APLICACIÓN DE TÉCNICAS DE REDUCCIÓN DE DIMENSIONALIDAD PARA DETERMINAR UN SCORE DE CONEXIÓN INTEGRAL EN ASOCIADOS DE LA COOPERATIVA COOMEVA

Daniel Arturo Sánchez Ortiz
Sergio Andrés Silva Guamanga
Juliana Valentina Niño Camargo

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

Director(a)
Orlando Joaqui Barandica

Codirector(a)
Oscar Orozco

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 19 DE 2025

TABLA DE CONTENIDO

1 DEFINICIÓN DEL PROBLEMA 6

1.1 PLANTEAMIENTO DEL PROBLEMA 6

1.2 FORMULACIÓN DEL PROBLEMA 6

2 OBJETIVOS DEL PROYECTO 7

2.1 OBJETIVO GENERAL 7

2.2 OBJETIVOS ESPECÍFICOS 7

3 MARCO TEÓRICO Y ANTECEDENTES 8

3.1 MARCO TEÓRICO 8

3.2 ANTECEDENTES 10

4 ANÁLISIS EXPLORATORIO DE DATOS *OBJETIVO ESPECIFICO 1* 11

4.1 DETALLE DE VARIABLES 12

4.2 ANÁLISIS VARIABLES NUMERICAS 14

4.2.1 CORRELACIÓN VARIABLES NUMERICAS 17

4.3 ANÁLISIS VARIABLES CATEGORICAS 18

5 APLICACIÓN DE TÉCNICAS DE REDUCCIÓN DE DIMENSIONALIDAD *OBJ. ESPECIFICO 2* 24

5.1 REVISIÓN DE METODOLOGÍAS DE REDUCCIÓN DE DIMENSIONALIDAD 24

5.2 SELECCIÓN DE LA METODOLOGÍA: ANÁLISIS DE COMPONENTES PRINCIPALES (ACP) 24

5.2.1 EVALUACIÓN DE LA ADECUACIÓN DEL ACP: TEST DE KMO 26
5.2.2 PREPROCESAMIENTO DE LOS DATOS: 26
5.2.3 CÁLCULO DE LOS COMPONENTES PRINCIPALES: 27
5.2.4 ANÁLISIS DE LA VARIANZA EXPLICADA: 27
5.2.5 EVALUACIÓN DE LAS CARGAS FACTORIALES: 27

5.3 RESULTADOS DE APLICACIÓN ACP. 28

5.3.1 ANÁLISIS DEL PRIMER COMPONENTE PRINCIPAL (PC1): 29
5.3.2 ANÁLISIS DEL SEGUNDO COMPONENTE PRINCIPAL (PC2): 30

6 CALCULO SCORE INTEGRAL DE CONEXIÓN OBJETIVO ESPECIFICO 3..... 32

6.1 OBTENCIÓN DE LOS COMPONENTES PRINCIPALES 32

6.2 PROYECCIONES SOBRE LOS COMPONENTES PRINCIPALES..... 32

6.3 NORMALIZACIÓN Y AJUSTE FINAL DEL SCORE 33

6.4 SELECCIÓN DEL NÚMERO DE COMPONENTES PRINCIPALES PARA EL SCORE 33

6.4.1 CRITERIOS PARA LA SELECCIÓN DEL NÚMERO DE COMPONENTES PRINCIPALES..... 33
6.4.2 JUSTIFICACIÓN DE LA NO SELECCIÓN DE OTROS NÚMEROS DE COMPONENTES..... 35

7 VISUALIZACIÓN E INTERPRETACIÓN SCORE INTEGRAL DE CONEXIÓN OBJ ESPECIFICO 4 35

7.1 SEGMENTACIÓN ASOCIADOS SEGUN SCORE 36

8 VALIDACIÓN Y DISCUSIÓN DE RESULTADOS 40

8.1 CORRELACIÓN DEL SCORE CON LAS VARIABLES ORIGINALES 40

8.2 ANÁLISIS DE VARIANZA (ANOVA) ENTRE CUARTILES DEL SCORE 41

8.3 ANÁLISIS COMPLEMENTARIO AGRUPAMIENTO MEDIANTE K-MEANS 43

8.3.1 CONCLUSIÓN COMPARATIVA ENTRE EL SCORE INTEGRAL Y EL MODELO K-MEANS 47

9 CONCLUSIONES Y TRABAJOS FUTUROS..... 48

9.1 CONCLUSIONES 48

9.2 TRABAJOS FUTUROS..... 50

10 REFERENCIAS BIBLIOGRÁFICAS 50

LISTA DE FIGURAS

Figura 1 Distribución Variables numéricas. 16
Figura 2. Correlación variables numéricas..... 17
Figura 3. Estado de los asociados. 19
Figura 4. Alertas comportamiento de pago externo. 19
Figura 5. Débito automático. 20
Figura 6. Hábito de pago estatutarios. 20
Figura 7. Estrato socioeconómico..... 21
Figura 8. Nivel académico por género. 21
Figura 9. Actividad laboral. 22
Figura 10. Tipo de vivienda..... 23
Figura 11. Círculo de correlación. 31
Figura 12. Representación simultánea. 32
Figura 13 Correlación entre índices sintéticos..... 34
Figura 14 Distribución del Score 36
Figura 15 Distribución de variables categóricas por segmento de conexión 39
Figura 16 Correlación del Score con variables numéricas 41
Figura 17 Método del Codo Para obtener el K óptimo..... 44
Figura 18 Metodo Silhouette 45
Figura 19 Descripción variables categóricas por Clusters..... 45

LISTA DE TABLAS

Tabla 1. Estadísticas Descriptivas de las Variables Numéricas. 15
Tabla 2. Estrato vs nivel académico. 22
Tabla 3. Actividad laboral versus Tipo de vivienda. 23
Tabla 4. Porcentaje de varianza acumulado en componentes. 28
Tabla 5. Carga factorial variables en componentes. 29
Tabla 6 Cuartiles del Score con cantidad de asociados 37
Tabla 7 Promedios de variables numéricas por cuartiles de conexión..... 37
Tabla 8 Resultados ANOVA 43
Tabla 9 Promedio Variables numéricas por Cluster 47

INTRODUCCIÓN

En los últimos años, la cooperativa Coomeva ha enfrentado una disminución significativa en su tasa de crecimiento de asociados, acompañada de una alta tasa de desvinculaciones. Este problema crítico no solo afecta la sostenibilidad de la cooperativa, sino que también compromete la cohesión social y el sentido de pertenencia entre sus miembros. A pesar de los esfuerzos realizados para atraer nuevos asociados, estos se ven contrarrestados por el número de retiros, lo que ha llevado a la necesidad urgente de comprender y abordar las causas subyacentes de este fenómeno.

Para resolver este problema, se desarrolló un score integral de conexión que permite medir y entender la relación de los asociados con los productos y servicios de Coomeva. Este score será una herramienta esencial para diseñar estrategias efectivas que mejoren la retención y fidelización de los miembros.

El proyecto siguió varias etapas clave: la comprensión del negocio y los datos, la preparación de los datos, el modelado, la interpretación y aplicación. Se utilizaron técnicas avanzadas de análisis multivariado y aprendizaje automático para identificar patrones y segmentos de asociados con diferentes niveles de conexión. Todo el procesamiento y análisis de datos se realizó utilizando el software estadístico **R** [1] aprovechando su amplia capacidad para el manejo de datos, modelado estadístico y visualización. Además, se desarrollaron visualizaciones que facilitaron la interpretación de los resultados y la identificación de factores críticos que afectan la relación de los asociados con la cooperativa.

Este proyecto no solo buscó resolver un problema crítico para Coomeva, sino también establecer un marco de trabajo robusto y sistemático para el análisis de datos, proporcionando herramientas y conocimientos que asegurarán la sostenibilidad y el crecimiento a largo plazo de la cooperativa.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En los últimos diez años, la cooperativa Coomeva ha enfrentado un estancamiento significativo en el crecimiento de su base de asociados. Mientras que en periodos anteriores (1991-2001 y 2002-2012) se lograron tasas anuales de crecimiento compuesto (TACC) del 21% y 6% respectivamente, el periodo más reciente (2013-2023) muestra un crecimiento promedio anual de apenas 0,5%. Estas cifras corresponden a un análisis propio realizado con base en los registros históricos de la cooperativa. Este estancamiento se debe, en gran parte, a la alta tasa de desvinculación, la cual ha contrarrestado los esfuerzos de vinculación de nuevos asociados.

Uno de los principales factores que explican este fenómeno es el bajo uso de los productos y servicios ofrecidos por la cooperativa. Muchos asociados no aprovechan los beneficios disponibles, ya sea por desconocimiento de la oferta, falta de interés o una comunicación inefectiva de los servicios. Esta desconexión con la cooperativa afecta la percepción de valor que los asociados tienen de su vínculo con Coomeva, lo que a su vez incrementa las tasas de retiro. Otro factor determinante en la desvinculación es la situación económica de los asociados, ya que se ha identificado que aquellos con moras superiores a 120 días tienen una probabilidad muy baja de regularizar su situación, lo que los lleva a estados de inactividad y, posteriormente, a la exclusión de la cooperativa.

Desde la perspectiva de la ciencia de datos, este problema representa un desafío complejo de segmentación y predicción, ya que la permanencia de los asociados depende de múltiples variables que interactúan de manera no lineal. Tradicionalmente, los análisis han sido descriptivos y basados en reglas generales, lo que limita la capacidad de anticipar comportamientos y diseñar estrategias personalizadas. La aplicación de técnicas avanzadas de ciencia de datos, como la reducción de dimensionalidad, aprendizaje automático y modelos de segmentación, permite identificar patrones ocultos en el comportamiento de los asociados y cuantificar su nivel de conexión con la cooperativa.

Al desarrollar un score integral de conexión, es posible sintetizar múltiples dimensiones en una única métrica que facilite la toma de decisiones estratégicas. Esto no solo optimiza la fidelización mediante intervenciones dirigidas, sino que también contribuye a la sostenibilidad de la cooperativa al reducir la tasa de desvinculación y mejorar la asignación de recursos en estrategias de retención.

1.2 FORMULACIÓN DEL PROBLEMA

Con este proyecto logramos dar respuesta a la siguiente interrogante: ¿Cómo puede la cooperativa Coomeva mejorar la retención y fidelización de sus asociados mediante el desarrollo de un score integral de conexión utilizando técnicas de análisis de datos multivariados?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un score integral para cada asociado de la cooperativa Coomeva, utilizando técnicas de análisis multivariado, que refleje su nivel de conexión con la cooperativa y su interacción con los productos y servicios ofrecidos, con el fin de mejorar la retención y fidelización de los asociados.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar un análisis descriptivo de los datos, abarcando información sobre el uso de productos y servicios, tenencia de productos, comportamiento de pagos y otros indicadores de valor y conexión con la cooperativa.
- Utilizar técnicas de reducción de dimensionalidad, para identificar las variables más significativas que influyen en la conexión de los asociados con la cooperativa.
- Calcular un score integral de conexión para cada asociado, basado en variables identificadas mediante la reducción de dimensionalidad y técnicas de análisis no supervisado. Este score considerará aspectos clave como la tenencia y uso de productos, comportamiento de pago y generación de valor, con el fin de proporcionar una medida completa de la relación de cada asociado con la cooperativa Coomeva.
- Generar visualizaciones que permita interpretar los resultados del cálculo del score integral de conexión, facilitando la identificación de segmentos de asociados con baja conexión y los factores clave que afectan su relación con la cooperativa.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés, Exploratory Data Analysis) es un enfoque estadístico utilizado para analizar conjuntos de datos y resumir sus características principales, a menudo mediante métodos visuales. Popularizado por John Tukey en la década de 1970, el EDA tiene como objetivos comprender la distribución de los datos, detectar anomalías, comprobar supuestos, generar hipótesis y explorar relaciones entre variables. Como Tukey afirmó: "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone" [2]. Las técnicas comunes de EDA incluyen la visualización de datos mediante gráficos como histogramas y diagramas de dispersión, el cálculo de estadísticas descriptivas como la media y la desviación estándar, y el análisis multivariante utilizando matrices de correlación y análisis de componentes principales. Behrens señala que el EDA es fundamental para comprender la estructura de los datos y detectar posibles sesgos o problemas en los datos. Este enfoque es esencial para proporcionar una comprensión profunda de los datos y preparar adecuadamente los datos para el modelado posterior [3].

- El machine learning o aprendizaje automático es una subdisciplina de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y hacer predicciones a partir de datos. A diferencia de los sistemas informáticos tradicionales, que siguen un conjunto de instrucciones predefinidas, los sistemas de machine learning identifican patrones y regularidades en los datos para mejorar su desempeño a lo largo del tiempo. Según Bishop (2006), el machine learning es un campo interdisciplinario que integra métodos de estadística, optimización y teoría de la información para descubrir patrones en los datos y construir modelos predictivos. Este enfoque multidisciplinario permite abordar una amplia gama de problemas y aplicaciones, desde el reconocimiento de patrones hasta la toma de decisiones automatizada [4]. Domingos (2012) destaca que el objetivo principal del machine learning es desarrollar métodos que puedan generalizar a 13 partir de ejemplos observados para hacer predicciones sobre datos nuevos y no vistos previamente. La capacidad de generalización es crucial para el éxito de los modelos de machine learning en aplicaciones del mundo real, donde los datos de prueba pueden diferir significativamente de los datos de entrenamiento [5]. Goodfellow, Bengio y Courville (2016) señalan que el machine learning abarca una amplia variedad de técnicas, desde modelos lineales simples hasta complejas redes neuronales profundas. La diversidad de métodos disponibles permite abordar problemas con diferentes niveles de complejidad y volúmenes de datos, adaptándose a las necesidades específicas de cada aplicación [6].
- El aprendizaje no supervisado se ocupa de datos sin etiquetas, y el objetivo es encontrar patrones o estructuras ocultas en los datos. Este enfoque es útil para la exploración y comprensión de los datos. Hastie, Tibshirani y Friedman (2009) explican que, en el aprendizaje no supervisado, no se proporcionan etiquetas para los datos de entrada. Los algoritmos deben descubrir las similitudes y

diferencias en los datos por sí mismos, comúnmente a través de métodos como el clustering y la reducción de dimensionalidad (Hastie, Tibshirani, & Friedman, 2009) [7].

- La reducción de dimensionalidad es una técnica empleada en el aprendizaje automático y análisis de datos que busca disminuir el número de variables (o dimensiones) en un conjunto de datos, preservando la mayor cantidad posible de información relevante. Los objetivos principales de este método incluyen la simplificación del modelo, lo que facilita su interpretación y análisis; la eliminación de redundancias, reduciendo características que no aportan nueva información; la mejora del rendimiento del modelo, al aumentar la eficiencia computacional y disminuir el riesgo de sobreajuste; y la facilitación de la visualización, permitiendo representar datos de alta dimensionalidad en espacios de menor dimensionalidad como 2D o 3D, haciendo posible la identificación de patrones y relaciones en los datos de forma más intuitiva [6] [8].
- El análisis de componentes principales (PCA, por sus siglas en inglés, Principal Component Analysis) es una técnica de reducción de dimensionalidad que transforma un conjunto de datos de alta dimensionalidad en uno de menor dimensionalidad, preservando la mayor cantidad posible de variabilidad original. Este método estandariza los datos, calcula la matriz de covarianza, determina los vectores y valores propios, selecciona los componentes principales más significativos y proyecta los datos en este nuevo espacio de componentes principales [4] [8]. El PCA se utiliza comúnmente para reducir el número de características, mejorar la visualización de datos y eliminar el ruido, facilitando el preprocesamiento y análisis exploratorio de conjuntos de datos complejos [9]. A continuación, se realiza la descripción de los Componentes Principales:
    - Primer componente principal (PC1): Es la dirección en el espacio de datos que tiene la mayor varianza. En otras palabras, es la combinación lineal de las variables originales que captura la mayor parte de la información contenida en el conjunto de datos.
    - Segundo componente principal (PC2): Es la dirección ortogonal (perpendicular) al primer componente y que captura la siguiente mayor cantidad de varianza posible. No tiene correlación con el primer componente, lo que asegura que se está extrayendo información nueva.
    - Componentes posteriores (PC3, PC4, ...): Cada componente posterior es ortogonal a los anteriores y explica la mayor cantidad posible de varianza restante en los datos. A medida que se avanza en los componentes, la varianza que explican va disminuyendo [10].
- Los autoencoders son redes neuronales diseñadas para aprender una representación eficiente de los datos, típicamente utilizada para la reducción de dimensionalidad o el aprendizaje no supervisado. Están compuestos por un codificador que comprime la entrada en una representación de menor dimensionalidad y un decodificador que reconstruye la entrada original a partir de esa representación comprimida. Entre las variantes de autoencoders se encuentran los denoising autoencoders, que eliminan el ruido de los datos, los sparse autoencoders, que promueven una representación latente esparsa, y los variational autoencoders, que generan nuevas muestras de datos similares a las de entrenamiento [5] [11] [12]. Estas técnicas son aplicables en la reducción de dimensionalidad, eliminación de ruido, generación de datos sintéticos y detección de anomalías.
- La retención de asociados en Coomeva se centra en mantener a los miembros actuales satisfechos y reducir la deserción. Esto se logra a través de la calidad de los servicios, una atención personalizada, beneficios exclusivos, comunicación eficiente, y facilidad de acceso. La cooperativa se esfuerza por

asegurar que los servicios ofrecidos, como salud, financieros, seguros, educación y recreación, cumplan con altos estándares de calidad, satisfaciendo así las necesidades de los asociados [13].

- La fidelización de asociados implica cultivar y mantener la lealtad de los miembros a largo plazo. Coomeva implementa programas de fidelización que recompensan la lealtad, mejora continuamente la experiencia del usuario, fomenta la participación en la cooperativa, ofrece programas de educación y formación, y promueve iniciativas de responsabilidad social. Estas estrategias buscan crear un sentido de pertenencia y comunidad entre los asociados, alineando los valores de la cooperativa con los intereses de sus miembros [14].

