MODELO DE ANALITICA DE DATOS PARA APOYAR LA COBERTURA DEL ASEGURAMIENTO EN SALUD EN EL DEPARTAMENTO DE CUNDINAMARCA

DERIAN JESÚS DORADO DAZA

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

DAVID ARANGO LONDOÑO

VALENTINA CORCHUELO GUZMÁN

JULIÁN GIL GONZÁLEZ

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Cali (Valle del Cauca), 14 de Agosto de 2023.

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 14 de Agosto de 2023

Autor: Derian Jesús Dorado Daza

Título del Trabajo de Grado: "MODELO DE ANALITICA DE DATOS PARA APOYAR LA COBERTURA DEL ASEGURAMIENTO EN SALUD EN EL DEPARTAMENTO DE CUNDINAMARCA"

Director: DAVID ARANGO LONDOÑO

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

_____

Firma del Director del Trabajo de Grado

Santiago de Cali, **14 de agosto de 2023**

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado “Modelo de analítica de datos para apoyar la cobertura del Aseguramiento en salud en el Departamento de Cundinamarca”, el cual fue realizado por el estudiante **Derian Jesús Dorado Daza** con código 8972749 perteneciente a la Maestría en Ciencia de Datos, bajo la dirección de **David Arango Londoño**.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Derian Jesús Dorado Daza
C.C. 76331763 de Popayán

David Arango Londoño
C.C. 1130586950 de Cali

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

FICHA RESUMEN

TÍTULO: Modelo de analítica de datos para apoyar la cobertura del Aseguramiento en Salud en el departamento de Cundinamarca.

1. ÁREA DE TRABAJO: **Sector Salud**
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): **Aplicado**
3. ESTUDIANTE(S): **Derian Jesús Dorado Daza**
4. CORREO ELECTRÓNICO: **djdorado@javerianacali.edu.co**
5. DIRECCIÓN Y TELÉFONO: **Calle 4 # 1-98 Cajicá, Cundinamarca. Telf: 3003464875**
6. DIRECTOR: **David Arango Londoño**
7. VINCULACIÓN DEL DIRECTOR: **Profesor de planta**
8. CORREO ELECTRÓNICO DEL DIRECTOR:**david.arango@javerianacali.edu.co**
9. GRUPO O EMPRESA QUE LO AVALA (Si aplica): **Gobernación de Cundinamarca**
10. PALABRAS CLAVE (al menos 5): **Aseguramiento en salud, analítica de datos, bases de datos, Prestadores de servicios de salud, Aprendizaje de Máquina No Supervisado, Clustering.**
11. FECHA DE INICIO: **4 de Julio de 2022**
12. DURACIÓN ESTIMADA (En meses): **12 meses**
13. RESUMEN:

Este trabajo aborda una problemática que con frecuencia se presenta en el procedimiento de Seguimiento a la Base de Datos del Aseguramiento en salud en el Departamento de Cundinamarca, que trata con la identificación de relaciones que no son evidentes por métodos tradicionales de análisis, entre distintas variables que caracterizan a los afiliados a los regímenes Subsidiado y Contributivo con el propósito de mejorar la toma de decisiones frente a la cobertura del aseguramiento y acceso a los servicios de salud. Plantea el diseño e implementación de un modelo de analítica de datos para mejorar la comprensión de estas relaciones recurriendo a conceptos y técnicas propias de la Ciencia de Datos.

MODELO DE ANALITICA DE DATOS PARA APOYAR LA COBERTURA DEL ASEGURAMIENTO EN SALUD EN EL DEPARTAMENTO DE CUNDINAMARCA

Derian Jesús Dorado Daza

Proyecto Aplicado para optar al título de
Magíster en Ciencia de Datos

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO DE 2023

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

El presente documento constituye el proyecto aplicado desarrollado en el programa de Maestría en Ciencia de Datos de la Universidad Javeriana - Cali. Este trabajo abordó una problemática que con frecuencia se presenta en el procedimiento de Seguimiento a la Base de Datos del Aseguramiento en salud en el Departamento de Cundinamarca y que tiene que ver con la identificación de relaciones que no son evidentes por métodos tradicionales de análisis, entre distintas variables que caracterizan a los afiliados al Sistema General de Seguridad Social en Salud (Regímenes Subsidiado y Contributivo) con el propósito de mejorar la toma de decisiones frente a la cobertura del aseguramiento y acceso a los servicios de salud.

El establecimiento de esta situación y propósito, justifican el desarrollo de este proyecto aplicado siendo necesario recurrir a los conceptos, técnicas y métodos que provee la Ciencia de Datos. Así pues, este trabajo planteó el diseño e implementación de un modelo de analítica que permita una comprensión más profunda del área de negocio y los datos, así como el desarrollo de los objetivos que apuntan a dar solución a la situación problema que se ha descrito. En este sentido, los resultados principales de este trabajo son la documentación de consulta en forma de monografía, el modelo de analítica de datos, y una herramienta de visualización de los resultados.

Se plantea el empleo de la metodología CRISP-DM para abordar el desarrollo general del proyecto. Puesto que es un referente de buenas prácticas que permite conducir ordenada y sistemáticamente las diferentes actividades que a su vez son coherentes con los objetivos general y específicos planteados en este trabajo.

1.- DEFINICIÓN DEL PROBLEMA

1.1.- Planteamiento del Problema

El Aseguramiento en Salud es un componente complejo del Sistema General de Seguridad Social en Salud que asume el mandato constitucional de garantizar el acceso universal a los servicios de salud establecido como derecho fundamental mediante los mecanismos de afiliación y permanencia en el sistema.

En este accionar, el Aseguramiento en Salud aborda distintos retos administrativos, normativos, operativos y técnicos de manejo de información. Es en este último aspecto que se enmarca la presente situación problema y propuesta de proyecto aplicado; cuyo contexto está determinado por las actividades y funciones de gestión de la información de los afiliados al sistema de salud del departamento de Cundinamarca.

Una situación problema que se presenta con frecuencia consiste en la necesidad de ampliar la comprensión y articulación de la información de los afiliados frente a cuestiones relevantes de caracterización y acceso a los servicios de salud; de tal modo que permita identificar y comprender relaciones existentes entre determinadas variables sociodemográficas, geográficas y de afiliación frente a los distintos actores que administran y prestan servicios de salud en el departamento. Dado el volumen, nivel de complejidad y factores de dinámica social de la información de los afiliados, estas relaciones ya no son evidentes u observables desde un enfoque tradicional de analítica, siendo necesario plantear un enfoque desde la Ciencia de Datos y sus técnicas de análisis para abordar esta situación.

En este punto es importante anotar que la prestación de los servicios de salud a los habitantes del territorio corresponde a los prestadores de servicios de salud que operan en el departamento y están constituidos principalmente como IPS (Instituciones Prestadoras de Servicios de Salud) de naturaleza publica, privada o mixta. Ahora bien, en este contexto las IPS públicas merecen atención especial pues constituyen la red pública de prestadores de servicios de salud distribuidos y configurados en las Regiones de Salud del departamento.

1.2.- Formulación del Problema

En este contexto, el presente proyecto plantea la siguiente cuestión sobre la situación descrita: ¿Cómo identificar y mejorar la comprensión de las relaciones entre determinadas variables de los afiliados al Sistema General de Seguridad Social en Salud de tal modo que apoye la toma de decisiones frente a la cobertura del Aseguramiento y acceso a los servicios de salud en el departamento de Cundinamarca?

2.- OBJETIVOS DEL PROYECTO

"El hombre, al contrario de lo que aparenta, no se inventa objetivos. Se los impone la época en que nació, puede estar a su servicio, o bien rebelarse contra ellos, pero tanto el objeto de la entrega como el de la rebelión vienen dados desde fuera. Para experimentar una plena libertad en la búsqueda de metas, tendría que vivir a solas y por ahí no hay salida, porque un hombre que no ha sido criado entre hombres no puede convertirse en ser humano. El que yo imagino… es un ser singular, privado de toda pluralidad"

STANISLAW LEM [Solaris]

2.1.- Objetivo General

Diseñar un modelo de analítica de datos que permita identificar relaciones existentes entre determinadas variables de los afiliados a los regímenes Subsidiado y Contributivo, y apoye la toma decisiones frente a la cobertura del Aseguramiento y acceso a los servicios de salud en el departamento de Cundinamarca.

2.2.- Objetivos Específicos

1.- Establecer y ejecutar actividades de planificación tendientes a la comprensión del área de negocio, identificación de variables para el modelo, exploración y consolidación de los datos.

2.- Construir un modelo de analítica a partir de los datos consolidados que permita identificar relaciones entre determinadas variables de los afiliados a los regímenes Subsidiado y Contributivo.

3.- Evaluar el modelo generado y verificar la validez de los resultados obtenidos.

4.- Desplegar el modelo en herramientas apropiadas para la publicación y socialización.

3.- MARCO TEORICO Y ANTECENTES

"- Propongo desconectar algunos de tus circuitos, en particular aquellos que implican tus funciones más elevadas ¿Te inquieta esto? [...]
- ¿Soñaré?
- Por supuesto que lo harás. Todas las criaturas inteligentes sueñan, aunque ninguna sabe el porqué. "

ARTHUR CLARKE [2010: Odisea dos]

3.1.- MARCO TEÓRICO

3.1.1.- Sistema General de Seguridad Social en Salud y el Aseguramiento en Salud.

La salud es un derecho fundamental establecido en la Constitución Política de Colombia, y corresponde de manera indelegable al Estado reglamentar, organizar y garantizar el acceso a los servicios de salud para los habitantes del territorio. Lo anterior en concordancia con los principios de eficiencia, universalidad y solidaridad. También es importante señalar que la constitución establece que es deber del Estado fijar "las políticas para la prestación de servicios de salud por entidades privadas, y ejercer su vigilancia y control. Así mismo, establecer las competencias de la Nación, las entidades territoriales y los particulares, y determinar los aportes a su cargo en los términos y condiciones señalados en la ley" [1].

En cumplimiento de este mandato constitucional se establece en Colombia el Sistema General de Seguridad Social en Salud (SGSSS) y uno de los actores principales está representado por las IPS (Instituciones Prestadoras de Servicios de Salud) del que hacen parte hospitales y clínicas (públicos y privados) que prestan los servicios de salud a los habitantes de los distintos territorios del país que están en condición de afiliados o incluso si no lo están, siendo obligación su afiliación al sistema por parte de las entidades territoriales municipales y las empresas promotoras de salud (EPS).

3.1.2.- Aseguramiento en Salud en el Departamento de Cundinamarca.

El contexto geográfico del presente trabajo es el Departamento de Cundinamarca, entidad territorial oficialmente establecida y reconocida dentro de la división político administrativa del país. Por su parte, la Gobernación de Cundinamarca es la entidad pública que administra los recursos del departamento en procura del mejoramiento continuo de la calidad de vida y bienestar de la población cundinamarquesa.

En este orden de ideas, el contexto organizacional de este trabajo está constituido por la Dirección de Aseguramiento de la Secretaría de Salud; entidad encargada de formular y dirigir las políticas de salud en el departamento. La figura 1 muestra el mapa de procesos de la Gobernación de Cundinamarca y la clasificación general en Procesos Estratégicos, Procesos Misionales, Procesos de Apoyo y Procesos de Evaluación [2]. Cada proceso a su vez está compuesto por diversos subprocesos y procedimientos que coadyuvan a la consecución de las metas institucionales trazadas en el Plan de Desarrollo Departamental.

En este escenario, la Dirección de Aseguramiento es la dependencia encargada de garantizar el acceso a los servicios de salud de los habitantes del departamento mediante la ejecución de acciones de vigilancia, seguimiento y control sobre el aseguramiento, y la administración de las distintas fuentes de financiación del Sistema General de Seguridad Social en Salud – SGSSS, para mantener y ampliar la cobertura de la afiliación y prestación de servicios con la provisión de la red de atención.

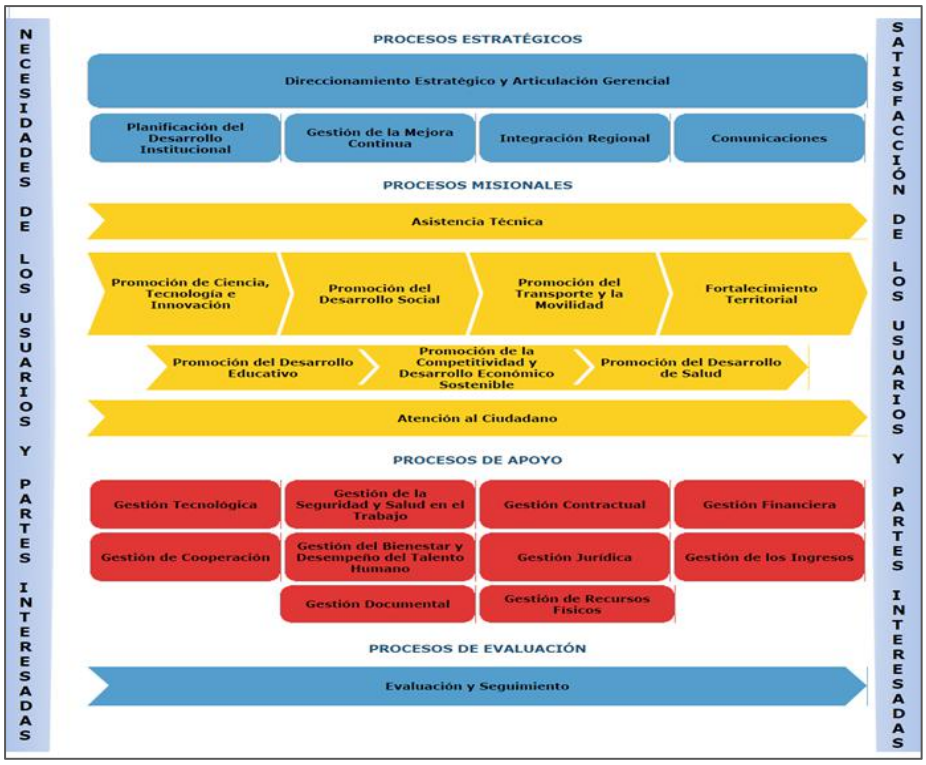


Figura 1. Mapa de procesos - Gobernación de Cundinamarca. [2]

En este punto, es necesario anotar que en la Dirección de Aseguramiento se ejecuta el procedimiento de Seguimiento a la Base de Datos del Aseguramiento que tiene como propósito principal promover la consolidación de la información de la población afiliada al Sistema de Salud, con el fin de realizar análisis de la cobertura del aseguramiento. En este procedimiento, las fuentes de información que alimentan la base de datos son un recurso fundamental para el desarrollo de las distintas actividades técnicas de consulta, cruce y análisis de información; la naturaleza de estos recursos también implica que deben ser gestionados con apego a la normatividad vigente sobre tratamiento de datos personales dado el volumen, el nivel se sensibilidad y detalle que manejan.

Así pues, la Base de Datos del Aseguramiento almacena *data* de considerable volumen pues contiene la información de más de 2.5 millones de afiliados en el departamento, caracterizados a través de distintas variables relacionadas con el Aseguramiento en salud. Ahora bien, en aras de apoyar el propósito fundamental del Aseguramiento en salud, este trabajo plantea que esta información masiva puede ser analizada desde un enfoque no tradicional y orientado hacia el descubrimiento de patrones, relaciones o agrupamientos que permitan mejorar la comprensión de

la dinámica del aseguramiento. Propuesta para la cual la Ciencia de Datos aporta técnicas que permiten abordar la complejidad de las múltiples variables involucradas generando un modelo de analítica que apoye la toma de decisiones frente al enfoque de programas y políticas de cobertura del aseguramiento y acceso a los servicios de salud a través de los distintos prestadores de servicios de salud del departamento.

3.1.3.- Ciencia de datos. Definición y principales conceptos

En este apartado es conveniente partir de la definición de Ciencia de Datos de tal modo que apoye desde lo conceptual el presente trabajo. La ciencia de datos es un área de conocimiento moderna nacida del avance de las tecnologías de la información y las comunicaciones, y su impacto en la ciencia y los fenómenos sociales, generadores de grandes volúmenes de datos cuyo análisis ya no es posible realizar por equipos humanos ni desde enfoques tradicionales de la computación.

Una definición muy completa encontrada en la literatura [3] indica que: “La Ciencia de Datos no es solo un concepto sintético para unificar estadísticas, análisis de datos y sus métodos relacionados, sino que comprende también sus resultados. La Ciencia de Datos intenta analizar y comprender un fenómeno real con ‘datos’. En otras palabras, el objetivo de la ciencia de datos es revelar las características o la estructura oculta de complicados fenómenos naturales, humanos y sociales desde un punto de vista diferente de la teoría y métodos tradicionales. Este punto de vista implica formas de pensamiento mutidimensional, dinámico y flexible.”
Desde un punto de vista más pedagógico, un diagrama bastante conocido que sintetiza los conceptos y habilidades con los que está relacionada la ciencia de datos es el que se muestra en la siguiente figura:

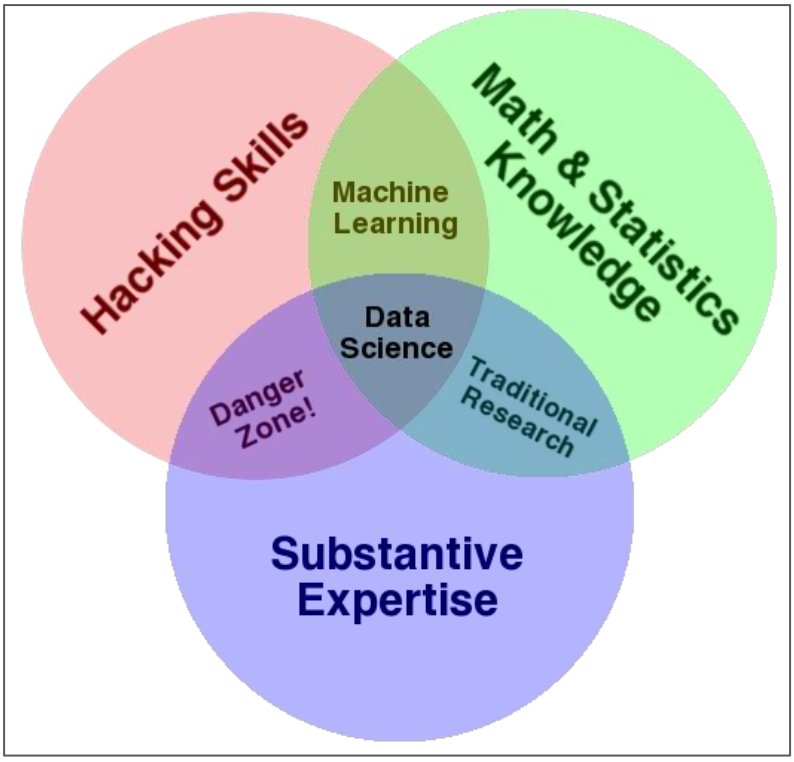


Figura 2. Diagrama de Venn de la Ciencia de Datos. [4]

En este diagrama de Venn, la Ciencia de Datos ocupa la intersección de varios conceptos y habilidades e ilustra con claridad las áreas con las que interactúa confiriéndole un carácter inherentemente multidisciplinar [4]. Además de las habilidades duras (Hacking, conocimiento estadístico y matemático) es de resaltar que confiere gran importancia a la experticia en un área de negocio o de conocimiento determinada. También ilustra con claridad que Ciencia de Datos no es exclusivamente Machine Learning (Aprendizaje de Máquina).

Por otro lado, en Ciencia de Datos, los datos por lo general provienen de entornos sobre los que se tiene conocimiento limitado de las condiciones bajo las cuales fueron generados, recolectados y preparados [5]. En este sentido también es importante anotar que gran parte de estos datos no son estructurados. Por ejemplo, las redes sociales y demás aplicaciones web de interacción, son fuentes donde no es clara la forma y dinámica como se generan los datos. En Ciencia de Datos, y más puntualmente en el fenómeno de Big Data, los datos presentan las dimensiones o características distintivas de volumen, velocidad, variedad, veracidad y valor. Es posible identificar 4 tipos de analítica de datos: Analítica Descriptiva, Analítica de Diagnóstico, Analítica

Predictiva y Analítica Prescriptiva [6]. Los dos primeros tipos se enfocan en la comprensión de los fenómenos o comportamientos sucedidos que generaron los datos, respondiendo en términos generales a las preguntas "¿Qué pasó?" y "¿Por qué sucedió?" respectivamente. Los dos últimos tipos de analítica se enfocan en el futuro de lo que sucederá con los datos y responden a las cuestiones generales "¿Qué pasa si?" y "¿Cómo hacer que suceda?" respectivamente. La figura 3 ilustra mejor estos conceptos;

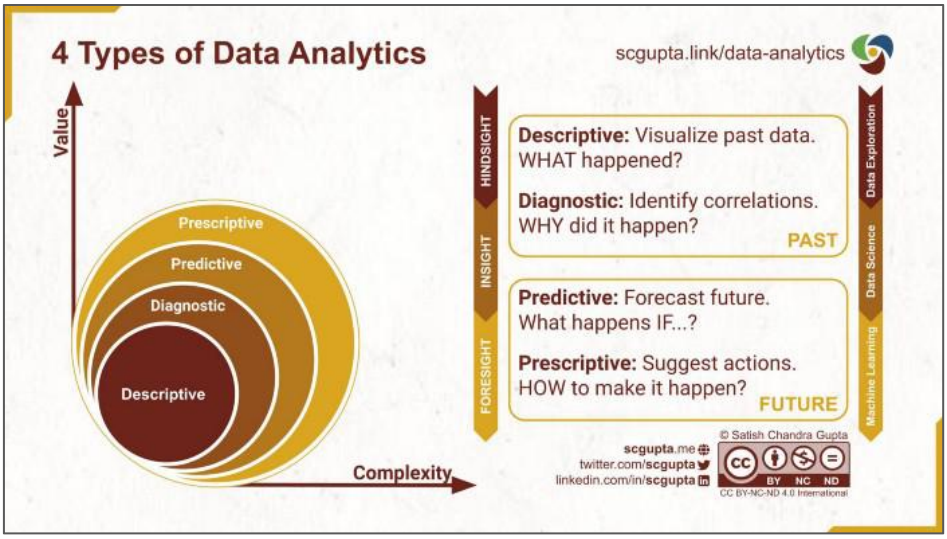


Figura 3. Tipos de analítica de datos. [6]

La Ciencia de Datos se enfoca principalmente en los dos últimos tipos de analítica: Predictiva y Prescriptiva.

En este punto y en el entendido que Machine Learning no es estricto sinónimo de Ciencia de Datos, si es posible anotar que la Ciencia de Datos recurre o se apoya, entre otras áreas de conocimiento, de modo intensivo en técnicas y modelos de Machine Learning para cumplir con su propósito principal descrito en su definición misma. En términos generales podemos clasificar los algoritmos de Machine Learning en las siguientes categorías [7]:

Aprendizaje Supervisado: En los algoritmos de aprendizaje supervisado la tarea principal es la de generación de un modelo, que a partir de un conjunto de datos etiquetados, pueda ser entrenado para clasificar nuevas observaciones o predecir futuros eventos.

Aprendizaje No Supervisado: En los algoritmos de Aprendizaje No Supervisado no se cuenta

con un conjunto de datos que posea un etiquetado o clasificación previa. De tal modo que la tarea o propósito principal es la búsqueda de un modelo que permita encontrar estructuras ocultas en los datos de entrada. Por tanto, es una tarea compleja definir una respuesta correcta para los resultados obtenidos porque no se conoce de antemano una muestra contra la cual comparar.

Aprendizaje Semi-Supervisado: Estos algoritmos combinan los dos enfoques anteriores. A partir de conjuntos de datos etiquetados y no etiquetados, el propósito es generar un modelo que permita realizar tareas de clasificación o predicción.

Aprendizaje por Refuerzo: En estos algoritmos el aprendizaje se logra a partir de un esquema de premios o penalidades según el nivel de error obtenido en los resultados. Un modelo generado a partir de estos algoritmos estará en capacidad de interactuar con un determinado ambiente y de acuerdo con la retroalimentación obtenida, reforzar su aprendizaje a partir de un proceso de ensayo y error.

El Machine Learning engloba y emplea un conjunto extenso de algoritmos y técnicas para construir modelos robustos de analítica de datos; la siguiente figura muestra una clasificación general de estos algoritmos para los tipos de Aprendizaje Supervisado, No Supervisado y por Refuerzo [8].

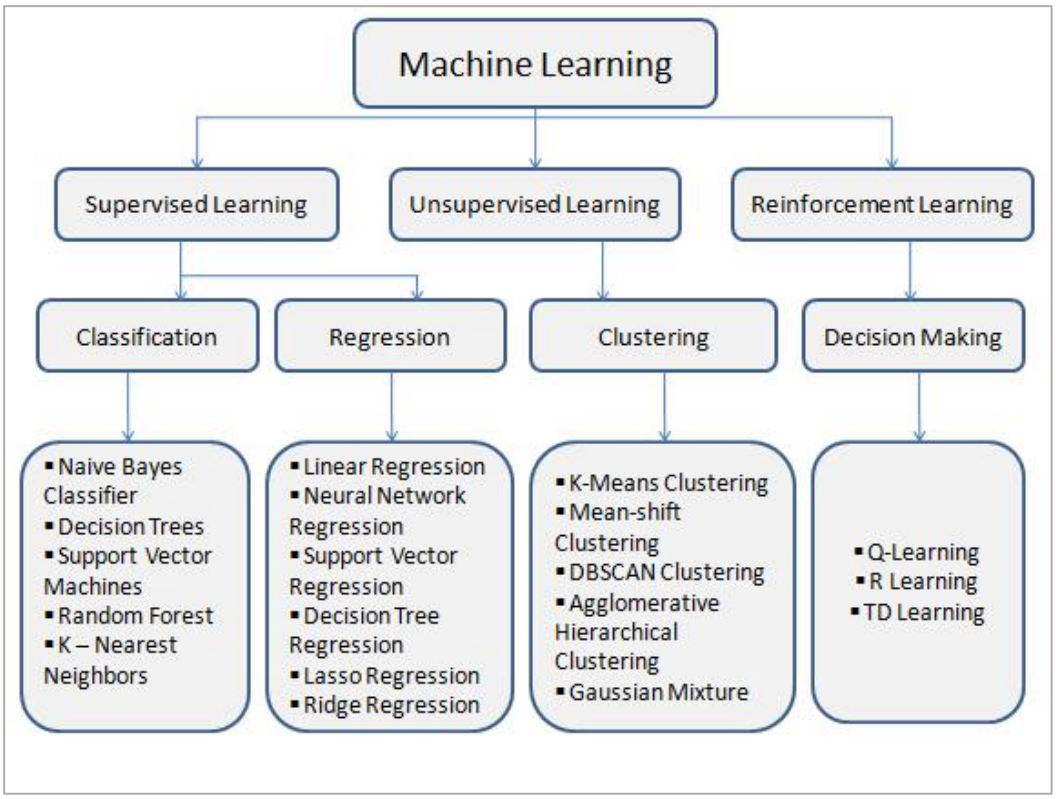


Figura 4. Tipos de algoritmos en Machine Learning. [8]

Dada la naturaleza del objetivo general propuesto, el presente trabajo queda enmarcado en los algoritmos de Aprendizaje de Máquina No Supervisado porque no se conoce de antemano los patrones o estructuras que subyacen a las variables de los afiliados frente a los prestadores de servicios de salud que les brindan la atención; En este orden de ideas, a continuación se describe de modo general la técnica de Clustering que se ajusta y contribuye con el diseño e implementación de un modelo de analítica propuesto.

3.1.4.- Clustering de datos

Esta técnica pertenece al paradigma de Aprendizaje de Máquina No Supervisado; El propósito principal es la organización de los datos no etiquetados en grupos o clusters que comparten características similares. Está orientado principalmente a la identificación de patrones o tendencias en los datos. En esta técnica la variable objetivo es desconocida, por tanto no es

posible ajustar el modelo buscando una salida específica o deseada. Así pues, lo que se busca es encontrar estructuras ocultas en los datos que no es posible a través de programación convencional [9]. Los indivíduos agrupados tienden a compartir a mayor cantidad posible de similaridades, mientras que los grupos o clúster creados tienden a ser los más disímiles posible.

