# PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR


FACULTAD INGENIERÍA

MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN CIENCIA DE DATOS

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE MÁSTER EN SISTEMAS DE INFORMACIÓN MENCIÓN CIENCIA DE DATOS

TEMA:

**IMPLEMENTACIÓN DE UN MODELO DE SEGMENTACIÓN DE SOCIOS PARA LA IDENTIFICACION DE CAPACIDADES DE PAGO Y RIESGO DE MOROSIDAD DE LOS SOCIOS DE LA COOPERATIVA DE AHORRO Y CREDITO JARDIN AZUAYO**

**AUTOR**: JUAN GABRIEL GOYO CANDO

**DIRECTOR**: VLADIMIR BONILLA, Ph.D.

QUITO, FEBRERO 2024

## DEDICATORIA

Becky, con tu simple acto de recostarte a mi lado mientras las horas se desvanecían frente a la pantalla del ordenador, me diste la fuerza para seguir adelante. Tu entusiasmo al recibirme, y acompañarme sin importar la hora, me recordaba que siempre hay razones para sonreír y compartir estos logros con los seres amados. Este logro, aunque académico en su naturaleza, lleva impreso tu espíritu y tu inquebrantable compañía.

Siempre serás parte de mi ser y de mi éxito en cada logro de mi vida...

## AGRADECIMIENTO

Agradezco a mi familia, principalmente a mi esposa por acompañarme directamente en este logro académico cada día en el que tuvimos que sacrificar el tiempo y organizar nuestras actividades para adquirir este conocimiento nuevo que nos ayudará en el futuro.

Un agradecimiento especial al Dr. Vladimir Bonilla, tutor de este trabajo de graduación por su guía y apoyo, y a todo el personal académico de la PUCE, que aportaron con su conocimiento para enseñarnos de la mejor manera las diferentes habilidades que ahora forman parte de nuestro perfil profesional.

Finalmente, gracias a los compañeros de la Cooperativa Jardín Azuayo y principalmente a mi equipo de trabajo, especialmente a la unidad de Ingería de Datos por facilitarme la información necesaria para este trabajo, compartiendo su tiempo y conocimiento para ayudarme a conseguir este logro.

# INDICE DE FIGURAS

INDICE DE TABLAS

# CAPITULO I: INTRODUCCION

## 1. Introducción

En el ámbito financiero, la identificación precisa de las capacidades de pago de los clientes es fundamental para garantizar una gestión eficiente de los recursos, minimizar los riesgos crediticios y ofrecer productos adecuados a diferentes segmentos de clientes. En el caso específico de la Cooperativa de Ahorro y Crédito Jardín Azuayo, una entidad financiera sólida y de gran envergadura, la automatización de un modelo de scoring de segmentación de clientes se presenta como una herramienta estratégica para mejorar la evaluación de la capacidad de pago de sus socios.

Actualmente, la cooperativa tiene múltiples oficinas donde los socios (clientes) de la cooperativa acceden a los diferentes productos y servicios, así como también varios canales digitales donde se pueden acceder a los mismos servicios. El análisis de la información de los socios es un proceso que actualmente se lo realiza para establecer su capacidad de endeudamiento y pago al momento de adquirir obligaciones crediticias con la Cooperativa. Los oficiales de crédito son los encargados de realizar este análisis y determinar los riesgos de morosidad de los socios y la capacidad de endeudamiento capaces de adquirir. Estos procesos están establecidos acorde con las políticas institucionales de seguridad financiera y uso responsable de los recursos, los mismos que lamentablemente se ejecutan de forma manual. Sin embargo, esta información obtenida no tiene mayor utilidad hasta el momento ya que no hay una iniciativa de explotación de la información que permita utilizar estos datos de tal manera que los mismos clientes tipificados dentro de ciertas capacidades crediticias, bien podrían también clasificarse en grupos o segmentos para que a través de ese conocimiento la Cooperativa pueda promocionar y colocar nuevos servicios financieros o refinar algunos ya existentes que se apeguen más a las necesidades y características reales de los socios.

La automatización de un modelo de scoring para la identificación de las capacidades de pago de los socios de la cooperativa impactará directamente en la efectividad y eficiencia del proceso de otorgamiento y promoción de los diferentes productos de la cooperativa ya que reduce el riesgo de morosidad y agiliza la toma de decisiones reduciendo el sesgo y sugiriendo productos adecuados de acuerdo con la capacidad de endeudamiento de los socios o su comportamiento de consumo. El uso de técnicas de análisis de datos y aprendizaje automático mejorarán la precisión y la rapidez con la cual las áreas de negocio

diseñan productos más personalizados a los socios, permitiendo una toma de decisiones más fundamentada y reduciendo los riesgos asociados a la provisión de estos servicios financieros, mejorando significativamente los ingresos y beneficios para la institución.

### 1.1. Antecedentes

El uso de herramientas y tecnología de analítica de datos en instituciones financieras no es un tema necesariamente reciente. Porras Castaño (2018), establece que los clientes de las instituciones financieras exigen la transformación digital en todos los productos y servicios que ellas proveen, y las compañías han reconocido los beneficios que esto conlleva para sus resultados financieros, mediante la mejora de los procesos internos y una toma de decisiones eficiente.

Las metodologías de análisis de datos e inteligencia artificial han estado presentes durante muchos años, pero su popularidad, sobre todo en bancos y cooperativas del país, se debe al surgimiento del Big Data, el mismo que facilita la gestión y procesamiento eficiente de grandes volúmenes de información. Cuanto mayor sea la cantidad de datos, más precisa será la detección de patrones y comportamientos, por lo que contar con cantidades significativas de información y la capacidad de procesarla rápidamente, incluso en tiempo real, resulta crucial (Porras Castaño, 2018).

En ese sentido, estas capacidades deben ser aprovechadas para poder brindarle a los socios de la cooperativa servicios adecuados a sus capacidades de endeudamiento o consumo, y sobre todo que sean de acceso y resultados inmediatos, sin necesidad de atarse a procesos presenciales y aprovechando las tecnologías que permiten acceder a los servicios de la cooperativa desde cualquier lugar y en cualquier momento.

Un ejemplo del uso de estas tecnologías es lo que está realizando el Banco Solidario. De acuerdo con Víctor Morales Oñate, Gerente de analítica de datos, Banco Solidario ha incursionado en el siguiente aspecto tecnológica de analítica de datos:

> El índice de cobertura espacial fue desarrollado por el área de analítica, y analiza el área de influencia de las agencias de este banco en distintas ciudades y la cobertura de potenciales clientes entregando información que servirá para decidir si se reubica las agencias o es necesario crear una nueva. En una dinámica que parte de la demanda, o también sobre propuestas propias tratamos de apoyar la decisión comercial a través de datos (Morales, 2021)

Por otro lado, el Banco Guayaquil, una de las más grandes entidades financieras, mediante la implementación del entorno Oracle BigData Services y la creación de SMART CAMPAIGN, buscan automatizar las campañas de marketing y fortalecer la interacción con sus clientes como parte de su estrategia empresarial (Oracle, 2015)

Todo esto indica que las instituciones financieras incursionan desde hace tiempo en la analítica de datos, en especial en la mejora de procesos internos que les brinden ventajas competitivas que aporten en el cumplimiento de objetivos estratégicos con una visión futurista.

### 1.2. Descripción del problema

Actualmente, en la Cooperativa de Ahorro y Crédito Jardín Azuayo, hay dificultad y lentitud en la toma de decisiones financieras y crediticias informadas, tanto de las sucursales como de las unidades de negocio que manejan tanto las estrategias de colocación de capital (crédito), como el diseño y promoción de productos y servicios, lo que aumenta el riesgo de morosidad de los clientes, pérdidas financieras para la cooperativa y una ineficiente promoción y entrega de servicios.

Se pueden definir otros problemas derivados de lo indicado anteriormente:

- Procesos de definición y promoción de productos y servicios son lentos y burocráticos
- Situación económica a nivel nacional aumenta el nivel de morosidad y posibles pérdidas debido a un análisis de información deficiente.
- Evaluación y segmentación manual de clientes propensa a errores y subjetividades
- Métricas inadecuadas para el seguimiento continuo del desempeño crediticio de los clientes y de los tiempos de gestión del acceso y contratación de servicios.
- Tiempo elevado para trámites de solicitud de servicios
- Herramientas tecnológicas inadecuadas no permiten automatizar el análisis y seguimiento de la capacidad de pago y el riesgo de morosidad

Podemos esquematizar esta sucesión de problemas a través de un diagrama causa-efecto (Diagrama Ishikawa) como se ve en la figura 1.

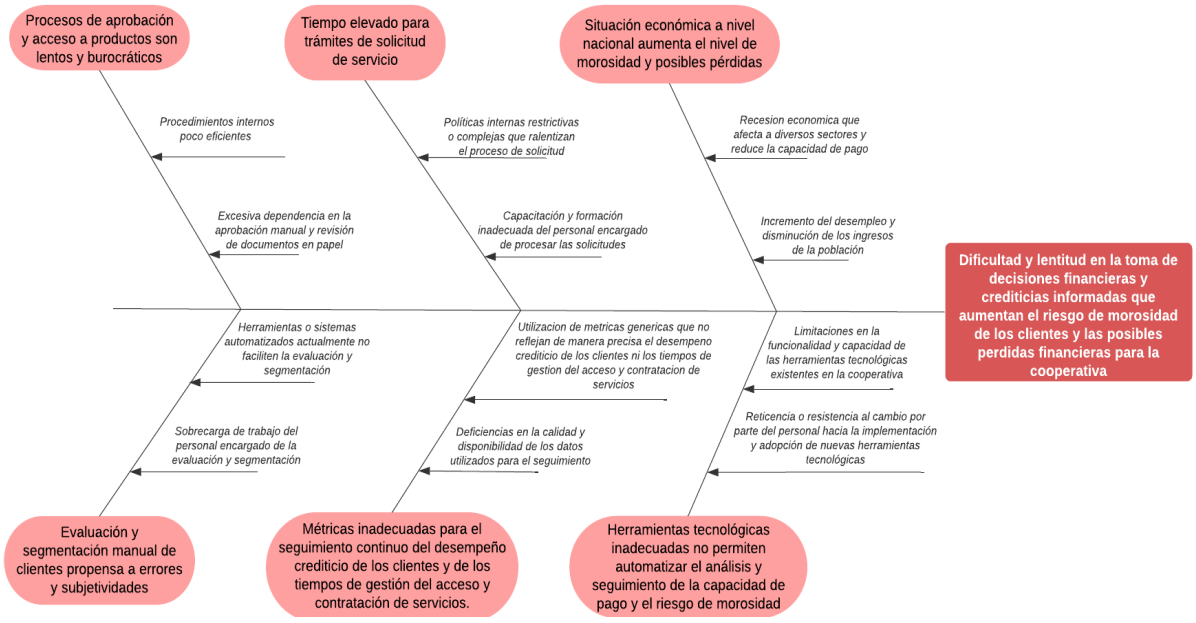



*Figura 1: Diagrama causa-efecto para la problemática establecida en COAC Jardín Azuayo. Fuente: Elaboración propia*

## 1.3. Objetivos

A continuación, se detallan los objetivos que persigue el presente trabajo de titulación

### *1.3.1. Objetivo general*

Automatizar un modelo de segmentación de clientes mediante técnicas de analítica de datos y aprendizaje automático

### *1.3.2. Objetivos específicos*

- Levantar información de los modelos actuales de tipologías de clientes, basado en el modelo actual de evaluación de riesgo crediticio.
- Identificar las técnicas y herramientas necesarias de analítica de datos aplicadas al análisis y perfilamiento de clientes
- Desarrollar modelos de aprendizaje automático basados en analítica de datos que automatice el proceso de toma de decisiones para la promoción y generación de productos y servicios financieros
- Evaluar la efectividad de las decisiones establecidas por parte de los algoritmos implementados comparándolas con las decisiones humanas en base a la norma actual.

## 1.4. Justificación

La analítica de datos hoy en día es un pilar fundamental en los procesos de toma de decisiones en todos los sectores de la industria. En el caso de las instituciones financieras no es la excepción y es uno de los sectores que actualmente están a la vanguardia en el uso cotidiano de tecnología para extraer, transformar y presentar los datos de las instituciones, para transformarse en información que les permite obtener una ventaja competitiva al tomar decisiones que apoyen al cumplir las estrategias empresariales, o a la generación de estas.

Para la Cooperativa de Ahorro y Crédito Jardín Azuayo, una de sus estrategias dentro del plan estratégico es precisamente mejorar y optimizar sus procesos internos y desarrollar procesos de minería de datos para la gestión de riesgos y mejorar la gestión de la calidad de sus productos y servicios. Ello permitirá superar la problemática descrita, ya que las técnicas actuales de analítica y minería de datos, así como las tecnologías actuales de automatización y aprendizaje automático mejorarán el proceso de segmentación de socios y su nivel de endeudamiento al basar la decisión en el análisis automatizado del posible riesgo de morosidad o capacidad de pago según la información personal, financiera e histórica de cada socio de la cooperativa.

Actualmente, la cooperativa no ha realizado un ejercicio real de determinar perfiles de clientes basado en la información interna que se dispone, ya que únicamente se trabaja en análisis de riesgo crediticio. Esto limita el accionar de las áreas de innovación y diseño de servicios puesto que al momento de generar o mejorar nuevos productos o servicios, no tienen una visión clara de cuáles son los socios o grupos de socios quienes pueden tener un interés real de consumo de dichos servicios, dentro de un nivel de riesgo adecuado que permita obtener los réditos suficientes que mejoren el y mantengan la sostenibilidad de los recursos que usa la Cooperativa para proporcionar dichos servicios financieros.

En este caso, las áreas de riesgos y análisis de datos requieren una solución que permita establecer de manera inicial un perfilamiento o segmentación de socios, para definir con un nivel inicial de certeza cuantas personas con determinadas características pueden ser sujetas de usar de manera adecuada los servicios financieros, que tengan las capacidades suficientes para cumplir con las obligaciones adquiridas con la Cooperativa, y que le

permitan a la institución mejorar sus ingresos y utilidades con los productos que se proporcionen a los socios

### 1.5. Alcance

El presente trabajo de titulación pretende explorar algunas estrategias de aprendizaje automático no supervisado (clustering), buscando gemnerar un modelo de segmentación de clientes que posteriormente permita profundizar el conocimiento para un perfilamiento de socios en base a ciertas características que ayuden a definir qué productos o servicios son los más apropiados según el grupo o segmento donde estos clientes sean clasificados. Con este fin, se pretende probar algunos modelos de aprendizaje no supervisado, donde cada enfoque comprende ciertas características y particularidades, y la elección del modelo más adecuado depende de la naturaleza de los datos obtenidos y objetivo final de esta exploración.

Tras seleccionar la estrategia o modelo que mejor se alinea con nuestros objetivos específicos y que tenga el mejor resultado al evaluarse, se construirá un modelo de segmentación usando una notebook para automatizar el proceso análisis de los socios.

La información que servirá como base de este análisis serán los socios personas naturales que tengan al menos una operación crediticia en la cooperativa a la fecha de corte del reporte (31 de noviembre de 2023), para ubicarlos en un segmento y luego definir que parámetros de negocio pueden aplicarse en cada segmento para determinar los posibles productos o servicios apropiados para cada persona según el resultado del análisis ejecutado.

Finalmente, la notebook con el modelo construido será socializada con los Departamentos de Riesgos y Análisis de Datos para constancia de la ejecución de este proyecto, y se entregará al Departamento de Análisis de Datos para su uso diario, una vez que se sean cpacitados y entrenados en el uso de la herramienta Jupyter Notebook y nociones de programacion Python para análisis de datos.

# CAPITULO II: MARCO TEORICO Y CONCEPTUAL

## 2. Marco teórico y conceptual

### 2.1. Importancia de la evaluación de riesgo en las instituciones financieras

La representación del mercado financiero a lo largo de los siglos refleja un desarrollo caracterizado por una aceleración en la innovación de instrumentos financieros y una creciente complejidad de la evaluación de riesgos.

De acuerdo con González-Duany (2021), este desarrollo corrió en paralelo a episodios disruptivos de crisis bursátil y financiera, que fueron desencadenados por una supervisión bancaria global inadecuada y una tendencia a una alta asunción de riesgos y dejaron profundas huellas en la estructura económica. Las diferentes crisis financieras mundiales han marcado un punto de inflexión y que puso de manifiesto la necesidad imperativa de mantener niveles óptimos de liquidez en las instituciones financieras para mantener la estabilidad del sistema financiero nacional y global (González-Duany, 2021).

A nivel de nuestro país, esta realidad reveló la fragilidad de muchas instituciones financieras condenadas al fracaso, lo que provocó acontecimientos con importantes pérdidas de empleos, absorciones y fusiones de varias instituciones, sobretodo del ámbito de la economía popular y solidaria; y finalmente, fuimos testigos también de la recapitalización forzada de muchas instituciones de la banca tradicional.

Esta situación, exacerbada por una supervisión bancaria inadecuada, un apalancamiento excesivo, una calidad inadecuada y deficiente del capital y la incapacidad del sector para adaptarse a las tensiones financieras y económicas, ha puesto sobre la mesa la vulnerabilidad de los mercados financieros (González-Duany, 2021). Por tanto, evaluar adecuadamente los riesgos de incumplimiento de obligaciones financieras, y proveer de productos y servicios acorde a las capacidades económicas de cada cliente de las instituciones, es vital hoy para minimizar las posibles pérdidas y gastos que reduzcan la liquidez necesaria para que bancos y cooperativas sigan prestando sus servicios de manera eficiente y efectiva a sus clientes.

## 2.2. Conceptos clave para la evaluación de riesgos y análisis de comportamiento de clientes

### 2.2.1. Riesgo de crédito

Elizondo & Altman (2004) definen el "riesgo de crédito" como la pérdida potencial ocasionada por el hecho de que un deudor o contraparte incumpla con sus obligaciones según los términos establecidos.

El riesgo de crédito puede verse desde dos puntos de vista: por un lado, desde la perspectiva de los activos financieros, ya que poseer un instrumento de deuda lo expone al riesgo de contraparte (riesgo de emisor), y por otro lado, desde la perspectiva de los activos crediticios, lo cual supone estar expuesto a un riesgo de incumplimiento (Elizondo & Altman, 2004).

De la misma manera, los autores establecen que, tal y como ocurre en todas las instituciones financieras, existe también el riesgo asociado a las denominadas carteras de crédito, el cual aborda la integralidad y diversidad de estos conjuntos financieros crediticios desde una perspectiva holística. Al analizar desde esta óptica las pérdidas potenciales de las operaciones crediticias, es imperativo contemplar la variabilidad y concentración de los créditos involucrados, así como las interrelaciones de riesgo existentes entre ellos (Elizondo & Altman, 2004).

Las instituciones financieras, al albergar en sus carteras de préstamos tanto a individuos como a entidades con perfiles semejantes, descubren en esta homogeneidad una herramienta simplificadora del análisis; es particular, cuando los deudores comparten aspectos comunes como su localización geográfica, sector de operación o escala de operaciones.

Elizondo & Altman (2004), son claros en señalar que la proyección de un escenario desfavorable al momento de analizar a un grupo o grupos de prestatarios sugiere una degradación paralela en la solvencia de aquellos de características afines, manifestando así una correlación palpable entre sus destinos financieros. La esencia del análisis de riesgo crediticio de cartera radica en discernir la presencia de concentraciones específicas, ya sea por actividad económica, ubicación geográfica, edad, estado civil, cargas familiares, etc., fundamentándose en las sinergias entre los miembros de la cartera. Tal

enfoque, incentiva a las entidades financieras a perseguir una diversificación más robusta en su portafolio de préstamos, como estrategia para mitigar riesgos inherentes.

### 2.2.2. Modelo de score crediticio

Rodríguez-Guevara et al. (2017), establecen que los modelos de score crediticio son métodos estadísticos o matemáticos especializados en pronosticar e identificar a un cliente que puede arriesgarse a incumplir un préstamo. Estos modelos se usan principalmente para aprobar préstamos, determinar la clasificación de los préstamos, asignar precios de los préstamos, generar alertas tempranas y estrategias de cobro de deudas. Cabe señalar que los métodos de evaluación crediticia utilizados en los bancos o cooperativas de ahorro y crédito se aplican directamente a las personas naturales, pero muy poco a las personas jurídicas.

La función principal del puntaje o score crediticio es identificar el riesgo de impago de un cliente y "discriminar" nuevos clientes en función del historial de impago de los antiguos clientes. Para ello es imprescindible un análisis de las variables personales de cada uno de los clientes, estos procesos se denominan análisis discriminante. El desarrollo de estos análisis varía en función de las necesidades de quienes los soliciten, del tipo y forma de los datos obtenidos, y de la exactitud y eficiencia de los modelos. Con los años, los modelos de calificación crediticia han adoptado aspectos de aplicación de modelos lineales, no lineales, paramétricos, estadísticos y econométricos, para encontrar una medición adecuada del riesgo. (Rodríguez-Guevara et al., 2017).

Finalmente, Rodríguez-Guevara et al., (2017) indican que, en el proceso de evaluación de la solvencia de los clientes, las instituciones financieras consideran factores cualitativos y cuantitativos. Sin embargo, la relevancia y aplicación de estas variables a menudo omite la especificidad del ámbito de negocio bajo análisis. En el caso de las evaluaciones individuales, se priorizan variables que reflejan las características personales y circunstancias socioeconómicas, resultando en un amplio espectrº como género, clase social, condiciones de vida, comunidad y etnia) y cuantitativos (incluyendo ingresos, egresos, patrimonio neto, salario, número de dependientes, y cantidad de vehículos). Por otro lado, la evaluación de la solvencia de entidades corporativas se centra en indicadores financieros, según Gonçalves & Braga (2008), abarcando aspectos como

el ratio de liquidez, activos totales, reservas bancarias, cobertura de deudas, provisiones para eventualidades, rentabilidad de los activos y patrimonio, entre otros.

### 2.2.3. Segmentación de mercado

Para Fernández & Aqueveque (2001), desarrollar estrategias comerciales centradas en los mercados objetivos de productos de consumo es una tarea desafiante, especialmente para las pequeñas y medianas empresas, un desafío que se manifiesta en lo que se conoce como *segmentación del mercado;* dicho concepto consiste en dividir al mercado objetivo en grupos definidos y relevantes para permitir a los especialistas en marketing adaptar su mezcla de marketing a necesidades de segmentos específicos. Este proceso se basa en identificar variables de consumo directamente observables y no directamente observables que permitan organizar a los consumidores en segmentos consistentes.

La implementación de una estrategia de segmentación de mercado ofrece a las empresas proveedoras de bienes o servicios la ventaja de evitar la competencia directa al diferenciar sus ofertas en términos de precio y en las características del producto, estrategias publicitarias y canales de distribución. A pesar de los costos asociados con la segmentación, que incluyen investigación de mercado y campañas publicitarias específicas; los beneficios de mayores ventas y márgenes a menudo compensan estos costos porque los productos finales se adaptan con mayor precisión a las necesidades detalladas de los consumidores. (Fernández & Aqueveque, 2001)

La figura 2 muestra las posibles combinaciones de estas bases de segmentación y ejemplos de variables que permiten establecer un esquema de segmentación de clientes.

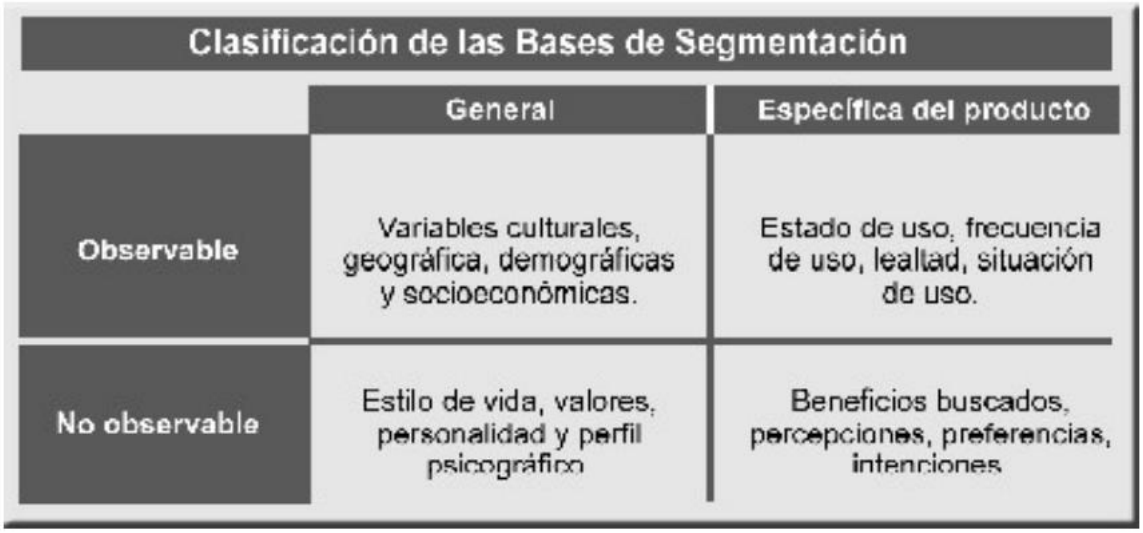



*Figura 2: Clasificación de criterios de segmentación. Fuente: (Fernández Robín, C., & Aqueveque Torres, C. (2001). Segmentación de mercados: buscando la correlación entre variables sicológicas y demográficas. P.3.)*

Para Fernández & Aqueveque (2001), la clave para una segmentación de mercado efectiva radica en la selección adecuada de variables que permitan dividir el mercado en grupos distintos. La premisa básica sostiene que existe una correlación directa entre las variables de segmentación elegidas y el comportamiento de los consumidores. Esta correlación se manifiesta de modo que, individuos con diferencias en las variables seleccionadas para la segmentación, tienden a mostrar diferencias en su interacción con los elementos de la mezcla de marketing.

Por el contrario, aquellos individuos con características similares en cuanto a variables de segmentación tienden a comportarse de manera similar en relación con la mezcla de marketing. Este enfoque establece la importancia de identificar con precisión aquellas variables que más influyen en las decisiones y preferencias de los consumidores, permitiendo así una estrategia de marketing más dirigida y efectiva (Fernández & Aqueveque, 2001).

### 2.3. Metodología de crédito de la Cooperativa de Ahorro y Crédito Jardín Azuayo

Para organizar el proceso de crédito que lleva a cabo la Cooperativa de Ahorro y Crédito Jardín Azuayo, se trabajó en cinco momentos que ayudan a definir como hacer operativo el proceso respectivo de la metodología de crédito existente. Esta perspectiva reconoce que el proceso de operaciones crediticias adecuadas depende no solo del momento de

aplicación y de la toma de decisiones, sino que también existen otras circunstancias directas e indirectas que influyen en el desarrollo de una cartera saludable y su gestión. Desde esta perspectiva, la metodología propone momentos o espacios de control que se resumen a continuación (Jardín Azuayo, 2022):

1. Análisis de las condiciones internas de la institución: entre los aspectos internos de la institución para llevar a cabo una buena gestión crediticia están:
    a. La gobernabilidad, reflejada en la dinámica entre personal y directivos, estabilidad y capacitación, afectando el riesgo de cartera.
    b. La capacidad administrativa, esta depende de la habilidad y conocimiento del equipo, crucial para la gestión y decisiones.
    c. El tiempo de operación y tamaño del mercado de una oficina influyen en su conocimiento y posicionamiento regional.
2. Análisis del contexto, área geográfica o sector económico al que atiende: Este control tiene como objetivo evaluar el riesgo de crédito enfocándose en el mercado objetivo y las condiciones específicas del entorno geográfico y socioeconómico de cada oficina. Involucra un análisis cuantitativo estadístico donde se levantan observaciones cuantitativas y se realizan reuniones para analizar estas, determinando condiciones de ponderación para la calificación crediticia de los socios y las características de riesgo, teniendo en cuenta factores como la fortaleza o debilidad, la concentración de la red de relaciones, y el impacto de la urbanización en el nivel de riesgo.

