Presentado por: Wendy Barreda, Harold Rodríguez, Jose Rivera y Carlos Andrés Rodríguez

ENTREGA FINAL DE PROYECTO DE APRENDIZAJE NO SUPERVISADO

Título: Arquetipos de clientes de banco Tibamoa

- **Resumen**

Los arquetipos de clientes son modelos que comparten características comunes entre los miembros que lo conforman. En termino generales, estas características pueden ir desde el estilo de vida hasta los comportamientos y necesidades de los clientes. En este caso la tarea se centra en información bancaria de 215 usuarios que se quieren clasificar de tal forma que el banco pueda entender las necesidades de cada subconjunto y planear estrategias futuras. Estas clasificaciones se basan en información como el género, estrato, actividad económica, estado civil, edad e información bancaria de los mismos.

- **Introducción**

Actualmente la definición de arquetipos es una técnica aplicada para el mejoramiento de Customer Experience (Experiencia de Clientes), que busca establecer estos subconjuntos y así poder generar experiencias positivas a los clientes y, que ellos de alguna manera sientan que son personalizadas. Nuestro cliente es una empresa consultora que realiza este tipo de análisis y, actualmente, los hace de manera cualitativa, casi, se podría decir que artesanal. Se nos suministró una base de datos real bancaria (sin datos sensibles y con un nombre ficticio de banco) para poder desarrollar este ejercicio.

Realizamos una exploración de literatura nacional e internacional para ejercicios similares donde encontramos, entre otros, los siguientes casos similares:

a) Proyecto de “Segmentación de clientes con afectación en sus servicios Área Analytics TIGO” [1]. Un proyecto realizado por alumnos de la Universidad de Antioquia donde se busca segmentar a los clientes con afectaciones por los servicios de la compañía TIGO con la finalidad de atender las necesidades y aumentar la satisfacción de estos. Para el caso del presente proyecto se implementó el algoritmo no supervisado K-Means al igual que en el caso de TIGO buscando segmentar los clientes y ofrecer mejores servicios a estos. Adicionando otros algoritmos para comparar resultados y tomar la mejor decisión.
b) Artículo publicado en la IEEE “Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning” [2]. Este artículo se publicó como parte de la 3rd International Conference on Advances in Computing, donde los autores se enfocan en la segmentación de clientes para la venta de productos, buscan conocer al cliente para descubrir tendencias que les ayuden a dirigir los productos adecuados a los clientes para el aumento de ventas. Este trabajo fue de gran utilidad para el presente proyecto debido a que se contaba con muchas variables y al final se debían clasificar y seleccionar la información demográfica y los datos de comportamiento más importantes tal como fue implementado en nuestro caso.
c) Oferta de este tipo de servicios por empresas que tienen información como es el caso de Brands by Rappi y su producto Behavioral Audiences [3], que la segmentación de los usuarios en 26 categorías generales y más de 280 subcategorías dependiendo de los datos. Otra compañía que ofrece soluciones de segmentación es Zendesk, una empresa de CRM dedicada especialmente a la atención al cliente. Zendesk en 2020 publicó un artículo “Segmentación de clientes: cómo llegar efectivamente a nuestro público objetivo” [4], donde habla de la importancia de la segmentación de clientes puntualizando que esto ayuda a mejorar el proceso de marketing, atraer y convertir leads, diferenciarse en el mercado, mejorar la experiencia del cliente, entre otras. Para el caso del proyecto fue de gran orientación este articulo para enfocar el objetivo principal descartando de un gran grupo de variables y seleccionando el público objetivo más acorde a las necesidades del cliente.

De todos los casos mencionado se puede observar que la segmentación de clientes ha sido aplicada en diversos ámbitos, tanto en el contexto nacional (Ej., de Tigo) e internacional y, que tiene un potencial importante de ser aplicado bajo el enfoque propuesto en este caso. Algo que se reafirma con la bibliografía es que la implementación de algoritmos no supervisados es adecuada para el objetivo de este proyecto.

