

Clustering para determinar los arquetipos de afiliado para una EPS y mejorar la experiencia del afiliado al brindar atención, priorización y comunicación especializada

03.09.2022

—

Yovany Gasca, Christian González, Jolman Salinas

Grupo 2

Aprendizaje no Supervisado

MIAD Universidad de los Andes

Resumen.

Este estudio busca determinar mediante criterios estadísticos de aprendizaje no supervisado, si es posible agrupar los afiliados de la EPS en grupos homogéneos, con el fin de identificar patrones de comportamiento que permitan mejorar la calidad de los servicios prestados por la EPS.

Mediante un ejercicio de clustering determinaremos los grupos de afiliados similares en términos de salud, riesgos, hábitos y uso de canales de atención, de esta manera podremos proponer estrategias de atención que mejoren la experiencia de los afiliados y la calidad de los servicios prestados por la EPS especializándose en cada arquetipo de manera independiente.

La EPS podrá medir el NEP (Net Promoter Score) percibido por sus afiliados en cada arquetipo, y así identificar las áreas de oportunidad para mejorar la experiencia de sus afiliados.

Introducción

Dada la naturaleza del negocio de las Entidades Promotoras de Salud (EPS), uno de los principales flancos de negocio es la retención de los afiliados. Según la Super Intendencia de Salud [1] el 40% de las multas recibidas en las EPS son causadas por fallas en la prestación de servicios de salud.

Dada esta situación, las entidades crean áreas robustas de Experiencia del Afiliado concentradas en la atención de los mismos, la comunicación acertada por los canales adecuados, la priorización de los servicios de salud, la atención de los afiliados en caso de emergencia, etc.

Dichas áreas saben que la experiencia del afiliado es un factor determinante para la satisfacción del cliente y que dicha experiencia debe ser medida y atendida de la forma más personalizada posible, especialmente en entornos donde se manejan millones de afiliados.

Los modelos de arquetipos de afiliado son una herramienta para la mejora de la experiencia del afiliado dado que logran agrupar a los afiliados teniendo en cuenta sus necesidades, riesgos, características y condiciones de salud.

Unos arquetipos que garanticen homogeneidad dentro de los grupos de afiliados y heterogeneidad entre los grupos de afiliados son la clave para que la experiencia del afiliado sea más individualizada y que las estrategias de atención, comunicación y priorización sean más eficientes.

[1]:

https://docs.supersalud.gov.co/PortalWeb/planeacion/Estadisticas%20SNS/BoletinEstadistico_2020.pdf

Revisión preliminar de la literatura

En la literatura hay aplicaciones de clustering a clientes con diversos objetivos, mejora en la toma de decisiones, mejora en los procesos comerciales, conocimiento del cliente entre otras, en Supporting healthcare management decisions via robust clustering of event logs [cita] vemos cómo se estudian los flujos de trabajo para la generación de informes que mejoren la toma de decisiones en la atención médica; la metodología utilizada es la minería de procesos que agrupa flujos de clientes y propone una métrica de similitud que funciona bien para minimizar el efecto del ruido de los valores atípicos. (Delias & Doumpos, 2015,203-213)

En "Survey of Clustering Algorithms for Categorization of Patient Records in Healthcare" se realiza la comparación de varios algoritmos de clusterización para determinar o clasificar información médica relacionada a problemas de salud, es importante para nuestro proyecto debido a que propone el algoritmo k-means, K Medias armónicas y medias armónicas Hybrid Fuzzy K (HFKHM) y maneja una cantidad importante de atributos como como la categorización del tumor, radio, textura, tersura y compacidad del tumor y los datos demográficos de los pacientes. (Narmadha & Balamurugan, 2016)

En un sector relacionado como es el de seguros se realizó clusterización de clientes en "Insurance customer segmentation using clustering approach" para determinar el nivel de rentabilidad en tres grandes grupos de tal manera que se los esfuerzos comerciales se optimicen en clientes "rentables", para ello utilizan k-means y se utiliza el paquete 'NbClust' de R para determinar la cantidad ideal de clusters. Los resultados de este estudio permiten establecer planes de marketing y comunicación adecuados para los 3 clusters hallados. (Abolmakarem et al., 2017, 18-39)

Un enfoque adicional lo aporta "Applied Research of Ant Colony Clustering Algorithms in Healthcare Consumer Segments" el cual tiene como objetivo la comprensión del mercado de los hospitales y la creación de estrategias de gestión de servicios y aplica un algoritmo de optimización de colonias de hormigas, que es un algoritmo evolutivo biónico emergente, tiene una gran robustez y adaptabilidad. Los resultados muestran eficiencia en la segmentación de clientes. (Jiang & Wang, 2011,335-342)

Descripción de los datos.