    Para conocer la situación financiera y social de los solicitantes de crédito, la metodología establece el uso de herramientas investigativas que dependen del conocimiento previo del socio. Entre las herramientas de investigación se incluyen la inspección del lugar de residencia o actividad económica, análisis del endeudamiento y historial crediticio, referencias e información sobre vivienda e ingresos, informes de comportamiento comunitario y situación patrimonial, y datos de otras instituciones.

3. Análisis del solicitante: Para un análisis crediticio efectivo, es esencial considerar aspectos clave: la red de relaciones del solicitante, su capacidad de pago, su situación económica actual y el objetivo específico del crédito solicitado. Estos elementos son cruciales para evaluar tanto el nivel de riesgo involucrado como la fiabilidad del solicitante de crédito. La gama de herramientas analíticas

disponibles es amplia, incluyendo desde el análisis del flujo de caja y la estabilidad de los ingresos, hasta la revisión del historial crediticio y las referencias, así como información detallada sobre ingresos y vivienda. En casos donde el riesgo percibido es elevado, resulta imprescindible llevar a cabo inspecciones detalladas y una revisión minuciosa de los registros públicos. Este enfoque holístico permite una comprensión profunda de la situación socioeconómica de los solicitantes, basándose en un conocimiento previo de sus circunstancias y abarcando desde visitas domiciliarias hasta el análisis financiero.

Para incrementar la precisión en la evaluación crediticia y asegurar un proceso de análisis adecuadamente ajustado al perfil de riesgo, se sugiere la creación de nueve categorías de clientes basadas en su nivel de riesgo crediticio y en la información previamente conocida sobre ellos. Dicha metodología posibilita la implementación de criterios analíticos específicos, adaptados al tipo de crédito solicitado, ya sea para consumo o fines productivos. Este enfoque se centra en un conocimiento directo de los solicitantes, evaluando su capacidad de pago y la situación de sus deudas actuales. Para una inclusión financiera efectiva y una valoración precisa de los riesgos, resultan fundamentales estrategias que fomenten la integración comunitaria, el fortalecimiento de capacidades locales y la formación en desarrollo cooperativo. La figura 3 muestra la estructuración de estas tipologías de socios que permiten articular los procesos de evaluación y análisis de los perfiles en los que se puede ubicar cada persona según las características evaluadas.

| NIVEL DE RIESGO<br><br>MONTO | Bajo Riesgo<br>Socios conocidos | Riesgo Medio<br>Socios medianamente conocidos | Alto Riesgo<br>Socios no conocidos |
|---|---|---|---|
| **Más de 40.000 dólares** | **B3**<br>1. Aprueba el Responsable de oficina o Gestor de Oficina.<br>2. Se requiere documentos de identificación del deudor principal y cónyuge.<br>3. Se requiere verificación telefónica de domicilio, trabajo o datos ciertos del grupo al que pertenece.<br>4. La garantía es hipotecaria, requiere inspección, el bien avaluado debe cubrir el 140% del monto solicitado.<br>5. Se podrá aceptar hipotecas que cubran el 120% del monto solicitado y adicionar otras garantías de las establecidas en el Reglamento de Crédito, hasta completar el valor de cobertura requerido.<br>6. El socio es conocido una vez que superó su tercer crédito y se conoce su red de relaciones, es posible identificar con claridad su calidad moral, su origen y fuentes de pago. | **M3**<br>1.Aprueba el Responsable de oficina o Gestor de Oficina.<br>2. Se requiere documentos de identificación del deudor principal y cónyuge.<br>3. Se requiere verificación telefónica de domicilio, trabajo o datos ciertos del grupo al que pertenece.<br>4. Se requiere un análisis del flujo de caja y diversificación de sus fuentes de ingresos.<br>5. La garantía es hipotecaria, requiere inspección, el bien avaluado debe cubrir el 140% del monto solicitado.<br>6. El socio es conocido una vez que superó su segundo crédito. | **A3**<br>1.Aprueba el Responsable de oficina o Gestor de Oficina.<br>2.Se requiere documentos de identificación del deudor principal y cónyuge.<br>3.Se requiere verificación telefónica del domicilio, trabajo o datos ciertos del grupo al que pertenece.<br>4.Se requiere un análisis del flujo de caja y diversificación de sus fuentes de ingreso.<br>5.La garantía es hipotecaria, requiere inspección, el bien avaluado debe cubrir el 140% del monto solicitado.<br>6.Los socios pueden o no ser conocidos. |
| **De 20.001– 40.000 dólares** | **B2**<br>1. Aprueba el Responsable de oficina o Gestor de Oficina.<br><br>2. Se requiere documentos de identificación del deudor principal y cónyuge, así como del garante.<br><br>3. Se requiere verificación telefónica del domicilio, trabajo o datos ciertos del grupo al que pertenecen tanto deudor principal como garante.<br><br>4. Evaluación de los garantes mediante la verificación de la fuente de ingresos, patrimonio que disponen y analizar la fortaleza de la red de relaciones con el deudor.<br><br>5. El socio es conocido una vez que superó su tercer crédito. | **M2**<br>1. Aprueba el Responsable de oficina o Gestor de Oficina.<br><br>2. Se requiere documentos de identificación del deudor principal y cónyuge, así como del garante.<br><br>3. Se requiere verificación telefónica del domicilio, trabajo o datos ciertos del grupo al que pertenecen tanto deudor principal como garante.<br><br>4. Evaluación de los garantes mediante la verificación de la fuente de ingresos, patrimonio que disponen y analizar la fortaleza de la red de relaciones con el deudor.<br><br>5. El socio es conocido una vez que superó su segundo crédito. | **A2**<br>1.Aprueba el Responsable de oficina o Gestor de Oficina.<br><br>2.Se requiere documentos de identificación del deudor y cónyuge, así como del garante.<br><br>3.Se requiere verificación telefónica del domicilio, trabajo o datos ciertos del grupo al que pertenecen tanto deudor principal como garante.<br><br>4.Se requiere un análisis del flujo de caja del deudor principal.<br><br>5. Evaluación de los garantes mediante la verificación de la fuente de ingresos, patrimonio que disponen y analizar la fortaleza de la red de relaciones con el deudor principal.<br><br>6.Los socios pueden o no ser conocidos. |
| **Hasta 20.000 dólares** | **B1**<br>1. Aprueba el Responsable de oficina o Gestor de Oficina.<br><br>2. Se requiere documentos de identificación del deudor y cónyuge,<br><br>3. El análisis fundamental tendrá que ver con la fortaleza de la red de relaciones del deudor,<br><br>4. Verificación de la fuente de ingresos, patrimonio que dispone.<br><br>5. El socio es conocido una vez que superó su tercer crédito. | **M1**<br>1. Aprueba el Responsable de oficina o Gestor de Oficina.<br>2. Se requiere documentos de identificación del deudor y cónyuge.<br>3. Se requiere verificación telefónica del domicilio, trabajo o datos ciertos del grupo al que pertenecen el deudor principal.<br>4. Análisis y verificación de la fuente de ingresos, patrimonio que disponen y analizar la fortaleza de la red de relaciones.<br>5. El socio es conocido una vez que superó su segundo crédito. | **A1**<br>1.Aprueba el Responsable de oficina o Gestor de Oficina.<br>2.Se requiere documentos de identificación del deudor principal y cónyuge.<br>3.Se requiere verificación telefónica del domicilio, trabajo o datos ciertos del grupo al que pertenecen el deudor principal.<br>4. Análisis de la estabilidad de la fuente de ingresos y verificación.<br>5. Análisis del flujo de caja (nivel de endeudamiento y patrimonio que dispone)<br>6.Los socios pueden o no ser conocidos. |

*Figura 3: Tipologías de socios en función de su riesgo de crédito y el nivel de conocimiento del socio. Fuente: (Jardín Azuayo (2022). Manual Metodológico de Crédito. p. 17.)*

En el contexto de los microcréditos y créditos para proyectos productivos, la cooperativa Jardín Azuayo se destaca por su compromiso con el mejoramiento de las condiciones de vida y laborales de sus miembros, enfocándose especialmente en el trabajo con colectivos organizados y en el fomento de una economía nacional solidaria. Al evaluar estas solicitudes de crédito, se toman en consideración factores como la estabilidad de los ingresos y los niveles de endeudamiento, empleando herramientas que incluyen análisis estadísticos y documentación de soporte.

La adjudicación de créditos se guía por las "5 C" de la calificación crediticia (Carácter, Capacidad, Capital, Condiciones y Colateral), adaptándose estas al perfil y necesidades del solicitante, así como a las características del mercado en el que se inserta. El objetivo final es apoyar iniciativas económicas que beneficien a la comunidad y contribuyan al desarrollo local, evidenciando un compromiso firme con la promoción de la prosperidad compartida.

4. Proceso de seguimiento y recuperación de cartera: La gestión y recuperación del crédito en la Cooperativa Jardín Azuayo involucra una consideración meticulosa de factores internos, el ambiente operativo de cada sucursal y el proceso crediticio en sí. Esta tarea se aborda desde tres ángulos críticos: la perspectiva del prestamista, el compromiso con la disciplina financiera y el enfoque institucional centrado en el bienestar de los socios. Estas dimensiones subrayan la importancia de administrar los recursos con responsabilidad, asegurar una recuperación eficaz del crédito y adoptar un enfoque proactivo en el seguimiento, que incluye notificaciones y esfuerzos de recuperación extrajudiciales, así como acciones legales cuando sean necesarias.

   El monitoreo de la cartera de créditos se realiza tanto a nivel institucional como a nivel de cada sucursal, abarcando la evaluación de la calidad de la cartera, revisiones mensuales de la disciplina financiera y análisis anuales sobre la probabilidad de incumplimiento. A nivel de sucursal, el enfoque está en las transacciones crediticias de mayor riesgo, implementando notificaciones tempranas, controles semanales de morosidad y seguimiento diario de los pagos atrasados.

   La estrategia de recuperación de créditos enfatiza la adherencia a procedimientos de notificación y promueve una cultura de pago puntual, involucrando a la comunidad y a organizaciones locales para fomentar un entorno de control social y educativo en la recuperación de créditos. La reestructuración y refinanciamiento de créditos se concibe con respeto hacia los socios y se considera un componente esencial en la creación de una cultura de cumplimiento de pagos. Para ejercer una presión social efectiva, es crucial activar las redes de relaciones e informar a los comités de crédito locales sobre las situaciones de interés.

Finalmente, la rehabilitación de créditos se lleva a cabo conforme a las normativas y procesos crediticios de la cooperativa, basándose en la calificación de riesgo de cada operación para determinar las acciones de rehabilitación adecuadas. La responsabilidad de este proceso recae en el gerente de la sucursal o su adjunto, quienes emplean herramientas como llamadas telefónicas y comunicaciones escritas para establecer contacto con los deudores y garantes, asegurando así un enfoque personalizado y eficiente en la gestión de recuperaciones.

## 2.4. Enfoques y modelos de aprendizaje automático no supervisado

Alpaydin, (2014) define el aprendizaje automático o machine learning (ML), como una rama de la inteligencia artificial que permite a los equipos computacionales tener la facultad de adquirir y generar conocimiento a partir de conjuntos de datos, y definir capacidades predictivas o de toma de decisiones sin necesidad de ser programadas de manera explícita para cada escenario específico. Alpaydin, (2014) indica además que este proceso de aprendizaje se realiza mediante la implementación de algoritmos capaces de distinguir y catalogar patrones dentro de los datos, permitiendo a los sistemas informáticos mejorar sus parámetros de rendimiento basándose en experiencias previas o muestras representativas de los datos. El aprendizaje automático abarca aplicaciones que van desde la predicción de comportamientos de consumidores hasta el reconocimiento de imágenes y el análisis avanzado del lenguaje natural.

El concepto de aprendizaje automático se sustenta en la aplicación de técnicas estadísticas avanzadas para el desarrollo de modelos matemáticos que permiten inferir y proyectar resultados a partir de datos muestrales (Alpaydin, 2014). Estos modelos pueden clasificarse como predictivos, cuando se desea anticipar ocurrencias futuras de un evento o valor, o descriptivos, orientados a la extracción de conocimiento valioso de los datos analizados, por ejemplo, una de las principales características del aprendizaje automático radica en su capacidad para procesar y asimilar volúmenes significativos de información, un atributo crucial en la denominada era del "big data". (Alpaydin, 2014).

Alpaydin (2014), señala que la eficacia de los algoritmos empleados para el entrenamiento y aplicación de modelos de aprendizaje automático es un aspecto crítico, tanto en términos de recursos temporales como de capacidad de almacenamiento; este aspecto subraya la relevancia del aprendizaje automático como un instrumento vital para la generación de conocimiento y para facilitar los procesos de toma de decisiones,

impulsando avances significativos en sectores tan variados como el retail, la banca y el ámbito sanitario entre muchos otros más. Mediante la transformación de los datos en acciones inteligentes y soluciones pioneras, el aprendizaje automático se erige como un pilar fundamental en la conformación de la próxima generación de innovaciones tecnológicas (Alpaydin, 2014).

### 2.4.1. Modelos de Aprendizaje No Supervisado

El aprendizaje automático no supervisado se diferencia del aprendizaje supervisado principalmente en que éste último funciona sin la guía de respuestas correctas (etiquetas) específicas que guían al modelo a predecir los resultados deseados; mientras en el aprendizaje supervisado se requiere aprender a asociar entradas con resultados correctos basándose en ejemplos, el aprendizaje no supervisado consiste en encontrar patrones y regularidades en datos que no están etiquetados, esto con el propósito de revelar la estructura oculta de los datos e identificar agrupaciones o patrones que surgen naturalmente (Alpaydin, 2014).

Alpaydin (2014), cita como un ejemplo de aprendizaje no supervisado la denominada "estimación de densidad", la cual nos permite comprender cómo se distribuyen los datos de los clientes en el entorno empresarial sin una clasificación previa. En este contexto, Alpaydin (2014) menciona los posibles usos del aprendizaje no supervisado los cuales van desde la segmentación de clientes hasta la compresión de imágenes. En un contexto empresarial, este enfoque permite agrupar a los clientes según características demográficas y de comportamiento, lo que ayuda a personalizar las tácticas de marketing y mejorar la gestión de las relaciones con los clientes; este análisis también es útil para descubrir segmentos de clientes principales y nichos de mercado sin explotar (Alpaydin, 2014). En el caso de la compresión de imágenes, este tipo de modelos agrupa píxeles similares para reducir la cantidad de datos necesarios para representar una imagen, preservando la esencia visual, pero con menos detalles.

El aprendizaje automático no supervisado, por lo tanto, se especializa en identificar estructuras y patrones ocultos en datos no etiquetados, lo que abre una gama de aplicaciones prácticas, desde su uso en estrategias de marketing, hasta bioinformática. Ya que no depende de datos de salida predefinidos, se convierte en una herramienta poderosa para analizar datos y descubrir agrupaciones, tendencias y anomalías naturales que pueden no ser evidentes de inmediato.

### 2.4.2. Algoritmos de Agrupamiento o Clustering

Los algoritmos de agrupamiento, también conocidos como algoritmos de "clustering", desempeñan un papel crucial en la exploración y análisis de grandes volúmenes de datos sin etiquetado previo. Su principal objetivo es identificar grupos o clústeres dentro de un conjunto de datos y agrupar los elementos según su similitud. A diferencia de otros métodos que asumen una distribución específica de los datos, como una distribución normal o gaussiana, el clustering se caracteriza por su capacidad de adaptarse a datos que tienen mayor diversidad y no se ajustan a patrones preestablecidos (Alpaydin, 2014).

Como indica Alpaydin (2014), esta adaptabilidad es invaluable en escenarios donde los datos se segmentan en diferentes subgrupos, como aplicaciones como el reconocimiento de caracteres o de voz; en estos casos, las variaciones en la escritura o la articulación de palabras dan como resultado diferentes categorías dentro del mismo grupo. Este método permite visualizar las clases como amalgamas de múltiples distribuciones, cada una de las cuales representa un subgrupo específico dentro del conjunto de datos. Este enfoque amplía la capacidad del modelo para capturar la riqueza y complejidad de los datos, lo que facilita la detección de patrones y estructuras complejas que de otro modo pasarían desapercibidas (Alpaydin, 2014).

Los algoritmos de agrupación son esenciales, no sólo para analizar y comprender grandes cantidades de información no clasificada, sino también para impulsar avances significativos en campos tan diversos como la bioinformática, el marketing y las ciencias sociales (Alpaydin, 2014). Al agrupar elementos con propiedades similares, estos algoritmos descubren estructuras subyacentes en los datos, abriendo puertas para identificar relaciones, tendencias y agrupaciones que no son visibles a simple vista (Alpaydin, 2014).

#### 2.4.2.1. Algoritmo K-Means

El algoritmo k-means es uno de los algoritmos de clustering más utilizados en el campo del aprendizaje automático por su simplicidad y notable eficiencia. Este algoritmo busca dividir un conjunto de datos en 'k' grupos, basándose en la similitud que pueda existir entre los datos. Esta organización se fundamenta en la cercanía de los datos a un conjunto

de centroides iniciales, que se ajustan de manera iterativa para capturar el promedio de los puntos dentro de cada grupo, como destaca Foster y Fawcett (2013).

El proceso de agrupamiento del algoritmo k-means inicia seleccionando de forma aleatoria, o mediante una selección previa, k puntos que servirán como centros (centroides) iniciales de los clústeres (Figura 4.). Posteriormente, los datos se asignan al centro más próximo, recalculando los centroides en cada iteración (Figura 5.), hasta que la asignación de puntos se estabiliza, señalando una optimización en la formación de los grupos (Foster & Fawcett, 2013).

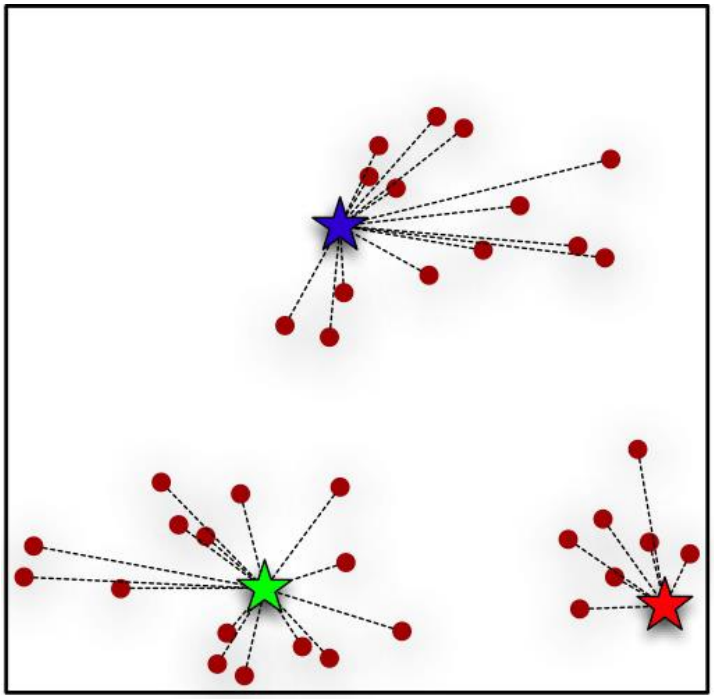


*Figura 4:Primer paso del algoritmo k-means, ubicar los puntos (datos) más cercanos a los centroides iniciales. Fuente: (Foster, & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking.)*

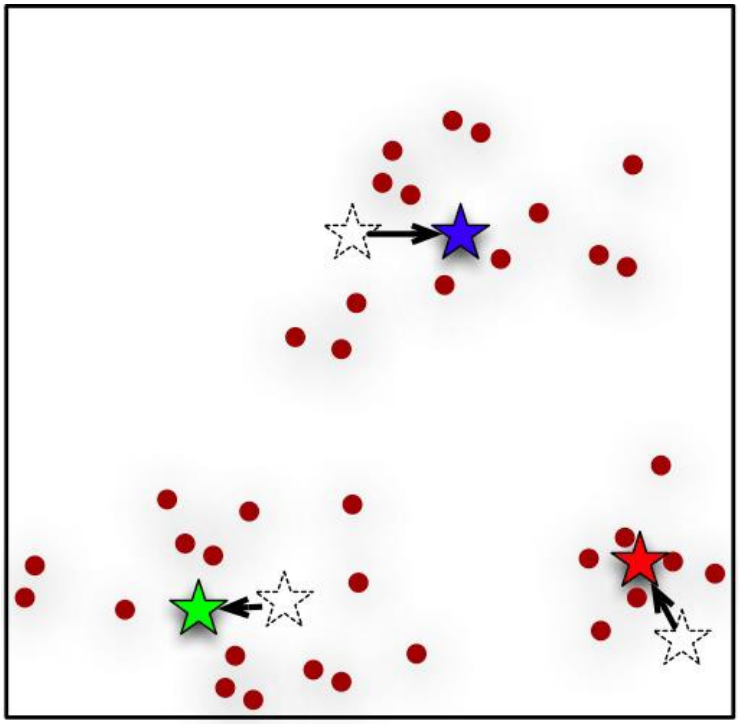


*Figura 5: Segundo paso del algoritmo k-means, encontrar el centro de los clusters hallados en el primer paso. Fuente: (Foster, & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media)*

Foster y Fawcett (2013), señalan que la elección del número de clústeres, k, es crítica para la eficacia de la agrupación. A menudo, se experimenta con diferentes valores de k para hallar aquel que mejor capta la estructura subyacente de los datos. Esta exploración permite identificar la configuración más adecuada que muestre con la mayor precisión las relaciones y distribución natural de los datos.

Como un ejemplo práctico, k-means demuestra su valor en la cuantificación de colores en imágenes digitales. Alpaydin (2014) ilustra cómo este algoritmo permite representar imágenes de alta resolución con una paleta reducida de colores, conservando la calidad visual de las imagenes originales. Esta capacidad de transformar espacios de datos continuos en representaciones discretas y manejables subraya la utilidad del k-means, encontrando aplicación en áreas como la segmentación de mercado (perfilamiento de clientes), la compresión de imágenes, y la detección de anomalías en conjuntos de datos. La versatilidad de k-means para identificar patrones y agrupaciones en diversos contextos, desde el marketing hasta la bioinformática, lo posiciona como una herramienta clave en la obtención de insights de conjuntos de datos extensos.

#### 2.4.2.2. Algoritmo DBSCAN

Los algoritmos de agrupación en clústeres basados en densidad, en particular DBSCAN (agrupación espacial de aplicaciones con ruido basada en densidad), ofrecen una alternativa importante a los enfoques y métodos de agrupación en clústeres tradicionales, como k-means.

Hahsler y cols. (2019) señalan que, a diferencia de los métodos que asumen una geometría de grupo determinada y se basan en la varianza para definir grupos, DBSCAN funciona sin asumir formas de grupo predeterminadas. Este algoritmo se centra en la densidad de datos alrededor de cada punto para identificar grupos. Esto permite detectar grupos de formas irregulares y procesar cambios en la densidad de datos. El algoritmo DBSCAN, según indica Hahsler et al. (2019), comienza seleccionando un punto de datos y explorando sus alrededores dentro de un radio definido por el parámetro ε (épsilon). Un punto se considera central si hay un número mínimo de otros puntos (minPts) cercanos, que marcan el inicio de un nuevo grupo; este grupo se expande para incluir todos los puntos conectados a cualquier punto central. Este proceso se repite hasta que no se identifican nuevos centros y los puntos que no pertenecen a ningún grupo se marcan como ruido.

La efectividad de DBSCAN depende en gran medida de los valores asignados a 'ε', 'minPts', y la selección de estos parámetros es crucial para el resultado final de la agrupación. Campello et al. (2015) proponen una versión mejorada de DBSCAN llamada DBSCAN*, que trata los puntos límite como ruido, resolviendo así el problema de su asignación ambigua en la versión original del algoritmo. Esta mejora contribuye a la coherencia del agrupamiento para que el resultado dependa menos del orden en que se procesan los datos.

De acuerdo con Hahsler et al. (2019), DBSCAN ha sido utilizado eficazmente en contextos donde la configuración del clúster es impredecible o donde los datos tienen anomalías, por ejemplo, en la detección de concentraciones de alta densidad en datos espaciales, segmentación de imágenes y minería de modelos en bioinformática. Su capacidad para detectar conglomerados en función de la proximidad y densidad de puntos lo convierte en una potente herramienta de análisis y extracción de datos en situaciones donde no se puede asumir la uniformidad de los datos (Hahsler et al., 2019).

### 2.4.3. Métodos de evaluación de modelos de agrupamiento o clustering

Evaluar la calidad o eficiencia de un algoritmo de clustering es esencial para comprender la relevancia de los grupos formados. Shahapure y Nicholas (2020) destacan al score de silueta (silhouette score) como una herramienta efectiva para esta evaluación, proporcionando un criterio numérico que refleja la precisión con la que los puntos de datos han sido agrupados. Este indicador resulta particularmente valioso para establecer el número óptimo de grupos en técnicas como el k-means, donde el valor de k no se define de antemano.

El silhouette score, según Heras Calvo (2023), mide tanto la cohesión dentro de cada grupo como la separación entre grupos distintos. Un score próximo a +1 sugiere una asignación adecuada del punto de datos a su cluster, indicando un buen ajuste. Por contraste, un valor cercano a -1 implica que el punto estaría mejor clasificado en otro grupo, y un score alrededor de 0 revela una asignación ambigua, apuntando a una posible confusión en la definición de los clusters.

Para Yadav (2023), el coeficiente de silueta es una métrica diseñada para cuantificar la adecuación con la que cada punto de datos se integra en el clúster que le ha sido asignado. Esta métrica incorpora análisis sobre dos aspectos fundamentales: la cohesión, que evalúa la proximidad de un punto de datos respecto a los demás miembros de su propio clúster, y la separación, que mide la distancia que separa un punto de datos de los integrantes de otros clústeres. Este coeficiente, por tanto, ofrece una visión comprensiva sobre la pertinencia de la asignación de los puntos a los clústeres, basándose en la premisa de que un ajuste óptimo se refleja tanto en una alta cohesión interna del clúster como en una clara separación entre clústeres distintos (Yadav, 2023).

Para determinar el silhouette score de un punto específico, se analizan dos distancias: la distancia promedio dentro del mismo cluster (a) y la distancia promedio al cluster más próximo (b). La fórmula (b - a) / max(a, b) permite calcular este score, brindando una perspectiva clara sobre la efectividad del agrupamiento (Shahapure & Nicholas, 2020; Heras Calvo, 2023).