3.2 ANTECEDENTES

Actualmente, se desarrollan múltiples investigaciones orientadas al uso de técnicas estadísticas y de aprendizaje automático para generar indicadores que faciliten la toma de decisiones en distintos contextos sociales y educativos, optimizando el análisis de grandes volúmenes de datos y mejorando la eficiencia en la interpretación de resultados. Un ejemplo representativo de este tipo de estudios tuvo lugar en Colombia, con el desarrollo de un proyecto orientado a construir un índice socioeconómico-familiar para los estudiantes que presentaron la prueba Saber 11 en el año 2012 [15]. Esta investigación utilizó información recolectada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), correspondiente a 424.916 estudiantes que además respondieron un cuestionario sociodemográfico de 58 preguntas sobre el entorno familiar, social, económico y cultural. El estudio se estructuró en dos fases principales: en la primera se realizó una revisión de literatura especializada para definir las dimensiones y variables que permitirían caracterizar el nivel socioeconómico; en la segunda fase, se aplicó la técnica de Análisis de Componentes Principales (ACP) como herramienta de reducción de dimensionalidad, logrando condensar la información relevante en un único componente interpretado como el índice socioeconómico. Dado que la mayoría de las variables eran categóricas, utilizaron el método de asignación óptima para atribuirles pesos cuantitativos. Finalmente, contrastaron el índice con el desempeño académico de los estudiantes en las pruebas Saber 11, encontrando una alta correlación entre ambos. Este estudio resulta relevante para la presente propuesta de investigación sobre el desarrollo de un score integral de conexión, ya que demuestra la utilidad del ACP para sintetizar información compleja y generar indicadores robustos, adaptables a contextos educativos, sociales o cooperativos. La metodología empleada ofrece un referente claro para el diseño de modelos que buscan identificar patrones latentes y construir herramientas de apoyo a la toma de decisiones estratégicas, en este caso, dirigidas a mejorar la gestión de los asociados en la cooperativa y fortalecer su vínculo institucional.

Durante el proceso de recolección de antecedentes, se encontró otra investigación que realizo el análisis del comportamiento de los clientes con tarjetas de crédito de una institución financiera colombiana con base en su calificación de riesgo de crédito [16], en este caso, utilizaron un modelo no supervisado llamado K-means, el cual es un método de agrupamiento que divide los datos en n observaciones dentro de agrupaciones diferentes, algunas variables dentro del modelo incluyeron información de saldos, uso de tarjeta, estado del cliente, entre otros. Durante el proceso se realizó la estandarización de variables aplicando el método de Elbow y el de average silhouette, que sirve para identificar la cantidad óptima de centroides (k) para definir la cantidad de grupos, una vez se aplica el algoritmo K-means se prueban cuatro

formas de este algoritmo: Lloyd, Forgy, MacQueen y Hartigan-Wong y se escoge el de mayor distancia intraclúster, para este proyecto el HartiganWong resultó con la mayor distancia intraclúster. Lo que permitió que en la investigación identificara patrones de comportamiento de cada grupo de clientes. La investigación sobre el análisis del comportamiento de los clientes con tarjetas de crédito de una institución financiera colombiana proporciona una perspectiva relevante para el proyecto de tesis sobre el desarrollo de un score integral para los asociados de la cooperativa. Esta investigación utilizó el modelo no supervisado K-means para identificar diferentes grupos de clientes basados en su comportamiento y calificación de riesgo de crédito. Al igual que en el proyecto propuesto, esta investigación se centró en el análisis de múltiples variables relacionadas con los clientes, como saldos, uso de tarjeta y estado del cliente, para comprender mejor su conexión con la institución financiera. Además, la aplicación de técnicas de estandarización de variables y la selección del algoritmo más apropiado para el análisis, en este caso Hartigan-Wong, demuestran un enfoque 16 metodológico riguroso similar al que se pretende implementar en el proyecto de tesis. Por lo tanto, esta investigación es sólida para desarrollar estrategias de segmentación y evaluación de los asociados de la cooperativa según su nivel de conexión y participación en los productos y servicios ofrecidos.

En el estudio "Customer Segmentation through Data Mining for E-commerce" [17], se aborda la segmentación de clientes utilizando técnicas de minería de datos. El objetivo principal es desarrollar una guía sistemática para implementar la segmentación de clientes utilizando datos transaccionales sin etiquetar de una plataforma de comercio electrónico. El estudio emplea métodos de aprendizaje no supervisado para segmentar a los clientes, comenzando con el modelo RFM (Recency, Frequency, Monetary Value) para crear características de comportamiento. Luego, se aplica el método TF-IDF a las descripciones de productos para generar categorías de productos. Posteriormente, se utiliza el algoritmo de clustering Kmeans para agrupar a los clientes, seguido del uso del algoritmo Apriori para analizar las reglas de asociación entre los productos comprados. Para reducir la dimensionalidad de los datos y crear visualizaciones, se utilizan PCA y T-SNE, proporcionando finalmente recomendaciones concretas para el negocio basadas en los resultados. Este estudio aporta significativamente al proyecto actual al ofrecer un enfoque sistemático para la segmentación de clientes mediante técnicas de análisis multivariado y aprendizaje no supervisado. El uso del modelo RFM y el algoritmo K-means puede adaptarse para calcular el score integral de conexión de los asociados de Coomeva. Las técnicas de reducción de dimensionalidad como PCA y T-SNE también pueden ser útiles para visualizar y entender mejor los datos de los asociados. Sin embargo, la diferencia principal radica en el contexto y los objetivos específicos. Mientras que el estudio se centra en la segmentación de clientes de una plataforma de comercio electrónico para mejorar estrategias de marketing, el proyecto actual se enfoca en la retención y fidelización de los asociados de una cooperativa. Además, el proyecto busca desarrollar un score integral de conexión que considere factores adicionales como el comportamiento de pagos y la generación de valor, aspectos no abordados en el estudio mencionado.

4 ANÁLISIS EXPLORATORIO DE DATOS *OBJETIVO ESPECIFICO 1*

El análisis exploratorio de datos (AED) constituye una etapa fundamental en cualquier proyecto analítico, ya que permite conocer en profundidad la estructura, calidad y características de los datos disponibles. En el presente caso, este análisis tiene como propósito proporcionar una comprensión inicial del conjunto de datos suministrado por el área de Analítica Comercial de Coomeva, con corte a junio de 2024. Este paso es

esencial para garantizar la calidad de los datos y sentar las bases para la construcción de un **score de conexión de los asociados con la cooperativa**, el cual busca identificar y optimizar las relaciones entre los asociados y la organización.

El conjunto de datos cuenta con **262,366 registros de asociados** activos a la fecha de corte, y se compone de **32 variables** seleccionadas cuidadosamente por su relevancia para el objetivo del estudio. Estas variables se han agrupado en bloques temáticos que representan dimensiones clave en el análisis:

1. **Salud de la cartera:** Información relacionada con el comportamiento de pago, probabilidades de fuga y manejo financiero de los asociados.
2. **Generación de valor:** Indicadores del aporte que los asociados generan para la cooperativa en términos de ingresos y beneficios.
3. **Variables demográficas:** Características como edad, género, nivel educativo, y lugar de residencia, que ayudan a contextualizar y segmentar a los asociados.
4. **Uso de productos y servicios transaccionales y no transaccionales:** Datos que reflejan la interacción de los asociados con los diferentes productos y servicios ofrecidos por la cooperativa.

Este análisis busca identificar patrones, relaciones y posibles problemas en los datos (como valores faltantes o inconsistencias) que puedan influir en la fase de modelado posterior. Además, permitirá realizar una caracterización inicial de los asociados que ayudará a diseñar estrategias más efectivas para fortalecer la relación entre los asociados y la cooperativa.

4.1 DETALLE DE VARIABLES

A continuación, se presenta una tabla descriptiva con las variables del conjunto de datos, agrupadas en los bloques temáticos previamente definidos. Estas variables han sido seleccionadas por su relevancia para el análisis y la construcción del **score de conexión de los asociados con la cooperativa**.

- **Salud de la cartera:**

Variable	Nombre	Tipo	Descripción	Unidad/Valores posibles
Estado	Estado	Categórica	Estado actual del asociado con la Cooperativa.	"Activo", "Inactivo", "Suspendido"
AEC_Externos	Alerta Estado Créditos Externos	Categórica	Señala alertas asociadas al estado de los créditos externos.	"Con alerta", "Sin alerta", "Sin información"
ACP_Externo	Alerta Capacidad Pago Externo	Categórica	Indica si hay alertas sobre la capacidad de pago externa del asociado.	"Con alerta", "Sin alerta", "Sin información"
AHP_Externo	Alerta Habito Pago Externo	Categórica	Indica si existen alertas externas relacionadas con los hábitos de pago del asociado.	"Con alerta", "Sin alerta", "Sin información"
DAutomatico	Débito Automático	Categórica	Indica si el asociado tiene	"Sí", "No"

			débito automático habilitado.	
SHPE_3M	Segmento Habito Pago Estatutario 3 Meses	Categórica	Clasificación basada en hábitos de pago estatutario en los últimos 3 meses.	"Excelente", "Bueno", "Aceptable", "Deficiente", "Muy deficiente", "Sin información"
Pde_fuga	Probabilidad de fuga	Numérica	Estimación de la probabilidad de que el asociado deje de utilizar los servicios.	Porcentaje (%)
Acierta	Calificación Acierta	Numérica	Puntaje de calificación crediticia otorgado por el sistema Acierta.	Escala numérica de 0 a 999
EdadCartera	Edad de Cartera	Numérica	Edad promedio de la cartera asociada al cliente.	Días
IRP_3M	Índice Recaudo Promedio GECC Últimos 3 Meses	Numérica	Promedio del índice de recaudo en los últimos 3 meses.	Porcentaje (%)
CVencidasE	Cuotas Vencidas Estatutaria	Numérica	Número de cuotas vencidas en la línea estatutaria del asociado.	Número

- **Generación de valor:**

Variable	Nombre	Tipo	Descripción	Unidad/Valores posibles
Ingresos	Ingresos económicos	Numérica	Total de ingresos económicos generados por el asociado.	Pesos (COP)
CLV	Puntaje CLV	Numérica	Valor del cliente (Customer Lifetime Value) calculado para el asociado.	Escala numérica
ValorVida	Valor Vida	Numérica	Valor seleccionado por línea vida del asociado.	Pesos (COP)
ValorPers	Valor Perseverancia	Numérica	Valor seleccionado por perseverancia del asociado.	Pesos (COP)
ReservaM	Reserva Matemática	Numérica	Total de la reserva matemática acumulada por el asociado.	Pesos (COP)
SaldoAportes	Saldo Aportes	Numérica	Saldo actual en aportes del asociado.	Pesos (COP)
SaldoSolid	Saldo Solidaridad	Numérica	Saldo actual en el fondo de solidaridad del asociado.	Pesos (COP)

- **Demografía:**

Variable	Nombre	Tipo	Descripción	Unidad/Valores posibles
Genero	Genero	Categórica	Género del asociado.	"Femenino", "Masculino"

Estrato	Estrato	Categórica	Estrato socioeconómico del asociado.	“Alto”, “Medio”, “Bajo”
NAcademico	Nivel Académico	Categórica	Nivel educativo alcanzado por el asociado.	"Profesional", "Técnico o Tecnólogo", "Ningúno"
ALaboral	Actividad Laboral	Categórica	Actividad económica principal del asociado.	"Ama de Casa", "Asalariado", "Estudiante", "Independiente", "Otro tipo de Actividad", "Pensionado"
Vivienda	Tipo de Vivienda	Categórica	Tipo de vivienda del asociado.	"Propia", "Familiar", "Alquilada", "Desconocida"
Colaborador	Colaborador GECC	Categórica	Indica si el asociado es colaborador de la cooperativa.	"Sí", "No"
Edad	Edad	Numérica	Edad del asociado en años.	Años
Antiguedad	Antigüedad	Numérica	Tiempo de vinculación del asociado con la cooperativa.	Años

- **Uso de productos:**

Variable	Nombre	Tipo	Descripción	Unidad/Valores posibles
CEH_Recre	Cantidad Eventos Recreación Histórico	Numérica	Número histórico de eventos recreativos asistidos por el asociado.	Número
CEH_Educ	Cantidad Eventos Educación Histórico	Numérica	Número histórico de eventos educativos asistidos por el asociado.	Número
CEH_Fund	Cantidad Eventos Fundación Histórico	Numérica	Número histórico de eventos organizados por la fundación asistidos por el asociado.	Número
CProductos	Cantidad Productos	Numérica	Cantidad de productos del GECC que posee el asociado.	Número

4.2 ANÁLISIS VARIABLES NUMERICAS

Inicialmente, se realizó un análisis exploratorio de las variables numéricas con el objetivo de identificar patrones, tendencias y posibles valores atípicos que puedan impactar en el desarrollo del modelo de score de conexión de los asociados. En la **Tabla 1**, se presentan las principales métricas descriptivas de estas variables, agrupadas en dimensiones como salud de la cartera, generación de valor, demografía y uso de productos. Las estadísticas incluyen el valor mínimo, el primer cuartil, la mediana, la media, el tercer cuartil y el valor máximo, proporcionando una visión integral sobre el comportamiento y la dispersión de los datos numéricos analizados.

Tabla 1. Estadísticas Descriptivas de las Variables Numéricas.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Pde_fuga	0,005	0,112	0,188	0,232	0,309	0,956
Acierta	0	0	600	436	740	999
EdadCartera	0	0	1000	2099	5000	7000
IRP_3M	0,0	0,0	0,7	0,7	1,0	216,8
CVencidasE	0,0	0,0	0,0	7186,0	5000,0	89000,0
ValorVida	-$ 49.455.824	$ 8.807.137	$ 13.080.000	$ 18.956.740	$ 20.000.000	$ 478.068.267
ValorPers	-$ 124.372.322	$ 12.329.996	$ 17.440.000	$ 35.496.505	$ 42.163.649	$ 1.179.999.996
ReservaM	-$ 28.240.000	$ 3.829.000	$ 6.561.000	$ 16.390.000	$ 15.750.000	$ 623.737.711
SaldoAportes	$ 0	$ 0	$ 2.505.281	$ 3.962.893	$ 7.404.390	$ 61.805.182
SaldoSolid	$ 0	$ 0	$ 153.172	$ 3.911.650	$ 4.599.502	$ 281.967.193
Ingresos	$ 0	$ 2.050.000	$ 3.824.000	$ 6.278.000	$ 6.843.000	$ 10.000.000.000
CLV	0,0	102868,0	241277,0	427979,0	497280,0	15710360,0
Edad	4,0	35,0	46,0	47,8	60,0	108,0
Antigüedad	0,0	2,7	8,7	12,1	20,7	60,4
CEH_Recre	0,0	2,0	8,0	28,5	24,0	3935,0
CEH_Educ	0,0	0,0	0,0	1,0	1,0	121,0
CEH_Fund	0,0	0,0	0,0	0,6	0,0	154,0
CProductos	0,0	1,0	3,0	4,0	6,0	25,0

SALUD DE LA CARTERA

- **Probabilidad de Fuga:** La media es 0.232, con un valor máximo de 0.956, lo que sugiere que la mayoría de los asociados tienen una probabilidad baja de dejar de utilizar productos, aunque existen algunos casos cercanos al 100% de fuga.
- **Edad de Cartera:** Presenta un rango amplio, con valores desde 0 hasta 7,000 días, lo que muestra una gran heterogeneidad en el tiempo que los asociados han tenido productos activos.
- **Índice de Recaudo Promedio y Cuotas Vencidas:** Muestran valores significativos en sus máximos, lo que podría indicar la necesidad de analizar con detalle los casos más críticos para entender las causas subyacentes.

GENERACIÓN DE VALOR

- **Valor Vida y Valor Perseverancia:** Ambos valores muestran distribuciones con amplias desviaciones, siendo sus máximos notablemente altos (478 millones y 1,179 millones, respectivamente). Estos valores reflejan el impacto económico generado por los asociados en la cooperativa.
- **Reserva Matemática:** Oscila entre -28 millones y 623 millones, mostrando que algunos asociados tienen balances negativos, lo cual es relevante para evaluar riesgos.
- **Saldo Aportes y Solidaridad:** Presentan comportamientos esperados, con medianas y máximos alineados con las características del portafolio.

DEMOGRAFÍA

- **Edad:** Varía entre 4 y 108 años, con una media de 47.8 años, lo que refleja una base de asociados madura. Este rango amplio sugiere analizar grupos etarios para identificar comportamientos diferenciados. También hace necesario excluir las personas jurídicas las cuales pueden generar confusión en diferentes variables.
- **Antigüedad en la Cooperativa:** Muestra un promedio de 12.1 años, lo cual indica una alta permanencia de los asociados en general.

USO DE PRODUCTOS

- **Eventos Recreativos, Educativos y de Fundación:** Se observa una alta dispersión, con valores máximos de hasta 154 eventos en educación, lo que sugiere que algunos asociados participan significativamente en actividades de la cooperativa.
- **Cantidad de Productos:** La mediana es 3, con un máximo de 25 productos contratados, lo que muestra un grado de vinculación diverso entre los asociados. El producto promedio de los asociados es de 4.

En general, estas estadísticas descriptivas destacan las diferencias entre los asociados en términos de permanencia, uso de productos y generación de valor. Además, los valores extremos en algunas variables sugieren la necesidad de realizar análisis complementarios para identificar posibles outliers y determinar su impacto en el modelo.

Con el fin de observar la dispersión relativa y posibles valores atípicos en una escala comparable, se construyó un gráfico tipo boxplot de todas las variables numéricas **ya estandarizadas**. Al aplicar la estandarización, se eliminaron las diferencias de escala entre variables, lo cual permite visualizar de forma más objetiva el comportamiento interno de cada una, sin que su magnitud original distorsione la comparación.