El dataset está compuesto por 1,044,692 observaciones y 79 variables, se trabajará en identificar las variables que sean de mayor interés para alcanzar el objetivo del proyecto y se desarrollará un análisis exploratorio y se creará a partir de estas variables un diccionario de datos.

Variables disponibles

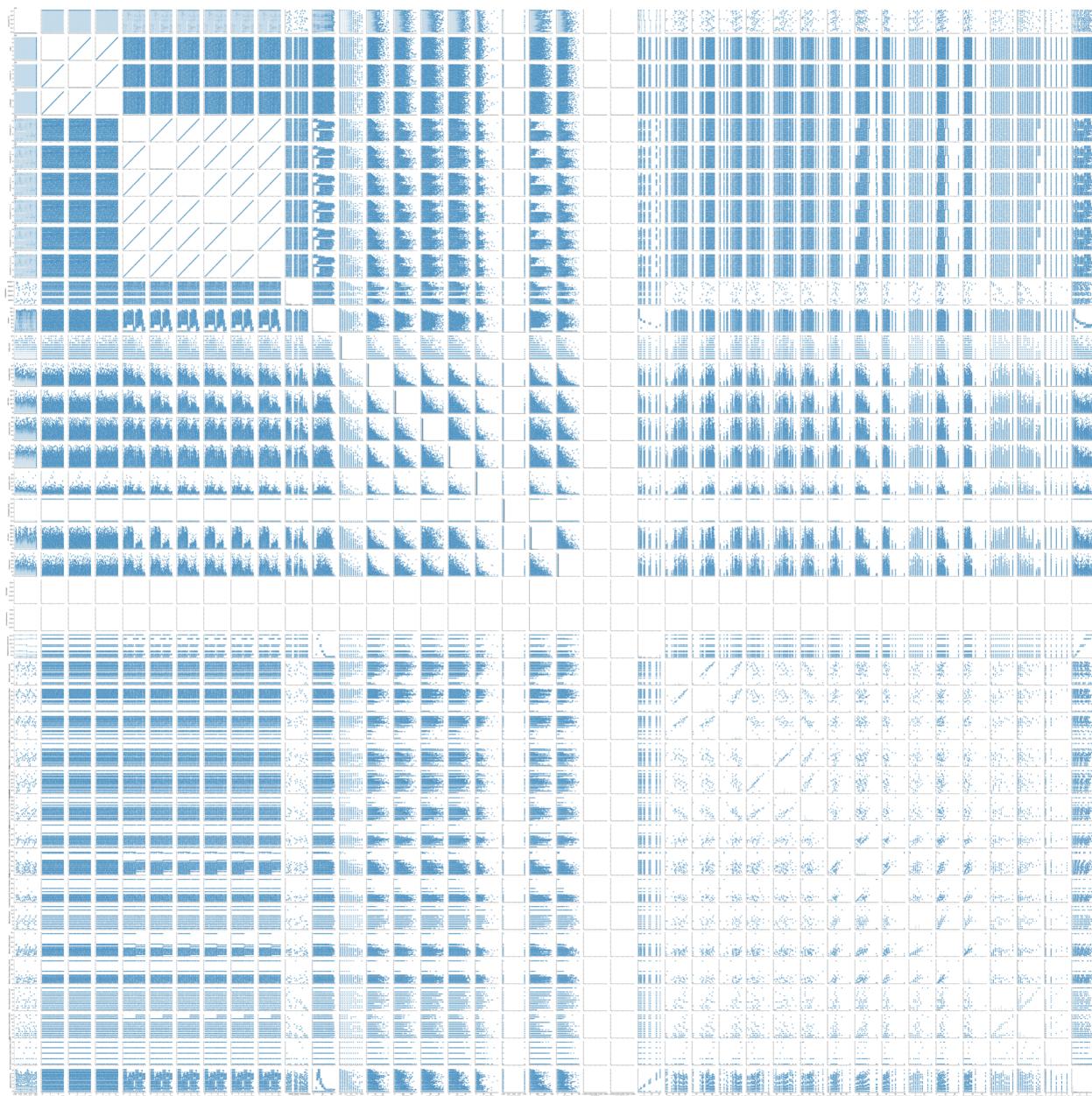
Variables			
Unnamed: 0.2	Tipoidentificacion_datos	Municipios	APP 2020
index	Nombre	Zonas	CallCenter 2020
Unnamed: 0	Apellido 1	CodigoDane	OAA 2020
Id_Afiliado	Apellido 2	REGIMEN	AsesorAClic 2020
Unnamed: 0.1	FechaNacimiento	AfilidoPAC	ViveDigital 2020
Unnamed: 0.1.1	Genero	Edades	IVR 2020
Unnamed: 0.1.1.1	TipoAfiliado	Generacion	ChatEnLinea
Unnamed: 0.1.1.1.1	Regionales	PQRS 2020	TELEFONO
Unnamed: 0.1.1.1.1.1	Zonales	Multiquejoso	CorreoElectronico
Unnamed: 0.1.1.1.1.1.1	Departamento	PortalWeb 2020	Digitalizacion generacion
Telfono celular Nacional	Tableta (Cabecera Mpal)	categoriaNivelDeUsoAsesorAClic	Etnia_datos