Silhouette score se aplica en diversos contextos, desde la segmentación de mercados hasta la organización de información genómica, siendo una herramienta crucial cuando es necesario agrupar datos sin categorías definidas previamente. Permite a los investigadores y profesionales de la ciencia de datos no solo comprobar la coherencia de los clústeres

identificados sino también ajustar los parámetros de los algoritmos de clustering para lograr un agrupamiento de datos más significativo y pertinente (Linkedin, 2023).

Además, la representación gráfica del silhouette score facilita enormemente la interpretación de los datos, permitiendo identificar anomalías o superposiciones entre los grupos y, por ende, mejorando la capacidad para tomar decisiones informadas basadas en el análisis de datos (Shahapure & Nicholas, 2020)

De acuerdo con Aman (2012), para un punto dado p, inicialmente se determina la distancia media entre p y todos los otros puntos dentro del mismo clúster (lo que constituye una medida de cohesión interna del grupo, señalada como A). Posteriormente, se mide la distancia media entre p y todos los puntos pertenecientes al clúster más próximo (esto representa la medida de separación respecto al grupo más cercano, indicada como B). El coeficiente de silueta para el punto p se calcula tomando la diferencia entre B y A, y dividiéndola por el mayor de ambos valores (max(A,B)). A través de este proceso, se obtiene el coeficiente de agrupamiento para cada punto, permitiendo calcular un coeficiente promedio "general" que refleja la calidad del agrupamiento en su conjunto. De forma intuitiva, este método busca cuantificar la separación entre clústeres. Un grupo bien cohesionado (donde A es bajo) y bien separado de los demás (donde B es alto) resultará en un score elevado, indicando una segmentación efectiva. Como ejemplo consideremos los siguientes ejemplos de clusters generados utilizando un modelo k-means:

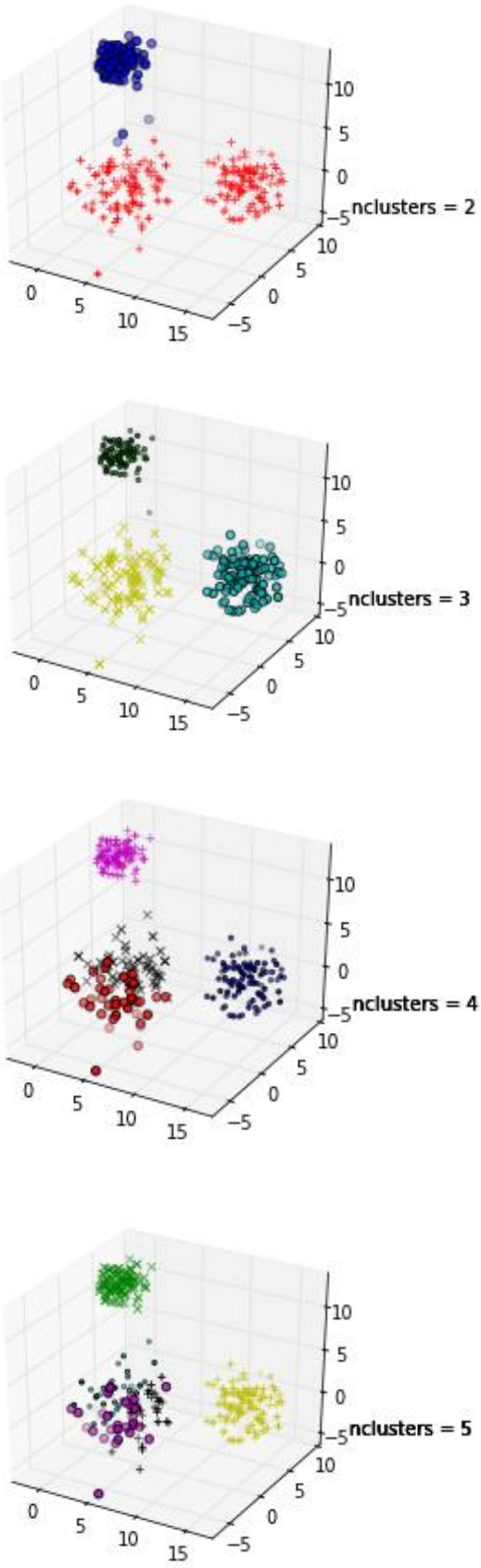



*Figura 6: Ejemplificación de agrupamiento de datos usando k-means. Fuente: Aman (2012). How to interpret mean of Silhouette plot? Cross Validated. https://stats.stackexchange.com/q/44653*

Los colores en el gráfico representan los grupos identificados mediante el método de agrupación de k-means, estableciendo el número de clústeres (k) en los valores de 1, 2, 3, 4 y 5 sucesivamente. Con esta indicación el algoritmo de agrupamiento empezó a segmentar los datos primero en dos grupos, y así sucesivamente, incrementando el número de grupos uno a uno (Aman, 2012). Si realizamos un análisis mediante el gráfico de silueta (Figura 7), este nos indica que el coeficiente de silueta alcanza su valor máximo cuando k es igual a 3, lo que implica que este es el número óptimo de clústeres para este conjunto de datos. En este caso particular, la visualización directa de los datos corrobora que tres grupos capturan de manera más precisa la segmentación del conjunto de datos analizado (Aman, 2012).

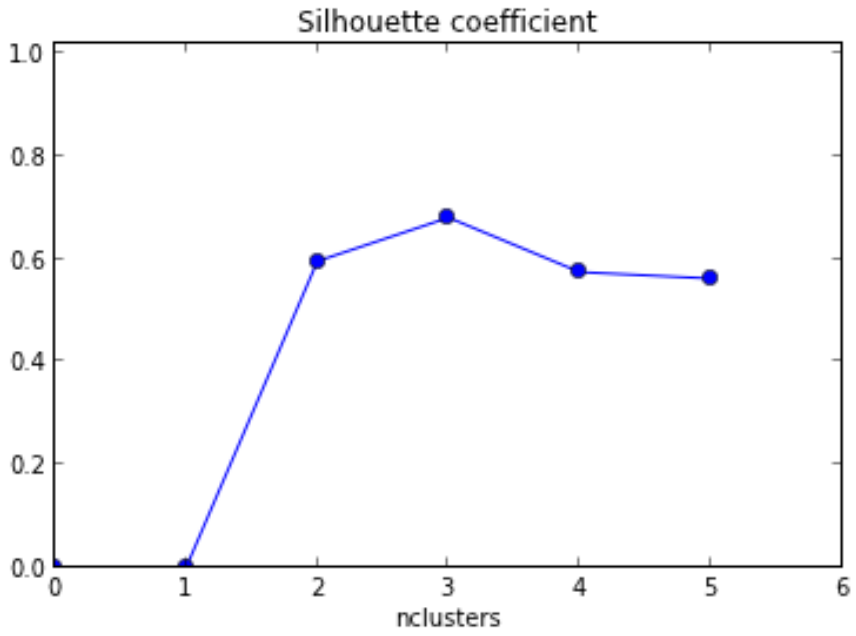



*Figura 7: Gráfico de coeficiente de silueta.. Fuente: Aman (2012). How to interpret mean of Silhouette plot? Cross Validated. https://stats.stackexchange.com/q/44653*

Como indica Aman (2012), en situaciones donde la visualización directa de los datos se ve obstaculizada por la alta dimensionalidad, un diagrama de silueta puede ofrecer indicaciones preliminares sobre el número adecuado de clústeres. Sin embargo, es importante recalcar que estas indicaciones pueden no ser siempre aplicables o precisas en todos los contextos. La "sugerencia" de un número óptimo de clústeres, aunque útil, podría resultar inadecuada o incorrecta en ciertos casos, subrayando la necesidad de una evaluación cuidadosa.

## 2.5. CRISP-DM como metodología para proyecto de minería de datos

CRISP-DM, acrónimo de CRoss-Industry Standard Process for Data Mining, es un modelo de proceso integral y estructurado que se utiliza ampliamente en la realización de proyectos de minería de datos. Este modelo trasciende los límites de la industria y la tecnología y está diseñado para su uso en una variedad de industrias y con diversas herramientas de minería de datos. Wirth y Hipp (2000) enfatizan que CRISP-DM tiene como objetivo hacer que los proyectos de minería de datos sean más eficientes en términos de costo, confiabilidad, reproducibilidad, gestión y velocidad. Además, este modelo se caracteriza por su adaptabilidad, permitiendo que personas con diferentes habilidades técnicas y limitaciones de tiempo exploren diferentes métodos.

La necesidad de adoptar una metodología como CRISP-DM surge de la complejidad y la naturaleza creativa de la minería de datos, ya que un proyecto de minería de datos requiere una amplia gama de habilidades y conocimientos (Wirth & Hipp, 2000). La falta de un marco estandarizado podría hacer que el éxito de los proyectos de minería de datos dependa demasiado de las habilidades individuales o del equipo responsable, lo que limitaría la capacidad de replicar prácticas exitosas en toda la organización (Wirth & Hipp, 2000).

Wirth & Hipp (2000) indican que CRISP-DM proporciona un enfoque estandarizado que facilita la transformación de problemas comerciales en tareas específicas de minería de datos, recomienda transformaciones y técnicas de análisis de datos apropiadas y proporciona estrategias para evaluar la efectividad de los resultados y documentar el proceso.

El modelo CRISP-DM se divide en seis fases principale, que no necesariamente se completan de forma secuencial en cada proyecto debido a posibles interdependencias entre ellas (Figura 6).

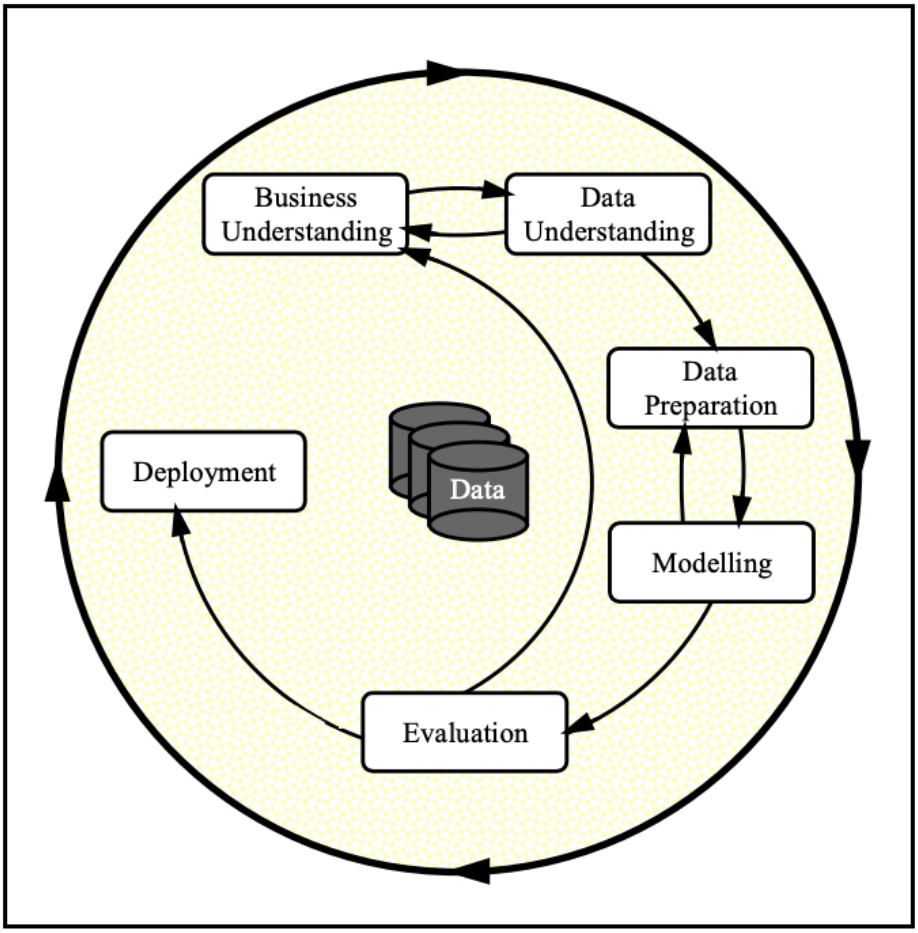



*Figura 8: Fases de la metodología CRISP-DM. Fuente: (Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf)*

El modelo CRISP-DM se compone de seis fases principales, cada una con actividades específicas (Figura 7), que contribuyen al desarrollo del proyecto (Wirth & Hipp, 2000):

- **Comprensión del negocio**: esta fase introductoria consiste en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial y traducir esta comprensión en una definición clara del problema de minería de datos y un plan preliminar para el proyecto.
- **Comprensión de los datos**: comienza con la recopilación inicial de datos y continúa con actividades que le permiten familiarizarse con los datos, identificar problemas de calidad, obtener conocimientos preliminares o descubrir conjuntos de datos interesantes.
- **Preparación de datos**: incluye todas las tareas necesarias para crear el conjunto de datos final utilizado en el modelado, incluida la selección del conjunto de datos, la limpieza de datos, la creación de nuevos atributos y la transformación de datos adecuada para las herramientas de modelado.

- **Modelado**: Aquí se seleccionan y aplican ciertas técnicas de modelado y sus parámetros se establecen en valores óptimos. Algunas técnicas pueden requerir formatos de datos específicos.
- **Evaluación**: esta fase evalúa el modelo desde una perspectiva de análisis de datos y revisa los pasos tomados para garantizar que se logren los objetivos comerciales.
- **Implementación**: La finalización del modelo generalmente no significa el final del proyecto. El conocimiento generado debe organizarse y presentarse de manera que sea útil para el cliente. Esto puede variar desde la creación de un informe hasta la implementación de un proceso de minería de datos operativo y repetible.

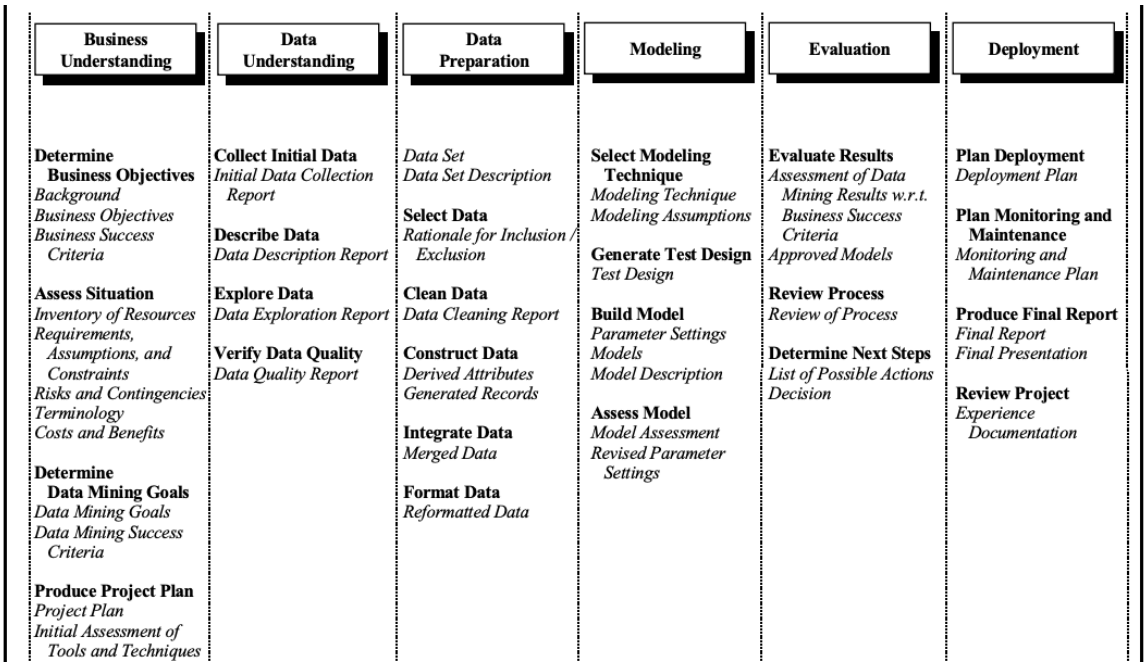



*Figura 9: Actividades y artefactos de cada fase del modelo CRISP-DM. Fuente: (Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf)*

Wirth & Hipp (2000) señalan que la metodología CRISP-DM resalta la imprtancia de una documentación y comunicación claras entre los miembros del equipo de proyecto y los stakeholders. La metodología ofrece recursos como listas de verificación, cuestionarios, técnicas y herramientas útiles incluso para analistas experimentados, y resulta particularmente beneficiosa para aquellos que se inician en el data mining. Además, destaca por su capacidad de adaptación a las necesidades de los usuarios finales y por su eficacia en abordar situaciones inesperadas y complejas (Wirth & Hipp, 2000).

## 2.6. Herramientas para la implementación de modelos de aprendizaje no supervisado

Rojas (2020), señala que en el ámbito de la implementación de proyectos de aprendizaje automático (ML), la selección del lenguaje de programación y herramientas adecuadas es una actividad crucial que ejerce una influencia determinante en el éxito de cualquier proyecto. Python, R y Matlab emergen como las opciones preeminentes, cada uno con sus particularidades y adaptabilidad a variadas situaciones en el ámbito del ML.

En ocasiones, el lenguaje elegido para el desarrollo se debe a la experiencia y comodidad del desarrollador; la ventaja de esta selección es que permite crear y desarrollar prototipos rápidamente, sin embargo, a largo plazo, esta puede no ser la mejor opción para desarrollar una aplicación de este tipo y una de las razones podría ser principalmente la cantidad de datos que se requieren procesar, lo que hace que las herramientas y lenguajes a utilizar sean especializados para este fin (Rojas, 2020).

En un resumen bastante sucinto por cada una de estas herramientas y lenguajes de programación, Python se distingue por su sintaxis intuitiva y un rico ecosistema de bibliotecas; R, en cambio, es apreciado por su orientación hacia la estadística y sus capacidades superiores en representación gráfica; finalmente, Matlab sobresale por su eficacia en el manejo de cálculos matemáticos y la optimización de algoritmos (Rojas, 2020)

Rojas (2020), enfatiza la relevancia de optar por la herramienta más adecuada con el fin de optimizar tanto la efectividad como la eficiencia en el desarrollo de soluciones innovadoras en el ámbito del ML. La decisión sobre qué lenguaje o herramienta adoptar debe estar alineada con las necesidades concretas del proyecto; la existencia de librerías soportadas y una comunidad activa son también factores importantes para considerar en esta elección Rojas (2020).

### 2.6.1. Python como lenguaje de programación para analítica de datos

Considerando a Python como una alternativa relevante, Rojas (2020) señala que este lenguaje de programación se distingue como un lenguaje sumamente flexible y orientado a objetos, con capacidad de operar en plataformas. Esta versatilidad lo hace apto para un amplio espectro de usos, que abarcan desde la ingeniería de software hasta el diseño de

páginas web. La creciente popularidad de Python se fundamenta en su diseño interpretado, lo que favorece un ciclo de desarrollo ágil; también su gran repertorio de bibliotecas, licencia gratuita para usarlo en el ámbito comercial, y su productividad, que requiere escribir menos cantidad de código en comparación con otros lenguajes (Rojas, 2020).

Dentro del campo del aprendizaje automático, Python se destaca por numerosas razones (Rojas, 2020):

- Su adopción generalizada en un rango variado de aplicaciones, desde desarrollos web hasta soluciones de automatización, establece una base sólida para su uso.
- Cuenta con una rica biblioteca de herramientas y marcos de trabajo que facilitan tanto la escritura del código como la optimización de los tiempos de desarrollo.
- Adicionalmente, su sintaxis, caracterizada por ser directa y sencilla, promueve un aprendizaje eficiente y la implementación efectiva de algoritmos de complejidad variable, elemento clave en el ámbito del aprendizaje profundo.
- La eficacia de Python para agilizar tanto el desarrollo como la evaluación de algoritmos, aunado a su carácter de código abierto y su propensión a favorecer el trabajo colaborativo, lo posiciona como la opción predilecta para proyectos dentro del aprendizaje automático.

### 2.6.2. Librerías especializadas para análisis y ciencia de datos en Python

Rojas (2020), enfatiza que Python se diferencia de otros lenguajes de programación por tener una comunidad de desarrolladores muy amplia y activa. Esta comunidad ha aportado al lenguaje una gran variedad de bibliotecas avanzadas, lo que ha mejorado notablemente sus funcionalidades. En particular, en el ámbito del aprendizaje automático, Python aprovecha la gran capacidad y complejidad de estas bibliotecas especializadas, listadas en la Tabla 1, que son esenciales para proyectos de analítica de datos o aprendizaje automático (Equipo Verne Academy, 2022).

| Librería | Descripción |
|---|---|
| **Pandas** | Utilizada para el manejo de datos, permite leer archivos de múltiples fuentes y realizar operaciones de datos complejas. |
| **Numpy** | Esencial para operaciones con matrices, permite crear y transformar estructuras numéricas. |
| **Plotly** | Facilita la creación de gráficas interactivas y visualizaciones de datos. |
| **Scikit-learn** | Proporciona herramientas para preprocesamiento de datos, creación de modelos de ML y construcción de Pipelines. |
| **Category Encoders** | Ofrece transformaciones para convertir variables categóricas en numéricas, mejorando el valor de los datos para ML. |
| **Imbalance Learning** | Ayuda a manejar conjuntos de datos desbalanceados mediante re-muestreo para clasificación. |
| **LightGBM / XGBoost** | Implementan técnicas de Gradient Boosting para modelos de árboles más potentes. |
| **Keras / Tensorflow** | Librerías de Deep Learning que simplifican la creación de redes neuronales con pocas líneas de código. |
| **Shap** | Permite interpretar los modelos de ML basándose en la Teoría de Juegos para entender el impacto de las variables. |
| **AzureML-sdk** | Proporciona herramientas para trabajar con Data Science y ML en la nube, permitiendo implementar servicios y automatizar tareas. |

*Tabla 1: Librerías Python para ciencia de datos y aprendizaje automático. Fuente: (Elaboración propia)*

# CAPITULO III: MARCO METODOLÓGICO

## 3. Marco metodológico

### 3.1. Modalidad de la investigación

Este trabajo utilizará un enfoque cuantitativo ya que, en este caso, se determina que es pertinente al tratar con soluciones de aprendizaje no supervisado, ya que se analizará grandes un gran conjunto de datos sin etiquetar para identificar patrones y agrupaciones naturales entre las variables. La metodología cuantitativa facilita la aplicación de algoritmos de agrupamiento que organizarán los datos de los socios en segmentos basados en similitudes, sin necesidad de intervención o interpretación subjetiva. Al emplear técnicas estadísticas y de modelado matemático, se puede cuantificar la relación entre diferentes características de los datos, permitiendo una segmentación precisa y objetiva. Este método es esencial para descubrir estructuras ocultas en los datos, ofreciendo una base sólida para tomar decisiones informadas y mejorar las estrategias de riesgo y crédito basadas en el análisis objetivo de la información.

### 3.2. Diseño de la investigación

El diseño de investigación que se plantea se enfoca en el desarrollo de un modelo de segmentación de socios (clientes), utilizando técnicas de aprendizaje automático no supervisado y clustering, enmarcado dentro de una metodología cuantitativa y no experimental. Este enfoque es el más adecuado para examinar el volumen significativo de datos recogidos en el almacén de datos de la cooperativa, con lo cual podremos automatizar la evaluación de la capacidad de pago y el riesgo de incumplimiento financiero de los socios a través del modelo de segmentación propuesto. La metodología cuantitativa juega un papel crucial en el procesamiento y análisis objetivo de los datos, facilitando la identificación de patrones y agrupaciones espontáneas entre los socios, sin necesidad de manipulación experimental de las variables.

Para validar el modelo, se realizará una comparación entre los resultados obtenidos en esta clasificación y aquellos de un estudio anterior que proponía un producto de crédito preaprobado a los socios. Este paso es fundamental para comprobar la eficacia del modelo en reconocer perfiles similares y verificar si la segmentación efectuada con el modelo propuesto se alinea con las decisiones anteriores sobre créditos preaprobados, garantizando así la relevancia y exactitud del modelo sugerido.

Este diseño de investigación no solo busca elevar la eficiencia en la promoción y provisión de servicios y en la gestión de riesgos por parte de la cooperativa, sino que también aspira a contribuir a la sostenibilidad financiera al minimizar el riesgo de incumplimiento y al identificar las maneras más adecuadas de ofrecer productos que se ajusten al perfil de los socios. Al perfeccionar los procesos internos y promover una toma de decisiones más ágil y fundamentada, el modelo propuesto emerge como una estrategia innovadora para afrontar retos financieros y potenciar los beneficios e ingresos de la cooperativa.

### 3.3. Población y muestra

La población objetivo serían los socios de la Cooperativa que hayan tenido menos una operación de crédito (activa o cancelada) desde 2021 hasta la fecha (ultimo corte realizado al 30 de noviembre de 2023, a los cuales se aplicará el modelo de aprendizaje no supervisado. No se realizará un muestro de los datos para ejecutar el análisis del modelo de clustering ya que este tipo de algoritmos requieren trabajar con la mayor cantidad de datos, por lo tanto, el modelo realizará un análisis exhaustivo de esta información para realizar la segmentación.

### 3.4. Métodos, técnicas y herramientas

En primera instancia, para este proyecto de titulación se usarán técnicas de análisis estadístico para analizar los datos recopilados y obtener información relevante sobre las características de los socios y ver la distribución de los datos. La información se extraerá desde el almacén de datos de la Cooperativa dando como resultado un archivo Excel que contiene la información a ser analizada. El almacén de datos tiene implementado de manera previa cubos de información y procesos ETL (Extract-Transform-Load) para obtener la información desde los diferentes orígenes de datos que forman parte del core informático de la Cooperativa.

### 3.5. Metodología de desarrollo del proyecto

El desarrollo del proyecto como tal, estará basado en la metodología CRISP-DM. Se ha seleccionado esta metodología para el desarrollo del proyecto por su nivel de detalle ya que profundiza en las actividades y artefactos que debe generar cada fase de la metodología y establece también una guía práctica para desarrollar y obtener cada artefacto. Además, actualmente hay mucha información y casos de uso relacionados con

el uso de CRISP-DM como metodología de proyectos de minería de datos, permitiendo tener múltiples referencias de su uso que sirven como guía para este proyecto de titulación.

Para este proyecto usaremos la metodología para refinar las necesidades y aplicación del modelo, para que se pueda aplicar de forma que responda a la necesidad planteada por la institución.

# CAPITULO IV: EJECUCIÓN E IMPLEMENTACION DEL MODELO

## 4. Ejecución e implementación del modelo

Usando el modelo CRISP-DM definiremos a continuación el marco de desarrollo del proyecto, de acuerdo con la guía propuesta por Chapman et al. (2000)

### 4.1. Comprensión del negocio

#### 4.1.1. Contexto

La Cooperativa de Ahorro y Crédito Jardín Azuayo ha identificado una oportunidad para mejorar la gestión de riesgo y la personalización de sus servicios a través de la segmentación de sus socios según sus capacidades de pago y riesgo de morosidad. La necesidad de esta iniciativa surge de la observación de patrones de comportamiento variados entre sus socios, lo que sugiere que una estrategia de gestión diferenciada podría mejorar la eficiencia operativa y la satisfacción del cliente.