En lenguaje más formal, el clustering de datos puede ser descrito del siguiente modo [10]:

Dado un conjunto de observaciones o indivíduos contables y finitos $\boldsymbol{X} = \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{N}\}$, el propósito de un algoritmo de clustering es el aprendizaje de una tarea que asigna un elemento del conjunto $\boldsymbol{Y} = \{\boldsymbol{y} \mid \boldsymbol{y} \in \mathbb{N}\}$, a cada indivíduo del conjunto $\boldsymbol{X}$. [1]
Se obtiene entonces que cada indivíduo $\boldsymbol{x}$ es etiquetado con un elemento $\boldsymbol{y}$, de tal modo que a los subconjuntos determinados por los elementos $\boldsymbol{y}$ se les llama *clusters* y pueden ser descritos en términos de un conjunto de parámetros $\boldsymbol{\theta}$ que miden el nivel similaridad entre indivíduos y disimilaridad entre clusters.
Es necesario entonces, la definición de una *función de costo o función objetivo* $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$ cuyos argumentos son los parámetros $\boldsymbol{\theta}$ y el conjunto $\boldsymbol{y}$. En el proceso de clustering se busca optimizar la función de costo, lo cual se traduce en un número óptimo $\boldsymbol{k}$ de clusters con indivíduos similares o bien cohesionados y clusters diferenciados o bien separados. Formalmente esto se expresa del siguiente modo:

$$\underset{(\boldsymbol{\theta}, \boldsymbol{y})}{\operatorname{argmin}} \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y}) \qquad (1)$$

3.1.5.- Análisis de Componentes Principales

El Análisis de Componentes Principales es un método estadístico que permite convertir un grupo de variables correlacionadas en un nuevo grupo de variables no correlacionadas ortogonales con el propósito principal de reducir la dimensionalidad del conjunto de datos [11].

En este proceso de transformación, se van generando iterativamente nuevas variables que son combinaciones lineales de las originales y que van recogiendo la mayor cantidad posible de

---

[1] $\mathbb{N}$ representa el conjunto de los números Naturales.

información o variabilidad de los datos. A este nuevo grupo de variables se le conoce como el conjunto de Componentes Principales.

En términos más formales, dado un conjunto de datos con una serie de variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ que caracterizan un grupo de individuos, se busca una nueva serie sde variables $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ que sean combinaciones lineales de las n variables originales, y que no estén correlacionadas; cada una recogiendo sucesivamente la mayor cantidad posible de varianza de los datos [12] .

Así pues, la primera componente principal de este nuevo conjunto está definida por la siguiente ecuación lineal:

$$\mathbf{y_1} = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \qquad (2)$$

y en términos generales, cada componente principal queda definida por la siguiente expresión:

$$\mathbf{y_j} = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jn}x_n \qquad (3)$$

Empleando conceptos de Álgebra Lineal, la ecuación (3) se puede escribir como el siguiente producto :

$$\mathbf{y_j} = \mathbf{a'_j x} \qquad (4)$$

donde $\mathbf{a'_j}$ es la transpuesta del vector $\mathbf{a}$, y representa el siguiente vector de constantes:

$$\mathbf{a'_j} = (a_{1j}, a_{2j}, \ldots, a_{nj}) \qquad (5)$$

y **X** está definida por la siguiente matriz :

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

De lo anterior es posible anotar que una cuestión clave es la determinación de los coeficientes a_{nj} del vector $\mathbf{a'_j}$, que dada la condición de mantener la ortogonalidad (no correlación) entre componentes principales, el módulo del vector $\mathbf{a'_j}$ debe ser igual a **1,** es decir:

$$\mathbf{a'_j a_j} = \sum_1^n \boldsymbol{a}^2_{nj} = 1 \qquad (6)$$

por otro lado, los coeficientes a_{nj} que operan la transformación lineal son los que definen la varianza de las componentes principales, asi por ejemplo la primera componente principal $\boldsymbol{y_1}$ posee los coeficientes de $\mathbf{a_1}$ que aportan la mayor varianza y deben cumplir con la condición de ortogonalidad.

En este punto, es necesario observar que el conjunto total de componentes principales representado por $\mathbf{y}$, queda definido por el siguiente producto entre la matriz $\boldsymbol{A}$ y el vector $\mathbf{x}$:

$$\boldsymbol{y} = \mathbf{Ax} \qquad (7)$$

Donde $\mathbf{A}$ es una matriz cuadrada que contiene todos los $\mathbf{a'_j}$ que escalan el vector $\mathbf{x}$ de las variables originales**.**

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \qquad (8)$$

Así pues, la expresión general queda definida del siguiente modo:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \qquad (9)$$

La matriz $\boldsymbol{A}$ se conoce como la matriz de autovectores o eigenvectores.

3.1.6.- Algoritmo K-means

El algoritmo K-means hace parte de los métodos de Clustering, y por tanto del paradigma de Aprendizaje de Máquina No Supervisado. El propósito general es el agrupamiento de las

observaciones o indivíduos del dataset a partir de una medida de similaridad entre sus características; los clusters o grupos resultantes deben cumplir con la premisa de ser lo más compactos en su interior y lo más aislados entre ellos.
Esta tarea de agrupamiento se lleva a cabo mediante la minimización de la suma de las distancias de los indivíduos al centroide del cluster más cercano al que quedan relacionados [13]. El número *K* de clusters se elige al principio del algoritmo y es una tarea no trivial que requiere ante todo conocimiento del área o contexto para el cual se está realizando el modelo de aprendizaje.
Los pasos generales de K-means son los siguientes:

1.- *Elección del número K de clusters*: De acuerdo con el requerimiento, la experticia y el contexto , se determina el número *K* de clusters para el conjunto de datos.
2.- *Inicialización:* A partir del número K de custers elegido, se definen K centroides iniciales, por lo general de manera aleatoria en el espacio definido por los datos.
3.- *Asígnación de puntos al cluster con centroide más cercano:* Dado que cada centroide define un cluster, *c*ada indivíduo u observación es asignado al cluster cuyo centroide sea más cercano. Esta tarea se realiza midiendo la distancia de los indivíduos a cada centroide; por tanto cada cluster contiene los indivíduos ubicados a menor distancia.
4.- *Actualización de centroides*: Se actualiza la posición de los centroides de cada grupo al promedio de todos los indivíduos del cluster. Puesto que la posición de cada centroide cambia, los clusters también se reconfiguran.
5.- Iteración de los pasos 3 y 4: Se repiten los pasos 3 y 4 hasta que los centroides ya no cambien. Este mínimo de variación de los centroides se obtiene cuando se alcanza el mínimo de la suma general de los cuadrados de las distancias de los indivíduos a cada cluster.

La siguiente figura ilustra el proceso K-means, las expresiones formales que determinan la conformación de los clusters y la función que indica la convergencia del algoritmo.

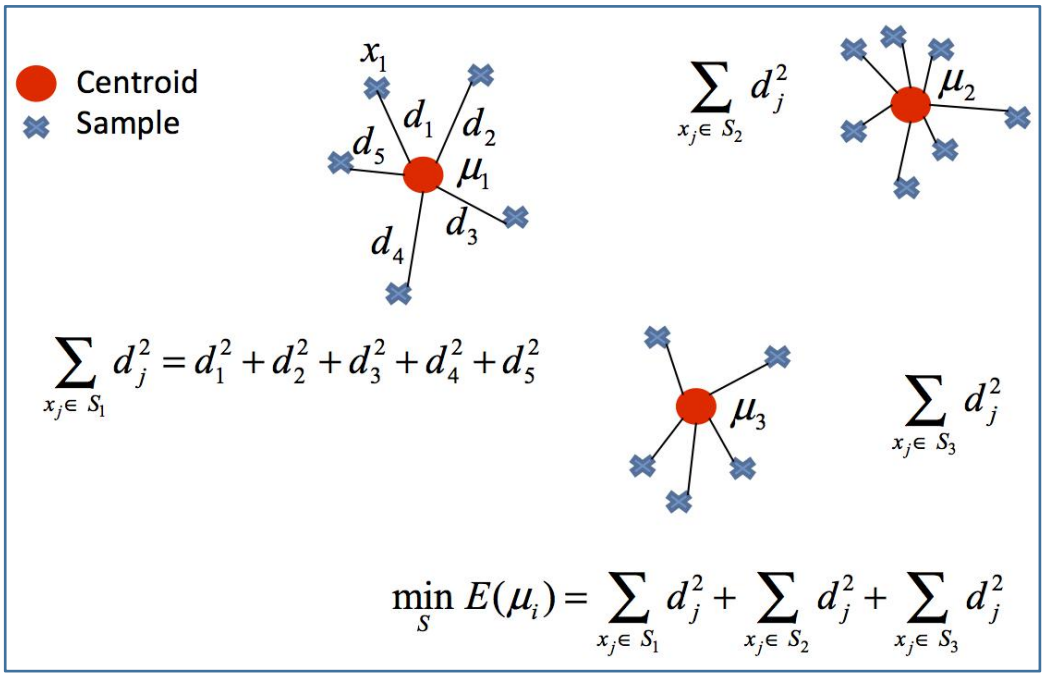


Figura 5. Algoritmo K-means. [13]

Así pues, K-means puede ser comprendido como un algoritmo de optimización en el que se busca como parámetro de convergencia la minimización de la suma de las distancias cuadráticas de cada indivíduo al respectivo centroide del cluster al cual está asociado.
Esta función de minimización se denomina función objetivo, que se expresa de la siguiente manera:

$$\min_{s} E(\mu_i) \;=\; \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - \mu_j\|^2 \qquad (10)$$

donde s representa un determinado cluster de los k construidos, y en cada cluster se busca la distancia mínima entre los individuos x_i y el respectivo centroide μ_j.

3.1.6.1.- Distancia: La distancia entre los indivíduos y entre los clusters es lo que determina el concepto de similaridad en los algoritmos de clustering. Es por tanto un elemento crítico a la hora de la ejecución del algoritmo y la conformación de los clusters.
Dado un conjunto de datos $n - dimensional$ (n variables) y un determinado número de observaciones o puntos de datos, la similaridad entre dos observaciones queda determinada por

la cercanía entre ellas [14], es decir por la menor distancia que las separa. Existen varias métricas utilizadas en clustering para la medición de la distancia.

- *Distancia Euclídea:* Es la métrica más intuitiva puesto que corresponde a la longitud del segmento de recta que une dos puntos. En un plano cartesiano, corresponde a la distancia que puede hallarse por el teorema de Pitágoras entre dos puntos p y q:

$$d\ (p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \qquad (11)$$

En términos generales, para un espacio o conjunto de datos multidimensional, cada punto queda definido por un vector con las coordenadas o variables: $\mathbf{p} = (x_p\ ,\ y_p\ ,\ \ldots,\ z_p)$ y $\mathbf{q} = (x_q\ ,\ y_q\ ,\ \ldots,\ z_q)$. De tal modo que la ecuación anterior queda expresada del siguiente modo general:

$$d\ (p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2 + \ldots\ +\ (z_p - z_q)^2} \qquad (12)$$

- *Distancia Manhattan:* En esta métrica la distancia entre dos puntos p y q, equivale a la sumatoria de las diferencias absolutas entre cada dimensión. La expresión que describe el concepto de la distancia Manhattan es la siguiente:

$$d_{manhattan}(p, q) = \sum_{i=1}^{n} |(p_i - q_i)| \qquad (13)$$

donde n representa el número de variables que describen los puntos o individuos, p_i representa el valor de la variable i en el punto p , y q_i representa el valor de la variable i en el punto q .

La siguiente figura ilustra mejor los conceptos de distancia Euclídea (en color rojo) y distancia Manhattan (en color verde) para dos puntos p y q en un espacio de dos dimensiones.

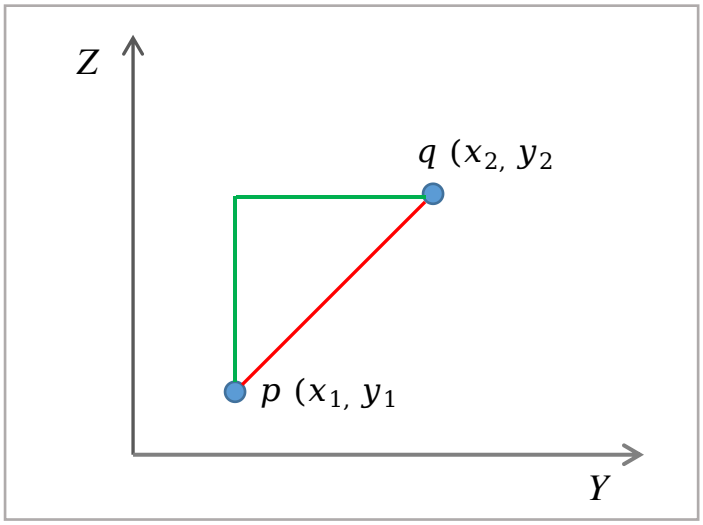


Figura 6. Distancia Euclídea (rojo) y distancia Manhattan (verde) (espacio 2D).

- *Distancia de Correlación:* En esta métrica se emplea el concepto de correlación para determinar el grado de similaridad entre los puntos u observaciones de un conjunto de datos. Dado un par de puntos p y q , caracterizados por los vectores de variables $\mathbf{p} = (x_p , y_p , \ldots , z_p)$ y $\mathbf{q} = (x_q , y_q , \ldots , z_q)$, la diferencia entre el nivel de correlación perfecta (1) y el grado de correlación entre ellos, indica el grado de similaridad entre los puntos. Siendo 0 la distancia más cercana y por tanto el mayor nivel de similaridad. La expresión que describe este concepto es el siguiente:

$$d_{corr}(p, q) = 1 - |cor(p, q)| \qquad (14)$$

donde $cor(p, q)$ representa la correlación entre los vectores $\mathbf{p}$ y $\mathbf{q}$.

Es posible aplicar diferentes tipos de correlación, según sea el contexto y la naturaleza del requerimiento para el modelo de clustering. Así pues, es posible aplicar la correlación de Pearson, Spearman, Kendall, etc.

3.1.6.2.- Determinación del número óptimo de clusters

Como se mencionó anteriormente, la determinación del número K óptimo de clusters es un paso fundamental en un algoritmo de clustering. En esta sección se repasan los métodos más conocidos para la aproximación a este criterio.

Método del Codo: El método del codo es una técnica que desde un enfoque heurístico basado en la experiencia, aporta un criterio para aproximarse al número óptimo de clusters. Emplea el concepto de la suma total de cuadrados de las distancias dentro del cluster (WSS - por sus siglas en inglés) que en un proceso iterativo de búsqueda debe ser la mínima. Dicho de otra manera, se busca que la variación de la distancia intra-cluster sea la mínima [15].
La expresión que describe este criterio es la siguiente:

$$min(\sum_{k=1}^{k} W\ (C_k)) \qquad (15)$$

donde el término $W(C_k)$ mide la variación intracluster y C_k hace referencia al k-ésimo cluster en el proceso de iteración. Empleando un elemento gráfico, la idea es identificar el punto de inflexión (una especie de 'codo') donde la suma de cuadrados comienza a estabilizarse.

Método de la Silueta: Este método aporta un criterio numérico conocido como el Coeficiente de la Silueta que puede adquirir valores en el rango entre -1 y 1. El coeficiente indica la medida de cuán similares son los indivíduos de un cluster (concepto de Cohesión) y que tan diferentes son respecto a los indivíduos de los otros clusters (concepto de Separación) [16].
La interpretación del score o puntaje obtenido es la siguiente:
Un *coeficiente=1* indica que el indivíduo está asignado de manera óptima a un cluster y los clusters están bién separados.
Un *coeficiente=0* indica traslapamiento entre los clusters.
Un *coeficiente=-1* indica que el indivíduo está asignado de manera incorrecta en los clusters.
La expresión que calcula el coeficiente de la silueta para cada indivíduo u observación, es la siguiente:

$$S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \qquad (16)$$

donde:

$S(i)$ es el coeficiente para el punto i .

$a(i)$ es la distancia promedio entre el punto i y los demás puntos que pertenecen al mismo cluster.

$b(i)$ es la distancia promedio desde el punto i a todos los clusters a los cuales no pertenece; más específicamente al cluster más cercano, pues garantiza que no pertenece al resto.

Método Gap Statistic (Brecha estadística): En este método se compara la variación intra-cluster para diferentes valores de *k* de la distribución real del conjunto de datos con los valores esperados para una distribución de referencia en la que no se evidencien clusters obvios (distribución uniforme). Esta distribución de referencia se obtiene mediante una simulación de Montecarlo. Para cada variable x_i en el conjunto de datos se calcula el rango $[min(x_i), max(x_j)]$ y mediante el proceso de simulación se generan los valores uniformes u homogéneos para los n puntos en cada intervalo $[min(x_i), max(x_j)]$.

La expresión que describe el proceso del método Gap Statisctic es la siguiente:

$$Gap_n(k) = E_n^* logW_k - logW_k \qquad (17)$$

Donde E_n^* representa la Esperanza estadística para una muestra de tamaño n de la distribución de referencia; se calcula mediante un proceso de remuestreo (bootstraping) generando B copias de las muestras de referencia y calculando el promedio $logW_k^*$ (observar que el asterico *W** indica la suma total de cuadrados dentro del cluster de las muestras de referencia). El proceso Gap Statistic mide la desviación del W_k observado versus el valor esperado en la distribución uniforme; se establece la verificación o rechazo de una hipótesis nula que supone la no existencia de clusters en el conjunto de datos.

3.1.7.- Algoritmo Clustering Jerárquico

El clustering jerárquico, hace parte de los algoritmos o técnicas del Aprendizaje de Máquina No Supervisado cuyo propósito general es también la búsqueda de agrupaciones de indivíduos u observaciones homogéneas en un conjunto de datos. En el clustering jerárquico no se requiere que se determine a priori un número k de clusters, y en el proceso se agrupan clusters o

indivíduos para formar nuevas agrupaciones (clustering jerárquico aglomerativo) o se busca separar un cluster ya existente para crear dos nuevas agrupaciones (clustering jerárquico divisivo).

Clustering jerárquico aglomerativo: También se denomina clustering jerárquico ascendente, es decir que el agrupamiento inicia a partir de todos los indivíduos que existen, considerando inicialmente un grupo por cada indivíduo. A partir de estas unidades se van construyendo agrupaciones, desde la base a la cima, culminando en una agrupación general que engloba a todos los clusters que se han ido generando en el proceso. Este enfoque de clustering Jerárquico también recibe el nombre de Anidamiento Aglomerativo (por sus siglas en inglés AGNES - Agglomerative Nesting).

Clustering Jerárquico Divisivo: También denominado clustering jerárquico descendente, es decir el proceso de clustering comienza con la concepción de un solo grupo general presente en el conjunto de datos. A partir de este metagrupo se van realizando divisiones sucesivas desde la cima a la base, que van generando clusters más pequeños. Este enfoque también recibe el nombre de Análisis Divisivo (por sus siglas en inglés DIANA- Divisive Analysis) .

La figura 7 ilustra los dos principales enfoques del clustering jerárquico, AGNES y DIANA. Los autores(*Kamande et al, 2018*) han empleado una representación en forma de árbol en el que el método AGNES va desde las hojas a la raiz y el método DIANA desde la raiz a las hojas; es decir siguen una dirección inversa en lo referente al proceso de clustering.

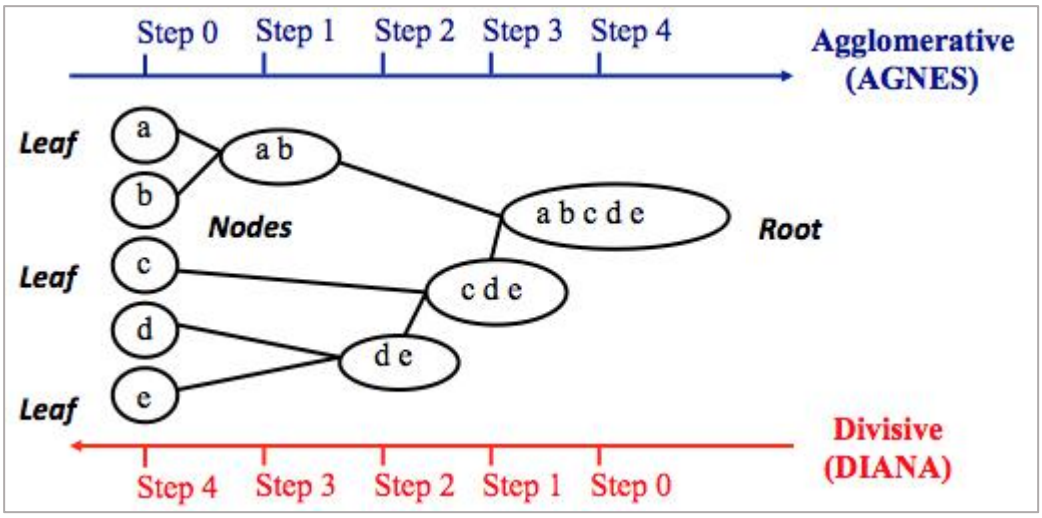


Figura 7. Clustering Jerárquico Aglomerativo. [17]

3.1.7.1.- Métodos de Similitud entre clusters

En el clustering jerárquico, además de las métricas de distancia entre los inviduos, es necesario establecer el criterio de *linkage* (enlace) para medir la distancia que determina la similitud entre los clusters. Los tipos de *linkage* más utilizados son los siguientes:

Único o Mínimo (Single or Minimun): En este método se calcula la distancia entre todos los pares de indivíduos de un cluster 1 y los indivíduos de un cluster 2. Se elige la menor distancia encontrada como la medida de disimilaridad entre los dos clusters; es decir, la mínima distancia de separación intercluster.

Completo o Máximo (Complete or Maximun): En este método de linkage se calcula la distancia entre todos los pares de indivíduos de un cluster 1 y los indivíduos de un cluster 2. Se elige la mayor distancia encontrada como la medida de disimilaridad entre los dos clusters; es decir, la máxima distancia de separación intercluster.

Promedio (Average): En este método de linkage se calcula la distancia entre todos los pares de indivíduos de un cluster 1 y los indivíduos de un cluster 2. Se elige la distancia promedio como la medida de disimilaridad entre los dos clusters; es decir, la distancia promedio de separación intercluster.

Centroide (Centroid): En este método se calcula la distancia entre el centroide de un cluster 1 y el centroide de un cluster 2.

Ward: En este método se busca minimizar la varianza total dentro del cluster. En cada iteración del proceso Ward, se busca aquellos dos clusters en los que una combinación conllevaría menor varianza dentro de cada cluster.

La siguiente figura ilustra cada método de linkage en un espacio de datos bidimensional (2 variables).

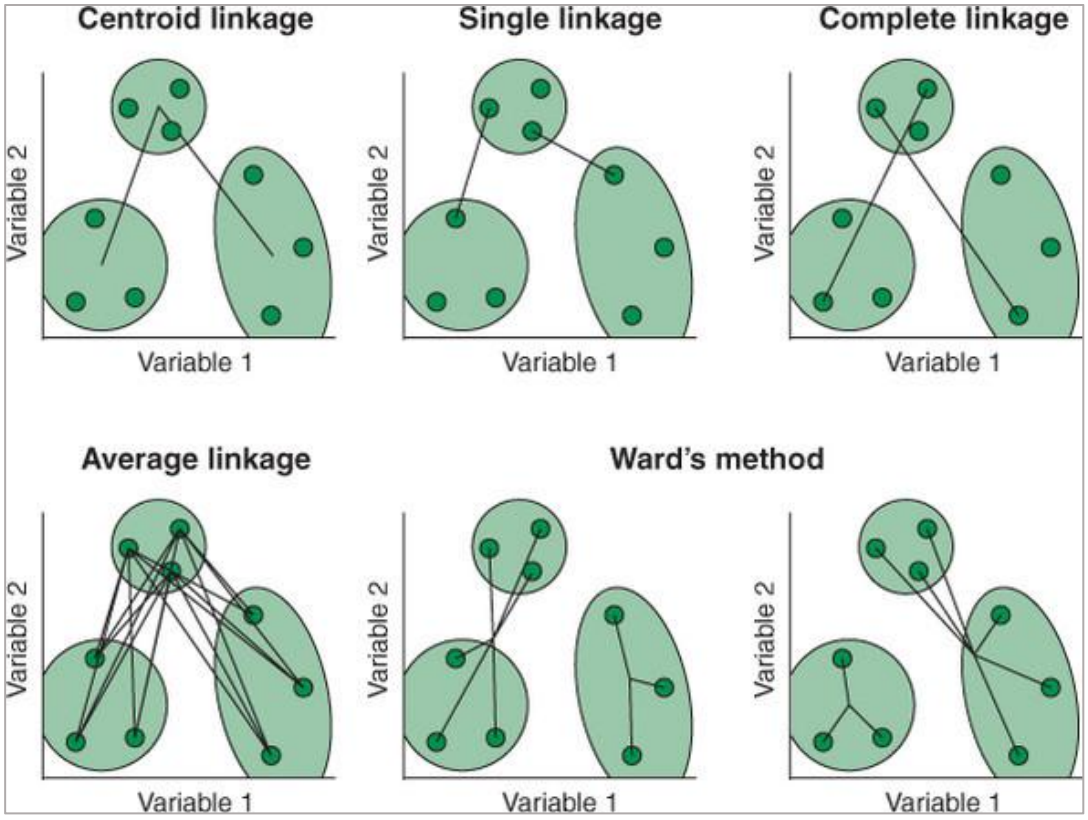


Figura 8. Tipos de linkage. [18]

3.1.8.- Algoritmo K-medoids

El algoritmo K-medoids es muy similar al algoritmo K-means; la diferencia radica en que los elementos centrales alrededor de los cuales convergen los clusters, en K-medoids corresponden a un individuo u observación; a este elemento se le denomina *medoid*. (En k-means este elemento corresponde a los centroides).

Una forma de comprender el concepto de *medoid* es asumirlo como el indivíduo más

representativo de un cluster; Así pues, en un determinado cluster, se busca que la distancia promedio entre el *medoid* y los demás indivíduos sea la mínima.

La implementación más conocida del modelo K-medoids se realiza mediante el algoritmo PAM (Partition Around Medoids). La ejecución del algoritmo considera el siguiente escenario general: Dado un conjunto de datos representados por *U*, el subconjunto *M* corresponde a los indivíduos seleccionados como *medoids* y el subcojunto *I* corresponde al resto de indivíduos no seleccionados; de tal modo que

$$I = U - M \qquad (18)$$

Para la determinación del subconjunto *M* y respectivos clusters asociados, el algoritmo PAM ejecuta en los siguientes pasos generales.

1.- En la primera fase, se elige de manera aleatoria una colección de k objetos candidatos para el subconjunto de medoids *M*.

2.- Asígna cada indivíduo al *medoid* más cercano. Este paso lo realiza mediante el cálculo de las distancias entre todos los indivíduos, asignando al medoid aquellos indivíduos con distancias más cercanas. En este paso quedan definidos los clusters preliminares.

3.- En este paso se mejora la calidad del clustering preliminar, cambiando iterativamente los *medoids* iniciales. Hasta conseguir *medoids* con los cuales se mejora la distancia promedio de los indivíduos.

La diferencia más notable entre K-means y K-medoids radica en el hecho de que los centroides no están asociados a una observación particular como los medoids; lo que implica, al igual que la mediana, que un medoid puede ser la mejor elección de medida de tendencia central cuando existen ciertos indivíduos u observaciones que se alejan del promedio.

3.2.- ANTECEDENTES

La ciencia de datos ha permitido abordar con éxito desafíos complejos del sector salud en distintas temáticas que van desde la investigación y diagnóstico médico hasta el apoyo a las labores administrativas y la formulación de políticas de salud en organismos gubernamentales. A continuación, se consolidan algunos antecedentes documentados en la literatura sobre este tema que permitieron fortalecer el propósito del presente trabajo.