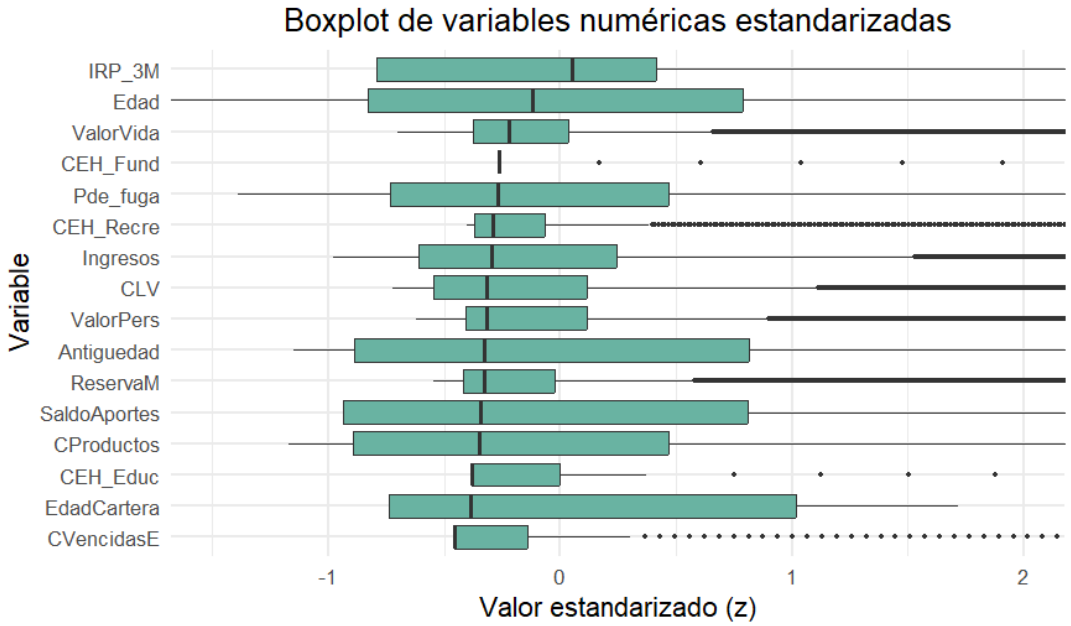

Figura 1 Distribución Variables numéricas.

En el gráfico se evidencian variables con fuerte concentración alrededor de la mediana y otras con presencia marcada de valores extremos, lo que es útil para interpretar su contribución futura en el análisis multivariado.

4.2.1 CORRELACIÓN VARIABLES NUMERICAS

Tras el análisis descriptivo de las variables numéricas, se procedió a evaluar la correlación entre estas, con el objetivo de identificar relaciones lineales que puedan aportar información relevante para el desarrollo del modelo de score de conexión de los asociados. Este análisis también busca detectar posibles problemas de multicolinealidad entre las variables, lo cual podría afectar la estabilidad y precisión del modelo.

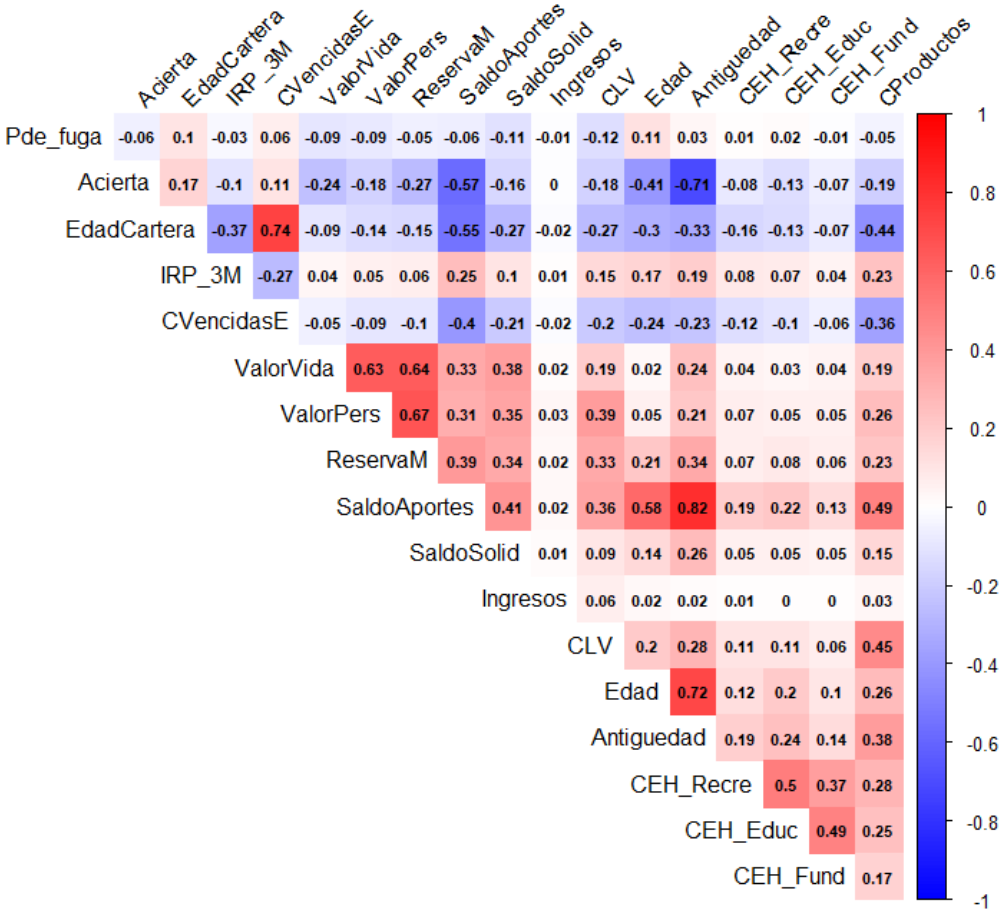

Figura 2. Correlación variables numéricas.

La Figura 1 presenta un mapa de calor que resume los coeficientes de correlación de Pearson para todas las variables numéricas incluidas en el estudio. Este gráfico permite identificar patrones y relaciones significativas que pueden ser útiles para comprender cómo se conectan las dimensiones de salud de cartera, generación de valor, demografía y uso de productos. Algunas relaciones destacadas son:

Relaciones Positivas Fuertes:

- Existe una alta correlación entre **ValorVida** y **ValorPers** (0.67), al igual que con la variable **Reserva matemática** lo que sugiere que estos valores están estrechamente relacionados en la generación de valor.
- **Saldo aportes** muestra una fuerte correlación positiva con **Antigüedad** (0.82), lo cual es esperable ya que el tiempo de antigüedad impacta directamente en el saldo acumulado de los aportes. Lo mismo ocurre con antigüedad y edad.

Relaciones Negativas Fuertes:

- **EdadCartera** tiene una correlación negativa notable con **Cantidad de productos** (-0.44), lo que podría indicar que, a mayor edad de la cartera, menor es la cantidad de productos.
- **Acierta** tienen una correlación negativa moderada entre las variables **Antigüedad** y **Saldo aportes** lo que podría reflejar diferencias en la forma de evaluar la capacidad de pago y las reservas acumuladas. Adicionalmente, se puede revisar la pertinencia de agregar esta variable ya que en algunos casos la calificación 0 no es del todo una calificación mala, ya que puede indicar que no se tiene el dato.

Baja o Nula Correlación:

- Variables como **CEH_Recre**, **CEH_Fund** y **CEH_Educ** presentan correlaciones bajas con la mayoría de las demás, indicando independencia en términos de eventos históricos y otras dimensiones analizadas.
- **Ingresos** muestra una relación baja con otras variables como **Pde_fuga** y **Edad**, lo que sugiere que no es un determinante directo en estas dimensiones.

Patrones Para Resaltar:

- **CLV** muestra correlaciones positivas moderadas con variables relacionadas con generación de valor, como **Valor Perseverancia** (0.39), al igual que con variables de uso de productos como la **cantidad de productos** (0.45), lo que subraya su relevancia en la evaluación de la conexión integral.

Este análisis no solo permite entender las relaciones entre las variables del modelo, sino que también ayuda a identificar cuáles podrían ser más relevantes para predecir el score de conexión de los asociados, así como potenciales ajustes en el modelado para reducir el impacto de multicolinealidad.

4.3 ANÁLISIS VARIABLES CATEGORICAS

Ahora procedemos a analizar las variables categóricas, inicialmente con las relacionadas con la salud de la cartera.

Estado: Respecto al estado actual de los asociados, el 74% de los asociados se encuentran activos, lo que indica una fuerte participación y un compromiso sostenido con la cooperativa. Por otro lado, un 25% de

los asociados están inactivos, lo que representa una porción significativa que se encuentra desconectada con los productos y servicios ofrecidos por la cooperativa. Finalmente, solo un 1% de los asociados está suspendido, lo cual refleja un nivel relativamente bajo de conflictos o incumplimientos dentro de la cooperativa.

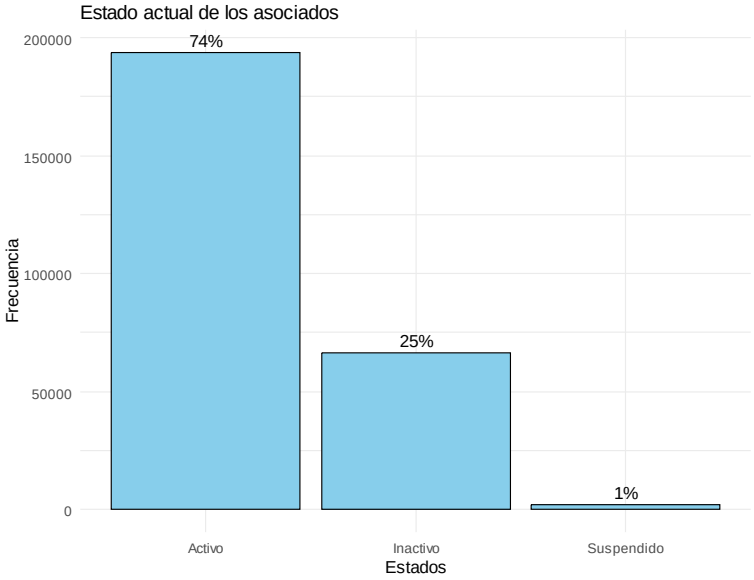

Figura 3. Estado de los asociados.

Alerta comportamiento de pago externo: Respecto a las alertas del comportamiento externo de los asociados se logra identificar que en las 3 variables revisadas el 78% de los no tienen información disponible, en el caso de las alertas de hábito de pago y estado créditos solamente el 3% y 2% cuentan con alertas, mientras que en el caso de la alerta de capacidad de pago el 16% de los asociados cuentan con esta alerta.

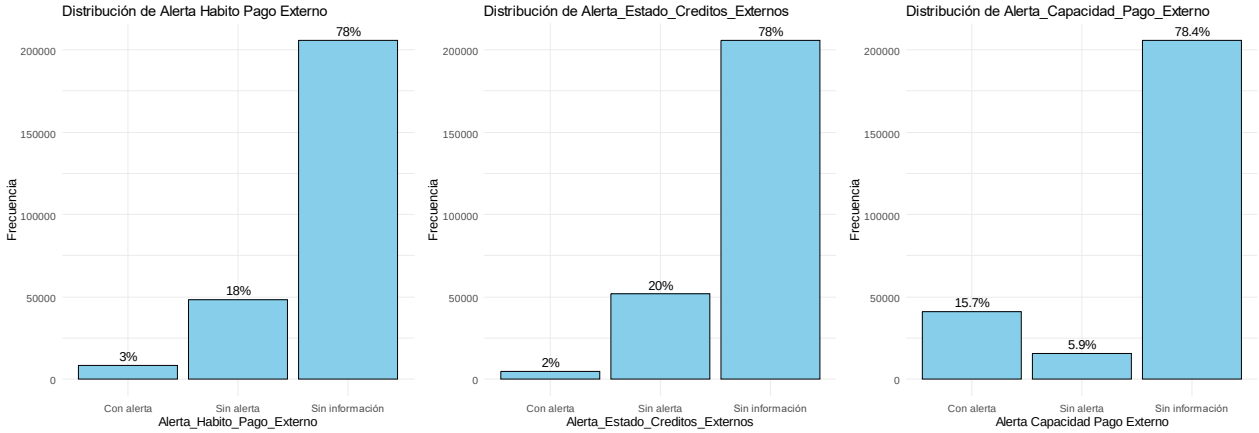

Figura 4. Alertas comportamiento de pago externo.

Débito automático activado: El 40% de los asociados cuentan con el débito automático activado.

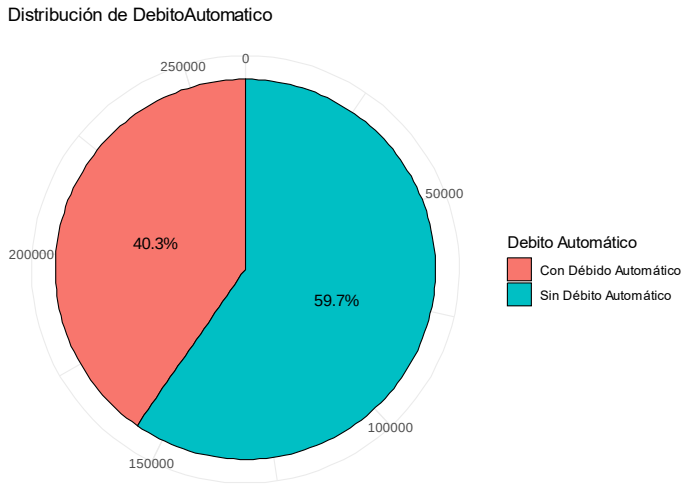

Figura 5. Débito automático.

Segmento Habito Pago Estatutario 3 Meses: Aproximadamente el 51% de los asociados se encuentran bien calificados por la segmentación de hábito de pago de los conceptos estatutarios. Se destaca además que el 17% se encuentran en la categoría **Muy deficiente.**

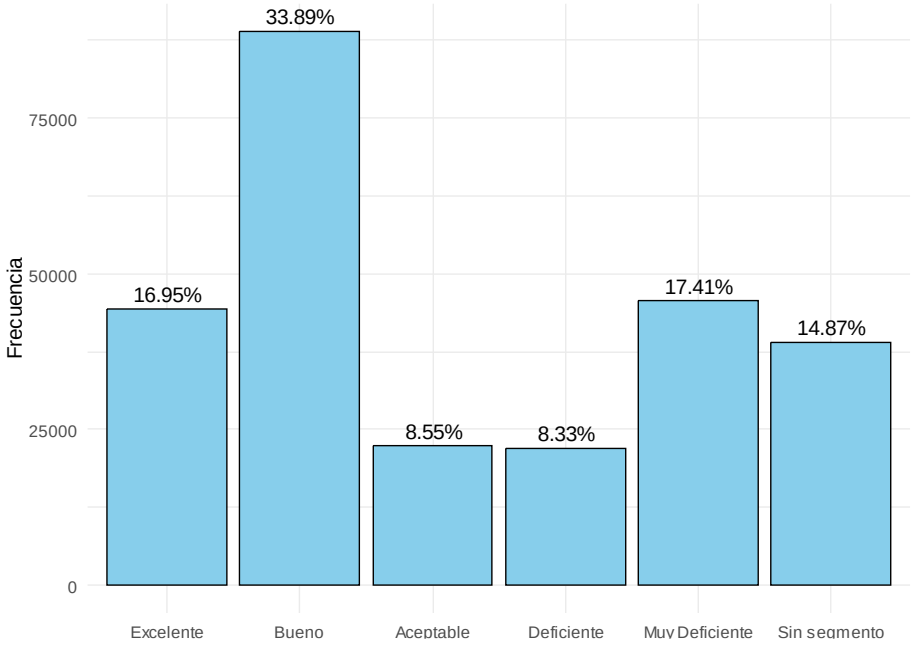

Figura 6. Hábito de pago estatutarios.

Estrato: La mayoría de los registros pertenecen al Estrato Medio (191,266 registros, 73%). Los Estratos Alto y Bajo están subrepresentados con 19% y 8%, respectivamente. Esto sugiere un sesgo hacia individuos de clase media.

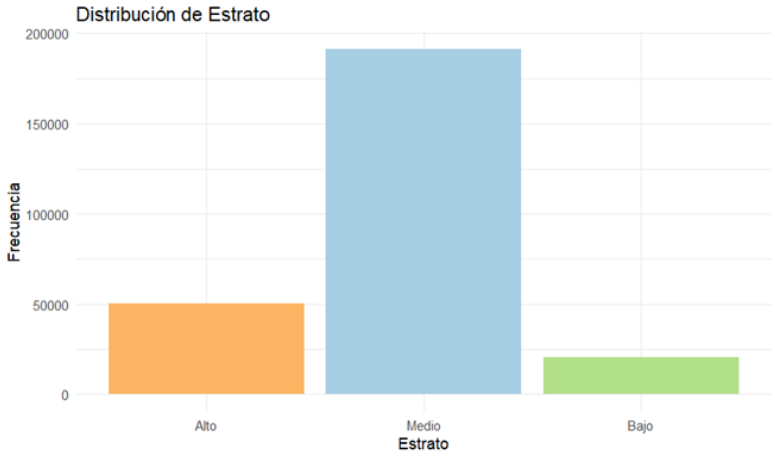

Figura 7. Estrato socioeconómico.

Nivel Académico: El 70% tiene nivel Profesional, lo que indica un enfoque en la vinculación de asociados con educación superior. Los niveles Técnico o Tecnólogo (16%) y Ninguno (4%) son significativamente menores.

