Análisis de Clúster para la generación de campañas de mercadeo en la distribución de la toxina US Botulinum en una compañía cosmética

Cristhian Amaya, Andres Beltrán, Gloria Ramos y Sergio Rojas

Maestría MIAD Uniandes

Resumen

En el trabajo se desarrolla un análisis de clústeres basados en datos de la herramienta Google Analytics, con el fin de segmentar posibles audiencias 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), ampliamente usado en el mercado de inyectables. En la actualidad, el mercado de Estados Unidos de esta toxina está valorado en US$4.6 Billones, y se espera un crecimiento año a año del 5% (Fortune Business Insights, 2023), lo que se traduce en oportunidades de mercado. La base de datos contiene información de las sesiones y los atributos demográficos de los usuarios que visitan la página web. Partiendo de esta base de datos se espera desarrollar la segmentación de audiencias con el fin de maximizar el impacto de las campañas de mercadeo que la compañía lanzará por medio de otras plataformas digitales. Para el análisis de clústeres se probaron una serie de modelos y métodos de reducción de dimensiones, y se evaluaron los resultados por medio del coeficiente de silhouette, el índice de Davies-Bouldin y un análisis de consistencia, con el fin de determinar el modelo más apropiado.

Introducción

De acuerdo con (Fortune Business Insights, 2023) en 2022 el mercado de la toxina U.S. botulinum estuvo valorado en US$4.6 billones, y para el 2030 se espera que crezca a US$6.7 billones, por lo que existen oportunidades de mercado creciente para la industria. Una compañía del sector belleza busca introducir su propio producto de esta toxina al mercado, y ha desarrollado campañas de 'Brand Awareness' con el fin de comenzar una relación entre potenciales clientes y la marca (Amazon, 2023), y llevar a estos usuarios a visitar su página web para conocer más del producto. Por medio de la herramienta 'Google Analytics' (Google, 2023), se ha recogido información de los usuarios que visitan la página web (como edad, genero, país, estado e intereses de marca) y se espera segmentar potenciales audiencias con base en estos datos. (Nigam & Nigam, 2010)

Al implementar análisis de Clúster, la compañía espera (1) segmentar a los consumidores basados en sus características, preferencias, comportamientos, o necesidades heterogéneas, permitiendo incrementar el entendimiento de la base de consumidores para adaptar los esfuerzos de mercadeo y maximizar el impacto de recursos limitados (Dibb, 1998; LinkedIn, 2023), (2) incrementar la eficiencia las estrategias de focalización (targeting), al añadir claridad al proceso de desarrollo de estrategias de mercadeo (Dibb, 1998; Fonseca, 2011), (3) ofrecer recomendaciones personalizadas, venta cruzada, cupones, descuentos, o promociones a los usuarios de manera enfocada (LinkedIn, 2023), e (4) identificar los segmentos no explorados para crear una ventaja competitiva, desarrollando productos y servicios llamativos únicos para dichas audiencias (Dibb, 1998; Fonseca, 2011).

De acuerdo con (Dibb, 1998) el análisis de clústeres aplicado a estrategias de mercadeo comprende tres fases: segmentación, focalización y posicionamiento (llamado STP por sus siglas en ingles). Dado el alcance de este trabajo, se esperar cubrir únicamente la fase de segmentación, esperando que el mismo sea un insumo de útilidad en camino de posicionar este nuevo producto en el mercado.

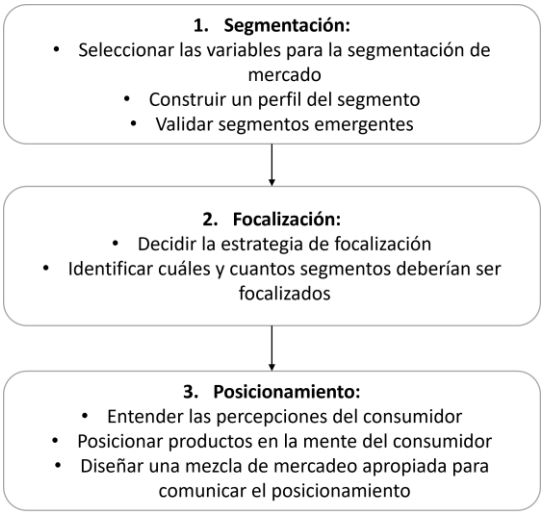

Ilustración 1 Fases de segmentación en mercadeo. Adaptado de (Dibb, 1998)

Cuando se estudian problemas similares de segmentación en la literatura, se evidencia que existe una variedad de metodologías dentro del aprendizaje no supervisado para desarrollar soluciones.

