Clusterización y Aplicación de Modelos Predictivos Para la Identificación de Factores que Conlleven a una Desafiliación de una Cooperativa Financiera

Moreno Tafur, William Dairo

Asesor

García García, Mireya

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería - ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Nota de Aceptación

García García, Mireya

Director de Trabajo de Grado

Gaitán Ospina, Rafael

Jurado

2024

Consideraciones Éticas

En cumplimiento de la Ley 1581 de 2012 sobre Protección de Datos Personales en Colombia, y con el objetivo de garantizar la confidencialidad y el respeto por la información sensible, se ha decidido cambiar el nombre de la empresa objeto de este estudio por “CoopFinan”. Este nombre será utilizado de manera consistente a lo largo del proyecto para referirse a la entidad analizada. Esta medida responde a la normativa vigente y a los principios de responsabilidad, transparencia y seguridad en el manejo de datos, buscando evitar cualquier vulneración de derechos o compromisos éticos relacionados con la divulgación de información no autorizada. Además, se alinea con la protección de la reputación y los intereses comerciales de la organización, garantizando un enfoque riguroso y ético en el desarrollo de este trabajo académico.

Resumen

Este proyecto tiene como objetivo identificar los factores que influyen en la desafiliación de clientes utilizando modelos de machine learning y análisis de clústeres. La retención es vital para la rentabilidad y crecimiento de una empresa, ya que conservar a los clientes existentes es más rentable que adquirir nuevos (Rana et al., 2023). Se realiza un análisis exploratorio de datos y se emplean técnicas de minería de datos para descubrir patrones y obtener información relevante a partir de datos históricos. Métodos de aprendizaje no supervisado como K-means contribuyen a la segmentación y personalización de estrategias de retención (Supraja & Sairamesh, 2023). Además, la integración de modelos de ensamble como Los Árboles de decisión y Random Forest, mejoran la precisión de las predicciones al manejar desequilibrios en los datos (Zhang et al., 2020; Allegue et al., 2020). La implementación de estos modelos predictivos en las empresas ha demostrado ser efectiva para prevenir la desafiliación de clientes y optimizar las estrategias de retención, aumentando la competitividad y rentabilidad (Wu & Li, 2018), lo que facilita una comprensión más profunda de las dinámicas de desafiliación y permite diseñar estrategias de intervención más efectivas y adaptadas a las necesidades específicas.

***Palabras claves*:** Clústeres, modelos predictivos, deserción, aprendizaje no supervisado.

Abstract

This project aims to identify the factors that influence customer disaffiliation using machine learning models and cluster analysis. Retention is vital for the profitability and growth of a company, since retaining existing customers is more profitable than acquiring new ones (Rana et al., 2023). Exploratory data analysis is performed and data mining techniques are used to discover patterns and obtain relevant information from historical data. Unsupervised learning methods such as K-means contribute to the segmentation and personalization of retention strategies (Supraja & Sairamesh, 2023). Furthermore, the integration of ensemble models such as Decision Trees and Random Forest improve the accuracy of predictions by handling imbalances in the data (Zhang et al., 2020; Allegue et al., 2020). The implementation of these predictive models in companies has proven to be effective in preventing customer disaffiliation and optimizing retention strategies, increasing competitiveness and profitability (Wu & Li, 2018), which facilitates a deeper understanding of the dynamics of disaffiliation and allows the design of more effective intervention strategies adapted to specific needs.

Keywords**:** Clusters, predictive models, churn, unsupervised learning.

Tabla de Contenido

Lista de Figuras

Introducción

La información utilizada en este proyecto proviene de “CoopFinan”, una cooperativa de ahorro y crédito de tamaño mediano perteneciente al sector financiero, con más de 60 años de trayectoria en el mercado. La entidad cuenta con presencia nacional a través de 22 oficinas, identificadas mediante un código único asignado por la organización, y gestiona una base de datos que abarca información de más de 44,000 clientes. Este amplio alcance proporciona un contexto sólido para analizar las dinámicas de desafiliación y diseñar estrategias efectivas de retención.

La retención de clientes es un aspecto fundamental para la sostenibilidad y el crecimiento de las empresas, especialmente en un entorno competitivo donde adquirir nuevos clientes implica costos significativamente más altos que mantener los existentes (Rana et al., 2023). En la empresa “CoopFinan”, siempre se ha presentado una alta tasa de desafiliación, lo que representa un desafío significativo para su modelo de negocio. Este fenómeno no solo genera pérdidas económicas, sino también la necesidad de desarrollar estrategias de retención más efectivas y personalizadas.

Este proyecto aplicado se centra en la identificación de los factores que influyen en la desafiliación de clientes a través del uso de técnicas avanzadas de análisis de datos y machine learning. Mediante la implementación de métodos de aprendizaje no supervisado como K-means y K-modes, se segmentan los clientes en clústeres que comparten características demográficas, de comportamiento y financieras. Esto permite no solo entender las dinámicas que conducen al abandono, sino también personalizar las estrategias de retención. Asimismo, se integran modelos predictivos supervisados, como Árboles de Decisión y Random Forest, los cuales ofrecen

precisión en la predicción del riesgo de deserción, incluso en conjuntos de datos desequilibrados (Zhang et al., 2020; Allegue et al., 2020).

A través del análisis exploratorio de datos y el uso de herramientas como el Análisis de Componentes Principales (PCA), el proyecto optimiza la interpretación y reduce la dimensionalidad de los datos, lo que facilita la extracción de insights clave. Estos modelos predictivos se validan con métricas como precisión, recall y el Área Bajo la Curva (AUC), asegurando su efectividad y capacidad para generalizar. Como resultado, se busca no solo predecir y prevenir la desafiliación, sino también proporcionar una guía práctica para diseñar intervenciones tempranas y específicas que mejoren la lealtad y permanencia de los clientes.

Este trabajo contribuye significativamente al campo de la ciencia de datos al demostrar cómo las técnicas de machine learning pueden aplicarse para abordar problemas reales de negocio, optimizando la toma de decisiones estratégicas en la gestión de clientes. Además, los hallazgos obtenidos servirán como base para futuras investigaciones en modelos predictivos y estrategias de fidelización en empresas de diversos sectores.

Planteamiento del Problema

La empresa “CoopFinan” enfrenta un desafío crítico en la retención de sus clientes, a pesar de los esfuerzos por ofrecer productos y servicios que atiendan las diversas necesidades de sus usuarios. La tasa de deserción, que alcanza el 10,93%, representa un problema significativo, ya que no solo disminuye la base de clientes, sino que también incrementa los costos operativos.

Este proyecto se propone analizar los factores que influyen en la desafiliación mediante un enfoque basado en machine learning y técnicas de análisis predictivo. Se utilizarán métodos de agrupación no supervisada, como K-means, para identificar patrones y segmentar a los clientes en función de sus características y comportamientos, lo que permitirá una mejor comprensión de los factores de riesgo asociados con la desafiliación. Además, los algoritmos de aprendizaje supervisado, como los Árboles de Decisión y Random Forest, serán fundamentales para predecir el riesgo de deserción con mayor precisión y gestionar los desequilibrios en los datos, garantizando que los resultados obtenidos sean representativos y confiables (Supraja & Sairamesh, 2023; Zhang & Qiu, 2020).

La relevancia de este análisis radica en su potencial para ofrecer insights valiosos que orienten la toma de decisiones estratégicas, permitiendo la implementación de intervenciones tempranas y personalizadas para reducir la desafiliación. Al identificar los factores específicos que contribuyen a la pérdida de clientes, La empresa podrá optimizar sus estrategias de retención y mejorar su oferta de valor, logrando así una relación más duradera y significativa con ellos.

Justificación

La retención de clientes en la empresa es crucial no solo para mantener una base de usuarios estable, sino también para asegurar el crecimiento económico de la entidad. Los clientes representan la principal fuente de ingresos a través de sus transacciones, productos y servicios. No obstante, la creciente desafiliación implica tanto una pérdida de volumen operativo como un incremento en los costos, ya que atraer nuevos prospectos requiere una inversión significativa en recursos de marketing y tiempo (Rana et al., 2023). Este fenómeno subraya la necesidad de contar con un enfoque proactivo y basado en datos para comprender y mitigar las causas de la desafiliación.

Mediante el uso de herramientas avanzadas de análisis de datos y modelos de machine learning, el presente proyecto busca identificar factores determinantes en la deserción, lo que permitirá a la empresa desarrollar estrategias personalizadas de retención. Los métodos de aprendizaje no supervisado, como K-means, facilitarán la segmentación según patrones de comportamiento y características demográficas, mientras que los modelos de machine learning, como los Árboles de Decisión y Random Forest, aumentarán la precisión en la predicción de deserción al manejar desequilibrios en los datos (Supraja & Sairamesh, 2023; Zhang & Qiu, 2020).

Al implementar estos modelos predictivos, la empresa no solo podrá reducir sus tasas de desafiliación, sino también optimizar sus estrategias de atención y retención, enfocándose en los usuarios con mayor riesgo de desafiliación. Este enfoque innovador no solo responde a una necesidad institucional inmediata, sino que también contribuye al desarrollo de estrategias de fidelización basadas en datos dentro del sector, consolidando la competitividad de la entidad y fortaleciendo su misión de generar valor para sus clientes.

Objetivos

Objetivo General

Analizar de manera sistemática los factores determinantes de la deserción de clientes en la empresa, con el fin de identificar patrones de desafiliación y generar insights que permitan la toma de decisiones estratégicas y el desarrollo de proyectos de automatización enfocados en mejorar la retención de sus usuarios.

Objetivos Específicos

Analizar patrones de comportamiento y las características de los clientes que han desertado utilizando herramientas estadísticas para determinar posibles causas y tendencias.

Estudiar el perfil sociodemográfico y financiero de los clientes que desertan, con el fin de determinar patrones de comportamiento y características comunes que puedan predecir futuros casos de desafiliación.

Definir indicadores clave de desempeño (KPIs) relacionados con la retención, que permitan medir de manera continua el éxito de las estrategias implementadas para mejorar la permanencia.

Desarrollar modelos predictivos para identificar clientes en riesgo de desertar con el fin de implementar estrategias de intervención temprana y mejorar la retención.

Marco Teórico

La retención de clientes se refiere a la capacidad de una empresa para mantener su base a lo largo del tiempo. Es una métrica crucial que impacta directamente en la rentabilidad y el potencial de crecimiento de una empresa (Rana et al., 2023). La retención de clientes es más rentable que la adquisición de nuevos clientes, ya que conservar a los existentes requiere una menor inversión en tiempo y recursos (Rana et al., 2023).