#### 4.1.2. Objetivos de negocio y criterios de éxito

Como parte de los objetivos estratégicos de la Cooperativa, se busca mejorar la eficiencia en la promoción y entrega de servicios financieros, personalizando la oferta según el perfil de cada socio. El éxito se definirá por la capacidad del modelo para identificar segmentos de socios con precisión, permitiendo ofrecer productos y servicios adecuados que mejoren la satisfacción del cliente y reduzcan el riesgo de morosidad.

#### 4.1.3. Inventario de recursos

- **Personal**: Equipo de análisis de datos, equipo de IT para soporte técnico, equipo de gestión de riesgos, y personal de atención al cliente.
- **Fuentes de Datos**: Data warehouse de la cooperativa, con información sobre ingresos, egresos, flujo de caja, créditos, estado civil, edad, nivel educativo, y pre-calificación de riesgo de los socios.
- **Instalaciones Técnicas**: Servidores para el procesamiento de datos, software de análisis de datos (e.g., Python, R), y herramientas de ETL para la extracción de datos.

### 4.1.4. Requerimientos, supuestos y restricciones

Los requerimientos principales para el desarrollo de este proyecto son los siguientes:

- El modelo debe arrojar una precisión aceptable en la segmentación para implementar estrategias diferenciadas eficazmente.
- El proyecto debe mantener la confidencialidad de los datos de los socios.

Como supuestos del proyecto, consideramos lo siguiente:

- Los datos disponibles son representativos de la población de socios.
- Suponemos que los datos en el datawarehouse son completos y precisos

Finalmente enumeramos las restricciones a saber:

- Consideraciones relacionadas con la privacidad y la Ley de Protección de Datos en el tratamiento de la información de los socios.

### 4.1.5. Riesgos y contingencias

El principal riesgo dentro del proyecto es la posibilidad de sesgos en los datos que podrían afectar la precisión del modelo de segmentación. Como contingencia, este riesgo de calidad de datos podría mitigarse mediante una limpieza y validación exhaustiva de los datos.

Otro riesgo es la resistencia al cambio y la posible desconfianza de los resultados del modelo automatizado, versus el trabajo manual que se ha venido haciendo actualmente en el área de Análisis de datos. Este riesgo se abordará con capacitaciones y demostraciones de los beneficios que se pueden obtener del modelo.

### 4.1.6. Terminología

Se definen los siguientes términos claves

- **Socio**: en el ámbito de las instituciones de la Economía Popular y Solidaria, el socio, a diferencia de un cliente, es un miembro activo que no solo consume bienes o servicios, sino que también participa en la gestión, toma de decisiones y dirección de una Cooperativa.
- **Segmentación**: es el procedimiento mediante el cual se fracciona una base de clientes o un mercado en conglomerados más reducidos, los cuales comparten características, necesidades, patrones de compra o datos demográficos similares.

- **Capacidad de pago**: alude a la evaluación de la aptitud que tiene una persona para hacer frente a sus compromisos financieros. Dicha evaluación contempla una serie de factores clave con el fin de determinar si el cliente posee la capacidad de cumplir de manera realista y sostenible con sus obligaciones de pago.
- **Morosidad**: describe la circunstancia en la que un cliente no logra respetar los tiempos estipulados para el pago de obligaciones o compromisos financieros. Un compromiso se considera retrasado cuando no se efectúa en la fecha límite previamente acordada o dentro del plazo de gracia concedido.
- **Perfilamiento**: es el proceso que se enfoca en identificar descripciones exhaustivas, o "perfiles", de los segmentos de mercado o clientes individuales identificados. A diferencia de la segmentación, que se ocupa de dividir el mercado en grupos distintos, el perfilamiento busca obtener, en base a los grupos definidos en la segmentación, un entendimiento preciso sobre las particularidades y el comportamiento de los segmentos o individuos reconocidos

### 4.1.7. Costos y beneficios

Los costos involucran entre otras cosas los recursos humanos y tecnológicos a utilizar. En el caso del desarrollo de este proyecto los costos se consideran los siguientes:

- **Costos relacionados a los recursos tecnológicos para el análisis (infraestructura):** estos son cubiertos por la cooperativa en su totalidad, lo cual incluye uso de servidores de base de datos, servidores de procesos ETL, recursos de red y servidores de aplicaciones
- **Costos de la construcción del modelo:** en este caso, los mismos son cubiertos por el maestrante de manera total. En este caso hablamos puntualmente del recurso humano involucrado en el proyecto y el tiempo invertido en el desarrollo y construcción del modelo
- **Capacitación del personal**: una vez entregado el modelo se prevé una capacitación técnica y operativa de las herramientas y técnicas utilizadas para este proyecto, lo cual corre por cuenta del maestrante

Los beneficios esperados incluyen entre otras cosas lo siguiente:

- Mejora en los tiempos de análisis y diseño de productos y servicios financieros
- Aumento en la eficiencia de entrega y promoción de productos y servicios
- Aumento significativo en la gestión de riesgos, relacionada a la toma de decisiones informadas al momento de ofrecer servicios a los socios
- Mejorar de la satisfacción del cliente, mediante las mejoras en la eficiencia operativa.
- Aumento los ingresos y utilidades de la cooperativa al minimizar los riesgos asociados a la entrega de productos ajustados al perfil de socios, y promocionar productos más acordes a las capacidades de pago de los socios, en lugar de negar el acceso a los mismos.

#### 4.1.8. Metas de la minería de datos y criterios de éxito

La meta del presente proyecto es desarrollar un modelo de clustering que permita un perfilamiento efectivo de socios, clasificándolos de tal manera que el equipo de riesgos pueda determinar su riesgo de morosidad y capacidad de pago de acuerdo a las reglas de negocio establecidas.

Como criterios de éxito del proyecto, establecemos los siguientes:

- Que el modelo propuesto tenga una precisión adecuada en la identificación de segmentos
- Que durante la presentación y socialización del modelo las áreas involucradas (Riesgos y Análisis de Datos) determinan que la aplicabilidad de los insights generados para el diseño de productos y servicios personalizados cumple con los requerimientos de la cooperativa.

### 4.1.9. Plan de proyecto

El presente proyecto se ha desarrollado basado en el siguiente plan:

| Fase | Semana | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Comprensión del negocio | | | | | | | | | | | | | | | | | | | | |
| Entendimiento de los datos | | | | | | | | | | | | | | | | | | | | |
| Preparación de los datos | | | | | | | | | | | | | | | | | | | | |
| Modelado | | | | | | | | | | | | | | | | | | | | |
| Evaluación | | | | | | | | | | | | | | | | | | | | |
| Despliegue | | | | | | | | | | | | | | | | | | | | |
| Presentación de resultados | | | | | | | | | | | | | | | | | | | | |
| Elaboración de informe | | | | | | | | | | | | | | | | | | | | |

*Tabla 2: Plan de proyecto. Fuente: (Elaboración propia)*

### 4.1.10. Evaluación inicial de técnicas y herramientas

Para el desarrollo del proyecto se anticipa el uso de Python como lenguaje de programación para análisis de datos y para la construcción y automatización del modelo. Se usarán principalmente las librerías como *Scikit-learn* para clustering, las librerías *pandas* y *numpy* para el manejo de datos y operaciones de normalización de datos. Finalmente usaremos la librería *matplotlib* para ilustrar la distribución de los clústeres.

En la fase de modelado, utilizaremos técnicas de aprendizaje automático no supervisado; puntualmente utilizaremos *k-means* y *DBSCAN* que son las técnicas que se adaptan más fácilmente a al objetivo de establecer la segmentación de clientes, buscando y estableciendo patrones en la información que permitan generar grupos o perfiles de socios.

Para la parte que implica el desarrollo del modelo, se utilizará Python como lenguaje de programación. Como herramienta para la construcción y automatización del modelo usaremos Jupyter Notebook para todo el proceso iterativo de procesamiento de datos, el modelado y evaluación del modelo.

Para el despliegue del modelo, se acordó con los funcionarios de la institución entregar directamente la notebook de desarrollo del modelo, con la cual ellos podrán refinar el modelo y expandir sus capacidades de acuerdo con los resultados socializados. El código Python generado en la notebook generada, se entregará al Departamento de Desarrollo de

Software de la Cooperativa para su revisión y sistematización en servicios web de carga de datos, entrenamiento del modelo y validación de datos nuevos, así como una interfaz gráfica para la invocación de dichos servicios, según los requerimientos de la plataforma tecnológica institucional, cuya competencia exclusiva del Departamento de Desarrollo de la Cooperativa.

Para el despliegue posterior, se solicita a nivel de infraestructura tecnológica un servidor de ejecución Linux (versión Rocky Linux) donde estará el servidor de despliegue de servicios Gunicorn y el motor de python que habilitará los servicios que posteriormente serán contruidos por el Departamento de Desarrollo según su planificación anual.

## 4.2. Entendimiento de los datos

### 4.2.1. Informe de Recolección de Datos Iniciales

Se obtuvo acceso a un conjunto de datos que contiene 55,687 registros correspondientes a socios de la cooperativa, con personería natural y jurídica, y que tengan al menos una operación de crédito a la fecha de corte del reporte. Los datos están estructurados en 15 columnas que representan diversas variables asociadas a los miembros de la Cooperativa de Ahorro y Crédito Jardín Azuayo. Este conjunto de datos se origina desde un cubo de datos denominado 'Base_Socios_Credito' localizado en el almacén de datos de la cooperativa, y fue recuperado a través de procesos ETL (Extracción, Transformación y Carga), los cuales integran información procedente de distintos sistemas informáticos de la empresa. La recolección y manejo inicial de estos datos se realiza en formato de archivo Excel, lo cual simplifica la ejecución del análisis preliminar.

Es importante destacar que, gracias a la implementación de procesos ETL estandarizados y a la eficaz consolidación de la información en el almacén de datos, no se identificaron inconvenientes significativos relacionados con la extracción de datos. Esta eficiencia en la gestión y recopilación de información subraya la solidez de las prácticas de manejo de datos adoptadas por la empresa.

### 4.2.2. Informe de Descripción de Datos

Las variables obtenidas en el conjunto de datos a analizar incluyen información relacionada a cada socio, proporcionando una visión integral de la situación financiera, demográfica y de riesgo de cada uno, lo cual, es esencial para el objetivo de segmentación del proyecto. Podemos clasificar las variables recopiladas de la siguiente manera:

- **Variables Financieras**: Ingresos individuales, egresos familiares, flujo de caja, y créditos solicitados y pagados proporcionan una visión completa de la salud financiera de los socios.
- **Variables Demográficas y de Riesgo**: Estado civil, edad, nivel de instrucción, y pre-calificación de riesgo ayudan a entender el perfil de riesgo y las necesidades de los socios.

Los ingresos se recopilan de forma individual, mientras que los egresos se consideran a nivel familiar, lo que requiere una normalización adecuada para el análisis. La siguiente tabla (Tabla 3) permite describir de manera detallada la información obtenida de acuerdo con la carga inicial:

| VARIABLE | DESCRIPCION | TIPO |
|---|---|---|
| CODIGO_SOCIO | Código identificador único para cada socio | Entero |
| ESTADO_CIVIL | Estado civil del socio (casado, divorciado, soltero, etc.) | Categórico |
| PROVINCIA_RESIDENCIA | Provincia de residencia del socio | Categórico |
| CARGAS_FAMILIARES | Número de cargas familiares reportadas por el socio | Entero |
| CREDITOS_TOTAL | Número total de créditos solicitados por el socio | Entero |
| CREDITOS_CANCELADOS | Número total de créditos cancelados o pagados por el socio | Entero |
| DIAS_MORA_PROMEDIO | Promedio de días de mora en pagos de créditos | Float |
| MONTO_PROMEDIO | Monto promedio solicitado en créditos | Float |
| TOTAL_INGRESO | Total de ingresos mensuales del socio | Float |
| EGRESO_SOCIO | Total de egresos mensuales del socio | Float |
| DIFERENCIA | Diferencia entre ingresos y egresos mensuales | Float |
| OFICINA | Código de la oficina que atiende al socio | Categórico |
| NIVEL_ESTUDIOS | Nivel de estudios alcanzado por el socio | Categórico |
| EDAD | Edad del socio | Entero |
| TIPO_PERSONA | Tipo de persona (natural, jurídica, etc.) | Categórico |

*Tabla 3: Tabla de descripcion de varaiables*

### 4.2.3. Informe de Exploración de Datos

La exploración preliminar busca identificar patrones, correlaciones entre variables como ingresos, egresos y capacidad de pago, y agrupaciones naturales que puedan indicar diferentes perfiles de riesgo y capacidad financiera. Especial atención se dará a la relación entre el flujo de caja y los días de mora promedio, así como a cómo factores demográficos y de nivel de educación pueden influir en la segmentación.

*Análisis descriptivo*

Se muestra en primera instancia en análisis estadístico descriptivo de las variables numéricas (Tabla4):

| VARIABLE | count | mean | std | min | 0,25 | 0,50 | 0,75 | max |
|---|---|---|---|---|---|---|---|---|
| **CODIGO_SOCIO** | 55687,00 | 551740,61 | 265153,18 | 10015,00 | 330964,50 | 569483,00 | 803721,00 | 909435,00 |
| **CARGAS_FAMILIARES** | 55687,00 | 1,03 | 1,11 | 0,00 | 0,00 | 1,00 | 2,00 | 12,00 |
| **CREDITOS_TOTAL** | 55687,00 | 5,09 | 10,66 | 1,00 | 1,00 | 2,00 | 5,00 | 338,00 |
| **CREDITOS_CANCELADOS** | 55687,00 | 3,69 | 9,86 | 0,00 | 0,00 | 1,00 | 4,00 | 305,00 |
| **DIAS_MORA_PROMEDIO** | 55687,00 | 7,13 | 22,88 | 0,00 | 0,00 | 2,00 | 7,00 | 1871,00 |

*Tabla 4: Análisis estadístico descriptivo inicial para variables numéricas. Fuente: (Elaboración propia)*

De manera inicial se puede ver que estos datos no son todos aquellos que están definidos como datos numéricos a nivel de la fuente de datos. Por lo tanto, se verifica en primera instancia los tipos de datos recuperados por el dataframe (Tabla 5.):

| VARIABLE | TIPO DE DATO |
|---|---|
| **CODIGO_SOCIO** | int64 |
| **ESTADO_CIVIL** | object |
| **PROVINCIA_RESIDENCIA** | object |
| **CARGAS_FAMILIARES** | int64 |
| **CREDITOS_TOTAL** | int64 |
| **CREDITOS_CANCELADOS** | int64 |
| **DIAS_MORA_PROMEDIO** | int64 |
| **MONTO_PROMEDIO** | object |
| **TOTAL_INGRESO** | object |
| **EGRESO_SOCIO** | object |
| **DIFERENCIA** | object |
| **OFICINA** | object |
| **NIVEL_ESTUDIOS** | object |
| **EDAD** | object |
| **TIPO_PERSONA** | object |

*Tabla 5: Tipos de datos recuperados en la carga del dataset. Fuente: (Elaboración propia)*

Como podemos ver, campos numéricos como ‘MONTO_PROMEDIO’, ‘TOTAL_INGRESOS’ y otros que se conoce previamente son numéricos, han sido recuperados desde el archivo de datos como campos de tipo ‘object’ por lo cual se hace

necesario primero un proceso de transformación o preprocesamiento de datos, para posteriormente tener una buena lectura de los estadísticos iniciales, esto será cubierto en la siguiente fase.

Sin embargo, realizamos un análisis exploratorio inicial de acuerdo con lo recuperado (Tabla 3.), encontrando las siguientes novedades:

- Los códigos de socio, como identificadores únicos, exhiben la esperada variabilidad amplia y no son de utilidad para el objetivo del proyecto, deberá eliminarse esta variable del proceso de modelado
- Se destaca que la mayoría de los socios tienen entre 0 y 2 cargas familiares, con casos que alcanzan hasta 12, lo que indica diferencias significativas en las responsabilidades familiares dentro de la población de socios, podría haber ciertos datos que pueden ser outliers, pero su validez se determinará con las áreas operativas que manejan directamente la información que se registra desde oficinas.
- En lo que respecta a los créditos, tanto totales como cancelados, la variación es notable, con una media de 5 y 3.69 respectivamente, pero con extremos que llegan a 338 créditos totales y 305 cancelados, esto principalmente se debe a que la cooperativa ha otorgado a sus socios un producto denominado 'línea de crédito' que simula una tarjeta de crédito donde el socio realiza avances en efectivo pudiendo diferir los mismos en un número seleccionado de meses. Esto hace que estos avances diferidos sean tratados como créditos pequeños y por ello la gran cantidad de operaciones crediticias analizadas.
- Los días de mora promedio, aunque con valores relativamente pequeños (7 días), muestra un rango amplio hasta 1871 días, revelando que, mientras la mayoría cumple en tiempo sus pagos, existe un grupo menor con morosidades significativas.

A continuación, realizamos la exploración inicial de las variables categóricas con lo que tenemos la siguiente información (Tabla 6):

| VARIABLE | count | unique | top | freq |
|---|---|---|---|---|
| ESTADO_CIVIL | 55555 | 5 | SOLTERO/A | 24029 |
| PROVINCIA_RESIDENCIA | 55687 | 52 | AZUAY | 20772 |
| MONTO_PROMEDIO | 55687 | 6939 | 10000 | 4340 |
| TOTAL_INGRESO | 55687 | 20236 | 600 | 1407 |
| EGRESO_SOCIO | 55687 | 22089 | #N/D | 409 |
| DIFERENCIA | 55687 | 28208 | #N/D | 409 |
| OFICINA | 55687 | 70 | 15 | 2347 |
| NIVEL_ESTUDIOS | 55687 | 8 | SECUNDARIA | 23509 |
| EDAD | 55687 | 79 | 30 | 1792 |
| TIPO_PERSONA | 55687 | 4 | PERSONA NATURAL | 55147 |

*Tabla 6: Análisis estadístico descriptivo inicial para variables categóricas. Fuente: (Elaboración propia)*

Aquí podemos ver cómo, efectivamente, existe un problema en la carga inicial de los datos ya que ciertas variables numéricas que de antemano se conoce que son numéricas, se están cargando como variables categóricas, con lo cual, es necesario en primera instancia ejecutar procesos de transformación de datos para clasificar las variables del modelo de forma adecuada. A pesar de ello, preliminarmente podemos establecer los siguientes hallazgos en esta exploración inicial:

- Existen 5 estados civiles únicos dentro de la base de socios, siendo "SOLTERO/A" el más prevalente con 24,029 registros lo cual puede indicar donde está concentrada la mayor parte de socios activos y con operaciones de crédito activas.
- Los socios provienen de 52 provincias distintas, destacando "AZUAY" como la más común con 20,772 socios. Dado que en Ecuador solo hay 24 provincias, es de suponer que existen datos de socios registrados en lugares geográficos fuera el Ecuador, ya que hay socios que, en su categoría de migrantes, igualmente tienen acceso a los servicios de la cooperativa. Esto deberá regularizarse de alguna manera con algun tipo de variable nueva que agrupe los valores de esta columna
- Se registran 70 oficinas, con la oficina "15"(Oficina La Troncal) mencionada 2,347 veces.
- Existen 8 niveles de estudios reportados, siendo "SECUNDARIA" el más común con 23,509 casos.
- En cuanto a SECTOR_DOMICILIO, se diferencian 4 sectores, siendo el sector "URBANO" el más frecuente (35,937 veces). Dado que los sectores son

solamente 2, esto requerirá depurar los datos restantes que pueden estar vacios o no definidos

- La EDAD, tratada como categórica debido a su formato original, identifica "30" años como la edad más reportada (1,792 veces).
- Hay 4 tipos de persona, siendo el más frecuente "PERSONA NATURAL" con 55,147 registros. Dado que aquí estamos manejando personas naturales y jurídicas, y que el proyecto incluso plantea únicamente el trabajo de modelado para personas naturales, probablemente existan datos vacíos o no definidos los cuales deberán ser tratados en la siguiente fase

En base a lo observado de manera inicial podemos establecer algunas consideraciones para la siguiente fase:

- Es necesario que se ejecute el proceso de transformación de datos que ajuste adecuadamente los tipos de datos para poder realizar el resto del análisis que debían ejecutarse en esta fase.
- Se espera correlacionar los ingresos y egresos con la capacidad de pago y los días de mora para identificar perfiles de riesgo.
- La concentración geográfica de socios en ciertas provincias y la prevalencia de domicilios urbanos, lo que podría influir en las estrategias de servicios y productos de la cooperativa, al existir varios lugares de residencia, lo más adecuado sería crear una variable nueva que agrupe esta variable;
- La edad y el estado civil podrían influir en la capacidad de pago y preferencias de productos financieros. Esta misma diversidad en estados civiles y niveles educativos sugiere que se podría desarrollar enfoques de promoción segmentados en la oferta de productos para estos socios
- Igualmente se hace necesario un proceso de normalización de ingresos y egresos, así como de codificación de variables categóricas

Como parte del proceso inicial verificamos la existencia de datos nulos o no definidos en cada columna, definiendo así cual puede ser el tratamiento que debería darse a estos registros (Tabla 7)

| VARIABLE | CANTIDAD NULOS |
|---|---|
| CODIGO_SOCIO | 0 |
| ESTADO_CIVIL | 132 |
| PROVINCIA_RESIDENCIA | 0 |
| CARGAS_FAMILIARES | 0 |
| CREDITOS_TOTAL | 0 |
| CREDITOS_CANCELADOS | 0 |
| DIAS_MORA_PROMEDIO | 0 |
| MONTO_PROMEDIO | 0 |
| TOTAL_INGRESO | 409 |
| EGRESO_SOCIO | 409 |
| DIFERENCIA | 409 |
| OFICINA | 409 |
| NIVEL_ESTUDIOS | 409 |
| EDAD | 409 |
| TIPO_PERSONA | 409 |

*Tabla 7: Conteo de variables con valores nulos o no definidos. Fuente: (Elaboración propia)*

### 4.2.4. Informe de Calidad de Datos

En esta sección la evaluación de los datos se centrará en aspectos fundamentales como la completitud, precisión y consistencia. Un elemento esencial de esta revisión será la verificación de la coherencia entre los ingresos individuales y los egresos familiares, además de la integridad de la información relacionada con créditos y morosidad.

Desde el inicio de este proyecto, se enfatizó la importancia de contar con información completa y precisa, lo cual es fundamental para garantizar un análisis confiable y efectivo. Es crucial que los datos representen de manera fiel las características financieras y demográficas de los socios de la cooperativa, evitando errores significativos o ausencias que puedan influir negativamente en los resultados del análisis.

Para evaluar la calidad de los datos, se implementaron diversas estrategias, incluyendo la revisión de valores faltantes para identificar registros incompletos o con marcadores de posición como '#N/D', y la verificación de los tipos de datos en cada columna para confirmar su adecuación (numérico, categórico).

La detección de outliers no pudo ser establecida en esta fase ya que se requiere primero la ejecución de procesos de limpieza y transformación de datos que corrijan el problema mencionado de inconsistencia en los tipos de datos del dataset.

Los resultados de esta evaluación revelaron la presencia de valores faltantes en diversas columnas, tipos de datos inapropiados que requerían normalización, lo cual podría indicar tanto errores de registro como diversidad auténtica en la información financiera de los socios. Para mejorar la calidad de los datos, se recomienda llevar a cabo una limpieza detallada que incluya estrategias para abordar los valores faltantes y especiales, validar y corregir outliers que no correspondan a situaciones reales, y convertir y estandarizar los formatos de todas las variables al tipo de dato adecuado, especialmente aquellas variables numéricas inicialmente reconocidas como tipo objeto, asegurando así la integridad y la fiabilidad de los datos para análisis subsiguientes.

## 4.3. Preparación de los datos

### 4.3.1. Selección de variables

En este paso definiremos la selección de las variables que serán usadas en procesamiento del modelo, considerando las variables que tiene el dataset obtenido desde el datawarehouse de la cooperativa, con las siguientes consideraciones:

- La variable ‘CODIGO_SOCIO’ se descarta ya que no aporta a establecer una relación o integración de datos significativa para el modelo.
- La variable 'OFICINA', no se considera dentro del estudio ya que, al tener algún tipo de afectación sobre el proceso de segmentación, la política institucional no implica brindar mayores o menores facilidades o beneficios a los socios que pertenecen a una determinada oficina, razón por la cual se descarta la variable.
- En el caso de la variable ‘PROVINCIA_RESIDENCIA’ se ha visto que al manejar múltiples sectores que pueden dividir demasiado los segmentos y teniendo en cuenta que hay registros de provincias que no pertenecen al Ecuador, a nivel de las partes de negocio que dan seguimiento a este proyecto (Dirección de Riesgos, Departamento de Análisis de Datos), se considera que es mejor agrupar las provincias en regiones (Costa, Sierra, Oriente, y Otros) para reducir esa variabilidad y tener un mejor campo de agrupamiento. Se creará la variable ‘REGION’ para este fin

- Se optará por limitar el estudio únicamente a personas naturales, por lo cual los registros de personas jurídicas serán eliminados del dataset para el proceso de modelado. Esto

### 4.3.2. Limpieza y estandarización de datos

En este paso, utilizaremos diferentes técnicas para mejorar el conjunto de datos y dar tratamiento a aquellos datos que pueden afectar de manera negativa el proceso de modelado. A continuación, se detalla el proceso seguido en función de los casos identificados en la fase anterior de Entendimiento de los Datos:

- Empezaremos eliminando aquellas variables que no vamos a utilizar, en este caso 'CODIGO_SOCIO' y 'OFICINA'
- En segunda instancia, se dará un tratamiento de transformación de datos para convertir aquellas variables que fueron importadas como 'object' al tipo de datos correspondiente y ejecutar un proceso de exploración más detallado al que se hizo inicialmente.
- Para los datos vacíos o, menores o iguales a cero, correspondientes a las variables financieras 'TOTAL_INGRESO', 'EGRESO_SOCIO', 'DIFERENCIA' y 'MONTO_PROMEDIO' se hará una eliminación de estos valores en base a lo acordado juntamente con una de las partes interesadas (Dirección de Riesgos), ya que se considera que al registrar un socio con ingresos de cero, o cuyo flujo de caja sea negativo, es un error operativo al momento de registrar o actualizar la información financiera de los socios, por tanto, dichos datos no deberían considerarse al momento de generar el modelo de segmentación., sino hasta que hayan sido corregidas las inconsistencias a nivel operativo.
- La variable 'DIFERENCIA' tendrá un cambio de nombre a 'FLUJO_CAJA'.
- En el caso de la variable 'EDAD', igualmente se hará una imputación de datos usando la moda (esto igualmente acordado con la parte de negocio interesada).
- En el caso del tratamiento de valores nulos que están relacionados a personas naturales, como es el caso de la columna 'ESTADO_CIVIL' o 'EDAD', estos deberán ser eliminados ya que corresponderían a registros de personas jurídicas. Dado que se eliminarán los registros de personas jurídicas del estudio en base a lo acordado con la parte interesada de negocio, esta acción como tal eliminaría directamente los registros con estado civil o edad en nulo.

- Finalmente se crea la nueva variable ‘REGION’ la cual agrupa los datos de las provincias de residencia de acuerdo con la divisón política actual del país. Aquellos datos registrados como provincias que no corresponden a la división política serán agrupados en una región que se denomina ‘Otros’.

