UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

Análisis de segmentación del cliente empresa de EPS SURA

Andrés Camilo Ospina Pérez

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Análisis de segmentación del cliente empresa de EPS SURA

Andrés Camilo Ospina Pérez

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

María Eugenia Castañeda López
Orientador Interno, Instituto de Matemáticas

Juan Luis Mejía Villa
Orientador externo, EPS SURA

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Agradecimientos

Agradezco a Dios, a mi madre Eliana Pérez, a mi Padre Nelson Ospina y a mi hermano Santiago Ospina, por todo su acompañamiento y apoyo incondicional en todos los momentos de mi vida, en especial para alcanzar este gran título.

Agradezco a todos los compañeros y docentes que hicieron parte de este proceso educativo, en especial a mi asesora de prácticas, profesora María Eugenia Castañeda López por su gran contribución de conocimiento a mi vida profesional.

Así mismo, expreso mis más sinceros agradecimientos a EPS SURA por permitir la realización de mis prácticas y por la gran acogida y aprendizaje que he tenido desde que comencé el proceso.

Resumen

Para las empresas, conocer sus clientes y entender sus necesidades hace parte de una gestión estratégica de mercado. Ese entendimiento puede hacerse a través de la segmentación de clientes donde se pueden conocer las características en común que tienen, permitiendo agruparlos de acuerdo con diferentes mediciones y teniendo en cuenta el comportamiento que han tenido a lo largo del tiempo. En este trabajo se presentan diferentes métodos estadísticos de reducción de dimensionalidad y clusterización tales como PCA, UMAP, K-MEANS y HDBSCAN, teniendo en cuenta variables numéricas y categóricas, con el objetivo de realizar la segmentación del cliente empresa de la Entidad Promotora de Salud en Colombia, EPS SURAMERICANA S.A.

Palabras clave: Empleadores, EPS, HDBSCAN, K-MEANS, PCA, Régimen de salud, Segmentación, Seguridad Social en Colombia, Tablero, UMAP.

Abstract

For companies, knowing their customers and understanding their needs is part of a strategic market management. This understanding can be done through customer segmentation where the characteristics in common that the agreement has can be known, allowing them to be grouped with different measurements and taking into account the behavior they have had over time. In this paper, different statistical methods of dimensionality reduction and clustering are presented, such as PCA, UMAP, K-MEANS and HDBSCAN, taking into account numerical and categorical variables, with the objective of performing the segmentation of the client company of the Health Promoting Entity in Colombia, EPS SURAMERICANA S.A.

Keywords: Dashboard, Employers, EPS, HDBSCAN, Health Regime, K-MEANS, PCA, Segmentation, Social Security in Colombia, UMAP.

Análisis de segmentación del cliente empresa de EPS SURA

Andrés Camilo Ospina Pérez [*]

12 de julio de 2021

[*]E-mail: acamilo.ospina@udea.edu.co, Instituto de Matemáticas, Universidad de Antioquia, Medellín, Colombia.

Resumen

Para las empresas, conocer sus clientes y entender sus necesidades hace parte de una gestión estratégica de mercado. Ese entendimiento puede hacerse a través de la segmentación de clientes donde se pueden conocer las características en común que tienen, permitiendo agruparlos de acuerdo con diferentes mediciones y teniendo en cuenta el comportamiento que han tenido a lo largo del tiempo. En este trabajo se presentan diferentes métodos estadísticos de reducción de dimensionalidad y clusterización tales como PCA, UMAP, K-MEANS y HDBSCAN, teniendo en cuenta variables numéricas y categóricas, con el objetivo de realizar la segmentación del cliente empresa de la Entidad Promotora de Salud en Colombia, EPS SURAMERICANA S.A.

Palabras clave: Empleadores, EPS, HDBSCAN, K-MEANS, PCA, Régimen de salud, Segmentación, Seguridad Social en Colombia, Tablero, UMAP.

Contenido

1. Introducción

En Colombia, el sistema de Seguridad Social, instituida por la ley 100 de 1993, integra un conjunto de entidades públicas y privadas, normas y procedimientos de protección laboral que se compone de los sistemas pensión, salud, riesgos laborales y servicios complementarios [3]. El sistema de Salud en Colombia está formado por dos sistemas coexistentes: El Régimen Contributivo y el Régimen Subsidiado. El Régimen Contributivo lo componen las personas que tienen una vinculación laboral, es decir, con capacidad de pago como los trabajadores de una empresa o los trabajadores independientes, los pensionados y sus familias. El Régimen Subsidiado lo componen la población de bajos recursos del país los cuales no tienen capacidad de pago para el sistema de salud y pueden acceder a los servicios a través de un subsidio que ofrece el Estado [13].

La regulación de la salud y el desarrollo de políticas para su ejecución está a cargo del Ministerio de Salud y Protección Social. A este ministerio pertenecen las EPS (Entidad Promotora de Salud) que son empresas públicas y privadas que representan a toda la población en su afiliación al Sistema de Seguridad Social en Salud. Estas empresas se encargan de implementar los objetivos fijados por el Ministerio de Salud en prestar los servicios sanitarios y médicos con el fin de garantizar el derecho de salud a todas las personas [15].

EPS SURAMERICANA S.A es una Entidad Promotora de Salud que ofrece los servicios de Plan de Beneficios en Salud (PBS), reglamentados por el Ministerio de Salud colombiano y además ofrece Planes Complementarios de Salud a sus afiliados. La afiliación de personas a la EPS se hace de manera principal por medio de los trabajadores de empresas que cotizan a la seguridad social del régimen contributivo, afiliando consigo a todo su grupo familiar.

Una de las formas de analizar la población afiliada al Régimen Contributivo de EPS SURA es a través de sus empleadores. Para EPS SURA es muy importante conocer las diferentes empresas a las cuales puede abordar para ofrecer el Plan de Beneficios en Salud. Este conocimiento del cliente empresa se encuentra enfocado en la necesidad de identificar cómo están constituidas las empresas teniendo en cuenta diferentes variables demográficas, socioeconómicas, de salud, comerciales, entre otras, y tener el detalle del comportamiento, tendencias, crecimiento de afiliados y organización.

Para comprender las características del cliente empresa de EPS SURA se propone como primera medida implementar una estadística descriptiva que den cuenta de la tendencia del número de afiliados pertenecientes a las empresas, análisis de relación entre variables de salud, demográficas y demás. Esta etapa inicial del entendimiento de las empresas implica la preparación de la información para visualizar, detectar, averiguar, calcular e identificar las características que componen el cliente empresa [5]. Esta etapa descriptiva tan importante se lleva a cabo por medio de tableros que permiten observar de forma dinámica la información del cliente empresa de manera general y detallada. Estos tableros son desarrollados con la herramienta Microstrategy. Los colaboradores de EPS SURA pueden acceder a estos tableros por medio de un enlace que genera el software, y los lleva a visualizar la información de manera general y conocer detalle de los empleadores conociendo sus características y la integración con el Cluster al que pertenecen, luego de hacer la segmentación.