El Centro de Datos de Familia de la Universidad de la Florida, creó mapas de densidad de puntos calientes con variables sociales y sanitarias para identificar disparidades en la región de interés de cubrimiento de servicios de salud. En este proyecto se emplearon intensivamente la tecnología GIS (Sistemas de Información Geográfica) y fuentes de datos públicas nacionales, estatales y locales con indicadores sociodemográficos (población total, grupos etarios, ingreso familiar, nivel educativo, conformación familiar), de salud (bajo peso al nacer, mortalidad infantil, tasas de infecciones de transmisión sexual, maltrato infantil) y tasas de muerte por causas seleccionadas (diabetes, enfermedades cardíacas, cáncer, accidente cerebrovascular, homicidio, suicidio, accidente automovilístico). La articulación de esta información y la interacción con los habitantes, permitió identificar disparidades en los servicios disponibles en el territorio para mejorar la inversión de recursos y el acceso a la atención médica [19].

Por otro lado, una iniciativa interesante en esta área de aplicación de la ciencia de datos, la constituye el proyecto sobre la “Predicción de futuros puntos calientes de Hospitalizaciones Potencialmente Prevenibles” [20]. En este proyecto se aplicaron modelos predictivos para identificar posibles áreas geográficas críticas con altas tasas de Hospitalizaciones Potencialmente Prevenibles, con el propósito de enfocar las intervenciones tempranas de los servicios de salud. Esto es importante porque permite optimizar recursos y se observa que con tales intervenciones tempranas los puntos críticos pueden ser regresados a comportamientos normales dentro del promedio. El estudio tiene como contexto geográfico Australia y como enfermedades con Hospitalizaciones Potencialmente Prevenibles se tomaron diabetes mellitus tipo II, insuficiencia cardíaca, enfermedad pulmonar obstructiva crónica y "pie de alto riesgo".

Centrando el enfoque en los prestadores de servicios de salud, Byrne et al. [21] implementaron un método para identificar pares de hospitales que permitan establecer comparaciones más

equilibradas frente a la calidad y financiamiento de estas entidades. La fuente de datos empleados provino de las bases de datos del Departamento de Asuntos para los Veteranos de los Estados Unidos. Estos grupos de pares, se determinaron considerando 133 centros médicos y 16 variables relacionadas con las facilidades de atención. El criterio inicial para la agrupación fue la distancia euclídea entre los indivíduos , esto es, la determinación de la matriz de distancias euclídeas entre todos los indivíduos del dataset. Posteriormente un indivíduo constituye el centro para su vecino más cercano, conformando de este modo los pares más homogéneos. También aplicaron métodos de clustering para poder establecer el modelo con el mejor desempeño frente a la generación de los pares. En este enfoque cada indivíduo o centro médico constituye el centro de su propio grupo, de tal modo que la comparación es equilibrada evitando que algunos centros médicos queden relegados o subestimados cuando los grupos son más grandes.

Por su parte, Delamater et al. [22] proponen una metodología de clustering de centros médicos con base en la similiaridad frente a los patrones de demanda de servicios de la comunidad y la ubicación geográfica. Con el propósito de apoyar la planificación de las políticas de salud a nivel regional en el estado de Michigan en los Estados Unidos. Para el desarrollo del modelo de clustering emplearon los algoritmos K-means y Clustering Jerárquico con el método de linkage Ward. Como resultado identificaron 33 clusters que agrupan los centros médicos frente a necesidades reales de demanda de servicios y con un enfoque en las áreas geográficas donde se ubican estos prestadores.

Con el propósito de identificar necesidades reales de salud, Nnoaham & Cann [23] proponen la determinación de segmentos o grupos de población bien diferenciados con base en el servicio de la atención primaria en salud en la región de South Wales Valleys en el Reino Unido. Los datos que usaron provienen de una base de datos de 79.607 pacientes que también incluyen variables socioeconómicas, comorbilidades, salud preventiva, e información de riesgo y costo. Para el desarrollo del modelo de clustering emplearon el algoritmo K-means que permitió la determinación de 10 clusters de pacientes con patrones similares de utilización del servicio de atención primaria; ingresaron 7 variables puntuales de uso para el modelo: Admisiones electivas de hospitalización, admisiones no electivas de hospitalización, atenciones ambulatorias, atenciones de seguimiento, visitas al departamento de emergencias, visitas y prescripciones al servicio de atención primaria.

En Colombia, un estudio interesante sobre el tema, aborda la determinación de clusters de prestadores de salud en el país [24]. En este trabajo se crearon grupos de prestadores con base en las características o variables de capacidad instalada y nivel de atención, con el propósito de lograr criterios más precisos de comparación, evaluar el desempeño y mejorar la calidad en la prestación del servicio. Para lograr este propósito emplearon un modelo de clustering, basado en K-means, generando 7 clusters con base en variables propias de los prestadores en las categorías Camas, salas, ambulancias, apoyo terapéutico.

4.- COMPRENSIÓN, EXPLORACIÓN Y PREPARACIÓN DE LOS DATOS

"El pequeño hexágono meditó un rato
sobre esto y luego me dijo:
—Pero tú has estado enseñándome a elevar números a
la tercera potencia: supongo que 9 elevado a la 3 tiene
que significar algo en geometría, ¿qué significa?"

EDWIN ABBOTT [Planilandia]

En este apartado se consignan los principales aspectos relacionados con la comprensión del área de negocio para la cual se desarrolló el modelo de analítica, las fuentes de datos que soportan el modelo y la preparación de los datos de tal modo que posean la estructura apropiada para el análisis.

4.1.- Área del negocio

En el área del Aseguramiento en Salud, uno de los principales propósitos es el de garantizar el acceso a los servicios de salud para los afiliados al Sistema General de Seguridad Social en Salud (SGSSS). En relación con este propósito, es completamente indispensable la consolidación y gestión de la información de los afiliados de tal modo que permita su adecuada caracterización y seguimiento.

En este orden de ideas, la afiliación al SGSSS es un evento que ocurre una sola vez y su duración es vitalicia; en otros términos, una persona se afilia una sola vez en su vida y los datos que aporta permiten su identificación en el sistema de salud. Sin embargo, algunos parámetros de información son objeto de diferentes eventos y actualizaciones, que ocurren a medida que transcurre la interacción del afiliado con el entorno geográfico, sociodemógráfico, laboral o familiar; lo que confiere a la información de afiliación una naturaleza altamente dinámica.

Los distintos parámetros de esta información capturan aspectos importantes que permiten la caracterización de los afiliados, facilitando la atención y acceso a los servicios y programas de salud.

4.2.- Fuentes de Datos

En este apartado se presentan las fuentes de datos que aportan información para la consolidación del conjunto de datos sobre el que se plantea el modelo.

Base de Datos Única de Afiliados (BDUA) : Es la base de datos que contiene la información que identifica plenamente a los afiliados al Sistema General de Seguridad Social en Salud. Esta información está estructurada y contiene los atributos que permiten (entre otros parámetros) identificar para cada afiliado, el municipio de afiliación, la EPS, la IPS de atención primaria, el estado de afiliación. Este recurso es administrado por la Administradora de los Recursos del Sistema General de Seguridad Social en Salud [25].

Registro de Prestadores de Servicios de Salud (REPS) : Este recurso contiene la información que identifica a los prestadores de servicios de salud habilitados en el departamento de Cundinamarca, y en el país. Entre otros atributos, contiene el código del prestador, sedes, la naturaleza jurídica del prestador (público, privado o mixto), el municipio de operación. Este recurso es administrado y provisto por el Ministerio de Salud [26].

Programa Sisbén: Este recurso contiene la información proveniente de la encuesta Sisbén que el Departamento Nacional de Planeación (DNP) aplica a los habitantes de una determinada entidad territorial y permite identificar el nivel socioeconómico del encuestado y su núcleo familiar, con miras a focalizar programas sociales del estado en los grupos de población más vulnerables de la sociedad [27].

A partir de las fuentes descritas se genera el conjunto de datos consolidado sobre el que se implementó el modelo de analítica planteado en este trabajo. A continuación, se consigna un esquema que ilustra esta relación.

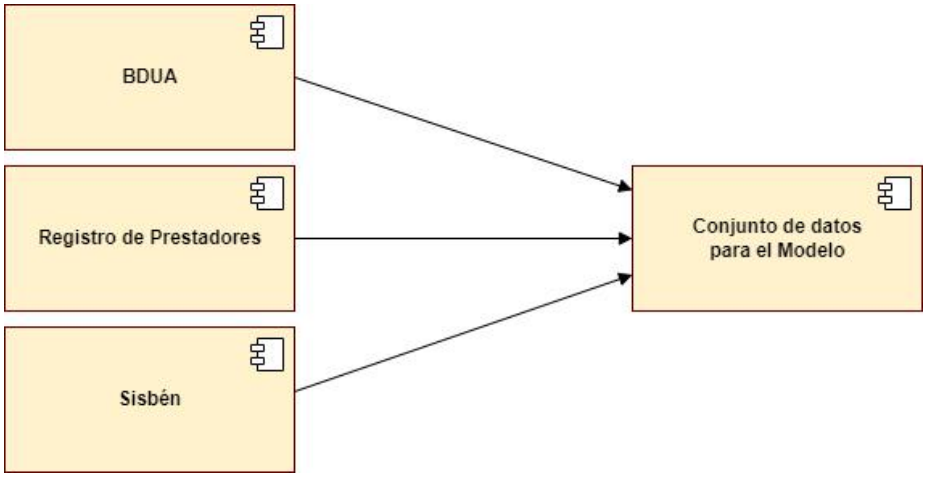


Figura 9. Fuentes del conjunto de datos para el modelo.

Es importante anotar que en el marco del presente trabajo, las fuentes anteriores tienen alcance departamental y corresponden a respectivas tablas dentro de la Base de Datos del Aseguramiento; un recurso administrado en la Dirección de Aseguramiento en el marco de sus funciones y competencias.

4.3.- Exploración de los datos

En este apartado se consignan los principales resultados relacionados con la etapa de exploración de los datos.

4.3.1.- Estructura del dataset

El dataset generado y que constituye la fuente para el modelo, contiene 240 registros (prestadores de servicios de salud) y 17 variables. Los registros corresponden a los prestadores de salud públicos y privados que cuentan con operación habilitada para el departamento de Cundinamarca.

De las 17 variables, 4 variables son tienen el propósito de mejorar la comprensión de los resultados y cobran relevancia en el componente de visualización: *Nombre_ips_prim, cod_ips, naturaleza_ips, Total_ips.* 13 variables son numéricas y son las que se emplean para la implementación del modelo: *Primera_infancia, Infancia, Adolescencia, Juventud, Adultez,*

Adulto_mayor, Femenino, Masculino, Nivel_1, Nivel_2, Nivel_3, Rural, Urbano. Estas variables son las que caracterizan a los afiliados que han registrado como institución de atención primaria uno de los 240 prestadores. Su elección obedece, por un lado, a las variables de los afiliados en las fuentes de información disponibles, por otro lado, a la articulación de estas variables con lineamientos metodológicos y técnicos de caracterización poblacional dispuestos por el Ministerio de Salud y Protección Social [28].

Tabla 1. Variables del conjunto de datos

Variable	Descripción	Tipo de variable
Nombre_ips_prim	Nombre del prestador	Categórica
Cod_ips	Codigo del prestador	Categórica
Naturaleza_ips	Naturaleza juridica del prestador	Categórica
Total_ips	Total de afiliados del prestador	Numérica
Primera_infancia	Número de afiliados de la Primera infancia	Numérica
Infancia	Número de afiliados Infantes	Numérica
Adolescencia	Número de afiliados Adolescentes	Numérica
Juventud	Número de afiliados Jóvenes	Numérica
Adultez	Número de afiliados Adultos	Numérica
Adulto_mayor	Número de afiliados Adultos mayores	Numérica
Femenino	Número de afiliados Mujeres	Numérica
Masculino	Número de afiliados Hombres	Numérica
Nivel_1	Número de afiliados Nivel socioeconomico 1	Numérica
Nivel_2	Número de afiliados Nivel socioeconomico 2	Numérica
Nivel_3	Número de afiliados Nivel socioeconomico 3	Numérica
Rural	Número de afiliados residentes en zona Rural	Numérica
Urbano	Número de afiliados residentes en zona Rural	Numérica

A continuación, se presenta una tabla sumaria de las variables, con los principales indicadores estadísticos:

Tabla 2. Indicadores estadísticos del dataset

```
##  nom_ips_prim         cod_ips            total_ips       Primera_infancia
##  Length:240         Length:240         Min.   :    1.0   Min.   :   0.00
##  Class :character   Class :character   1st Qu.:  727.5   1st Qu.:  48.75
##  Mode  :character   Mode  :character   Median : 3734.5   Median : 271.00
##                                        Mean   : 8477.4   Mean   : 580.92
##                                        3rd Qu.:10355.0   3rd Qu.: 632.25
##                                        Max.   :94785.0   Max.   :6484.00
##     Infancia        Adolescencia       Juventud        Adultez
##  Min.   :   0.00   Min.   :   0.0   Min.   :    0   Min.   :    0.0
##  1st Qu.:  57.75   1st Qu.:  54.5   1st Qu.:  114   1st Qu.:  250.2
##  Median : 302.00   Median : 351.5   Median :  610   Median : 1456.0
##  Mean   : 698.72   Mean   : 787.1   Mean   : 1492   Mean   : 3535.8
##  3rd Qu.: 772.75   3rd Qu.: 916.0   3rd Qu.: 1671   3rd Qu.: 4162.0
##  Max.   :8492.00   Max.   :9576.0   Max.   :18743   Max.   :41843.0
##   Adulto_mayor           F                M             Nivel_1
##  Min.   :    0.0   Min.   :    0.0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:  127.8   1st Qu.:  357.2   1st Qu.:  333.2   1st Qu.:  227.5
##  Median :  701.0   Median : 1903.0   Median : 1894.0   Median : 1431.0
##  Mean   : 1382.9   Mean   : 4297.9   Mean   : 4179.5   Mean   : 2577.6
##  3rd Qu.: 1619.5   3rd Qu.: 5212.8   3rd Qu.: 5049.5   3rd Qu.: 3068.8
##  Max.   :10692.0   Max.   :47933.0   Max.   :46852.0   Max.   :20491.0
##     Nivel_2           Nivel_3            Rural             Urbano
##  Min.   :    0.00   Min.   :    0.0   Min.   :    0.00   Min.   :    0
##  1st Qu.:   75.25   1st Qu.:   20.0   1st Qu.:   86.75   1st Qu.:  422
##  Median :  649.00   Median :  144.0   Median :  900.00   Median : 2208
##  Mean   : 1672.38   Mean   :  664.9   Mean   : 1632.28   Mean   : 6845
##  3rd Qu.: 2142.50   3rd Qu.:  629.8   3rd Qu.: 2229.00   3rd Qu.: 6527
##  Max.   :24509.00   Max.   :12125.0   Max.   :11019.00   Max.   :92322
##      nat
##  Length:240
##  Class :character
##  Mode  :character
```

La tabla de indicadores estadísticos generales del dataset muestra detalles interesantes que es necesario destacar: Existe un prestador con un total de 94.785 afiliados activos registrados y a su vez un prestador con 1 afiliado activo registrado. Esta observación no puede ser considerada como ‘outlier’ (valor atípico) porque refleja un hecho real en el Aseguramiento en Salud, en el que existen prestadores que por su tamaño, ubicación urbana y capacidad presentan acumulaciones de afiliados activos que los han registrado como prestadores de atención primaria. Así como prestadores que por su ubicación rural o distante de las principales ciudades del departamento, atienden pocos usuarios. En este trabajo se ha considerado que ambos tipos de prestadores deben ser considerados en el modelo, puesto que la cifras representan en realidad personas afiliadas con con pleno derecho a la cobertura en los servicios de salud.

Otro dato interesante es el hecho de que algunas variables presentan el valor cero (0), esto no implica que el cero sustituye un valor indefinido, implica que en realidad el prestador no registra usuarios afiliados según una determinada variable. Este valor tampoco se omite en razón al argumento anterior.

4.3.2.- Análisis descriptivo de las Variables.

A continuación, se presentan los resultados obtenidos por cada variable del dataset y que constituyeron la entrada para el modelo.

1.- Total de usuarios afiliados en los prestadores: En la siguiente figura se muestra la distribución de los afiliados activos según la naturaleza jurídica de los prestadores que les brindan la atención primaria en salud.

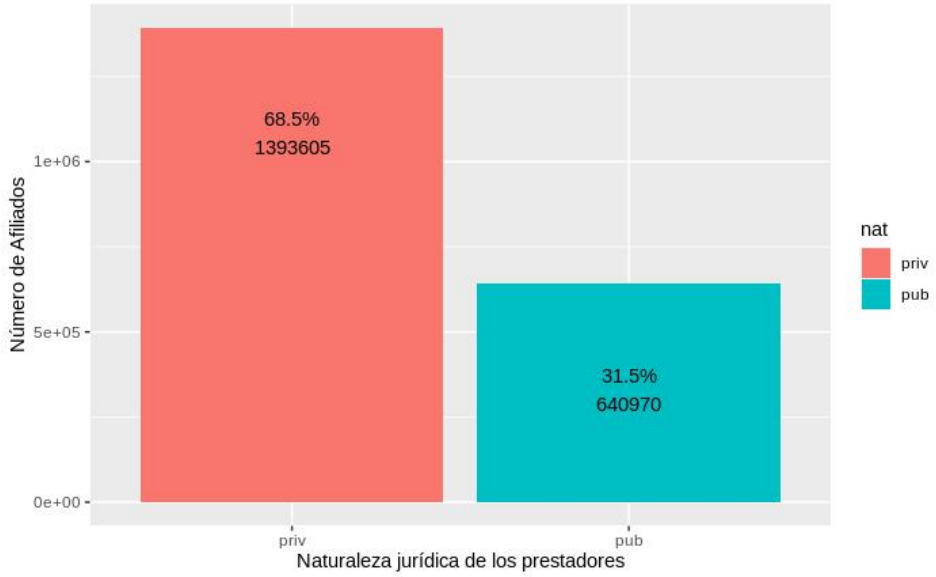


Figura 10. Distribución de los afiliados según la naturaleza juridíca de los prestadores.

Es interesante notar que el 68.5 % de los afiliados activos tienen como institución de atención primaria un prestador privado (Tabla 3); lo cual implica que en la formulación y el seguimiento a las políticas y programas de salud, es esencial la articulación entre prestadores de distinta naturaleza para lograr las metas trazadas en materia de cobertura del aseguramiento y acceso a los servicios de salud.

Tabla 3. Total de afiliados en los prestadores públicos y privados a febrero de 2023.

Tipo de prestador	número de prestadores	Número de afiliados activos	porcentaje
Prestadores públicos	124	640.970	31,5%
Prestadores privados	116	1.393.605	68,5%
Total	**240**	**2.034.575**	

A partir de este punto es necesario indicar que las variables se analizan segun su clasificación general en variables de Grupo Etario, Género, Nivel Socioeconómico y Zona; con el fin de facilitar el análisis y la presentación de los resultados.

2.- Variables de Grupo Etario: El dataset contiene 5 variables relacionadas con el grupo etario en el que se encuentran los afiliados: Primera infancia (0 a 5 años), Infancia (6 a 11 años), Adolescencia (12 a 17 años), Juventud (18 a 28 años), Adultez (29 a 59 años) y Adulto Mayor (mayor de 60 años). La siguiente figura presenta la distribución del número de afiliados por cada grupo etario.

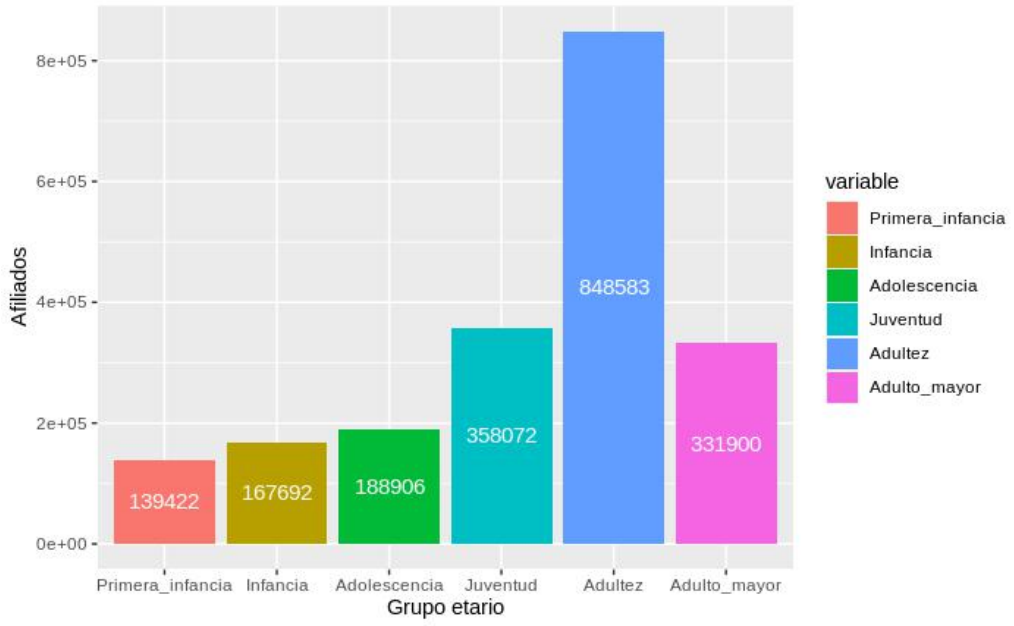


Figura 11. Distribución de afiliados por grupos etarios.

Una observación interesante es el hecho de que los afiliados en las etapas de Juventud, Adultez y Adulto mayor, son las más númerosas. A los cursos de vida Juventud y Adultez pertenece la población económicamente activa, y dado que muchos afiliados en grupos etarios poseen una vinculación laboral, realizan aportes parafiscales al sistema de salud a través del Régimen Contributivo. Al curso de vida Adulto Mayor corresponde la población que en virtud de su edad posiblemente requiera un enfoque diferencial frente a los programas y políticas de salud.

Tabla 4. Distribución de los afiliados por grupo etario

Grupo etario	Número de afiliados activos	porcentaje
Primera infancia	139.422	6,85%
Infancia	167.692	8,24%
Adolescencia	188.906	9,28%
Juventud	358.072	17,6%
Adultez	848.583	41,71%
Adulto mayor	331.900	16,31%
Total	**2.034.575**	

Por otro lado, la siguiente figura presenta un gráfico de caja y bigotes (Boxplot) que muestra los principales indicadores estadísticos por cada variable de grupo etario. Como se indicó anteriormente, en el gráfico se observa la presencia de valores que se ubican por fuera de las marcas o hitos estadísticos y que representan cifras de afiliados a los prestadores de mayor tamaño.

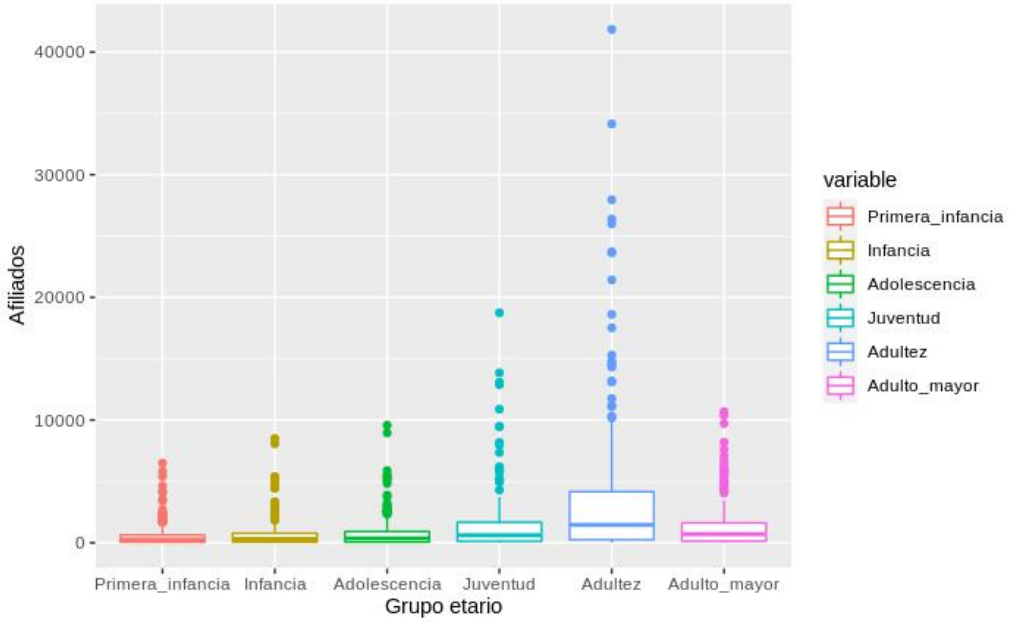


Figura 12. Gráfico boxplot por grupo etario.

3.- Variables de Género: El dataset presenta una distribución balanceada frente a estas dos variables (Femenino [F] y Masculino [M]).

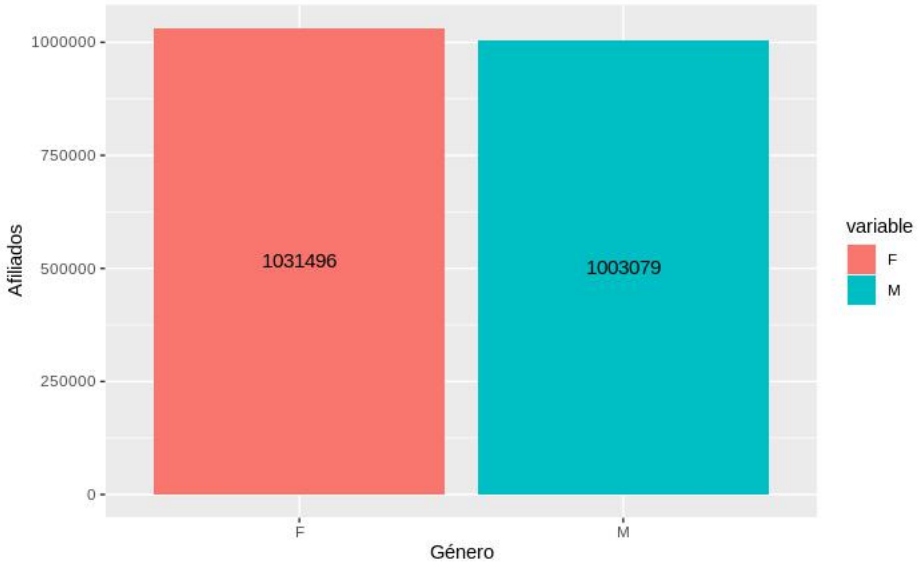


Figura 13. Distribución de afiliados por género

Tabla 5. Distribución de los afiliados por género

Género	Número de afiliados activos	porcentaje
Femenino	1.031.496	50,7%
Masculino	1.003.079	49,3%
Total	**2.034.575**	

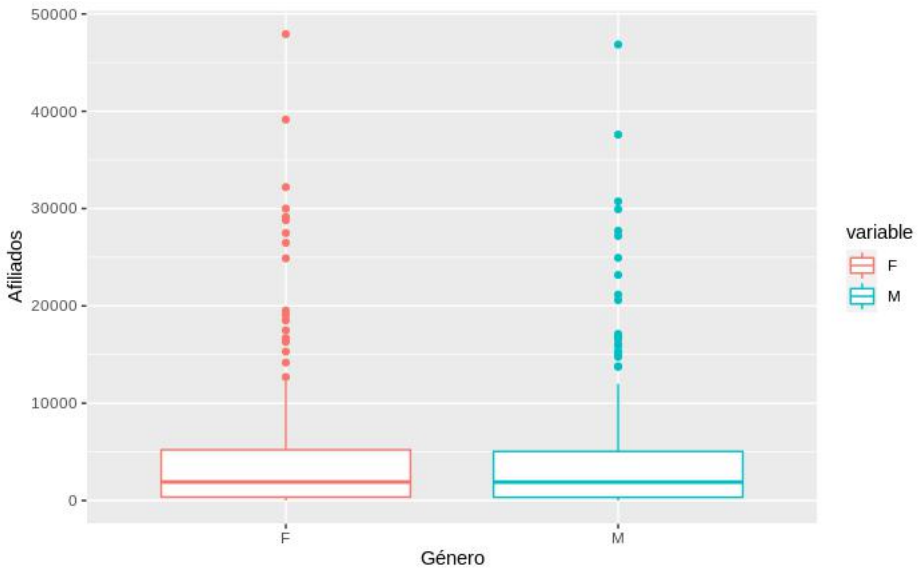


Figura 14. Gráfico boxplot por Género.