La evolución de las técnicas de análisis de datos y toma de decisiones ha sido significativa a lo largo de la historia, partiendo de métodos manuales y matemáticos clásicos hasta llegar a las sofisticadas técnicas de machine learning actuales. En sus inicios, se utilizaban métodos matemáticos clásicos como el cálculo diferencial e integral para modelar cambios y optimizar funciones continuas (Hamming, 1986), y el álgebra lineal para resolver sistemas de ecuaciones y comprender transformaciones lineales (Kolmogorov, 1950). La teoría de probabilidades sentó las bases para muchos métodos estadísticos y de inferencia (Kolmogorov, 1950). Con el tiempo, se desarrollaron métodos analíticos y heurísticos, como el método de Newton-Raphson para encontrar raíces de funciones no lineales (Hamming, 1986) y técnicas de interpolación para aproximar funciones (Hamming, 1986). Los métodos gráficos y manuales, como los gráficos de barras y el análisis de series temporales, permitieron una mejor comprensión visual de los datos (Polya, 1945). El razonamiento lógico y deductivo, incluyendo la lógica proposicional y el método de resolución de problemas de Polya (1945), fue fundamental en la resolución de problemas y la demostración de teoremas. Las técnicas de optimización clásica, como la programación lineal y no lineal (Bishop, 2006) y los métodos de optimización por gradiente (Hamming, 1986), también fueron esenciales. Con la llegada del siglo XX, las técnicas estadísticas clásicas como la regresión lineal y no lineal, el análisis de varianza

(Breiman, 2001) y el análisis de componentes principales (Bishop, 2006)) se convirtieron en la base del análisis de datos. La minería de datos combinó técnicas estadísticas y algoritmos para descubrir patrones en grandes conjuntos de datos, utilizando herramientas como árboles de decisión (Naoui et al., 2020), algoritmos de agrupamiento (Naoui et al., 2020) y redes Bayesianas (Bishop, 2006). Los sistemas expertos y algoritmos de búsqueda fueron los precursores de la inteligencia artificial moderna (Bishop, 2006). Finalmente, el machine learning ha revolucionado el análisis de datos y la toma de decisiones, permitiendo a los sistemas aprender y mejorar con la experiencia.

Las técnicas de machine learning han emergido como herramientas poderosas para prevenir la deserción de clientes y mejorar las estrategias de retención. Según Supraja y Sairamesh (2023), los métodos de aprendizaje no supervisado, como K-means y DBSCAN, son efectivos para segmentar clientes basándose en diversos atributos, lo cual permite a las organizaciones personalizar sus estrategias de marketing y retención, mejorando así la satisfacción y lealtad del cliente (Supraja & Sairamesh, 2023). Este enfoque ayuda a identificar patrones de comportamiento que pueden predecir la probabilidad de que un cliente abandone el servicio, permitiendo a las empresas tomar medidas proactivas.

La integración de modelos de ensamble ha demostrado mejorar significativamente la precisión en la predicción de la deserción de clientes. Zhang, Zhang y Qiu (2020) destacan que la combinación de múltiples algoritmos, como Random Forest, SVM y KNN, aumenta la robustez de las predicciones, especialmente cuando se utiliza en conjunto con técnicas de balanceo de clases como SMOTE (Zhang et al., 2020). Allegue et al., 2020, añaden que los modelos híbridos que combinan diferentes algoritmos pueden manejar mejor los desequilibrios en los datos y

capturar la diversidad de comportamientos de los clientes, proporcionando insights más profundos y accionables para la retención de clientes (Allegue et al., 2020).

En el contexto de la banca digital, los modelos de machine learning han sido particularmente efectivos para predecir la deserción de clientes. Un estudio realizado por Wu y Li (2018) aplicó métodos de regresión logística combinados con técnicas de ensamblaje, logrando una alta precisión en la predicción del abandono de clientes (Wu & Li, 2018). La personalización de los servicios y la respuesta rápida a las necesidades de los clientes son estrategias clave derivadas de estos modelos predictivos, las cuales no solo ayudan a retener a los clientes existentes, sino que también atraen a nuevos usuarios, aumentando la competitividad y rentabilidad de las instituciones financieras.

Durante la pandemia, la importancia de la gestión de relaciones con clientes (CRM) basada en machine learning se hizo aún más evidente. Según Supraja y Sairamesh (2023), la segmentación de clientes mediante análisis RFM (Recencia, Frecuencia y Valor Monetario) y el uso de algoritmos como K-means permiten a las empresas optimizar sus recursos de marketing y mejorar las estrategias de retención (Supraja & Sairamesh, 2023). Este tipo de análisis permite a las empresas comprender mejor las necesidades y comportamientos de sus clientes, facilitando la implementación de tácticas de retención más efectivas.

La adopción de modelos de machine learning pueden presentar problemas como el desequilibrio de clases en los datos, lo que puede llevar a modelos sesgados. Para abordar este problema, Ahmad et al. (2020) emplearon técnicas como SMOTE y GAN para equilibrar los conjuntos de datos, mejorando así la fiabilidad de los modelos predictivos (Ahmad et al., 2020).

Estos enfoques avanzados no solo ayudan a prever con mayor precisión la deserción de clientes, sino que también proporcionan una comprensión más profunda de los patrones de datos, permitiendo intervenciones más personalizadas y efectivas.

Los algoritmos de Random Forest y K-means proporcionan una poderosa metodología dual que abarca tanto el aprendizaje supervisado como el no supervisado. Este nos asegura un análisis exhaustivo y preciso, que puede manejar la complejidad y diversidad de los datos, al tiempo que proporciona resultados interpretables y útiles para la toma de decisiones informadas en el proyecto.

Metodología

La base de datos donde se encuentran almacenada la información es Sybase, la cual contiene datos históricos y actuales de los clientes, incluyendo variables demográficas, de comportamiento y su historial de afiliación. La recolección de estos datos se realiza mediante procedimientos almacenados diseñados específicamente para extraer información relevante a través de consultas estructuradas. Estos procedimientos seleccionan los registros pertinentes, los almacenan temporalmente en una tabla dentro de la misma base de datos y, posteriormente, se exportan a un archivo en formato CSV. Este dataset generado constituye la base principal para la implementación del proyecto. Se va a utilizar una metodología cuantitativa para analizar y predecir la desvinculación de clientes de la empresa, combinando técnicas de análisis descriptivo y analítico, se llevará a cabo una investigación descriptiva para resumir y describir las características principales de los datos históricos, incluyendo variables demográficas, de comportamiento y de historial de afiliación. Posteriormente, se empleará una investigación analítica para identificar patrones, realizar hipótesis e identificar relaciones significativas entre estas variables y la desvinculación.

Fase 0 Visualización y Análisis Descriptivo de Datos

Esta fase incluye gráficos que permiten comprender cómo se distribuyen los retiros de clientes a lo largo del tiempo (mensual y diario) y por diferentes características demográficas y organizacionales, como sexo, edad, estado civiles, tipo de canal y profesión.

Frecuencia de Retiros Mensual y Diaria

Como se puede observar en la Figura 1 y Figura 2, la frecuencia de los retiros presenta patrones significativos tanto a nivel mensual como diario durante el periodo comprendido entre agosto de 2023 y julio de 2024.

En la Figura 1, la frecuencia de retiros por mes muestra que los meses de abril, mayo y junio concentran la mayor cantidad de retiros, superando los 250 casos cada uno. Por otro lado, en la Figura 2, Se observa una concentración significativa de desafiliaciones en los primeros días del mes, y podría estar relacionado con la intención de evitar un impacto negativo en las cifras comerciales correspondientes al periodo anterior, asegurando así un cierre adecuado de los indicadores comerciales.

Figura 1

Frecuencia de Desafiliaciones por Mes del 08/01/2023 al 07/31/2024

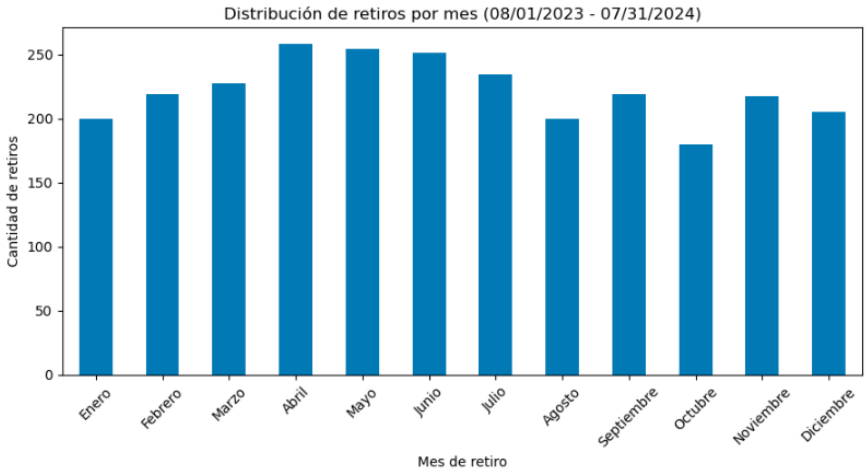
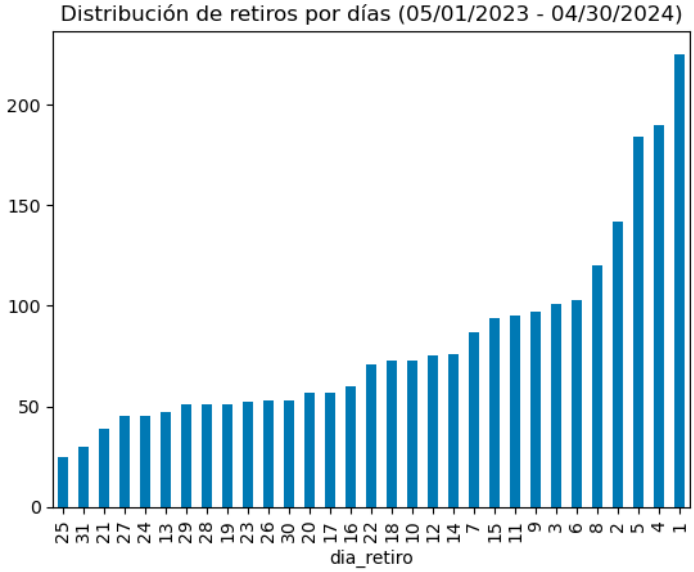


Figura 2

Frecuencia de Desafiliaciones por Día del 08/01/2023 al 07/31/2024



Frecuencia de Desafiliaciones Análisis por Categorías

Como se puede observar en las siguientes figuras, la frecuencia de las desafiliaciones presenta patrones significativos al analizar variables clave como oficina, sexo, rango de edad, estado civil, tipo de canal y motivo de desafiliación.

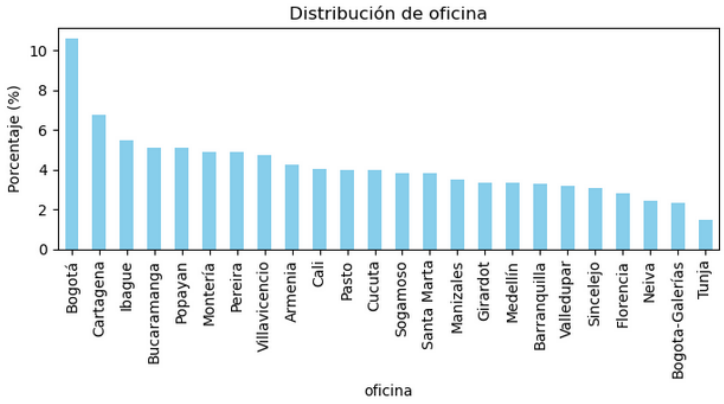
En la Figura 3, la frecuencia de retiros por oficina muestra que la oficina Bogotá concentra el mayor porcentaje de desafiliaciones, con un 10.57%, seguida por las oficinas Cartagena e Ibagué. Este comportamiento indica que ciertas oficinas tienen una mayor incidencia de retiros, lo cual puede estar asociado a factores específicos como atención al cliente, ubicación geográfica o cantidad de asociados.