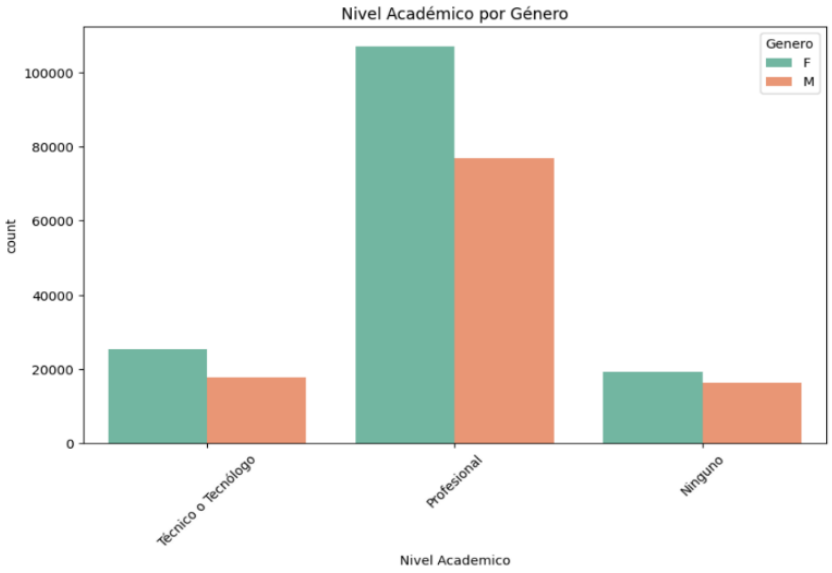

Figura 8. Nivel académico por género.

El gráfico muestra una clara tendencia hacia el nivel académico "Profesional" como el más predominante entre los géneros, especialmente en mujeres, quienes tienen una mayor representación en esta categoría en comparación con los hombres. Sin embargo, en los niveles "Técnico o Tecnólogo" y "Ninguno", aunque las diferencias entre géneros son menos marcadas, sigue habiendo una ligera mayor presencia femenina.

Esto podría reflejar una tendencia hacia mayores niveles de educación formal entre las mujeres en este grupo de datos, mientras que los hombres están menos representados en el nivel profesional y de forma equilibrada en los demás niveles.

Estrato | Nivel académico	Ninguno	Profesional	Técnico o Tecnólogo
Alto	13,0%	**81,0%**	6,0%
Medio	13,8%	67,5%	**18,7%**
Bajo	13,2%	66,0%	**20,8%**

Tabla 2. Estrato vs nivel académico.

El análisis del porcentaje de nivel académico por estrato revela diferencias significativas en la distribución educativa. En el Estrato Alto, el 81% de las personas tienen nivel profesional, lo que refleja un acceso considerable a la educación superior, mientras que los niveles técnicos (6%) y sin estudios formales (13%) son minoritarios. Por el contrario, el Estrato Bajo tiene una menor proporción de personas con nivel profesional (66%), destacando un mayor porcentaje en nivel técnico o tecnólogo (21%) y sin estudios (13%), lo que podría indicar limitaciones en el acceso educativo.

Actividad Laboral: La mayoría son Asalariados (55%) e Independientes (29%), lo que cubre el 84% de los registros. Categorías como Ama de Casa y Estudiante son poco representadas.

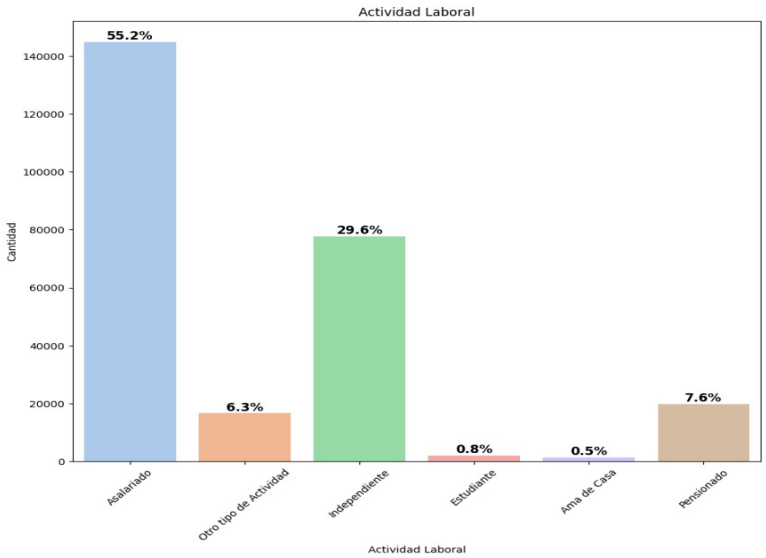

Figura 9. Actividad laboral.

La Figura 9 muestra la distribución porcentual de diferentes actividades laborales. La categoría predominante es "Asalariado", que representa el 55.20% del total, seguida por "Independiente" con un 29.62%. Este análisis sugiere que la mayoría de las personas en este grupo de datos están empleadas en trabajos asalariados, con una significativa proporción trabajando de forma independiente, mientras que las demás ocupaciones tienen una participación minoritaria.

Tipo de Vivienda: El gráfico de pastel muestra la distribución de los tipos de vivienda en la población analizada. La mayoría de las personas, el 65.90% (172,900 individuos), viven en viviendas clasificadas como "Familiar", mientras que el 25.22% (66,156 individuos) poseen una vivienda propia. Un porcentaje menor, el 6.82% (17,886 individuos), vive en viviendas de "Alquiler", y solo el 2.07% (5,424 individuos) pertenece a la categoría "Desconocida". Esto indica que la mayor parte de la población reside en viviendas familiares, seguido de una proporción significativa con viviendas propias, mientras que las opciones de alquiler y desconocida tienen una menor incidencia.

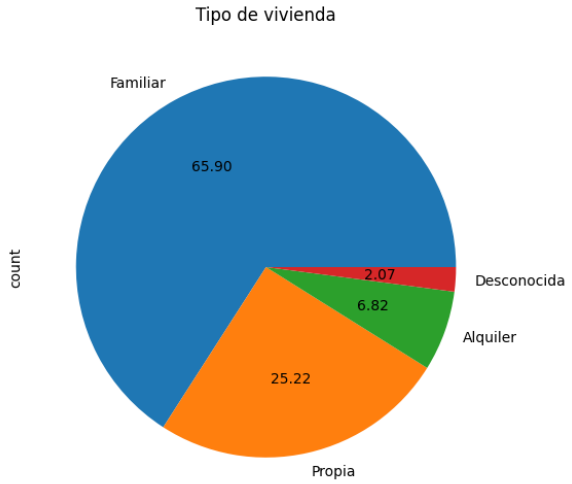

Figura 10. Tipo de vivienda.

las personas con actividad laboral ama de casa y pensionadas se destacan por tener principalmente vivienda propia, en comparación con el resto de las actividades laborales donde predomina la vivienda familiar.

Actividad laboral | Vivienda	Alquiler	Desconocida	Familiar	Propia
Ama de Casa	6,7%	5,2%	36,0%	52,0%
Asalariado	6,7%	2,4%	69,8%	21,1%
Estudiante	7,5%	2,9%	61,9%	27,8%
Independiente	7,1%	1,3%	67,6%	23,9%
Otro tipo de Actividad	4,7%	0,5%	75,8%	19,1%
Pensionado	8,0%	3,9%	24,9%	63,2%

Tabla 3. Actividad laboral versus Tipo de vivienda.

El análisis exploratorio nos permitió conocer el comportamiento de las variables numéricas y categóricas, identificando su distribución, tendencias y posibles inconsistencias en los datos. Por otra parte, permitió evaluar las correlaciones entre las variables, detectando relaciones significativas que pueden influir en la conexión de los asociados con la cooperativa.

5 APLICACIÓN DE TÉCNICAS DE REDUCCIÓN DE DIMENSIONALIDAD *OBJ. ESPECIFICO 2*

5.1 REVISIÓN DE METODOLOGÍAS DE REDUCCIÓN DE DIMENSIONALIDAD

La reducción de dimensionalidad es una técnica fundamental en el análisis de datos, ya que permite transformar un conjunto de variables en un espacio de menor dimensión, conservando la mayor cantidad de información posible [18]. En este estudio, el objetivo era identificar las variables más significativas que influyen en la conexión de los asociados con la cooperativa, reduciendo la redundancia y mejorando la interpretabilidad del modelo [19].

Inicialmente, se evaluaron distintas metodologías de reducción de dimensionalidad. Se consideró el uso de Autoencoder, una técnica basada en redes neuronales que permite aprender una representación comprimida de los datos [20]. Sin embargo, esta metodología no resultó adecuada debido a dos razones principales: en primer lugar, su manejo de variables categóricas no fue tan efectivo, lo que generaba una representación menos precisa de los datos [21]; en segundo lugar, su interpretación era compleja, ya que los Autoencoders operan como una "caja negra", dificultando la comprensión de cómo se generaban los componentes a partir de las variables originales [22].

Posteriormente, se exploró el Dual Multiple Factor Analysis (DMFA), una técnica utilizada para analizar múltiples tablas de datos heterogéneos y encontrar estructuras comunes entre ellas [23]. No obstante, al aplicar DMFA a la base de datos de la cooperativa, se encontró que los primeros componentes no capturaban suficiente variabilidad, lo que dificultaba la representación efectiva de las variables y su impacto en la conexión de los asociados.

Dado que ninguna de estas metodologías cumplía con los requerimientos del estudio, se optó por utilizar el Análisis de Componentes Principales (ACP), una técnica ampliamente utilizada en la reducción de dimensionalidad que permite transformar variables correlacionadas en nuevas variables no correlacionadas llamadas componentes principales [18], [24].

5.2 SELECCIÓN DE LA METODOLOGÍA: ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

El Análisis de Componentes Principales (ACP), es una técnica estadística multivariada ampliamente utilizada para la reducción de dimensionalidad y la extracción de características en conjuntos de datos con múltiples variables correlacionadas. Su objetivo principal es transformar un conjunto original de variables posiblemente correlacionadas en un conjunto más pequeño de variables no correlacionadas llamadas componentes principales, que retienen la mayor cantidad de varianza posible del sistema original [18].

Este proceso resulta particularmente útil en contextos donde se desea simplificar el análisis sin perder información esencial, permitiendo representar estructuras subyacentes complejas de forma más manejable. En el contexto de esta investigación, el ACP fue seleccionado como metodología para identificar las variables más influyentes en la conexión de los asociados con la cooperativa, facilitando el análisis exploratorio posterior y la construcción de modelos más eficientes.

El ACP es una técnica no supervisada, ya que no requiere etiquetas o variables de salida para su aplicación. Su fundamento radica en encontrar direcciones (vectores propios) en el espacio de las variables que capturen la mayor varianza posible del conjunto de datos. Matemáticamente, esto se realiza a través de la descomposición en valores propios de la matriz de covarianza de los datos estandarizados:

$$\boldsymbol{S} = \frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})^{\boldsymbol{T}}$$

donde:
$\boldsymbol{x_i}$representa el vector de observaciones de la muestra i,
$\overline{\boldsymbol{x}}$ es el vector media de las variables,
$\boldsymbol{S}$ es la matriz de covarianza del conjunto de datos.

La descomposición espectral de $\boldsymbol{S}$ permite obtener los autovalores (λ_i) y autovectores (v_i), los cuales definen la varianza explicada por cada componente y su dirección respectiva. Cada componente principal $Z_k$se expresa como una combinación lineal de las variables originales:

$$\boldsymbol{Z_k} = \boldsymbol{v_k^t X}$$

El número de componentes seleccionados se determina evaluando la varianza explicada acumulada, buscando conservar un alto porcentaje (por ejemplo, el 80% o 90%) de la información original con la menor cantidad de componentes posibles [19].

El ACP fue elegido por las siguientes razones:

- Reducción de la complejidad: Permite trabajar con menos variables sin sacrificar demasiada información, lo cual es clave en conjuntos de datos de alta dimensión [25].
- No requiere etiquetas: Ideal para análisis exploratorio donde no se dispone de una variable objetivo clara.
- Mejora la interpretación: Los componentes extraídos permiten observar relaciones entre variables y estructuras latentes en los datos [26].
- Preprocesamiento útil: En muchos casos, el ACP se utiliza como etapa previa a modelos de clasificación, segmentación o clustering, ayudando a eliminar ruido o multicolinealidad.

Etapas del proceso de ACP implementado:

- Estandarización de las variables: Se normalizaron las variables para tener media cero y desviación estándar uno, eliminando el efecto de escalas distintas.
- Cálculo de la matriz de covarianza: Se calculó la matriz de covarianza de los datos estandarizados.
- Descomposición espectral: Se obtuvieron los autovalores y autovectores de la matriz.
- Selección de componentes: Se seleccionaron los componentes que explicaban el mayor porcentaje de la varianza acumulada.
- Proyección de los datos: Se transformaron los datos originales a un nuevo espacio reducido definido por los componentes seleccionados.

Este enfoque ha sido validado ampliamente en investigaciones de análisis exploratorio de datos, especialmente cuando se busca detectar patrones ocultos, agrupar observaciones o reducir el ruido en sistemas multivariados complejos [24]. A continuación, se detallan las etapas seguidas para la implementación del ACP en el análisis de los datos:

5.2.1 EVALUACIÓN DE LA ADECUACIÓN DEL ACP: TEST DE KMO

Como parte del análisis exploratorio y previo a la aplicación del Análisis de Componentes Principales (ACP), se calculó el índice de Kaiser-Meyer-Olkin (KMO). Este índice evalúa la adecuación de los datos para técnicas de reducción de dimensionalidad, midiendo qué tanto de la varianza puede atribuirse a una estructura común entre variables.

El valor global obtenido fue de **0.67**, lo cual, según la escala de Kaiser, se interpreta como **aceptable**. Esto sugiere que, si bien el conjunto de datos es adecuado para aplicar ACP, no se encuentra en un nivel óptimo.

A nivel individual, la mayoría de las variables presentó un MSA (Measure of Sampling Adequacy) superior a 0.5, destacándose aquellas con mayor adecuación como IRP_3M (0.97), CVencidasE (0.93), y SaldoAportes (0.91). No obstante, algunas variables mostraron valores bajos, por ejemplo, ALaboral.Otro.tipo.de.Actividad (0.10) o SHPE_3M.Excelente (0.39), lo cual indica baja participación en la estructura común de correlación. Este aspecto fue considerado posteriormente para evaluar su aporte a los componentes principales.

En conjunto, el índice de KMO refuerza la decisión metodológica de aplicar ACP, aportando una validación cuantitativa a la estructura subyacente en el conjunto de datos.

5.2.2 PREPROCESAMIENTO DE LOS DATOS:

- **Transformación de variables categóricas**: Las variables categóricas fueron transformadas mediante one-hot encoding, una técnica que convierte cada categoría en una nueva variable binaria (0 o 1). Esto permitió incorporar las variables categóricas en el análisis sin generar sesgos en las relaciones entre las categorías. Al transformar las variables de esta manera, se asegura que la información relevante se conserve sin distorsionar la estructura de los datos.
- **Eliminación de variables con baja variabilidad**: Se descartaron las variables cuya varianza era muy baja, ya que estas no aportaban suficiente diferenciación entre las observaciones y, por ende, no tenían impacto significativo en la construcción del modelo. En este paso, el conjunto de datos pasó de 38 a 33 variables, eliminando las variables Colaborador.1, Vivienda.desconocida, ALaboral.Estudiante, Estado.Suspendido y Pde_Fuga.
- **Revisión de variables correlacionadas y exclusión de la variable "Acierta":** Aunque en el análisis se identificaron variables con alta correlación, no todas fueron eliminadas, ya que su inclusión aportaba información valiosa y complementaria para el estudio. Sin embargo, se decidió excluir la variable *Acierta* por dos razones principales: en primer lugar, su calificación de cero no

necesariamente representa un mal desempeño, sino que, en algunos casos, indica ausencia de información; en segundo lugar, presentaba una correlación negativa moderada con variables como *Antigüedad* y *Saldo aportes*, lo que podía introducir ruido en el análisis. Esta decisión permitió mantener la calidad del análisis sin eliminar información potencialmente útil de otras variables correlacionadas.

- **Escalamiento de variables numéricas**: Para evitar que las variables con diferentes escalas de medición afectaran de manera desigual el análisis, todas las variables numéricas fueron **escaladas** a una distribución estándar, con **media 0** y **varianza 1**. Esto asegura que todas las variables contribuyan de manera equitativa al análisis de los componentes principales, evitando que las variables con mayor rango de valores dominen el proceso.

5.2.3 CÁLCULO DE LOS COMPONENTES PRINCIPALES:

Una vez preprocesados los datos, se aplicó el Análisis de Componentes Principales sobre la base de datos procesada. Este procedimiento genera un conjunto de componentes principales ordenados según la cantidad de varianza que explican en los datos. Los primeros componentes capturan la mayor parte de la variabilidad, lo que es fundamental para entender las relaciones subyacentes en el conjunto de datos.

5.2.4 ANÁLISIS DE LA VARIANZA EXPLICADA:

En este paso, se evaluaron los componentes principales obtenidos para seleccionar aquellos que explicaban un porcentaje significativo de la varianza total. Este análisis permite determinar cuántos componentes son necesarios para representar adecuadamente la estructura de los datos. Generalmente, se busca un umbral de varianza explicada acumulada (por ejemplo, 80-95%) para garantizar que se conserve la mayor parte de la información relevante en el conjunto reducido de componentes.

5.2.5 EVALUACIÓN DE LAS CARGAS FACTORIALES:

Las cargas factoriales representan la correlación entre las variables originales y los componentes principales, es decir, cuánto influye cada variable en cada componente. Se analizó la contribución de cada variable a los componentes principales, identificando las variables que tenían mayor peso en la estructura latente de los datos. Estas variables fueron consideradas las más relevantes en la identificación de patrones y conexiones de los asociados con la cooperativa.

El objetivo principal de este proceso fue reducir la dimensionalidad del conjunto de datos sin perder información clave, permitiendo así una identificación más eficiente de las variables más significativas que influyen en la relación de los asociados con la cooperativa. Este enfoque no solo mejora la interpretabilidad del modelo, sino que también facilita la construcción de modelos predictivos o descriptivos más eficientes, al centrarse en las características más relevantes y eliminando redundancias.

El ACP, al reducir la cantidad de dimensiones necesarias para describir los datos, proporciona una visión clara y simplificada de las relaciones y permite a los analistas identificar las variables que tienen un impacto real en la estructura latente de los datos.