Realizados estos procesos de limpieza de datos, podemos volver a generar las estadisticas descriptivas en las variables numéricas y categóricas como se muestra en las tablas 8, 9 y 10

| VARIABLE | TIPO DE DATO |
|---|---|
| CODIGO_SOCIO | int64 |
| ESTADO_CIVIL | object |
| PROVINCIA_RESIDENCIA | object |
| CARGAS_FAMILIARES | int64 |
| CREDITOS_TOTAL | int64 |
| CREDITOS_CANCELADOS | int64 |
| DIAS_MORA_PROMEDIO | int64 |
| MONTO_PROMEDIO | float64 |
| TOTAL_INGRESO | float64 |
| EGRESO_SOCIO | float64 |
| FLUJO_CAJA | float64 |
| NIVEL_ESTUDIOS | object |
| SECTOR_DOMICILIO | object |
| EDAD | float64 |
| TIPO_PERSONA | object |
| NIVEL_RIESGO_PREC | object |

*Tabla 8: Tipos de datos corregidos en limpieza de datos. Fuente: (Elaboración propia)*

| VARIABLE | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **CARGAS_FAMILIARES** | 54404,00 | 1,03 | 1,11 | 0,00 | 0,00 | 1,00 | 2,00 | 12,00 |
| **CREDITOS_TOTAL** | 54404,00 | 5,12 | 10,69 | 1,00 | 1,00 | 2,00 | 5,00 | 338,00 |
| **CREDITOS_CANCELADOS** | 54404,00 | 3,72 | 9,89 | 0,00 | 0,00 | 1,00 | 4,00 | 305,00 |
| **DIAS_MORA_PROMEDIO** | 54404,00 | 7,17 | 23,05 | 0,00 | 0,00 | 2,00 | 7,00 | 1871,00 |
| **MONTO_PROMEDIO** | 54404,00 | 8922,46 | 10645,58 | 18,67 | 2010,71 | 5000,00 | 12000,00 | 700000,00 |
| **TOTAL_INGRESO** | 54404,00 | 6754,91 | 1070439,29 | 2,00 | 728,55 | 1200,00 | 2144,80 | 249673666,70 |
| **EGRESO_SOCIO** | 54404,00 | 1424,67 | 5440,25 | 0,00 | 390,00 | 676,00 | 1357,61 | 535292,29 |
| **FLUJO_CAJA** | 54404,00 | 5330,24 | 1070284,55 | 0,00 | 246,00 | 418,02 | 766,25 | 249640097,60 |
| **EDAD** | 54404,00 | 40,51 | 13,73 | 18,00 | 30,00 | 38,00 | 50,00 | 94,00 |

*Tabla 9: Análisis estadístico descriptivo corregido en limpieza de datos. Fuente: (Elaboración propia)*

| VARIABLE | count | unique | top | freq |
|---|---|---|---|---|
| **ESTADO_CIVIL** | 54404 | 5 | SOLTERO/A | 23623 |
| **NIVEL_ESTUDIOS** | 54404 | 7 | SECUNDARIA | 23156 |
| **TIPO_PERSONA** | 54404 | 1 | PERSONA NATURAL | 54404 |
| **REGION** | 54404 | 4 | SIERRA | 34640 |

*Tabla 10: Análisis estadístico descriptivo corregido con limpieza de datos. Fuente: (Elaboración propia)*

Con las acciones realizadas, ya no es necesario tener la columna 'TIPO_PERSONA' porque existe una sola frecuencia, por tanto, también se procede a eliminar dicha columna. Se procede entonces al análisis grafico de variables numericas y categóricas en busca de outliers, problemas de distribución de las variables, etc.

### 4.3.3. Análisis gráfico y tratamiento de los datos

Este análisis inicial nos permitirá obtener insights sobre la necesidad de transformación de datos, limpieza y cualquier otro pre-procesamiento necesario, así como conclusiones relacionadas con los objetivos de minería de datos y objetivos de negocio. A continuación, visualizamos el gráfico de cajas relacionado con las variables numéricas (Figura 10)

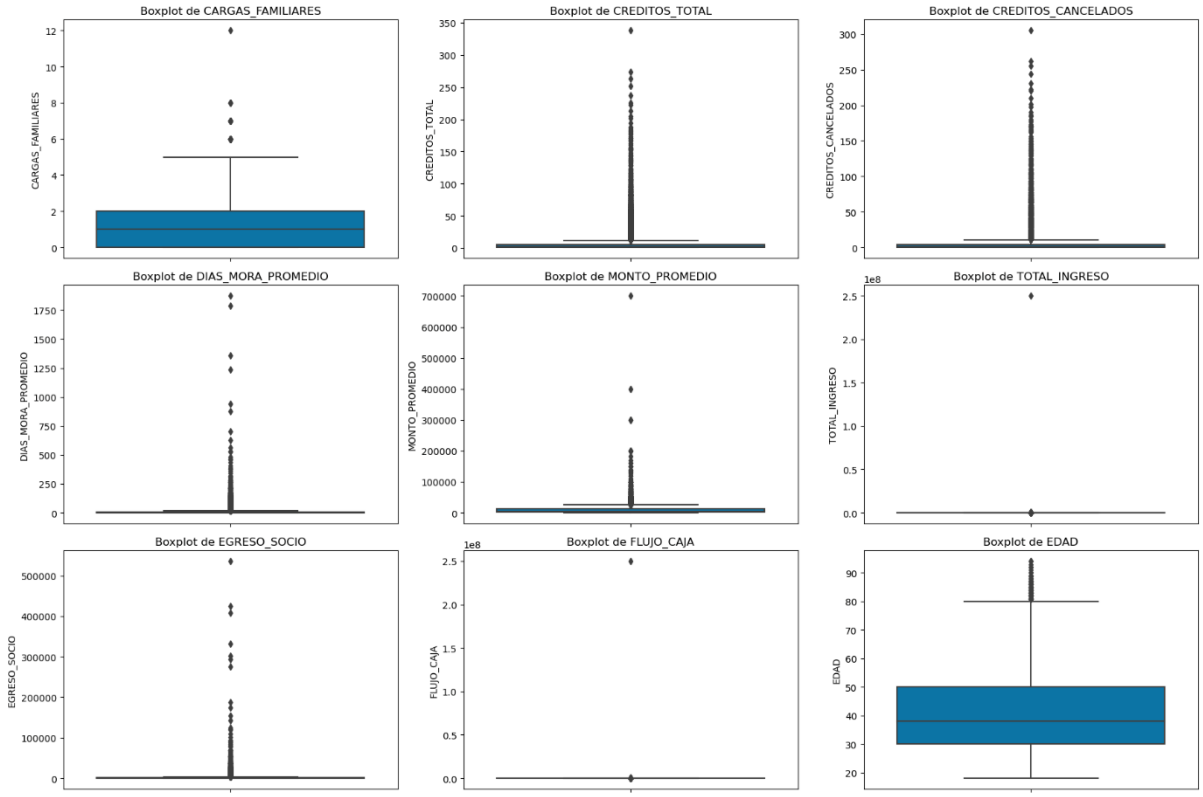



*Figura 10: Diagramas de caja para variables numéricas. Fuente: (Elaboración propia)*

En función de los gráficos mostrados, podemos realizar las siguientes observaciones:

- En el gráfico de las cargas familiares se observa que la mayoría de los socios reportan tener pocas responsabilidades familiares, aunque existen casos aislados con un número considerablemente más alto de estas.
- En cuanto a los créditos totales y cancelados, ambas variables exhiben un patrón similar, donde la tendencia general es que los socios poseen una cantidad moderada de créditos, no obstante, se identifican outliers que destacan por tener un volumen elevado de créditos.
- Respecto a los días de mora promedio, se nota que la mayor parte de los socios presenta pocos días de mora, pero se encuentran excepciones significativas que muestran morosidades notablemente altas.

- Las variables de Monto Promedio, Total Ingreso, Egreso del Socio y Flujo de Caja, demuestran una diversidad considerable en los montos financieros manejados por los socios, con la presencia de varios valores atípicos que subrayan una variabilidad importante en la situación financiera de los mismos. Sin embargo, de acuerdo con la descripción de las estadísticas descriptivas, al parecer la mayoría de los datos estan ubicados dentro de los 3 primeros cuartiles, lo que da a entender que los datos de los socios son muy parecidos entre sí, lo que podría significar que la segmentación podría dificultarse al no tener muchos elementos diferenciadores.
- Por otro lado, es visible la existencia de un dato que sale de la escala completa de los valores máximos en los graficos de caja de 'TOTAL_INGRESO' y 'FLUJO_CAJA'. Al revisar este dato en particular, se trata de un caso especial de una persona natural con ingresos extraordinariamente altos. Aunque se valida con personal de negocio que los datos posiblemente son reales según el conocimiento del cliente, es posible que esto genere por sí mismo un segmento propio al estar tan alejado del resto de los casos, por lo cual trataremos este dato como outlier y será eliminado
- La distribución de la edad de los socios muestra un balance razonable, cubriendo un espectro amplio desde individuos jóvenes hasta personas mayores, aunque se detectan outliers principalmente en el extremo superior de la escala de edades.

Este panorama refleja la heterogeneidad en las condiciones económicas y demográficas de los socios, resaltando la presencia de casos excepcionales en diversas dimensiones analizadas. Dada la cantidad de registros existente, no realizaremos un tratamiento de outliers en el resto de los casos ya que se verifican a nivel de negocio los casos más representativos y al ser datos posiblemente reales y que no conllevan una afectación significativa al proceso de segmentación, no tendrán ningún tratamiento ya que son datos reales.

Por otro lado, analizando las variables categóricas (Figura 11.) podemos observar las siguientes novedades respecto al conjunto de datos a analizar:

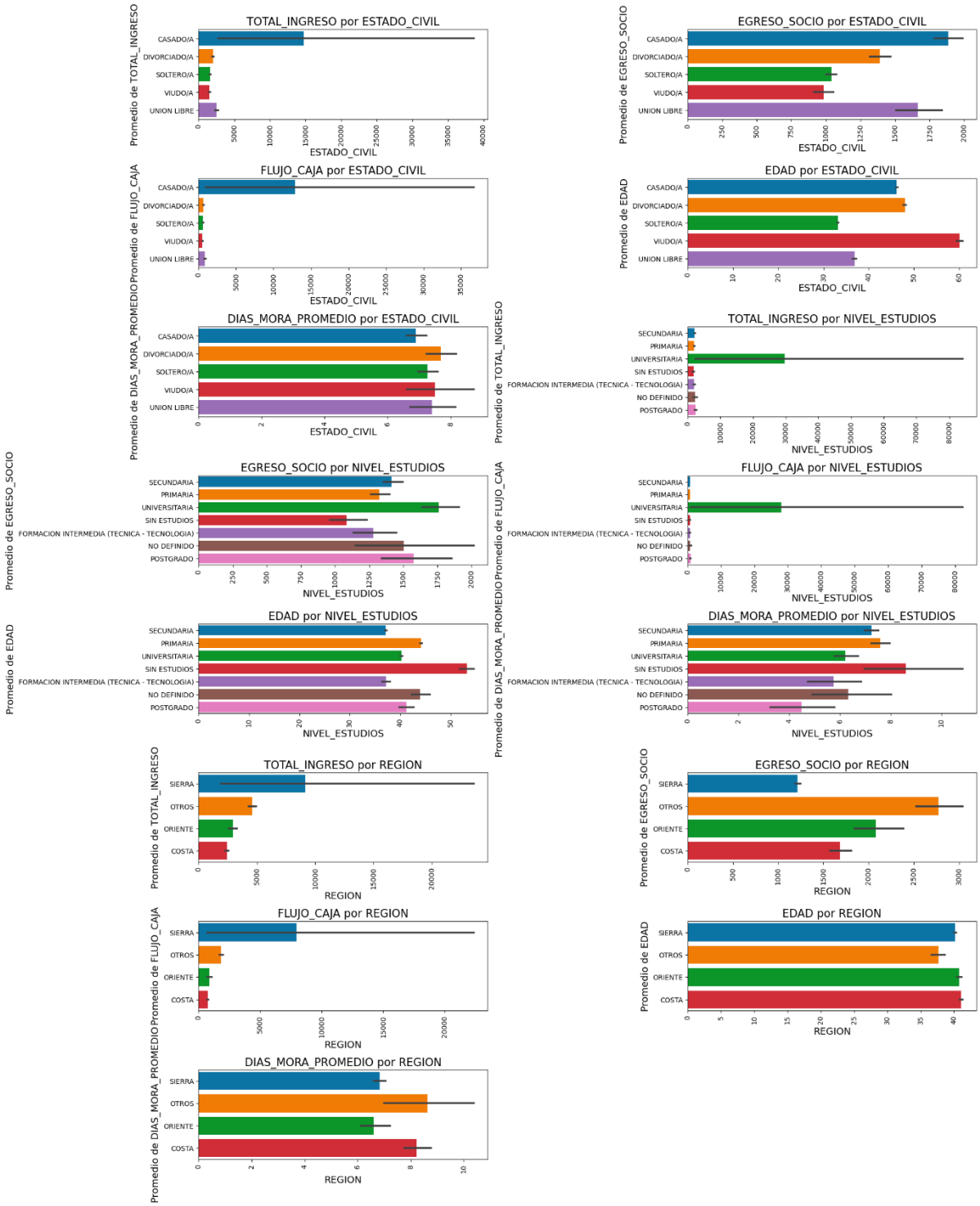



*Figura 11: Diagramas de barra para variables categóricas. Fuente: (Elaboración propia)*

Los análisis de los ingresos y egresos revelan una notable variabilidad en función de las variables categóricas, mostrando que distintos segmentos demográficos y geográficos poseen capacidades y obligaciones financieras que varían significativamente. En lo que respecta al flujo de caja, definido como la diferencia entre los ingresos y los egresos, se

obtiene una perspectiva crucial sobre la solidez financiera de los socios distribuidos en diversas categorías.

Es notorio que ciertas categorías exhiban un flujo de caja más restringido, sugiriendo una demanda para productos financieros diseñados específicamente para gestionar o mejorar su liquidez. Por otro lado, el análisis según la edad demuestra patrones distintivos en relación con el estado civil y el nivel educativo, reflejando posiblemente diferentes fases de la vida y las correspondientes necesidades financieras que acompañan a cada una de estas fases.

En observaciones más detalladas, se percibe que el nivel educativo ejerce una influencia notable en los ingresos, egresos y el flujo de caja, señalando potenciales oportunidades para la propuesta de productos de ahorro o inversión que se ajusten a este criterio.

Finalmente, las disparidades observadas en función de la región de residencia en todas las variables numéricas enfatizan la importancia de adoptar estrategias financieras que contemplen las particularidades económicas de cada región, subrayando la necesidad de enfoques financieros que sean sensibles a las condiciones económicas locales.

### 4.3.4. Normalización y estandarización de datos

Para culminar el proceso de preparación de los datos, procederemos a generar la normalización del conjunto de datos, donde básicamente efectuaremos un tratamiento de estructuración de las variables categóricas, y una estandarización de las variables numéricas. Para ello usaremos las técnicas ‘One-Hot Encoder’ y ‘Standard Scaler’ disponibles en Python de tal manera que con la primera transformemos las variables categóricas en variables numéricas, y que las variables numéricas estandaricen sus rangos de valores de tal forma que tengan una media de 0 y una desviación estándar de 1. El resultado de esta aplicación es el siguiente:

- Se generan 5 nuevas variables a partir de los diferentes valores de la columna ‘ESTADO_CIVIL’
- Se crean 7 nuevas columnas basadas en cada una de las categorias de la variable ‘NIVEL_ESTUDIOS’
- Finalmente, se generan 4 nuevas características derivadas de la variable ‘REGION’

- Las variables numéricas estandarizan su valor para estar dentro de un rango entre 0 y 1

Este procesamiento queda confirmado al observar las nuevas dimensiones del dataframe normalizado donde el número de columnas es mayor a la inicial, contando con 25 variables dentro del dataframe que será usado para el proceso de modelado (Figura 12.).

```
# Aplicando el preprocesamiento a los datos
datos_normalizados = preprocesador.fit_transform(clientes_copia1)
datos_normalizados.shape
```

```
(54773, 25)
```

*Figura 12: Resultado de normalización usando técnicas 'One-Hot Encoder' y 'Standard Scaler'. Fuente: (Elaboración propia)*

Con la ejecución de este último paso, la etapa de preparación de datos ha incrementado notablemente la calidad y aplicabilidad del conjunto de datos para alcanzar los objetivos establecidos en el proyecto. Las medidas implementadas se han centrado en resolver problemas fundamentales relativos a la calidad de los datos, sentando una base sólida para proceder a un modelado eficaz. Este proceso meticuloso de preparación asegura que el dataset esté optimizado para el análisis y la interpretación subsiguientes, lo que es crucial para extraer insights valiosos y construir modelos lo más precisos posible.

## 4.4. Modelado

### 4.4.1. Supuestos para ejecutar el modelado

Partimos de la premisa de que el conjunto de datos, una vez preprocesado, refleja de manera fiel y exacta la población de socios pertenecientes a la cooperativa. En lo que concierne a la aplicación de algoritmos de clustering, existe la suposición inicial de que las características seleccionadas para el análisis mantienen una independencia relativa entre sí. No obstante, se contempla llevar a cabo evaluaciones específicas destinadas a verificar la validez de esta suposición de independencia. Este paso es esencial para confirmar que las interrelaciones entre variables no sesguen los resultados del agrupamiento, garantizando así la integridad y la fiabilidad del modelado estadístico.

### 4.4.2. Diseño de prueba

El proceso de modelado se enfocará en la utilización de dos métodos de clustering: K-Means y DBSCAN, con el objetivo de segmentar a los socios basándose en sus

características financieras y demográficas. En el caso de K-Means, se llevará a cabo una exploración variando el número de clústeres desde 2 hasta 9 para determinar el número óptimo que ofrece el mejor resultado de acuerdo al silhouette score,

Por otro lado, en el método DBSCAN, se experimentará ajustando los parámetros de eps (distancia máxima entre dos muestras para ser consideradas en el mismo vecindario) y min_samples (número mínimo de muestras para formar un clúster) para hallar la configuración que optimiza el silhouette score.

El silhouette score será utilizado como la principal herramienta de evaluación para ambos modelos, proporcionando una estimación de qué tan similares son los objetos dentro de su propio clúster en contraste con aquellos pertenecientes a diferentes clústeres, facilitando así la identificación de la estructura de agrupamiento más coherente y diferenciada.

Para este modelo de clustering, no se establece una división entre datos de prueba y entrenamiento, en este caso se utiliza toda la información obtenida del datawarehouse para entrenar el modelo y obtener los centroides que permiten agrupar los datos en los diferentes clústeres

### 4.4.3. Descripción del modelo

Durante esta etapa del proyecto, se han implementado 2 modelos de agrupamiento: K-Means y DBSCAN, con el propósito de generar segmentos de socios con ciertas características comunes que puedan posteriormente identificarse como perfiles de clientes. Mientras el modelo K-means obedece a las configuraciones del número de clústeres con los que queremos iniciar el proceso de agrupamiento, en el caso de DBSCAN, éste localiza regiones de alta densidad que están separadas entre sí por regiones de baja densidad mediante la medición de la distancia entre los puntos cercanos entre si (épsilon).

En primera instancia, se ejecuta el algoritmo K-Means. En la configuración de este modelo, se experimentó variando el número de clústeres de 2 a 9, para determinar el número óptimo de segmentos. A nivel de parámetros, la inicialización del algoritmo, este se efectuó utilizando el método 'k-means++', y el resto de hiperparámetros del modelo es mantienen aquellos por defecto. Para tener una métrica comparación con el modelo DBSCAN, implementamos este modelo almacenando los resultados de la evaluación del

silhouette score de cada iteración de numero de clúster y con ello generar un gráfico que nos permita ver el comportamiento con cada intento al variar el número de clústeres.

Ejecutando pruebas sucesivas de manera individual a nivel de cambio de hiperparámetros como el número de iteraciones del modelo, el algoritmo de ejecucion (‘lloyd’ o ‘elkan’), el método de elección del centroide de inicio (‘kmeans++’ o ‘random’), o la tolerancia (‘tol’), no vemos mayores cambios en los resultados del score de silueta.

Por otro lado, en el caso de DBSCAN, este se distingue por su capacidad para identificar agrupaciones de formas irregulares y destacar puntos de datos aislados como outliers, lo que lo convierte en una herramienta valiosa para discernir comportamientos financieros altamente atípicos entre los socios. La configuración de parámetros del modelo, específicamente los valores de ‘eps’ (distancia máxima entre dos puntos para ser considerados vecinos) y min_samples (el número mínimo de puntos requeridos para formar un clúster), se ajustó para optimizar la identificación de agrupaciones significativas, siendo esta selección determinante para el desempeño efectivo del modelo, como buena práctica, se considera que este valor de ‘min_samples’ sea siempre 5.

En cuanto a las variaciones de hiperparámetros, se trabajó igual que con k-means facilitando en primera instancia los datos por defecto, y luego realizar modificaciones de los parámetros como ‘min_samples’ colocando valores mayores a 5, en ‘metric’ cambiando entre los metodos ‘euclidean’, ‘manhattan’ y ‘cosine’, y finalmente en ‘algoritm’ donde hicimos variaciones entre ‘kd_tree’ o ‘brute’ (por defecto se maneja ‘auto’), sin embargo, no se ven mayores cambios a nivel del score de silueta y el número de clústeres detectados por el algoritmo (siempre genera 1 solo clúster.)

### 4.4.4. Evaluación del modelo

La evaluación del modelo se realizará a través del Silhouette Score, una medida que permite estimar la cohesión interna de los clústeres y su separación mutua ya que lo que nos interesa es que cada clúster este delimitado adecuadamente. Según lo conversado a nivel de negocio con las áreas interesadas, manejaremos un umbral para el score de silueta de 0.5; todo modelo con un score de silueta mayor a 0.5 es adecuado y se apega a las necesidades de la institución.

Para hacer una evaluación inicial generamos una gráfica comparativa respecto a cómo se comporta el score de silueta a medida que se van aumentando la cantidad de clústeres

según el algoritmo utilizado. Las figuras 13 y 14 muestran el comportamiento de cada algoritmo y según el número de clústeres identificado y el valor del score de silueta en cada iteración.

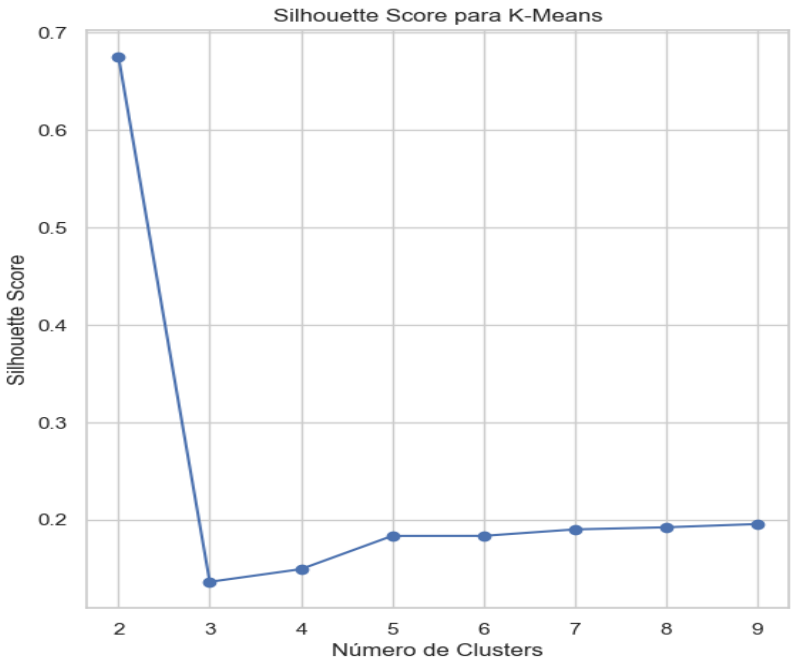



*Figura 13: Evaluación de Silouhette Score para modelo K-Means. Fuente: (Elaboración propia)*

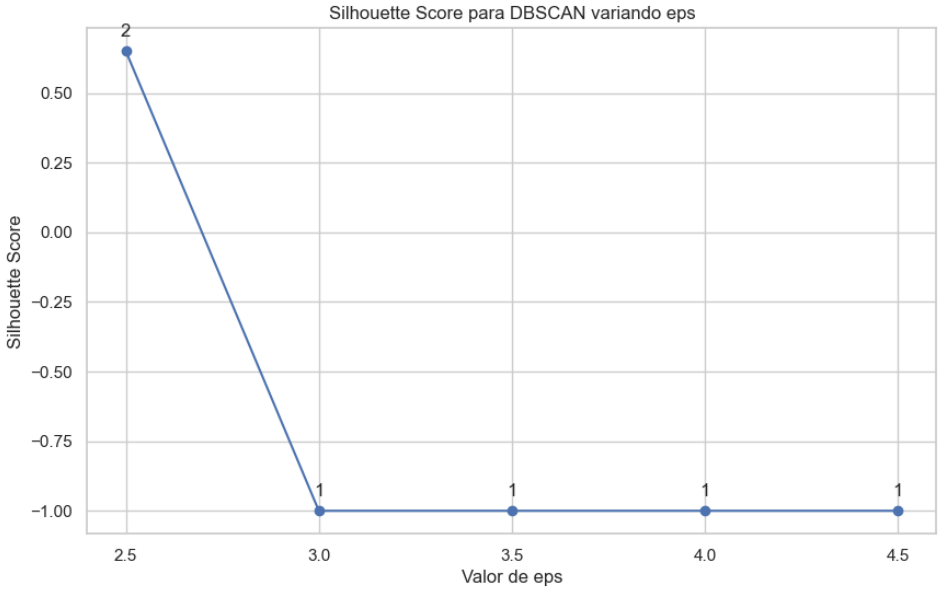



*Figura 14: Evaluación de Silouhette Score para modelo DBSCAN. Fuente: (Elaboración propia)*

Como puede verse en las gráficas, mientras el algoritmo k-means reconoce e identifica al menos 2 clústeres con un score de silueta dentro del umbral mínimo definido, el algoritmo DBSCAN no es capaz de identificar más de 2 clústers con un valor aceptable de score de silueta (casi cercano al umbral mínimo) lo cual hace que descartemos este modelo.

A pesar de que el modelo k-means ha sido capaz de identificar 2 grupos de socios definidos de manera adecuada, tener únicamente 2 grupos de socios también puede resultar insuficiente, por lo cual, generaremos una iteración de refinamiento del modelo con las siguientes consideraciones acordadas tanto con los interesados de las unidades de negocio de la cooperativa, como con el director de este proyecto de titulación:

- Buscaremos a nivel del cubo de información del datawarehouse de la cooperativa si es posible integrar nuevas variables que permitan impactar de mejor manera el conjunto de datos, aumentando clústeres obtenidos sin disminuir la eficiencia del modelo de acuerdo con el umbral mínimo definido.
- Tomando en cuenta lo anterior, se definen 2 nuevas variables dentro delo conjunto de datos:
    - 'SECTOR_DOMICILIO': esta variable define si el socio se encuentra domiciliado en el área urbana o rural
    - 'NIVEL_RIESGO_PREC': relacionada con la calificación actual de confianza que tiene un socio y en base a la cual se determinan los prerrequisitos (documentación) que debe cumplir un socio cuando solicita el producto de crédito.
- Se realiza el mismo proceso de depuración y limpieza de estas nuevas variables para que se integren al dataset que ahora será parte del proyecto.
- En vista de que los resultados del modelo k-means solamente definen 2 segmentos de socios, se propone una solución que implica generar niveles de profundidad en el proceso de descubrimiento de los segmentos. De tal manera que, una vez que el modelo cree los 'k' clusters con un score de silueta dentro de los umbrales mínimos, el algoritmo deberá replicarse a sí mismo, pero sobre uno de los grupos o clústeres descubiertos en la primera iteración y que deberá ser aquel con la mayor cantidad de registros ya que se entiende que tiene una concentración o densidad tales que posiblemente dentro de si pueda segmentarse nuevamente para

crear clústeres mejor definidos desde la perspectiva de ese segmento y ya no desde el dataset completo.