Figura 3

Frecuencia de Desafiliaciones por Oficina

Tabla de frecuencias para la variable: oficina:

oficina	Frecuencia	Porcentaje (%)
Bogotá	479	10.573951
Cartagena	305	6.732892
Ibague	248	5.474614
Bucaramanga	230	5.077263
Popayan	230	5.077263
Montería	222	4.900662
Pereira	222	4.900662
Villavicencio	215	4.746137
Armenia	192	4.238411
Cali	182	4.017660
Pasto	181	3.995585
Cucuta	180	3.973510
Sogamoso	172	3.796909
Santa Marta	172	3.796909
Manizales	159	3.509934
Girardot	152	3.355408
Medellín	150	3.311258
Barranquilla	148	3.267108
Valledupar	143	3.156733
Sincelejo	138	3.046358
Florencia	128	2.825607
Neiva	110	2.428256
Bogota-Galerías	106	2.339956
Tunja	66	1.456954



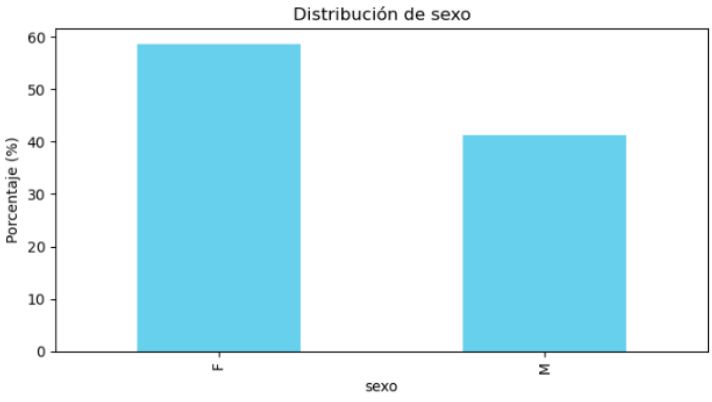
La Figura 4 presenta la frecuencia de retiros por sexo, donde se observa que el 58.65% de las desafiliaciones corresponden a clientes del sexo femenino, mientras que el 41.35% corresponde a clientes del sexo masculino.

Figura 4

Frecuencia de Desafiliaciones por Sexo

Tabla de frecuencias para la variable: sexo

sexo	Frecuencia	Porcentaje (%)
F	2657	58.653422
M	1873	41.346578



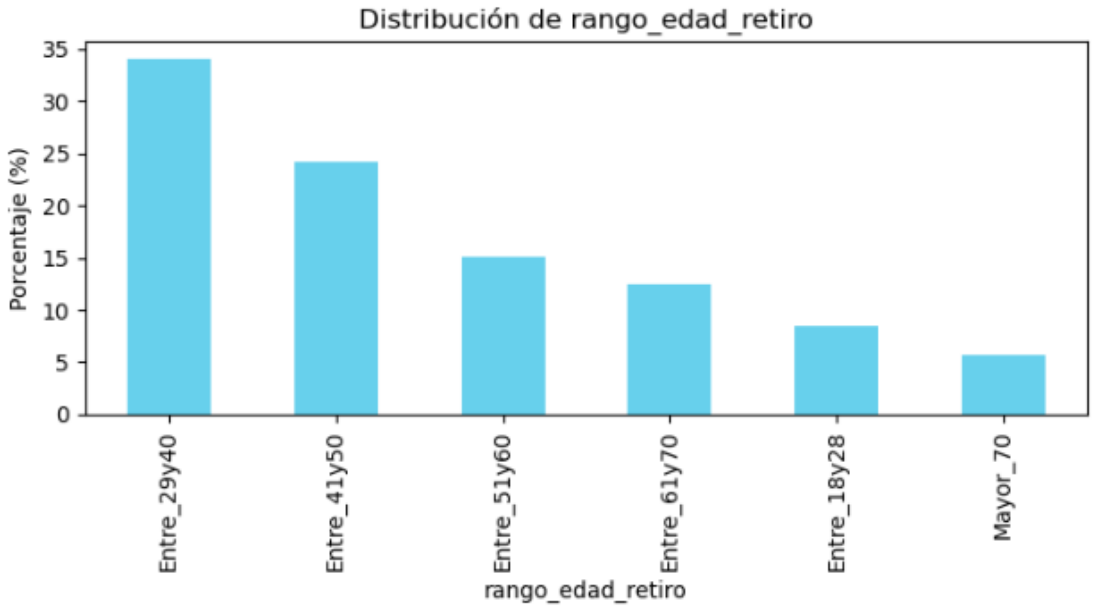
En la Figura 5, la frecuencia por rango de edad revela que el grupo de 29 a 40 años concentra el 34.01% de las desafiliaciones, seguido por el grupo de 41 a 50 años con un 24.15%. Los clientes mayores de 70 años y aquellos entre 18 y 28 años muestran porcentajes más bajos, lo que indica que la edad media tiene una mayor incidencia en las decisiones de retiro.

Figura 5

Frecuencia de Desafiliaciones por Rango de Edad

Tabla de frecuencias para la variable: rango_edad_retiro

rango_edad_retiro	Frecuencia	Porcentaje (%)
Entre_29y40	1541	34.017660
Entre_41y50	1094	24.150110
Entre_51y60	686	15.143488
Entre_61y70	567	12.516556
Entre_18y28	383	8.454746
Mayor_70	259	5.717439



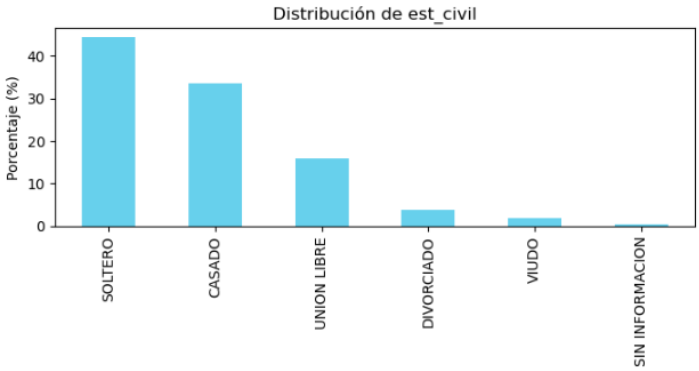
La Figura 6 muestra la frecuencia por estado civil, donde los clientes solteros representan la mayor proporción de desafiliaciones con un 44.33%, seguidos por los casados con un 33.53%. En contraste.

Figura 6

Frecuencia de Desafiliaciones por Estado Civil

Tabla de frecuencias para la variable: est_civil

est_civil	Frecuencia	Porcentaje (%)
SOLTERO	2008	44.326711
CASADO	1519	33.532009
UNION LIBRE	716	15.805740
DIVORCIADO	178	3.929360
VIUDO	87	1.920530
SIN INFORMACION	22	0.485651



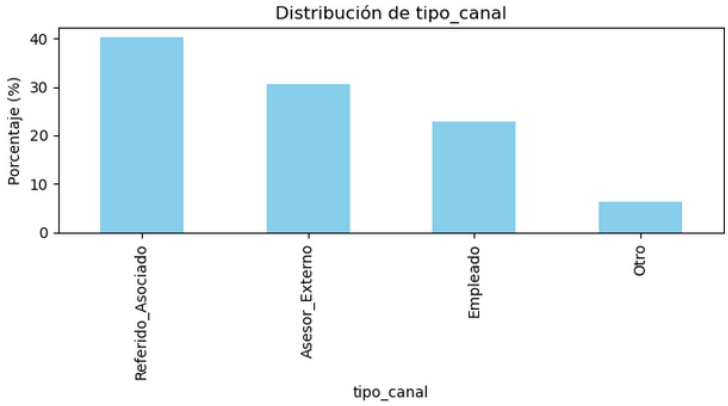
En la Figura 7 presenta la frecuencia de desafiliaciones por tipo de canal, donde se observa que los asociados vinculados a través del canal Referido-Asociado, registran el mayor porcentaje de desafiliación con un 40.15%, seguidos por aquellos ingresados mediante el canal Asesor Externo, que representan un 30.61%, mientras que otros canales como Empleados y Otro tienen menor porcentaje de desafiliación.

Figura 7

Frecuencia de Desafiliaciones por Tipo de Canal

Tabla de frecuencias para la variable: tipo_canal:

tipo_canal	Frecuencia	Porcentaje (%)
Referido_Asociado	1819	40.154525
Asesor_Externo	1387	30.618102
Empleado	1037	22.891832
Otro	287	6.335541



En la Figura 8 nos muestra la frecuencia por motivo de desafiliación, donde se destaca que la "necesidad de aportes" es la causa más frecuente con un 42.05%, seguida por el "cruce de aportes" (11.15%) y la condición de "no sujeto de crédito" (6.84%). Otros motivos, como traslados, negación de crédito y pérdida de empleo, tienen una menor incidencia, pero también aportan a la dinámica general de retiros.

Figura 8

Frecuencia de Desafiliaciones por Motivo de Desafiliación (Top 15)

Tabla de frecuencias para la variable: motivo_desafiliacion

motivo_desafiliacion	Frecuencia	Porcentaje (%)
NECESIDAD DE APORTES	1905	42.052980
CRUCE DE APORTES	505	11.147903
NO SUJETO DE CREDITO	310	6.843267
TRASLADO O VIAJE	251	5.540839
NEGACION DE CREDITO	168	3.708609
BAJA CAPACIDAD PARA PAGO APORTES	158	3.487859
NO CUMPLIO CON LAS EXPECTATIVAS	138	3.046358
DISMINUCION DE INGRESO	118	2.604857
OTRO	118	2.604857
VOY A VIVIR EN EL EXTERIO	99	2.185430
SIN CAPACIDAD DE AHORRO	97	2.141280
PERDIDA DE EMPLEO	71	1.567329
NO USO DEL PORTAFOLIO DE PRODUCTOS SERVICIOS Y ...	67	1.479029
DESEMPLEO	66	1.456954
DISMINUCION DE INGRESOS	64	1.412804

Al analizar la matriz de correlación respecto a la variable "abandono", como se observa en la Figura 9, se evidencia que las variables "ingresos" (0.67), "activos" (0.58) y "cartera" (0.57) muestran una correlación positiva moderada con el abandono, lo que sugiere que niveles más altos en estas características están asociados con una mayor probabilidad de desafiliación. De manera similar, variables como "tarjeta débito" (0.40) y "biometría" (0.07) también presentan correlaciones positivas, aunque de menor magnitud. En contraste, variables con correlación negativa, como "saldo_aho" (-0.07), "edad" (-0.05) y " tiempo_afil (Bishop, 2006)" (-0.03), indican que un incremento en estos valores se asocia con una reducción en la probabilidad de abandono.

Figura 9

Análisis de Correlaciones Ordenadas Respecto a la Variable Abandono

```
# Ordenar las correlaciones respecto a la columna "abandono"
correlacion_abandono = correlaciones['abandono'].sort_values(ascending=False)

# Mostrar las correlaciones ordenadas
correlacion_abandono
```

```
abandono                      1.000000
ingresos                      0.670633
activos                       0.578278
cartera                       0.570997
tarjeta_debito                0.405976
biometria                     0.077275
val_cuota_apo                 0.075246
cuenta_pac                    0.069545
enrrollado                    0.048404
extracto_digital              0.039913
personas_cargo                0.020338
cdats                         0.015941
amnistia                      0.008725
Aportes_Pendientes            0.007906
educacion_coope               0.006182
cuenta_coasmedito             0.000565
antigu_laboral               -0.002134
cuenta_ahorros               -0.004357
reafiliacion                 -0.008149
estrato                      -0.009184
num_hijos                    -0.016942
coaspatitas                  -0.019950
tiempo_afil                  -0.033087
capital_vencido              -0.040116
cartera_castigada            -0.044806
saldo_pfijo                  -0.047554
total_utilizacion_tcredito   -0.050236
total_auxilios               -0.050289
en_trata_datos               -0.050852
total_dep_pfijo              -0.051966
dec_renta                    -0.053026
edad                         -0.056470
reclam_auxilios              -0.064461
mora_cartera                 -0.064928
saldo_aho                    -0.074701
libranza                     -0.079546
afiliacion_virtual           -0.087449
saldo_apo                    -0.090948
total_dep_aho                -0.107431
as_registro_app              -0.131026
tarjeta_credito              -0.154212
monto_desembolsado           -0.195463
Name: abandono, dtype: float64
```

Fase 1 Análisis de Cohortes

Un cohorte hace referencia a un grupo de asociados que comparten una característica común dentro de un periodo específico, en este caso, la fecha de afiliación. Por ejemplo, un cohorte puede estar compuesto por los asociados que ingresaron en enero de 2024, quienes se analizarán de forma conjunta para evaluar su comportamiento de pagos de aportes a lo largo del tiempo. Este análisis permite identificar patrones de retención, cumplimiento y continuidad en los pagos, facilitando así la toma de decisiones estratégicas para mejorar la fidelización y gestión de los asociados.

