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

Curso Aprendizaje no supervisado
Maestría en Inteligencia Analítica de Datos

uniandes

Problema y Contexto

Posibilidades de Mercado

La toxina Botulinum es una neurotoxina producida por la bacteria *Clostridium Botulinum*. Está bacteria elabora exotoxinas que interfieren con la transmisión neural al bloquear la liberación de acetilcolina, lo que causa parálisis muscular (Nigam & Nigam, 2010).

De acuerdo con la revista Fortune Business Insights (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**.

Lanzamiento de un nuevo producto

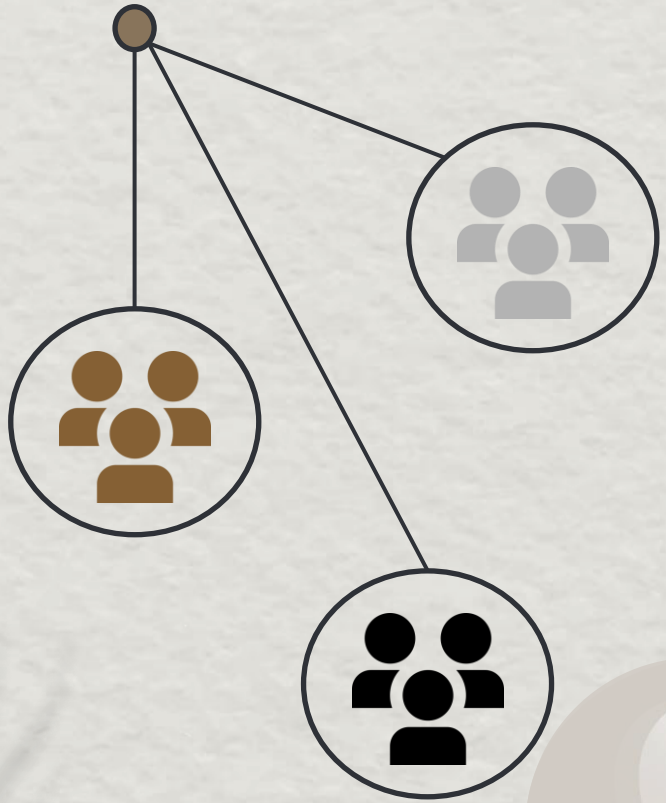
Una compañía del sector belleza busca introducir su propio producto de la toxina US Botulinum y ha desarrollado una campaña de *Brand Awareness* con el fin de **comenzar una relación entre potenciales clientes y la marca** (Amazon, 2023), y llevar a los usuarios a visitar su página web para conocer más del producto.

Por medio de la herramienta *Google Analytics* se ha recogido información de los usuarios que visitan la página web y se espera segmentar poten[illegible] audiencias con base en estos datos.

¿Por qué se aborda este problema desde el aprendizaje no supervisado?

La compañía espera:

1. **Segmentar** a los consumidores basados en sus características, preferencias, comportamientos o necesidades heterogéneas.
2. **Incrementar la eficiencia** las estrategias de focalización.
3. Ofrecer recomendaciones personalizadas, venta cruzada, cupones, descuentos o promociones a los usuarios.
4. **Identificar los segmentos no explorados** para crear una ventaja competitiva.



Por esto y teniendo en cuenta la naturaleza de los datos, se ha decidido que el análisis de clúster, una metodología del aprendizaje no supervisado es la mejor opción para segmentar estas audiencias.