5.3 RESULTADOS DE APLICACIÓN ACP.

Uno de los primeros aspectos a evaluar en el Análisis de Componentes Principales (ACP) es la cantidad de variabilidad que logran capturar los primeros componentes. La **inercia acumulada**, que representa el porcentaje de variabilidad explicada de manera acumulativa a medida que se incluyen más componentes, se presenta en la siguiente tabla para los primeros 10 componentes:

Componentes	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Desviación Estándar	2,62	1,68	1,54	1,37	1,32	1,27	1,18	1,16	1,11	1,08
% de varianza	20,9%	8,6%	7,2%	5,7%	5,2%	4,9%	4,2%	4,0%	3,7%	3,6%
% Varianza Acumulada	20,9%	29,4%	36,7%	42,4%	47,6%	52,5%	56,7%	60,8%	64,5%	68,1%

Tabla 4. Porcentaje de varianza acumulado en componentes.

Se observa que los primeros 10 componentes capturan el 68.07% de la variabilidad total de los datos, lo que indica que el ACP logra una reducción significativa de la dimensionalidad al concentrar una gran proporción de la información en un número reducido de componentes. Sin embargo, los dos primeros componentes explican únicamente el **29.44%** de la variabilidad, lo que sugiere que la estructura de los datos es compleja y que se requiere un mayor número de componentes para describir adecuadamente la información. A pesar de esto, se utilizarán estos dos primeros componentes para comprender la estructura de los datos y la relación entre variables, permitiendo visualizar agrupaciones, correlaciones y posibles patrones que ayuden a interpretar la contribución de cada variable en la variabilidad explicada. No obstante, para el cálculo del score se tomarán los componentes necesarios que permitan capturar una mayor proporción de la variabilidad, garantizando así un resultado más robusto y representativo de la conexión de los asociados con la cooperativa.

El siguiente paso en el análisis consiste en examinar las cargas factoriales de las variables en los componentes seleccionados para identificar cuáles tienen mayor peso en la estructura de los datos y, por ende, una mayor influencia en la conexión de los asociados con la cooperativa.

Componentes	PC1	PC2
IRP_3M	0,16906	0,17386
CVencidasE	-0,22126	-0,37484
EdadCartera	-0,26895	-0,34755
ValorVida	0,12864	-0,19841
ValorPers	0,14789	-0,17439
ReservaM	0,20155	-0,23529
SaldoAportes	0,32688	-0,07540
Ingresos	0,11449	-0,02768
CLV	0,18610	-0,04059
Edad	0,25038	-0,13395
Antigüedad	0,29534	-0,20672

CEH_Recre	0,11671	-0,01373
CEH_Educ	0,12453	-0,04981
CEH_Fund	0,08193	-0,04516
CProductos	0,24391	0,05494
Estado.Inactivo	-0,25978	-0,34442
DAutomatico.Sin.Débito.Automático	-0,17963	0,01204
SHPE_3M.Bueno	0,15185	0,08594
SHPE_3M.Deficiente	-0,02316	-0,05182
SHPE_3M.Excelente	0,13231	0,01072
SHPE_3M.Muy.Deficiente	-0,21240	-0,36766
SHPE_3M.Sin.segmento	-0,12772	0,31427
Genero.M	0,01097	-0,04419
Estrato.Bajo	-0,01430	-0,16462
Estrato.Medio	-0,12605	0,21373
NAcademico.Profesional	0,09530	-0,05529
NAcademico.Técnico.o.Tecnólogo	-0,11510	0,06462
ALaboral.Asalariado	-0,03039	0,05656
ALaboral.Independiente	-0,00636	-0,04947
ALaboral.Otro.tipo.de.Actividad	-0,05092	-0,01285
ALaboral.Pensionado	0,10900	-0,01030
Vivienda.Familiar	-0,25775	0,18035
Vivienda.Propia	0,22034	-0,15339

Tabla 5. Carga factorial variables en componentes.

5.3.1 ANÁLISIS DEL PRIMER COMPONENTE PRINCIPAL (PC1):

Representa una combinación de factores que reflejan la estabilidad financiera y el nivel de conexión de los asociados con la cooperativa, permitiendo una diferenciación entre aquellos con buen comportamiento financiero y aquellos que enfrentan problemas de pago. Este componente se interpreta como una medida general de la relación positiva o negativa de los asociados con la cooperativa, considerando tanto aspectos financieros como de comportamiento.

- Variables con mayor peso positivo (mayor influencia en el componente):
  - SaldoAportes (0.32688): Indica que los asociados con mayores aportes tienen una fuerte influencia positiva en este componente.
  - Antigüedad (0.29534) y Edad (0.25038): Asociados con mayor tiempo en la cooperativa también tienen un impacto positivo.
  - Cantidad de productos (CProductos - 0.24391) y Vivienda Propia (0.22034): Sugiere que los asociados con más productos y vivienda propia están representados positivamente en PC1.
  - IRP_3M (0.16906), CLV (0.18610) y ReservaM (0.20155): También tienen cargas positivas,

lo que podría sugerir que este componente representa estabilidad y conexión con la cooperativa.

- Variables con mayor peso negativo (mayor influencia en sentido opuesto):
    - Estado.Inactivo (-0.25978) y CVencidasE (-0.22126): Indican que los asociados con estados inactivos o con vencimientos de crédito contribuyen en sentido negativo.
    - SHPE_3M.Muy.Deficiente (-0.21240) y EdadCartera (-0.26895): Relacionadas con mal comportamiento de pago, lo que sugiere que este componente diferencia entre asociados con buena y mala conexión.
    - Vivienda Familiar (-0.25775): Puede indicar una menor estabilidad financiera en comparación con tener vivienda propia.

El PC1 refleja una combinación de factores que permiten identificar a los asociados con una buena estabilidad financiera y una conexión fuerte con la cooperativa, contrastando con aquellos que muestran signos de inactividad o problemas de pago. Este componente es útil para diferenciar entre los asociados con buen comportamiento financiero y los que podrían estar enfrentando dificultades económicas. Las variables con cargas positivas indican características como altos aportes, mayor antigüedad, y estabilidad financiera, mientras que las cargas negativas destacan factores asociados con comportamiento deficiente en términos de pagos y compromiso con la cooperativa.

5.3.2 ANÁLISIS DEL SEGUNDO COMPONENTE PRINCIPAL (PC2):

El Segundo Componente Principal (PC2) captura información adicional sobre la segmentación y los perfiles de los asociados. Representa otro eje de variabilidad en los datos, proporcionando una dimensión complementaria a la del primer componente. Mientras que el PC1 se centra en la estabilidad financiera y la conexión con la cooperativa, el PC2 se enfoca más en aspectos relacionados con la segmentación de los asociados y sus perfiles característicos.

- Representa otro eje de variabilidad en los datos.
- Variables con mayor peso positivo:
    - SHPE_3M.Sin.segmento (0.31427) y Estrato.Medio (0.21373): Podrían indicar una separación de los asociados que no tienen una clasificación clara.
    - Vivienda Familiar (0.18035): Puede estar relacionada con un perfil diferente al de los asociados con vivienda propia.
- Variables con mayor peso negativo:
    - CVencidasE (-0.37484), Estado.Inactivo (-0.34442) y SHPE_3M.Muy.Deficiente (-0.36766): De nuevo resaltan los asociados con problemas de pago o estado inactivo.
    - EdadCartera (-0.34755): Refuerza la idea de que este componente también capta información sobre estabilidad financiera.

El PC2 se interpreta como un eje adicional de segmentación y perfilado de los asociados. A diferencia del PC1, que se centra en la estabilidad financiera y la conexión con la cooperativa, el PC2 captura información sobre la clasificación socioeconómica de los asociados y su comportamiento en términos de pago. Las variables con carga positiva indican perfiles menos definidos o asociados con una clasificación intermedia, como los que no tienen una segmentación clara o aquellos con vivienda familiar. En cambio, las variables con carga negativa resaltan a los asociados con problemas de pago, inactividad o dificultades financieras,

sugiriendo que este componente también está relacionado con la estabilidad económica y el compromiso financiero de los asociados.

Por otro lado, tenemos La nube de variables, esta es un espacio de menor dimensión la cual conserva al máximo las correlaciones originales entre ellas. El ángulo entre cada par de flechas representa la correlación entre esas dos variables. Ángulos pequeños significan una alta correlación positiva, ángulos cercanos a 90° ausencia de correlación y ángulos cercanos a 180° una alta correlación negativa.

La Figura 11 muestra el círculo de correlaciones en función de la contribución a la formación de los ejes de cada variable.

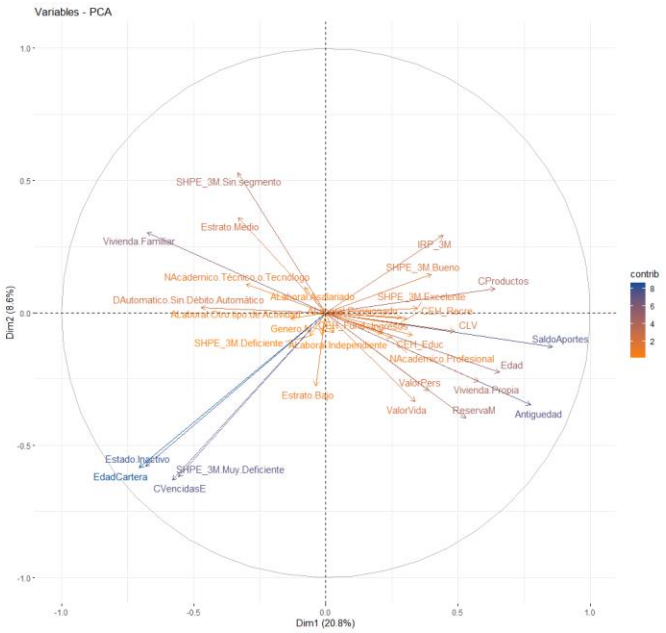

Figura 11. Círculo de correlación.

De la figura anterior podemos interpretar que:

- Grupo en la parte derecha (Dim1 positivo): Variables como Ingresos, CLV, CProductos, CEH_Educ están fuertemente correlacionadas con Dim1. Esto sugiere que clientes con mayor ingreso y educación tienen una mejor relación con la cooperativa.
- Grupo en la parte izquierda (Dim1 negativo): Variables como Estado Inactivo, EdadCartera y CVencidasE, Estrato Bajo están en la dirección opuesta. Esto sugiere que clientes con historial de cartera vencida y menor interacción con la cooperativa tienen valores negativos en Dim1.
- Grupo en la parte superior (Dim2 positivo): Variables como SHPE_3M_SinSegmento, Vivienda Familiar y Estrato Medio. Pueden representar clientes con un perfil intermedio en términos de relación con la cooperativa.

Ahora procedemos a analizar el gráfico de representación simultánea, esta nos permite interpretar a los individuos (asociados) en términos de similitudes de comportamiento con respecto a las variables:

Figura 12. Representación simultánea.

Podemos visualizar conjuntos de asociados en los 3 grupos identificados en el análisis de las variables. Asociados con buen comportamiento en cuanto a hábitos de pago y uso de productos, asociados con un nivel intermedio de conexión y asociados con poca conexión al tener hábitos de pago deficientes, estar inactivos y con cuotas vencidas.

6 CALCULO SCORE INTEGRAL DE CONEXIÓN OBJETIVO ESPECIFICO 3.

El Score Integral de Conexión es una métrica compuesta que permite evaluar la relación global de un asociado con la cooperativa, basándose en los componentes principales extraídos del análisis PCA. Este score integra las diferentes dimensiones identificadas en el PCA en un único valor que permite clasificar a los asociados según su grado de conexión y compromiso con la cooperativa.

A continuación, se describe el proceso para calcular el Score Integral de Conexión:

6.1 OBTENCIÓN DE LOS COMPONENTES PRINCIPALES

Tras realizar el **Análisis de Componentes Principales (PCA)** sobre el conjunto de datos preprocesados, se obtuvieron los primeros **componentes principales** $(PC_1, PC_2, \ldots., PC_k)$ que explican la mayor parte de la variabilidad presente en los datos. Estos componentes representan combinaciones lineales de las variables originales y reflejan las dimensiones más relevantes relacionadas con la conexión de los asociados con la cooperativa.

6.2 PROYECCIONES SOBRE LOS COMPONENTES PRINCIPALES

Para cada asociado, se calculó un score sintético a partir de las proyecciones sobre los componentes

principales seleccionados. Primero, se tomó la matriz de observaciones proyectadas sobre los componentes principales $X_{PC1:PCk}$, obtenida a partir de la descomposición PCA del conjunto de datos estandarizado. Luego, cada proyección fue ponderada multiplicándola por la desviación estándar correspondiente a su componente S_{PCi}, con el fin de reflejar la importancia relativa de cada dimensión en la variabilidad total de los datos. El score de cada asociado se calculó de la siguiente forma:

$$\boldsymbol{Score_k = X_{PC1:PCk} \times S_{PC1:PCk}}$$

donde:

- $X_{PC1:PCk}$ es la matriz de coordenadas de los asociados en el espacio definido por los primeros *k* componentes principales,
- $S_{PC1:PCk}$ es el vector de desviación estándar de los componentes principales seleccionados.

De acuerdo con Becerra Avella (2010), el índice sintético generalmente se construye a partir de la primera componente principal [27]. Sin embargo, en el presente estudio se observó que la primera componente no explicaba una proporción suficiente de la variabilidad total del conjunto de datos, por lo que se decidió incluir componentes adicionales. Esta decisión busca garantizar que el score sintetice adecuadamente la información contenida en las distintas dimensiones relevantes para el análisis.

6.3 NORMALIZACIÓN Y AJUSTE FINAL DEL SCORE

Para mejorar la interpretabilidad del score, se aplicó una **normalización min-max**, transformándolo a una escala de 0 a 100.

La normalización se realizó mediante la siguiente fórmula:

$$score_normalizado = \left(\frac{score - \boldsymbol{min}(score)}{\mathbf{max}(score) - \mathbf{min}\,(score)}\right) \times 100$$

Esta transformación asegura que:

- El valor mínimo del score corresponda a 0,
- El valor máximo corresponda a 100.

De esta forma, se facilita la interpretación y comparación entre asociados, siendo los valores más altos indicativos de un mayor nivel de conexión y compromiso con la cooperativa.

6.4 SELECCIÓN DEL NÚMERO DE COMPONENTES PRINCIPALES PARA EL SCORE

6.4.1 CRITERIOS PARA LA SELECCIÓN DEL NÚMERO DE COMPONENTES PRINCIPALES

La elección del número óptimo de componentes principales a incluir en el Score Integral de Conexión se basó en un análisis combinado de técnicas estadísticas y criterios de parsimonia.

Inicialmente, se calcularon distintos índices sintéticos utilizando:

- Únicamente el primer componente principal (PC1),
- La combinación de las tres primeras componentes (PC3),
- Las cinco primeras componentes (PC5),
- Y las diez primeras componentes principales (PC10).

Cada índice se obtuvo proyectando los datos estandarizados sobre las componentes seleccionadas y ponderando dichas proyecciones según la importancia relativa de cada componente (medida a través de sus desviaciones estándar), para luego ser normalizados a una escala de 0 a 100.

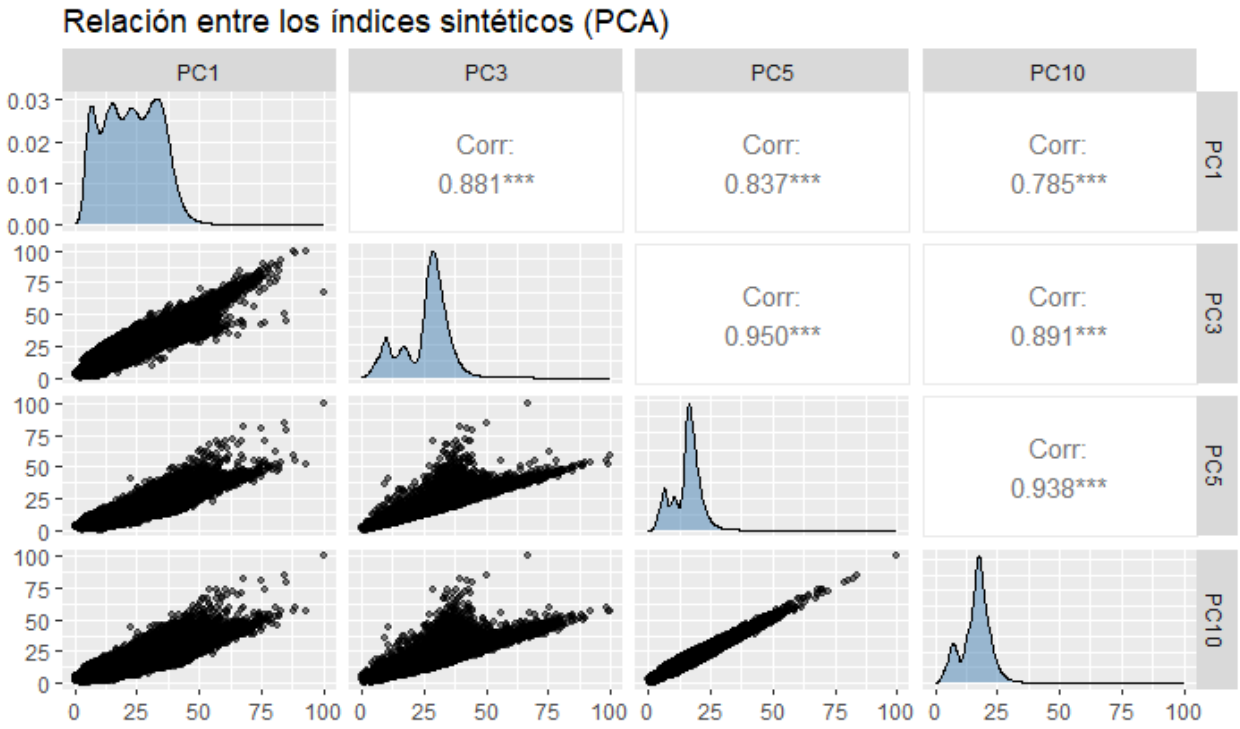

Figura 13 Correlación entre índices sintéticos

Con el fin de comparar estos índices y determinar cuál reflejaba de manera más adecuada la información del conjunto de datos original, se llevaron a cabo los siguientes análisis a partir de la Figura 12:

- **Matriz de Correlaciones:** La matriz de correlaciones de Pearson entre los diferentes índices (PC1, PC3, PC5 y PC10) muestra que todos los índices presentaban correlaciones positivas y elevadas entre sí, indicando que capturaban información similar. En particular, el índice PC5 mostró una correlación muy alta tanto con PC3 (r = 0,95) como con PC10 (r = 0,94), lo que sugiere que el uso de cinco componentes permite una representación robusta y muy cercana a la obtenida con diez componentes, pero con una menor complejidad.
- **Análisis Gráfico:** Los diagramas de dispersión entre los índices mostraron una alineación casi lineal entre PC5 y PC10, reforzando la evidencia de que utilizar cinco componentes es suficiente para

preservar la estructura de la información contenida en el espacio de diez componentes. Por su parte, los histogramas de densidad ubicados en la diagonal ofrecieron una visión de la distribución marginal de cada índice.