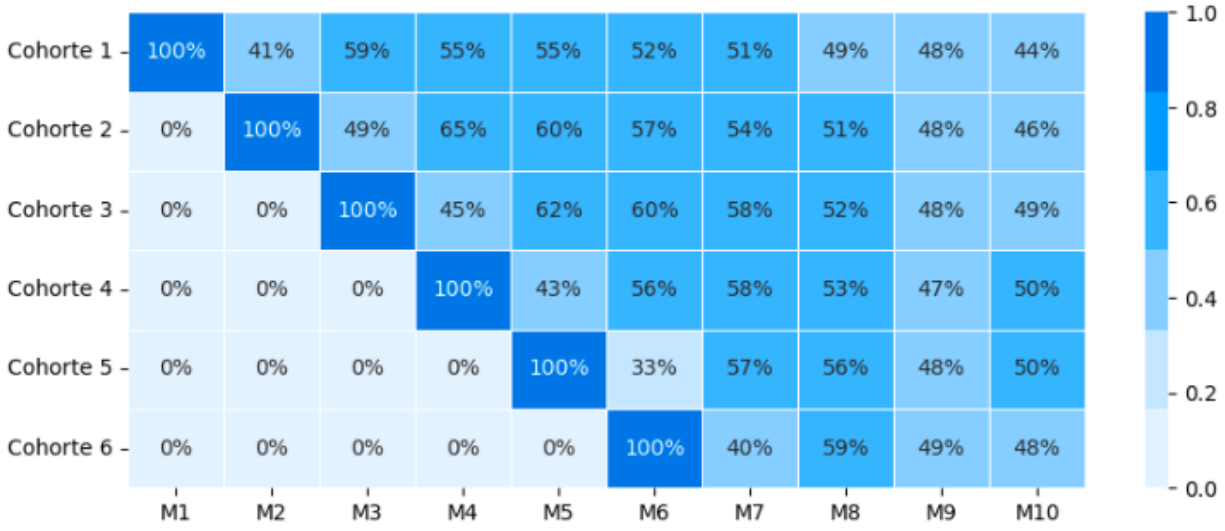
Como se observa en la Figura 10, Los labels en el eje X representan los meses posteriores a la afiliación del asociado. Por ejemplo: M1, Corresponde al primer mes después de la afiliación, donde todos los asociados del cohorte se encuentran activos y realizan sus aportes iniciales. M2, M3, M4...: Representan el segundo, tercer y cuarto mes consecutivo, y así sucesivamente. Cada columna indica la proporción de asociados activos que continúan realizando sus pagos en los meses posteriores, permitiendo visualizar el comportamiento de retención de cada cohorte a lo largo del tiempo.

El análisis de cohortes muestra el comportamiento de los pagos de aportes de los asociados ingresados entre el 01/01/2024 y el 05/31/2024. Se observa que, en el primer mes (M1), la totalidad de los asociados en cada cohorte inicia con un 100% de participación. Sin embargo, a medida que avanzan los meses, se evidencia una disminución progresiva en la proporción de asociados que continúan realizando los pagos. Por ejemplo, en la Cohorte 1, la retención cae del 100% en M1 al 41% en M2 y continúa descendiendo hasta alcanzar un 44% en M10. Esta tendencia decreciente se repite en las demás cohortes, aunque con variaciones en los porcentajes de retención mes a mes. Estos resultados nos indican que, si bien existe un

compromiso inicial, la retención de pagos disminuye significativamente en los meses siguientes, por esta razón es necesario implementar estrategias de fidelización y seguimiento para mantener la participación de los asociados a lo largo del tiempo.

Figura 10

Comportamiento de Pagos de Aportes por Cohortes



Fase 2 Algoritmos No Supervisados

El modelo K-means es ideal para datos numéricos, ya que utiliza la distancia euclidiana como medida de similitud para agrupar a los individuos según variables como la edad, montos de aportes o antigüedad de afiliación (Wu et al., 2023). Por otro lado, el K-modes, como extensión del K-means, permite manejar datos categóricos, utilizando la moda como medida central, lo que resulta adecuado para variables como estado civil, canal de afiliación y tipo de producto o servicio (Huang et al., 2023).

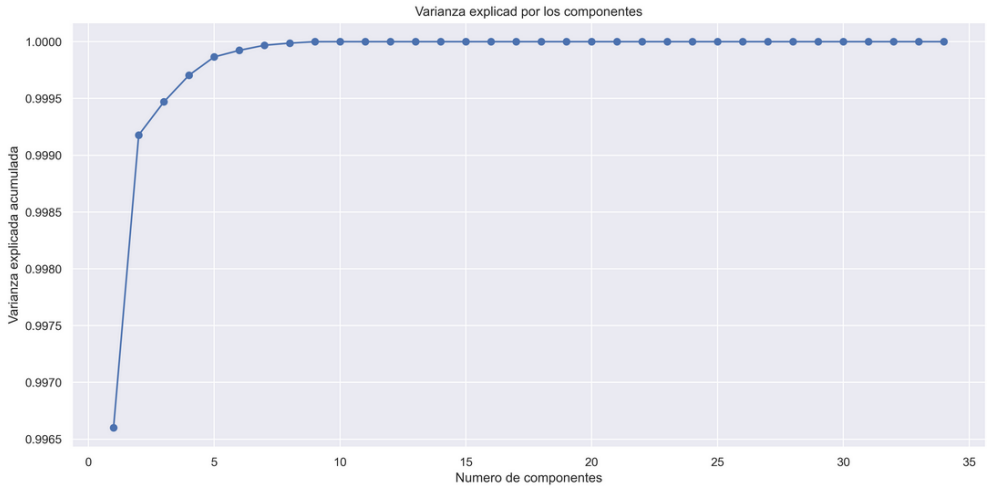
Modelo Kmeans

Al trabajar con datasets de alta dimensionalidad, los algoritmos de clustering, como K-means, pueden verse afectados por el "problema de la maldición de la dimensionalidad", lo cual dificulta la identificación de patrones y la eficiencia computacional. PCA permite resolver este problema al generar componentes principales, que son combinaciones lineales de las variables originales, ordenadas según la cantidad de varianza que capturan (Jolliffe & Cadima, 2016).

En la Figura 11, se observa la varianza explicada acumulada en función del número de componentes principales seleccionados mediante PCA. El gráfico muestra que los primeros componentes capturan la mayor parte de la varianza presente en los datos. En particular, se evidencia un punto de inflexión alrededor del componente 5, donde la tasa de crecimiento de la varianza explicada se reduce significativamente, acercándose al 100% con un número limitado de componentes. La selección de este número óptimo de componentes no solo mejora la eficiencia computacional del modelo K-means, sino que también facilita la interpretación de los clústeres generados en el espacio reducido.

Figura 11

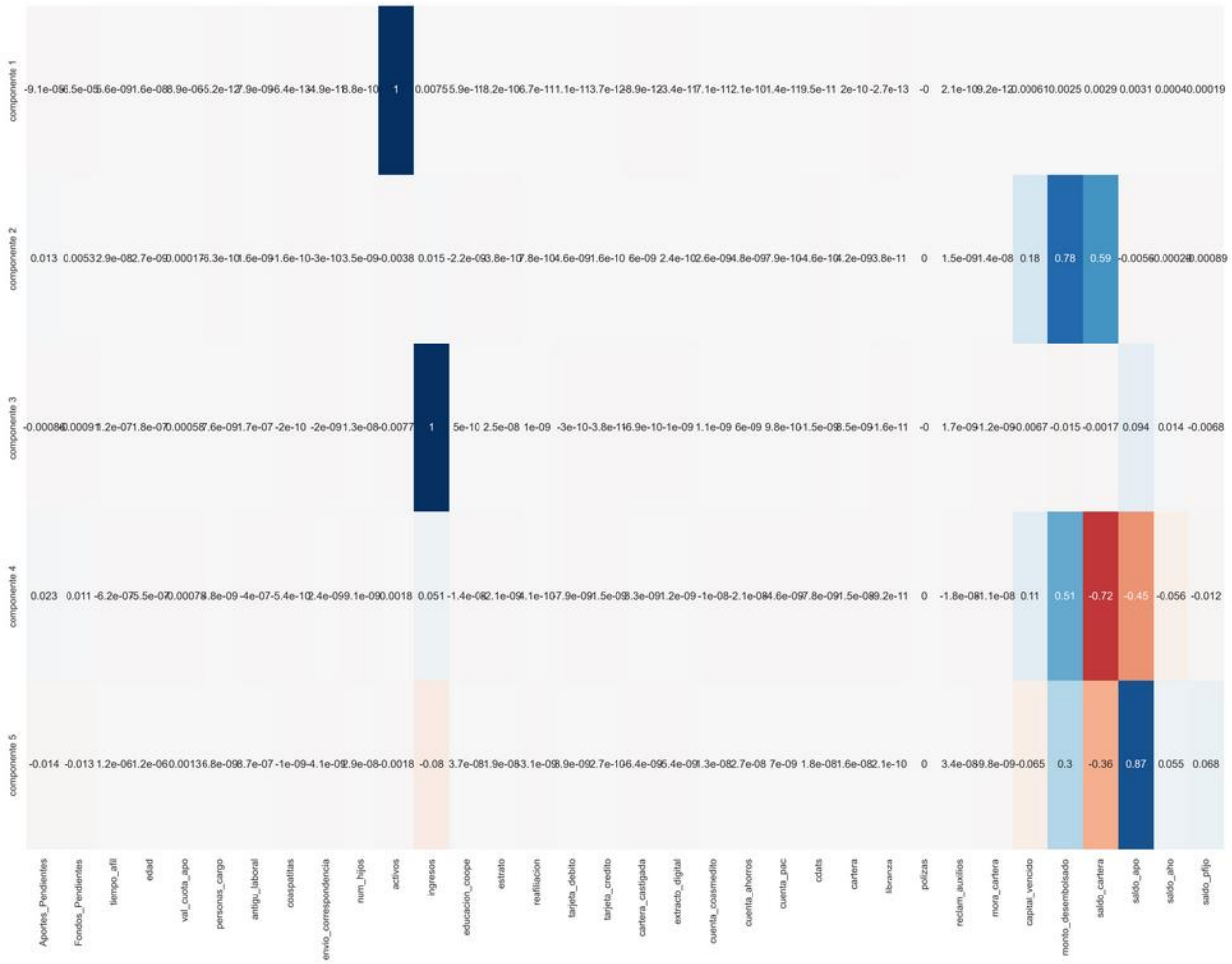
Curva de Varianza Explicada por Componentes Principales



En la figura 12, muestra la matriz de componentes principales (PCA), donde se destacan las variables más influyentes en cada componente principal según los colores y valores de las celdas. Los valores altos, positivos o negativos, sugieren correlaciones fuertes entre las variables y los componentes, lo que puede ayudar a identificar patrones o factores subyacentes en el análisis, como características financieras o de riesgo. Este análisis facilita reducir la dimensionalidad del conjunto de datos, destacando las variables más relevantes para interpretaciones y decisiones posteriores.