Algoritmo y Datos

Análisis descriptivo de los datos

4.503
Registros

206.696
Sesiones de GA

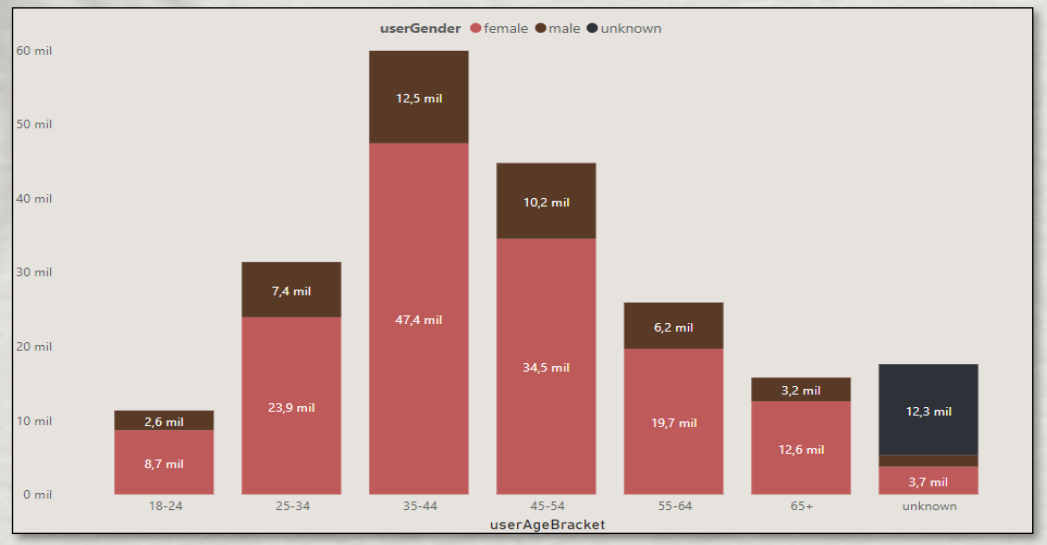
349
Registros unknown*

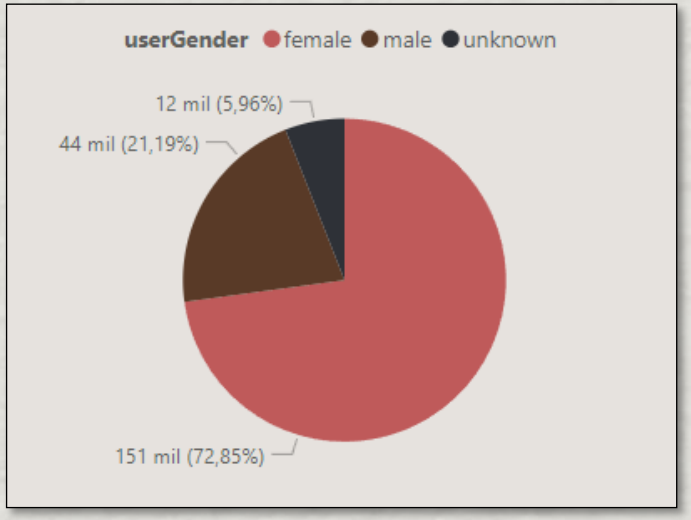
Variable	Tipo	Valores únicos	Moda	Min	Max	Mediana	Media
'country'	Texto	1	'United States'				
'region'	Texto	20	'California'				
'userAgeBracket'	Texto	7	35-44				
'userGender'	Texto	3	'Female'				
'brandingInterest'	Texto	116	'Media & entertainment'				
'sessions'	Numérica			2	217	9	14,44