- **Materiales y Métodos**

El dataset proporcionado por la consultora perteneciente al Banco Tibamoa (empresa ficticia), cuenta con 215 registros y 70 variables crudas. El preprocesamiento de información consistió en eliminación de columnas con índices generales, creación de variables dummie para las variables categóricas e imputación de valores faltantes en las variables numéricas. Cabe destacar que los valores faltantes provenían del resultado de medir el Net Promotor Score (NPS), donde el cliente manifestó que no conocía o no usaba el producto. Una vez procesada la información se obtuvieron 192 variables.

Debido a la cantidad de variables se buscó una reducción de la dimensionalidad aplicando estrategias como SVD y PCA. Ambos métodos sugieren 62 componentes que explican al menos el 90% de la información. Para el proceso de implementación del algoritmo de clasificación se eligió PCA.

Posterior a la implementación de PCA, se analizaron los pesos las primeras 10 componentes para obtener cuáles eran las variables con mayor impacto que explicaban aproximadamente el 40% de la varianza. Para ello ponderamos el peso en valor absoluto de cada variable en cada componente y lo sumamos a lo largo de los primeros componentes. Esto nos arrojó 11 variables, todas numéricas las cuales describimos a continuación:

Variables relacionadas con NPS

calificación	NPS Probabilidad		NPS Web		NPS App		NPS Banca en línea		NPS whatsapp		NPSTeléfono		NPS Cajeros electrónicos		NPS Correo electrónico		NPS Oficinas	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%		
0	10	4,4%																
1	6	3,4%	10	4,9%	9	4,1%	10	4,4%	20	10,2%	21	9,7%	8	3,5%	14	6,7%	15	7,5%
2	2	1,2%	3	2,0%	2	1,4%	4	2,2%	9	4,7%	6	3,0%	4	2,5%	5	2,5%	10	5,6%
3	3	1,9%	5	2,6%	7	3,8%	5	2,3%	7	4,0%	6	3,5%	3	1,5%	4	2,1%	7	3,6%
4	9	4,3%	6	3,0%	6	3,3%	4	1,9%	11	4,6%	1	0,6%	2	1,0%	3	1,3%	7	3,4%
5	7	3,9%	14	6,5%	8	3,5%	9	4,4%	11	6,0%	11	4,5%	7	2,5%	7	3,5%	17	6,6%
6	4	2,3%	5	2,3%	4	2,6%	3	2,0%	6	2,0%	5	2,4%	5	2,2%	8	2,6%	10	5,2%
7	13	6,3%	11	5,3%	5	1,6%	15	7,5%	15	6,3%	16	8,1%	8	3,3%	6	3,2%	13	4,6%
8	21	8,9%	22	9,5%	28	12,0%	17	7,2%	20	8,5%	15	7,4%	23	12,7%	9	5,3%	21	9,4%
9	36	15,6%	30	14,3%	34	15,7%	30	13,9%	18	8,7%	7	4,0%	22	11,1%	11	4,4%	25	12,5%
10	104	47,7%	69	31,0%	91	40,3%	66	30,8%	59	24,2%	39	16,6%	81	33,3%	38	15,9%	75	34,7%
No conozco este canal			2	1,33%	2	1,22%	5	2,40%	2	1,28%	10	5,25%	3	1,87%	13	5,54%		
No uso este canal			38	17,32%	19	10,62%	47	21,12%	37	19,55%	78	35,04%	49	24,65%	97	46,99%	15	6,93%
Total general	215	100%	215	100%	215	100%	215	100%	215	100%	215	100%	215	100%	215	100%	215	100%