Estadísticamente, otro de los aspectos claves para realizar un entendimiento del cliente empresa de EPS SURA es realizar una segmentación y reducción de dimensionalidad de las variables que describen a los empleadores. Esta segmentación permitirá a EPS SURA avanzar en la medición de características

y permitir a la fuerza de ventas entregar una asesoría y acompañamiento adecuado a las empresas y potencializar las afiliaciones de sus empleados al régimen contributivo de EPS SURA. La segmentación permite clasificar y dividir en grupos afines el cliente empresa con el fin de identificar claramente las características, necesidades y comportamientos del grupo de cliente llamado Clúster. Para llevar a cabo el proceso de Segmentación se disponen de varias técnicas Estadísticas y de Machine Learning tales como el Análisis de Componentes Principales (PCA) que se encarga de reducir un conjunto de variables correlacionadas de los datos en términos de nuevas variables incorrelacionadas logrando reducir dimensionalidad [14]; el Método de K-means, que es un algoritmo no supervisado de clasificación que genera k puntos aleatorios (centroides) y luego toma la distancia euclideana de cada muestra a estos centroides, y posteriormente asigna un clúster a esta muestra teniendo en cuenta la mínima distancia [14]; el método de aproximación y proyección de variedad uniforme (UMAP) usado para reducir dimensionalidad a partir de sus algoritmos basados en gráficos teniendo en cuenta una estructura topológica difusa [12], y el algoritmo de agrupación espacial jerárquica basada en densidad de aplicaciones con ruido (HDBSCAN)[1].

El presente informe cuenta con el siguiente esquema. En la sección II, encontrará el marco teórico donde se presentan los fundamentos teóricos y el funcionamiento de los algoritmos de reducción de dimensionalidad UMAP y PCA, y los algoritmos de clusterización HDBSCAN y K-MEANS. En la sección III, se presenta la metodología empleada en el presente estudio, en el que da cuenta de los detalles de la preparación de la información y la aplicación de los algoritmos. En la sección IV, se presentan los resultados de la segmentación del cliente empresa teniendo en cuenta la comparación de los diferentes algoritmos de reducción de dimensionalidad y clusterización, mencionados anteriormente. Finalmente, en la sección V, se presentan las conclusiones y recomendaciones del estudio, realizando una recapitulación del análisis, hallazgos y sugerencias para análisis posteriores.

2. Marco Teórico

2.1. UMAP

Uniform Manifold Approximation and Projection (UMAP) es un método de reducción de dimensionalidad no lineal propuesto por Leland Mcinnes et al. (2018) [12]. UMAP usa un algoritmo que permite reducir gráficamente una alta dimensión de datos en una más simple. Esta técnica de aprendizaje múltiple se construye a partir de un marco teórico basado en geometría riemanniana y topología algebraica.

El algoritmo se basa en tres supuestos:

1. Los datos se distribuyen uniformemente en una variedad de Riemann
2. La métrica de Riemann es localmente constante (o puede aproximarse como tal)
3. La variedad está conectada localmente

El procedimiento que UMAP emplea para reducir dimensionalidad y graficar comienza con una ilustración denominada complejo simplicial difuso, el cual es un gráfico ponderado con los pesos de los bordes que representan la probabilidad de que dos puntos estén conectados. Para lograr esta conexión con los puntos, UMAP extiende un radio hacia afuera desde cada punto, conectando puntos cuando esos radios se superponen. La elección de este radio es muy importante ya que una elección demasiado pequeña dará lugar a clústeres pequeños y aislados, mientras que una elección demasiado grande conectará todo. Para llevar a cabo esta decisión de la longitud del radio, UMAP elige un radio localmente, basado en la distancia a cada punto n-Vecino más cercano. Luego, UMAP hace que el gráfico sea difuso al disminuir la probabilidad de conexión a medida que aumenta el radio. Finalmente, al estipular que cada punto debe estar conectado al menos a su vecino más cercano, UMAP asegura que la estructura local se mantenga en equilibrio con la estructura global. Una vez que se construye el gráfico de alta dimensión, UMAP optimiza el diseño de un análogo de baja dimensión para que sea lo más similar posible [2].

El algoritmo UMAP es una de las técnicas más recientes y robustas que soportan grandes volúmenes de información y es usado generalmente para clasificar imágenes, textos y cualquier tipo de datos numéricos y categóricos, bajo transformaciones por medio de variables indicadoras. Este algoritmo es muy similar al de t-SNE (Incrustación de vecinos estocásticos distribuidos en t), el cual es usado para reducir datos de alta dimensión en el espacio bidimensional o tridimensional basado en la asignación de probabilidad de pares de puntos u observaciones [16]. La diferencia con t-SNE radica en que UMAP es un algoritmo basado en gráficos, es más veloz computacionalmente y mantiene la estructura global de los datos, a pesar del gran volumen de datos que puede manejar, lo que para t-SNE se le dificulta hacer la reducción para esta cantidad. Teniendo en cuenta estas características y la no restricción en dimensionalidad, UMAP puede ser usado como técnicas de aprendizaje automático.

Similar a t-SNE, el procedimiento que emplea UMAP es considerar primero los datos de entrada $X = \{x_1, x_2, \ldots, x_N\}$, $x_i \in \mathbb{R}^M$ y busca una representación dimensional baja óptima $\{y_1, \ldots, y_N | y_i \in \mathbb{R}^k\}$ tal que $k < M$. El primer paso es la construcción de grafos de k-vecinos ponderados, ya que UMAP usa el algoritmo de descendencia del vecino más cercano [12]. Se define una métrica $d : X \times X \to \mathbb{R}^+$ dada una entrada del hiperparámetro k y para cada x_i del conjunto de datos de entrada se calcula el k vecino más cercano bajo la métrica d donde se define un ρ_i tal que:

$$\rho_i = min\{d(x_i, x_j) | 1 \leq j \leq k, d(x_i, x_j) > 0\}$$

Se elige ρ_i para asegurarse de que al menos un punto de datos está conectado a x_i y tiene un peso de borde de 1.

Luego se establece σ_i como un parámetro de escala de longitud, tal que:

$$\sum_{j=1}^{k} \exp\left(\frac{-max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k$$

Se define un grafo dirigido ponderado $\bar{G} = (V, E, \omega)$ donde V es el conjunto de vértices del conjunto de datos X, E es el conjunto de bordes $E = \{(x_i, x_j) | 1 \leq h \leq k, 1 \leq i \leq N)\}$ y ω es una función de pesos de los bordes tal que:

$$\omega(x_i, x_j) = \exp\left(\frac{-max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right)$$

UMAP define un gráfico ponderado G a partir $\bar{G}$ a través de simetrización. Es decir, busca combinar gráficos simpliciales y difusos en una representación topológica unificada donde se tiene una matriz simétrica por medio de:

$$B = A + A^T - A \otimes A^T$$

Donde A es la matriz de adyacencia ponderada de $\bar{G}$ y $\otimes$ es el producto de Hadamard, donde se toman 2 matrices de igual dimensión y se produce otra con las mismas dimensiones. Por lo tanto, B_{ij} es la probabilidad de que exista al menos uno de los bordes dirigidos de x_i a x_j y de x_j a x_i.

Es decir, el algoritmo primero describe cómo están los datos en una estructura topológica para luego transformarlos en una estructura reducida. Primero visualiza cada dato como si fuera un espacio métrico (conjunto donde se define una distancia), luego transforma esos espacios para que se puedan conectar y se pueda visualizar esa estructura topológica inicial. Luego calcula distancias (probabilidades) entre conjuntos simpliciales y construye un vector de probabilidades.