4.- Variables de Nivel Socioeconómico: Estas variables indican la clasificación socioeconómica de los afiliados que han solicitado o aceptado la encuesta Sisbén del Departamento Nacional de Planeación. Actualmente, en el sector salud se definen 3 niveles socioeconómicos que permiten enfocar los programas y asistencia financiera del Estado para la atención en salud; siendo en su orden las poblaciones de nivel 1 y 2 las que presentan vulnerabilidad socio-económica, y la población de nivel 3 la que posee capacidad de pago de los servicios de salud.

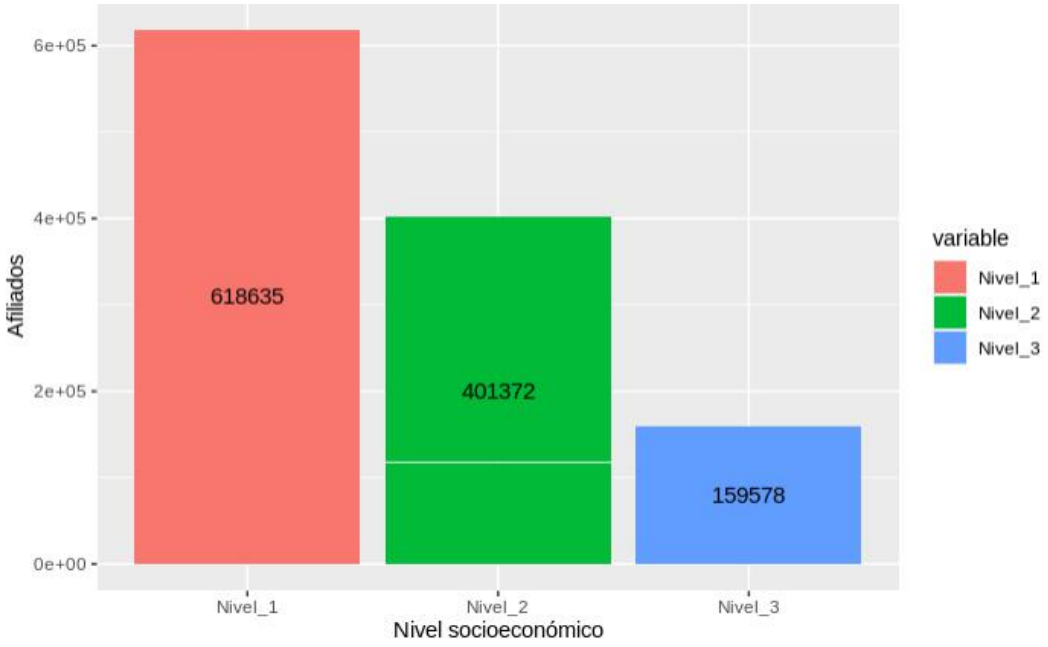


Figura 15. Distribución de afiliados por nivel socioeconómico

La tabla 6 muestra que los niveles socioeconómicos 1 y 2 son los más númerosos entre los afiliados con encuesta Sisbén, lo que resulta de interés en varios temas del sector salud a la hora de enfocar metas y políticas sobre la población más vulnerable o de escasos recursos económicos.

Tabla 6. Distribución de los afiliados por género

Nivel socioeconómico	**Número de afiliados con encuesta Sisbén**	**porcentaje**
Nivel 1	618.635	52,45%
Nivel 2	401.372	34,03%
Nivel 3	159.578	13,53%
Total	**1.179.585**	

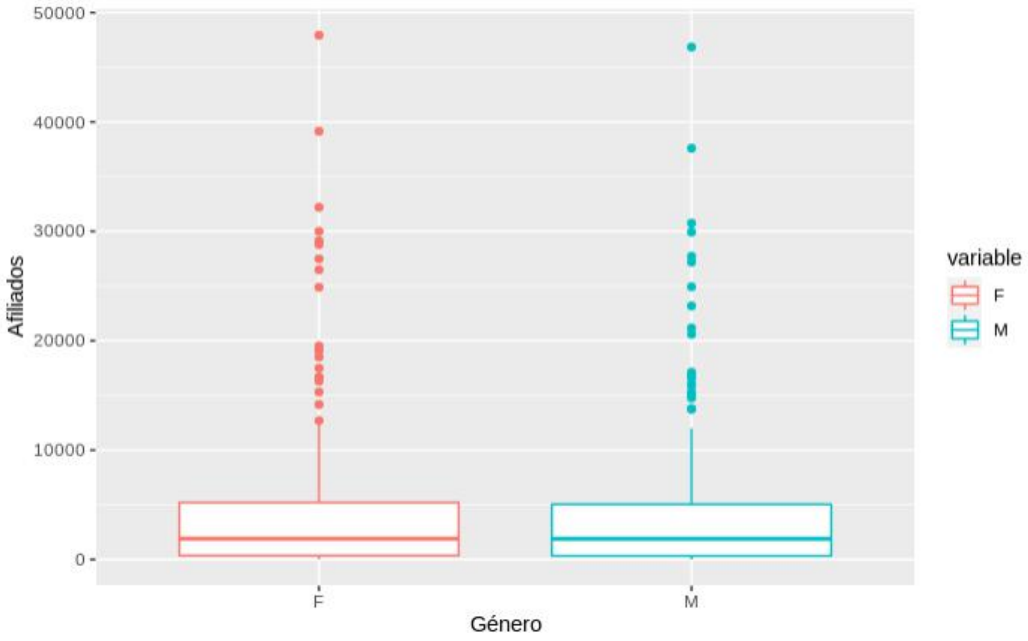


Figura 16. Gráfico boxplot por Nivel Socioeconómico.

5.- Variables de Zona: La figura 17 presenta la distribución de afiliados según la zona de residencia. Claramente se observa la tendencia hacia la residencia en la zona urbana del departamento, lo que también constituye un detalle interesante para los programas y políticas del relacionadas con la atención y servicios de salud.

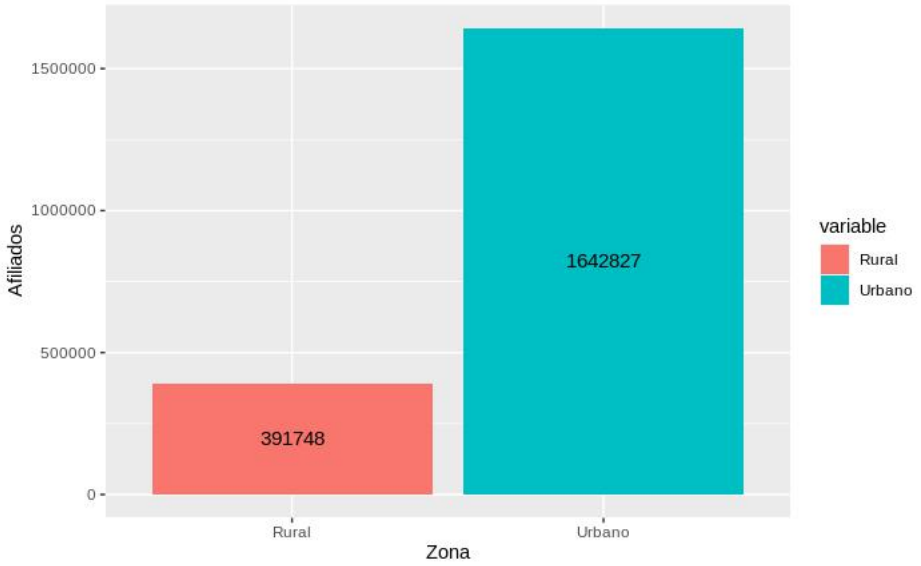


Figura 17. Distribución de afiliados por zona de residencia

Tabla 7. Distribución de los afiliados por zona de residencia

Zona de residencia	Número de afiliados	porcentaje
Urbano	1.642.827	80,75%
Rural	391.748	19,25%
Total	**2.034.575**	

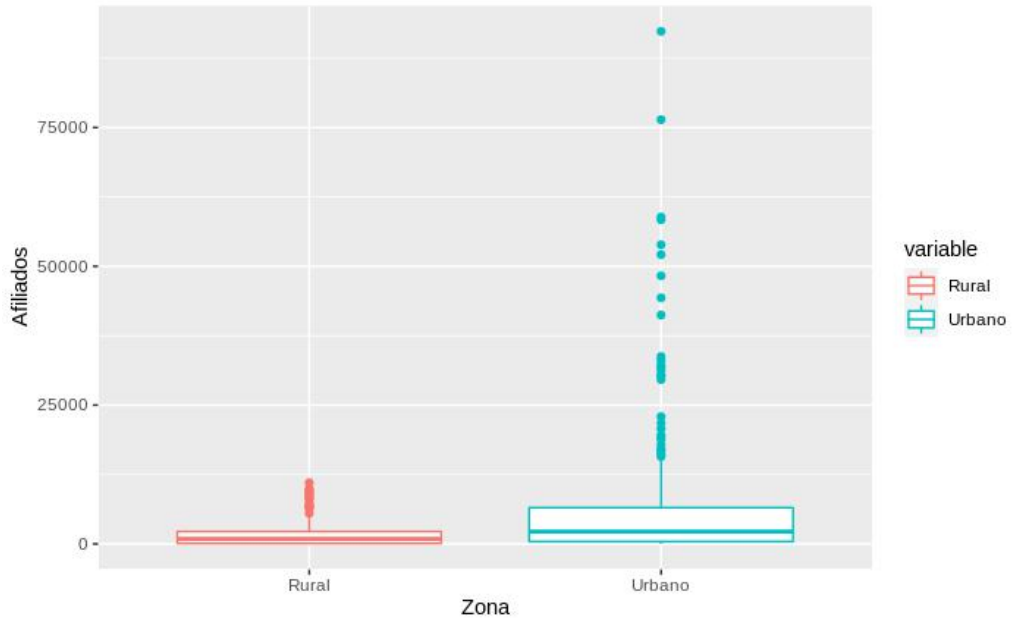


Figura 18.- Gráfico boxplot por Zona

4.3.3.- Normalización de las variables y verificación de tendencias en los datos

Se aplicó el proceso de normalización a las variables del dataset llevándolas a la condición de estandarización que implica *media=0* y *desviación estándar=1*; esto permitió una escala común para las variables y atenuar la situación de los valores altos que aportan algunos prestadores al restringirlos a un rango específico.

Por otro lado, se realizó un ejercicio de verificación preliminar de existencia tendencias en los datos. Es decir, verificar que en la estructura de los datos existan indicios de agrupaciones reales de tal modo que un modelo de clustering no realice agrupaciones al azar. Para esta tarea se empleó el estadístico de Hopkings [28] que en términos generales mide el nivel de uniformidad y aleatoriedad en la distribución de los datos. valores cercanos a 0 indican que los datos presentan posibles agrupaciones, mientras que valores cercanos a 0.5 indican que los datos presentan una distribución uniforme que posiblemente no tenga agrupaciones.

Para el conjunto de datos, el valor obtenido fue H= 0.067. Lo cual es un buen indicador de agrupaciones naturales en los datos.

4.3.3.- Medida de correlaciones

El dataset consolida el conteo de afiliados activos por cada prestador, de acuerdo con las variables que los caracterizan. Esto permite inferir que las variables empleadas están altamente correlacionadas. Para verificar esta situación se recurrió a la medición de correlaciones empleando el coeficiente de Correlación de Pearson [29] que está definido por la siguiente expresión:

$$r_{xy} = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}} \quad (19)$$

donde x y y representan la variable independiente y dependiente respectivamente, que pueden tomar valores usualmente en el conjunto de los números reales $\mathbb{R}$. El coeficiente r_{xy} representa el nivel de correlación entre las variables x y y que puede variar entre -1 y 1 , siendo $r_{xy} = -1$ una correlación perfectamente negativa y $r_{xy} = 1$ una correlación perfectamente positiva. n representa el número de observaciones *(x,y)*.

A continuación, se presenta una representación gráfica que muestra alta correlación entre las variables. Sin embargo, es interesante observar la baja correlación entre la zona rural y las demás variables; especialmente entre la zona rural y el nivel socioeconómico 3. Lo cual es consistente con el hecho de que la mayor parte de los afiliados se encuentra en el sector urbano del departamento.

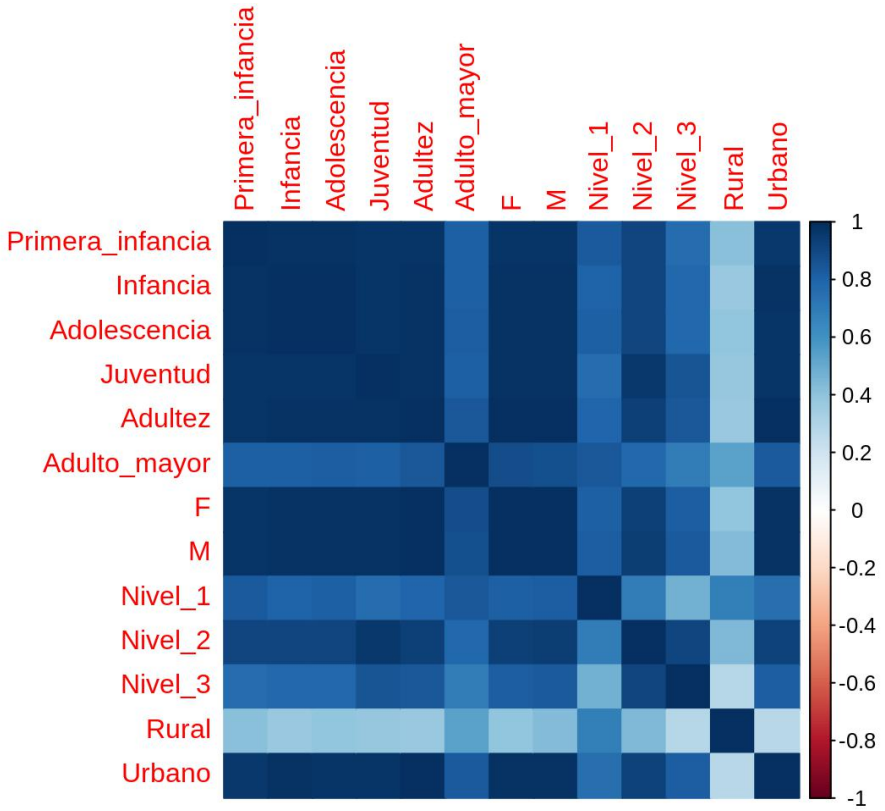


Figura 19. Matriz gráfica de correlaciones.

4.4.- Preparación de los datos

En esta etapa, se realizaron las tareas relacionadas con la preparación de los datos para el modelo. Se recurrió a la técnica de Análisis de Componentes Principales (PCA por sus siglas en inglés) para profundizar en la comprensión de la estructura de los datos, pero también para efectos de reducir la dimensionalidad teniendo en cuenta que tratamos con un dataset que presenta múltiples variables.

Para el caso de este trabajo, teniendo en cuenta que cada dimensión corresponde a la combinación lineal de las 13 variables incluidas en el análisis, a continuación se presenta la varianza explicada por cada una las dimensiones.

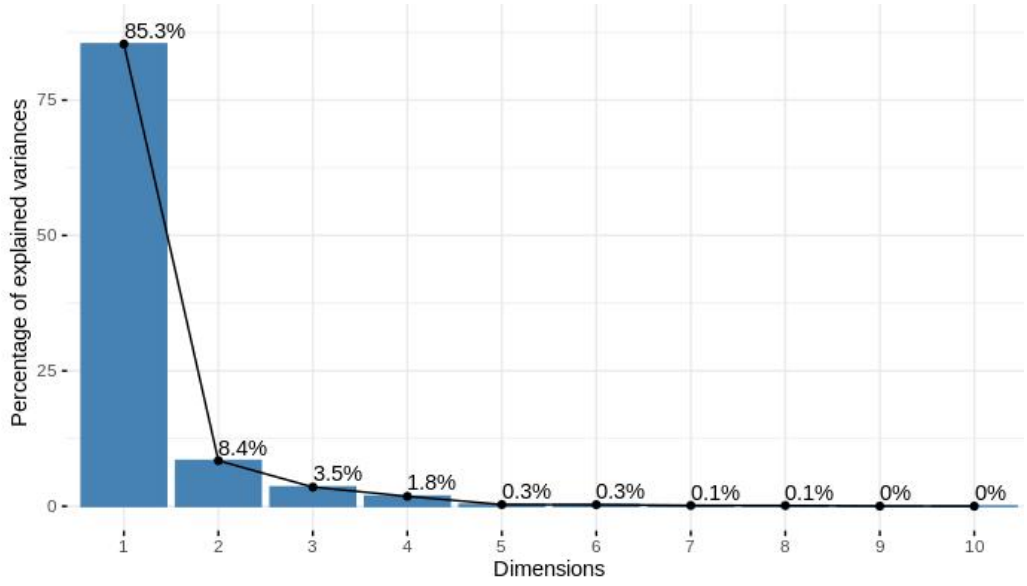


Figura 20. Componentes principales y varianza explicada

Se observa que las dos primeras dimensiones explican más del 93% de la varianza en los datos. Este dato es relevante en la etapa de desarrollo del modelo porque constituyen la entrada para los algoritmos de clustering.

También se empleó el concepto de *coseno cuadrado (cos2)* para verificar la calidad de representación de las variables en las dos dimensiones principales.
El coseno cuadrado indica la contribución de un componente al cuadrado de la distancia de la observación al origen [30]. Lo que se corresponde con el cuadrado del coseno del ángulo del triángulo rectángulo que se forma entre el origen, la observación y su proyección sobre el componente. El parámetro *cos2* adquiere valores entre 0 y 1; de tal modo que valores iguales o cercanos a 1 indican la mejor calidad de la representación en las dimensiones.

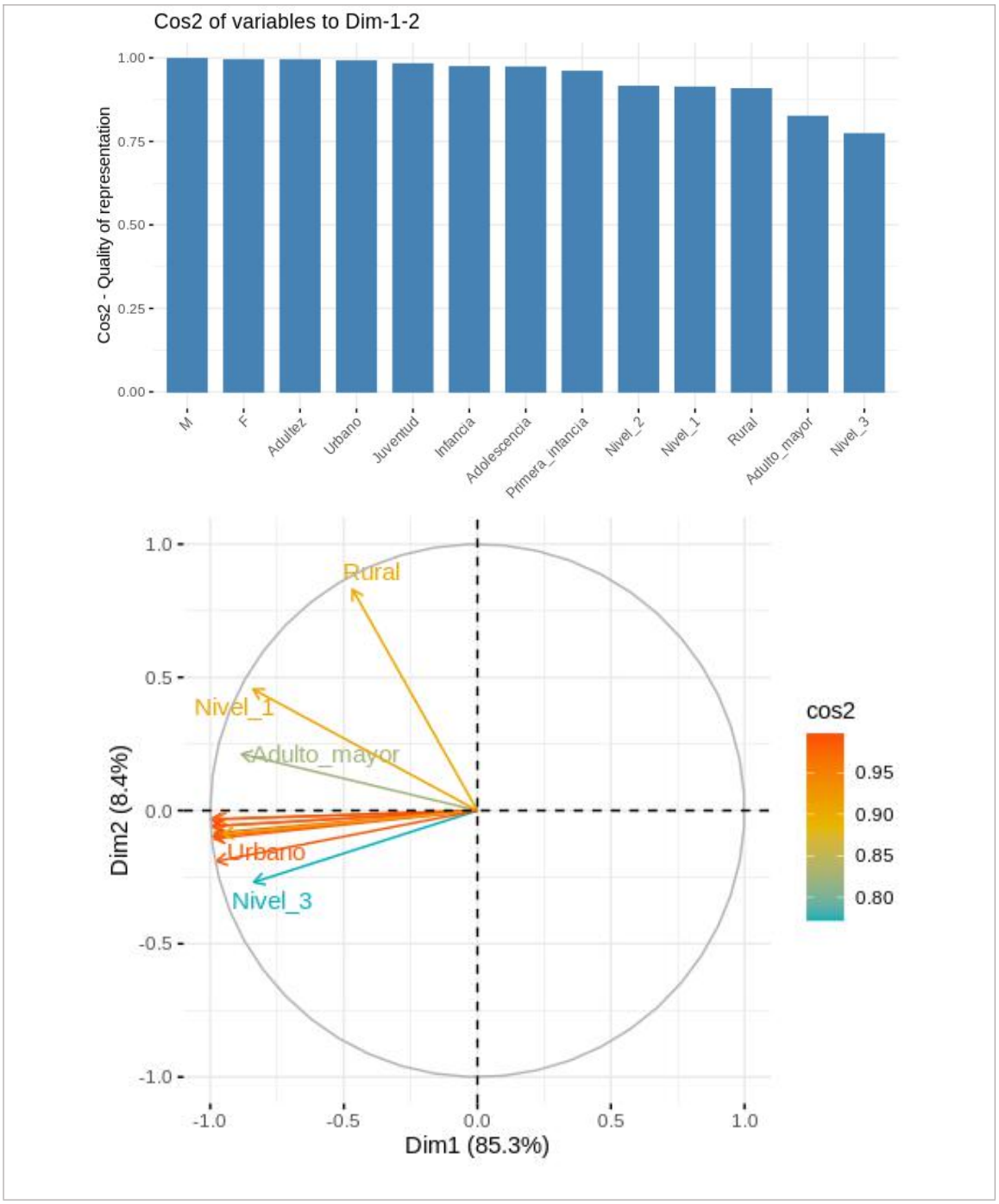


Figura 21. Calidad de representación de las variables - Coseno cuadrado.

La figura 22 presenta el resultado de la distribución de los prestadores en las dos dimensiones principales. Con la abreviatura ***ip_*** están representados prestadores privados y con la abreviatura ***e_*** están representados los prestadores públicos.

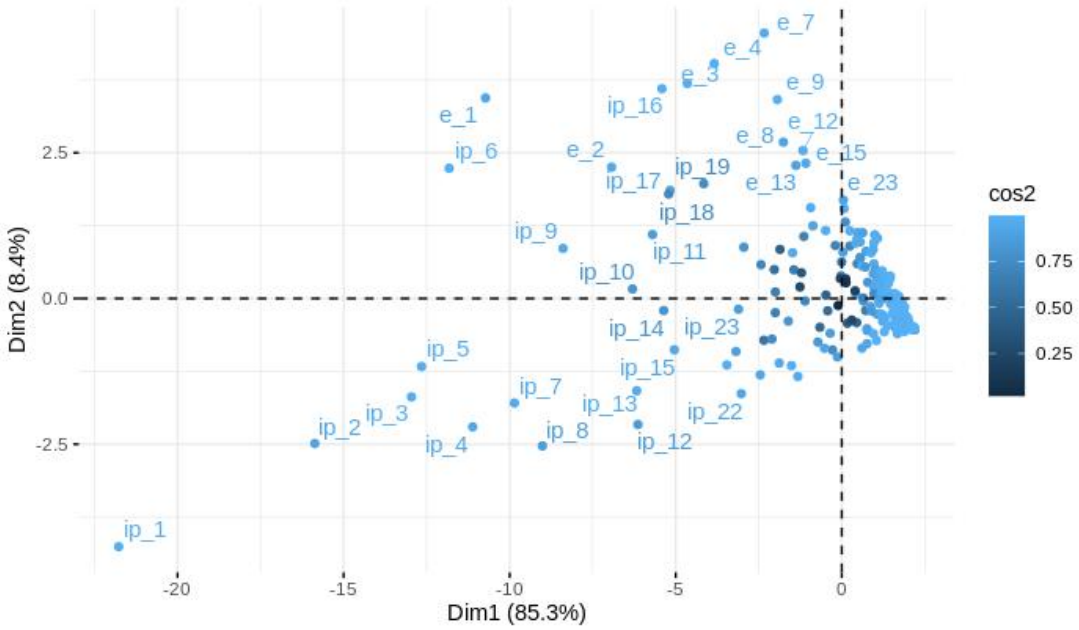


Figura 22. Distribución de prestadores (indivíduos) en las dimensiones 1 y 2

Es necesario anotar que e este gráfico la escala es adimensional, es decir no representa las unidades de medida o magnitudes de las variables.
Se observa una distribución con forma triangular en la que los prestadores con menos afiliados tienden a agruparse en uno de los vertices, a partir del cual tiende a proyectarse o expandirse los prestadores de mayor tamaño.

Por su parte, la figura 23 muestra en un solo elemento gráfico la distribución de los prestadores y variables en el plano de las dos dimensiones principales. En este gráfico es posible identificar la tendencia de los prestadores hacia la cobertura de determinadas variables de los afiliados.

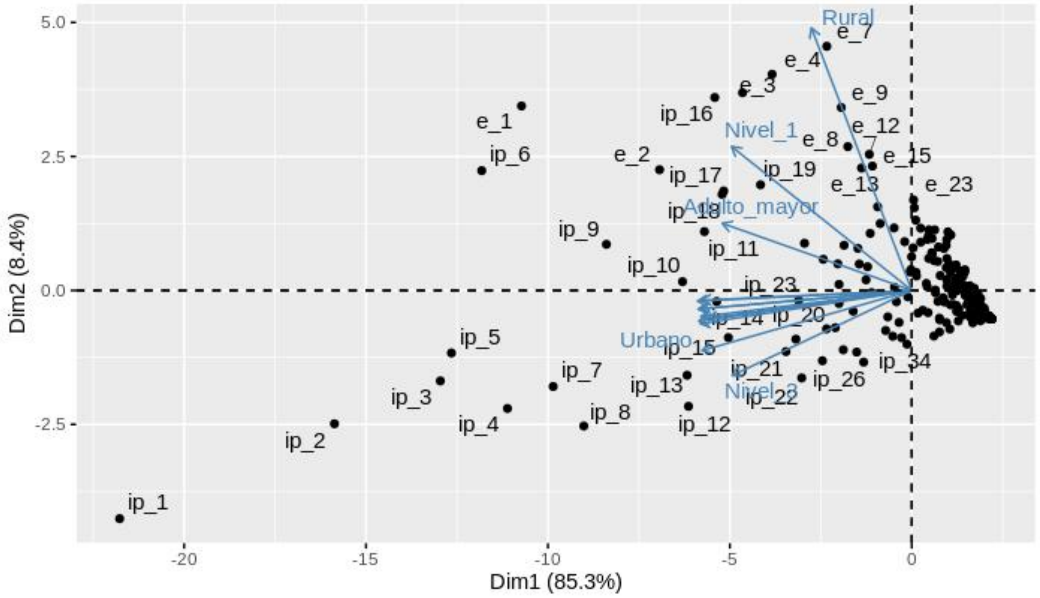


Figura 23. Distribución de prestadores y variables en las dimensiones 1 y 2

En el gráfico de la figura 24 se empleó la variable categórica ***Naturaleza_ips*** representada en color naranja para los prestadores públicos y en color azul para los prestadores privados. Esto permitió mejorar la comprensión de la distribución de los prestadores en las dos dimensiones principales y como observación interesante se encontró que los prestadores públicos presentan tendencia a cubrir usuarios de ***Nivel_1***, ubicados en la ***zona rural***.