- **Criterio de Parsimonia:** Dado que aumentar el número de componentes implica un incremento en la complejidad del modelo y una potencial pérdida de interpretabilidad, se priorizó el principio de parsimonia. Así, se seleccionó el número mínimo de componentes que permitiera explicar una proporción adecuada de la varianza total sin sacrificar representatividad.

Con base en estos análisis, se decidió utilizar las cinco primeras componentes principales ($Score_5$) para el cálculo del Score Integral de Conexión, ya que ofrecen un equilibrio óptimo entre simplicidad, poder explicativo y representación fiel de la información original, adicionalmente con 5 componentes se recoge casi el 50% de la varianza acumulada de los datos.

6.4.2 JUSTIFICACIÓN DE LA NO SELECCIÓN DE OTROS NÚMEROS DE COMPONENTES

Aunque el uso exclusivo del primer componente principal (PC1) habría simplificado aún más el modelo, esta opción fue descartada debido a que la correlación de PC1 con los índices construidos con más componentes, aunque alta (r ≈ 0,88 con PC3 y r ≈ 0,84 con PC5), era inferior a la obtenida entre índices que consideraban más componentes. Esto indica que PC1, aunque captura la mayor proporción individual de varianza, no es suficiente por sí solo para representar adecuadamente la complejidad de las múltiples dimensiones presentes en los datos.

Por otro lado, aunque utilizar diez componentes principales (PC10) permitía una representación aún más completa de la varianza total, se consideró que la ganancia marginal en la calidad del índice no justificaba la pérdida de interpretabilidad ni el aumento de la complejidad analítica. Además, los altos niveles de correlación observados entre PC5 y PC10 (r ≈ 0,94) indicaron que la información adicional capturada por las componentes del sexto al décimo era redundante respecto a la capturada por las primeras cinco componentes.

En consecuencia, la elección de cinco componentes principales se fundamentó en lograr un balance entre:

- Representatividad adecuada de la varianza
- Simplicidad interpretativa
- Reducción del riesgo de sobreajuste
- Coherencia con los principios de parsimonia en la modelación estadística.

7 VISUALIZACIÓN E INTERPRETACIÓN SCORE INTEGRAL DE CONEXIÓN OBJ ESPECIFICO 4

Con el propósito de facilitar la interpretación de los resultados obtenidos en el cálculo del score integral de conexión, en este capítulo se presentan diversas visualizaciones que permiten analizar la distribución de los asociados en función de su nivel de conexión con la cooperativa Coomeva. Además, se identifican los segmentos de asociados con baja conexión y los factores clave que afectan su relación con la organización.

Inicialmente se realizó una revisión de la distribución del **Score Integral de Conexión** con el objetivo de comprender cómo se encontraban agrupados inicialmente los asociados según su nivel de conexión con la cooperativa.

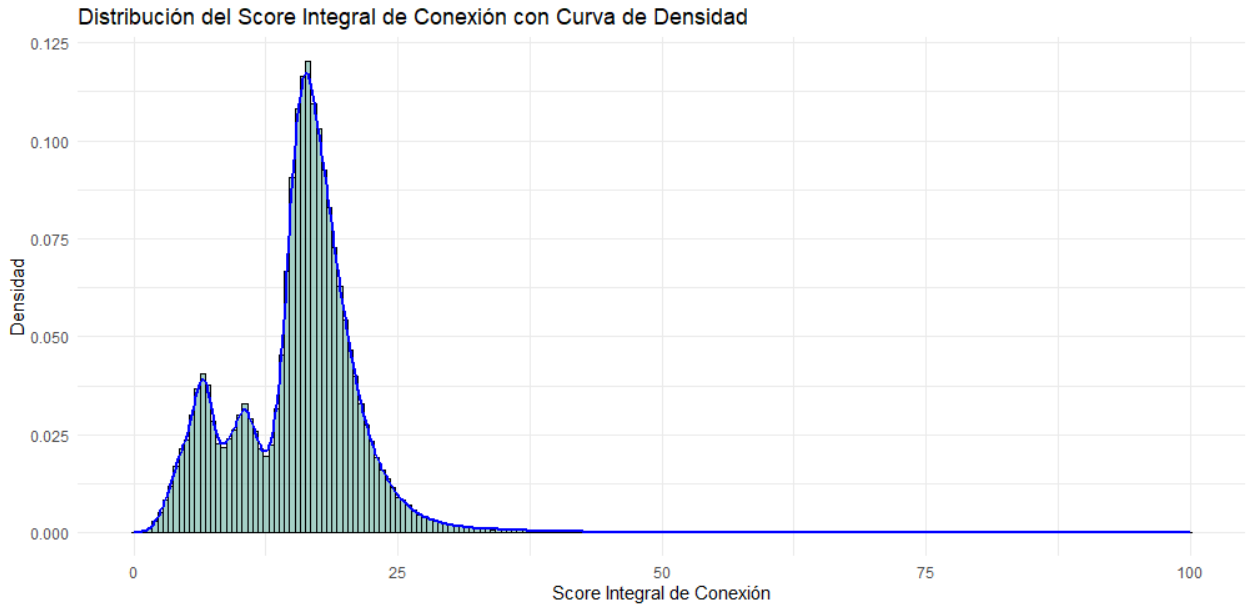

Figura 14 Distribución del Score

En el gráfico de la Figura 14 se evidencia que la mayoría de los asociados presenta scores bajos a intermedios, concentrándose principalmente entre los 6 y 20 puntos. La distribución mostró una forma asimétrica positiva, con una alta densidad de asociados en valores relativamente bajos del score y una disminución progresiva hacia valores más altos.

Además, se identificaron varios picos secundarios en la distribución, lo que sugiere la posible existencia de subgrupos de asociados con distintos niveles de conexión. En particular, se evidenció que son pocos los asociados que alcanzan scores cercanos al máximo de 100 puntos, indicando que una conexión integral elevada es un fenómeno poco común dentro de la población estudiada.

Estos hallazgos resaltan la importancia de realizar una segmentación adecuada, que permita diferenciar claramente a los asociados con baja, media y alta conexión, facilitando así el diseño de estrategias de intervención específicas orientadas a fortalecer la relación de los asociados con la cooperativa.

7.1 SEGMENTACIÓN ASOCIADOS SEGUN SCORE

Con base en la distribución observada del Score Integral de Conexión, se procedió a realizar una segmentación de los asociados utilizando cuartiles. Esta metodología permite dividir la población en cuatro grupos de tamaño similar, facilitando la interpretación y comparación de los niveles de conexión.

Los cuartiles se definieron de la siguiente manera:

- **Primer cuartil (Q1)**: Asociados con los scores más bajos, representando el grupo con menor nivel de conexión con la cooperativa.
- **Segundo cuartil (Q2)**: Asociados con un nivel de conexión ligeramente superior, pero aún por debajo del promedio general.
- **Tercer cuartil (Q3)**: Asociados con un buen nivel de conexión, por encima del promedio.
- **Cuarto cuartil (Q4)**: Asociados con los scores más altos, indicando una conexión sólida e integral con la cooperativa.

Esta segmentación permite identificar no solo a los asociados con baja conexión, en quienes sería prioritario enfocar acciones de fortalecimiento de la relación, sino también reconocer aquellos perfiles de alta conexión, que pueden servir como referencia para diseñar estrategias de retención y fidelización.

A continuación, se presenta el análisis detallado de cada segmento, junto con las principales características que los distinguen.

Tabla 6 Cuartiles del Score con cantidad de asociados

Segmento	Mínimo	Máximo	N° Asociados	Porcentaje
Q1 - Baja Conexión	0	12.3	65.592	25
Q2 - Conexión Moderada	12.3	16.2	65.591	25
Q3 - Buena Conexión	16.2	18.7	65.591	25
Q4 - Alta Conexión	18.7	100	65.592	25

Una vez segmentados los asociados en cuartiles según su **Score Integral de Conexión**, se procede a caracterizar cada uno de los grupos con base en las variables numéricas relevantes del estudio. Esta caracterización permite identificar las diferencias sustanciales entre los niveles de conexión y facilitar la comprensión de los factores asociados al compromiso de los asociados con la cooperativa.

La Tabla 7 muestra el valor promedio de cada variable numérica dentro de cada cuartil. En la visualización, los cuadros en tonalidades rojas representan valores que son considerados menos favorables para la cooperativa, mientras que los cuadros en tonalidades verdes indican valores positivos o deseables. Esta codificación por color facilita la identificación visual de los patrones que distinguen a los grupos con mayor o menor nivel de conexión.

Tabla 7 Promedios de variables numéricas por cuartiles de conexión

Segmento	Q1 - Baja Conexión	Q2 - Conexión Moderada	Q3 - Buena Conexión	Q4 - Alta Conexión
Pde_fuga	0,27	0,23	0,22	0,21
IRP_3M	0,04	0,62	0,91	1,06
CVencidasE	26,78	1,24	0,40	0,33
EdadCartera	6,58	1,10	0,40	0,32
ValorVida	$ 13.686.431	$ 10.341.139	$ 15.489.259	$ 36.309.947
ValorPers	$ 19.811.730	$ 16.779.062	$ 26.956.768	$ 78.438.044

ReservaM	$ 6.485.423	$ 7.105.198	$ 12.032.230	$ 39.924.235
SaldoAportes	$ 129.075	$ 2.801.658	$ 4.989.170	$ 7.931.666
Ingresos	$ 4.285.776	$ 4.992.536	$ 6.580.421	$ 9.251.873
CLV	$ 165.039	$ 262.113	$ 400.866	$ 883.894
Edad	40	48	50	53
Antiguedad	6	10	13	19
CEH_Recre	7,46	11,57	22,02	73,12
CEH_Educ	0,32	0,41	0,76	2,53
CEH_Fund	0,21	0,19	0,36	1,67
CProductos	1,38	2,70	4,89	8,14

A continuación, se destacan los hallazgos más relevantes:

- **Probabilidad de fuga (Pde_fuga):** Los asociados con la conexión mas baja tienen 6 puntos porcentuales más alta la probabilidad de retirarse en comparación con los asociados con Alta conexión.
- **Recaudo (IRP_3M):** Los asociados en el cuartil más bajo tienen un valor promedio de solo 0.04, mientras que en el Q4 este valor se eleva a 1.06, lo que indica un mayor recaudo reciente en el grupo de alta conexión.
- **Morosidad:** Los indicadores de cartera vencida (CVencidasE y EdadCartera) son mucho más altos en los segmentos de baja conexión. Por ejemplo, el Q1 presenta una media de 26,78 cuotas vencidas, en contraste con solo 0,33 en el Q4.
- **Ingresos y rentabilidad:** El ingreso promedio se incrementa de 4,2 millones en Q1 a más de 9,2 millones en Q4. A su vez, el CLV también crece progresivamente, alcanzando su valor más alto (más de $883 mil) en el cuartil de mayor conexión.
- **Participación en programas:** El uso de beneficios como recreación, educación y Fundación es considerablemente mayor en los asociados con alta conexión, indicando un mayor nivel de integración con las iniciativas de la cooperativa.
- **Antigüedad y edad:** Se observa que los asociados más conectados tienden a ser de mayor edad y antigüedad, lo que sugiere una mayor permanencia e involucramiento con el tiempo.
- **Productos y servicios utilizados:** El número promedio de productos (CProductos) aumenta de forma consistente desde el Q1 (1,38 productos) hasta el Q4 (8,14 productos), lo que sugiere que la tenencia de productos está directamente relacionada con el nivel de conexión del asociado.

Se evidencian que los resultados son consistentes con la lógica del score integral de conexión, ya que los asociados ubicados en el cuartil más alto (Q4 - Alta Conexión) presentan, en general, características más favorables para la cooperativa, como menores probabilidades de fuga, mejor comportamiento de pago, mayor tenencia y uso de productos, así como mayores niveles de ingresos y valor económico. Por el contrario, los asociados en el cuartil más bajo (Q1 - Baja Conexión) exhiben valores asociados a mayor

riesgo o desvinculación.

A continuación, se realiza el análisis de las variables categóricas con el objetivo de identificar diferencias estructurales en la composición de los grupos según su nivel de conexión integral. Las variables analizadas incluyen aspectos demográficos, educativos, laborales y del vínculo con la entidad.
La Figura 15 muestra la distribución proporcional de las categorías de cada variable, desagregadas por los cuartiles de conexión: Q1 (Baja Conexión), Q2 (Conexión Moderada), Q3 (Buena Conexión) y Q4 (Alta Conexión).

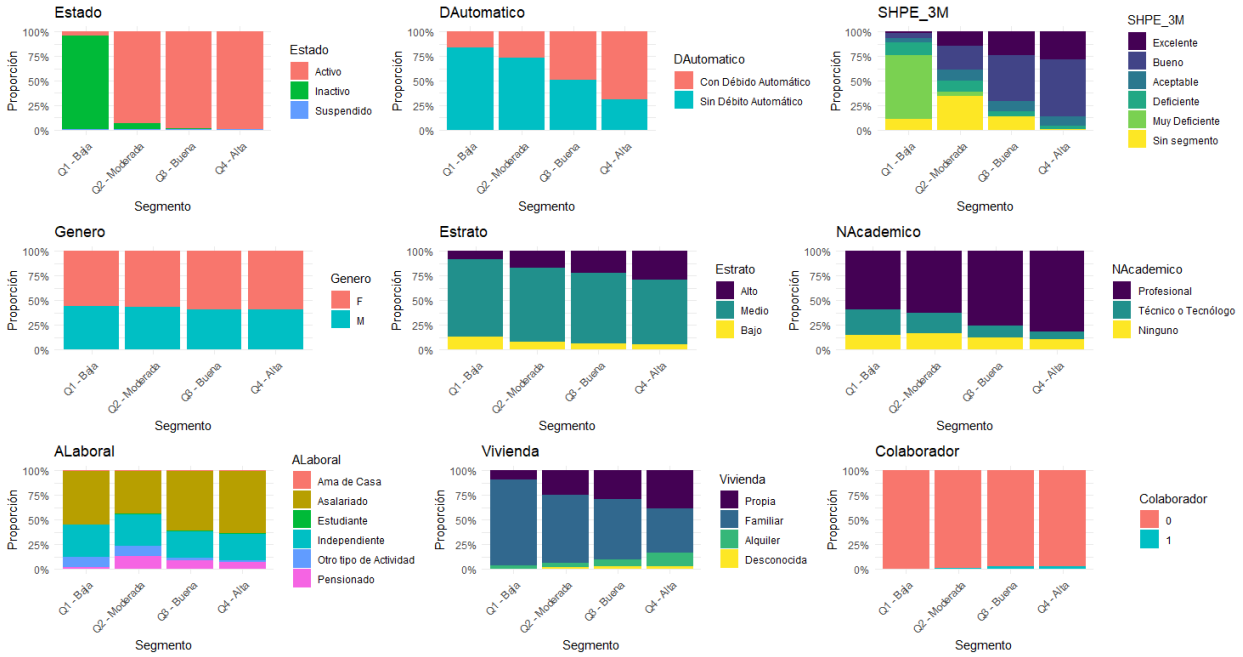

Figura 15 Distribución de variables categóricas por segmento de conexión

Se evidencian diferencias relevantes entre los segmentos, que complementan y refuerzan los hallazgos del análisis de variables numéricas:

- **Estado**: La proporción de asociados activos aumenta notablemente en los cuartiles más altos de conexión, siendo casi la totalidad del grupo Q4, mientras que en el grupo de conexión más bajo los Inactivos representan el 95%. Esto sugiere una relación directa entre la permanencia activa del asociado y su conexión con la entidad.
- **Débito Automático (DAutomatico)**: Los niveles Q3 y Q4 presentan una mayor proporción de asociados con débito automático, lo cual puede reflejar una mayor formalización en su relación financiera con la cooperativa.
- **Hábito de pago (SHPE_3M)**: En los segmentos Q3 y Q4 se incrementa la proporción de asociados con perfiles financieros clasificados como "Bueno" o "Excelente", indicando un mejor habito de pago en los últimos meses.
- **Género**: La distribución por género es relativamente estable entre los segmentos, sin diferencias

marcadas entre hombres y mujeres, con una mayor participación de mujeres en todos los grupos.

- **Estrato**: Se observa un aumento gradual de la proporción de asociados de estratos medios y altos a medida que se avanza hacia segmentos de mayor conexión, lo que podría estar asociado a mayor capacidad adquisitiva o estabilidad económica.
- **Nivel Académico (NAcademico)**: Los grupos con mayor conexión presentan una mayor proporción de asociados con nivel academico profesional.
- **Actividad Laboral (ALaboral)**: El segmento de alta conexión tiene una mayor participación de personas asalariadas y una menor proporción de actividades económicas informales o no activas como amas de casa o estudiantes.
- **Tipo de Vivienda**: La tenencia de vivienda propia incrementa significativamente en los grupos Q3 y Q4, lo cual puede interpretarse como una mayor estabilidad económica y patrimonial.
- **Colaborador**: Aunque la mayoría de los asociados no son colaboradores de la entidad, se observa una leve mayor participación relativa en los segmentos de mayor conexión.