Luego de haber definido el gráfico de k vecinos ponderados, se procede al diseño de un gráfico de menor dimensión en el que UMAP utiliza fuerzas atractivas y repulsivas cuyas coordenadas y_i y y_j están dadas por:

$$\frac{-2ab \left\|y_i - y_j\right\|_2^{2(b-1)}}{1 + \left\|y_i - y_j\right\|_2^2} w(x_i, x_j)(y_i - y_j)$$

donde a y b son hiperparámetros, y la fuerza repulsiva está dada por:

$$\frac{2b(1 - w(x_i, x_j))(y_i - y_j)}{\left(\varepsilon + \left\|y_i - y_j\right\|_2^2\right)\left(1 + a\left\|y_i - y_j\right\|_2^{2b}\right)}$$

Aquí ε se toma como un valor pequeño de modo que el denominador no sea cero. El objetivo es encontrar las coordenadas óptimas de baja dimensión $\{y_i\}_{i=1}^N$, $y_i \in \mathbb{R}^k$, que minimizan la entropía cruzada

de borde con los datos originales en cada punto [8].

Así entonces, luego de tener el primer espacio definido, se define un segundo espacio en $\mathbb{R}^2$, en el cual el algoritmo lleva los datos de forma aleatoria construyendo otra estructura topológica en $\mathbb{R}^2$ y comienza a optimizar la posición de los datos, teniendo en cuenta un vector de probabilidades minimizando la distancia entre los dos vectores de probabilidad (ley de entropía cruzada) y luego compara esas probabilidades obtenidas y las originales. Si en esa iteración los puntos quedaron distantes, el algoritmo vuelve e itera hasta que logre que esa distancia entre probabilidades sea mínima.

A nivel computacional, en Python [6], el algoritmo UMAP tiene varios hiperparámetros que generan un impacto importante al momento de reducir dimensionalidad y graficar, algunos de ellos son: n_neighbors, min_dist, n_components, metric [12].

- **n_neighbors:** Controla el número de vecinos más cercanos a un punto para construir el gráfico inicial de alta dimensión. Este hiperparámetro equilibra la estructura local frente a la global de los datos. Los valores bajos hacen que UMAP se centre más en la estructura local ya que se restringe el número de puntos vecinos, mientras que valores altos hacen que el algoritmo se centre principalmente en la estructura global de los datos para su graficación, haciendo que se pierdan algunos detalles.
- **min_dist:** Tiene en cuenta la distancia mínima entre puntos y controla qué tan estrictamente se le permite al algoritmo agrupar puntos. Valores bajos de este hiperparámetro indican incrustaciones más agrupadas de los puntos, lo que es ideal al momento de clusterizar. Y los valores altos hacen que los puntos se unan de manera más flexible, enfocándose en una estructura general amplia.
- **n_components:** Permite determinar la dimensionalidad del espacio a la que se desea reducir los datos.
- **metric:** Corresponde a la métrica de distancia d mencionada anteriormente y permite controlar cómo se calcula la distancia en el espacio de los datos de entrada X. Estas distancias pueden ser Euclidiana, Manhattan, Minkowski, Chebyshev, entre otras, dependiendo de la aplicación de los datos.

La Figura 1 representa el comportamiento del algoritmo al variar los parámetros de la distancia mínima y los n vecinos. Note que valores bajos de la distancia implica que los puntos estén más incrustados, y valores altos muestran un poco más la dispersión entre estos a nivel local debido a que se aumenta la distancia. De igual manera ocurre con los n vecinos, al aumentar su valor hace que se centre más en la estructura global para formar los grupos, haciendo que, para este caso, la forma de estrella vaya mejorando su estructura.

No hay una manera estándar de establecer el valor de los hiperparámetros, la elección de estos depende de los datos y el objetivo de la segmentación para el caso que se esté tratando. Para ello, el interés podrá estar en el detalle de la agrupación, para lo cual se eligen valores pequeños de los hiperparámetros, o si el interés se centra en encontrar grupos heterogéneos entre sí, permitiendo observar claramente los grupos de manera general, se puede enfocar en valores altos de los hiperparámetros.

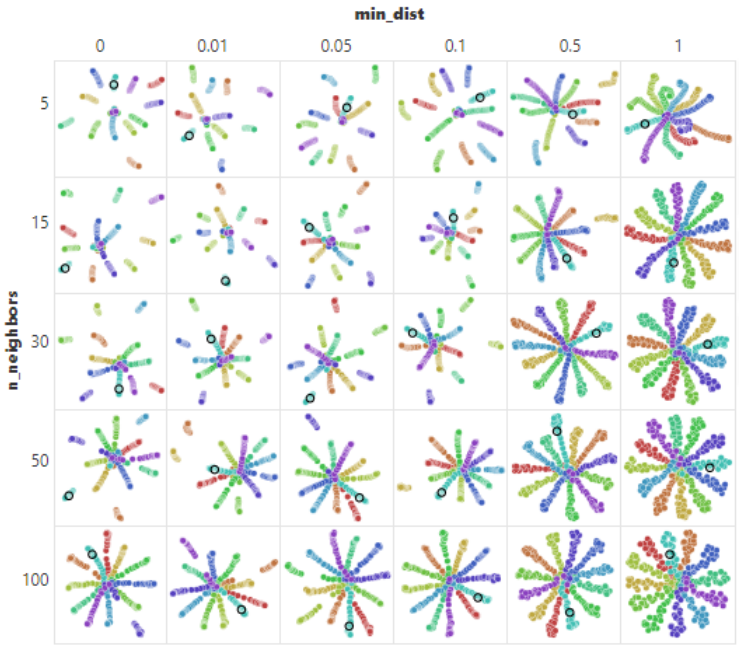


Figura 1: Proyección UMAP bajo diferentes valores de n_neighbors y min_dist. Tomado de [2]

2.2. HDBSCAN

HDBSCAN por sus siglas en inglés Hierarchical Density - Based Spatial Clustering of Applications with Noise es un algoritmo de clusterización no supervisado usado para encontrar grupos o regiones densas de un conjunto de datos, propuesto por Ricardo J. et al. [1]. Este algoritmo es una mejora del método del cual está basado como lo es el DBSCAN (Density - Based Spatial Clustering of Applications with Noise) el cual encuentra muestras centrales de alta densidad y expande grupos a partir de ellas [4]. El algoritmo se basa en la noción intuitiva de clusters y ruido, en el cual, para cada punto de un conglomerado, la vecindad de un radio determinado debe contener al menos un número mínimo de puntos. Esto se refiere a que requiere de parámetros como un tamaño mínimo de puntos y un umbral de distancia épsilon [10]. Este valor de épsilon determina la distancia máxima entre dos grupos que se consideran del mismo grupo, y los puntos mínimos que definen el valor en el que una vecindad de puntos se considera densa.

La diferencia entre el DBSCAN y el HDBSCAN es que este último, en vez de contar puntos en un radio épsilon, lo hace más efectivo y emplea DBSCAN para diferentes valores de épsilon, de modo que sólo requiere un sólo parámetro y es el tamaño mínimo de cluster y aprovecha el hecho que, al disminuir el valor de épsilon, los clusters se fragmentan en clusters más pequeños o permanecen igual [11].

2.3. PCA

El análisis de componentes principales (PCA) es una de las técnicas de reducción dimensional más utilizadas para el análisis exploratorio de datos de alta dimensión [9]. Dado un conjunto de datos con muchas variables, el objetivo será reducirlas a un menor número, perdiendo la menor cantidad de información posible [7].

