & planning*, 113-119.

Fonseca, J. R. (2011). Why does segmentation matter? Indentifying Market segments through a mixed methodology. *European Retail Research*, 1-26.

Fortune Business Insights. (2023). *U.S. Botulinum Toxin Market Size, Share & Industry Analysis, By Application (Therapeutics), and Aesthetics, By Type, By End User and Forecast, 2023-2030.* USA: Fortune Business Insights.

Google. (2023). *How Google Analytics works*. Obtenido de Google Support: https://support.google.com/analytics/answer/12159447?hl=en

Lin, C.-F. (2002). Segmenting customer brand preference: demographic or psychographic. *Journal of Product and Brand Management*, 249-268.

LinkedIn. (2023). *How do you use cluster analysis to segment your customers*. Obtenido de LinkedIn advice: https://www.linkedin.com/advice/3/how-do-you-use-cluster-analysis-segment-your

Nigam, P., & Nigam, A. (2010). Botulinum toxin. *National Library of Medicine*, 8-14.

Saunders, J. (1980). Cluster Analysis for Market segmentation. *European Journal of Marketing*, 422-435.