Ilustración 1. Estadísticas descriptivas de variables de NPS

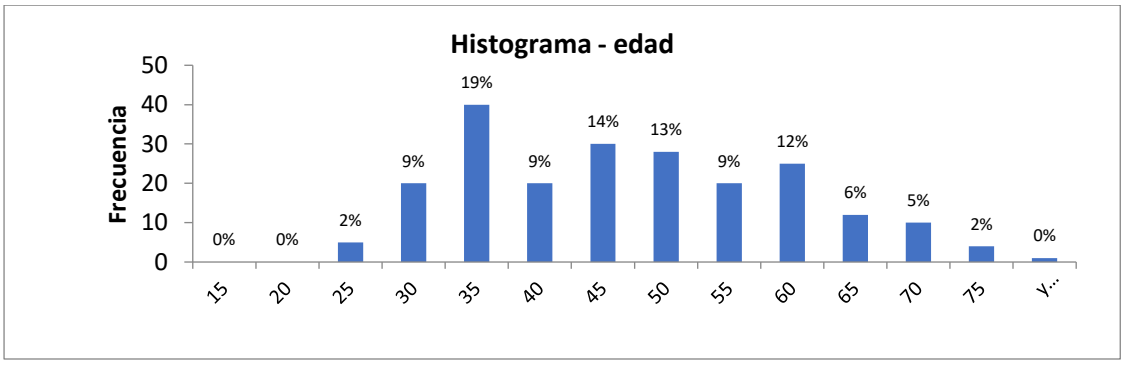


Ilustración 2. Histograma de edad

Dado que Los algoritmos de Aprendizaje no Supervisados se utilizan para agrupar los datos no estructurados según sus similitudes y patrones distintos en el conjunto de datos, se implementan K-medias, K-medoides y Clúster Jerárquico para comparar los resultados.

K-medias: El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster.

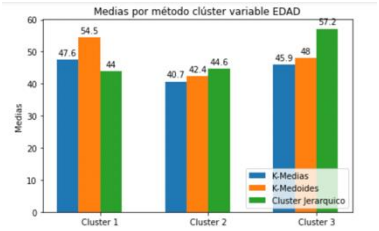
K-medoides: Divide la muestra, al igual que K-medias, en grupos minimizando la distancia entre ellos, pero asigna como centros del clúster (medoides) un punto de la muestra. Como parte del algoritmo se revisa iterativamente si los clústeres creados mejoran o no al intercambiar medoides internamente.

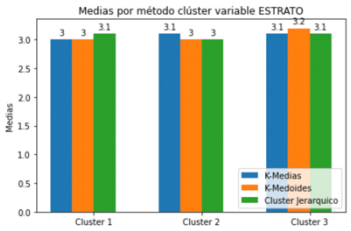
K-medias y K-medoides requieren como parámetro a priori el número de clusters hecho por el cual se utilizó el método del codo que utiliza la Varianza intra-clúster y el Índice de Silhouette para determinar el número óptimo de clusters dando como resultado 3.

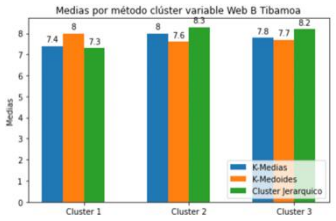
Clustering Jerárquico Aglomerativo: Estos métodos de jerarquías son excelentes para organizar información ya que se enfocan en relaciones anidadas entre los puntos de datos, el proceso inicia con cada uno de los puntos como su propio clúster y luego se van agrupando con los puntos más cercanos formando nuevos clústeres. Para determinar el número de clústeres y la configuración de los parámetros affinity y linkage se analizaron diferentes dendogramas buscando los valores más adecuados, obteniendo los mejores resultados con la distancia Euclidiana y el método Ward, así como un numero de clústeres de cuatro para los datos reducidos con PCA y de tres clústeres para los datos con SVD.