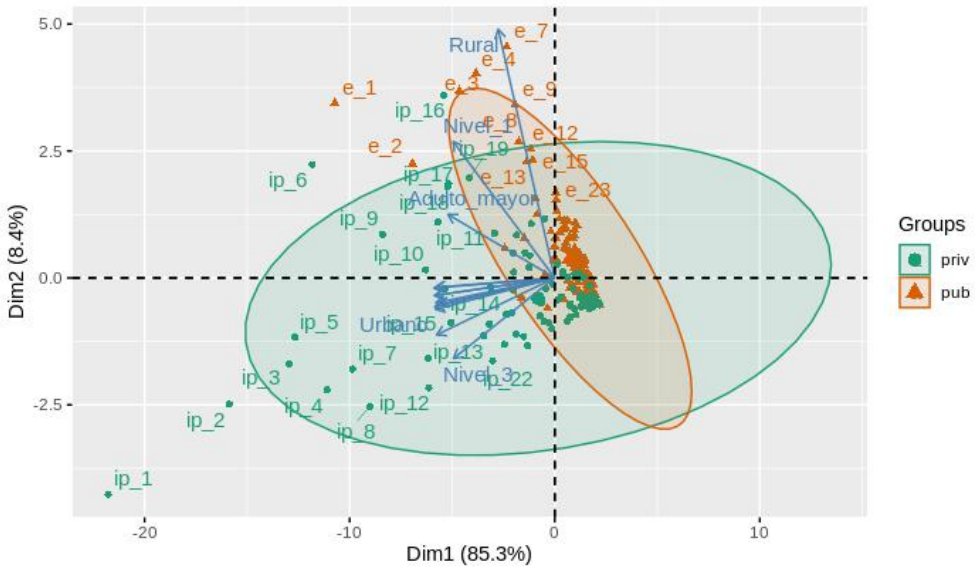


Figura 24. Distribución de prestadores públicos y privados en las dimensiones 1 y 2

El gráfico de la figura 25 presenta la distribución de los prestadores en las tres primeras dimensiones principales. Se observa una distribución en forma cónica, que constituye la extensión tridimensional de la distribución en las dos primeras dimesiones. También se observa la concentración de los prestadores pequeños en el origen de este gráfico.

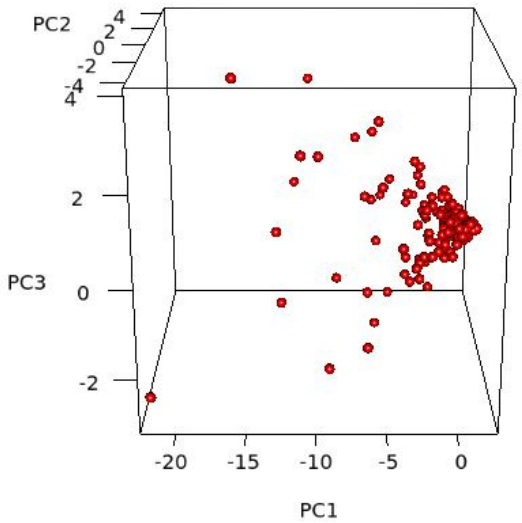


Figura 25. Representación de los prestadores en las tres primeras dimensiones.

Sumario del capítulo

Los resultados estadísticos y gráficos obtenidos en este capitulo muestran observaciones importantes que es es necesario discutir: Se observó que es notablemente mayor la cobertura de los prestadores privados frente a los prestadores públicos; y que algunos prestadores presentan de modo natural un mayor número de afiliados. Frente a las variables que caracterizan a los afiliados, el análisis y exploración estadística permitió, por un lado, identificar que ciertas variables presentan tendencia a agrupar mayor número de afiliados, es el caso de los afiliados Adultos, afiliados de niveles socioeconómicos 1 y 2, y afiliados residentes en zona Urbana. Por otro lado, la distribución del número de afiliados en las variables, refleja también la existencia de prestadores con mayor tamaño que en el marco del Aseguramiento en Salud no pueden ser considerados valores atípicos. Para efectos de reducir la dimensionalidad del dataset y mejorar la comprensión de los datos, se aplicó la técnica de Análisis de Componentes Principales cuyos resultados indicaron tendencias a agrupar los prestadores frente a ciertas variables de los afiliados. Asi pues, en las actividades y resultados obtenidos en esta etapa de comprensión, exploración y preparación de los datos, se estudiaron desde la estadística descriptiva las variables que caracterizan a los afiliados y mediante el Análisis de Componentes Principales la distribución que presentan los prestadores frente a estas variables. De este modo se da cumplimiento al primer objetivo de este trabajo; aportando los insumos para la siguiente etapa de desarrollo del modelo de analítica.

5.- DISEÑO DEL MODELO DE ANALÍTICA DE DATOS

"Si deseamos que la mayor parte de las cosas buenas de este mundo vaya a manos de alguna élite racial, el hombre nórdico o los miembros de un partido o una aristocracia, los métodos que habríamos de emplear son los mismos que asegurarían una distribución igualitaria."

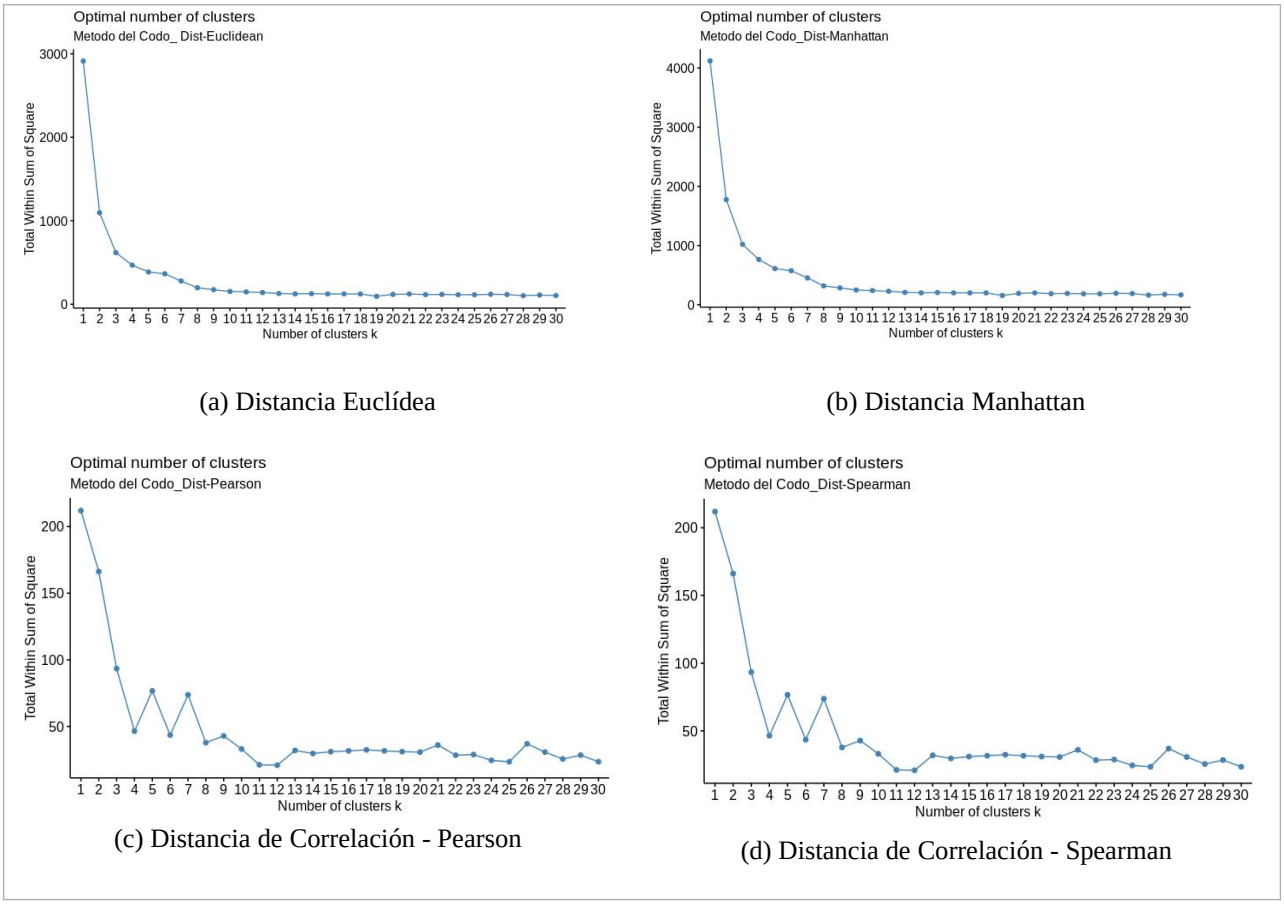
FRIEDRICH HAYEK [Camino de Servidumbre]

En este apartado se desarrolla el modelo de analítica basado en técnicas propias de la Ciencia de Datos que tiene como propósito principal mejorar la comprensión de la dinámica del Aseguramiento en Salud en el Departamento de Cundinamarca frente a los prestadores de servicios de salud habilitados. Como se mencionó en el capítulo anterior, este dataset considera, por un lado, los prestadores en relación con las principales características de los afiliados a los que prestan servicios de salud en el departamento; por otro lado, constituye el insumo para desarrollar el propósito principal del trabajo, cual es indagar por relaciones o tendencias no evidentes por métodos convencionales de analítica, de tal modo que permita agrupar o segmentar los prestadores y mejorar la toma de decisiones frente a distintos programas y políticas de salud en el departamento.

Así pues, este trabajo aborda el enfoque de Aprendizaje de Máquina No Supervisado para la implementación del modelo planteado. En este sentido, se recurrió a la implementación del modelo empleando los algoritmos K-Means, Clustering Jerárquico y K-medoids para evaluar y elegir el modelo con el mejor desempeño.

5.1.- Desarrollo del modelo basado en K-means

Con base en los conceptos del algoritmo k-means, se tomó como entrada para el modelo, las dos primeras componentes identificadas en la etapa de Análisis de Componentes Principales. Posteriormente se procedió a determinar el número óptimo de clusters para el modelo. A continuación, se presentan los resultados obtenidos al emplear varios métodos de determinación del número óptimo de clusters y variando el criterio de distancia en cada uno de ellos.



(a) Distancia Euclídea

(b) Distancia Manhattan

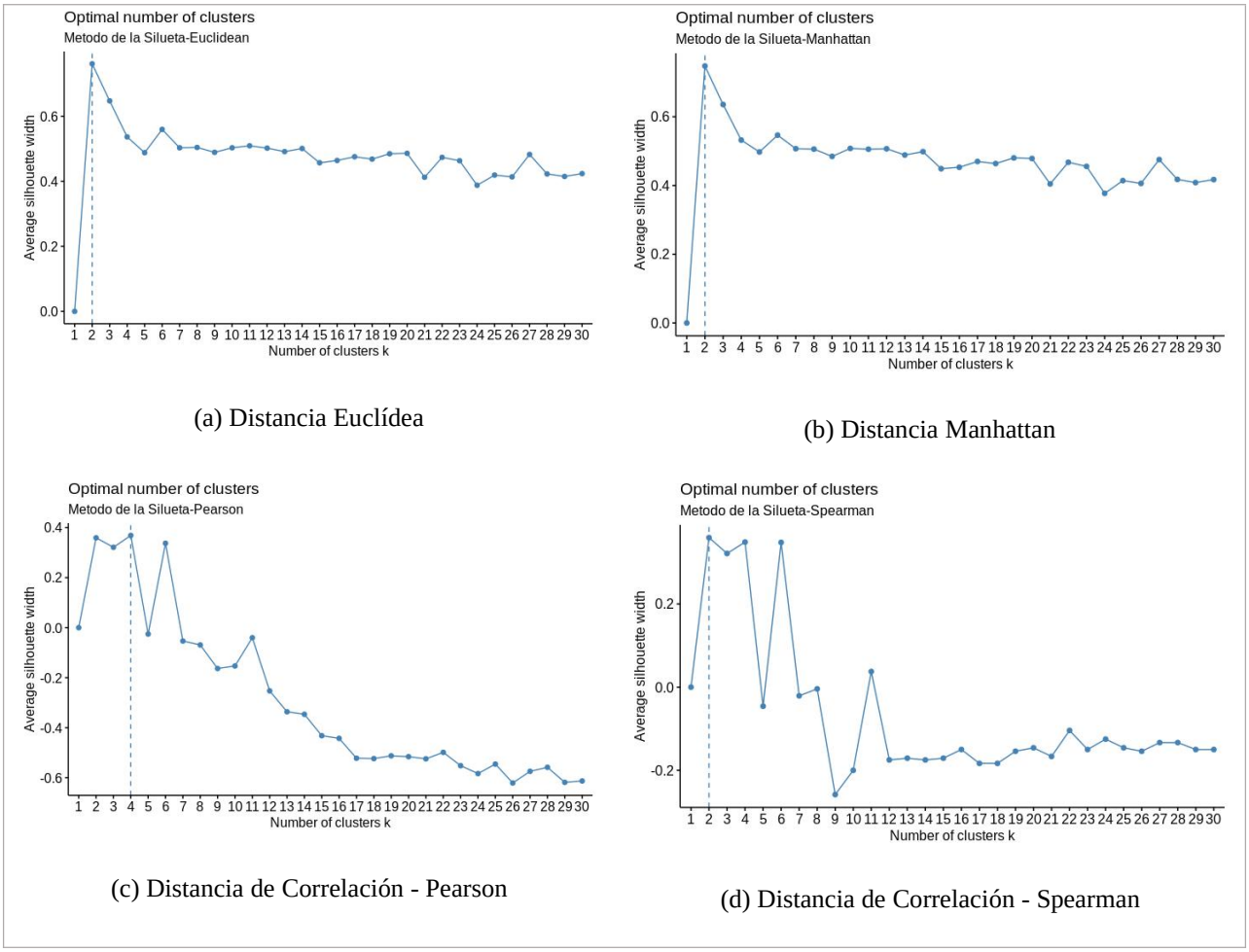
(c) Distancia de Correlación - Pearson

(d) Distancia de Correlación - Spearman

Figura 26. Número óptimo de clusters mediante el Método del Codo - Modelo K-means.

La figura 26 muestra los resultados aplicando el método del codo con el criterio Suma Total de Cuadrados dentro de los clusters, y variando el parámetro distancia (Euclídea, Manhattan, Pearson y Spearman). El resultado de este método indica que la distancia más óptima es la distancia Euclídea y Manhattan. Es razonable estimar el número óptimo alrededor de los 4 clusters.

La figura 27 presenta los resultados aplicando el método de la silueta variando el hiper-parámetro distancia (Euclídea, Manhattan, Pearson y Spearman). El resultado de este método indica que el número óptimo podría ubicarse alrededor de los 2 clusters.



(a) Distancia Euclídea

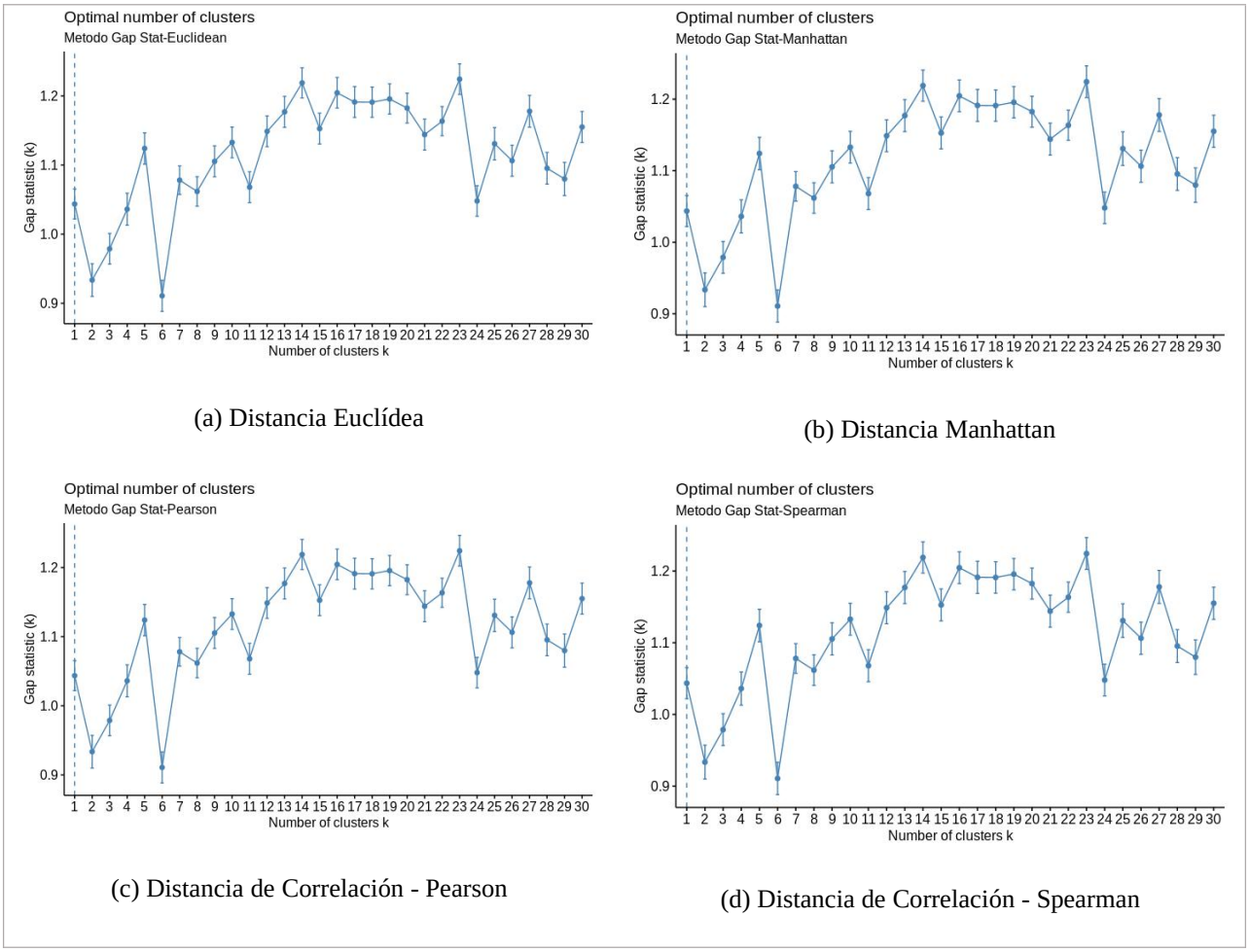
(b) Distancia Manhattan

(c) Distancia de Correlación - Pearson

(d) Distancia de Correlación - Spearman

Figura 27. Número óptimo de clusters mediante el método de la Silueta

La figura 28 presenta el resultado del método Gap Statistic variando el parámetro distancia (Euclídea, Manhattan, Pearson y Spearman). El resultado de este método indica que el número óptimo podría ubicarse alrededor de los 5 clusters.



(a) Distancia Euclídea

(b) Distancia Manhattan

(c) Distancia de Correlación - Pearson

(d) Distancia de Correlación - Spearman

Figura 28. Número óptimo de clusters mediante el método de Gap Statistic.

Los métodos para determinar el número óptimo de clusters constituyen un buen referente para tomar una decisión frente a este criterio. Sin embargo, la experiencia en el tema o área de negocio es un factor clave para decidir cual es el resultado que más se ajusta al requerimiento y contexto. En el caso de este trabajo, por un lado es necesario garantizar la representación y diversidad de los prestadores de salud, y por otro lado es necesario ser coherente con los resultados obtenidos pero considerando un número de clusters ciertamente manejable;

Así pues, se decidió considerar el resultado de 4 clusters en el método K-means con distancia Euclídea.

A continuación, se presenta de manera gráfica el resultado de los clusters obtenidos con el modelo K-means en las dos primeras dimensiones.

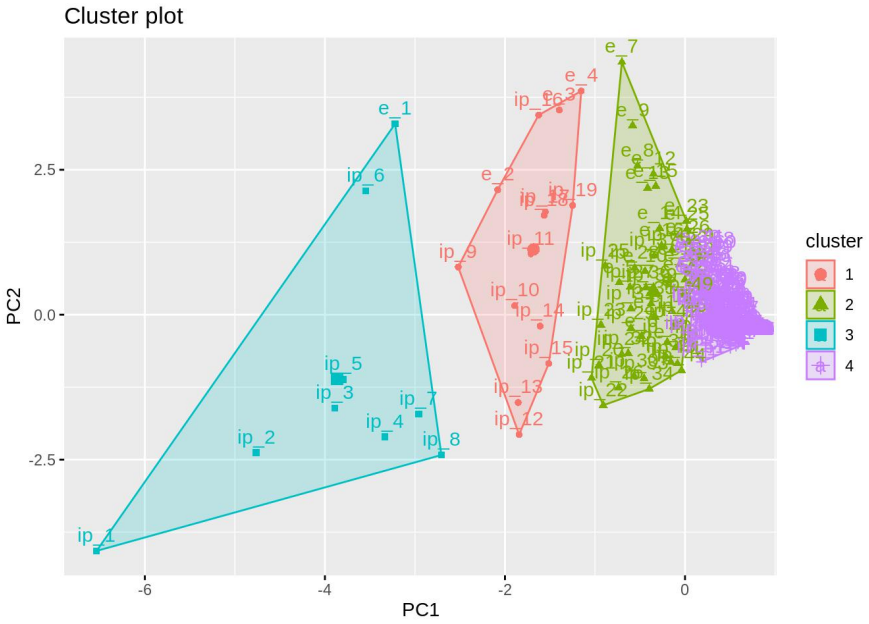


Figura 29. Representación de clusters en dos dimensiones - Clustering K-means

Las figuras geométricas trazadas delimitan los cuatro clusters en los que converge de manera estable el algoritmo K-means y la distribución de los prestadores en cada uno de ellos. Con el propósito de explorar y comprender el significado de cada cluster, a continuación, se presenta un consolido general de la composición de cada cluster en cuanto a número y naturaleza de los prestadores, y cobertura de afiliados.

Tabla 8. Distribución de los prestadores en los clusters K-means

Cluster	Prestadores Públicos	Prestadores Privados	Total Prestadores	Número de afiliados	Porcentaje de cobertura
Cluster 1	3	11	14	423.771	20,83%
Cluster 2	25	28	53	681.308	33,49%
Cluster 3	1	8	9	555.377	27,30%
Cluster 4	95	69	164	374.119	18,39%

En la tabla anterior se observa que el algoritmo creó grupos heterogéneos en cuanto al número de

prestadores; asi por ejemplo, el cluster 4 es el que incluye notablemente mayor cantidad (164 prestadores), pero al considerar la cobertura en cuanto al número de afiliados, el cluster 4 es el que presenta el menor valor (18,39 %); Esta observación es interesante porque permite inferir que existe variación intercluster y que el algortimo no se ha inclinado por la creación de clusters con igual número de prestadores en detrimento de clusters que incluyan los prestadores mas similares frente a las variables de los afiliados a quienes prestan el servicio de atención primaria. Sin embargo, estos aspectos que se abordan de manera mas formal y detallada mediante las métricas de validación en la etapa de Evaluación de los modelos.

5.2.- Desarrollo del modelo basado en Clustering Jerárquico.

Con base en los conceptos del Clustering Jerárquico, se ingresaron al modelo los dos primeros componentes principales identificados en la etapa de comprensión y preparación de los datos. Posteriormente se procedió a determinar el número óptimo de clusters para el modelo.
En el clustering jerárquico el modo visual para estimar el número de clusters es a través del gráfico de un dendrograma, de acuerdo con los siguientes criterios:

1.- Con el propósito de verificar que método de linkage y distancia son los óptimos, se calcula la correlación que existe entre una matriz de distancias de los individos del dataset versus la matriz de distancias del dendrograma. Un coeficiente cercano a 1 indica cual estructura del dendrograma es la mejor.

2.- Sobre el dendrograma encontrado en el paso anterior, se identifica el intervalo de altura donde no cambia notablemente el número de ramas. Un elemento de corte de ramas, como una linea horizontal, puede ser utilizado para determinar el número de clusters para el modelo.

Así pues, se realizaron las combinaciones de distancia y método, y se encontró que la mejor es distancia Euclídea y linkage Single.

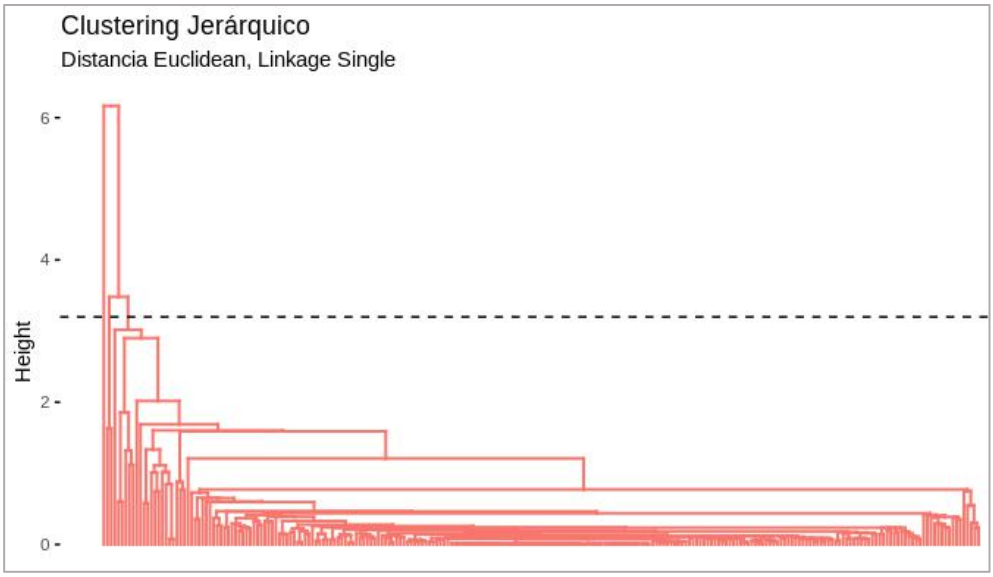
Tabla 9. Resultado de coeficiente de correlación de distancias.

	Método linkage			
Distancia	Single	Complete	Average	Ward
Euclídea	0.946	0.922	0.93	0.812
Manhattan	0.945	0.898	0.93	0.766

Con base en los resultados presentados en la tabla 9, se construyó el dendrograma con distancia

Euclídea y linkage Single.

Figura 30. Dendrograma con distancia Euclídea y linkage Simple



Fuente: Autor

En el dendrograma obtenido, visualmente es posible proponer un intervalo entre 2 y 5 para un posible número óptimo de clusters.

Con el fin de complementar y brindar más precisión a los resultados anteriores, se utilizó el paquete NbClust() disponible en el ecosistema del Lenguaje R para aproximarse al número óptimo de clusters [31]. A continuación, se presentan los resultados empleando esta prueba exhaustiva, variando la distancia (Euclídea y Manhattan) y el método de linkage.

Tabla 10. Número óptimo de clustes - Clustering Jerárquico

	Método linkage			
Distancia	Single	Complete	Average	Ward
Euclídea	2	2	2	4
Manhattan	2	3	2	4

De acuerdo con la tabla anterior, el resultado más conveniente lo aporta el método linkage Ward

que sugiere 4 clusters; La distancia Manhattan aporta un número conveniente de 3 y 4 clusters.

Como se puede observar, los resultados son coherentes con el método visual basado en el dendrograma. Pero desde un punto de vista del conocimiento del contexto o área de negocio, es necesario atender nuevamente la condición sobre la necesidad de garantizar la representación y diversidad de los prestadores al tiempo que sea coherente con los resultados y la facilidad en el manejo de la información. Así pues, para el modelo de Clustering Jerárquico se eligió el resultado de 4 clusters, empleando el método linkage Ward y la distancia Manhattan.

A continuación, se presenta el resultado del clustering en forma de dendrograma y de Árbol Filogenético.