- Definiremos hasta 2 niveles más de profundad contados tras el primer agrupamiento del modelo, con lo que esperamos obtener al menos 6 grupos identificados de clientes según lo que requiere el interesado de negocio. En cada nivel, el score de silueta deberá mantenerse por encima del umbral mínimo definido.
- Finalmente, aplicaremos el analisis de componentes principales (PCA) al conjunto de datos normalizado de tal manera que, por una parte, el modelo se enfoque en las direcciones de máxima varianza en los datos las cuales a menudo contienen la información más relevante, y por otro lado, ayude a reducir posibles afectaciones al modelo debido a la multicolinealidad, así como equilibrar la contribución de cada variable al análisis, tomando en cuenta que las variables originales tienen diferentes unidades o rangos de variación como ya hemos visto durante la fase de entendimiento de los datos

Con estos ajustes, el conjunto de datos queda compuesto de la siguiente manera (Tabla 11):

| VARIABLE | count | unique | top | freq |
|---|---|---|---|---|
| ESTADO_CIVIL | 54774 | 5 | SOLTERO/A | 23820 |
| NIVEL_ESTUDIOS | 54774 | 7 | SECUNDARIA | 23324 |
| SECTOR_DOMICILIO | 54774 | 3 | URBANO | 35575 |
| TIPO_PERSONA | 54774 | 1 | PERSONA NATURAL | 54774 |
| NIVEL_RIESGO_PREC | 54774 | 9 | A1 | 21780 |
| REGION | 54774 | 4 | SIERRA | 34640 |

*Tabla 11: Análisis estadístico descriptivo con variables adicionales. Fuente: (Elaboración propia)*

Al ejecutar nuevamente el modelo k-means con las variables adicionales y aplicando PCA a los datos normalizados, obtenemos los siguientes resultados:

- Al visualizar las diferentes iteraciones de ejecución del modelo, podemos ver que incluso hay una mejoría significativa del score de silueta el cual incluso con 4 clusters (Figura 15.) mantiene un valor significativamente alto (superior al 85%), lo cual indica que estos 4 clusters tienen bien demarcados sus límites y la agrupación de cada segmento es también adecuada. El gráfico incluso muestra

que un agrupamiento de hasta 7 clusters, cumple con los valores del umbral mínimo de silhouette score que se ha establecido.

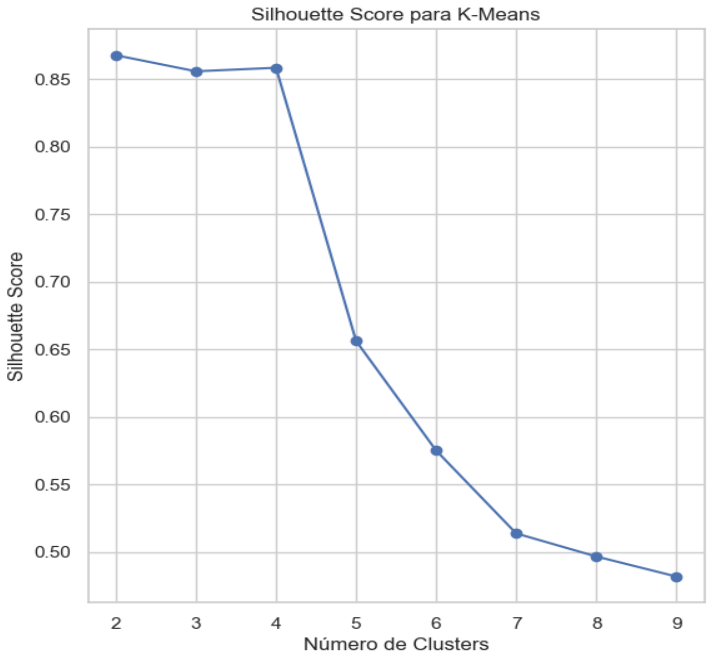



*Figura 15: Evaluación de Silouhette Score para modelo K-Means con ajustes de variables y aplicación de PCA. Fuente: (Elaboración propia)*

- A nivel de la gráfica de segmentación de observaciones en los diferentes clústeres, podemos ver que en efecto se identifican adecuadamente cada uno de los grupos (Figura 16), sin embargo, al analizar el detalle de frecuencia por clúster, vemos que los clústeres generados si bien es cierto, algunos de ellos están plenamente identificados, no contienen un número significativo de casos, mientras que al parecer la mayor cantidad de observaciones se encuentran sobrecargadas en un solo clúster (Tabla 12).
- A nivel de la gráfica de segmentación (Figura 16), igualmente pueden notarse ciertos sectores alrededor del cluster 0 (en color rojo), donde efectivamente se denota que existe una sobrecarga de las observaciones.

| CLÚSTER | FRECUENCIA |
|---|---|
| 0 | 53576 |
| 1 | 982 |
| 2 | 207 |
| 3 | 8 |

*Tabla 12: Tabla de frecuencias por cluster para modelo k-means con valor k=4. Fuente: (Elaboración propia)*

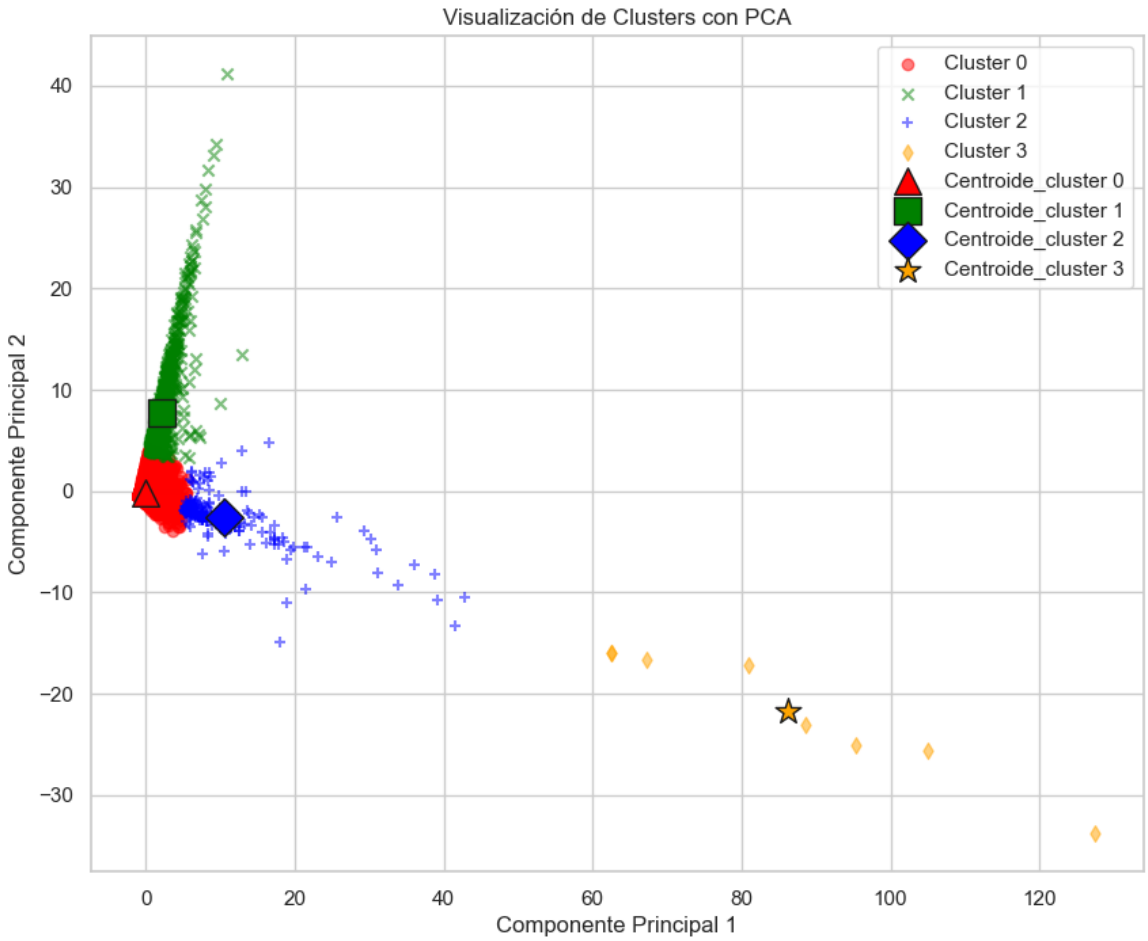



*Figura 16: Segmentación de observaciones usando k-means con k=4. Fuente: (Elaboración propia)*

Para mejorar esta situación del modelo procedemos a generar una segmentación inicial de 2 clústeres, identificando en cada uno de los grupos cuantas observaciones existen (Tabla 13), y cuáles son las características de cada segmento en función de la media (Tabla 14). La distribución de esta segmentación inicial puede verse en la figura 17.

| CLUSTER | FRECUENCIA |
|---|---|
| 0 | 53742 |
| 1 | 1031 |

*Tabla 13: Tabla de frecuencias por clúster para modelo k-means con valor k=2. Fuente: (Elaboración propia)*

| VARIABLES | CLUSTER | |
|---|---|---|
| | 0 | 1 |
| CARGAS_FAMILIARES | 1,023203454 | 1,264791465 |
| CREDITOS_TOTAL | 3,97655465 | 64,5014549 |
| CREDITOS_CANCELADOS | 2,669662461 | 58,34529583 |
| DIAS_MORA_PROMEDIO | 7,258866436 | 3,242483026 |
| MONTO_PROMEDIO | 9125,56868 | 2414,563967 |
| TOTAL_INGRESO | 2051,084297 | 9014,265538 |
| EGRESO_SOCIO | 1334,566992 | 6573,260941 |
| FLUJO_CAJA | 716,5173049 | 2441,004597 |
| EDAD | 40,51596517 | 39,18234724 |

*Tabla 14: Tabla caracterización de segmentos para modelo k-means con valor k=2. Fuente: (Elaboración propia)*

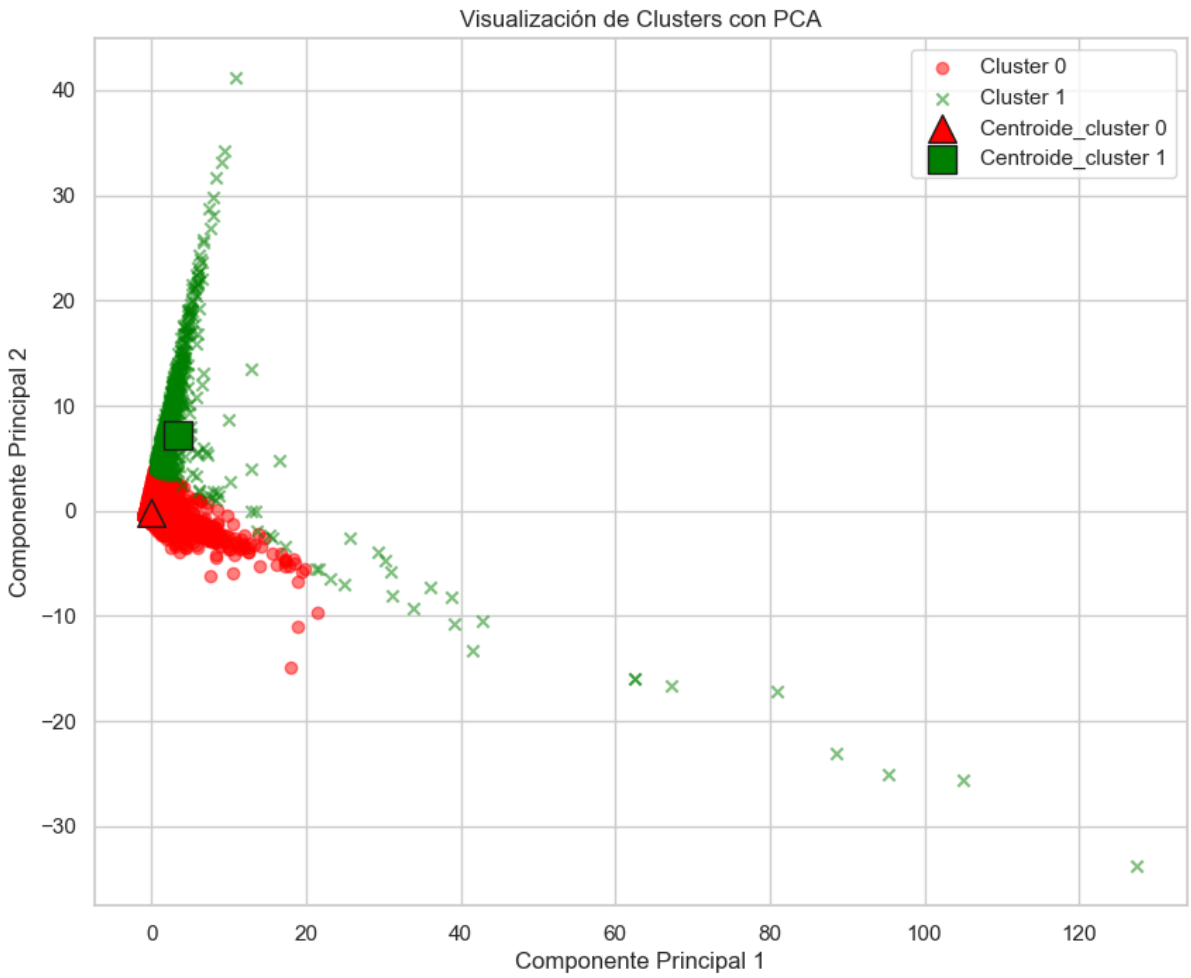



*Figura 17: Segmentación de observaciones usando k-means con k=2. Fuente: (Elaboración propia)*

Ahora generaremos la primera sub-segmentación donde trabajaremos como nuevo origen de datos a los datos del clúster con la mayor cantidad de observaciones para subdividirlo en sub-clusters según el valor más adecuado de 'k' y verificando que los nuevos subsegmentos se mantengan sobre el umbral mínimo del score de silueta. Partimos del cluster '0' (subsegmento C0) como base para ejecutar nuevamente el modelo de segmentación. Para eso veremos primero el grafico de score de silueta (Figura 18) para determinar el número de clústeres más adecuado que permita mantener el umbral mínimo definido.

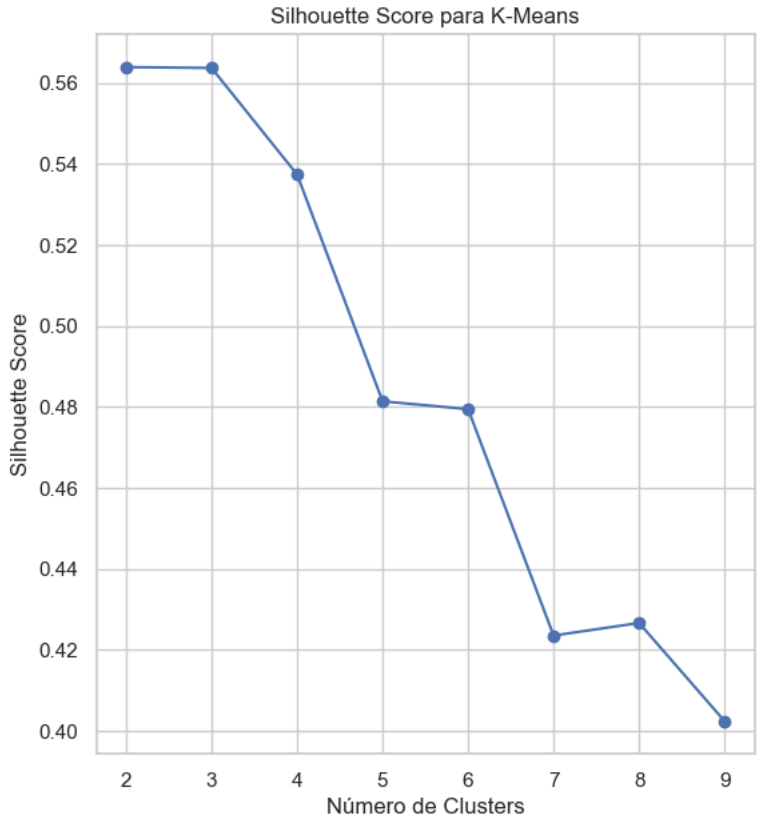



*Figura 18: Evaluación de Silouhette Score para modelo K-Means correspondiente al subsegmento C0. Fuente: (Elaboración propia)*

El gráfico muestra que el modelo aún se mantiene por encima del umbral mínimo acordado y nos permite establecer que en este caso podemos trabajar con 2 o máximo 3 clusters o subsegmentos. Para este caso consideraremos trabajar con 2 clusters o nuevos subsegmentos los cuales estarán descritos por su respectiva tabla de frecuencias (Tabla 15), su descripción del segmento (Tabla 16) y el respectivo grafico de segmentación de observaciones (Figura 19)

| CLUSTER | FRECUENCIA |
|---|---|
| 0 | 9354 |
| 1 | 44388 |

*Tabla 15: Tabla de frecuencias por clúster para modelo k-means correspondiente a Subsegmento C0 con valor k=2. Fuente: (Elaboración propia)*

| VARIABLES | Sub-Cluster | |
|---|---|---|
| | C0-S0 | C0-S1 |
| CARGAS_FAMILIARES | 1,21 | 0,98 |
| CREDITOS_TOTAL | 11,39 | 2,41 |
| CREDITOS_CANCELADOS | 9,21 | 1,29 |
| DIAS_MORA_PROMEDIO | 8,12 | 7,08 |
| MONTO_PROMEDIO | 9178,01 | 9114,52 |
| TOTAL_INGRESO | 4186,18 | 1601,15 |
| EGRESO_SOCIO | 2893,08 | 1006,14 |
| FLUJO_CAJA | 1293,10 | 595,01 |
| EDAD | 48,06 | 38,93 |

*Tabla 16: Tabla caracterización de segmentos para modelo k-means correspondiente al subsegmento C0 con valor k=2. Fuente: (Elaboración propia)*

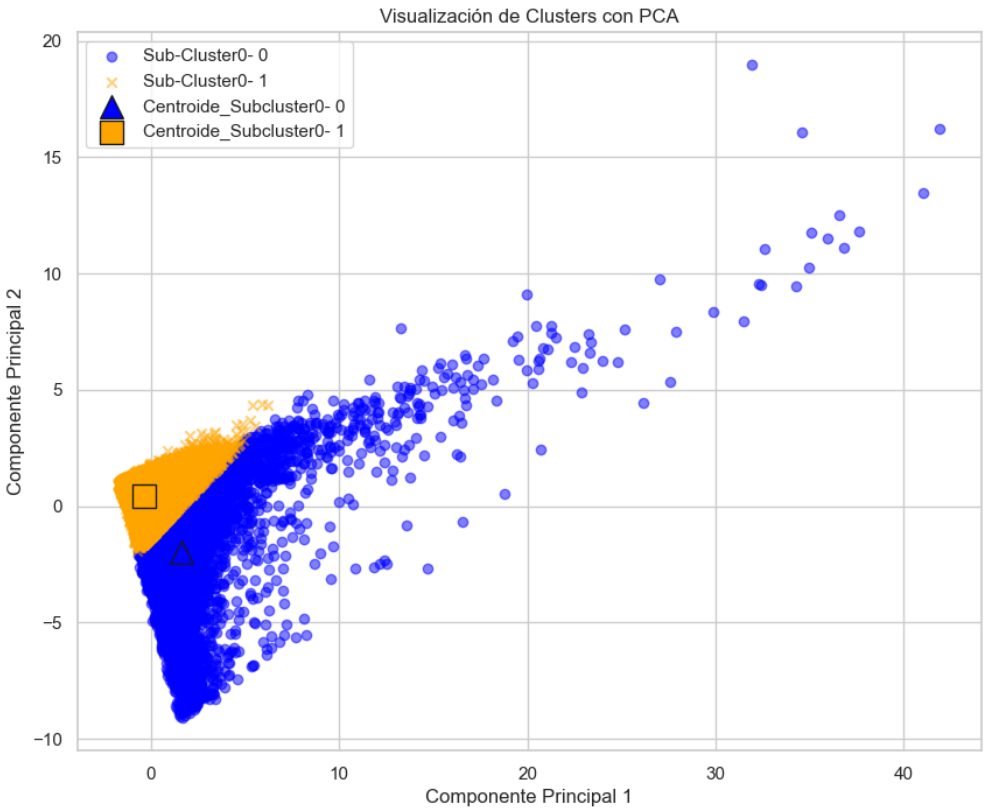



*Figura 19: Segmentación de observaciones usando k-means con k=2 en subsegmento C0. Fuente: (Elaboración propia)*

Como puede observarse, la segmentación realizada en el subsegmento C0 ha generado 2 nuevos clústeres donde podemos ver que la frecuencia entre los grupos empieza a ser un poco más equilibrada y se empiezan a descubrir nuevos grupos con determinadas características en función de las variables utilizadas.

Para finalizar este proceso de construcción de segmentos de socios, generaremos una nueva sub-segmentación, esta vez sobre el subsegmento C0-S1 dentro de los mismos lineamientos de buscar el número de clústeres adecuado que se mantenga por encima del umbral mínimo. A continuación, mostramos el grafico de score de silueta para determinar los valores adecuados para generar los nuevos sub-segmentos (Figura 19).

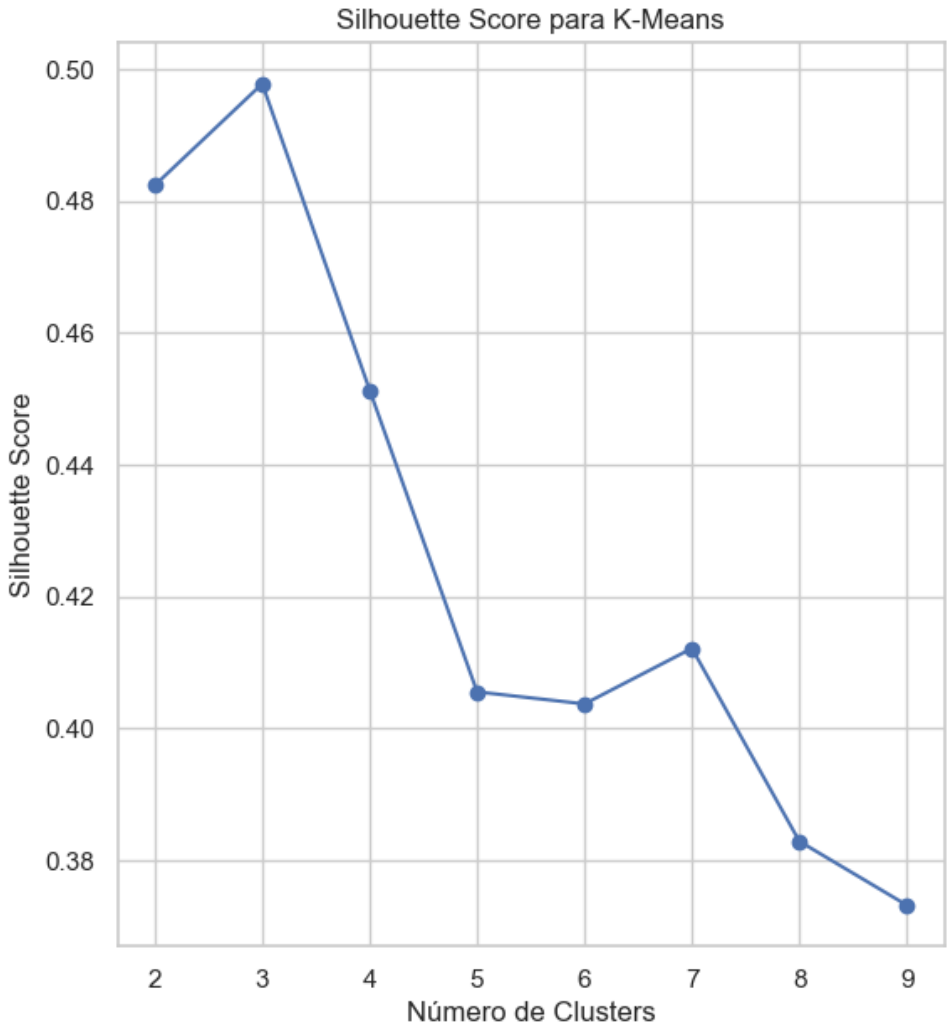



*Figura 20: Evaluación de Silouhette Score para modelo K-Means correspondiente al subsegmento C0-S1. Fuente: (Elaboración propia)*

Como puede apreciarse en el grafico del score, el numero apropiado de clústeres para el ultimo nivel de segmentación es de 3 clústeres, sin embargo, para este caso el score de silueta al parecer se encuentra por debajo del umbral mínimo, aunque la diferencia es bastante corta en relación al umbral procederemos a la generación de los subsegmentos según la idea inicial. Posteriormente se establecerá con la reunión de socialización si el negocio puede manejarse con 4 clústeres o con los 7 clúster resultantes de la última segmentación. A continuación, se presentan la tabla de frecuencias de la última generación de segmentos (Tabla 17), la tabla de características de los nuevos subsegmentos (Tabla 18), y el gráfico de segmentación de observaciones (Figura 21)