Así entonces, dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Se buscan variables incorrelacionadas donde un pequeño número de ellas explique la mayor variabilidad posible [14]. Este algoritmo permite identificar posibles variables ocultas o no observadas, que están generando la variabilidad de los datos; permite transformar las variables originales, en general correlacionadas, en nuevas variables incorrelacionadas, facilitando la interpretación de los datos; permite observar la contribución de cada variable; permite identificar grupos o clusters de variables y de datos al momento de su interpretación.

La idea de obtener nuevas variables con varianza mayor surge de manera natural como una forma de perder menos información. Inicialmente solo se buscan nuevos ejes, con una importancia decreciente que permitirá escoger algunos de ellos. El PCA tiene como objetivo encontrar ejes ortogonales no correlacionados linealmente, que también se conocen como Componentes Principales (PC) en el espacio dimensional *m* para proyectar los puntos de datos en esos PC. La primera PC captura la mayor variación en los datos [14].

2.4. K-MEANS

Es uno de los algoritmos no supervisados de clasificación más populares en el aprendizaje automático [8], que tiene como objetivo agrupar o particionar un conjunto de datos $\{x_1,\ldots,x_N\}$, $x_i \in \mathbb{R}^M$ en k clusters, $\{C_1,\ldots,C_k\}$, $k \leq N$. El algoritmo comienza con la generación de k puntos aleatorios (centroides) y luego toma la distancia euclidiana de cada muestra a estos centroides, y posteriormente asigna un clúster a esta muestra teniendo en cuenta la mínima distancia [14]. Los centroides se actualizan minimizando la suma de cuadrados incluidos (WCSS) definidos como:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|_2^2$$

Donde μ_j es el promedio de datos de los puntos en el cluster j. μ_j es dada por:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

El método encuentra el centroide más óptimo dado un número fijo j. Para obtener la cantidad de clusters se puede hacer uso del método del codo, el cual consiste en trazar WCSS contra el número de clusters y se elije la cantidad j donde se observe un punto de inflexión en el gráfico.

3. Metodología

Para llevar a cabo el presente trabajo se tuvieron en cuenta cinco etapas, la primera es la extracción y preparación de los datos. La segunda incluye un análisis descriptivo de los empleadores por medio de un tablero (Dashboard). En la tercera etapa se desarrollan los algoritmos de reducción de dimensionalidad. Luego de ello, se desarrollan los algoritmos para la segmentación de empresas y, en la última etapa, se hace la validación y comparación de los modelos.

3.1. Extracción y preparación de la información

Luego del entendimiento de la necesidad con las áreas de la compañía, se dispone a hacer la extracción de los datos teniendo en cuenta el enfoque de segmentación del cliente empresa de EPS SURA. Para ello se definen con los analistas de las áreas, las variables de interés que podrían aportar al entendimiento de las empresas para realizar su segmentación. Inicialmente se plantearon 17 variables entre numéricas y categóricas que representan las empresas con cotizantes activos durante el mes de abril de 2021. Por políticas de confidencialidad de la información, no se menciona la cantidad de empleadores afiliados a EPS SURA y tampoco el detalle de las variables que se tomaron para su análisis. Estas 17 variables planteadas inicialmente corresponden a información demográfica, socioeconómica, de salud, utilizaciones, entre otras.

En este contexto, se procede a extraer la información de la bodega de datos que maneja la compañía, por medio del software Teradata SQL Assistant, a través de sentencias en lenguaje SQL. En esta extracción se consideran 2 bases de datos, una de ellas contiene el detalle de cada una de las características de los empleadores para ser visualizada y analizada en el tablero en la segunda etapa. La otra base de datos contiene la información agrupada de estas empresas para luego llevarla a Python y ejecutar los modelos de reducción de dimensionalidad y segmentación, que se presentan en la tercera y cuarta etapa, respectivamente.

Para las dos bases de datos se realiza un análisis de calidad y consistencia de la información, que consiste en la evaluación de valores faltantes, y veracidad de la información. Para ello se seleccionan algunas empresas aleatoriamente y se compara la información de la extracción con la información disponible en los aplicativos de la compañía.

Teniendo disponible la información, se procede a realizar la segunda etapa que corresponde al análisis descriptivo del cliente empresa de EPS SURA.

3.2. Análisis descriptivo

Con la información preparada, se procede a realizar el análisis descriptivo de los empleadores por medio del programa Microstrategy, este es un software enfocado en inteligencia de negocios que permite crear informes y análisis de datos almacenados en una base de datos relacional y otras fuentes. Para ello se carga la información al programa y se procede con la creación del tablero (Dashboard). Este tablero se crea con dos enfoques, el primero es mostrar la información general del cliente empresa a nivel país, en el cual se conocen los indicadores agrupados por cada una de las ciudades y demás variables que se tuvieron en cuenta y son de gran importancia para el análisis desarrollado en la compañía. La segunda parte permite mostrar el detalle de cada uno de los empleadores con todas su descripción y características

definidas en la primera etapa.

Este tablero les permite a los colaboradores de EPS SURA navegar de forma interactiva con la información, ya que, una vez ingresen a este encontrarán la información general del cliente empresa permitiendo hacer los filtros por las variables de interés. Una vez visualizada la información a nivel general, la persona puede dirigirse a otra pestaña del tablero donde se encuentra el detalle de la cada una de las empresas, así entonces podrá filtrar por el nombre o número de identificación de la empresa e inmediatamente le saldrá todo el detalle de la información asociada a esa empresa.

Este tablero es de fácil acceso, por lo que el colaborador de EPS SURA puede entrar en cualquier momento a visualizar la información desde el celular o el computador, y si requiere tomar alguna información del tablero, puede descargar los gráficos o tablas resumen a través de un botón que dispone la herramienta, permitiendo que la información siempre esté disponible para los colabores de EPS SURA.

Adicional al tablero, se prepara un informe con un análisis general de la información del cliente empresa. Este informe se complementa al final luego de la última etapa de la validación de los modelos, cuando se tiene el cluster al que pertenece cada empresa. Esta información de la segmentación se retorna al tablero para que esté disponible a la hora de realizar algún análisis, permitiendo que, cuando se consulte de manera general o detallada la información del cliente empresa, también se muestre el segmento al que pertenece. Así entonces, la compañía podrá tener a disposición las empresas que correspondan a un evento de interés y que esté relacionada con la descripción del segmento. Por ejemplo, si se requiere hacer énfasis en las empresas que tengan un alto potencial de venta, se pueden seleccionar teniendo en cuenta la definición de los clusters.

Teniendo claridad en la información y características que describen a cada cliente empresa, se procede con la tercera etapa de reducción de dimensionalidad.

3.3. Reducción de dimensionalidad

Para realizar la segmentación del cliente empresa se requieren de dos fases fundamentales: la reducción de dimensionalidad y la clusterización. En esta tercera etapa de la metodología, se lleva a cabo la reducción de dimensionalidad, la cual consiste en representar la información de las 17 variables definidas en la primera etapa, en 2 dimensiones, para luego dar paso a la segmentación.

En esta etapa de reducción de dimensionalidad se consideran 2 modelos: UMAP y PCA. Estos modelos son ejecutados en Python, y para ello, primero se carga la base de datos que se definió en la primera etapa, que contiene la información agrupada de los empleadores. Primero se ejecuta el algoritmo UMAP, para ello se extraen de la base de datos las variables numéricas y categóricas.