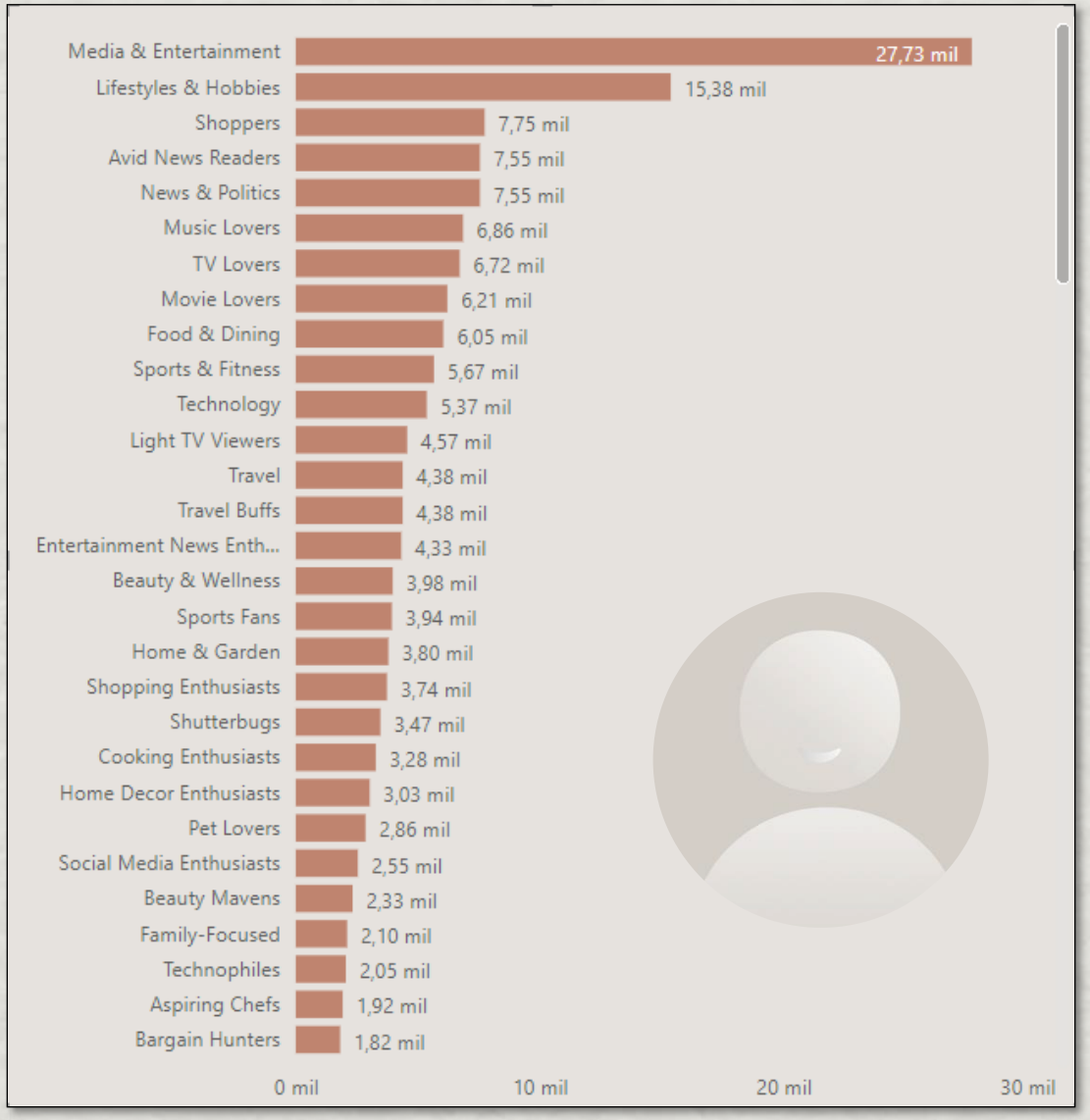
El análisis de calidad de los datos evidencio que no existen registros nulos en los datos.

*Registros unknown se refiere a usuarios que prefirieron no declarar su sexo o edad, pero si otros valores como intereses