Figura 12

Relación de Componentes Principales con Variables Originales

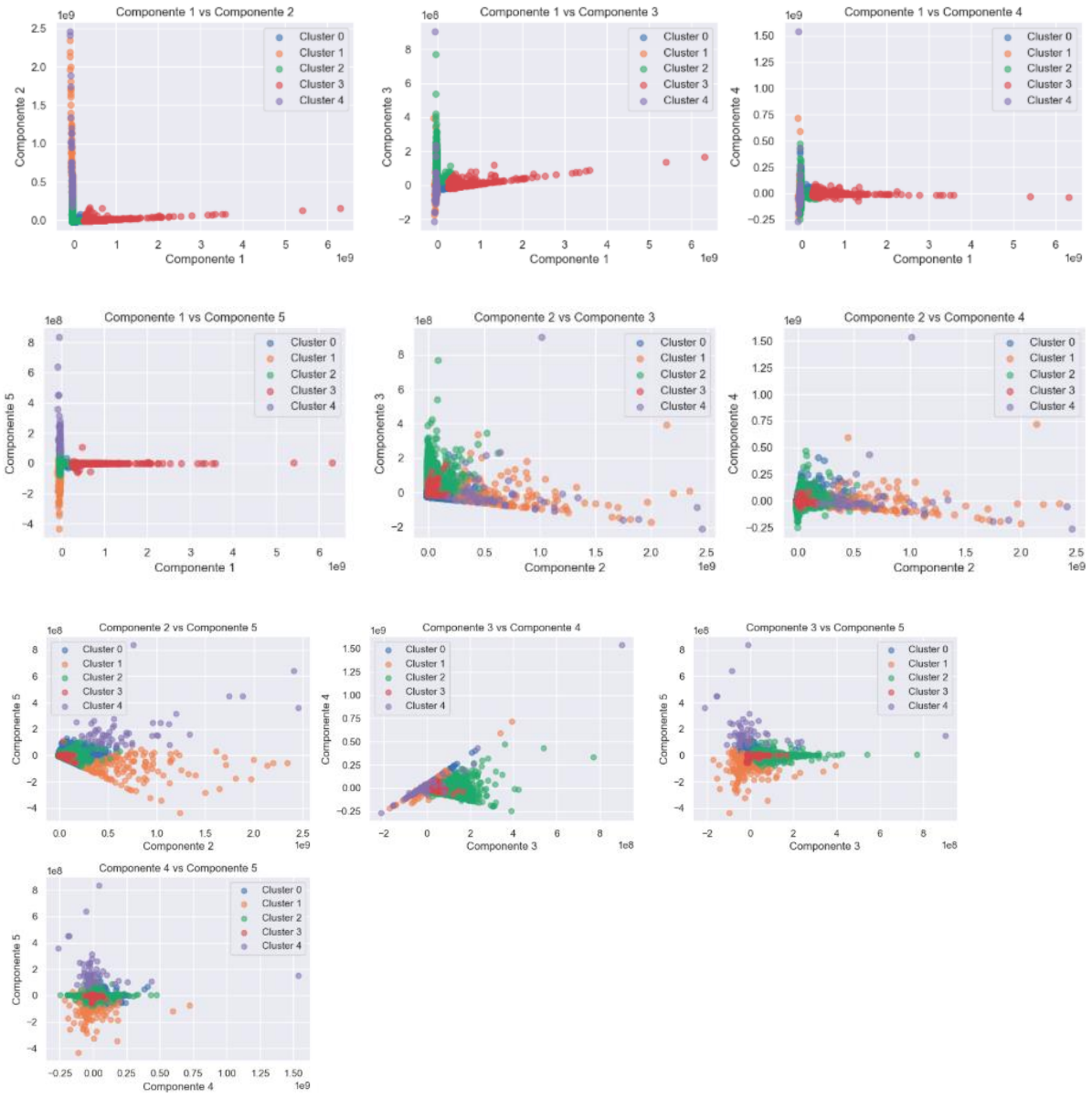


En los gráficos observados en la figura 13, se muestran relaciones bidimensionales entre los diferentes componentes principales (PCA) coloreados por clústeres, lo que facilita identificar patrones y agrupaciones en los datos. Cada combinación de componentes muestra la distribución de puntos pertenecientes a los clústeres 0 a 4, con concentraciones claras en algunos casos y dispersiones amplias en otros. Los clústeres parecen estar bien diferenciados en ciertos componentes, como los pares de “Componente 1 vs Componente 2” y “Componente 3 vs Componente 4”, mientras que, en otras combinaciones, como “Componente 4 vs Componente

5”, hay una mayor superposición. Este análisis ayuda a visualizar cómo los clústeres se separan en el espacio de componentes y qué combinaciones aportan mayor discriminación, lo que puede guiar la interpretación de factores clave en el agrupamiento.

Figura 13

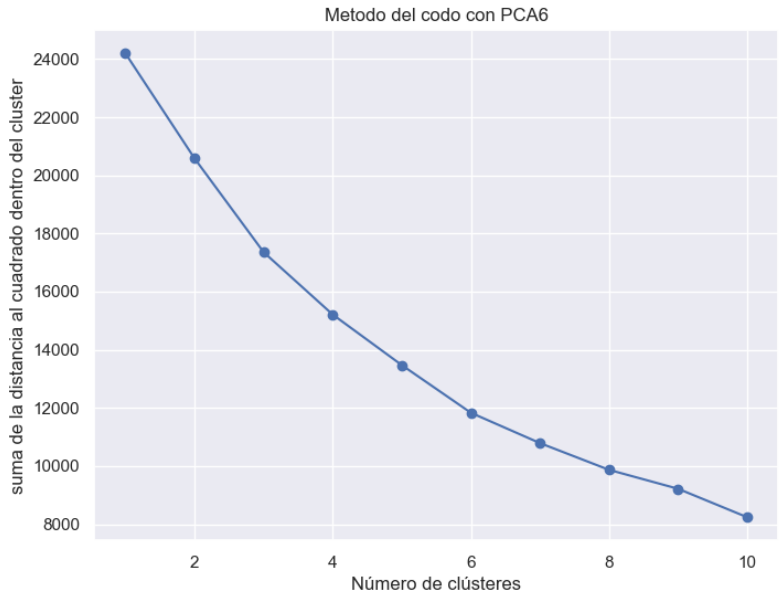
Distribución de Clústeres según Combinaciones de Componentes



A medida que aumenta el número de clústeres, la suma de las distancias disminuye, ya que cada clúster captura mejor los datos. El método del codo (figura 14), identifica el número óptimo de clústeres como el punto donde la disminución de la suma de distancias se vuelve menos pronunciada, formando una especie de “codo” en la curva. Según el gráfico, 4 o 5 clústeres son una buena elección, ya que representan un balance entre simplicidad del modelo y calidad del agrupamiento.

Figura 14

Método del Codo para Determinar la Cantidad de Clústeres



En la figura 15, encontramos el código con los parámetros para el entrenamiento del modelo, donde definimos el número de clústeres según la interpretación de la técnica del codo, el algoritmo se inicializa 10 veces (n_init=10) con diferentes posiciones aleatorias para los centroides, seleccionando la mejor solución que minimice la distancia dentro de cada clúster. Además, el parámetro random_state=99 asegura la reproducibilidad del modelo, es decir, que los resultados sean consistentes cada vez que se ejecute el código.

Figura 15

Implementación del Modelo K-Means con Datos Reducidos por PCA

```
# de acuerdo a la interpretación del grafico el codo está en k = 5
cluster_k5_pca5 = KMeans(n_clusters = 5, n_init = 10, random_state = 99)
```

```
# hacemos el fit del modelo con los scores pca6 estándarizados
cluster_k5_pca5.fit(pca5_std)
#cluster_k5_pca5.fit(5)
```

```
KMeans
KMeans(n_clusters=5, n_init=10, random_state=99)
```

```
KMeans(n_clusters=5,random_state = 99)
```

```
KMeans
KMeans(n_clusters=5, random_state=99)
```

```
# concatenar datafame con las variables originales y las latentes
df_k5_pca5 = pd.concat([df_matriz4.reset_index(drop = True), pd.DataFrame(scores_pca5)],axis = 1)
```

En la Figura 16, se observa la calificación del modelo con un Silhouette Score de 0.591, lo cual indica que los clústeres presentan una definición aceptable, con una separación moderada entre grupos y un solapamiento mínimo. Si bien no alcanza un valor óptimo (cercano a 1), este resultado nos indica que el modelo de clustering, ha logrado una segmentación coherente y funcional, permitiendo identificar patrones relevantes en los datos.

Figura 16

Cálculo del Silhouette Score para Evaluación del Modelo

```
print(f"Silhouette Score calculado por bloques: {silhouette_avg:.3f}")
```

```
Silhouette Score calculado por bloques: 0.591
```

Como se observa en la figura 17, el Clúster 0 concentra la mayor cantidad de abandonos con 2923 registros, lo que lo convierte en el grupo más crítico y de mayor riesgo. Por otro lado,

el Clúster 3 presenta 1012 registros, ubicándose en una posición intermedia, mientras que el Clúster 2 muestra solo 25 registros, indicando una baja incidencia de deserción y representando un grupo más estable. Estos resultados sugieren la necesidad de implementar estrategias de retención específicas para los clientes del Clúster 0 y profundizar en las causas del abandono en el Clúster 3.

Figura 17

Cantidad de Clientes que Han Desertado por Cada Clúster

```
# Conteo de registros por combinación de cluster y abandono
df_abandono_1.groupby(['clusters5', 'abandono']).size().reset_index(name='cantidad')
```

	clusters5	abandono	cantidad
0	0	1	2923
1	2	1	25
2	3	1	1012

Modelo K-Modes (Variables Categóricas)

La implementación del modelo K-Modes en este proyecto tiene como objetivo realizar una segmentación de los asociados, considerando variables de tipo categórico que son críticas para comprender los patrones de comportamiento y posibles factores de deserción. K-Modes está diseñado específicamente para manejar datos categóricos. Esta característica hace que sea una herramienta ideal para analizar información como sexo, profesión, clase de vivienda, nivel de estudios, tipo de contrato, ocupación, actividad económica, estado civil, tipo de canal, atributos cualitativos que son relevantes para el proyecto.

En la Figura 18, se presenta la evaluación del modelo K-Modes, donde el análisis de cohesión, calculado mediante la distancia de Hamming, arroja un valor de 0.0941. Este resultado

nos muestra que los puntos dentro de cada clúster son cercanos entre sí y exhiben una alta homogeneidad categórica, lo que indica que los clústeres están bien definidos internamente. En consecuencia, se evidencia una buena calidad en la segmentación de los datos, validando la capacidad del modelo para agrupar observaciones con características similares.