Las variables categóricas que se consideraron fueron binarias, por lo que se hace una transformación numérica de ceros y unos, tal como lo recomienda el autor del método [12]. Con las variables organizadas se procede a hacer un análisis de correlación para observar si existe algún grado de asociación lineal entre las variables. Luego se procede con la estandarización de las variables, con el fin de manejar una sola escala de unidades y de esta manera controlar los supuesto del modelo para que la reducción de dimensionalidad no se vea influenciada por valores grandes o pequeños de las variables al manejar diferentes escalas. Con estas variables estandarizadas se procede a ejecutar el algoritmo UMAP para

diferentes valores de sus hiperparámetros de la cantidad de vecinos y la distancia mínima. Con ello se construye una matriz gráfica con variación en los parámetros para observar el comportamiento de los datos para cada uno de los valores definidos. El objetivo se centra en encontrar estructuras globales bien delimitadas y enfatizar en pequeñas regiones, debido a que se trata de un aprendizaje no supervisado y se buscar encontrar una cantidad moderada de clusters. Luego de esta matriz de gráficos, se escogen los parámetros apropiados para continuar con la etapa de segmentación.

En esta etapa de reducción de dimensionalidad también se emplea el método de componentes principales, para ello observa la cantidad de variación de los datos en cada una de las componentes. Para este trabajo se consideraron dos componentes principales con el fin de realizar una comparación de los métodos y evaluar su eficiencia. Con las dos primeras componentes definidas, se procede a realizar un Biplot que permite ver el grado de asociación de cada una de las variables con las componentes principales.

3.4. Segmentación

Luego de haber realizado la reducción de dimensionalidad, se procede con la implementación de los modelos para definir los cluster de las empresas. El primer método que se empleó fue el HDBSCAN teniendo en cuenta el modelo UMAP. Similar al procedimiento empleado con la definición de parámetros para UMAP, para HDBSCAN se realizan diferentes gráficos con diferentes valores de parámetros. Para la elección de los parámetros de este modelo, se tuvo en cuenta la cantidad de ruido, es decir, la cantidad de puntos que el algoritmo no pudo asociarlos a un cluster y que clasifican como error o ruido. Así mismo se tuvo en cuenta el gráfico de árbol condensado, este gráfico es similar a un árbol jerárquico o dendrograma en el que se observan la cantidad de cluster que pueden existir en el modelo y dan cuenta de cuántos y cuáles son recomendados elegir teniendo en cuenta la cantidad de puntos asociados al cluster y si perduran en el tiempo, es decir, que no se ramifiquen en otros cluster, priorizando clusters tempranos, es decir las primeras ramas del árbol.

Otro método de clusterizacioón implementado fue K-MEANS bajo el método de reducción de dimensionalidad UMAP seleccionado en la etapa tres. Para seleccionar la cantidad de cluster se hace uso del método del codo para evaluar la cantidad de cluster adecuada para el modelo. Bajo esta misma metodología del gráfico de codo se implementó el modelo de K-MEANS con un modelo de reducción de dimensionalidad PCA.

A continuación se presenta el resumen de los tres modelos implementados de UMAP, PCA, HDBSCAN y K-MEANS:

Modelo	**Reducción de dimensionalidad**	**Segmentación**
1	UMAP	HDBSCAN
2	UMAP	K-MEANS
3	PCA	K-MEANS

Tabla 1: Modelos de segmentación empleados

3.5. Validación de los modelos

Luego de implementar los 3 modelos definidos en la Tabla 1, se procede con la evaluación de los modelos. Para ello se tienen en cuenta los siguientes criterios:

1. **Forma gráfica:** Se busca un modelo que separe adecuadamente los puntos, formando grupos definidos y distantes, teniendo en cuenta la cantidad de cluster que se puedan formar.
2. **Cantidad de Cluster:** Se busca un modelo que contenga una cantidad moderada de cluster. Aunque se debe tener en cuenta el criterio gráfico, no es apropiado seleccionar ni muy pocos cluster ni muchos cluster, debido a que se pierde el objetivo de la segmentación de empleadores para observar características comunes entre ellos y así realizar un entendimiento de las necesidades.
3. **Ruido:** Se busca un modelo que haya clasificado la mayoría de los puntos a un cluster y no a un ruido. Así mismo se busca que el ruido no se convierta en un cluster más, es decir, que existan más puntos clasificados como ruido y no como cluster.

Así mismo se deben tener en cuenta los criterios propios de validación de cada modelo, definidos en las etapas 4 y 5. Con esta validación de los modelos se procede a escoger uno de ellos y a devolver a la base de datos el cluster para el cual pertenece cada empleador. Luego, esta información será llevada al informe y al tablero para realizar el análisis correspondiente y observar su relación con las demás variables.

4. Resultados

4.1. Análisis Descriptivo

Para realizar un entendimiento de cómo estaban conformadas las empresas, cuáles eran sus características y cómo era su comportamiento frente a las diferentes variables de la compañía, se realiza un tablero en el software Microstrategy. Por confidencialidad en la información, los valores que se muestran en la Figura 2 fueron simulados y las variables fueron homologadas en códigos. Este tablero representa la cantidad de empresas expresadas en cada una de las variables demográficas, socioeconómicas, de salud, comerciales, entre otras. Allí también se muestra el Top 10 de empresas con mayor valor en la variable V8, la cual es de importante medición en la compañía. En la parte superior del tablero se encuentran los filtros de algunas variables con el fin de observar y entender la información de forma dinámica.

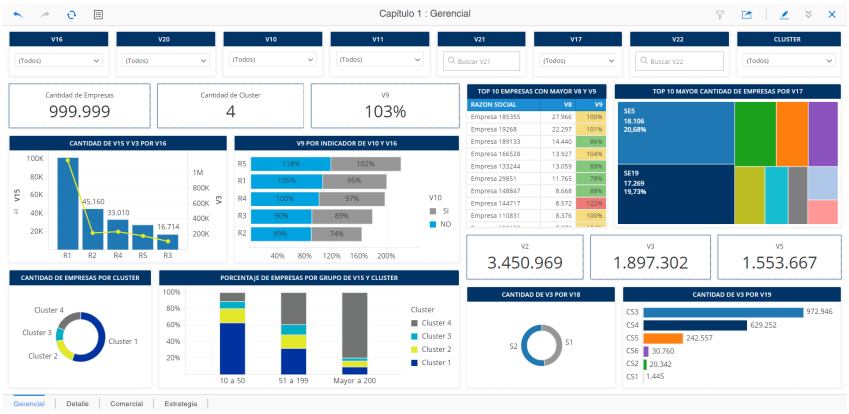


Figura 2: Tablero

4.2. Modelo 1: UMAP y HDBSCAN

Inicialmente, para el modelo de reducción de dimensionalidad UMAP, se realiza en Python un gráfico del método con la variación de los hiperparámetros de n-vecinos y distancia mínima.

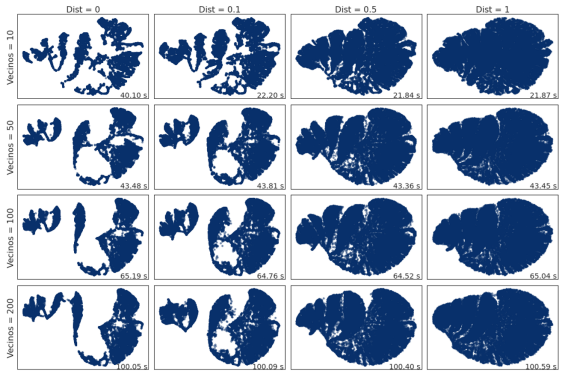


Figura 3: Variación hiperparámetros UMAP

Teniendo en cuenta la Figura 3, se selecciona el modelo UMAP con 100 vecinos y una distancia de 0. Para estos hiperparámetros, el modelo presenta, visualmente, una buena separación entre grupos de datos, los cuales serían los cluster más adelante. Para este análisis de las empresas, sería ideal no escoger una cantidad muy grande o pequeña de cluster. Con una cantidad pequeña de vecinos o distancia, se forman muchos grupos pequeños, mientras que, para valores grandes, no se observa una buena separación de grupos ya que todos se forman en una sola nube de puntos.