Este análisis de las variables categóricas complementa los hallazgos de las variables numéricas al revelar que los asociados más conectados tienden a ser activos, tener débito automático, pertenecer a estratos medios o altos, poseer vivienda propia, contar con niveles educativos más altos y estar empleados formalmente. Estos perfiles también presentan una mejor salud financiera y mayor formalización en su relación con la entidad.

8 VALIDACIÓN Y DISCUSIÓN DE RESULTADOS

A continuación, se presentan tres análisis complementarios realizados para validar la calidad, coherencia interna y utilidad del score integral de conexión. La validación se llevó a cabo mediante técnicas estadísticas y exploratorias que permiten evaluar la capacidad del índice para sintetizar información y segmentar perfiles de asociados de forma consistente. Este enfoque se justifica dado que, al tratarse de una metodología no supervisada, no se dispone de un criterio externo de verdad o una variable objetivo que permita aplicar métricas tradicionales de precisión, exactitud o AUC propias de modelos supervisados. Asimismo, no se cuenta con un score previo construido y documentado de manera consistente que pueda servir como referencia de comparación objetiva. Por tal motivo, la validación se basó en analizar si el score refleja patrones diferenciados en variables clave, si sus segmentos son estadísticamente distintos y si los agrupamientos obtenidos son coherentes con el conocimiento del negocio, permitiendo así sustentar su aplicabilidad práctica como herramienta de segmentación estratégica.

8.1 CORRELACIÓN DEL SCORE CON LAS VARIABLES ORIGINALES

Con el objetivo de verificar que el score tiene relación significativa con los indicadores numéricos que reflejan comportamiento, uso y valor, se calculó la correlación de Pearson entre el score y cada variable numérica.

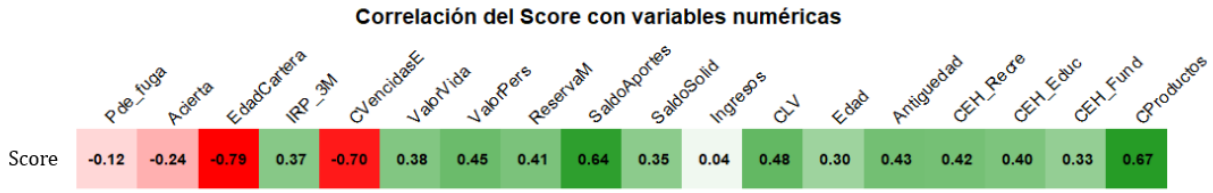

Figura 16 Correlación del Score con variables numéricas

El análisis de correlación permite evidenciar la relación entre el Score Integral de Conexión y cada una de las variables numéricas incluidas en el modelo. Entre los principales resultados se destacan:

Las correlaciones más elevadas y positivas se observaron con:

- Cantidad de Productos contratados (r=0,67), indicando que los asociados con mayor score tienden a tener más productos activos.
- Saldo de Aportes (r=0,64) y Valor de Vida (CLV) (r=0,48), variables que reflejan mayor generación de valor y compromiso económico con la cooperativa.
- Antigüedad (r=0,43) y variables de uso de beneficios (CEH_Recreativo, CEH_Fundacional, CEH_Educativo), que muestran que los asociados con mayor permanencia y participación presentan un score más alto.

Las correlaciones moderadas y negativas se identificaron con:

- Edad de Cartera (r=-0,79) y Cartera Vencida Estatutaria (r=-0,70), evidenciando que los asociados con historial de mora o antigüedad prolongada en la deuda tienden a registrar scores bajos.
- La variable Acierta (r=-0,24), asociada a riesgos crediticios, también mostró una relación negativa, aunque de menor magnitud.
- Variables como Ingresos (r=0,04) presentaron correlación muy baja, sugiriendo que no todas las dimensiones económicas se asocian de forma directa con el nivel de conexión medido por el score.

En conjunto, estos resultados confirman que el score refleja principalmente:

- Niveles de uso de productos y beneficios.
- Aportes y generación de valor.
- Buen comportamiento de pago (baja morosidad).

Por tanto, la correlación consistente con variables clave permite concluir que el score sintetiza adecuadamente patrones que son estratégicos para la cooperativa, diferenciando perfiles de alto y bajo compromiso con claridad.

8.2 ANÁLISIS DE VARIANZA (ANOVA) ENTRE CUARTILES DEL SCORE

Con el objetivo de validar la hipótesis de que los segmentos definidos por el score presentan diferencias estadísticamente significativas en las medias de las variables relevantes, realizamos un análisis de varianza (ANOVA) univariado para cada una de las variables numéricas consideradas en el estudio.
Este procedimiento permite comparar de manera formal si las medias de cada variable difieren entre los distintos cuartiles del score, evaluando la existencia de variaciones sistemáticas en función del nivel de conexión estimado por el índice. En otras palabras, el ANOVA ayuda a identificar si efectivamente los cuartiles definidos son representativos de perfiles diferenciados de comportamiento, uso de productos, generación de valor y riesgo crediticio.

La hipótesis estadística planteada en cada análisis fue la siguiente:

- **Hipótesis nula (H0):** Las medias de la variable son iguales en todos los cuartiles del score.
- **Hipótesis alternativa (H1):** Al menos un cuartil presenta una media distinta en comparación con los demás.

El contraste de hipótesis se realiza mediante el **estadístico F**, que es el cociente entre la variabilidad observada entre los grupos (en este caso, los cuartiles) y la variabilidad interna dentro de los grupos. Formalmente:

$$F = \frac{Varianza\ entre\ grupos}{Varianza\ dentro\ de\ los\ grupos}$$

Valores elevados del estadístico F indican que las diferencias entre medias no se explican únicamente por el azar, sino que son consistentes con un efecto real de segmentación. El p-valor asociado permite determinar la significancia estadística de estas diferencias en cada variable.

Para cada variable, se estimó el p-valor asociado al estadístico F, considerando un nivel de significancia de $\alpha = 0{,}05$ como criterio para rechazar la hipótesis nula. Valores p menores a este umbral indican que existen diferencias significativas en la media de la variable entre los grupos definidos por el score. A continuación, se presentan los resultados obtenidos:

Variable	F	Valor P
EdadCartera	439956,94	0,00E+00
CVencidasE	93108,67	0,00E+00
CProductos	83216,68	0,00E+00
SaldoAportes	73462,87	0,00E+00
CLV	24083,6	0,00E+00
Antiguedad	23490,48	0,00E+00
ValorPers	20922,47	0,00E+00
SaldoSolid	18245,91	0,00E+00
IRP_3M	14768,36	0,00E+00
ReservaM	14753,55	0,00E+00
ValorVida	14478,43	0,00E+00
CEH_Recre	13669,83	0,00E+00

CEH_Educ	11187,45	0,00E+00
Edad	9044,45	0,00E+00
Acierta	6805,57	0,00E+00
CEH_Fund	6761,38	0,00E+00
Pde_fuga	1641,6	0,00E+00
Ingresos	125,34	3,95E-81

Tabla 8 Resultados ANOVA

Los resultados obtenidos muestran que en todas las variables analizadas los p-valores fueron extraordinariamente bajos, en la mayoría de los casos inferiores a 1e-10, confirmando la existencia de diferencias estadísticamente significativas en sus medias entre los segmentos. A modo de ejemplo, el p-valor más elevado fue observado en la variable Ingresos ($p \approx 3{,}95e\text{-}81$), que sigue siendo sustancialmente inferior al umbral de significancia convencional.

Estos hallazgos indican que los cuartiles del score integral son capaces de discriminar perfiles de asociados con características marcadamente diferentes en dimensiones clave como el comportamiento de pago (Edad de Cartera, Cartera Vencida), el nivel de uso de productos y beneficios, la generación de valor (Saldo de Aportes, CLV) y otros indicadores financieros y transaccionales.

En conclusión, la evidencia obtenida a partir de las pruebas de ANOVA respalda de forma contundente la validez del score como herramienta de segmentación, demostrando que su estructura ordinal refleja diferencias significativas y coherentes en las variables relevantes para el análisis del nivel de conexión de los asociados con la cooperativa.

8.3 ANÁLISIS COMPLEMENTARIO AGRUPAMIENTO MEDIANTE K-MEANS

Como parte del proceso de validación y enriquecimiento del análisis, se incorporó un método no supervisado de agrupamiento basado en K-means. Este enfoque complementario permite identificar segmentos de asociados que comparten características similares a partir de la información contenida en los componentes principales seleccionados. El propósito principal de esta técnica no es sustituir el score integral de conexión, sino ofrecer una perspectiva adicional que ayude a contrastar y describir patrones latentes en los datos.

El algoritmo de K-means clasifica las observaciones en un número predefinido de grupos, minimizando la variabilidad interna de cada clúster y maximizando las diferencias entre clústeres. En este caso, el análisis se aplicó sobre las proyecciones de los registros en el espacio de los cinco primeros componentes principales, que representan una síntesis de la información multidimensional utilizada en la construcción del índice.

Mediante esta aproximación se busca evaluar si los grupos formados muestran coherencia interna y diferenciación respecto a variables clave, así como explorar la relación entre los clústeres obtenidos y los niveles de conexión definidos por los cuartiles del score. Este contraste contribuye a validar la utilidad del índice como herramienta de segmentación y a identificar oportunidades de caracterización adicional de los perfiles de los asociados.

Para determinar el número óptimo de clústeres a utilizar en el modelo K-means, se aplicó el método del codo directamente sobre el conjunto completo de datos proyectado en las primeras cinco componentes principales del Análisis de Componentes Principales (ACP). Este método evalúa la reducción progresiva de la suma de cuadrados intra-cluster (WSS) al incrementar el número de grupos, buscando identificar el punto donde agregar más clústeres deja de mejorar sustancialmente la segmentación.

En el gráfico obtenido se observa una disminución pronunciada de la variabilidad intra-cluster hasta aproximadamente K = 4 o 5, momento a partir del cual la curva se aplana, indicando rendimientos decrecientes en la mejora del modelo. Este comportamiento sugiere que una configuración con 4 o 5 clústeres logra un balance adecuado entre simplicidad del modelo y segmentación efectiva. Con base en este análisis, se seleccionó **K = 5** como valor final para aplicar el modelo K-means sobre las proyecciones de los datos.

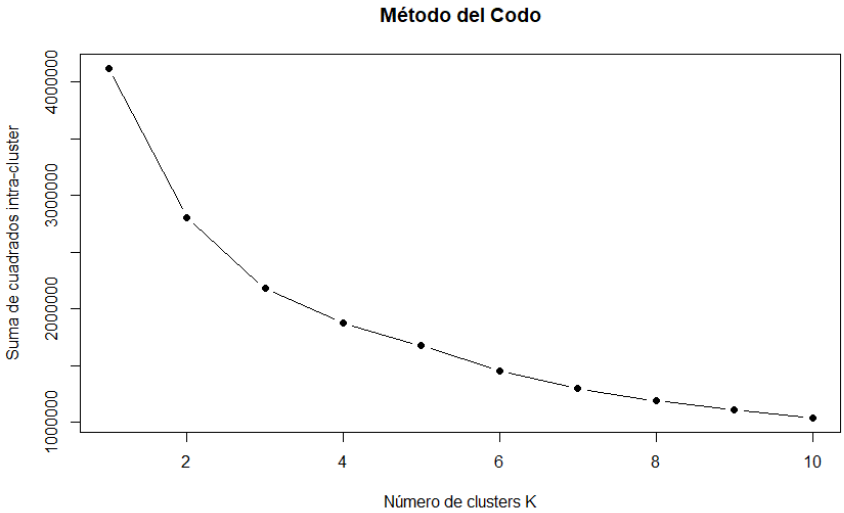

Figura 17 Método del Codo Para obtener el K óptimo

Además del método del codo, se aplicó el método del coeficiente de Silhouette sobre una muestra aleatoria de 50.000 observaciones con el fin de identificar el número óptimo de clústeres a utilizar en el modelo K-means. Esta métrica evalúa tanto la cohesión interna de los grupos como su separación respecto a otros grupos, otorgando valores cercanos a 1 cuando los clústeres están bien definidos.

En la Figura 18 se muestra que el valor promedio de Silhouette alcanza su punto máximo en **K = 5**, con un valor cercano a **0.33**, superior al obtenido con otros valores de K. Aunque el valor absoluto no es elevado (lo cual es esperable dada la complejidad de los datos), la comparación relativa entre configuraciones sugiere que **cinco clústeres** ofrecen una segmentación más coherente. Esta evidencia complementa la obtenida mediante el método del codo y respalda la selección de **K = 5** como valor final para aplicar el modelo K-means sobre los datos proyectados en el espacio de componentes principales.

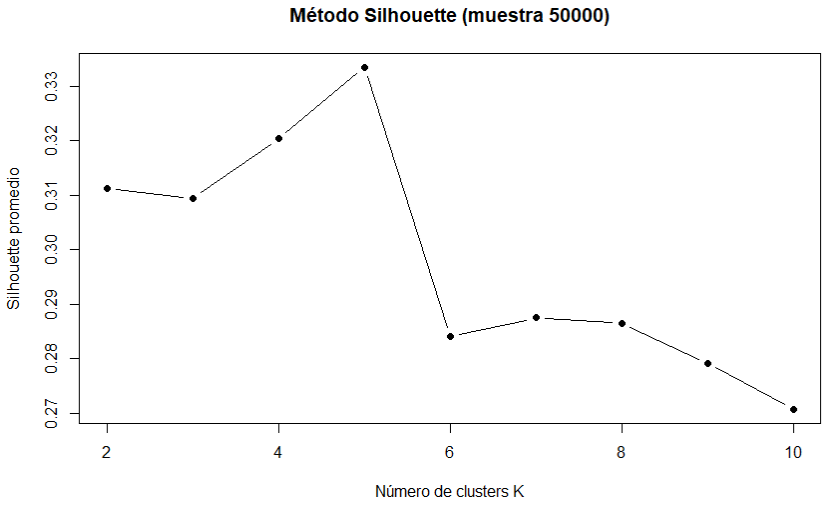

Figura 18 Metodo Silhouette

Con el objetivo de comprender el perfil sociodemográfico y conductual de los asociados agrupados mediante el modelo K-means, se realizó un análisis descriptivo de las variables categóricas, visualizando su distribución proporcional dentro de cada clúster. Este análisis permite identificar características diferenciales entre los grupos y contrastar la coherencia de los patrones emergentes.

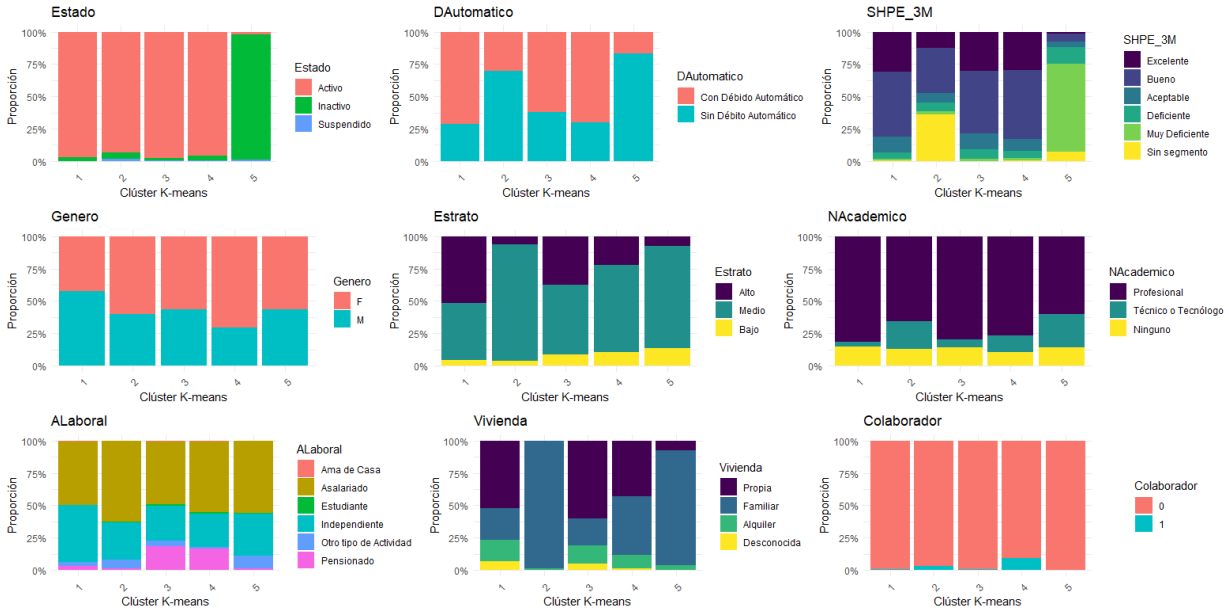

Figura 19 Descripción variables categóricas por Clusters

Los resultados se muestran en la Figura 19, y permiten extraer las siguientes observaciones clave:

- Clúster 5 se destaca negativamente por concentrar la mayor proporción de asociados inactivos (más del 90%), así como los niveles más altos de asociados sin débito automático, sin segmentación

o en condiciones muy deficientes de hábito de pago (SHPE_3M), estrato bajo, y nivel académico técnico o tecnólogo. También sobresale por tener alta proporción de personas en situación de pensión o inactividad laboral, y un acceso limitado a vivienda propia o de carácter familiar. En conjunto, este clúster representa un grupo con escasa vinculación, bajo compromiso y perfil de riesgo elevado, lo que lo convierte en un segmento crítico para intervenciones específicas.