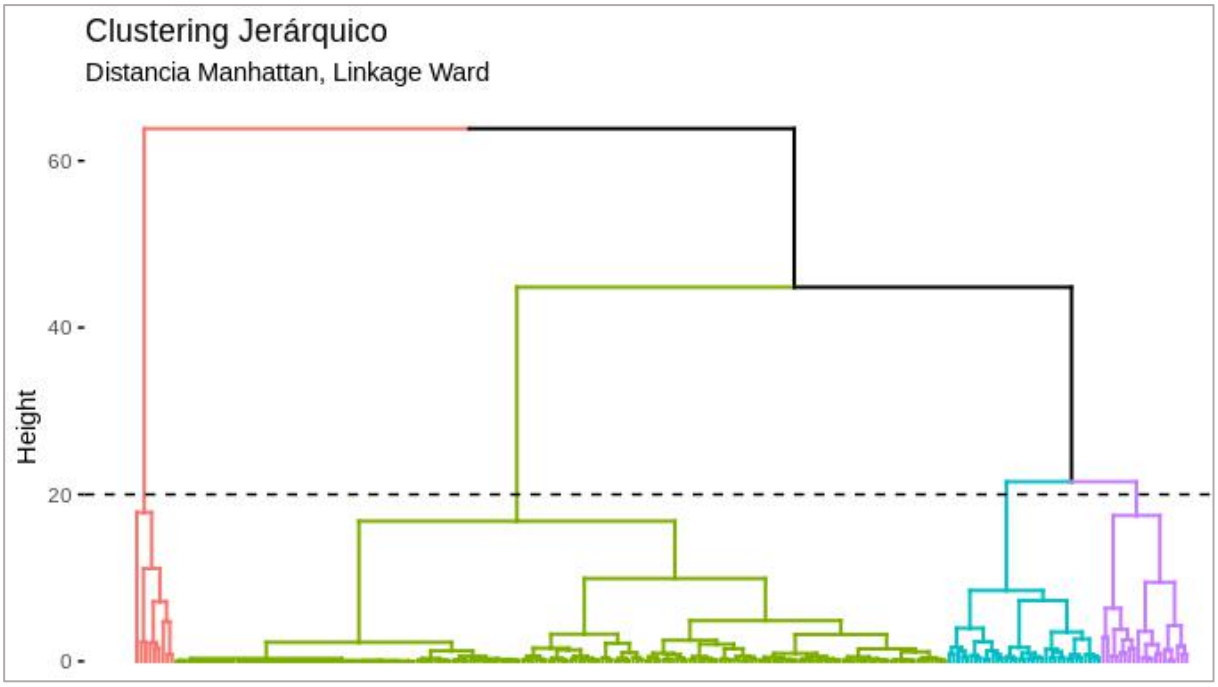


Figura 31. Representación de clusters mediante dendrograma (Manhattan,Ward)

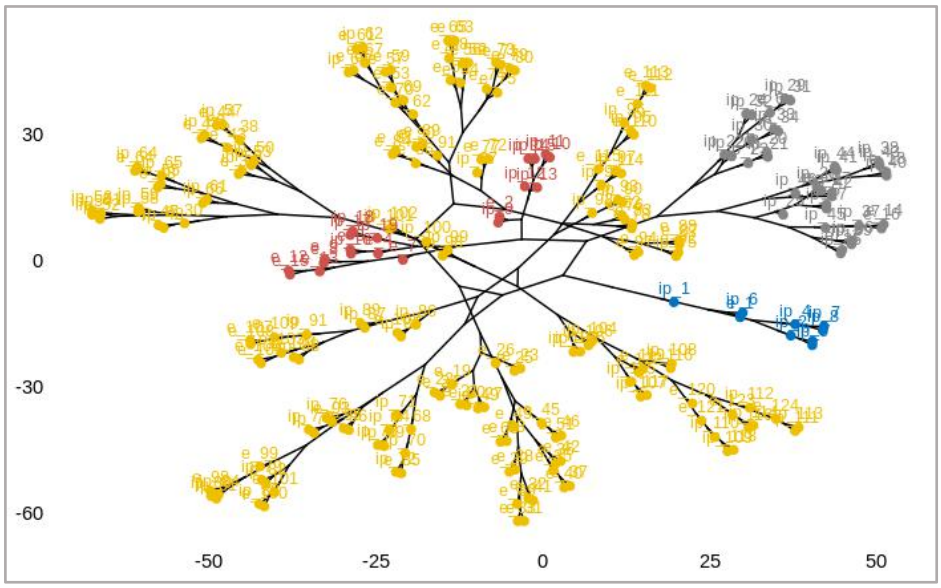


Figura 32. Representación de clusters en árbol filogenético - Clustering Jerárquico

La representación gráfica de los clusters obtenidos mediante Clustering Jerárquico en las dos dimensiones principales, es la siguiente:

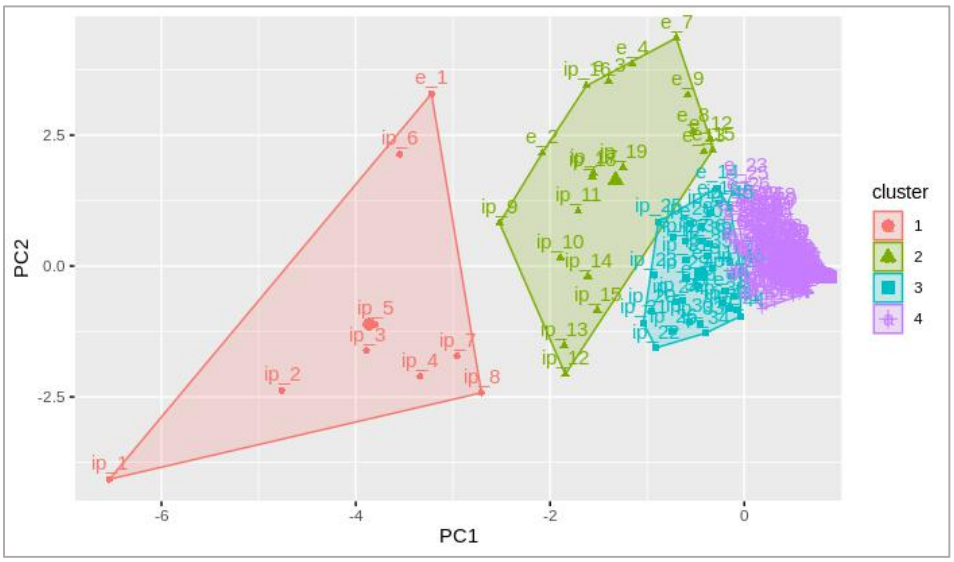


Figura 33. Representación de clusters en dos dimensiones - Clustering Jerárquico

Con el propósito de explorar y comprender el significado de cada cluster, a continuación, se presenta un consolido general de la composición de cada cluster en cuanto a número y naturaleza de los prestadores, y cobertura de afiliados.

Tabla 11. Distribución de los prestadores en los clusters del Clustering Jerárquico

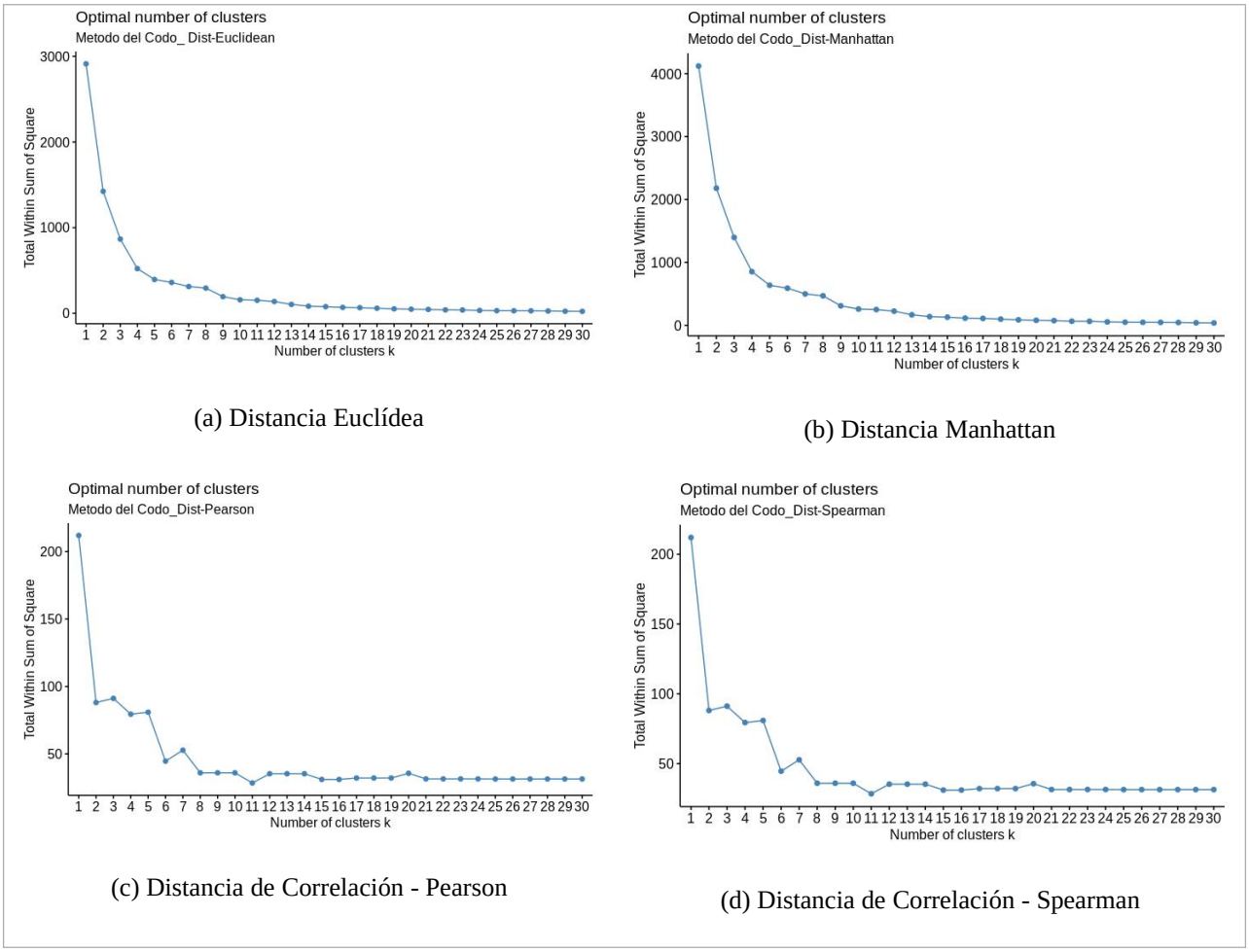
Cluster	Prestadores Públicos	Prestadores Privados	Total prestadores	Número de afiliados	Porcentaje de cobertura
Cluster 1	1	8	9	555.377	27,30%
Cluster 2	9	11	20	501.605	24,65%
Cluster 3	8	27	35	511.470	25,14%
Cluster 4	106	70	176	466.123	22,91%

En la tabla anterior se observa que el algoritmo de Clustering Jerárquico tambien creó grupos heterogéneos en cuanto al número de prestadores; Observación nuevamente interesante porque permite inferir que existe variación intercluster y que el algortimo no se ha inclinado por la creación de clusters con igual número de prestadores en detrimento de clusters que incluyan los prestadores mas similares frente a las variables de los afiliados a quienes prestan el servicio de atención primaria. Sin embargo, estos aspectos que se abordan de manera mas formal y detallada mediante las métricas de validación en la etapa de Evaluación de los modelos.

5.3.- Desarrolo del modelo basado en K-medoids

Al igual que para los modelos anteriores, para el modelo basado en K-medoids se aplicaron también los pasos generales de determinación del número óptimo de clusters, y a partir de este criterio se procedió con la ejecución del modelo .

Método del Codo: Se aplicó este método variando el parámetro distancia (Euclídea, *Manhattan*, *Pearson* y *Spearman*) , obteniendo resultados similares a los del modelo K-means. La siguiente figura consolida los gráficos obtenidos para cada tipo de distancia y es posible también proponer razonablemente un número de 4 clusers para este modelo.



(a) Distancia Euclídea

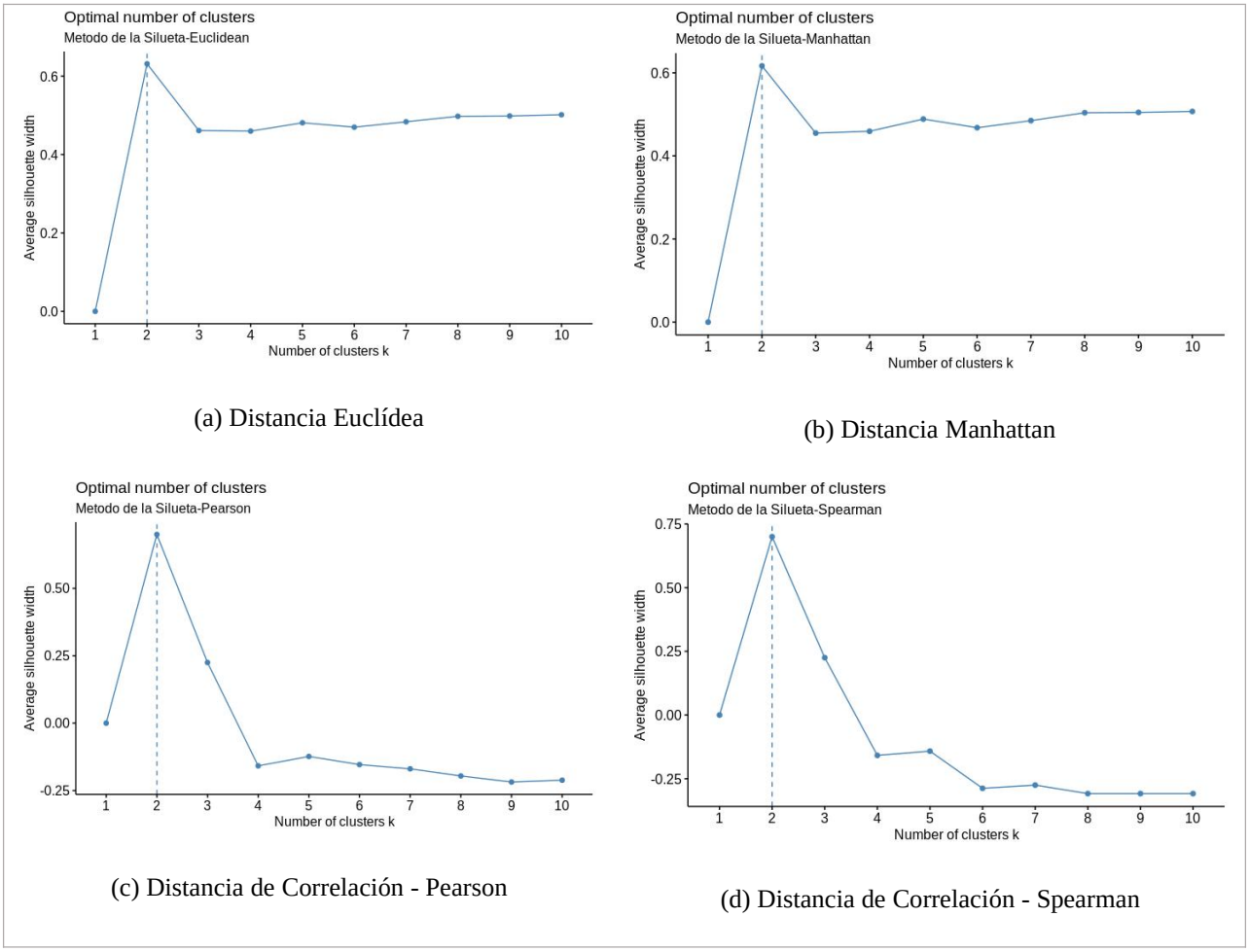
(b) Distancia Manhattan

(c) Distancia de Correlación - Pearson

(d) Distancia de Correlación - Spearman

Figura 34. Número óptimo de clusters mediante el Método del Codo - Modelo K-medoids.

Método de la silueta: Se aplicó este método variando el parámetro distancia (Euclídea, *Manhattan*, *Pearson* y *Spearman*), obteniendo resultados similares a los del modelo K-means. La siguiente figura consolida los gráficos obtenidos para cada tipo de distancia. Con este método también es razonable proponer un número de 2 clusters para el modelo basado en K-medoids.



(a) Distancia Euclídea

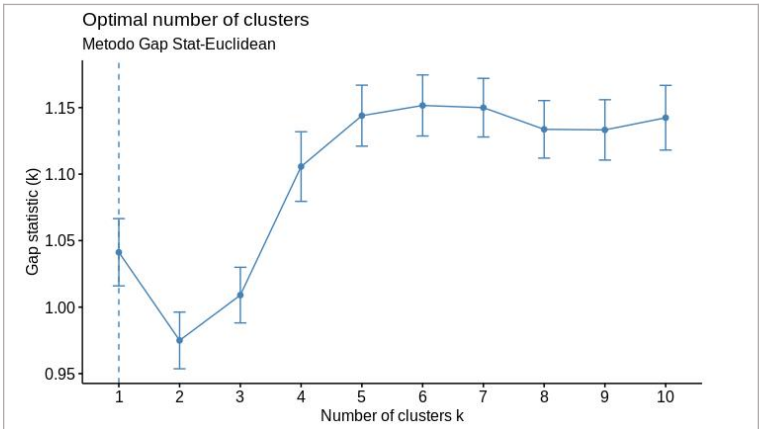
(b) Distancia Manhattan

(c) Distancia de Correlación - Pearson

(d) Distancia de Correlación - Spearman

Figura 35. Número óptimo de clusters mediante el Método de la Silueta - Modelo K-medoids.

Método Gap Statistic: Se aplicó este método variando el parámetro distancia (Euclídea, *Manhattan*, *Pearson* y *Spearman*). Dos observaciones resultan interesantes para este método: Por un lado, su ejecución requiere más recursos de máquina. Por otro lado, se obtuvo el mismo gráfico para todos los tipos de distancia. La siguiente figura presenta el gráfico obtenido para el caso de la distancia euclídea. Este método no fue claro para identificar un número razonable de clusters para el modelo basado en K-medoids.



(a) Distancia Euclídea

Figura 36. Número óptimo de clusters Método Gap Statistic - Modelo K-medoids.

De acuerdo con los resultados anteriores, sobre la determinación del número *k* óptimo de clusters para el modelo basado en K-medoids, y considerando un número apropiado acorde con el contexto de este trabajo, se optó por fijar en 4 el número de clusters.

Con base en los resultados anteriores, se procedió a ejecutar el algoritmo PAM considerando k=4 y distancia Manhattan. La figura 37 presenta el resultado gráfico de los clusters distribuidos en las dos primeras componentes principales.

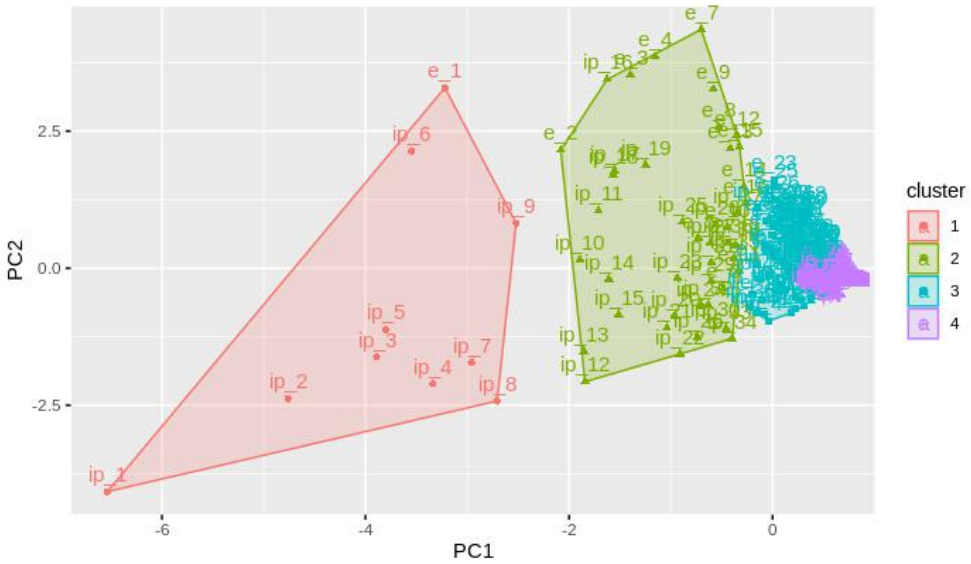


Figura 37. Representación de clusters en dos dimensiones - K-medoids

El gráfico anterior presenta en las dos primeras dimensiones o componentes principales, los cuatro cluster en los cuales converge de manera estable el algoritmo K-medoids y la distribución de los prestadores en cada uno de ellos.

Tabla 12. Distribución de los prestadores en los clusters K-medoids

Cluster	Prestadores Públicos	Prestadores Privados	Total prestadores	Número de afiliados	Porcentaje de cobertura
Cluster 1	1	9	10	595.497	29,27%
Cluster 2	15	29	44	867.613	42,64%
Cluster 3	67	28	95	513.509	25,24%
Cluster 4	41	50	91	57.956	2,85%

En la tabla anterior se observa que el algoritmo K-medoids tendió a equilibrar el cluster 3 y 4 en cuanto al número de prestadores; sin embargo, al considerar la cobertura se observa un desbalance significativo; Tambien se puede observar que el cluster 2 incrementa significativamente el número de prestadores. Observación interesante porque permite inferir que que el algortimo se ha inclinado por la creación de clusters con similar número de prestadores. Aspectos que se abordan de manera mas formal y detallada mediante las métricas de validación

en la etapa de Evaluación de los modelos.

Sumario del capítulo

En este capítulo se plantearon los algoritmos de clustering K-means, Clustering Jerárquico y K-medoids para el modelo de analítica. Los modelos K-means y Clustering Jerárquico aportaron resultados similares con grupos heterogeneos en cuanto al número de afiliados y cobertura de afiliados. Sin embargo, se pudo observar que el algoritmo K-medoids difiere de los dos anteriores y tiende a igualar el número de prestadores en los grupos; resultados y observaciones que constituyen entradas para las métricas de validación abordadas en la etapa de Evaluacion.

Así pues, de acuerdo con los modelos de clustering planteados y los resultados obtenidos se da cumplimiento al segundo objetivo específico de este trabajo. Dando paso a la siguiente etapa en la que se establecen los criterios de evaluación sobre los modelos planteados para efectos de evaluar el mejor desempeño, y determinar el modelo que permita obtener las mejores agrupaciones o segmentación de los prestadores de servicios de salud y afiliados en el departamento.

6.- EVALUACIÓN DEL MODELO

"Sin embargo, las entidades de baja entropía eran distintas: reducían su entropía e incrementaban su orden, como columnas de brillos fosforescentes alzándose sobre el mar negro azabache. Eso era significado, el significado supremo, más elevado que el placer."

CIXIN LIU [El fin de la muerte]

En este apartado se exponen los criterios de evaluación para los modelos de clustering planteados en la etapa anterior; En términos generales podemos identificar los siguientes criterios de validación.

Validación interna de los clusters: En la que se mide que tan homogéneos son los indivíduos de un cluster (distancia intra-cluster) y que tan separados están los clusters (distancia inter-clusters).
Para validar este criterio se recurrió a las métricas: Ancho promedio de la Silueta, índice de Dunn, índice de Davis-Bouldin, Conectividad y Estabilidad.

Validación Externa de los clusters: En este tipo de validación se comparan los resultados de un modelo de clustering con los resultados de un modelo conocido en un área de negocio relacionada o similar. Es decir, se cuenta con el resultado de un modelo de clustering que ha agrupado y etiquetado cada indivíduo de un dataset similar.
Sin embargo, en este trabajo no se cuenta con casos previos de clustering que guarden estrecha relación o semejanza con el área de negocio que corresponde a los prestadores de servicios de salud que brinda cobertura a afiliados que demandan estos servicios en un determinado territorio.

Validación Relativa: En este enfoque de validación se evalúa la estructura de los clusters variando diferentes parámetros del algoritmo en el que se basa el modelo de clustering. Así pues, el proceso de obtener un número óptimo de clusters variando parámetros como k, distancia,

método linkage, entra en la tipificación de validación relativa. En este trabajo se empleó este enfoque para identificar el número óptimo de clusters con técnicas como el método del codo, la silueta y análisis de brechas (Gap Stastic).

6.1.- Evaluación de los modelos

6.1.1.- Ancho promedio de la Silueta

Como se indicó anteriormente en el método de la Silueta se calcula el coeficiente $s(i)$ que permite verificar que tan bién fue asignado un indivíduo a su cluster a partir de la comparación de la distancia de los indivíduos que conforman su propio cluster y con los indivíduos de otros clusters. Por otra parte, el concepto del ancho promedio de la silueta, una vez se ha ingresado al modelo el número óptimo de clusters, indica si los clusters están bién conformados e incluyen de modo óptimo a los indivíduos. Esto implica que un valor alto del ancho de la silueta sea un indicador de que los clusters están bien conformados.

A continuación, se presentan los valores y respectivas gráficas aplicando el método del Ancho promedio de la Silueta para los modelos planteados.