Para el modelo de segmentación HDBSCAN, se realiza en Python un gráfico del método con la variación de los hiperparámetros del mínimo de muestras y un tamaño mínimo de cluster.

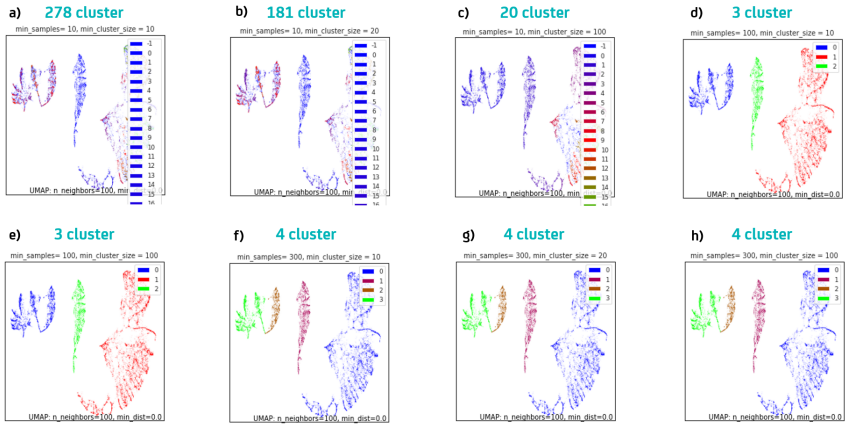


Figura 4: Variación hiperparámetros HDBSCAN

Teniendo en cuenta la Figura 4, se selecciona el modelo HDBSCAN con un mínimo de 300 muestras y un tamaño mínimo de cluster de 10, el cual corresponde al gráfico f). Bajo estos parámetros, el modelo propone 4 cluster y ausencia de ruido. Se observa que, con un tamaño mínimo de muestras y de cluster, el modelo arroja muchos clusters, incluso con una alta participación de ruido, lo cual no sería ideal considerar. Estos modelos con gran cantidad de grupos se ilustran en las gráficas a), b) y c). Luego para un valor de 100 muestras y un tamaño de cluster 10, sólo se consideran 3 cluster, como ocurren en los gráficos d) y e). Visualmente, la separación de 3 a 4 clusters se asocia a diferencias importantes que puede haber entre las empresas y que, teniendo en cuenta las unidades de los hiperparámetros, cambia rápidamente.

La Figura 5 corresponde al modelo seleccionado con un mínimo de 300 muestras y un tamaño mínimo de cluster de 10, obteniendo como resultado 4 grupos o segmentos, los cuales indican, visualmente, una muy buena separación de los datos y una cantidad de cluster apropiadas para el entendimiento del cliente empresa.

Para validar los criterios de selección anteriores se hace énfasis en el árbol condensado. Teniendo en cuenta la Figura 6, la selección del modelo fue pertinente ya que, a la larga, estos clusters no se subdividen en otros cluster en una gran medida. La subdivisión de los cluster 2, 3 y 4 ocurre para muy pocos puntos, y debido a su magnitud y color no representan una influencia importante. Aunque la subdivisión del cluster 1 ocurre en varios clusters, estos indican que, a una corta distancia se forman.

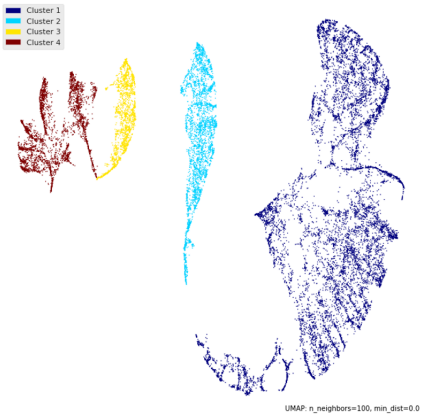


Figura 5: Modelo seleccionado

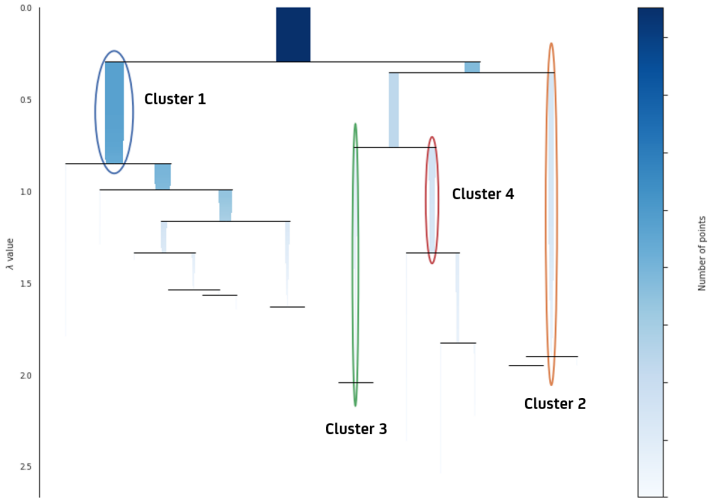


Figura 6: Árbol Condensado

4.3. Modelo 2: UMAP y K-Means

Basado en el modelo UMAP seleccionado anteriormente, con 100 vecinos y una distancia 0, se propone un modelo de segmentación de K-Means con 3, 4 y 10 clusters para realizar la comparación con el Modelo 1.

De acuerdo con la Figura 7, los clusters propuestos por k-means no presentan un buen ajuste con los puntos agrupados por UMAP, incluso la cantidad de puntos por cluster no es la más indicada. Por ejemplo, para el modelo de 4 clusters, el método asocia la mayoría de los puntos al cluster azul, mientras que para el Modelo 1, estos puntos equivalen a 3 cluster diferentes.

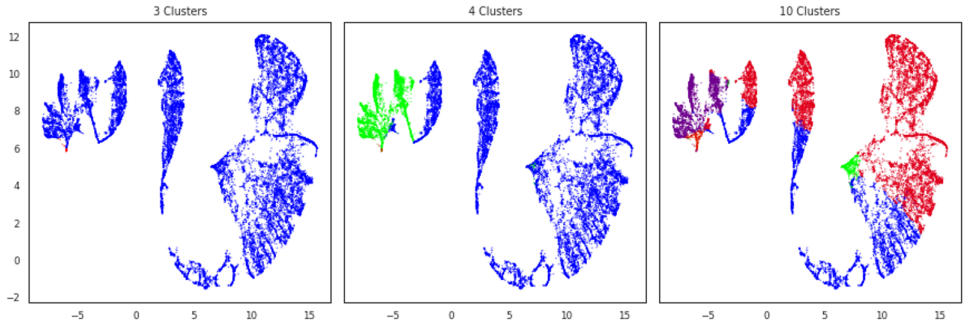
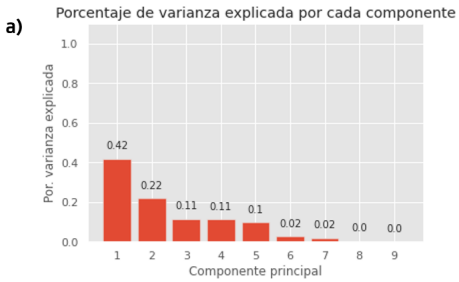
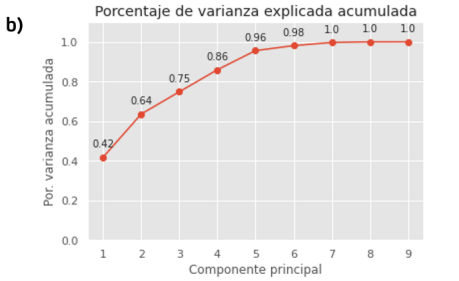