- El clúster 2 presenta ciertas similitudes con el clúster 5, aunque con un matiz diferenciador importante. Comparte con el clúster 5 algunas características de baja vinculación, como un porcentaje elevado de asociados sin débito automático, niveles educativos inferiores y segmentaciones medias o deficientes. Sin embargo, se distingue por una menor proporción de asociados inactivos, lo que sugiere que aún podría haber oportunidad de fortalecer su relación con la cooperativa. Este grupo podría representar un segmento en riesgo de desconexión, pero aún recuperable mediante intervenciones adecuadas.
- Los clústeres 1, 3 y 4 agrupan a los asociados con los perfiles más sólidos y estrechamente integrados con la cooperativa. Estos clústeres concentran altas proporciones de asociados activos, con presencia significativa de débito automático, segmentación buena o excelente, nivel educativo profesional y vivienda propia o familiar. También presentan una mayor estabilidad en su actividad laboral, lo que se traduce en un mayor compromiso y participación con los servicios y beneficios ofrecidos por la organización.
- Entre estos grupos, el clúster 1 destaca como el de mejor perfil general. Se observa en este clúster una combinación óptima de características: mayor proporción de asociados activos, fuerte presencia de débito automático, mayor porcentaje de individuos con educación profesional, mejores calificaciones en el hábito de pago, y un entorno de vivienda más estable. En conjunto, este grupo representa el perfil ideal de asociado para la cooperativa, con mayor potencial de permanencia, fidelización y generación de valor.

Este análisis complementa el uso del score integral al demostrar que los clústeres formados de manera independiente con base en variables latentes proyectadas (ACP) muestran patrones claros y coherentes, lo que refuerza la validez del proceso de segmentación. En particular, la identificación del clúster 5 como grupo de menor conexión y alta vulnerabilidad permite enfocar acciones específicas para su retención, reactivación o gestión diferencial.

Posteriormente, se analizó el comportamiento de las variables numéricas utilizando el promedio por cada clúster identificado mediante el modelo K-means. Esta comparación permitió una caracterización más profunda de los grupos de asociados, revelando patrones claros en términos de riesgo, uso de productos, generación de valor y vinculación con la cooperativa.

Cluster Kmeans	1	2	3	4	5
Pde_fuga	0,20	0,21	0,23	0,25	0,27
IRP_3M	0,90	0,74	0,96	0,93	0,03
CVencidasE	0,78	0,85	0,51	0,88	28,41
EdadCartera	0,52	0,84	0,51	0,62	6,78
ValorVida	$ 98.898.564	$ 13.755.790	$ 19.829.128	$ 21.052.341	$ 14.319.075
ValorPers	$ 241.996.067	$ 25.746.168	$ 35.152.833	$ 44.928.152	$ 21.169.892
ReservaM	$ 128.417.961	$ 6.538.962	$ 22.081.872	$ 23.165.869	$ 6.880.674
SaldoAportes	$ 10.147.048	$ 1.897.497	$ 8.012.040	$ 6.739.149	$ 58.924
Ingresos	$ 15.079.555	$ 5.906.264	$ 7.189.690	$ 6.527.147	$ 4.254.526
CLV	$ 1.671.828	$ 333.603	$ 571.772	$ 664.722	$ 163.404
Edad	57	39	61	55	40

Antiguedad	25	5	22	19	6
CEH_Recre	37,06	17,23	33,83	291,22	8,65
CEH_Educ	1,32	0,42	1,31	10,69	0,37
CEH_Fund	0,80	0,26	0,61	8,55	0,26
CProductos	8,43	3,89	5,94	8,41	1,39

Tabla 9 Promedio Variables numéricas por Cluster

- **Clúster 1** se destaca como el grupo con mejores condiciones generales, registrando los valores más altos en ingresos promedio ($15 millones), CLV ($1,67 millones), uso de productos (8,4), y beneficios acumulados (e.g., CEH_Recre, CEH_Educ). Asimismo, presenta un buen hábito de pago (IRP_3M = 0,90) y bajos indicadores de riesgo (CVencidasE = 0,78).
- **Clúster 2**, por el contrario, exhibe condiciones considerablemente menos favorables, con ingresos y saldos muy por debajo del promedio, baja tenencia de productos (3,89), menor educación financiera y generación de valor (e.g., ValorPers = $25 millones, CLV = $333 mil). Aunque sus niveles de riesgo no son tan altos como los del clúster 5, refleja un perfil vulnerable y en riesgo de desconexión.
- **Clúster 5** concentra los valores más bajos en la mayoría de las variables evaluadas: ingresos ($4,2 millones), CLV ($163 mil), beneficios, aportes y uso de productos (1,39). Presenta los peores indicadores de cartera (CVencidasE = 28,41; EdadCartera = 6,78), lo que refuerza su perfil de alto riesgo y bajo compromiso.
- **Clústeres 3 y 4** presentan características intermedias. El clúster 3 tiene una combinación interesante de ingresos aceptables y baja morosidad, mientras que el clúster 4 muestra buenos indicadores de reservas y aportes, además de una destacada participación en programas no transaccionales (por ejemplo, CEH_Fund = 8,55).

8.3.1 CONCLUSIÓN COMPARATIVA ENTRE EL SCORE INTEGRAL Y EL MODELO K-MEANS

A partir del análisis realizado, es posible comparar los resultados obtenidos por el Score Integral de Conexión y los generados a través del modelo no supervisado K-means. Aunque ambos métodos ofrecen una segmentación útil de los asociados, existen diferencias importantes tanto en su propósito como en su capacidad de caracterización.

El modelo K-means permitió agrupar a los asociados según similitudes en múltiples dimensiones, revelando grupos con patrones compartidos. Esta metodología resulta útil para identificar perfiles emergentes y para construir estrategias específicas basadas en la similitud de comportamientos y características. Sin embargo, una limitación observada es que no todos los clusters presentaron una diferenciación tan clara o gradual en cuanto a su valor estratégico para la cooperativa. Por ejemplo, aunque el clúster 5 mostró perfiles de bajo compromiso, otros grupos como el 2 presentaban una mezcla de características que dificultan una lectura lineal de valor.

Por otro lado, el Score Integral de Conexión, al ser una métrica continua derivada del Análisis de Componentes Principales, permitió clasificar a los asociados de forma ordenada y cuantitativa en función de su nivel de conexión. Este enfoque ofrece ventajas claras en términos de interpretación y priorización, ya que los valores más altos del score se asocian consistentemente con características estratégicas para Coomeva: mayor uso de productos, mejores indicadores de pago, más ingresos y beneficios, así como un

mayor nivel educativo y de antigüedad.

Además, la segmentación por cuartiles del score presentó una evolución progresiva en la calidad del perfil de los asociados, lo cual facilita su uso práctico para la toma de decisiones comerciales, priorización de campañas, asignación de recursos y evaluación de impacto.

En resumen, aunque ambos enfoques ofrecen valor analítico, el uso del Score Integral demostró una mayor capacidad para representar la relación entre los asociados y la cooperativa de manera numérica, continua y alineada con los objetivos estratégicos de fidelización y retención.

9 CONCLUSIONES Y TRABAJOS FUTUROS

Este trabajo de grado logró desarrollar un score integral de conexión para los asociados de la cooperativa Coomeva, cumpliendo con el objetivo general planteado de evaluar y cuantificar el nivel de relación de cada asociado con las distintas dimensiones de la organización. A través del uso de herramientas de análisis de datos, exploración descriptiva, técnicas de reducción de dimensionalidad y segmentación, se logró identificar patrones relevantes que permiten entender la conexión actual de los miembros con la cooperativa. A continuación, presentamos las principales conclusiones encontradas a lo largo del estudio.

9.1 CONCLUSIONES

- El análisis exploratorio permitió comprender en profundidad la estructura y calidad de los datos disponibles sobre los asociados, revelando una alta variabilidad en aspectos clave como salud de la cartera, generación de valor, características demográficas y uso de productos. Se identificaron patrones relevantes, como la baja probabilidad de fuga en la mayoría de los casos, la alta permanencia de los asociados y una relación directa entre la antigüedad y el saldo de aportes. La correlación entre variables mostró relaciones esperadas en algunos bloques temáticos, así como independencia entre otros indicadores, lo que aporta valor al análisis multivariado. Finalmente, el análisis de variables categóricas reveló que, aunque la mayoría de los asociados están activos y bien calificados en hábitos de pago, existe una proporción considerable con alertas o inactividad, aspectos clave para considerar en la construcción del score de conexión. En conjunto, estos hallazgos constituyen una base sólida para el diseño de estrategias analíticas más precisas y enfocadas en mejorar la vinculación de los asociados con la cooperativa.

- La aplicación de técnicas de reducción de dimensionalidad permitió identificar y seleccionar las variables más relevantes que inciden en la conexión de los asociados con la cooperativa, mejorando la interpretación del modelo y reduciendo la redundancia de información. A través de una revisión exhaustiva de diferentes metodologías, se concluyó que el Análisis de Componentes Principales (ACP) era la técnica más adecuada para el contexto del estudio, dada su capacidad explicativa, facilidad de interpretación y adaptabilidad a datos numéricos y categóricos debidamente transformados.

- El preprocesamiento cuidadoso de los datos permitió construir una base robusta para el análisis. Posteriormente, el ACP logró reducir significativamente la dimensionalidad del conjunto de datos, reteniendo el 68.07% de la variabilidad total en los primeros 10 componentes. Aunque los dos primeros componentes explicaron solo el 29.44% de la varianza, su análisis permitió obtener una comprensión inicial de las estructuras latentes, agrupaciones y patrones presentes en los datos. En particular, el primer componente principal (PC1) resultó ser altamente interpretativo, reflejando una combinación de factores asociados a la estabilidad financiera y nivel de conexión de los asociados con la cooperativa. Variables como *SaldoAportes*, *Antigüedad*, *Edad*, *Cantidad de productos* y *Vivienda propia* mostraron una alta carga positiva, indicando su papel fundamental en la construcción del índice de conexión. Estos resultados resaltan que los asociados con mayor permanencia, comportamiento financiero sólido y mayor participación en productos tienden a estar más conectados con la cooperativa. En resumen, el ACP permitió no solo reducir la complejidad del conjunto de datos, sino también identificar un grupo clave de variables que explican con claridad el nivel de relación de los asociados con la organización. Esta información resulta crucial para el diseño de estrategias de fidelización, segmentación y gestión del vínculo con los asociados, sirviendo como base para futuros modelos predictivos más eficientes y precisos.

- El cálculo del Score Integral de Conexión permitió condensar de manera efectiva la información multivariada sobre la relación de los asociados con la cooperativa en una métrica única, continua e interpretable. A través del Análisis de Componentes Principales (PCA), se identificaron las dimensiones clave que explican la mayor parte de la variabilidad en los datos, y mediante la proyección ponderada sobre los primeros cinco componentes principales, se construyó un índice robusto que refleja con precisión el grado de conexión y compromiso de cada asociado. La posterior normalización del score a una escala de 0 a 100 facilitó su interpretación y comparación. La elección de cinco componentes principales se justificó tanto por criterios estadísticos como por principios de parsimonia, al lograr un equilibrio óptimo entre capacidad explicativa (capturando cerca del 50% de la varianza acumulada), simplicidad e interpretabilidad. Este score se consolida así como una herramienta estratégica clave para segmentar, priorizar y fortalecer la gestión relacional con los asociados.

- El análisis integral de las variables numéricas y categóricas según los segmentos de conexión evidencia un patrón claro y consistente: los asociados con mayor nivel de conexión (Q4) presentan características significativamente más favorables tanto en términos financieros como sociodemográficos. Esto valida la utilidad del score de conexión como herramienta para identificar perfiles de alto valor y compromiso con la cooperativa. Desde el punto de vista numérico, los asociados del cuartil más alto muestran un mejor desempeño en indicadores clave como ingresos, saldos en productos, estados óptimos de pago, valor de vida del cliente (CLV) y número de productos. Estas variables reflejan una mayor profundidad en el uso de los servicios, mejores

hábitos financieros y una relación más estable y rentable con la entidad. Por otro lado, el análisis de las variables categóricas complementa estos hallazgos al revelar que los asociados más conectados tienden a ser activos, tener débito automático, pertenecer a estratos medios o altos, poseer vivienda propia, contar con niveles educativos más altos y estar empleados formalmente. Estos perfiles también presentan una mejor salud financiera y mayor formalización en su relación con la entidad. En conjunto, estos resultados permiten concluir que el score de conexión no solo agrupa a los asociados según su vinculación con los productos y servicios, sino que además captura diferencias estructurales en su perfil financiero, educativo y social. Esto abre oportunidades valiosas para diseñar estrategias de fidelización, segmentación de productos y políticas de retención más focalizadas y efectivas.

9.2 TRABAJOS FUTUROS

Para próximos ejercicios, sería valioso diferenciar a los asociados activos de los inactivos con el fin de construir scores específicos para cada grupo. Esta segmentación permitiría aplicar estrategias diferenciadas: enfocadas en la fidelización y fortalecimiento de la relación con los asociados activos, y en la recuperación y reactivación de los asociados inactivos. Adicionalmente, el score integral de conexión desarrollado en este estudio puede constituirse en una variable explicativa clave dentro de modelos predictivos, como los de fuga o cancelación de productos, aportando una visión más completa del comportamiento y compromiso del asociado. Desde la perspectiva de mercadeo, los resultados obtenidos abren la posibilidad de diseñar estrategias personalizadas por grupo, orientadas a mantener, fidelizar y promover el uso de productos y servicios según las características de cada segmento. Finalmente, se sugiere implementar un monitoreo dinámico del score, que permita identificar cambios en el comportamiento de los asociados en el tiempo, anticipar riesgos de desvinculación o detectar nuevas oportunidades de vinculación y crecimiento.

10 REFERENCIAS BIBLIOGRÁFICAS

[1] R Core Team, «R: A Language and Environment for Statistical Computing.,» Vienna, 2025.

[2] J. W. Tukey, Exploratory Data Analysis., Addison-Wesley., 1977.

[3] J. T. Behrens, de *Principles and procedures of exploratory data analysis*, Psychological Methods, 1997, pp. 131-160.

[4] C. M. Bishop, de *Pattern Recognition and Machine Learning*, Springer., 2006, p. 20.

[5] P. Domingos, de *A few useful things to know about machine learning*, Communications of the ACM, 2012, pp. 78-87.

[6] I. B. Y. &. C. A. Goodfellow, de *Deep Learning.*, MIT Press., 2016.

[7] T. T. R. &. F. J. Hastie, de *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*,

(2nd ed.). Springer, 2009.

[8] I. T. Jolliffe, de *Principal Component Analysis*, Springer Series in Statistics, 2002.

[9] S. E. K. &. G. P. Wold, de *Principal component analysis.*, Chemometrics and Intelligent Laboratory Systems, 1987, pp. 37-52.

[10] IBM, «IBM,» 8 diciembre 2023. [En línea]. Available: https://www.ibm.com/es-es/think/topics/principal-component-analysis. [Último acceso: 15 mayo 2025].

[11] G. E. &. S. R. R. Hinton, de *Reducing the Dimensionality of Data with Neural Networks*, Science, 2006, pp. 504-507.

[12] P. L. H. B. Y. &. M. P. A. Vincent, de *Extracting and Composing Robust Features with Denoising Autoencoders.*, In Proceedings of the 25th International Conference on Machine Learning, 2008, pp. (pp. 1096-1103.

[13] C. Cooperativa., «Informe de Gestión Anual,» Coomeva, 2023.

[14] J. Gómez, Estrategias de Retención y Fidelización en Cooperativas Colombianas., Editorial Cooperativa., 2021.

[15] F. O. y. D. F. L. L. M. Cruz Castro, «Construcción de un índice socioeconómico familiar para los estudiantes que presentan la prueba Saber 11,» 2016. [En línea]. Available: https://revistas.usantotomas.edu.co/index.php/estadistica/article/view/2644/3236 . [Último acceso: 14 03 2025].

[16] D. Barragán Garnica, «Patrones de comportamiento de clientes con tarjetas de crédito de consumo con deterioro de calificación por riesgo utilizando K-means. ODEON,» 22 julio 2023. [En línea]. Available: https://revistas.uexternado.edu.co/index.php/odeon/article/view/8873 . [Último acceso: 15 02 2025].

[17] B. Shen, «E-commerce Customer Segmentation via Unsupervised Machine Learning,» de *In The 2nd International Conference on Computing and Data Science (CONF-CDS 2021)*, New York, NY, USA, 2021.

[18] I. T. J. a. J. Cadima, "Principal component analysis: a review and recent developments,", Philosophical Transactions of the Royal Society A, vol. 374 no. 2065, 2016.

[19] A. C. R. a. W. F. Christensen, Methods of Multivariate Analysis, 3rd ed. Wiley, 2012.

[20] Y. B. a. A. C. I. Goodfellow, Deep Learning, MIT Press, 2016.

[21] F. Chollet, Deep Learning with Python,, 2nd ed., Manning Publications, 2021.

[22] W. S. a. K.-R. M. G. Montavon, de *Methods for interpreting and understanding deep neural networks*, Digital Signal Processing, vol. 73, 2018, p. 1–15.

[23] B. E. a. J. Pagès, de *Multiple factor analysis (AFMULT package)*, computational Statistics & Data Analysis, vol. 18, 1994, p. 121–140.

[24] H. A. a. L. J. Williams, de *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, 2010, p. 433–459.

[25] D. W. T. H. a. R. T. G. James, An Introduction to Statistical Learning: with Applications in R, Springer, 2013.

[26] W. C. B. B. J. B. a. R. E. A. J. F. Hair, Multivariate Data Analysis, 8th ed. Cengage Learning, 2019.

[27] M. O. B. Avella, «Comparación del análisis factorial múltiple (AFM) y del análisis en componentes principales para datos cualitativos (Prinqual), en la construcción de índices,» 2010. [En línea].

Available:
https://repositorio.unal.edu.co/bitstream/handle/unal/6804/832213.2010.pdf?sequence=1&isAllowed=y. [Último acceso: 2025 enero 30].