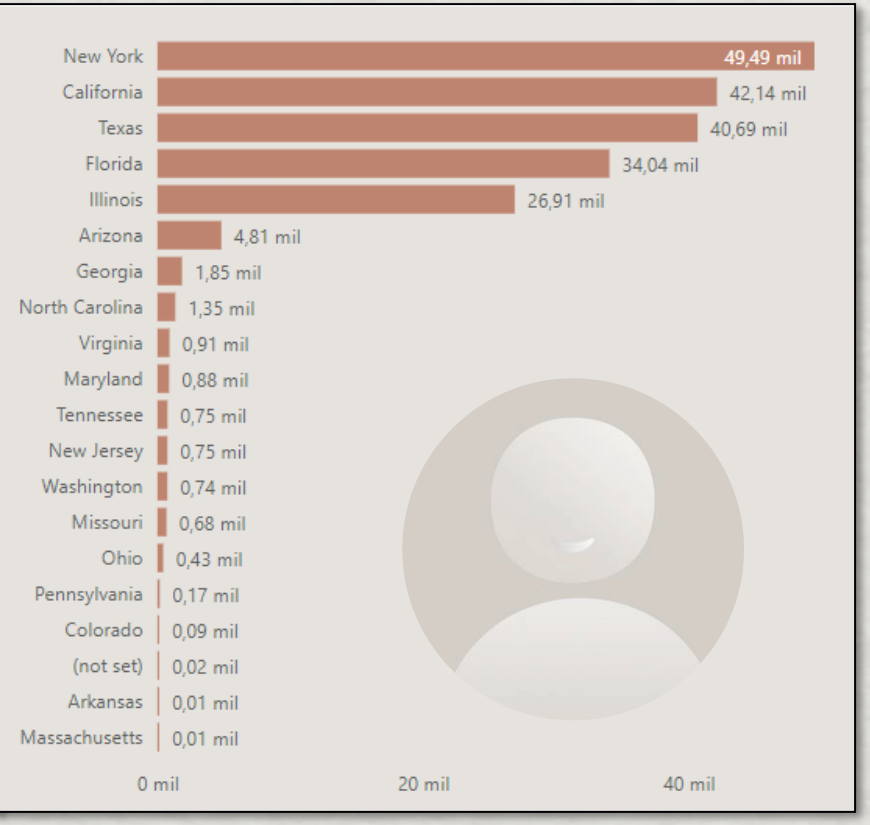
Análisis descriptivo de los datos







Análisis descriptivo de los datos



Casos de estudio

(La Cruz, Severeyn, Matute, & Estrada, 2021)

Análisis de Clúster

- K-medias en una base de datos preprocesada
- Se determina un número aproximado de clústeres por el método del codo
- Por PCA se determinaron las variables que explicaban la mayor proporción de varianza
- De nuevo se aplica K-medias para encontrar el número de segmentos de mercado
- Por medio del coeficiente de Silhouette se evaluaron los resultados

(Kansal, Bahuguna, Singh, & Choudhury, 2018)