Modelo basado en K-means:

Tabla 13. Resultado Ancho de la Silueta - Kmeans

K-means (*distancia euclídea*)		
Cluster	Tamaño	Ancho promedio de la Silueta por cluster
1	14	0.42
2	53	0.26
3	9	0.28
4	164	0.69
		Ancho promedio general: 0.57

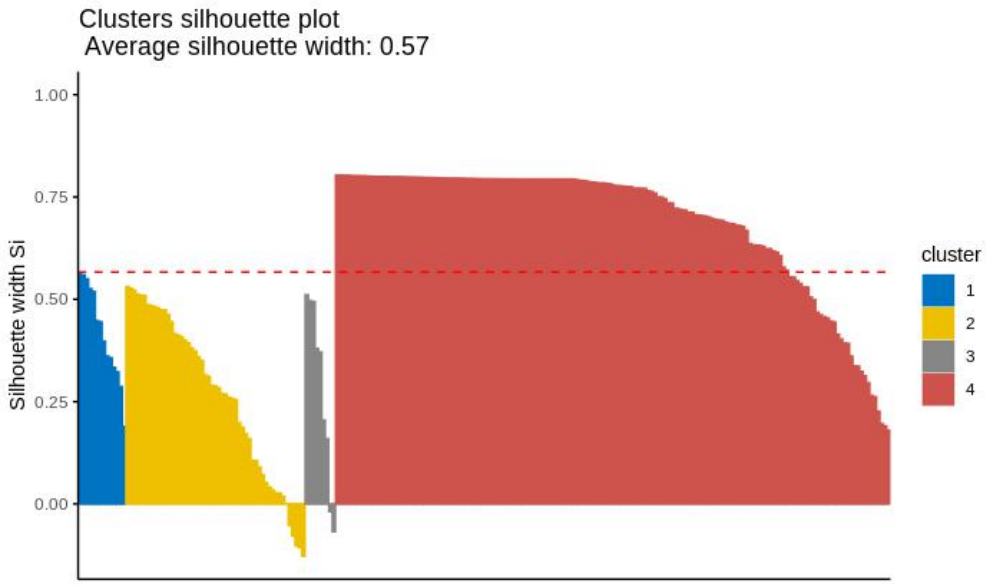


Figura 38. Ancho promedio de la silueta - Modelo K-means

Modelo basado en Clustering Jerárquico:

Tabla 14. Resultado Ancho de la Silueta - Clustering Jerárquico

Clustering Jerárquico (*distancia Manhattan, linkage Ward*)		
Cluster	Tamaño	Ancho promedio de la silueta por cluster
1	9	0.40
2	20	0.12
3	35	0.40
4	176	0.63
		Ancho promedio general: 0.54

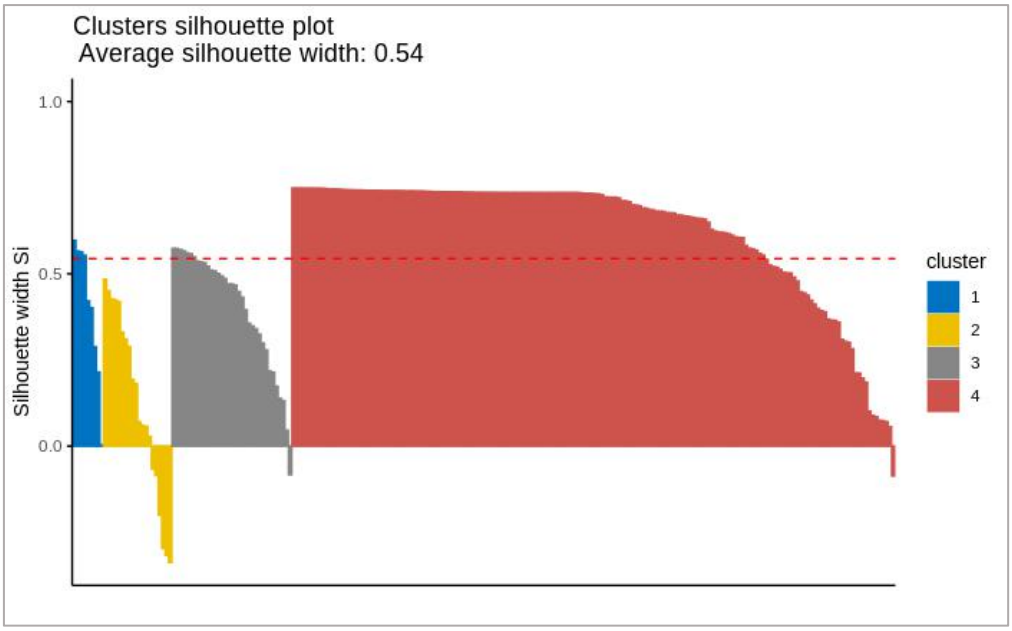


Figura 39. Ancho promedio de la silueta - Modelo Clustering Jerárquico

Modelo basado en K-medoids:

Tabla 15. Resultado Ancho de la Silueta - K-medoids

Clustering Jerárquico (*distancia Manhattan, linkage Ward*)		
Cluster	Tamaño	Ancho promedio de la silueta por cluster
1	10	0.42
2	44	0.17
3	95	0.27
4	91	0.81
		Ancho promedio general: 0.46

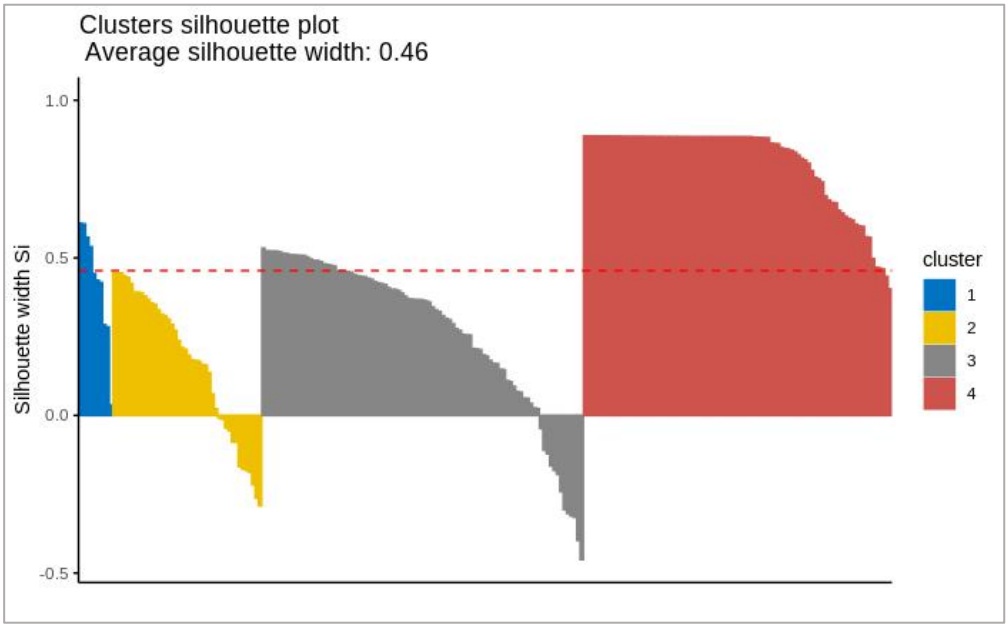


Figura 40. Ancho promedio de la silueta - Modelo K-medoids

De acuerdo con este método, los clusters del modelo basado en K-means presentan mejores resultados tanto en el promedio general del Ancho de la Silueta como en el promedio de cada cluster; Por un lado, el promedio general del ancho de la silueta indica que K-means presenta mejor desempeño, por otro lado la gráfica permite verificar que los clusters de K-means presentan menor número de prestadores con valores negativos para el ancho de la silueta, lo que indica que los clusters incluyen correctamente a los prestadores con características similares.

6.1.2.- Índice de Dunn

Este método también brinda una métrica para la validación interna de los clusters (Dunn, 1973); En este método se realiza en dos pasos generales: En el primer paso se elige la distancia mínima que se obtiene de la comparación entre las distancias de los indivíduos de un cluster versus los indivíduos de los demás clusters. Esta es la mínima distancia intercluster.

En el segundo paso se elige la máxima distancia que se obtiene al comparar las distancias de los indivíduos a su respectivo cluster. Esta es la máxima distancia intracluster.

La expresión que describe el método del índice Dunn es la siguiente:

$$D = \frac{mínima\ distancia\ intercluster}{máxima\ distancia\ intracluster} \qquad (20)$$

A partir de la expresión anterior, se puede observar que si los clusters están bién conformados entonces el índice D será alto, en caso contrario el índice D será bajo.

La tabla que se presenta a continuación, muestra los resultados del cálculo del índice Dunn para los dos modelos.

Tabla 16. Resultado índice Dunn.

Modelo	Índice Dunn
Kmeans	0.013
Clustering Jerárquico	0.028
K-medoids	0.0058

De acuerdo con este índice, el mejor desempeño corresponde al modelo basado en Clustering Jerárquico.

6.1.3.- Índice de Davies- Bouldin (DB)

El propósito de esta métrica de evaluación es medir el tamaño o dispersión de los clusters versus la distancia promedio entre los clusters. Un índice pequeño indica clusters mejor estructurados.
La expresión que describe el índice de Davies-Bouldin es la siguiente:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} \left(\frac{s_i + s_j}{d_{ij}} \right) \qquad (21)$$

donde k es el número de clusters, s_i es la distancia promedio entre cada punto del cluster C_i al centro del propio cluster C_i , s_j es la distancia promedio entre cada punto del cluster C_j al centro del propio cluster C_j, y d_{ij} es la distancia entre los centros de los clusters C_i y C_j.

Así pues, se realizó la medición del índice de Davies-Bouldin para los modelos y a continuación, se presentan los resultados:

Tabla 17. Resultado índice de Davies-Bouldin

Modelo	Índice Davies-Bouldin
Kmeans	0.851
Clustering Jerárquico	0.958
K-medoids	0.975

De acuerdo con este índice, el mejor rendimiento corresponde al modelo basado en K-means.

6.1.4. Validación de Conectividad

El concepto de Conectividad en Clustering, indica el grado o nivel de conexión que existe al interior de los clusters empleando el criterio de medición de los k-vecinos más cercanos. El índice de Conectividad adquiere un rango entre 0 e infinito, siendo las medidas óptimas aquellas cercanas a 0.

Se aplicó este concepto sobre los modelos planteados, y a continuación se presentan los resultados obtenidos:

Tabla 18. Resultado índice de Conectividad

Modelo	Índice Conectividad
Kmeans	26.678
Clustering Jerárquico	23.864
K-medoids	28.369

De acuerdo con este índice, el mejor desempeño corresponde al modelo basado en Clustering Jerárquico.

6.1.5.- Validación de Estabilidad

Las métricas de evaluación de la estabilidad de los clusters aportan un criterio valioso a la hora de comprobar si los clusters obtenidos se mantienen sin mayor variación frente al clustering obtenido en un proceso iterativo de eliminar cada columna o variable del conjunto de datos.

Estas métricas son las siguientes [32]:

Proporción Promedio de No traslapamiento (Average Proportion of Non-overlap - APN -): Indica la proporción promedio de indivíduos que no se asignan al mismo cluster en el proceso iterativo de eliminación de una columna frente al clustering completo que incluye todas las columnas.

Distancia Promedio (Average Distance -AD-): Indica el promedio de las distancias intra-cluster entre los individuos asignados al mismo cluster, obtenidas en el proceso iterativo de eliminación de una columna frente al clustering completo que incluye todas las columnas.

Distancia Promedio entre los centroides (Average Distance between Means - ADM-): Indica el promedio de las distancias entre los centroides y los individuos asignados al respectivo cluster; obtenidos en el proceso iterativo de eliminación de una columna frente al clustering completo que incluye todas las columnas.

Figura de Merito (Figure of Merit - FOM): Mide la varianza intra-cluster promedio de los individuos u observaciones en la columna eliminada, en el escenario de clustering del proceso iterativo de eliminación de una columna.

La medición de APN, ADM y FOM toma un rango de valores entre 0 y 1 , siendo los valores cercanos a 0 aquellos que indican un Clustering estable y consistente. La métrica AD toma valores entre 0 e infinito, siendo los valores óptimos aquellos cercanos a 0.

A continuación. se presentan los resultados obtenidos para las métricas de estabilidad:

Tabla 19. Resultado métricas de Estabilidad

Modelo	APN	AD	ADM	FOM
Kmeans	0.004	1.795	0.042	0.465
Clustering Jerárquico	0.111	4.915	0.384	0.534
k-medoids	0.043	4.602	0.171	0.501

De acuerdo con estos resultados, el modelo basado en K-means presenta el mejor desempeño frente a las 4 métricas de estabilidad. Frente a la métrica APN indica que K-means generó clusters que incluyen prestadores similares, aún en condiciones de variación del conjunto de datos (eliminación iterativa de una columna). El resultado obtenido en la métrica AD indica que K-means generó clusters cuyos individuos son cercanos entre si, es decir son similares, aún en el caso de variación del conjunto de datos (eliminación iterativa de una columna). Por su parte, el resultado de la métrica ADM indica que k-means generó clusters compactos cuyos individuos son cercanos al respectivo centroide, aún en condiciones de variación del conjunto de datos (eliminación iterativa de una columna). Finalmente, la métrica FOM indica que K-.means generó clusters con la menor varianza al eliminar iterarativamente una columna o variable en el conjunto de datos.

6.1.6.- Elección del modelo

En este punto es necesario determinar el modelo de clustering con el mejor de desempeño a partir de los resultados obtenidos con las distintas métricas empleadas. Para ello la siguiente tabla consolida los valores de estas mediciones y permite observar que el modelo con el mejor desempeño es el basado en el algoritmo K-means.

Tabla 20. Consolidado de las métricas de validación

Modelo	Ancho Sil.	Dunn	DB	Conect.	APN	AD	ADM	FOM
Kmeans	0.57	0.013	0.851	26.678	0.004	1.795	0.042	0.465
Clustering Jerárquico	0.54	0.028	0.958	23.864	0.111	4.915	0.384	0.534
k-medoids	0.46	0.0058	0.975	28.369	0.043	4.602	0.171	0.501

Atendiendo los principios de separación (máxima distancia intercluster) y cohesión (mínima distancia intracluster) que deben considerarse en el clustering, la tabla anterior permite verificar que el algortimo K-means presentó mejor desempeño en las métricas que permiten validar la separación intercluster: Ancho promedio de la silueta, el índice de Davies-Bouldin y APN. Por otro lado, (a excepción del índice Dunn y Conectividad) K-means presentó mejor desempeño en las métricas que permiten validar la cohesión intracluster: AD, ADM. En cuanto a la métrica FOM para verificar la menor varianza, K-means también presentó mejor desempeño.

6.2.- Análisis de los clusters

En este punto se consigna el análisis de los clusters generados con el modelo basado en K-means. Se verificó como se distribuye la cobertura o peso de los 4 clusters frente a las características de los afiliados a los distintos prestadores de servicios de salud en el departamento de Cundinamarca. Esto con el fin de buscar relaciones o tendencias que permitan optimizar y enfocar las metas y acciones para mejorar la cobertura y acceso a los servicios de salud. Lo cual esta en línea con el objetivo principal planteado en este trabajo.

Distribución de los prestadores en las componentes principales: La figura 41 muestra la distribución de los prestadores en las dos y tres primeras componentes principales permitiendo distinguir que los 4 clusters están bien delimitados y son coherentes con la distribución que fue encontrada en el Análisis de Componentes Principales (PCA).

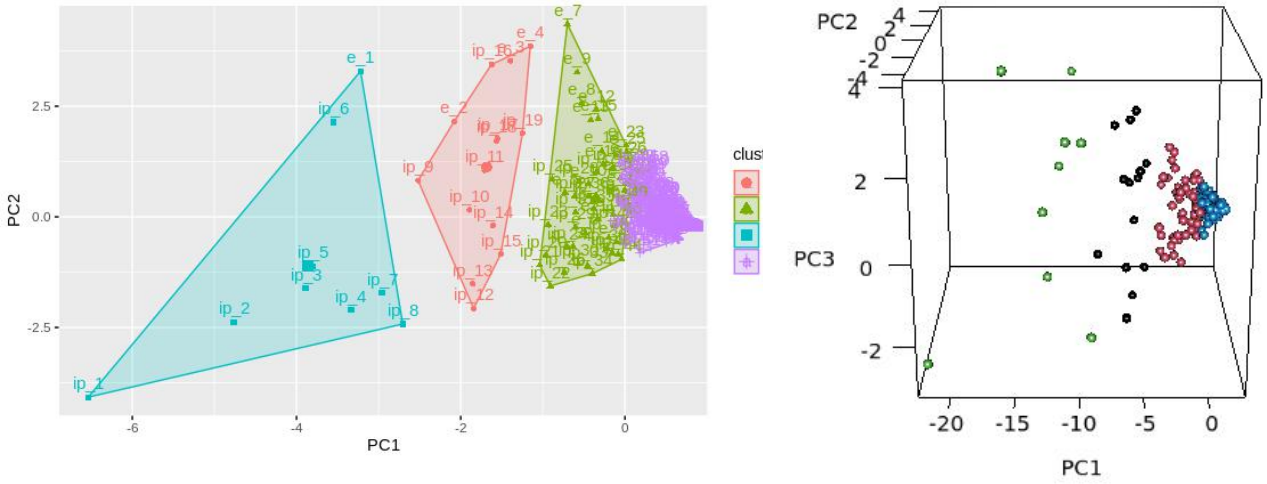


Figura 41. Clusters y prestadores en dos y tres primeras componentes principales.

Clusters y total de afiliados: Es importante abordar este aspecto toda vez que permitió analizar el tamaño de los clusters en número de afiliados que incluye cada una de estas agrupaciones. La figura 42 muestra que los clustes son equilibrados; no existe una marcada dominancia de un cluster sobre los otros. Esto es favorable porque indica que el modelo no presentó un favorecimiento hacia los prestadores más grandes en cuanto al número afiliados que atienden. Este hecho es relevante porque todos los afiliados presentan igual derecho al acceso a los servicios de salud, y el modelo no favorece a los prestadores más grandes con más recursos en detrimento de los más pequeños.

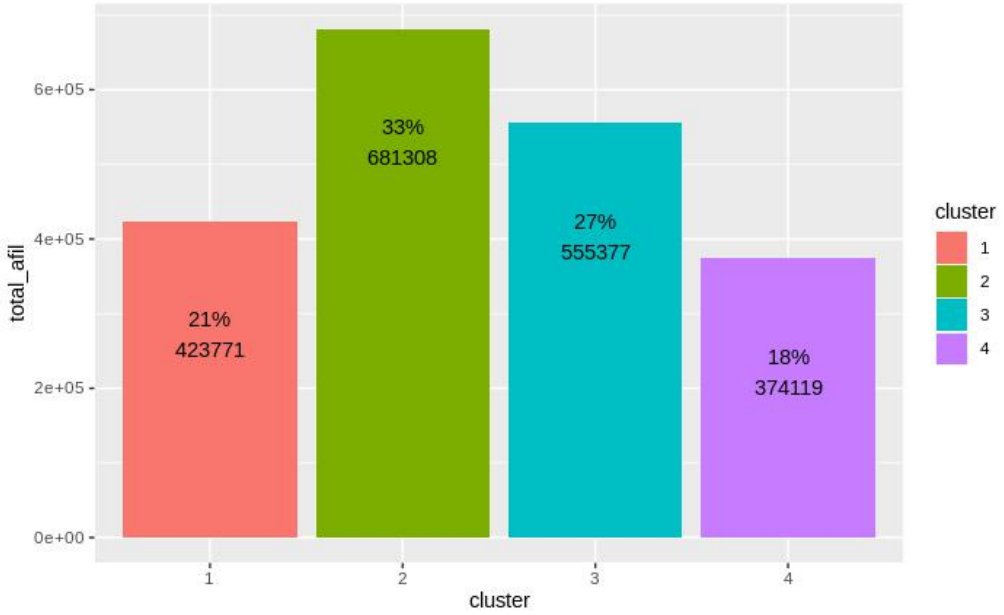


Figura 42. Tamaño de los clusters en número de afiliados

Clusters y naturaleza de los prestadores: Este aspecto también es relevante pues, es necesario conocer la distribución de los prestadores públicos y privados en los clusters dado que los prestadores de naturaleza pública constituyen la Red de Prestadores del Departamento y son objeto directo de políticas de cobertura y acceso a los servicios de salud para los afiliados. Sin desconocer que los prestadores de naturaleza privada también son articulados en las políticas y programas de la entidad.

La figura 43 indica que los clusters frente a este criterio presentan una tendencia espontánea e interesante porque esta variable categórica no fue incluida dentro de las variables para el modelo: Los clusters 1 y 3 presentan una marcada mayoría de prestadores privados; mientras que los clusters 2 y 4 incluyen un notable porcentaje de prestadores públicos. Esto claramente facilita el enfoque en las políticas y programas de salud dirigidos principalmente a los prestadores de la Red de Prestadores del departamento.

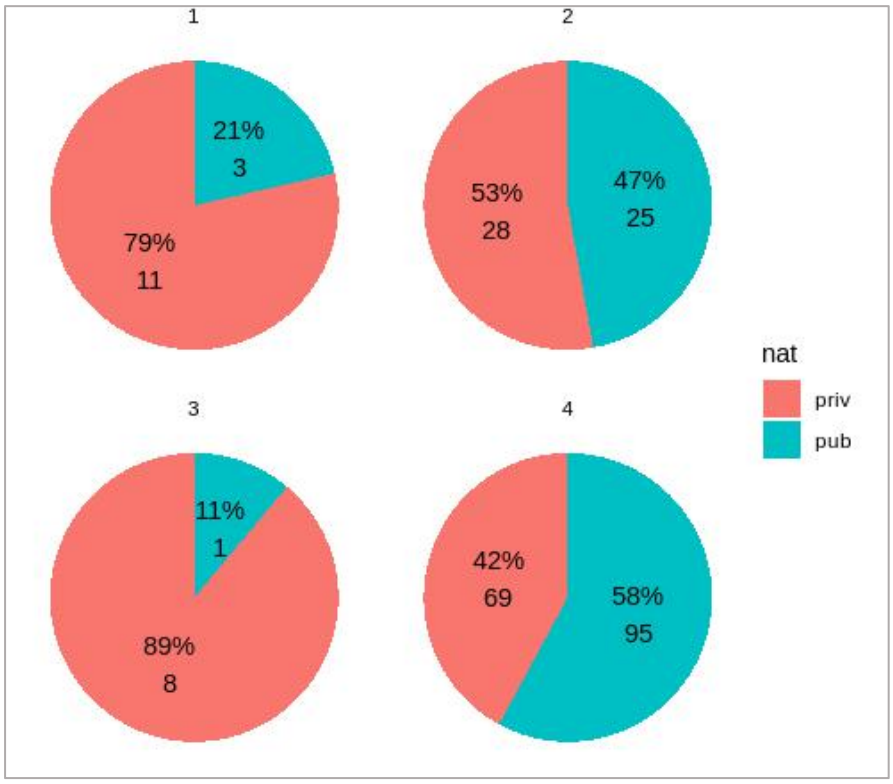


Figura 43. Clusters y naturaleza de los prestadores

Distribución de los clusters en las distintas variables de afiliados: Este aspecto del análisis es el más interesante porque permitió revisar en detalle el modo en que los clusters responden a las variables que caracterizan a los afiliados. Permitiendo refinar aun más el enfoque de programas específicos dirigidos a ciertos sectores poblacionales de los afiliados, por ejemplo: Programas o políticas de salud con enfoque de género o edad, ruralidad y difícil acceso, sectores socioeconómicos en condición de pobreza o vulnerabilidad, entre otros casos relevantes. En este punto es importante destacar que las distribuciones o porcentajes específicos de los clusters en las variables constituyen un buen referente a la hora de elaborar o proyectar los recursos presupuestales, logísticos y administrativos.

Variables de Grupo Etario: Se observa que los clusters tienen un peso porcentual frente a estas variables que es acorde con la distribución general para el total de afiliados (figura 44). Sin embargo, como anotación interesante, se observa que en la variable Adulto Mayor, los clusters 2 y 4 cobran mayor relevancia. Esto resulta interesante porque en los clusters 2 y 4 tienen notable

representatividad los prestadores públicos.

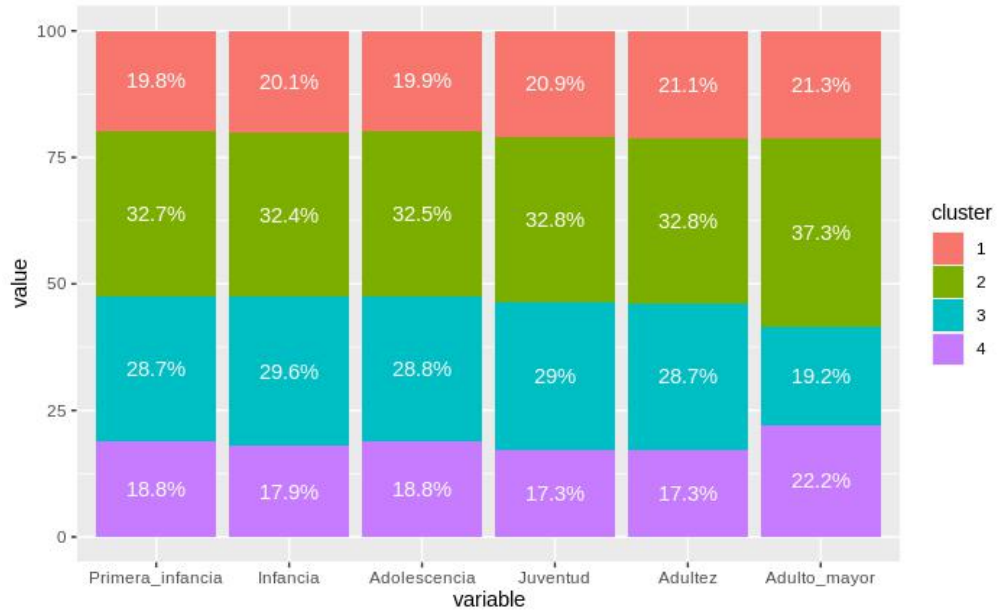


Figura 44. Distribución de los clusters en las variables de Grupo Etario

Variables de Género: Frente a estas variables los clusters conservan el peso porcentual acorde con la distribución general para el total de afiliados que se muestra en la figura 45. Este comportamiento es el esperado e indica la buena conformación de los clusters porque en las variables de género la población de hombres es similar a la de mujeres, independientemente del territorio o sociedad.

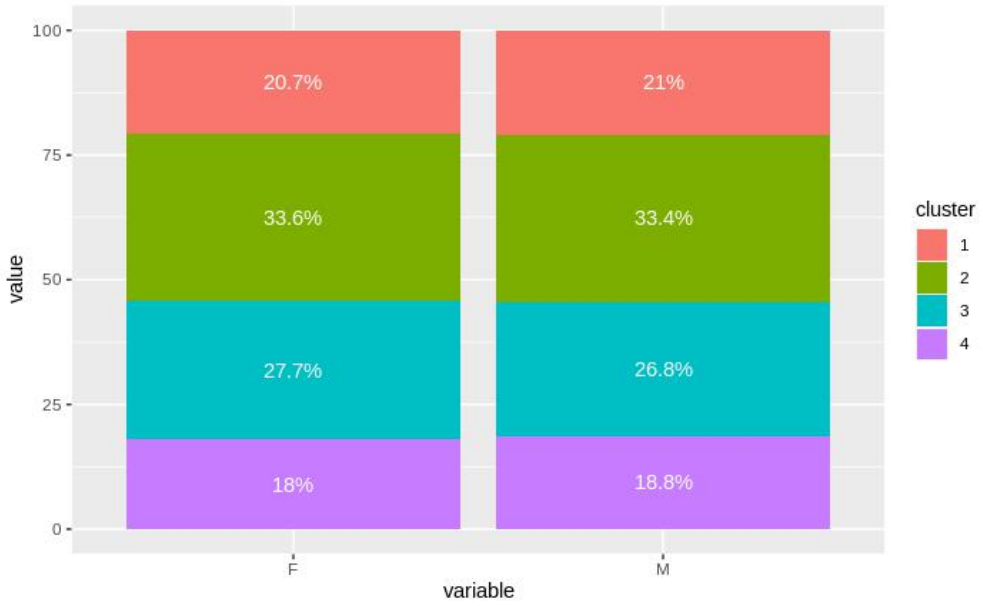


Figura 45. Distribución de los clusters en las variables de Género.

Variables de Nivel Socioeconómico: Los clusters presentan un comportamiento que en términos generales es coherente con la distribución general para el total de afiliados que se muestra en la figura 46. Sin embargo, en este punto también surge una observación interesante relacionada con el hecho de que para el Nivel 1 cobran relevancia los prestadores agrupados en los clusters 4 y 2 que tienen una notable presencia de prestadores públicos. Se observa también que para el Nivel 3 el cluster 4 disminuye su peso o influencia.

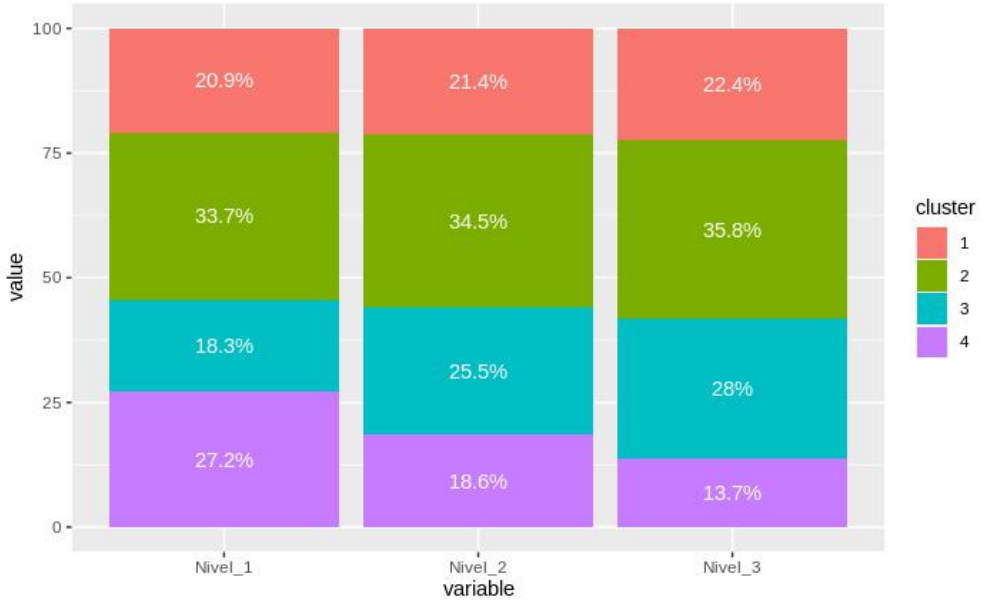


Figura 46. Distribución de los clusters en las Variables de Nivel Socioenómico.

Variables de Zona: Frente a estas dos variables se obtiene una observación interesante relacionada con el hecho de que para la zona Rural cobran relevancia los clusters 2 y 4 que tienen gran presencia de prestadores de naturaleza pública. Este hecho en articulación con las variables de Nivel Socioeconómico cobra relevancia porque indica que los prestadores públicos presentan una tendencia a la cobertura en la zonas rurales o alejadas de los mayores centros poblados del departamento, y cubren población de escasos recursos o en condición de vulnerabilidad. Por otra parte se observa que para la población con Nivel 3, que no es pobre ni vulnerable, cobra relevancia el cluster 3 que esta principalmente compuesto por prestadores privados.