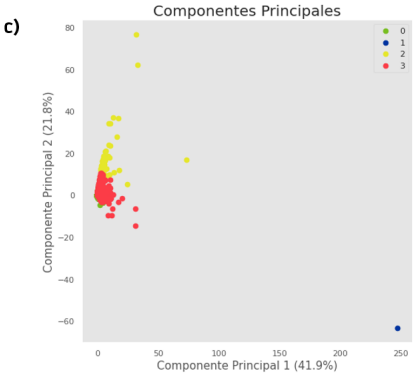
Figura 7: Variación cantidad de cluster

4.4. Modelo 3: PCA y K-Means

Para implementar el modelo de reducción de dimensionalidad, inicialmente se hace uso del gráfico de codo para analizar la varianza explicada por cada componente. En la Figura 8, el gráfico a) indica la varianza explicada por cada una de las componentes, en este caso, la componente 1 aporta el 42 % de la variabilidad de los datos; la componente 2, el 22 %; la componente 3 y 4, el 11 %; la componente 5 cerca del 10 %; y las demás componentes aportan muy poca variabilidad.







d)

Cluster	% Empresas Modelo 1	% Empresas Modelo 3
1	55.59%	81.84%
2	16.76%	0.00%
3	8.86%	0.23%
4	18.79%	17.92%

Figura 8: Modelo PCA y K-Means

El gráfico b) de esta Figura 8 indica la varianza acumulada en cada una de las componentes, en este caso, el método indica en seleccionar 5 componentes, en los cuales se encuentra la flexión de la línea. Sin embargo, se seleccionan 2 componentes principales, los cuales representan una varianza acumulada del 64 %, y parar la segmentación, se seleccionan 4 cluster con el fin de ilustrar y comparar con los otros métodos. El Modelo 3, expresado en el gráfico c) de esta Figura, asocia la mayoría de los puntos a 2 cluster y no se observa una separación definida de los puntos, como ocurre con el Modelo 1. En el

elemento d) de la Figura, corresponde a una tabla donde se presenta el porcentaje de empresas por cada uno de los cluster en los Modelos 1 y 3.

Luego de la comparación de los 3 modelos, se selecciona el Modelo 1, el cual corresponde a un modelo de reducción de dimensionalidad UMAP y segmentación HDBSCAN. Para la selección de este modelo se tuvo en cuenta que la representación gráfica tuviera una separación adecuada de los puntos; que la cantidad de clusters fuera moderada; que la cantidad de ruido fuera mínima y que la cantidad de empresas en cada cluster fuera proporcional. Incluso para el Modelo 1, se validó la selección de los hiperparámetros con el Árbol Condensado.

Teniendo en cuenta este modelo seleccionado, se hace un análisis de los 4 clusters con algunas variables que se consideraron para el modelo. En la Figura 9 se observa el porcentaje de empresas en la variable 1 asociadas a cada uno de los cluster y su representación visual en el modelo UMAP. Para esta variable, el Cluster 1 representa las empresas con valores de 1, mientras que, para el Cluster 4, se encuentran las empresas con mayor valor en V1, de 4 a 12.

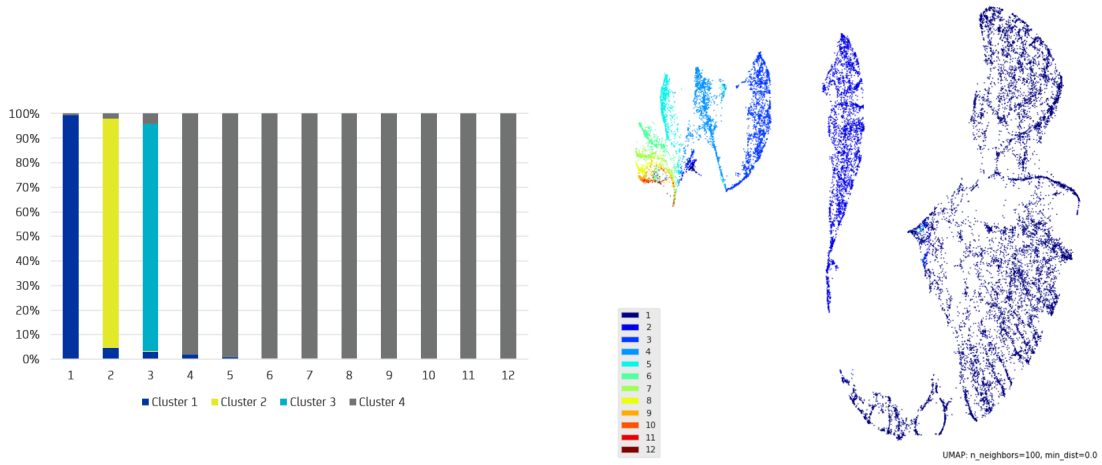


Figura 9: Representación de V1 por cluster

En la Figura 10 se encuentran los indicadores de algunas variables por los 4 cluster. Este análisis se ve representado en la Tabla 2, que indica las características más relevantes de las variables en cada uno de los cluster. La información detallada con las descripciones de las variables se presenta a la compañía, y son homologadas por códigos para presentar en este informe.