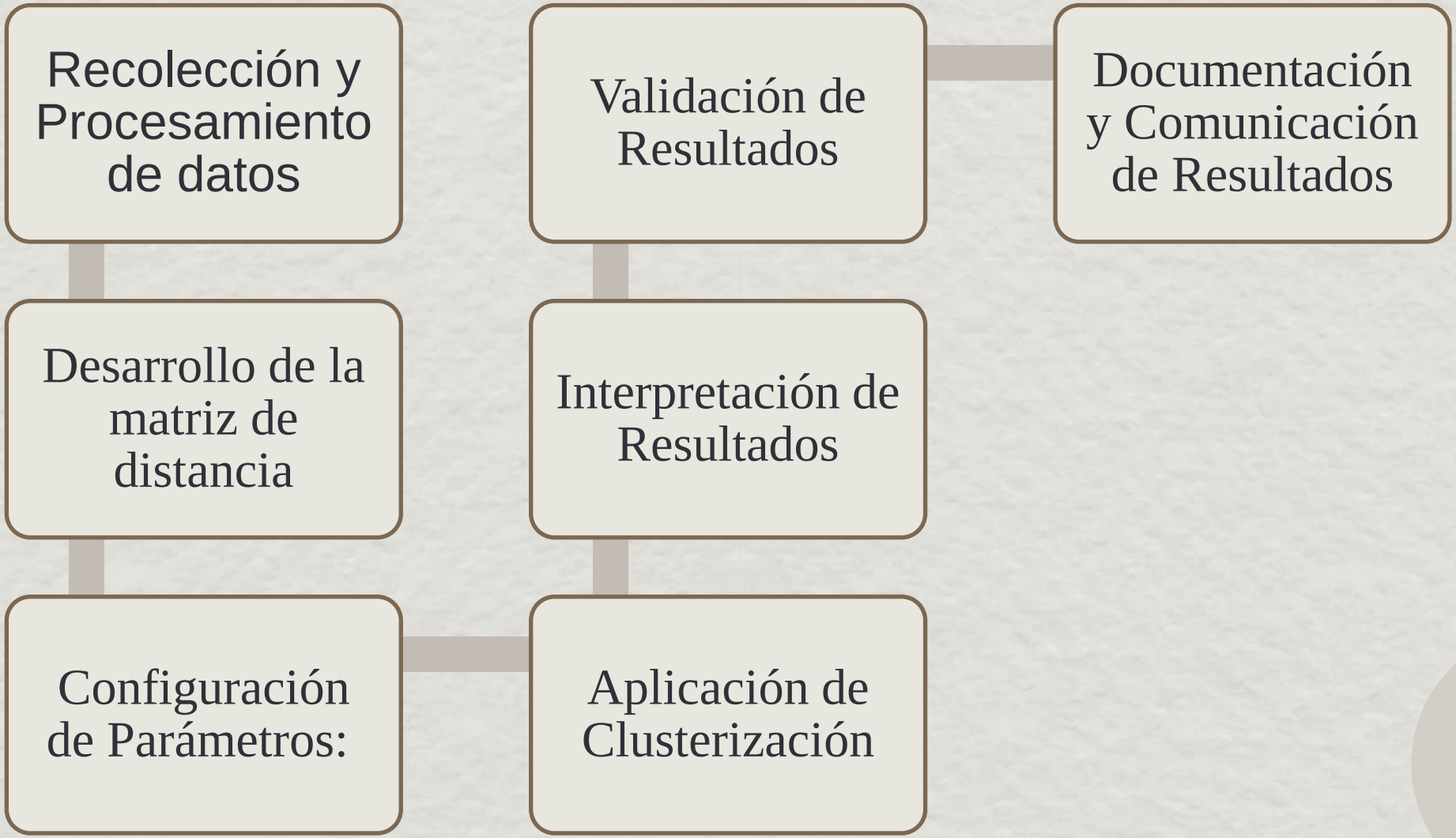
Análisis de Clúster

- Se escalan las variables de la base de datos
- Se realiza la agrupación por k-medias
- Se realiza también agrupación por método de aglomeraciones
- Se realiza agrupación por turnos medios (mean shift clustering)
- Por medio del coeficiente de Silhouette se comparan los resultados de cada agrupación