Por ejemplo, (La Cruz, Severeyn, Matute, & Estrada, 2021) desarrollaron un análisis de clúster con base en datos de 'Google Analytics' donde (1) usaron k-medias en una base de datos preprocesada para determinar el número de clústeres por el método de codo ('elbow'), (2) determinaron cuáles variables explicaban la mayor parte de la varianza por medio de una reducción por componentes principales (PCA), (3) luego, dados los componentes principales y el número de segmentos determinado anteriormente, utilizaron de nuevo k-medias para encontrar el número final de segmentos de mercado, y (4) por medio del coeficiente de 'silhouette' evaluaron los resultados. A pesar de que existen muchas similitudes con el presente trabajo, la naturaleza de los datos de este último está enfocada en las características de los usuarios que visitaron una página web, mientras que en (La Cruz, Severeyn, Matute, & Estrada, 2021) se agrupan usuarios por las acciones (eventos) que estos realizaron en una página web. No obstante, los elementos de la metodología propuesta como PCA y k-medias han sido consideradas para el desarrollo del presente trabajo.

De manera similar, (Kansal, Bahuguna, Singh, & Choudhury, 2018) proponen una metodología basada en k-medias para segmentar clientes con base en un histórico de visitas y compras anuales por cliente de un negocio de 'retail'. En este ejercicio, después de recoger los datos, se usaron (1) escalamiento de variables, (2) agrupación por k-medias, (3) agrupación por aglomeraciones, (4) agrupación por turnos medios ('mean shift clustering'), y (5) el coeficiente de 'silhouette' para evaluar resultados.

Por otro lado, (Kamthania, Pahwa, & Madhavan, 2018) utilizan un análisis de clúster diferente llamando agrupación por k-modas, complementado con PCA, para un negocio de comercio electrónico, con el fin de segmentar audiencias dado la marca de productos que los usuarios compraron en un histórico de ventas. En este ejercicio se aborda un ejemplo del algoritmo de k-modas, y se realiza como producto final una visualización de la distribución geográfica de los clientes. El coeficiente de 'silhouette' es también usado para analizar los resultados.

Al considerar los ejemplos citados, existen varias similitudes y metodologías propuestas que han sido consideradas en el presente ejercicio, sin embargo, otras metodologías también se han usado en otros ejercicios que también se han evaluado en este trabajo. Entre ellas, agrupamiento jerárquico, aglomeración, agrupamiento restringido (Saunders, 1980), modelos latentes de clase (Fonseca, 2011), y DBSCAN (Sander, 1997).

Materiales y Metodos

Basados en la información proveída por Google Developer (Google Dev, 2023), se describen a continuación las variables consideradas en el presente ejercicio.

- **'brandingInterest':** Esta variable muestra el tipo de interés que tuvieron los usuarios que se encuentran en etapas avanzadas del proceso de compra.
- **'userAgeBracket':** Esta variable se utiliza para agrupar a los usuarios en categorías o rangos de edad.
- **'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.
- **'country':** Esta variable indica el país de origen o ubicación geográfica de los usuarios.
- **'region'**: Muestra la zona geográfica (Estado o Departamento) de donde viene el usuario, basándose en su IP.
- **'sessions':** Muestra el número de sesiones que superaron los 10 segundos, lograron una conversión o tuvieron 2 o más visitas.

Variable	Tipo	Valores únicos	Moda	Min	Max	Mediana	Media
'country'	Texto	1	'United States'				
'region'	Texto	20	'California'				
'userAgeBracket'	Texto	7	35-44				
'userGender'	Texto	3	'Female'				
'brandingInterest'	Texto	116	'Media & entertainment'				
'sessions'	Numérica			2	217	9	14,44

Tabla 1 Distribución de variables

En total se cuentan con 4.503 registros, realizando un análisis detallado no tiene ningún valor nulo. Cuando se analizan la distribución de visitas por género, es evidente que los usuarios de género femenino contienen la mayor cantidad de visitas al sitio.