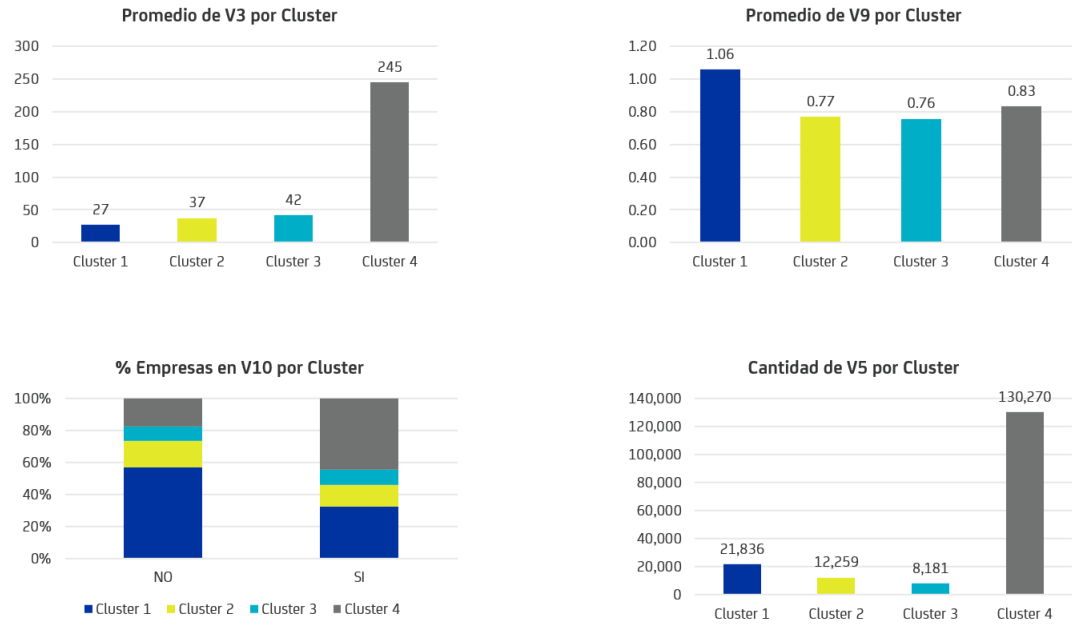


Figura 10: Análisis de Variables por Cluster

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
V1	1	2	3	4 a 12
V3	10 a 50	10 a 50	51 a 199	Mayor a 200
V4	0.5	0.5	0.6	0.6
V8	0 a 50	51 a 199	51 a 199	Mayor a 200
V10	No	No	Sí	Sí
V11	Sí	No	No	No
V12	T1	T1	T3	T3
V14	S26	S24	S24	S6

Tabla 2: Aspectos más relevantes de las Variables por Cluster

Luego del entendimiento de la segmentación de las empresas, estos datos son llevados al tablero para conocer detalladamente a qué cluster corresponde cada empresa y poder enfocar estrategias particulares a cada una ellas por medio de su agrupación, tal como se muestra en la Figura 11. El tablero funciona de forma dinámica, si se desea conocer el detalle de una empresa se puede hacer la búsqueda en los filtros, tal como se muestra en el elemento a) de la Figura 11. Inmediatamente se muestran las características de esta empresa filtrada como se muestra en el elemento b), incluso el cluster al que pertenece, como se muestra en el elemento d). Si se desean conocer la empresas que pertenecen a un cluster específico, se puede filtrar el cluster, como se muestra en el elemento c) de la Figura, e inmediatamente se muestran todas las empresas pertenecientes, con sus características. Esta información puede ser descargada en Excel o pdf y a partir de esta información se puede realizar un análisis detallado de las empresas con el fin de conocerlas e implementar estrategias comerciales, enfocadas a un potencial de venta.

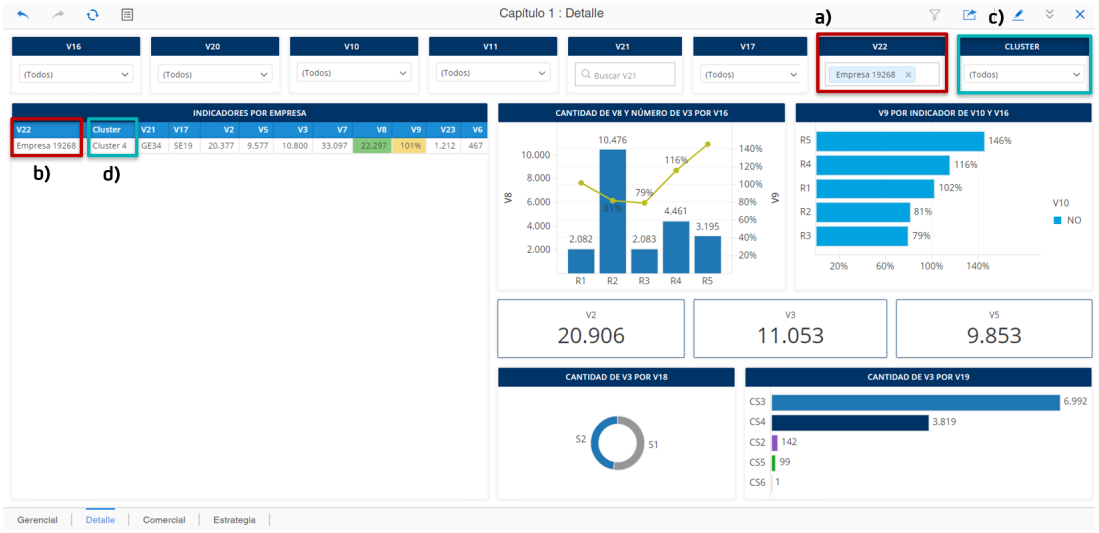


Figura 11: Detalle de empresas con cluster

5. Conclusiones y Recomendaciones

Para analizar cómo estaba constituido el cliente empresa de EPS SURA, se implementaron 3 modelos de clusterización; el Modelo 1, corresponde al método de reducción de dimensionalidad UMAP con el método de segmentación HDBSCAN; el Modelo 2, corresponde al método UMAP y K-Means y, el Modelo 3, corresponde al método PCA con K-Means. Teniendo en cuenta varios criterios de validación como la cantidad cluster, presencia de ruido, agrupación de los datos, y demás, se selecciona el Modelo 1.

Con la implementación del Modelo 1 (UMAP y HDBSCAN) se generaron 4 clusters, los cuales permitieron agrupar cada una de las empresas con ausencia de ruido y una alta probabilidad de clasificación. Este modelo, permitió conocer las características en común que tenían las empresas, y a partir de su agrupación se logró identificar tendencias, comportamientos, necesidades e información de valor para la compañía.

El método de reducción de dimensionalidad UMAP arroja muy buenos resultados si se implementa con el método de segmentación basado en densidad HDBSCAN, en comparación con K-Means.

Para el método UMAP, considerar variables categóricas y homologarlas a numéricas puede afectar el modelo debido a su escala, como valores muy altos o bajos.

Teniendo como base este primer acercamiento para el análisis de segmentación del cliente empresa, para trabajos futuros se recomienda un análisis en la definición de las variables y una inclusión de otros indicadores de las empresas que maneja la compañía.

Para trabajos posteriores se recomienda realizar un análisis entre la cantidad de observaciones y variables, ya que grandes volúmenes de información podrían ser computacionalmente costoso implementar algunos modelos.

Referencias

[1] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[2] Andy Coenen and Adam Pearce. Understanding umap. *Google PAIR*, 2019.

[3] Departamento Nacional de Planeación. `www.dnp.gov.co`, 2020.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[5] Santiago Fernández Fernández, José María Cordero Sánchez, Alejandro Córdoba, and Alejandro Córdoba Largo. *Estadística descriptiva*. Esic Editorial, 2002.

[6] Guido van Rossum. *Python*. Python Software Foundation, 3.7.10.

[7] Manuel Gurrea. Análisis de componentes principales. *Proyecto e-Math Financiado por la Secretaría de Estado de Educación y Universidades (MECD)*, 2000.

[8] Yuta Hozumi, Rui Wang, Changchuan Yin, and Guo-Wei Wei. Umap-assisted k-means clustering of large-scale sars-cov-2 mutation datasets. *Computers in biology and medicine*, 131:104264, 2021.

[9] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[10] Alboukadel Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda, 2017.

[11] CHA Logan and Sotiria Fotopoulou. Unsupervised star, galaxy, qso classification-application of hdbscan. *Astronomy & Astrophysics*, 633:A154, 2020.

[12] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[13] Ministerio de Salud. `https://www.minsalud.gov.co`, 2021.

[14] Daniel Peña. *Análisis de datos multivariantes*, volume 24. McGraw-hill Madrid, 2002.

[15] Secretaría Distrital de Planeación. `www.sdp.gov.co`, 2020.

[16] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.