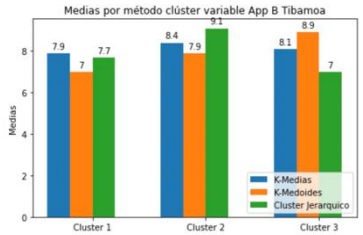
- **Resultados y discusión:**

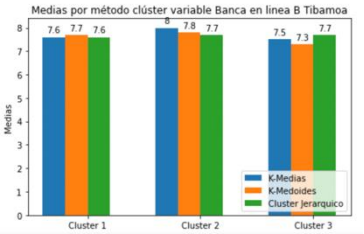
Se implementaron los algoritmos y se compararon el comportamiento de variables que arrojaban cada uno, teniendo en cuenta cada una de las 11 variables que definimos para revisar con más detalle, dados los pesos encontrados en el PCA, a continuación, se muestra el resultado de este análisis

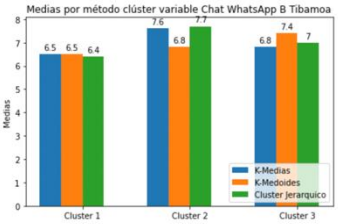


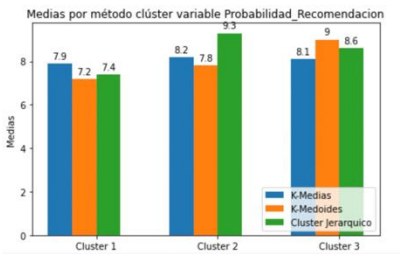


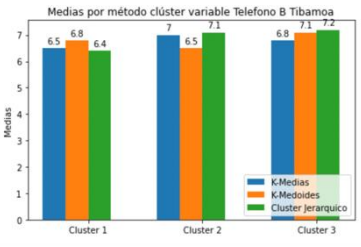


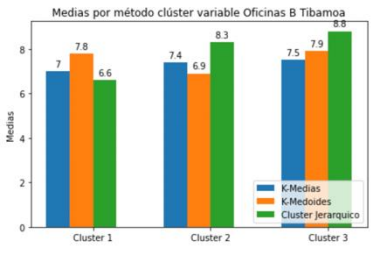


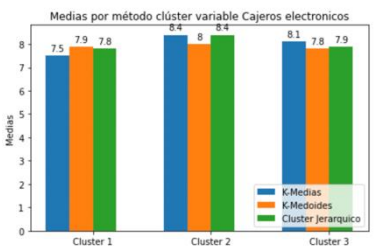


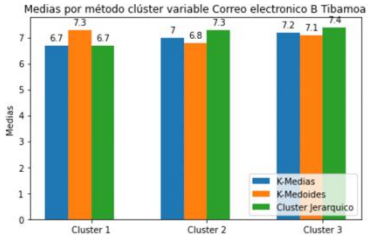












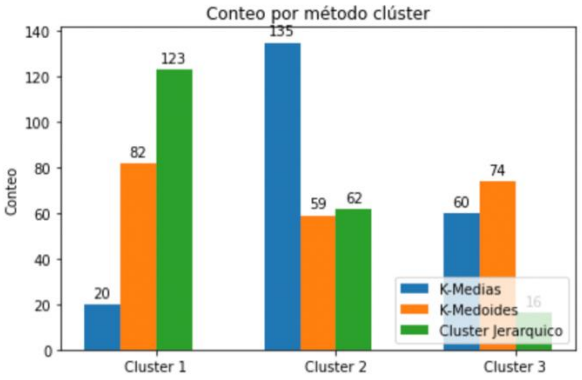


Ilustración 3. Conteos por clúster

Se revisó también la cantidad de individuos en cada clúster y con esa información y, observando los comportamientos de las variables, se piensa que el mejor método para hacer la clústerización es el de k-medoides. Se tiene en cuenta que clasifica de manera más homogénea los grupos y muestra de manera aceptable diferencias en las variables que nos pueden servir para el objetivo. Estas diferencias se detallan en el siguiente cuadro como características distintivas de cada clúster que dan pie a aprovechar las oportunidades descritas también en el cuadro de la siguiente página.