Figura 18

Cohesión del Modelo K-Modes Evaluada con Distancia de Hamming

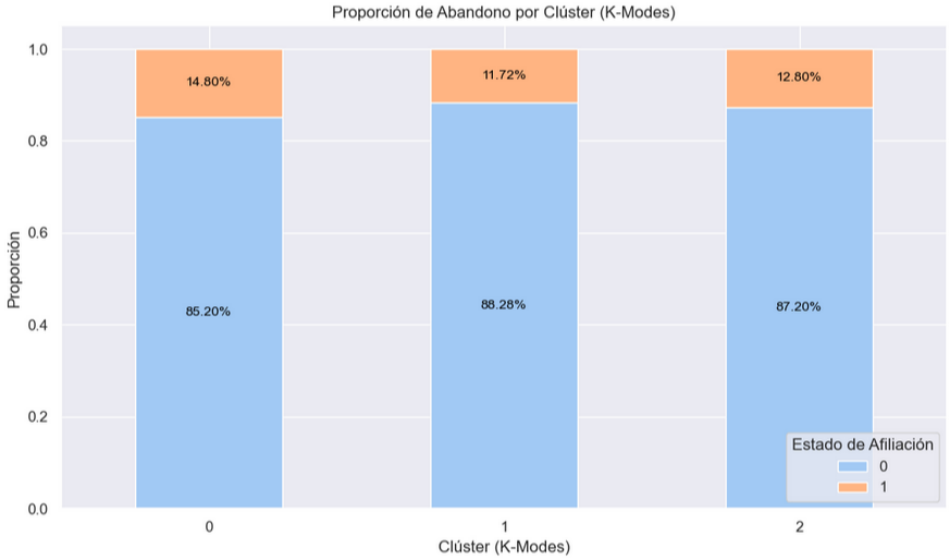
```
# Evaluación basada en cohesión
cohesion = np.mean([hamming_distancias[i, cluster] for i, cluster in enumerate(clusters_kmodes)])
print(f"Cohesión de K-Modes (Hamming): {cohesion:.4f}")
```

```
Cohesión de K-Modes (Hamming): 0.0941
```

En la figura 19, muestra la proporción de abandono en cada clúster identificado mediante el modelo K-Modes. Se observan tres clústeres (0, 1 y 2) con una distribución similar en términos de abandono y permanencia. El Clúster 1 presenta el menor porcentaje de abandono, con un 11.72%, los asociados en este grupo tienen características más estables en comparación con los otros clústeres. Por otro lado, el Clúster 0 registra un 14.80% de abandono, siendo el grupo con el mayor riesgo de desafiliación. Finalmente, el Clúster 2 muestra una proporción de abandono del 12.80%, ubicándose en una posición intermedia.

Figura 19

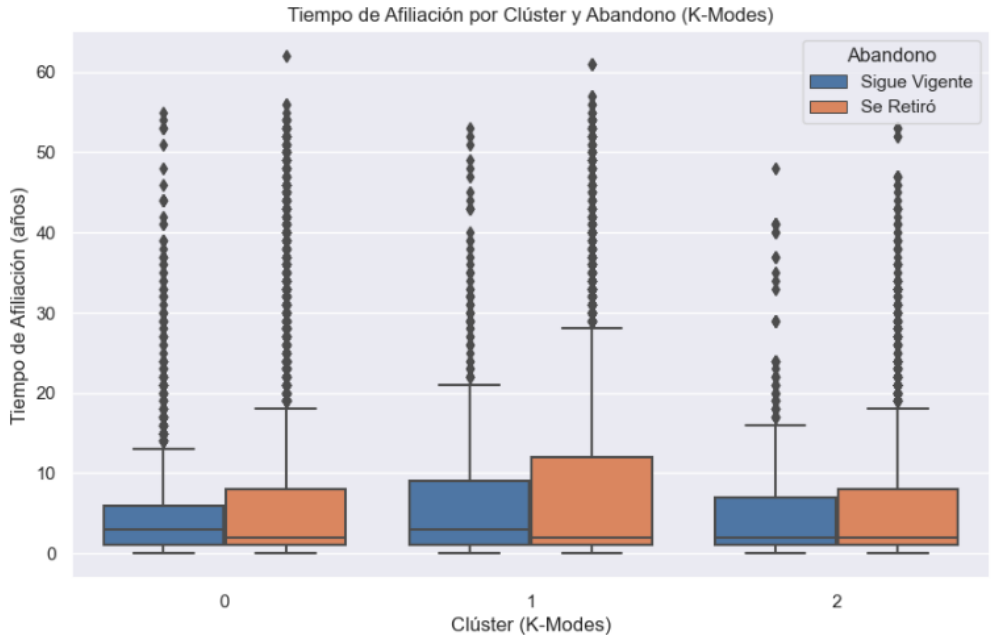
Distribución de Abandono y Permanencia porClúster



En la Figura 20, se observa la comparación del tiempo de afiliación entre los clústeres generados mediante K-Modes y el estado de abandono. El gráfico revela que los asociados que se retiraron tienden a tener tiempos de afiliación ligeramente mayores en comparación con aquellos que siguen vigentes, aunque existen traslapes significativos entre ambos grupos. Las distribuciones de los clústeres son similares, con medianas cercanas y la presencia de valores atípicos en tiempos de afiliación más largos, particularmente en el grupo de abandonos. Este resultado sugiere que, si bien el tiempo de afiliación podría tener cierta influencia en el abandono, no es un factor determinante.

Figura 20

Distribución del Tiempo de Afiliación



Fase 3 Algoritmos Supervisados

La aplicación de algoritmos cómo el Árbol de Decisión y Random Forest, permite no solo segmentar a los asociados de manera eficiente, sino también generar insights valiosos sobre los posibles desencadenantes de la deserción, facilitando así la toma de decisiones orientadas a la retención de clientes.

Algoritmo Árbol de Decisión

Para construir un Árbol de Decisión es fundamental realizar el balanceo de clases, ya que un desbalance significativo puede provocar que el modelo se sesgue hacia la clase mayoritaria, reduciendo su capacidad para identificar correctamente la clase minoritaria y afectando la calidad de las predicciones.

En la Figura 21, se observa el porcentaje original de clases para la variable objetivo abandono, donde la clase "0" (no abandono) representa un 89.06% de los datos, mientras que la clase "1" (abandono) corresponde apenas al 10.93%, evidenciando un desbalance significativo entre ambas categorías. Este desbalance podría sesgar el modelo predictivo hacia la clase mayoritaria, reduciendo su capacidad para identificar correctamente los casos de abandono. Por otro lado, en la Figura 22, se muestra la distribución balanceada de clases después de aplicar técnicas de balanceo, logrando una distribución equitativa con 24,866 registros en cada clase (0 y 1). Este balance garantiza que el modelo tenga una representación adecuada de ambas clases, lo que permite mejorar la precisión en la predicción de abandonos y mitigar el sesgo hacia la clase mayoritaria.

Figura 21

Porcentaje de Clases Originales

```
porcentaje_clase = (conteo_clases / len(df['abandono'])) * 100
```

```
porcentaje_clase
```

```
abandono
0    89.063841
1    10.936159
Name: count, dtype: float64
```

Figura 22

Distribución de Clases Equilibradas

```
Y_equilibrado.value_counts()
```

```
abandono
0    24866
1    24866
Name: count, dtype: int64
```

En el reporte de clasificación observado en la Figura 23, El modelo presenta un rendimiento excepcional, con una precisión, recall y F1-score de 1.00 para ambas clases (No abandono y Abandono), lo que indica que clasifica correctamente casi todos los casos. La matriz de confusión muestra únicamente 6 errores (3 falsos positivos y 3 falsos negativos) en un total de 8649 muestras, lo que representa menos del 0.07% del total. Además, el AUC de 0.9995 refleja una capacidad casi perfecta para discriminar entre las clases. Estas métricas sugieren que el modelo está altamente ajustado a los datos, pero podría estar sobreajustado, por esta razón se debe validar su rendimiento en un nuevo conjunto de datos y adicional se realizará un nuevo modelo con los componentes principales resultado de PCA.

Figura 23

Reporte de Clasificación

```
Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      7477
           1       1.00      1.00      1.00      1172

    accuracy                           1.00      8649
   macro avg       1.00      1.00      1.00      8649
weighted avg       1.00      1.00      1.00      8649

Matriz de Confusión:
 [[7474    3]
 [   3 1169]]
AUC: 0.9995499851421492
```

Algoritmo Árbol de Decisión con PCA

El PCA desempeña un papel crucial al reducir la dimensionalidad, condensando la información de las variables originales en un número menor de componentes principales, sin perder una cantidad significativa de información. La combinación de PCA y Árbol de Decisión facilita la creación de un modelo más eficiente y robusto, mejorando la capacidad de generalización al eliminar ruido y redundancias presentes en los datos originales.

En el código presentado en la figura 24, el valor de pca = PCA(n_components=0.95) permite aplicar el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos, garantizando la conservación del 95% de la varianza. Este enfoque es eficiente para manejar datasets con un alto número de características, especialmente aquellas que son redundantes o altamente correlacionadas, facilitando así un procesamiento más rápido y la mejora del rendimiento del modelo.

Figura 24

Reducción de Dimensionalidad por PCA

```
# Aplicar PCA para reducir la dimensionalidad
pca = PCA(n_components=0.95)  # Retener el 95% de la varianza
X_train_pca = pca.fit_transform(X_train_bal_scaled)
X_test_pca = pca.transform(X_test_scaled)

# Verificar el número de componentes seleccionados
print(f"Componentes seleccionados por PCA: {pca.n_components_}")
```

```
Componentes seleccionados por PCA: 76
```

El reporte de clasificación del modelo de Árbol de Decisión entrenado con datos transformados por PCA (figura 25) muestra un desempeño sólido, con una precisión global del 92% y un AUC de 0.905. La clase 0 presenta una precisión de 95%, un recall de 96% y un f1-score de 96%, lo que indica que el modelo identifica correctamente la mayoría de los casos de esta clase mayoritaria. Por otro lado, la clase 1, aunque tiene un rendimiento menor debido a su menor representación en los datos, presenta una precisión de 72%, un recall de 71% y un f1-score de 71%, lo que sugiere que el modelo aún logra captar correctamente una parte considerable de los casos de abandono. La matriz de confusión revela que existen 320 falsos positivos y 342 falsos negativos, lo cual implica que el modelo tiene ligeras dificultades para predecir la clase minoritaria. Sin embargo, el resultado general indica una buena capacidad predictiva, con un balance entre precisión y recall adecuado para ambas clases.