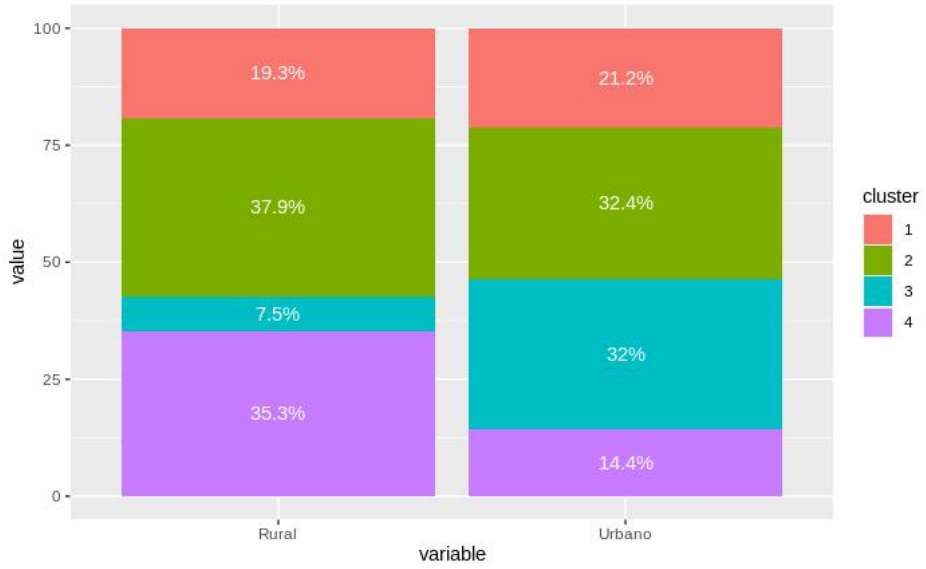


Figura 47. Distribución de los clusters en las Variables de Zona.

Sumario del capítulo

En este capítulo se revisaron y aplicaron las métricas de evaluación para los modelos de clustering, y los resultados permitieron elegir el modelo con mejor desempeño que correspondió a K-means. La ejecución de este modelo permitió la generación de 4 clusters de prestadores de servicios de salud en el departamento de Cundinamarca que fueron analizados frente a las diferentes variables o características de los afiliados. De este modo se da cumplimiento al tercer objetivo específico de este trabajo.

7.- DESPLIEGUE DEL MODELO

"Finalmente, quiero, Sancho, me digas lo que acerca de esto ha llegado a tus oídos, y esto me has de decir sin añadir al bien ni quitar al mal cosa alguna, que de los vasallos leales es decir la verdad a sus señores en su ser y figura propia, sin que la adulación la acreciente o otro vano respeto la disminuya; y quiero que sepas, Sancho, que si a los oídos de los príncipes llegase la verdad desnuda, sin los vestidos de la lisonja, otros siglos correrían"

MIGUEL DE CERVANTES [Don Quijote de la Mancha]

Para el despliegue del modelo se construyó una visualización que consolida los principales resultados del trabajo. El Diagrama de Despliegue que se presenta a continuación, muestra los componentes, nodos y artefactos que integran la solución técnica que se construyó en este trabajo.

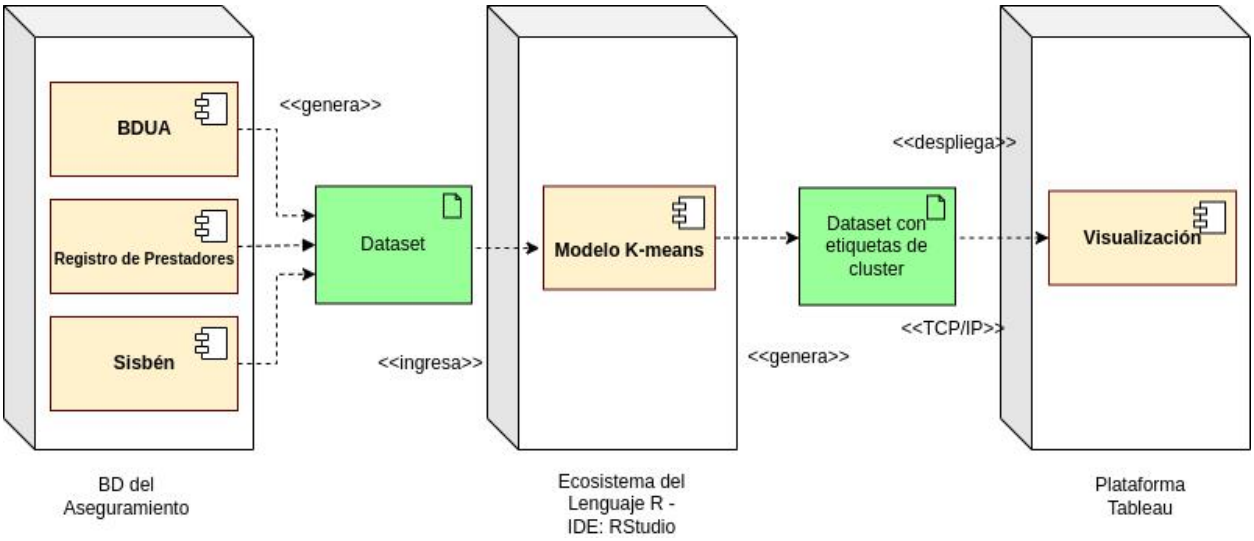


Figura 48. Diagrama de Despliegue

La Base de Datos del Aseguramiento es el componente que incluye las tablas que mediante consultas SQL generan el Dataset en formato CSV (Valores Separados por Coma) para el Modelo, que en este trabajo está basado en el algoritmo K-means. El algoritmo se desarrolló en el lenguaje R [33] empleando el IDE R Studio como herramienta de desarrollo [34]. La ejecución del modelo genera un archivo CSV con las etiquetas de cluster asiganado a cada prestador. Este archivo CSV fue desplegado en la plataforma para visualizaciones Tableau [35].

El código del modelo puede ser consultado en el siguiente enlace:

https://rpubs.com/djdorado/modelo_PrestadoresAfil_Cund

7.1.- Visualización

La visualización fue diseñada y publicada en la plataforma Tablaeu Public, con un cariz de narrativa (*storytelling*) mediante preguntas orientadoras. Considerando que una visualización debe ser clara y concisa en la comunicación de la idea principal, se construyó en una vista unificada que mediante distintas gráficas y consultas interactivas aporta información sobre el propósito general del trabajo y la aplicabilidad de los resultados obtenidos con el modelo de analítica planteado.
La visualización puede ser consultada en el siguiente enlace:

https://public.tableau.com/app/profile/jesus.dorado/viz/Viz_20230621/Intro

y en ella se pueden distinguir las siguientes secciones:

1.- Introducción a modo de narrativa y estadística general de los prestadores públicos y privados que operan en el departamento.

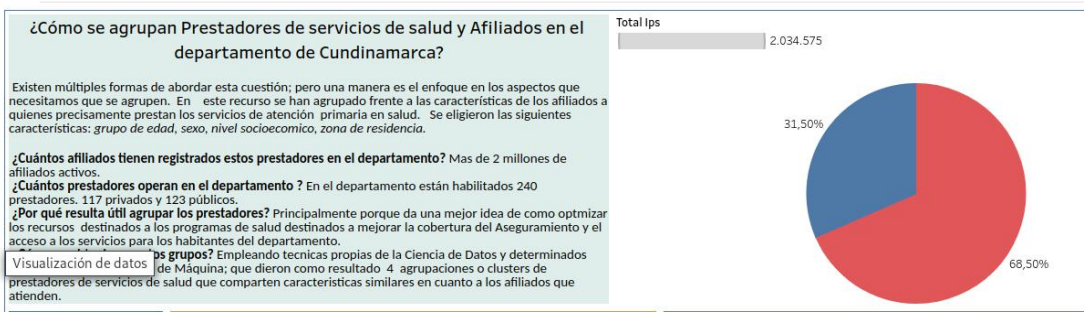
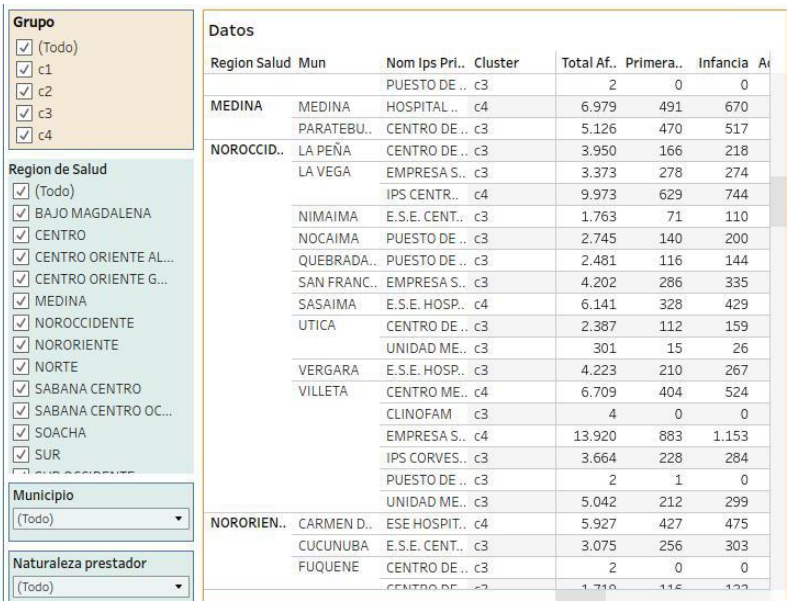


Figura 49. Sección de Introducción

2.- Consulta completa al dataset con los clusters a los que fueron asignados los prestadores de salud. Esta consulta fue enriquecida con información importante para la Secretaría de Salud y sus dependencias como la Región de Salud y el municipio donde se ubica el prestador, y la naturaleza jurídica.



Region Salud	Mun	Nom Ips Pri..	Cluster	Total Af..	Primera..	Infancia
		PUESTO DE ..	c3	2	0	0
MEDINA	MEDINA	HOSPITAL ..	c4	6.979	491	670
	PARATEBU..	CENTRO DE ..	c3	5.126	470	517
NOROCCID..	LA PEÑA	CENTRO DE ..	c3	3.950	166	218
	LA VEGA	EMPRESA S..	c3	3.373	278	274
		IPS CENTR..	c4	9.973	629	744
	NIMAIMA	E.S.E. CENT..	c3	1.763	71	110
	NOCAIMA	PUESTO DE ..	c3	2.745	140	200
	QUEBRADA..	PUESTO DE ..	c3	2.481	116	144
	SAN FRANC..	EMPRESA S..	c3	4.202	286	335
	SASAIMA	E.S.E. HOSP..	c4	6.141	328	429
	UTICA	CENTRO DE ..	c3	2.387	112	159
		UNIDAD ME..	c3	301	15	26
	VERGARA	E.S.E. HOSP..	c3	4.223	210	267
	VILLETA	CENTRO ME..	c4	6.709	404	524
		CLINOFAM	c3	4	0	0
		EMPRESA S..	c4	13.920	883	1.153
		IPS CORVES..	c3	3.664	228	284
		PUESTO DE ..	c3	2	1	0
		UNIDAD ME..	c3	5.042	212	299
NORORIEN..	CARMEN D..	ESE HOSPIT..	c4	5.927	427	475
	CUCUNUBA	E.S.E. CENT..	c3	3.075	256	303
	FUQUENE	CENTRO DE ..	c3	2	0	0

Figura 50. Sección de Consulta

3.- Se incluyó un elemento de información geográfica que interactúa con las opciones de consulta. Este elemento es importante porque permite visualizar la ubicación de los prestadores en la geografía del departamento. Una observación interesante que mostró este elemento fue la

distribución de los prestadores públicos y privados, que indica que los prestadores públicos tienden a cubrir la atención en salud de los municipios y regiones más apartados del departamento. También muestra a los prestadores con una marca de tamaño lo que permite distinguir y ubicar a los prestadores más grandes (en cuanto al número de afiliados).

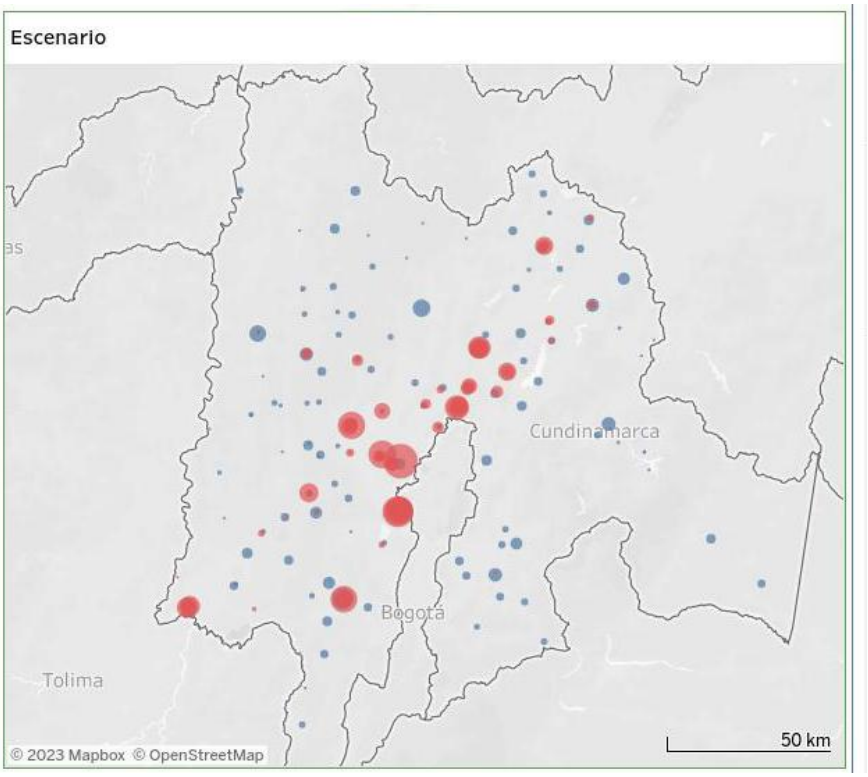


Figura 51. Sección de Información Geográfica

4.- En una cuarta sección se presentan las estadísticas de los clusters frente a las variables de los afiliados. Este recurso presenta a los analistas y tomadores de decisiones la posibilidad de mejorar el enfoque de los programas de salud, pues muestra la cobertura o peso de cada cluster en las variables de los afiliados.

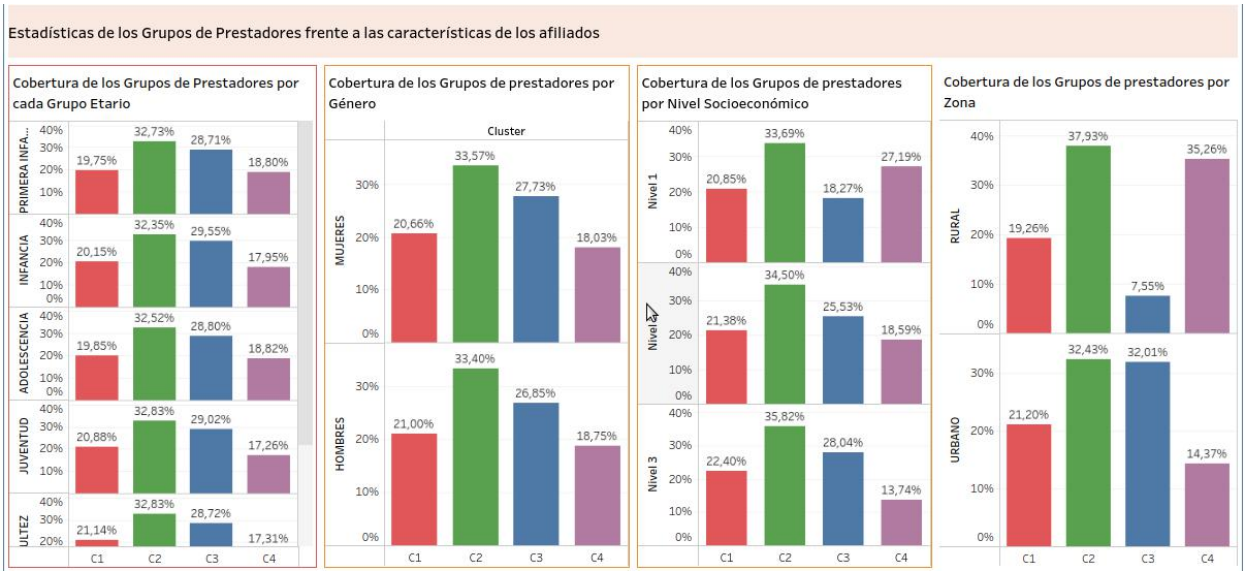


Figura 52. Sección de estadísticas de los clusters

Sumario del capítulo

En esta etapa se implementó un recurso para el despliegue de los principales resultados obtenidos en la etapa de evaluación y elección del modelo con el mejor desempeño, que correspondió a K-means; en este sentido, se afianzó la importancia de la visualización como elemento esencial para la comunicación de los principales aspectos y resultados de un modelo de analítica de datos; Para el diseño y despliegue de este recurso se empleó las facilidades y herramientas de la plataforma Tableau; dando cumplimiento al cuarto objetivo específico de este trabajo de grado.

8.- CONCLUSIONES Y TRABAJOS FUTUROS

"Y sigues ahi sentada, en el patio. No te has movido desde que comencé a escribir esta carta. Estás, siplemente, pensando. Ruego a Dios que algún día encuentres eso que buscas con tanto afán."
CARL SAGAN [Contacto]

Este trabajo se propuso la identificación de tendencias o relaciones entre las variables de los afiliados al Sistema de Salud en el departamento de Cundinamarca mediante el desarrollo de un modelo de analítica de datos que permita apoyar decisiones tendientes al mejoramiento de la cobertura y acceso a los servicios de salud; Se identificaron tres aspectos generales en este propósito, por un lado las variables que caracterizan a los afiliados, por otro lado el hecho de que los servicios de salud están asociados a los prestadores que operan en el departamento y brindan, entre otros, la atención primaria en salud a los afiliados. En la articulación de estos dos aspectos, se consolidó un dataset multidimensional con 13 características de los afiliados y el número de afiliados en cada variable que atiende cada prestador de salud.

En este punto, se atendió el tercer aspecto del propósito general cual es identificar las tendencias y relaciones en el dataset, para lo cual se recurrió a los algoritmos de clustering que propone la Ciencia de Datos. Después de un proceso de evaluación de los algoritmos K-means, Clustering Jerárquico, y K-medoids, las métricas empleadas permitieron determinar que el algoritmo K-means presenta el mejor desempeño. Como resultado se obtuvieron 4 clusters estables que reflejan las relaciones o tendencias entre características de afiliados y prestadores de servicios de salud. Este resultado fue desplegado en una plataforma de visualización, lo que responde a la necesidad de apoyar la toma de decisiones desde un contexto administrativo.

Así pues, se concluye que el proceso sistemático de implementación de este modelo permitió abordar y responder positivamente la situación problema y el objetivo general planteados; y constituye un referente para abordar casos de estudio similares con mayor alcance territorial.

La experiencia que aportó el desarrollo de este trabajo permite plantear trabajos futuros desde distintos puntos de vista:

Desde un punto de vista administrativo, se plantea como trabajo futuro aplicar los resultados hallados en este trabajo en un escenario real. Se propone la elección de un programa piloto con campañas de salud dirigidas a poblaciones específicas, que permita aplicar las distribuciones de los clusters frente a a características de los afiliados. Por ejemplo, programas de salud orientados a mitigar riesgos de salud mental en la población adolescente de la zona urbana y rural del departamento, con el concurso y articulación decisiva de los prestadores de servicios de salud.

Desde el punto de vista técnico, se propone para futuros trabajos o proyectos el empleo de algoritmos de clustering para la segmentación o agrupamiento de los afiliados en relación con variables médicas que permitan mejorar la atención en un contexto clínico. Se trata de un escenario mucho más complejo pero con el potencial de contribuir al fortalecimiento del Sistema de Salud en la región y el país.

9.- REFERENCIAS BIBLIOGRÁFICAS

[1] Constitución Política de Colombia. Art. 49. (1991). Disponible en: https://www.corteconstitucional.gov.co/inicio/Constitucion%20política%20de%20Colombia%20-%202015.pdf . Consultado en abril de 2023.

[2] Gobernación de Cundinamarca. “Sistema Integrado de Gestión y Control – SIGC. Manual del Sistema Integral de Gestión y Control”. Función Pública, Gobernación de Cundinamarca. Disponible en: http://isolucion.cundinamarca.gov.co/Isolucion. Consultado en abril de 2023.

[3] Hayashi, Chikio. "What is Data Science? Fundamental Concepts and a Heuristic Example". In Hayashi, Chikio; Yajima, Keiji; Bock, Hans-Hermann; Ohsumi, Noboru; Tanaka, Yutaka; Baba, Yasumasa (eds.). Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51. (1 January 1998). doi:10.1007/978-4-431-65950-1_3. ISBN 9784431702085.

[4] Conway, D. “The Data Science Venn Diagram”. 2010. Portal web Drew Conway. Disponible en: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram . Consultado en abril de 2023.

[5] Brodie, M. “What’s Data Sciencie?”. Applied Data Science (pp.101-130). 2019. http://dx.doi.org/10.1007/978-3-030-11821-1_8 .

[6] Gupta, S. C. “Actionable Insights from 4 Types of Data Analytics”. Toward Data Science. 2021. Recurso disponible en: https://towardsdatascience.com/actionable-insights-from-descriptive-diagnostic-predictive-prescriptive-data-analytics-drivetrain-approach-f4e08828cc7. Accedido en abril de 2023.

[7] Ayodele, Taiwo. (2010). “Types of Machine Learning Algorithms”. 10.5772/9385.

[8] Rajbanshi, S. “Everything you need to know about Machine Learning”. Analytics Vidhya. 2021. Recurso disponible en https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/. Consultado en abril de 2023.

[9] Lerena, O. “Métodos y aplicaciones de la ciencia de datos para las políticas de CTI: redes sociales, minería de textos y clustering”. Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación. 2019.

[10] Görür, Dilan. “Introduction to Clustering”. University of California, Irvine. 2011. Recurso disponible en: https://www.math.uci.edu/icamp/summer/research_11/gorur/clustering_iCAMP.pdf . Consultado en junio de 2023.

[11] Almenara J, González JL, García C, Peña P. “¿Qué es el Análisis de Componentes Principales?” 1998; 1268:58-60.

[12] Marin, J. “Analisis de Componentes Principales”. Universidad Carlos III de Madrid. 2006. Recurso disponible en: https://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema3am.pdf. Consultado en abril de 2023.

[13] Galiano Casas, G. & García Gonzalo, E.. “El algoritmo k-means aplicado a clasificación y procesamiento de imágenes”. Computación Numérica. Departamento de Matemáticas. Universidad de Oviedo. 2023. Recurso disponible en: https://www.unioviedo.es/compnum/laboratorios_py/new/kmeans.html . Consultado en mayo de 2023.

[14] Amat Rodrigo, Joaquín. “Clustering y heatmaps: Aprendizaje no supervisado”. cienciadedatos.net. 2017. Recurso disponible en: https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps#Medidas_de_distancia . Accedido en mayo de 2023.

[15] University of Cincinnati. UC Business Analytics R Programming Guide. 2018. Recurso disponible en: https://uc-r.github.io/kmeans_clustering. Consultado en mayo de 2023.

[16] Banerji, Ankita. “K-Mean: Getting the Optimal Number of Clusters”. Analytics Vidhya. Recurso disponible en: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/ . Consultado en mayo de 2023.

[17] Kamande, Samuel & Miriti, Kenneth & Ahishakiye, Emmanuel. “Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations”. International Journal of Computer Applications. 181. 33-42. 2018. 10.5120/ijca2018917609.

[18] Rhys, H. (2020). “Machine Learning with R, the tidyverse, and mlr”. Manning Publications. ISBN 9781617296574.

[19] Hardt, N. S., Muhamed, S., Das, R., Estrella, R., & Roth, J. “Neighborhood-level hot spot maps to inform delivery of primary care and allocation of social resources”. The Permanente journal, 17(1), 4–9. 2013. https://doi.org/10.7812/TPP/12-090

[20] Tuson, M., Turlach, B., Murray, K., Kok, M. R., Vickery, A., & Whyatt, D. “Predicting Future Geographic Hotspots of Potentially Preventable Hospitalisations Using All Subset Model Selection and Repeated K-Fold Cross-Validation”. International journal of environmental research and public health, 18(19), 10253. 2021. https://doi.org/10.3390/ijerph181910253

[21] Byrne, M. M., Daw, C. N., Nelson, H. A., Urech, T. H., Pietz, K., & Petersen, L. A. “Method to develop health care peer groups for quality and financial comparisons across hospitals”. Health services research (44), 577-592. 2009. doi:10.1111/j.1475-6773.2008.00916.x . Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677055/.

[22] Delamater, P. L., Shortridge, A., & Messina, J. P. “Regional health care planning: a methodology to cluster facilities using community utilization patterns”. BMC health services research, 13:333. 2013. doi: 10.1186/1472-6963-13-333 . Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3766152/

[23] Nnoaham, K.E., Cann, K.F. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population?. BMC Public Health 20, 798. 2020. https://doi.org/10.1186/s12889-020-08930-z

[24] Bejarano, J.M. “Creación de clusters en el mercado de prestadores de salud de Colombia”. Universidad de los Andes. Facultad de Economía. Bogotá. 2016.

[25] Administradora de los Recursos del Sistema General de Seguridad Social en Salud. ADRES. “Base de Datos Unica de Afiliados (BDUA)”. 2023. Recurso disponible en: https://www.adres.gov.co/eps/procesos/bdua . Consultado en mayo de 2023.

[26] Ministerio de Salud y Protección Social. “Registro Especial de Prestadores de Servicios de Salud - REPS”. 2023. Recurso disponible en: https://prestadores.minsalud.gov.co/habilitacion/ . Consultado en mayo de 2023.

[27] Departamento Nacional de Planeación. “¿Que es el Sisbén?”. 2023. Disponible en: https://www.sisben.gov.co/Paginas/que-es-sisben.aspx . Consultado en mayo de 2023.

[28] Hopkins, Brian; Skellam, John Gordon. "Un nuevo método para determinar el tipo de distribución de plantas individuales". Anales de botánica . Annals Botany Co. 18 (2): 213-227. 1954.

[29] Obilor, E & Amadi, E. “Test for Significance of Pearson’s Correlation Coefficient (r)”. International Journal of Innovative Mathematics, Statistics & Energy Policies 6(1):11-23, Jan-Mar, 2018.

[30] Abdi, H. and Williams l. J. “Principal component analysis”. John Wiley & Sons, Inc. WIREs Comp Stat 2010 2 433–459. 2010.

[31] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs Azam. (2022). “NbClust Package for determining the best number of clusters”. The Comprehensive R Archive Network - CRAN. Recurso disponible en: https://cran.r-project.org/web/packages/NbClust/NbClust.pdf . Consultado en mayo de 2023.

[32] Statistical tools for high-throughput data analysis - STHDA-. “How to choose the appropriate clustering algorithms for your data? - Unsupervised Machine Learning”. 2023. Recurso disponible en: http://www.sthda.com/english/wiki/wiki.php?id_contents=7932. Consultado en mayo de 2023.

[33] R Core Team. “R: A language and environment for statistical computing”. R Foundation for Statistical Computing. Vienna, Austria. 2023. Recurso disponible en: https://www.R-project.org/. Consultado en junio de 2023.

[34] RStudio Team. RStudio: Integrated Development for R. RStudio. PBC, Boston, MA. 2020. Recurso disponible en: http://www.rstudio.com/ . Consultado en junio de 2023.

[35] Tableau Team. Tableau version Public. 2023. Recurso disponible en: https://public.tableau.com/ . Consultado en junio de 2023.