Telfono celular (Cabecera Mpal)	Tableta (Centro Poblado y Rural)	categoriaNivelDeUsoChatEnLinea	NivelEducativo_datos
Telfono celular (Centro Poblado y Rural)	Otros dispositivos Nacional	categoriaNivelDeUsoIVR	Tipoidentificacion_idFalsos
Computador de escritorio Nacional	Otros dispositivos (Cabecera Mpal)	digitalizacionOficial	Regimen
Computador de escritorio (Cabecera Mpal)	Otros dispositivos (Centro Poblado y Rural)	mujerEnEdadFertil	Condicion_Cronica_idSalud
Computador de escritorio (Centro Poblado y Rural)	categoriaNivelDePQRS	categoriaNivelDeDigitalizacionOficial	NivelRiesgo_idSalud
Computador porttil Nacional	categoriaNivelDeUsoPortalWeb	idFalso	Discapacidad_idSalud
Computador porttil (Cabecera Mpal)	categoriaNivelDeUsoAPP	Condicion_Cronica_datos	Etnia_idSalud
Computador porttil (Centro Poblado y Rural)	categoriaNivelDeUsoCallCenter	NivelRiesgo_datos	NivelEducativo_idSalud
Tableta Nacional	categoriaNivelDeUsoOAA	Discapacidad_datos	