Figura 25

Evaluación del Modelo de Árbol de Decisión con PCA

```
# Configuración del árbol con hiperparámetros ajustados
tree = DecisionTreeClassifier(
    max_depth=10,
    min_samples_split=2,
    min_samples_leaf=20,
    class_weight='balanced',
    random_state=42
)


# Entrenar el modelo
tree.fit(X_train_pca, y_train_bal)

# Predicciones
y_pred = tree.predict(X_test_pca)
y_pred_proba = tree.predict_proba(X_test_pca)[:, 1]

# Evaluación
print("Reporte de Clasificación:\n", classification_report(y_test, y_pred))
print("Matriz de Confusión:\n", confusion_matrix(y_test, y_pred))
print("AUC:", roc_auc_score(y_test, y_pred_proba))
```

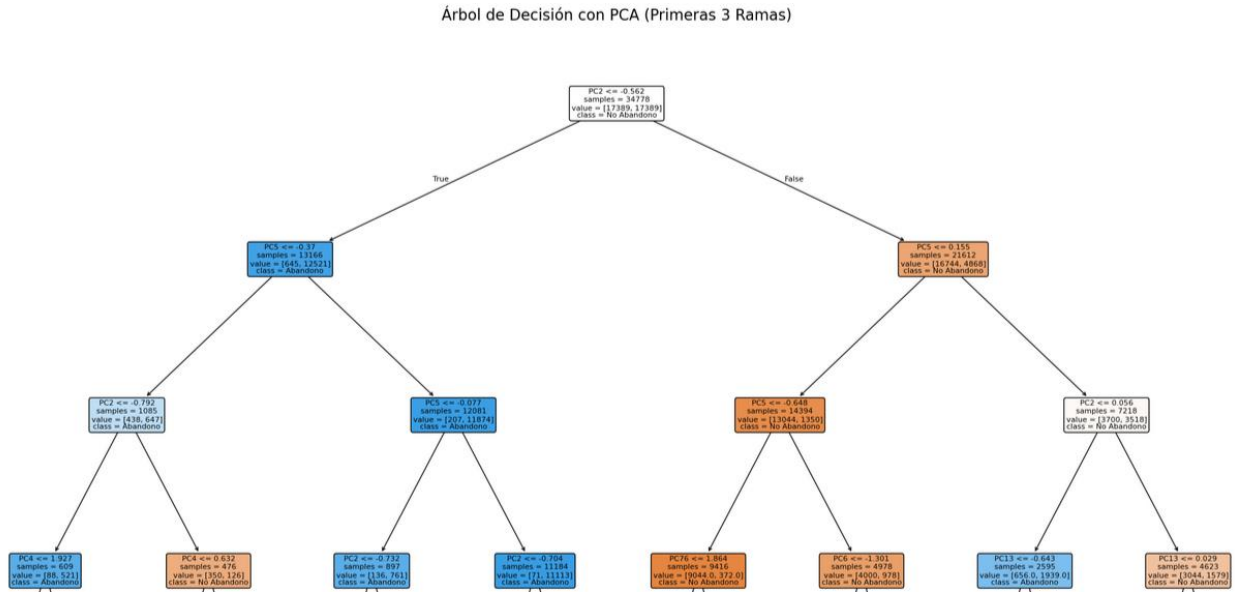
```
Reporte de Clasificación:
               precision    recall  f1-score   support

           0       0.95      0.96      0.96      7477
           1       0.72      0.71      0.71      1172

    accuracy                           0.92      8649
   macro avg       0.84      0.83      0.84      8649
weighted avg       0.92      0.92      0.92      8649

Matriz de Confusión:
 [[7157  320]
 [ 342  830]]
AUC: 0.9052438285143838
```

En la Figura 26 se observan las primeras tres ramas del árbol de decisión, donde la componente PC2 es la principal variable de división inicial, separando los datos en dos grupos en función del abandono y no abandono. Los valores más bajos de PC2 (≤ -0.562) se asocian predominantemente con casos de abandono, mientras que en la rama derecha (PC2 > -0.562), la componente PC5 toma relevancia al generar nuevas divisiones que refuerzan la separación entre ambas clases. En las ramas siguientes, PC4 y PC13 contribuyen a la segmentación más específica, destacando una concentración mayor de abandono en ciertos nodos con valores críticos de estas componentes. Este análisis nos indica que PC2 y PC5 capturan patrones

importantes para identificar el abandono, permitiendo al modelo clasificar con mayor precisión los diferentes grupos de afiliados en función de su comportamiento.

Figura 26

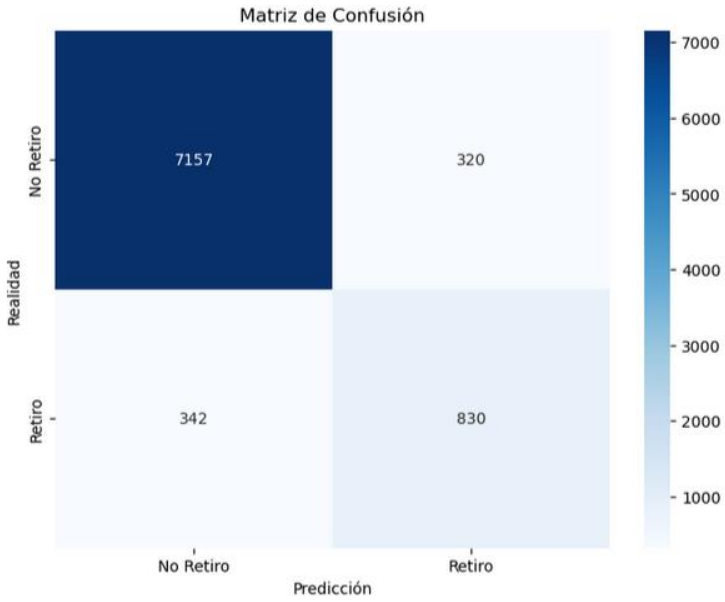
Árbol de Decisión con PCA



La Matriz de Confusión presentada en la Figura 27 muestra el desempeño del modelo en la clasificación de los estados de Retiro y No Retiro. El modelo ha logrado predecir correctamente 7157 casos de No Retiro y 830 casos de Retiro, evidenciando un buen rendimiento en la identificación de la clase mayoritaria (No Retiro). Sin embargo, se observan 342 falsos negativos, es decir, casos de Retiro que fueron incorrectamente clasificados como No Retiro, y 320 falsos positivos, que representan casos donde el modelo predijo Retiro cuando realmente correspondían a No Retiro. Aunque el modelo muestra una alta precisión general en la clase predominante, el rendimiento en la predicción de la clase minoritaria (Retiro) es más limitado, se requiere la necesidad de ajustar parámetros del modelo para mejorar la sensibilidad hacia esta categoría.

Figura 27

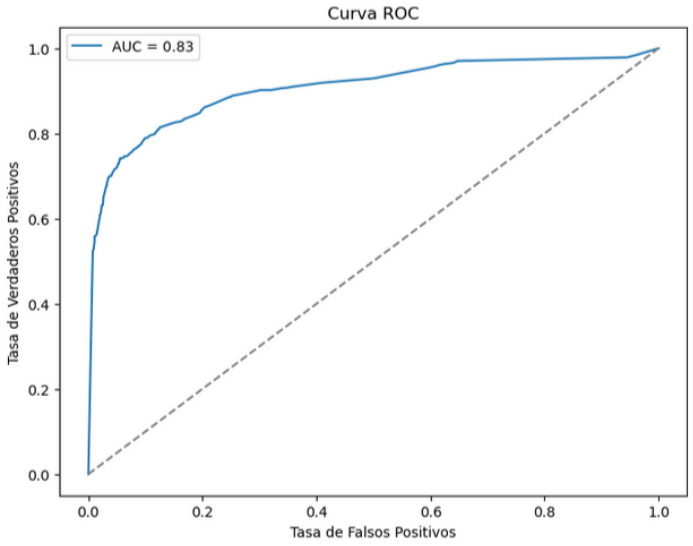
Matriz de Confusión del Modelo de Árbol de Decisión



El gráfico de la curva ROC (Figura 28) muestra el desempeño del modelo al clasificar correctamente las clases en diferentes umbrales de probabilidad. El AUC (Área Bajo la Curva) es de 0.83, lo que indica un buen rendimiento del modelo en la discriminación entre las clases positivas y negativas. La curva está notablemente por encima de la línea diagonal, esto confirma que el modelo tiene una capacidad razonable para diferenciar entre clases. Sin embargo, hay margen de mejora para acercarse a un AUC más alto (cercano a 1.0), lo que podría lograrse ajustando hiperparámetros, el balance de clases o probando otros modelos.

Figura 28

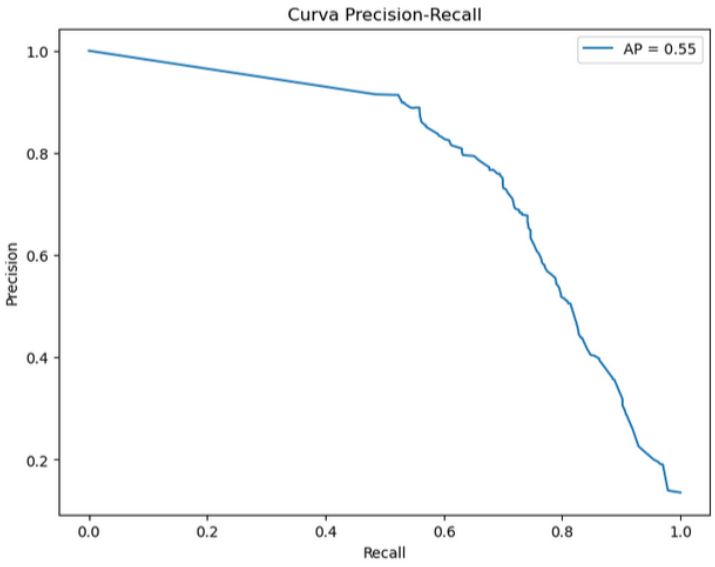
Curva ROC del Modelo del Árbol de Decisión



La curva Precisión-Recall (Figura 29) evalúa el equilibrio entre la precisión y el recall del modelo en diferentes umbrales. Un AP (Average Precisión) de 0.55 nos indica que el modelo tiene un rendimiento moderado para identificar correctamente la clase positiva (Abandono). La curva comienza con alta precisión y recall bajos, pero a medida que aumenta el recall, la precisión disminuye considerablemente, lo que indica que el modelo enfrenta dificultades para mantener un buen balance entre estas métricas, especialmente cuando intenta capturar más casos positivos. Se pueden implementar ajustes, como optimizar el umbral de clasificación o probar modelos más complejos para lograr un mejor rendimiento en la clase positiva.

Figura 29

Curva Precisión-Recall del Modelo Árbol de Decisión



Algoritmo Random Forest

Para este modelo, se genera un conjunto de variables dummy a partir de las variables categóricas, permitiendo transformar dichas variables en un formato numérico adecuado para su procesamiento por el algoritmo.

El reporte de clasificación (figura 30) y la matriz de confusión (figura 31), muestran una evaluación detallada del desempeño del modelo. El modelo logró una precisión general del 94% y un área bajo la curva (AUC) de 0.949, lo cual indica una excelente capacidad discriminativa entre las clases. La clase "No Retiro" presenta una precisión de 97%, un recall de 96% y un f1-score de 96%, demostrando que el modelo identifica de manera efectiva la mayoría de los casos de no retiro. Por otro lado, la clase "Retiro" obtuvo una precisión de 75%, un recall de 79% y un f1-score de 77%, si bien la identificación de retiros es aceptable, aún hay margen de mejora. La matriz de confusión refuerza estos resultados, mostrando 925 predicciones correctas para la clase

de "Retiro" y 247 falsos negativos, lo que indica que algunos casos de retiro no fueron detectados. El modelo presenta un desempeño robusto en la clasificación general, aunque se debe seguir optimizando para mejorar la detección de la clase minoritaria (Retiro).