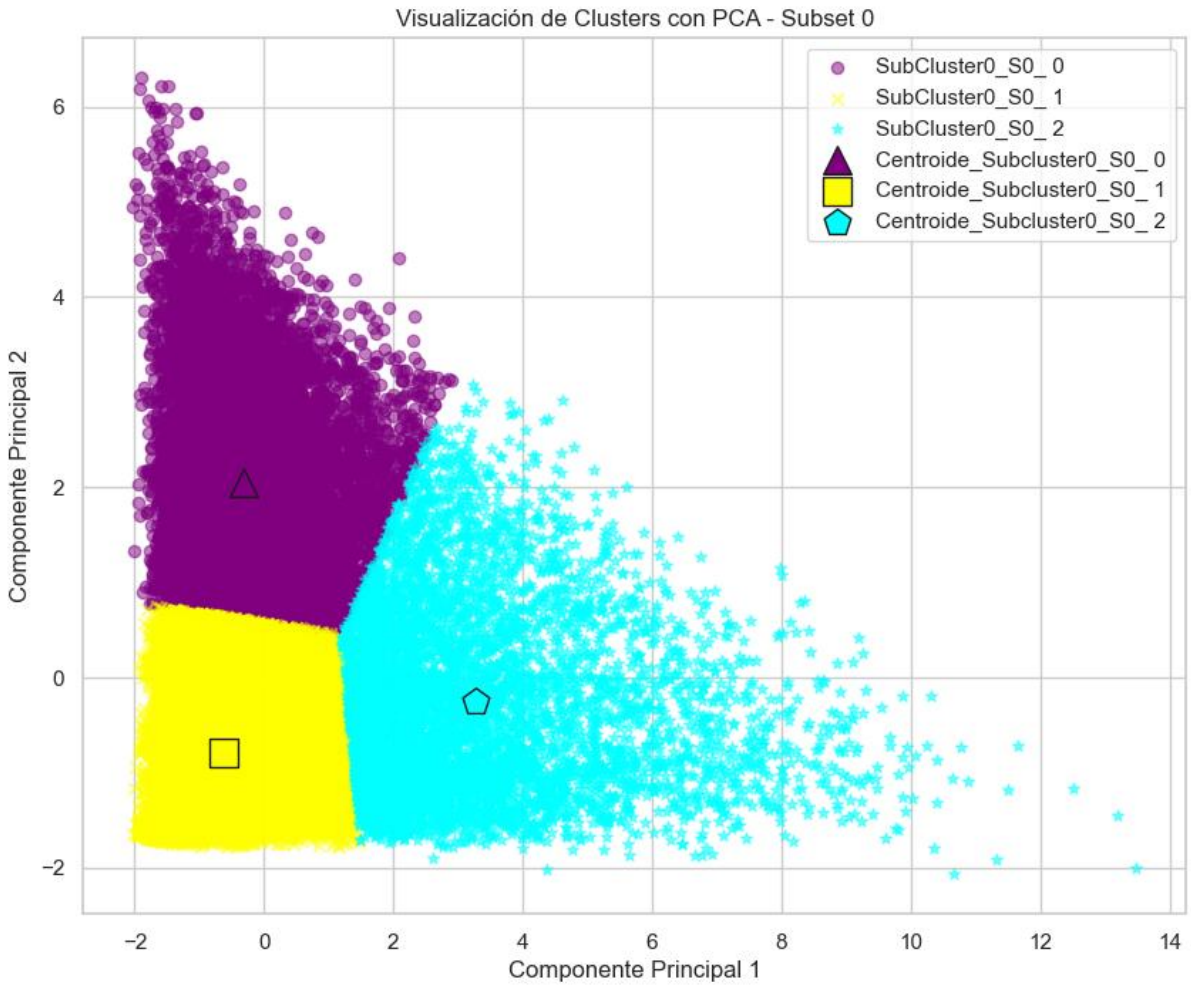



*Figura 21: Segmentación de observaciones usando k-means con k=3 en subsegmento C0-S1. Fuente: (Elaboración propia)*

| CLUSTER | FRECUENCIA |
|---|---|
| **0** | 11272 |
| **1** | 26927 |
| **2** | 6189 |

*Tabla 17: Tabla de frecuencias por clúster para modelo k-means correspondiente a Subsegmento C0-S1 con valor k=3. Fuente: (Elaboración propia)*

| VARIABLES | Sub-Cluster | | |
|---|---|---|---|
| | **C0-S1-s0** | **C0-S1-s1** | **C0-S1-s2** |
| **CARGAS_FAMILIARES** | 1,21 | 0,86 | 1,11 |
| **CREDITOS_TOTAL** | 4,69 | 1,54 | 2,08 |
| **CREDITOS_CANCELADOS** | 3,35 | 0,51 | 0,95 |
| **DIAS_MORA_PROMEDIO** | 9,34 | 6,06 | 7,39 |
| **MONTO_PROMEDIO** | 7172,02 | 7400,49 | 20109,74 |
| **TOTAL_INGRESO** | 1320,35 | 1036,69 | 4568,43 |
| **EGRESO_SOCIO** | 857,35 | 623,02 | 2943,97 |
| **FLUJO_CAJA** | 463,00 | 413,66 | 1624,46 |
| **EDAD** | 43,30 | 36,66 | 40,81 |

*Tabla 18: Tabla caracterización de segmentos para modelo k-means correspondiente al subsegmento C0-S1 con valor k=3. Fuente: (Elaboración propia)*

# CAPITULO V: PRESENTACION DE RESULTADOS

## 5. Presentación de resultados

### 5.1. Resultados de la ejecución del modelo

Los resultados obtenidos con el modelo K-Means muestran la existencia de diversos perfiles de socios, que van desde aquellos con una elevada capacidad de pago y un bajo riesgo de morosidad, hasta situaciones completamente contrarias. Las conclusiones derivadas del análisis indican que es posible identificar patrones donde ciertas combinaciones de ingresos, flujo de caja, morosidad, etc., que se asocian con diferentes grados de riesgo de morosidad, proporcionando así insights valiosos para la toma de decisiones en la gestión de riesgos y el desarrollo de estrategias de marketing personalizadas.

La Tabla 19, muestra en resumen cuales son los segmentos descubiertos por el modelo.

| VARIABLES | Clúster - Segmento | | | | | | |
|---|---|---|---|---|---|---|---|
| | **C0** | **C1** | **C0-S0** | **C0-S1** | **C0-S1-s0** | **C0-S1-s1** | **C0-S1-s2** |
| **CARGAS_FAMILIARES** | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **CREDITOS_TOTAL** | 4 | 65 | 11 | 2 | 5 | 2 | 2 |
| **CREDITOS_CANCELADOS** | 3 | 58 | 9 | 1 | 3 | 1 | 1 |
| **DIAS_MORA_PROMEDIO** | 7 | 3 | 8 | 7 | 9 | 6 | 7 |
| **MONTO_PROMEDIO** | 9126 | 2415 | 9178 | 9115 | 7172 | 7400 | 20110 |
| **TOTAL_INGRESO** | 2051 | 9014 | 4186 | 1601 | 1320 | 1037 | 4568 |
| **EGRESO_SOCIO** | 1335 | 6573 | 2893 | 1006 | 857 | 623 | 2944 |
| **FLUJO_CAJA** | 717 | 2441 | 1293 | 595 | 463 | 414 | 1624 |
| **EDAD** | 41 | 39 | 48 | 39 | 43 | 37 | 41 |

*Tabla 19: Tabla caracterización de segmentos de socios de la cooperativa Jardín Azuayo. Fuente: (Elaboración propia)*

En la socialización de estos resultados a los Departamentos de Análisis de Datos y la Dirección de Riesgos (Anexo 1), las áreas involucradas determinaron que aceptan los resultados mostrados por el estudio, tomando como base los 7 segmentos de socios identificados mostrados en la figura 22 y detallados en el Anexo II.

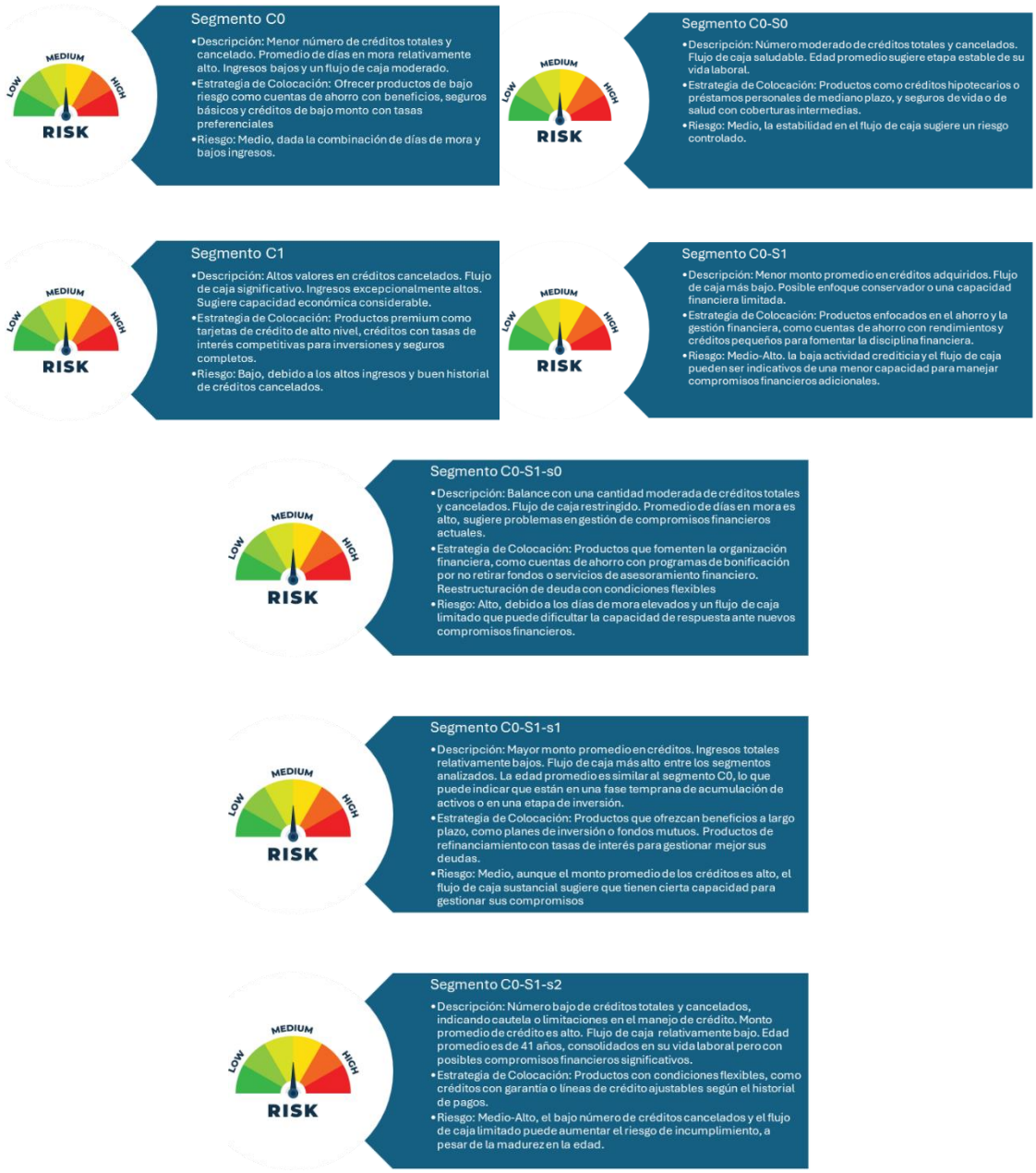



*Figura 22: Perfilamiento de socios COAC Jardín Azuayo. Fuente: (Elaboración propia)*

El resultado del modelo de clustering y los diferentes el análisis de cada segmento obtenido dan como resultado el gráfico de la figura 22. De manera detallada, existen 7 perfiles de socios que, dentro del análisis de negocio, permiten diseñar y promocionar productos y servicios que se apeguen al perfil del socio, haciendo más eficiente este proceso y reduciendo el riesgo de otorgar productos o servicios que probablemente generen una obligación financiera para personas más aptas para otro producto completamente diferente.

## 5.2. Evaluación de resultados

La aplicación del modelo de segmentación basado en aprendizaje automático no supervisado ha facilitado la identificación de grupos con riesgos altos, medios o bajos de morosidad o incumplimiento de obligaciones financieras, lo cual es fundamental para el desarrollo de estrategias preventivas orientadas a mitigar este riesgo. Si bien los resultados obtenidos son prometedores, la efectividad real en la reducción de la morosidad estará condicionada por la implementación adecuada de medidas proactivas derivadas de la segmentación realizada.

Adicionalmente, de acuerdo al criterio emitido por las áreas de negocio interesadas, esta segmentación podrá mejorar el proceso de asignación de créditos mediante la identificación precisa de perfiles de riesgo de forma anticipada, aunque la pertinencia del modelo para alcanzar este fin deberá ser evaluada en condiciones reales de mercado.

En cuanto a la satisfacción del cliente, las áreas de negocio han indicado que el modelo construido y la propuesta de perfilamiento realizada (Ver Anexo II), abre la puerta a la oferta de servicios más personalizados que podrían, en teoría, incrementar la satisfacción del cliente y mejorar los ingresos y utilidades para la cooperativa.

Al revisar el éxito de este proyecto, se ha logrado un aporte significativo para lograr los objetivos de negocio planteados. Si bien es cierto, la validación definitiva dependerá de la aplicación práctica de las estrategias de segmentación y su posterior evaluación en el ámbito empresarial, es el inicio para el uso y adopción de una herramienta que determinará su impacto real en la operación diaria y la estrategia de la cooperativa.

Por lo pronto, desde las áreas de negocio interesadas y que han dado seguimiento a este proyecto, han indicado su satisfacción con los resultados presentados (Ver Anexo I).

## 5.3. Revisión del proceso

En cuanto a la revisión del proceso llevado a cabo, se reconoce la efectividad de las herramientas de implementación como Python, la guía de la metodología CRISP-DM, y la disponibilidad de la información procesada de las distintas fuentes de datos de la cooperativa, ya que ello facilito el proceso de obtención e integración de las diversas fuentes de datos que entregan las variables a analizar. Hay oportunidades de mejora,

especialmente en lo que respecta a la automatización del proceso de limpieza de datos y la evaluación de los modelos.

Fue igualmente importante mantener una comunicación fluida y constante con los interesados del negocio para garantizar que los resultados obtenidos a través del modelado estén en concordancia con las necesidades empresariales, las cuales pueden variar a lo largo del tiempo.

### 5.4. Acciones por seguir

Dentro de las iniciativas propuestas para capitalizar los hallazgos de este proyecto, se incluyen entre otras las siguientes:

- La implementación de estrategias de segmentación diseñadas para mitigar el riesgo de morosidad y optimizar el proceso de asignación de créditos: esta acción implica el desarrollo y puesta en práctica de medidas específicas que se apoyen en la segmentación detallada de los socios.
- Establecimiento de un mecanismo de monitoreo continuo y ajuste de los segmentos identificados: esto con el objetivo de adaptar estas categorizaciones a las variaciones en el comportamiento y las necesidades de los socios a lo largo del tiempo.
- Exploración de nuevas técnicas de modelado y la incorporación de variables adicionales: con ello se podrá contribuir a afinar y enriquecer la segmentación realizada, razón por la cual en la fase de despliegue.
- Construir un servicio de integración que permita consumir los datos del modelo para evaluar personas y devolver su perfil en línea

### 5.5. Despliegue

El despliegue del modelo corresponde a la entrega de la notebook donde se construyó el modelo al Departamento de Análisis de Datos para que forme parte de sus herramientas de operación diaria, tanto el código Python, como la herramienta conocida como Jupyter Notebook y el servidor Anaconda.

Con estas acciones se materializa la incorporación de las prácticas de segmentación en los procesos de análisis de información de socios para generar actividades de análisis de segmentos de socios y definir donde puede o no ser util un nuevo servicio o producto financiero. Esto requerirá tanto la capacitación del personal del Departamento de Análisis

de Datos en el uso de la herramienta Jupyter Notebook y el uso de lenguaje Python para análisis de datos en estos procesos, como la entrega el resultado de los perfiles obtenidos para que formen parte de los procesos de análisis de nuevos productos y servicios realizados por el personal del área de innovación y diseño de servicios, socializando el proceso realizado para obtener los perfiles de socios y como podrían integrarlos en sus procesos.

En cuanto al plan de monitoreo y mantenimiento, el mismo estará a cargo del Departamento de Análisis de Datos, donde se ha previsto un esquema de seguimiento continuo que permitirá evaluar la eficacia de las estrategias de segmentación, especialmente en términos de su impacto en la obtención de grupos de socios para promocionar productos o créditos preaprobados.

# CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES

## 6. Conclusiones y recomendaciones

### 6.1. Conclusiones

- El modelo K-Means ha demostrado ser una herramienta sumamente eficaz en la identificación de diversos perfiles de socios dentro de la cooperativa, abarcando desde aquellos con elevada capacidad de pago y reducido riesgo de morosidad hasta aquellos con perfiles opuestos. Esta segmentación minuciosa resulta fundamental para entender a profundidad las necesidades y comportamientos de los socios, posibilitando una personalización y ajuste más preciso de los servicios financieros ofrecidos. La variedad en los perfiles detectados enfatiza la complejidad inherente al comportamiento financiero de los socios y resalta la importancia de adoptar estrategias diferenciadas para gestionar eficientemente la relación entre los clientes y la entidad financiera.
- La implementación del modelo ha desvelado patrones de comportamiento financiero complejos, proporcionando insights valiosos que tienen el potencial de cambiar sustancialmente tanto la gestión de riesgos como el desarrollo de estrategias de marketing. La habilidad para obtener estos patrones y asociarlos con diversos niveles de riesgo no solo optimiza la toma de decisiones, sino que también personaliza la oferta de productos, un aspecto crítico para una Jardín Azuayo, que aspira a mantener su relevancia y competitividad en el mercado actual.
- Una segmentación precisa de los socios es clave para mejorar el proceso de asignación de créditos y potencialmente elevar la satisfacción del cliente. Al reconocer de antemano los perfiles de riesgo, la cooperativa puede proporcionar servicios más adecuados y personalizados, incrementando no solo la eficiencia operativa sino también fomentando la fidelidad de los socios. Esta orientación hacia el cliente (socio) puede traducirse en un incremento de los ingresos y la rentabilidad para la cooperativa al alinear de manera más efectiva sus ofertas con las necesidades y posibilidades de sus socios.
- La verdadera prueba de la efectividad del modelo de segmentación en la reducción de la morosidad y la mejora en la asignación de créditos dependerá de su aplicación práctica en el entorno real del mercado. Se requiere un compromiso

firme por parte de la cooperativa hacia una validación continua y meticulosa del modelo bajo condiciones de mercado auténticas, para confirmar que las estrategias de segmentación generen un impacto positivo y concreto en la operativa diaria y en la estrategia global de la entidad.

- Como pudo verse en los gráficos de silhouette score, el número de clúster optimo en una sola corrida el modelo puede no se suficiente ya que hemos visto que para este caso los datos indicaban un numero de clúster bastante reducido con un buen porcentaje de eficiencia, pero a medida que aumentaba el número de clústeres, el porcentaje de eficiencia del silhouette score disminuía. Esto puede explicarse debido a la similaridad de los socios de la cooperativa en cuanto a su información, razón por la cual no existían muchas variables diferenciadoras que pudieran mejorar el score. Esto hizo que se opte por usar variables categóricas, en este caso la variable 'NIVEL_RIESGO_PREC' que de alguna manera permitía que se diferencien mejor los registros al momento de ejecutar el modelo, lo que permitió que el silhouette score mejore sustancialmente en la primera ejecución
- Se evidencia también en el caso de la información de la cooperativa, ninguno de los algoritmos propuestos, DBSCAN y K-means, pudieron establecer de primera instancia una segmentación adecuada, por la razón mencionada en la conclusión anterior. Es por esta razón que se ejecuta una ejecución del modelo por niveles de profundidad tomando los clústeres donde se veía que se agrupaba la mayor cantidad de observaciones. Ello permitió que se puedan diferenciar de mejor manera los diferentes segmentos obteniendo los 7 grupos identificados al momento.
- Durante la presentación de los resultados del modelo se vió que, a nivel de negocio, dado que el score de silueta era adecuado, pero disminuyó en el tercer nivel de profundidad, la decisión de adopción del modelo fue su utilidad en la segmentación más que en establecimiento de un scoring para uso directo en la parte de crédito o nivel de riesgo. Esto porque el proceso de crédito de la cooperativa tiene múltiples variables incluso de carácter subjetivo que influyen en la decisión de crédito. Por eso, la cooperativa utilizará los resultados del modelo exclusivo para la segmentación de clientes creando perfiles que permitan promocionar adecuadamente los productos y servicios según cada perfil. Posteriormente, y como una mejora, se explotarán los resultados del modelo para refinar sus resultados y poder establecer un posible score de crédito basado en la

segmentación inicial, y luego integrado con modelos de predicción basados en el perfil de cada socio, como regresión logística, o redes neuronales tanto para personas naturales, jurídicas, o socios nuevos sin historial en la cooperativa.

- Es importante recalcar que la calidad de la información es crucial para proyectos de machine learning ya que garantiza la precisión, la generalización y la confianza en los modelos desarrollados. Datos de baja calidad pueden llevar a predicciones erróneas, reducir la capacidad del modelo para aplicarse a nuevos datos, y consumir recursos valiosos en limpieza y preprocesamiento. Además, tiene implicaciones éticas, sociales y legales significativas, afectando la equidad de los resultados y el cumplimiento normativo. Por lo tanto, asegurar datos de alta calidad es fundamental para el éxito y la aplicabilidad responsable de los modelos de machine learning. En este caso la información de socios personas naturales con ingresos sumamente altos general una dispersión de la información, y puntualmente, un dato en particular salido de la escala generaba una distorsión en la segmentación, por lo cual se decidió tratarlo como outlier a pesar de que con la parte de negocio se valida que es un dato real.

### 6.2. Recomendaciones

- Considero que es mandatorio no solo el desarrollo sino también la implementación y monitoreo constante de estrategias de segmentación que busquen mitigar el riesgo de morosidad. Estas estrategias deben fundamentarse en datos actualizados, íntegros, y lo suficientemente adaptables a los cambios en el comportamiento financiero de los socios. La cooperativa debe asegurar que estas estrategias sean específicas, medibles y modificables, lo cual exige una colaboración estrecha entre los departamentos de riesgo, marketing y operaciones.
- Para conservar la pertinencia de la segmentación, es esencial establecer un protocolo para el monitoreo y ajuste regular de los segmentos identificados. Esto implica evaluar la efectividad de las estrategias de segmentación y la capacidad de adaptación del modelo ante cambios económicos y en el comportamiento de los socios. Actualizaciones y ajustes periódicos son cruciales para preservar la exactitud del modelo y su utilidad a largo plazo. De esto deberá encargarse el Departamento de Análisis de Datos como actor fundamental en la gestión y creación de modelos de aprendizaje basados en datos, asi como en la administración de los sistemas de información institucionales

- Finalmente, la cooperativa debe continuar explorando y adoptando nuevas técnicas de modelado, así como considerar la integración de variables adicionales que puedan enriquecer aún más el modelo actual. En un campo tan dinámico como la analítica de datos, la incorporación de innovaciones y las mejores prácticas disponibles pueden ofrecer ventajas competitivas significativas y profundizar el entendimiento sobre la clientela.

# CAPITULO VII: REFERENCIAS

## 7. REFERENCIAS


Alpaydin, E. (2014). Introduction to Machine Learning (3rd ed.). MIT Press.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2013). A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. Data Mining and Knowledge Discovery, 27(3), 344–371. https://doi.org/10.1007/s10618-013-0311-4

Elizondo, A., & Altman, E. I. (2004). Medición integral del riesgo de crédito. Editorial Limusa.

Equipo Verne Academy. (2022, September 12). 10 Librerías Python para Data Science y Machine Learning. Verne academy. https://verneacademy.com/blog/articulos-ia/10-librerias-python-data-science-machine-learning/

Fernández, C., & Aqueveque, C. (2001). Segmentación de mercados: buscando la correlación entre variables sicológicas y demográficas. Revista Colombiana de Marketing, 2(2), 15. https://www.redalyc.org/pdf/109/10900204.pdf

Foster, & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

González-Duany, A. (2021). Metodología para la evaluación del riesgo de liquidez en el Banco de Crédito y Comercio. Revista Estudios Del Desarrollo Social: Cuba y América Latina, 9(1). http://scielo.sld.cu/scielo.php?pid=S2308-01322021000100016&script=sci_arttext&tlng=en

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). DBSCAN: Fast density-based clustering with R. Journal of Statistical Software, 91(1). https://doi.org/10.18637/jss.v091.i01

Heras Calvo, D. (2023). Biblioteca para la evaluación sistemática de algoritmos de clustering [ETSI_Informatica]. https://oa.upm.es/75555/

Linkedin. (2023, November 2). How can you calculate the silhouette score for a clustering algorithm? Linkedin.com; www.linkedin.com. https://www.linkedin.com/advice/0/how-can-you-calculate-silhouette-score-clustering-algorithm-w9bcc?

Rodríguez-Guevara, D. E., Becerra-Arévalo, J. A., & Cardona-Valencia, D. (2017). Modelos y metodologías de credit score para personas naturales: una revisión literaria. Rev. CEA, 3(5), 13–28. https://doi.org/10.22430/24223182.645

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, 28, 586–599. https://www.proquest.com/scholarly-journals/machine-learning-análisis-de-lenguajes/docview/2388304894/se-2

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). https://doi.org/10.1109/dsaa49011.2020.00096

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Yadav, S. (2023, June 14). Silhouette coefficient explained with a practical example: Assessing cluster fit". Medium. https://medium.com/@Suraj_Yadav/silhouette-coefficient-explained-with-a-practical-example-assessing-cluster-fit-c0bb3fdef719

# ANEXOS

## ANEXO I – Acta de socialización y aceptación de resultados

JARDÍN AZUAYO | COOPERATIVA DE AHORRO Y CRÉDITO JARDÍN AZUAYO
CONTROLADA POR LA SUPERINTENDENCIA DE ECONOMÍA POPULAR Y SOLIDARIA

Cuenca, 22 de febrero de 2024

**Para:** Srs. Pontificia Universidad Católica del Ecuador (PUCE)
**Asunto:** Socialización y presentación del proyecto de tesis "IMPLEMENTACIÓN DE UN MODELO DE SEGMENTACIÓN DE SOCIOS PARA LA IDENTIFICACION DE CAPACIDADES DE PAGO Y RIESGO DE MOROSIDAD DE LOS SOCIOS DE LA COOPERATIVA DE AHORRO Y CREDITO JARDIN AZUAYO"

De nuestras consideraciones:

Junto con un atento y afectuoso saludo, por el medio del presente, damos a conocer a ustedes que el miércoles 21 de febrero de 2024, el Ing. JUAN GABRIEL GOYO CANDO socializó los resultados de su proyecto de tesis. Luego de dicha presentación, quienes participamos como interesados en el resultado del estudio, manifestamos en este documento **nuestra conformidad con los resultados** presentados en base a los requerimientos de la institución, y establecemos que dicho modelo presentado será de utilidad para establecer una **segmentación y perfilamiento de nuestros socios para la promoción de productos y servicios financieros** ajustados a las características de clasificación mostradas en este proyecto de tesis.

**Atentamente,**

**Jhon Machuca**
**Director de Riesgos**
**Cooperativa Jardín Azuayo**

**Adrián Limaico**
**Responsable de Análisis de Datos**
**Cooperativa Jardín Azuayo**

JARDÍN AZUAYO
COOPERATIVA DE AHORRO Y CRÉDITO
COORDINACIÓN

www.jardinazuayo.fin.ec
Dir.: Benigno Malo 9-75 entre Gran Colombia y Simón Bolivar | Teléfono PBX: 07 2 833 255 / Cuenca - Ecuador

## ANEXO II – Perfilamiento de socios basado en Aprendizaje Automático no Supervisado para Cooperativa de Ahorro y Crédito Jardín Azuayo


### Segmento C0

- Descripción: Menor número de créditos totales y cancelado. Promedio de días en mora relativamente alto. Ingresos bajos y un flujo de caja moderado.
- Estrategia de Colocación: Ofrecer productos de bajo riesgo como cuentas de ahorro con beneficios, seguros básicos y créditos de bajo monto con tasas preferenciales
- Riesgo: Medio, dada la combinación de días de mora y bajos ingresos.