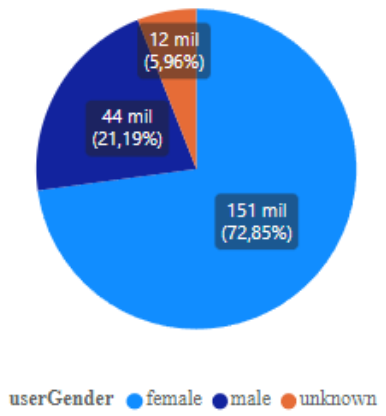

Ilustración 2 Distribución de Genero por sesiones

El 93% de las sesiones están contenidas en 5 (New York, California, Texas, Florida e Illinois) de los 20 estados que registran visitan en la página web.

Los usuarios de entre 35-44 (29%) años visitan más la página web, sin embargo, el volumen de visitas de usuarios entre 25-34 (15%) y 45-54 (22%) años también representan una significativa cantidad de visitas.

En la variable 'brandingInterest' se puede observar que "Media & Entertrainment" es el interés más alto, seguido por "Lifestyles & Hobbies". El top 10 comprende el 47% del total de las sesiones.

Metodología Propuesta

1. Recopilación y Preprocesamiento de Datos: Se añaden la latitud y la longitud a la base de datos según la variable región que define los estados asociados al ingreso de los usuarios. Se retiran las variables de 'sessions', 'country' y 'region'. Se escalan las variables numéricas.

2. Desarrollo de la matriz de distancia: Se emplean tres métodos para el cálculo de las distancias, empleamos distancia gower, otro se aplica distancia Euclidiana a las variables categóricas codificadas a binarias aplicando PCA para reducir dimensionalidad y aplicar Análisis Factorial de datos mixtos (FAMD) (Kassambara, 2017) directamente a las variables para calcular las distancias euclidianas.

3. Configuración de Parámetros: Se configuran parámetros según el modelo a utilizar:

- Kmedias: cálculo iterativo del índice del Silhouette con base al mejor número de cluster definir como parámetro
- Jerárquico: cálculo iterativo del índice del Silhouette para el cálculo de la distancia (t) a seleccionar.
- DBSCAN: parámetro eps calculado por medio del método de máxima curvatura mostrada por la gráfica y el número mínimo de puntos para que una región se considere densa (min_samples = 2*D+1).

4. Aplicación de Clusterización: Se utiliza el algoritmo de DBSCAN, Kmedias y Jerarquico para agrupar los datos.

5. Interpretación de Resultados: Se evalúan los clústeres identificados, por medio del coeficiente de Silhouette, y el índice de Davies-Bouldin.

6. Validación de Resultados: Se validan los resultados y la coherencia de los clústeres al someterlos a un análisis de impacto por medio del porcentaje de sesiones que abarcan sus clústeres más significativos y las características de dichos clústeres.

7. Documentación y Comunicación de Resultados: Se generan las recomendaciones pertinentes al negocio para segmentar audiencias, e iniciar con la fase de focalización.

Resultados y Discusión

Después de evaluar los resultados de los modelos probados por medio de los índices de Silhouete y Davies-Bouldin, y validando la coherencia de los clústeres con respecto al número de sesiones, se llegó a la conclusión que el modelo más apropiado para esta segmentación es el modelo que utiliza análisis jerárquico basado en una matriz de distancia de Gower.

Para este modelo se escalaron las variables numéricas y se creó la matriz de distancia de Gower. Después con la ayuda del coeficiente de silhouette, se optimizo el parámetro de distancia (t), y con dicho criterio optimo se parametrizo el algoritmo de clúster jerárquico.

Este modelo obtuvo el coeficiente de silhouette más alto al compararlo con los demás modelos (0.3), sugiriendo que la segmentación tiene una calidad aceptable para el propósito del ejercicio con instancias separadas apropiadamente. También obtuvo el índice de Davies-Bouldin más bajo, lo que indicaría una buena calidad de clústeres, con una separación más clara entre los clusters, y una baja dispersión interna.

Al observar el impacto de los clusters segmentados por este análisis jerárquico en términos de sesiones se puede observar que de los 62 clusteres propuestos, 28 representan al menos el 80% de las

sesiones, lo que se traduce en oportunidades de segmentación importantes para los mercaderistas que avancen con la fase de focalización.