Figura 30

Evaluación del Modelo Random Forest

```
# Evaluación
print("Reporte de Clasificación:\n", classification_report(y_test, y_pred_nuevo))
print("Matriz de Confusión:\n", confusion_matrix(y_test, y_pred_nuevo))
```

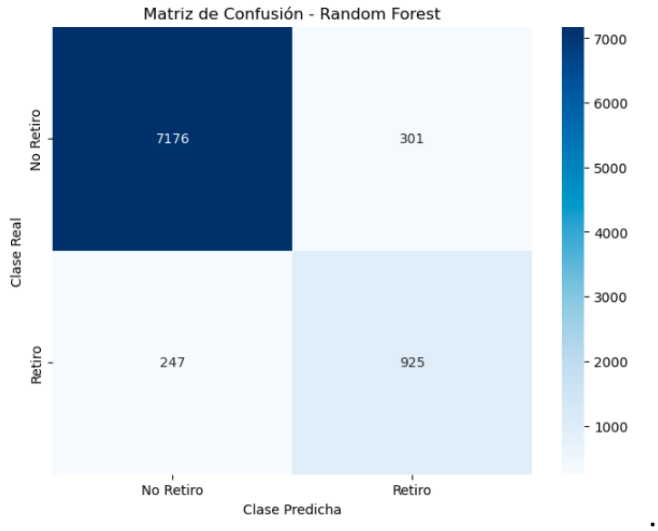
```
Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.97      0.96      0.96      7477
           1       0.75      0.79      0.77      1172

    accuracy                           0.94      8649
   macro avg       0.86      0.87      0.87      8649
weighted avg       0.94      0.94      0.94      8649

Matriz de Confusión:
 [[7176  301]
 [ 247  925]]
```

```
print("AUC:", roc_auc_score(y_test, y_pred_proba))
```

```
AUC: 0.9491899161980699
```

Figura 31

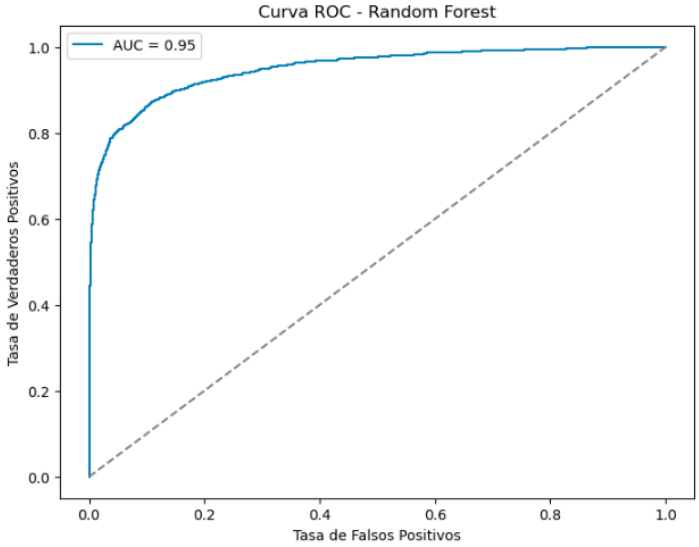
Matriz de Confusión del Modelo Random Forest



La Curva ROC presentada en la figura 32, evalúa el desempeño del modelo Random Forest para predecir el abandono de clientes. La curva muestra una alta separación entre las tasas de verdaderos positivos y falsos positivos, lo que indica una excelente capacidad del modelo para distinguir entre las clases. El área bajo la curva (AUC) es de 0.95, un valor cercano a 1, lo que sugiere que el modelo tiene un desempeño muy robusto y una alta precisión en la clasificación. Esto significa que el modelo puede identificar correctamente a los clientes propensos al abandono con un bajo nivel de falsos positivos, proporcionando resultados confiables para la toma de decisiones en estrategias de retención.

Figura 32

Curva ROC del Modelo de Random Forest

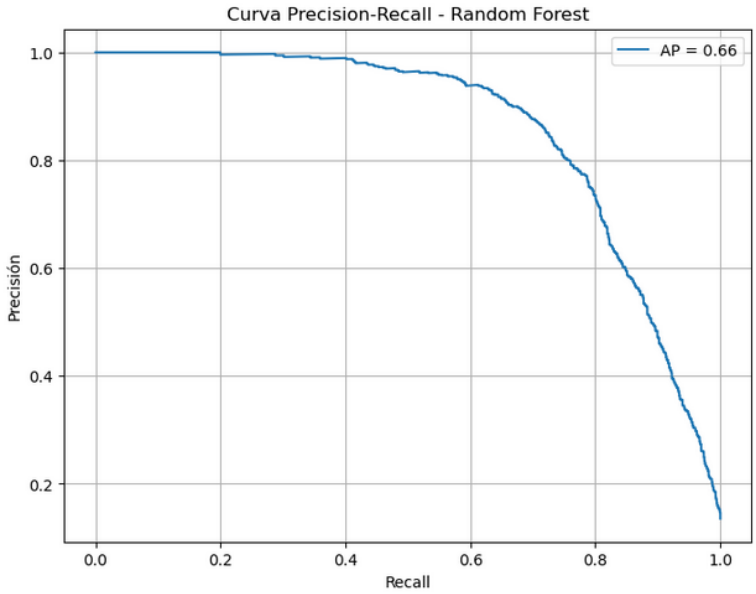


En la figura 33, se observa que la curva Precision-*Recall* muestra un buen rendimiento inicial con una alta precisión para valores bajos de recall, lo que significa que cuando el modelo clasifica menos casos como “Retiro, los identifica con alta precisión. Sin embargo, a medida que aumenta el recall, la precisión disminuye, indicando un incremento en los falsos positivos. Resalta la importancia de seleccionar un umbral que logre un balance adecuado según las

prioridades del negocio, como maximizar la identificación de retiros reales sin generar demasiadas falsas alarmas.

Figura 33

Curva de Precisión - Recall del Modelo Random Forest



Insights

- ✓ La "necesidad de aportes" destaca como la principal causa de desafiliación, seguida de “No sujeto a Crédito”. Esto sugiere que la situación económica de los asociados juega un papel crítico en la decisión de retiro.
- ✓ El tiempo de afiliación, la morosidad en cartera, y el nivel de interacción con los productos financieros de la empresa se destacan como variables clave en la predicción de desafiliación. Esto demuestra que factores financieros y de engagement son cruciales para comprender el abandono.
- ✓ Se identificó que los patrones de abandono no son homogéneos, sino que varían según el tiempo de afiliación y el ciclo de vida del cliente, lo que subraya la importancia de un monitoreo continuo de las cohortes.
- ✓ Los clientes en riesgo utilizan menos los productos ofrecidos por la empresa, lo que indica una desconexión con los servicios o una falta de percepción de valor.
- ✓ El análisis sugiere que incluir variables cualitativas, como encuestas de satisfacción, podría mejorar la comprensión de los motivos de desafiliación, ofreciendo una perspectiva más amplia sobre los factores psicológicos o emocionales.

KPIs

- ✓ Porcentaje de clientes que se retiran y permanecen afiliados en un período determinado.
- ✓ Porcentaje de clientes que utilizan los productos de la empresa.
- ✓ Tiempo promedio que permanece el cliente antes de retirarse.
- ✓ Identificar qué clústeres tienen mayores índices de desafiliación.
- ✓ Porcentaje de clientes en riesgo que responden positivamente a las estrategias implementadas.
- ✓ Porcentaje de clientes que se retiran y que fueron correctamente identificados como "en riesgo" por el modelo.
- ✓ Impacto financiero de los retiros en términos de pérdida de ingresos y costos de captación de nuevos clientes.
- ✓ Indicador basado en encuestas que mide la percepción de los clientes sobre los servicios ofrecidos.

Conclusiones

Se logró identificar patrones de comportamiento y factores clave que influyen en la desafiliación de los clientes. Entre los principales factores se encuentran variables como la cartera, los ingresos y el tiempo de afiliación, los cuales presentan una correlación significativa con el abandono.

La implementación de algoritmos como Random Forest, Árboles de Decisión y K-means demostró ser efectiva para predecir y segmentar el riesgo de abandono. El uso de PCA permitió reducir la dimensionalidad de los datos sin perder información relevante, mejorando la eficiencia computacional y la calidad del análisis.

Los clústeres generados mediante técnicas no supervisadas permitieron identificar subgrupos de clientes con características específicas, facilitando la personalización de estrategias de retención. Se observó que el clúster predominante agrupa a la mayoría de los clientes desertores, destacando la importancia de enfoques dirigidos en este segmento.

La integración de técnicas avanzadas de machine learning y análisis de datos no solo mejoró la precisión en la predicción de deserción, sino que también proporcionó insights clave para el desarrollo de estrategias proactivas y personalizadas de retención.

Aunque los modelos presentan buenos desempeños en términos de precisión y AUC, algunos como el Árbol de Decisión muestran señales de sobreajuste. Esto subraya la necesidad de validar los modelos en datos no vistos para garantizar su generalización.

Recomendaciones

Crear programas de beneficios personalizados para jóvenes y nuevos clientes, como educación financiera, microcréditos accesibles y actividades de integración, para mejorar su percepción de valor y compromiso.

Monitorear el comportamiento de las cohortes a lo largo del tiempo para identificar tendencias de abandono y evaluar la efectividad de las estrategias implementadas, ajustándolas en función de los resultados observados.

Continuar aplicando técnicas para abordar el desbalance de clases y explorar modelos híbridos que combinen diversos algoritmos para mejorar la predicción de la clase minoritaria (abandonos).

Ajustar hiperparámetros y probar técnicas adicionales como ensambles más complejos (XGBoost o LightGBM) para mejorar el rendimiento en la identificación de retiros, especialmente en la clase minoritaria.

Analizar variables adicionales como satisfacción del cliente, interacción en canales digitales y uso de productos específicos para enriquecer los modelos y mejorar la capacidad predictiva.

Establecer métricas continuas de seguimiento como la tasa de retención, tasas de abandono por clúster y efectividad de las intervenciones, que permitan evaluar el impacto de las estrategias implementadas en tiempo real.

Crear sistemas automáticos que alerten sobre clientes en riesgo y ofrezcan recomendaciones personalizadas para retenerlos, basados en los patrones identificados por los modelos predictivos.

Referencias Bibliográficas

Ahmad, N., et al. (2020). *Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques. IEEE International Conference on Artificial Intelligence and Machine Learning (AIML*).

Allegue, S., Abdellatif, T., & Bannour, K. (2020). *RFMC: A Spending Categories Segmentation Using Machine Learning and Multiple Criteria Techniques*. 29th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms*. MIT Press.

Dawane, V., Waghodekar, P., & Pagare, J. (2021). *RFM Analysis Using K-means Clustering for Revenue and Customer Retention Improvement. Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

Hamming, R. W. (1986). *Numerical Methods for Scientists and Engineers*. Dover Publications.

Huang, Z., Chen, L., & Wang, X. (2023). *A comparative study of clustering algorithms for categorical data analysis*. Journal of Data Science and Analytics, 12(3), 243–258. https://doi.org/10.1007/s13253-023-01435-1

Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: A review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability*. Chelsea Publishing Company.

Naoui, M. A., Ayad, M., & Lejdel, B. (2020). *Modelos de regresión para sistemas de big data: Enfoques basados en MapReduce*. Revista Cubana de Ciencias Informáticas, 14(2), 34-48. Recuperado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992020000200034&lang=pt

Polya, G. (1945). How to Solve It: A New Aspect of Mathematical Method. Princeton University Press.

Quiñones, A. & Huiman, H. (2021). *Resolución de problemas con el método matemático de polya: la aventura de aprender*. HTTPS://WWW.REDALYC.ORG/JOURNAL/280/28071845006/.

Rana, A., Salman, R., Sabeer, S., Ansari, S. K., Aarif, M., & Das, L. (2023). *Customer Retention Using Machine Learning*. 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON).

Supraja, P., & Sairamesh, L. (2023). *Customer segmentation using machine learning*. Proceedings of the Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1–9). IEEE. https://doi.org/10.1109/ICAECT57570.2023.10117924

Wu, L., & Li, M. (2018). *Aplicar el método de regresión logística CG para predecir el problema de pérdida de clientes*. Quinta Conferencia Internacional sobre Sistemas de Economía Industrial e Ingeniería de Seguridad Industrial (IEIS), IEEE.

Wu, H., Li, T., Zhang, Q., & Xu, F. (2023). *K-means clustering for large-scale numerical data: And optimization perspective*. Applied Soft Computing, 130, 109721. https://doi.org/10.1016/j.asoc.2023.109721

Zhang, M., Zhang, Z., & Qiu, S. (2020). *A Customer Segmentation Model Based on Affinity Propagation and Improved K-means Algorithm. International Conference on Intelligent* Information Processing. Springer.