Muestra de los valores

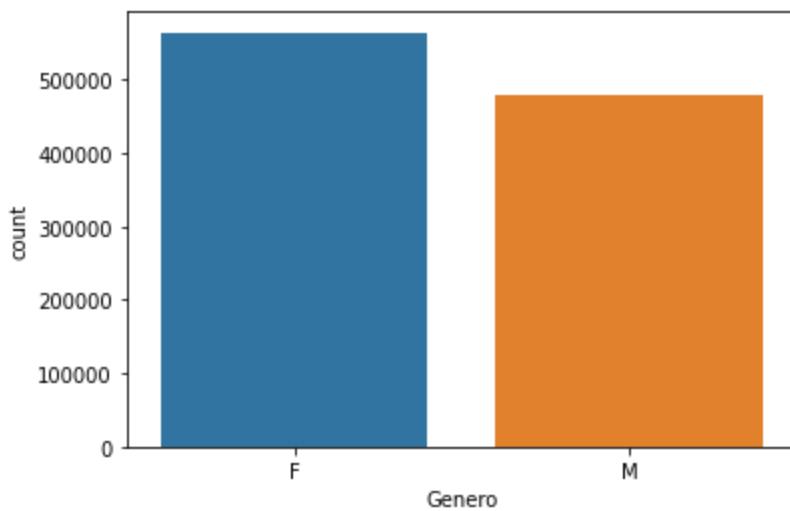
Computador porttil Nacional	Computador porttil (Cabecera Mpal)	Computador porttil (Centro Poblado y Rural)	Tableta Nacional	Tableta (Cabecera Mpal)	Tableta (Centro Poblado y Rural)	Otros dispositivos Nacional	Otros dispositivos (Cabecera Mpal)	Otros dispositivos (Centro Poblado y Rural)	digitalizacionOficial
1.044692e+06	1.041771e+06	1.039839e+06	1.044692e+06	1.041771e+06	1.039839e+06	1.044692e+06	1.041771e+06	1.039839e+06	1.044692e+06
2.998010e+01	2.407271e+01	1.939992e+01	6.680784e+00	5.474807e+00	5.093012e+00	5.488804e+00	2.765130e+00	7.868680e-01	1.552472e+00
5.223671e+00	7.370981e+00	4.720112e+00	2.559585e+00	2.977734e+00	3.905681e+00	3.169081e+00	2.790274e+00	1.080974e+00	1.655563e+00
1.700000e+01	1.300000e+01	1.300000e+01	2.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
2.800000e+01	1.700000e+01	1.600000e+01	5.000000e+00	4.000000e+00	2.000000e+00	3.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
3.000000e+01	2.400000e+01	2.000000e+01	6.000000e+00	6.000000e+00	5.000000e+00	5.000000e+00	1.000000e+00	0.000000e+00	8.830500e-01
3.100000e+01	3.000000e+01	2.300000e+01	7.000000e+00	7.000000e+00	6.000000e+00	7.000000e+00	5.000000e+00	1.000000e+00	2.698150e+00
5.700000e+01	5.900000e+01	5.900000e+01	1.400000e+01	2.400000e+01	2.400000e+01	1.100000e+01	1.200000e+01	4.000000e+00	6.208800e+00

Correlograma

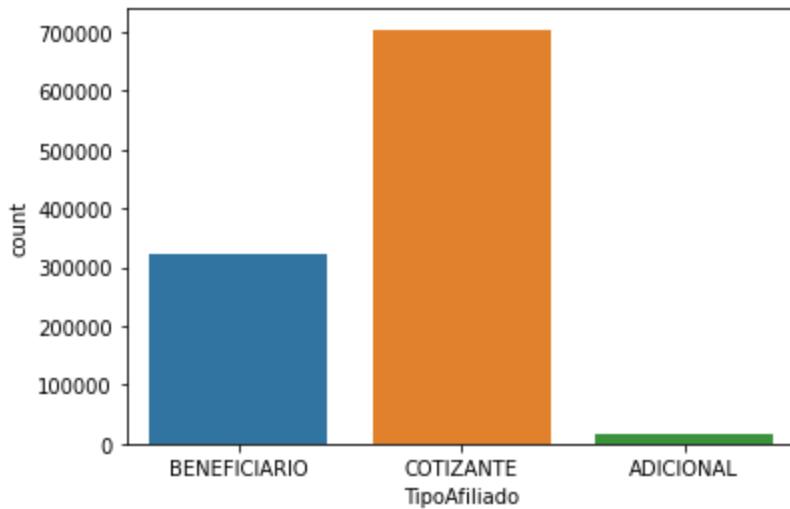
La imagen original se puede descargar en la siguiente ruta [Correlograma](#)