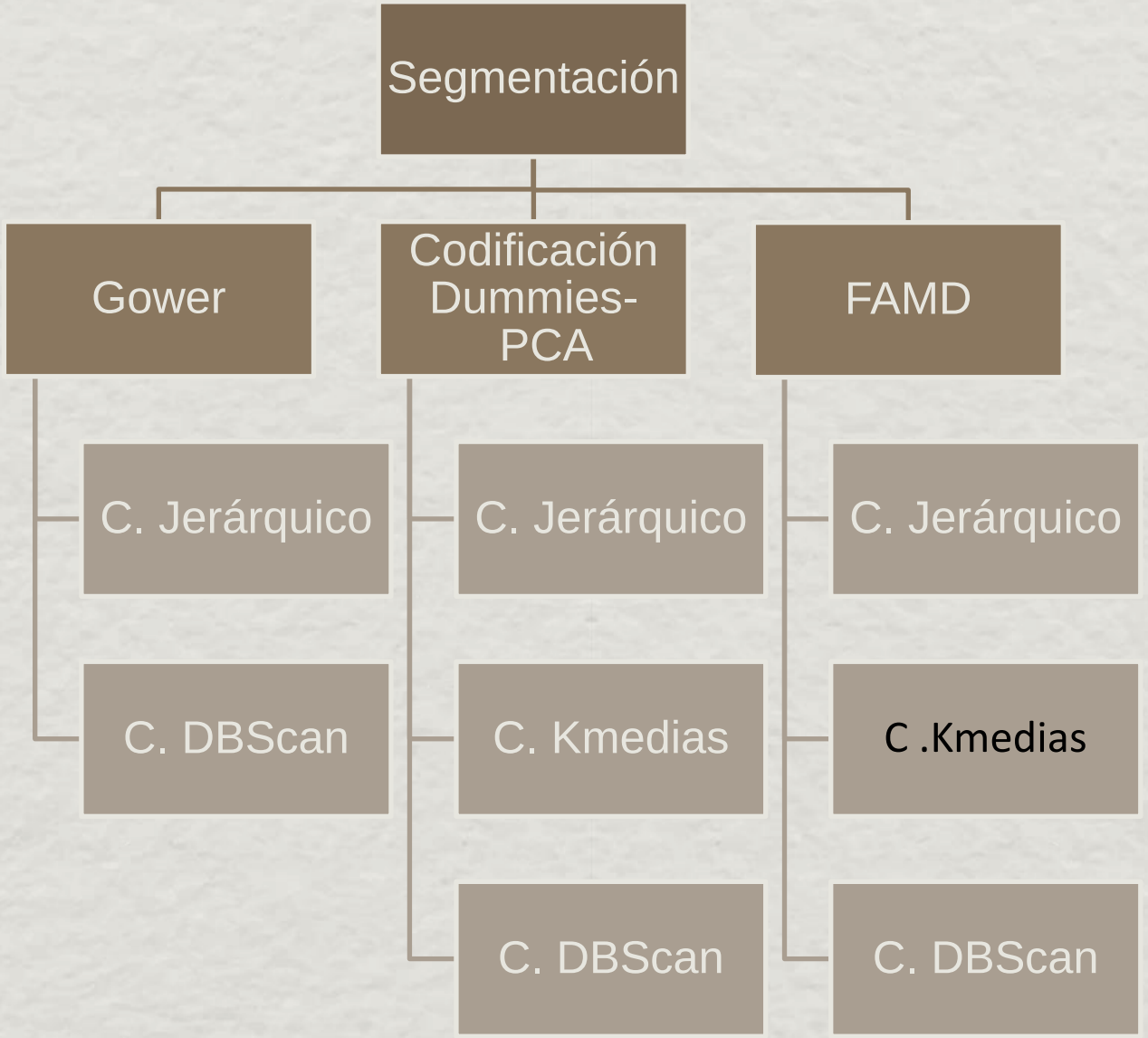
Otros casos de estudio

- Agrupación por k modas
- Agrupamiento jerárquico
- Agrupación por método de aglomeración
- Agrupamiento restringido
- Modelos latentes de clase

Metodología propuesta



Modelos de clúster a evaluar



Índices de evaluación

Coeficiente de Silhouette

1 Alta calidad, con instancias bien separadas

0 Indica que puede haber solapamiento entre los clústeres

-1 Los clústeres tienen una mala calidad

Davies-Bouldin

Cercano a **0** indica una buena calidad de los clústeres, con una separación clara entre ellos y una baja dispersión interna

Análisis de impacto

Se evalúa la calidad de las audiencias generadas por el número de sesiones asociadas, y la consistencia con las expectativas del ejercicio

Resultados y Discusión

Modelo	Silhouette	Davies-Bouldin
Clúster Jerárquico Gower	0,3	1,425
Clúster DBSCAN Gower	0,251	1,708
Clúster Kmeans Dummies con PCA	0,249	1,604
Clúster Jerárquico Dummies con PCA	0,041	3,416
Clúster DBSCAN Dummies con PCA	-0,024	2,962
Clúster Kmeans FAMD	0,005	6,448
Clúster Jerárquico FAMD	-0,252	8,361
Clúster DBSCAN FAMD	-0,044	7,647

Modelo	# Clusters	#Clusters <80%	Primer Cluster	Sesiones 1º Cluster	Sesiones por Cluster
Clúster Jerárquico Gower	62	28	31	6,33	1,61
Clúster DBSCAN Gower	65	28	0	6,33	1,54
Clúster Kmeans Dummies con PCA	46	23	4	7,18	2,17
Clúster Jerárquico Dummies con PCA	51	29	51	12,13	1,96
Clúster DBSCAN Dummies con PCA	141	63	-1	18,27	0,71
Clúster Kmeans FAMD	55	28	13	16,95	1,82
Clúster Jerárquico FAMD	98	40	27	6,08	1,02
Clúster DBSCAN FAMD	85	40	-1	6,15	1,18

Clúster # 31

Genero: Female

Rango de Edad: 35-44

Región: California

Intereses de Marca: News & Politics', 'Media & Entertainment', 'Lifestyles & Hobbies', 'Technology', 'Shoppers', 'Travel', 'Food & Dining', 'Beauty & Wellness', 'Home & Garden', 'Sports & Fitness', 'Banking & Finance', 'Vehicles & Transportation', 'Avid News Readers', 'Light TV Viewers', 'Movie Lovers', 'TV Lovers', 'Shutterbugs', …

Conclusión

Conclusiones

- El análisis jerárquico demostró ser adecuado para esta base de datos con variables categóricas, aunque DBSCAN y K-medias también mostraron resultados prometedores, respaldando hallazgos previos en la literatura.
- Las características del clúster más grande, en términos de sesiones, son coherentes con el análisis descriptivo previo, lo que sugiere que estos resultados son esperados y proporcionan una base sólida para la segmentación.
- El número de audiencias obtenido a través del análisis de clúster es suficiente para iniciar campañas digitales focalizadas de inmediato, lo que puede generar resultados positivos en las próximas etapas de posicionamiento y focalización en los resultados del negocio
- Se recomienda refinar el análisis de clúster en futuras etapas, especialmente cuando se disponga de más datos y atributos de usuarios.

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