Método	# Clústeres	Índice Silhouette	Índice Davies-Bouldin	# Clústeres <80%	Primer Clúster	Sesiones de primer clúster	Sesiones primer clúster %	Sesiones por Clúster %
Clúster jerárquico Gower	62	0.3	1.425	28	31	9,719	6.33	1.61
Clúster DBSCAN Gower	65	0.251	1.708	28	0	9,719	6.33	1.54
Cluster KMeans Dummies con PCA	46	0.249	1.604	23	4	11,014	7.18	2.17
Clúster jerárquico Dummies con PCA	51	0.041	3.416	29	51	18,606	12.13	1.96
Clúster DBSCAN Dummies con PCA	141	-0.024	2.962	63	-1	28,033	18.27	0.71
Clúster KMeans FAMD	55	0.005	6.448	28	13	26,008	16.95	1.82
Clúster jerárquico FAMD	98	-0.252	8.361	40	27	9,336	6.08	1.02
Clúster DBSCAN FAMD	85	-0.044	7.647	40	-1	9,441	6.15	1.18

Tabla 2 Resumen de resultados de los metodos probados

Al evaluar el clúster con mayor número de sesiones resultado del método jerárquico con Gower, y realizar una prueba de consistencia de los datos, se encuentra que:

- Este clúster agrupa alrededor del 6.3% de todas las sesiones del sitio, para la base de datos de prueba
- Este cluster contiene las siguientes características: Genero: Femenino, Rango de edad: 35-44, Región: California, intereses: 'Media & Entertainment', 'Lifestyles & Hobbies', 'News & Politics', entre otros 100 intereses.

Conclusión

Al considerar las características de clúster más grande, en términos de sesiones, del análisis jerárquico, se percibe que estas son esperadas y consistentes con el análisis descriptivo de los datos realizado previamente. Vale la pena resaltar que, si bien la cantidad de intereses de estos clústeres es bastante alta, esto no representa una desventaja para los mercaderistas que desarrollan campañas por medio de redes sociales, sino que es un input aceptable para segmentar posibles audiencias.

El número de audiencias resultado del análisis de clúster es aceptable para realizar campañas en medios digitales de manera focalizada, de manera que los resultados presentados pueden someterse a prueba inmediatamente.

Para ser un primer ejercicio de segmentación, se considera que el resultado tendrá un impacto positivo en las siguientes etapas de focalización y posicionamiento, que serán implementados por la compañía. Si bien este primer ejercicio tiene hallazgos prometedores, se recomienda refinar el análisis de cluster una vez se tenga una mayor cantidad de datos, y un mayor número de atributos de los usuarios que visitan su página, como 'ciudad' y otros campos que se pueden habilitar dentro de la herramienta de Google Analytics.

Para finalizar, la agrupación por análisis jerárquico mostro ser un modelo apropiado para esta base de datos caracterizada en mayor parte por contener variables categóricas. Sin embargo, tanto DBSCAN como K-medias mostraron resultados prometedores para este ejercicio, confirmando hallazgos similares encontrados en la literatura.

Bibliografía

Amazon. (2023). *Guide: Brand Awareness*. Obtenido de https://advertising.amazon.com/library/guides/brand-awareness

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

Google Dev. (16 de Ago de 2023). *Google Analytics*. Obtenido de Google Analytics: https://developers.google.com/analytics/devguides/reporting/data/v1/api-schema?hl=es-419

Kamthania, D., Pahwa, A., & Madhavan, S. S. (2018). Market Segmentation Analysis and Visualization Using K-mode Clustering Algorithm for E-commerce Business. *CIT. Journal of Computing and Information Technology, 26*, 57-68. doi:10.20532/cit.2018.1003863

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS).*

La Cruz, A., Severeyn, E., Matute, R., & Estrada, J. (2021). Users Segmentation Based on Google Analytics Income Using K-means. *Information and Communication Technologies, 9th conference of Ecuador, TITEC 2021* (págs. 225-226). Guayaquil, Ecuador: Springer.

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.

Sander, J. (1997). *Density-Based clustering in Spatil Databases:The Algorithm GBSCAN and Its Applications.* Muchen.

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