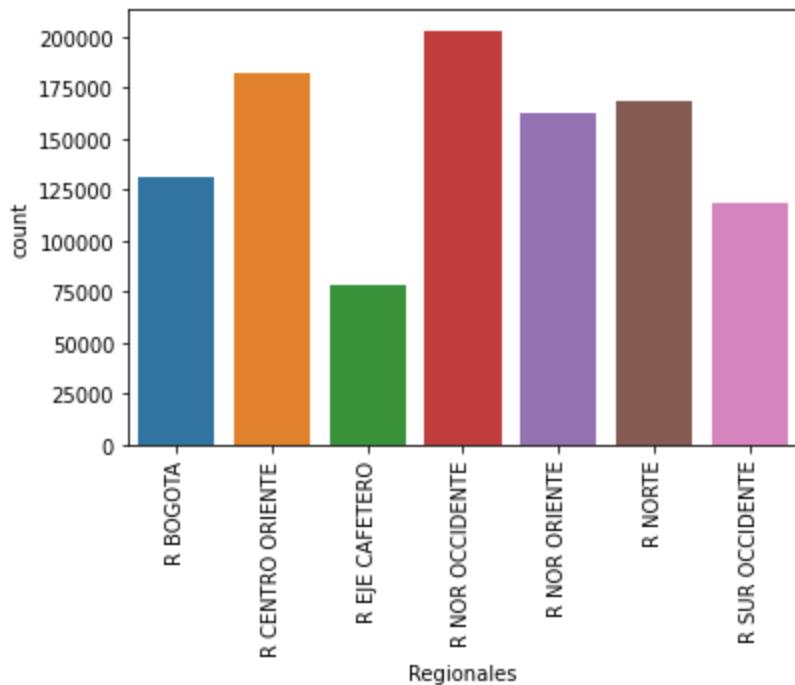
Contamos con un 54% de registros de género femenino del total de 1044692 registros de afiliados únicos contenidos en la base de datos.



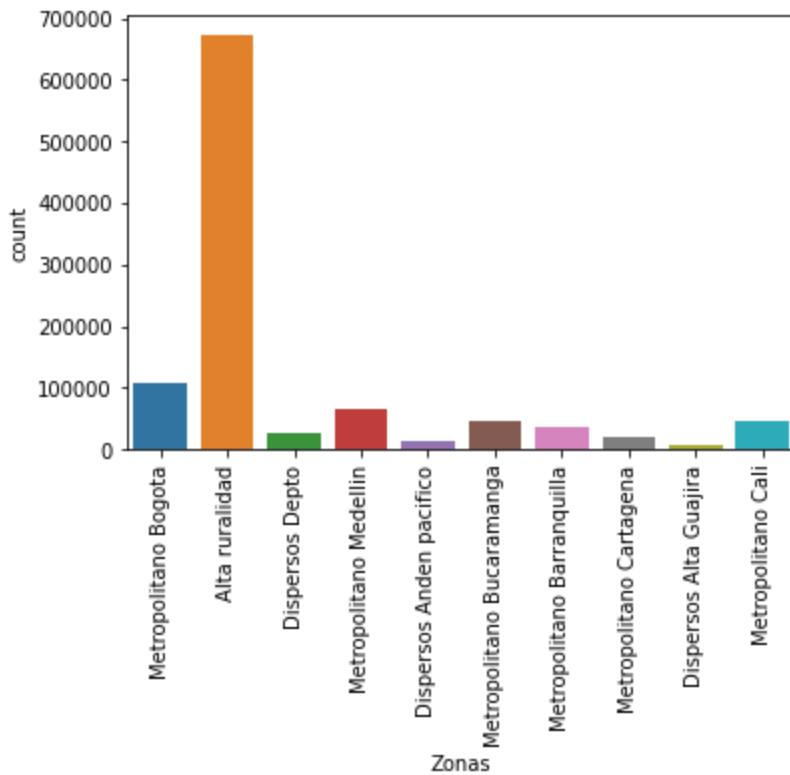
El tipo de afiliado cotizante es el más frecuente en los datos con un total de 704503 registros conformando el 67% del total de registros, esto es importante dado que indica las intenciones prestacionales de la EPS a clientes que en su mayoría cuentan con capacidad adquisitiva para pagar los servicios prestados, El tipo "Adicional" está conformado por personas adscritas a servicios militares, policiales, penitenciarios, entre otros.