Una limitación de la implementación fue no conocer todas las posibilidades que nos ofrecían las metodologías del curso a tiempo porque también se habría podido incluir información de encuestas que no estaba estructurada. Por otro lado, la manera como se recogió la información en variables categóricas no ordenadas tampoco permitía mucho aprovechar esa información. Al parecer, al procesar esas variables como binarias (dummies), se perdía información y se volvía casi irrelevante entre el resto de las variables. Para futuros ejercicios vale la pena revisar las encuestas para que su diseño tenga tanta información de ese tipo.

Características y oportunidades de líneas de acción por clúster

Clúster	Características distintivas	Oportunidades de líneas de acción
1	• % de individuos en la muestra 38% • Promedio de edad 54,5 años • NPS de APP = 7 • NPS chat WhatsApp= 6,5 • NPS oficinas = 7,8 • Probabilidad de recomendación 7,2.	Bajo NPS o neutro, especialmente en temas relacionados con canales tecnológicos. Oportunidad de brindar atención más personalizada para que realice sus transacciones o, explicar los canales digitales. El NPS mejora para el tema de oficinas
2	• % de individuos en la muestra 27% • Promedio de edad 42,4 años • NPS de APP = 7,9 • NPS chat WhatsApp= 6,8 • NPS oficinas = 6,9 Probabilidad de recomendación 7,8.	NPS o neutro cerca de ser promotor, valora mejor los canales tecnológicos. No valora bien oficinas presenciales. Presenta la oportunidad de desarrollar más los canales digitales para evitar que debe desplazarse a oficinas.
3	• % de individuos en la muestra 34% • Promedio de edad 48 años • NPS de APP = 8,9 • NPS chat WhatsApp= 7,4 • NPS oficinas = 7,9 Probabilidad de recomendación 9.	NPS promotor, valora muy bien la APP y mejor el WhatsApp que el resto. También valora bien oficinas presenciales. Presenta la oportunidad de ofrecer nuevos productos y darle incentivos por recomendar nuevos clientes al banco.

- **Conclusión:**

De acuerdo con lo expuesto vemos como se analizó la información y se aplicaron técnicas para limpiar, procesar y clústerizar la información de las encuestas. El ejercicio arroja conclusiones útiles tanto para aprovechar las características de los clústers como para realizar de mejor manera ejercicios similares en el futuro. Las principales conclusiones son:

a) Vale la pena revisar para futuros ejercicios no recoger tantas variables que se requieran recategorizar de manera binaria (dummies), porque tener tantas variables de este tipo parece que no ayuda mucho para lograr más información de cara a clústerizar.
b) En este caso se observa que el algoritmo k-medoides genera unos clústers más balanceados. Seguramente tiene que ver con que maneja mejor los outliers.
c) Se generan 3 clústers y tiene sentido dado que en un ejercicio realizado de manera cualitativa por la consultora se definieron 4 arquetipos, aunque, el ejercicio cualitativo tuvo en cuenta más variables de comportamiento. El punto es que este ejercicio no se considera tan confiable dada su naturaleza cualitativa.
d) Las recomendaciones por cada clúster se consideran valiosas y permitirían atacar las necesidades de cada cliente de cara a mejorar el NPS de cada clúster.

Bibliografía

[1] S. López Stan, “Segmentación de clientes con indisponibilidad o afectación en sus servicios, Área Data Analytics TIGO”, trabajo de grado profesional, Ingeniería de Sistemas, pregrado, Universidad de Antioquia, Ciudad Universitaria, 2022. Recuperado 1 de septiembre de 2022, de https://bibliotecadigital.udea.edu.co

[2] M. F. Alam, R. Singh and S. Katiyar, "Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning," *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 94-98, Recuperado 1 de septiembre de 2022, de https://ieeexplore.ieee.org

[3] Brands by Rappi. (2022). For Agencies. Recuperado 1 de septiembre de 2022, de https://brands.rappi.com/for-agencies

[4] Douglas Da Silva, D. S. (2020, 25 septiembre). Segmentación de clientes: cómo llegar efectivamente a nuestro público objetivo. Blog de Zendesk. Recuperado 1 de septiembre de 2022, de https://www.zendesk.com.mx/blog/segmentacion-de-clientes/