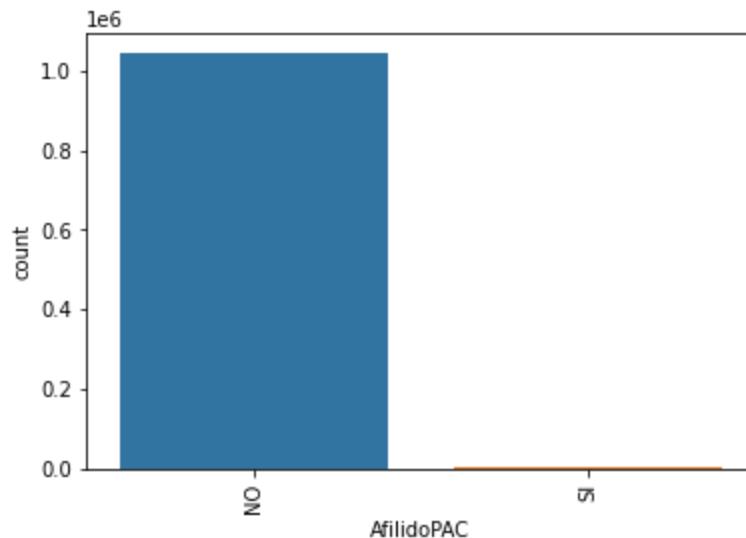
Esta EPS cuenta con cobertura por todo el territorio nacional, por lo que una buena segmentación debería considerar las variaciones sociodemográficas que se presentan en cada región del país.



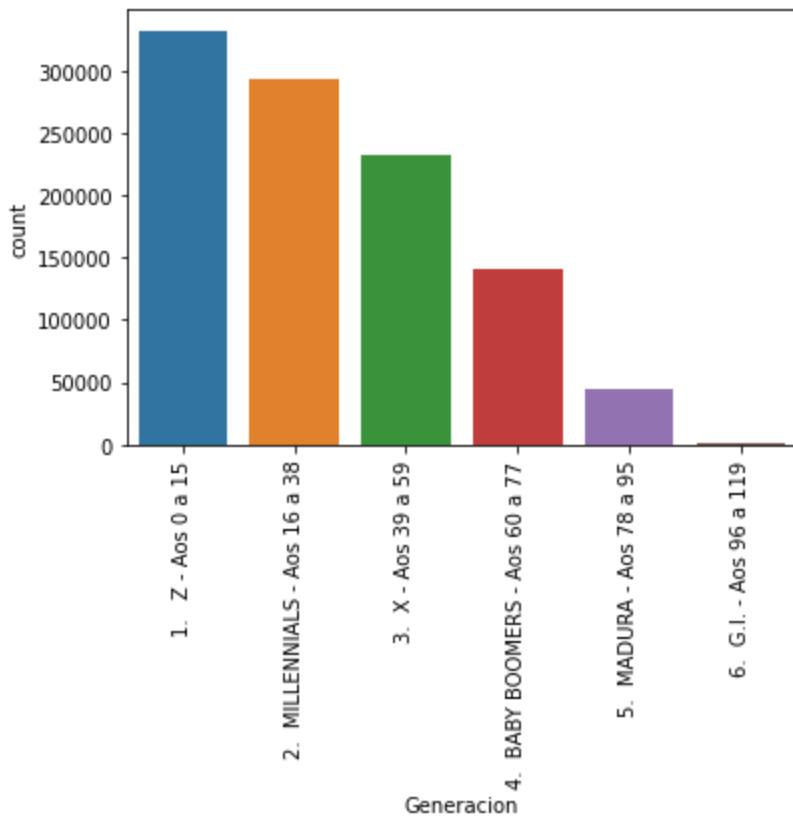
La variable Zonas debe ser una variable de segmentación principal ya que indica si el afiliado vive en una zona metropolitana de alta densidad o, por el contrario, vive en una zona rural más alejada de cabeceras municipales y por lo tanto con menor acceso a servicios de salud, educación, etc. Esto también se podrá ver reflejado en la manera en que usa los canales de la EPS siendo este un afiliado menos 'Digital'.



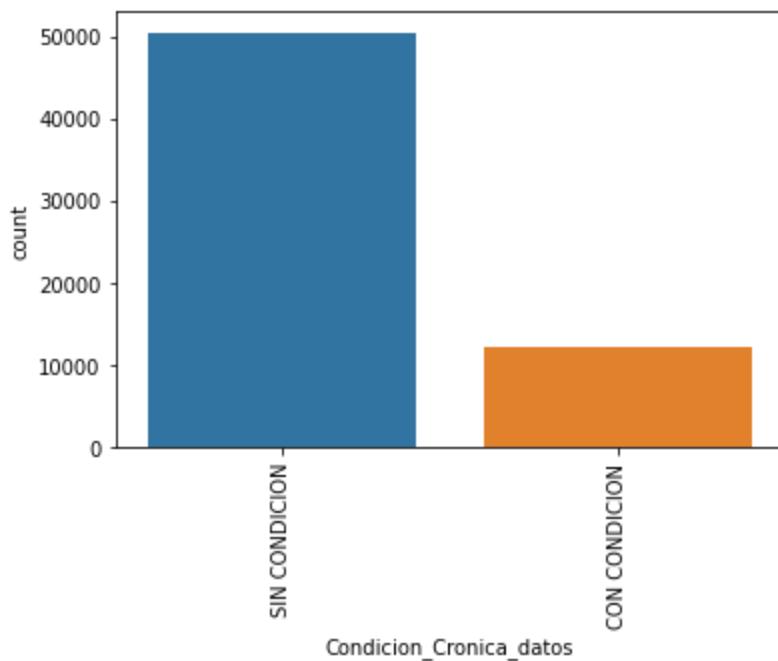
La EPS cuenta con un PAC (Plan Complementario de Salud) dirigido a afiliados con mayor poder adquisitivo, son un total de 1413 afiliados. El PAC tiene un costo mensual de \$360.000 COP, el cual es pagado por el afiliado y cada uno de los miembros del grupo familiar que estén inscritos.



En cuanto a la generación, La población Joven y Joven Adulta marca la mayor participación de afiliados, consecuentemente, al ser el régimen contributivo la mayoría de los afiliados, se puede concluir que la población joven y joven adulta está siendo sostenida en su mayoría por un menor porcentaje de la población adulta y adulta mayor.



Son 12127 (12%) afiliados que cuentan con una condición crónica, por lo tanto son quienes frecuentemente visitan los servicios de salud y resultan ser un costo importante para el sistema de salud. Entre las condiciones crónicas más frecuentes se encuentran: hipertensión arterial, diabetes mellitus, enfermedad pulmonar obstructiva crónica, enfermedad cardiovascular, enfermedad renal crónica, enfermedad hepática crónica, cáncer, enfermedad mental y otras enfermedades crónicas.



Propuesta metodológica

Este ejercicio está enfocado a identificar las características en común de los afiliados de las EPS, con el objetivo de identificar perfiles que permitan personalizar la atención recibida y mejorar su experiencia de usuario.

Para poder identificar a los afiliados que comparten características emplearemos algoritmos de **clustering** y sin embargo y debido a que la información a trabajar presenta una combinación tanto de variables categóricas y numéricas se plantea utilizar un algoritmo adecuado a la fuente de datos, nuestra aproximación inicial será por medio de la implementación del algoritmo de K-Prototype. K-Prototype es un método de agrupamiento basado en particiones. Su algoritmo es una mejora del algoritmo de agrupamiento K-Means y K-Mode para manejar el agrupamiento con tipos de datos mixtos.

Sin embargo, procesar millones de afiliados y decenas de características, puede resultar computacionalmente complejo, por lo que se consideró implementar un **algoritmo de reducción de dimensionalidad**, que nos permita retener el mayor porcentaje de variación de los datos en la menor cantidad de componentes principales con el objetivo de optimizar el desempeño de los algoritmos de ML, por lo que consideramos implementar **PCAMix** o **PCAmixdata** que nos permita el análisis tanto de variables continuas como categóricas.

En resumen la metodología que implementaremos se reduce a los siguientes pasos:

- 
1. Reducción de dimensionalidad
 2. Identificación de Clusters
 3. Medición de desempeño
 4. Comparación resultados
 5. Iteración
 6. Conclusiones.

Bibliografía

- Abolkarem, S., Abdi, F., & Khalili-Damghani, K. (2017, 09 1). Insurance customer segmentation using clustering approach. *International Journal of Knowledge Engineering and Data Minin*, 4(2016), 18-38.
<https://www.inderscienceonline.com/doi/abs/10.1504/IJKEDM.2016.082072>
- Delias, P., & Doumpos, M. (2015). Supporting healthcare management decisions via robust clustering of event logs. *ELSEVIER*, 84(2015), 203-213. Retrieved 09 03, 2022, from
<https://www.sciencedirect.com/science/article/abs/pii/S0950705115001501>
- Jiang, Q.W., & Wang, J. (2011). Applied Research of Ant Colony Clustering Algorithms in Healthcare Consumer Segment. *Communications in Computer and Information Science*, 210, 335-342.
https://link.springer.com/chapter/10.1007/978-3-642-23065-3_49
- Narmadha, D., & Balamurugan, A. a. (2016, 02 1). Survey of Clustering Algorithms for Categorization of Patient Records in Healthcare. *Indian Journal of Science and Technology*, 9(8).
<https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2016/Issue-8/Article29.pdf>



- **Potencial cliente interesado:**

EPS del sector privado que busca retener y fidelizar a sus afiliados.

- **Describir la fuente de donde los obtendría:**

Gracias al trabajo de apoyo en una EPS de la región de la cual se trabaja, los datos de los afiliados se pueden obtener de manera directa, se solicitará acceso a una muestra anónima de los afiliados para el desarrollo del proyecto.

- **Evaluar si existen barreras o no para obtenerlos:**

No existen barreras para obtener los datos de los afiliados.

- **Definir el rol que cada uno de los miembros del grupo ocupará en el desarrollo del proyecto**

- Christian González: Encargado de generación y limpieza de datos de los afiliados.

- Yovany Gasca: Encargado del procesamiento y segmentación de los datos de los afiliados.

- Jolman Salinas: Encargado de la investigación y documentación del proyecto.

- Elegir una de las preguntas utilizando la retroalimentación recibida en el avance anterior y expandir sobre su motivación a responderla.
- Describir si el problema pertenece a una tarea de reducción de dimensión, clustering, o una combinación de los dos, explicando el por qué.
- Proponer al menos un algoritmo/técnica de aprendizaje no supervisado que ustedes crean que es la adecuada para responder la pregunta.
- Presentar estadísticas descriptivas utilizando tablas y/o visualizaciones de los datos crudos que tengan a su disposición. Describir el plan que van a seguir para tener los datos listos. Por ejemplo, cómo van a limpiarlos, que van a hacer con los datos faltantes, etc.
- Actualmente se trabaja en la extracción de los datos, se tendrán disponibles el lunes en el transcurso del día.
- Por lo que las actividades planeadas son las siguientes:
 - a. Identificar el layout y tipos de datos.
 - b. Estadísticas de las dimensiones.
 - c. Distribución de los valores de cada dimensión.
 - d. Identificar valores atípicos que pudieran alterar los resultados esperados.

- 
- e. Si se identifican datos faltantes, se contemplan dos acciones dependiendo del porcentaje, sin es un volumen alto, poblar los datos faltantes con valores aleatorios que mantengan la distribución de los datos reales.
 - Con base en los roles definidos en el documento de la semana anterior, delinear las actividades que se llevaron y llevarán a cabo para la primera entrega calificada del proyecto. Ser preciso, y tomar este punto como un contrato entre los miembros del equipo.

- Christian González:

Encargado de generación y limpieza de datos de los afiliados.

- Iniciamos con la extracción y carga de los datos en el ambiente de trabajo..
- Se realizará un análisis exploratorio de datos que nos permita:
 - Identificar el layout y tipos de datos.
 - Estadísticas de las dimensiones.
 - Distribución de los valores de cada dimensión.
 - Identificar valores atípicos que pudieran alterar los resultados esperados.
- Si es necesario se tratarán las dimensiones que contengan datos faltantes

- Yovany Gasca:

Encargado del procesamiento y segmentación de los datos de los afiliados.

- Despues del análisis exploratorio inicial se procesarán los datos de acuerdo a los algoritmos definidos.
- Se trabajará en la reducción de dimensiones.
- Se elegirán los componentes principales
- Se ajustarán los datos por un modelo de clustering
- Y se evaluarán los resultados obtenidos con métricas de desempeño.

- Jolman Salinas:

Encargado de la investigación y documentación del proyecto.

A partir de la información generada tanto por el análisis exploratorio se buscará lo siguiente:

- Se documentará y presentarán los hallazgos obtenidos de cada una de las etapas.
 - Exportación de datos

- 
- Reducción de dimensionalidad
 - Y clustering
 - Se hará un análisis integral y se presentarán las conclusiones de la investigación