### Segmento C1

- Descripción: Altos valores en créditos cancelados. Flujo de caja significativo. Ingresos excepcionalmente altos. Sugiere capacidad económica considerable.
- Estrategia de Colocación: Productos premium como tarjetas de crédito de alto nivel, créditos con tasas de interés competitivas para inversiones y seguros completos.
- Riesgo: Bajo, debido a los altos ingresos y buen historial de créditos cancelados.


### Segmento C0-S0

- Descripción: Número moderado de créditos totales y cancelados. Flujo de caja saludable. Edad promedio sugiere etapa estable de su vida laboral.
- Estrategia de Colocación: Productos como créditos hipotecarios o préstamos personales de mediano plazo, y seguros de vida o de salud con coberturas intermedias.
- Riesgo: Medio, la estabilidad en el flujo de caja sugiere un riesgo controlado.


### Segmento C0-S1

- Descripción: Menor monto promedio en créditos adquiridos. Flujo de caja más bajo. Posible enfoque conservador o una capacidad financiera limitada.
- Estrategia de Colocación: Productos enfocados en el ahorro y la gestión financiera, como cuentas de ahorro con rendimientos y créditos pequeños para fomentar la disciplina financiera.
- Riesgo: Medio-Alto. la baja actividad crediticia y el flujo de caja pueden ser indicativos de una menor capacidad para manejar compromisos financieros adicionales.


### Segmento C0-S1-s0

- Descripción: Balance con una cantidad moderada de créditos totales y cancelados. Flujo de caja restringido. Promedio de días en mora es alto, sugiere problemas en gestión de compromisos financieros actuales.
- Estrategia de Colocación: Productos que fomenten la organización financiera, como cuentas de ahorro con programas de bonificación por no retirar fondos o servicios de asesoramiento financiero. Reestructuración de deuda con condiciones flexibles
- Riesgo: Alto, debido a los días de mora elevados y un flujo de caja limitado que puede dificultar la capacidad de respuesta ante nuevos compromisos financieros.


### Segmento C0-S1-s1

- Descripción: Mayor monto promedio en créditos. Ingresos totales relativamente bajos. Flujo de caja más alto entre los segmentos analizados. La edad promedio es similar al segmento C0, lo que puede indicar que están en una fase temprana de acumulación de activos o en una etapa de inversión.
- Estrategia de Colocación: Productos que ofrezcan beneficios a largo plazo, como planes de inversión o fondos mutuos. Productos de refinanciamiento con tasas de interés para gestionar mejor sus deudas.
- Riesgo: Medio, aunque el monto promedio de los créditos es alto, el flujo de caja sustancial sugiere que tienen cierta capacidad para gestionar sus compromisos


### Segmento C0-S1-s2

- Descripción: Número bajo de créditos totales y cancelados, indicando cautela o limitaciones en el manejo de crédito. Monto promedio de crédito es alto. Flujo de caja relativamente bajo. Edad promedio es de 41 años, consolidados en su vida laboral pero con posibles compromisos financieros significativos.
- Estrategia de Colocación: Productos con condiciones flexibles, como créditos con garantía o líneas de crédito ajustables según el historial de pagos.
- Riesgo: Medio-Alto, el bajo número de créditos cancelados y el flujo de caja limitado puede aumentar el riesgo de incumplimiento, a pesar de la madurez en la edad.

## ANEXO III – Código fuente de construcción del modelo

# ScoreJA

February 29, 2024

### 0.1 MODELO DE APRENDIZAJE AUTOMATICO NO SUPERVISADO PARA SEGMENTACIÓN DE CLIENTES

#### 0.1.1 AUTOR: JUAN GABRIEL GOYO CANDO

**TRABAJO DE TITULACION - IMPLEMENTACIÓN DE UN MODELO DE SEGMENTACIÓN DE SOCIOS PARA LA IDENTIFICACION DE CAPACIDADES DE PAGO Y RIESGO DE MOROSIDAD DE LOS SOCIOS DE LA COOPERATIVA JA**

[1]:
```
# Cargar librerías necesarias
import pandas as pd
import numpy as np

from pandas import read_csv
from pandas.plotting import scatter_matrix
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from pyod.models.knn import KNN   # K Nearest Neighbors detector

import seaborn as sns
# Librerías para gráficos específicos y manejo de os
import plotly as py
import plotly.graph_objs as go
from sklearn.cluster import KMeans
import warnings
import os
print('Librerías cargadas.....')
```

```
Librerías cargadas…
```

[137]:
```
# Cargar dataset
#clientes_data = pd.read_csv(data_dir)
clientes_data = pd.read_csv('DataSetTesisv5.csv',delimiter=';',na_values='#N/D'␣
 ↪)
# Ver valores superiores
print("Muestra de los datos:")
clientes_data.head(10)
```

Muestra de los datos:

[137]:

| | CODIGO_SOCIO | ESTADO_CIVIL | PROVINCIA_RESIDENCIA | CARGAS_FAMILIARES |
|---|---|---|---|---|
| 0 | 12467 | CASADO/A | AZUAY | 0 |
| 1 | 13411 | CASADO/A | AZUAY | 1 |
| 2 | 12838 | DIVORCIADO/A | AZUAY | 1 |
| 3 | 12142 | CASADO/A | AZUAY | 0 |
| 4 | 12191 | SOLTERO/A | AZUAY | 0 |
| 5 | 13739 | SOLTERO/A | ILLINOIS | 0 |
| 6 | 13462 | CASADO/A | AZUAY | 1 |
| 7 | 15674 | SOLTERO/A | MORONA SANTIAGO | 1 |
| 8 | 14993 | CASADO/A | AZUAY | 0 |
| 9 | 12581 | CASADO/A | AZUAY | 0 |

| | CREDITOS_TOTAL | CREDITOS_CANCELADOS | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO |
|---|---|---|---|---|
| 0 | 4 | 3 | 10 | 3500 |
| 1 | 14 | 13 | 8 | 3000 |
| 2 | 4 | 3 | 0 | 10000 |
| 3 | 17 | 14 | 60 | 472,65 |
| 4 | 15 | 14 | 10 | 6000 |
| 5 | 71 | 64 | 1 | 1279,05 |
| 6 | 1 | 0 | 0 | 10000 |
| 7 | 10 | 8 | 9 | 1500 |
| 8 | 10 | 9 | 1 | 3000 |
| 9 | 3 | 3 | 5 | 1000 |

| | TOTAL_INGRESO | EGRESO_SOCIO | DIFERENCIA | OFICINA | NIVEL_ESTUDIOS |
|---|---|---|---|---|---|
| 0 | 2746,55 | 2505,23 | 241,32 | 1.0 | SECUNDARIA |
| 1 | 1125 | 708 | 417 | 20.0 | PRIMARIA |
| 2 | 986 | 598,59 | 387,41 | 1.0 | UNIVERSITARIA |
| 3 | 2050 | 1580 | 470 | 2.0 | UNIVERSITARIA |
| 4 | 500 | 207 | 293 | 1.0 | SECUNDARIA |
| 5 | 1701,7 | 1027 | 674,7 | 19.0 | UNIVERSITARIA |
| 6 | 1798,83 | 1003,15 | 795,68 | 19.0 | SECUNDARIA |
| 7 | 3674 | 3510,15 | 163,85 | 13.0 | UNIVERSITARIA |
| 8 | 1000 | 567 | 433 | 20.0 | PRIMARIA |
| 9 | 800 | 540 | 260 | 1.0 | PRIMARIA |

| | SECTOR_DOMICILIO | EDAD | TIPO_PERSONA | NIVEL_RIESGO_PREC |
|---|---|---|---|---|
| 0 | URBANO | 65.0 | PERSONA NATURAL | M1 |
| 1 | RURAL | 53.0 | PERSONA NATURAL | B1 |
| 2 | RURAL | 41.0 | PERSONA NATURAL | M1 |
| 3 | URBANO | 60.0 | PERSONA NATURAL | B1 |
| 4 | URBANO | 55.0 | PERSONA NATURAL | B1 |
| 5 | URBANO | 43.0 | PERSONA NATURAL | B2 |
| 6 | URBANO | 43.0 | PERSONA NATURAL | A1 |
| 7 | URBANO | 39.0 | PERSONA NATURAL | B1 |

```
8            RURAL  63.0  PERSONA NATURAL              B1
9           URBANO  58.0  PERSONA NATURAL              M1
```

[138]:
```
# Dimensión del dataset
print("Dimensión: ")
clientes_data.shape
```

```
Dimensión:
```

[138]:
```
(55687, 17)
```

[139]:
```
# Breve descripción estadística
print("EDA: ")
clientes_data.describe()
```

```
EDA:
```

[139]:

| | CODIGO_SOCIO | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | \ |
|---|---|---|---|---|---|
| count | 55687.000000 | 55687.000000 | 55687.000000 | 55687.000000 | |
| mean | 551740.614973 | 1.026416 | 5.085568 | 3.690610 | |
| std | 265153.176746 | 1.114226 | 10.655189 | 9.861535 | |
| min | 10015.000000 | 0.000000 | 1.000000 | 0.000000 | |
| 25% | 330964.500000 | 0.000000 | 1.000000 | 0.000000 | |
| 50% | 569483.000000 | 1.000000 | 2.000000 | 1.000000 | |
| 75% | 803721.000000 | 2.000000 | 5.000000 | 4.000000 | |
| max | 909435.000000 | 12.000000 | 338.000000 | 305.000000 | |

| | DIAS_MORA_PROMEDIO | OFICINA | EDAD |
|---|---|---|---|
| count | 55687.000000 | 55278.000000 | 55278.000000 |
| mean | 7.130623 | 28.613535 | 40.371359 |
| std | 22.878394 | 20.389692 | 13.823352 |
| min | 0.000000 | 1.000000 | 0.000000 |
| 25% | 0.000000 | 10.000000 | 30.000000 |
| 50% | 2.000000 | 26.000000 | 38.000000 |
| 75% | 7.000000 | 49.000000 | 50.000000 |
| max | 1871.000000 | 70.000000 | 94.000000 |

[140]:
```
import pandas as pd

# Suponiendo que 'df' es tu DataFrame
# df = pd.read_csv('tu_archivo.csv')

# Usar el método describe() para obtener el resumen estadístico del DataFrame
descripcion = clientes_data.describe()

# Escribir el DataFrame resultante a un archivo Excel
```

```
descripcion.to_excel('/Users/produccion-ja/Documents/MAESTRIA_PUCE/TITULACION/
 ↪resultado.xlsx', engine='openpyxl')
descripcion
```

[140]:

| | CODIGO_SOCIO | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | \ |
|---|---|---|---|---|---|
| count | 55687.000000 | 55687.000000 | 55687.000000 | 55687.000000 | |
| mean | 551740.614973 | 1.026416 | 5.085568 | 3.690610 | |
| std | 265153.176746 | 1.114226 | 10.655189 | 9.861535 | |
| min | 10015.000000 | 0.000000 | 1.000000 | 0.000000 | |
| 25% | 330964.500000 | 0.000000 | 1.000000 | 0.000000 | |
| 50% | 569483.000000 | 1.000000 | 2.000000 | 1.000000 | |
| 75% | 803721.000000 | 2.000000 | 5.000000 | 4.000000 | |
| max | 909435.000000 | 12.000000 | 338.000000 | 305.000000 | |

| | DIAS_MORA_PROMEDIO | OFICINA | EDAD |
|---|---|---|---|
| count | 55687.000000 | 55278.000000 | 55278.000000 |
| mean | 7.130623 | 28.613535 | 40.371359 |
| std | 22.878394 | 20.389692 | 13.823352 |
| min | 0.000000 | 1.000000 | 0.000000 |
| 25% | 0.000000 | 10.000000 | 30.000000 |
| 50% | 2.000000 | 26.000000 | 38.000000 |
| 75% | 7.000000 | 49.000000 | 50.000000 |
| max | 1871.000000 | 70.000000 | 94.000000 |

[141]:
```
# Tipos de datos
print("Tipos de datos: ")
clientes_data.dtypes
```

```
Tipos de datos:
```

[141]:

| | |
|---|---|
| CODIGO_SOCIO | int64 |
| ESTADO_CIVIL | object |
| PROVINCIA_RESIDENCIA | object |
| CARGAS_FAMILIARES | int64 |
| CREDITOS_TOTAL | int64 |
| CREDITOS_CANCELADOS | int64 |
| DIAS_MORA_PROMEDIO | int64 |
| MONTO_PROMEDIO | object |
| TOTAL_INGRESO | object |
| EGRESO_SOCIO | object |
| DIFERENCIA | object |
| OFICINA | float64 |
| NIVEL_ESTUDIOS | object |
| SECTOR_DOMICILIO | object |
| EDAD | float64 |
| TIPO_PERSONA | object |
| NIVEL_RIESGO_PREC | object |

```
dtype: object
```

[142]:
```
# PASAR INFORMACION DE TABLAS JUPYTER A EXCEL

# Usar el método describe() para obtener el resumen estadístico del DataFrame
type_data = clientes_data.dtypes

# Escribir el DataFrame resultante a un archivo Excel
type_data.to_excel('/Users/produccion-ja/Documents/MAESTRIA_PUCE/TITULACION/
 ↪typedata.xlsx', engine='openpyxl')
```

[143]:
```
# Valores faltantes por columna
print("Valores faltantes: ")
clientes_data.isnull()
print(clientes_data.isnull().sum())
```

```
Valores faltantes:
CODIGO_SOCIO              0
ESTADO_CIVIL            132
PROVINCIA_RESIDENCIA      0
CARGAS_FAMILIARES         0
CREDITOS_TOTAL            0
CREDITOS_CANCELADOS       0
DIAS_MORA_PROMEDIO        0
MONTO_PROMEDIO            0
TOTAL_INGRESO           409
EGRESO_SOCIO            409
DIFERENCIA              409
OFICINA                 409
NIVEL_ESTUDIOS          409
SECTOR_DOMICILIO        409
EDAD                    409
TIPO_PERSONA            409
NIVEL_RIESGO_PREC         0
dtype: int64
```

[144]:
```
# PASAR INFORMACION DE TABLAS JUPYTER A EXCEL

# Usar el método describe() para obtener el resumen estadístico del DataFrame
datos_nulos = clientes_data.isnull().sum()

# Escribir el DataFrame resultante a un archivo Excel
datos_nulos.to_excel('/Users/produccion-ja/Documents/MAESTRIA_PUCE/TITULACION/
 ↪datos_nulos.xlsx', engine='openpyxl')
```

[145]:
```
# Para obtener un resumen descriptivo de todas las columnas categóricas
resumen_categoricas = clientes_data.describe(include=['object'])
```

```
resumen_categoricas
```

[145]:

| | ESTADO_CIVIL | PROVINCIA_RESIDENCIA | MONTO_PROMEDIO | TOTAL_INGRESO | \ |
|---|---|---|---|---|---|
| count | 55555 | 55687 | 55687 | 55278 | |
| unique | 5 | 52 | 6939 | 20235 | |
| top | SOLTERO/A | AZUAY | 10000 | 600 | |
| freq | 24029 | 20772 | 4340 | 1407 | |

| | EGRESO_SOCIO | DIFERENCIA | NIVEL_ESTUDIOS | SECTOR_DOMICILIO | \ |
|---|---|---|---|---|---|
| count | 55278 | 55278 | 55278 | 55278 | |
| unique | 22088 | 28207 | 7 | 3 | |
| top | 300 | 200 | SECUNDARIA | URBANO | |
| freq | 349 | 245 | 23509 | 35937 | |

| | TIPO_PERSONA | NIVEL_RIESGO_PREC |
|---|---|---|
| count | 55278 | 55687 |
| unique | 3 | 9 |
| top | PERSONA NATURAL | A1 |
| freq | 55147 | 22220 |

[146]:
```
#Borrar columnas que no entran por el momento en el analisis
clientes_copia1= clientes_data.drop(['OFICINA'], axis=1)
clientes_copia1.rename(columns={'DIFERENCIA': 'FLUJO_CAJA'}, inplace=True)

# Convertir columnas numéricas con comas en decimales a formato flotante
# Reemplazando '#N/D' con NaN y convirtiendo a flotante
num_cols_to_convert = ['MONTO_PROMEDIO', 'TOTAL_INGRESO', 'EGRESO_SOCIO',␣
 ↪'FLUJO_CAJA']
clientes_copia1[num_cols_to_convert] = clientes_copia1[num_cols_to_convert].
 ↪replace('#N/D', np.nan)
clientes_copia1[num_cols_to_convert] = clientes_copia1[num_cols_to_convert].
 ↪apply(lambda x: x.str.replace(',', '.').astype(float))


# Imputar datos faltantes con la media (podría ser la mediana u otro método␣
 ↪dependiendo del análisis)
clientes_copia1[num_cols_to_convert] = clientes_copia1[num_cols_to_convert].
 ↪fillna(clientes_copia1[num_cols_to_convert].mean())

# Ahora, eliminar registros donde 'TOTAL_INGRESO' <= 0
clientes_copia1 = clientes_copia1[clientes_copia1['TOTAL_INGRESO'] > 0]

# Eliminar registros donde 'FLUJO_CAJA' < 0
clientes_copia1 = clientes_copia1[clientes_copia1['FLUJO_CAJA'] >= 0]
```

[147]:
```
# Lista de nombres de columnas a transformar
clientes_copia1['EDAD'] = clientes_copia1['EDAD'].replace('#N/D', np.nan)
clientes_copia1['EDAD'].replace([np.inf, -np.inf], np.nan, inplace=True)
clientes_copia1['EDAD'].fillna(clientes_copia1['EDAD'].mode()[0], inplace=True)

# Lista de nombres de columnas a transformar
columnas_a_transformar = ['EDAD']
#clientes_copia1[columnas_a_transformar] =␣
 ↪clientes_copia1[columnas_a_transformar].replace('#N/D', np.nan)
# Iterar a través de las columnas y aplicar la conversión
for columna in columnas_a_transformar:
    clientes_copia1[columna] = clientes_copia1[columna].astype('int64')

num_cols_to_convert = ['DIAS_MORA_PROMEDIO', 'MONTO_PROMEDIO', 'TOTAL_INGRESO',␣
 ↪'EGRESO_SOCIO', 'FLUJO_CAJA', 'EDAD']
```

[148]:
```
# Mapeo de provincias a sus respectivas regiones
provincia_region = {
    'ESMERALDAS': 'COSTA', 'MANABI': 'COSTA', 'LOS RIOS': 'COSTA',
    'GUAYAS': 'COSTA', 'SANTA ELENA': 'COSTA', 'EL ORO': 'COSTA',
    'SANTO DOMINGO DE LOS TSACHILAS': 'COSTA', 'CARCHI': 'SIERRA',
    'IMBABURA': 'SIERRA', 'PICHINCHA': 'SIERRA', 'COTOPAXI': 'SIERRA',
    'TUNGURAHUA': 'SIERRA', 'BOLIVAR': 'SIERRA', 'CHIMBORAZO': 'SIERRA',
    'CAÑAR': 'SIERRA', 'AZUAY': 'SIERRA', 'LOJA': 'SIERRA',
    'SUCUMBIOS': 'ORIENTE', 'NAPO': 'ORIENTE', 'ORELLANA': 'ORIENTE',
    'PASTAZA': 'ORIENTE', 'MORONA SANTIAGO': 'ORIENTE', 'ZAMORA CHINCHIPE':␣
 ↪'ORIENTE'
}

# Creando la columna 'REGION' en el DataFrame
clientes_copia1['REGION'] = clientes_copia1['PROVINCIA_RESIDENCIA'].str.upper().
 ↪map(provincia_region, na_action='ignore')


# Rellenar los valores NaN (provincias no encontradas en el mapeo) con 'OTROS'
clientes_copia1['REGION'].fillna('OTROS', inplace=True)

#finalmente eliminar las columnas 'PROVINCIA_RESIDENCIA' y 'CODIGO_SOCIO'
clientes_copia1= clientes_copia1.drop(['PROVINCIA_RESIDENCIA','CODIGO_SOCIO'],␣
 ↪axis=1)
```

[149]:
```
clientes_copia1 = clientes_copia1.dropna()
print('DF transformado: \n',clientes_copia1.dtypes)
tipos_corregido=clientes_copia1.dtypes
descripcion_corregida=clientes_copia1.describe()
descripcion_corregida
```

[147]:
```
# Lista de nombres de columnas a transformar
clientes_copia1['EDAD'] = clientes_copia1['EDAD'].replace('#N/D', np.nan)
clientes_copia1['EDAD'].replace([np.inf, -np.inf], np.nan, inplace=True)
clientes_copia1['EDAD'].fillna(clientes_copia1['EDAD'].mode()[0], inplace=True)

# Lista de nombres de columnas a transformar
columnas_a_transformar = ['EDAD']
#clientes_copia1[columnas_a_transformar] =␣
 ↪clientes_copia1[columnas_a_transformar].replace('#N/D', np.nan)
# Iterar a través de las columnas y aplicar la conversión
for columna in columnas_a_transformar:
    clientes_copia1[columna] = clientes_copia1[columna].astype('int64')

num_cols_to_convert = ['DIAS_MORA_PROMEDIO', 'MONTO_PROMEDIO', 'TOTAL_INGRESO',␣
 ↪'EGRESO_SOCIO', 'FLUJO_CAJA', 'EDAD']
```

[148]:
```
# Mapeo de provincias a sus respectivas regiones
provincia_region = {
    'ESMERALDAS': 'COSTA', 'MANABI': 'COSTA', 'LOS RIOS': 'COSTA',
    'GUAYAS': 'COSTA', 'SANTA ELENA': 'COSTA', 'EL ORO': 'COSTA',
    'SANTO DOMINGO DE LOS TSACHILAS': 'COSTA', 'CARCHI': 'SIERRA',
    'IMBABURA': 'SIERRA', 'PICHINCHA': 'SIERRA', 'COTOPAXI': 'SIERRA',
    'TUNGURAHUA': 'SIERRA', 'BOLIVAR': 'SIERRA', 'CHIMBORAZO': 'SIERRA',
    'CAÑAR': 'SIERRA', 'AZUAY': 'SIERRA', 'LOJA': 'SIERRA',
    'SUCUMBIOS': 'ORIENTE', 'NAPO': 'ORIENTE', 'ORELLANA': 'ORIENTE',
    'PASTAZA': 'ORIENTE', 'MORONA SANTIAGO': 'ORIENTE', 'ZAMORA CHINCHIPE':␣
 ↪'ORIENTE'
}

# Creando la columna 'REGION' en el DataFrame
clientes_copia1['REGION'] = clientes_copia1['PROVINCIA_RESIDENCIA'].str.upper().
 ↪map(provincia_region, na_action='ignore')


# Rellenar los valores NaN (provincias no encontradas en el mapeo) con 'OTROS'
clientes_copia1['REGION'].fillna('OTROS', inplace=True)

#finalmente eliminar las columnas 'PROVINCIA_RESIDENCIA' y 'CODIGO_SOCIO'
clientes_copia1= clientes_copia1.drop(['PROVINCIA_RESIDENCIA','CODIGO_SOCIO'],␣
 ↪axis=1)
```

[149]:
```
clientes_copia1 = clientes_copia1.dropna()
print('DF transformado: \n',clientes_copia1.dtypes)
tipos_corregido=clientes_copia1.dtypes
descripcion_corregida=clientes_copia1.describe()
descripcion_corregida
```

```
DF transformado:
 ESTADO_CIVIL            object
CARGAS_FAMILIARES        int64
CREDITOS_TOTAL           int64
CREDITOS_CANCELADOS      int64
DIAS_MORA_PROMEDIO       int64
MONTO_PROMEDIO         float64
TOTAL_INGRESO          float64
EGRESO_SOCIO           float64
FLUJO_CAJA             float64
NIVEL_ESTUDIOS          object
SECTOR_DOMICILIO        object
EDAD                     int64
TIPO_PERSONA            object
NIVEL_RIESGO_PREC       object
REGION                  object
dtype: object
```

[149]:

| | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS |
|---|---|---|---|
| count | 54774.000000 | 54774.000000 | 54774.000000 |
| mean | 1.027787 | 5.115894 | 3.717695 |
| std | 1.114284 | 10.671237 | 9.874067 |
| min | 0.000000 | 1.000000 | 0.000000 |
| 25% | 0.000000 | 1.000000 | 0.000000 |
| 50% | 1.000000 | 2.000000 | 1.000000 |
| 75% | 2.000000 | 5.000000 | 4.000000 |
| max | 12.000000 | 338.000000 | 305.000000 |

| | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO |
|---|---|---|---|---|
| count | 54774.000000 | 54774.000000 | 5.477400e+04 | 54774.000000 |
| mean | 7.183226 | 8999.238238 | 6.740365e+03 | 1433.762376 |
| std | 23.017511 | 10729.302298 | 1.066818e+06 | 5426.888185 |
| min | 0.000000 | 18.670000 | 2.000000e+00 | 0.000000 |
| 25% | 0.000000 | 2015.515000 | 7.300000e+02 | 390.000000 |
| 50% | 2.000000 | 5000.000000 | 1.200000e+03 | 680.000000 |
| 75% | 7.000000 | 12000.000000 | 2.180000e+03 | 1378.225000 |
| max | 1871.000000 | 700000.000000 | 2.496737e+08 | 535292.290000 |

| | FLUJO_CAJA | EDAD |
|---|---|---|
| count | 5.477400e+04 | 54774.000000 |
| mean | 5.306603e+03 | 40.491018 |
| std | 1.066664e+06 | 13.710070 |
| min | 0.000000e+00 | 18.000000 |
| 25% | 2.475500e+02 | 30.000000 |
| 50% | 4.200000e+02 | 38.000000 |
| 75% | 7.758750e+02 | 50.000000 |
| max | 2.496401e+08 | 94.000000 |

[150]:
```
# Para obtener un resumen descriptivo de todas las columnas categóricas
resumen_categoricas = clientes_copia1.describe(include=['object'])
resumen_categoricas
```

[150]:

| | ESTADO_CIVIL | NIVEL_ESTUDIOS | SECTOR_DOMICILIO | TIPO_PERSONA |
|---|---|---|---|---|
| count | 54774 | 54774 | 54774 | 54774 |
| unique | 5 | 7 | 3 | 1 |
| top | SOLTERO/A | SECUNDARIA | URBANO | PERSONA NATURAL |
| freq | 23820 | 23324 | 35575 | 54774 |

| | NIVEL_RIESGO_PREC | REGION |
|---|---|---|
| count | 54774 | 54774 |
| unique | 9 | 4 |
| top | A1 | SIERRA |
| freq | 21780 | 34640 |

[151]:
```
# PASAR INFORMACION DE TABLAS JUPYTER A EXCEL

# Usar el método describe() para obtener el resumen estadístico del DataFrame

# Escribir el DataFrame resultante a un archivo Excel
tipos_corregido.to_excel('/Users/produccion-ja/Documents/MAESTRIA_PUCE/
 ↪TITULACION/typedata.xlsx', engine='openpyxl')
descripcion_corregida.to_excel('/Users/produccion-ja/Documents/MAESTRIA_PUCE/
 ↪TITULACION/resultado.xlsx', engine='openpyxl')
resumen_categoricas.to_excel('/Users/produccion-ja/Documents/MAESTRIA_PUCE/
 ↪TITULACION/categoricas.xlsx', engine='openpyxl')
```

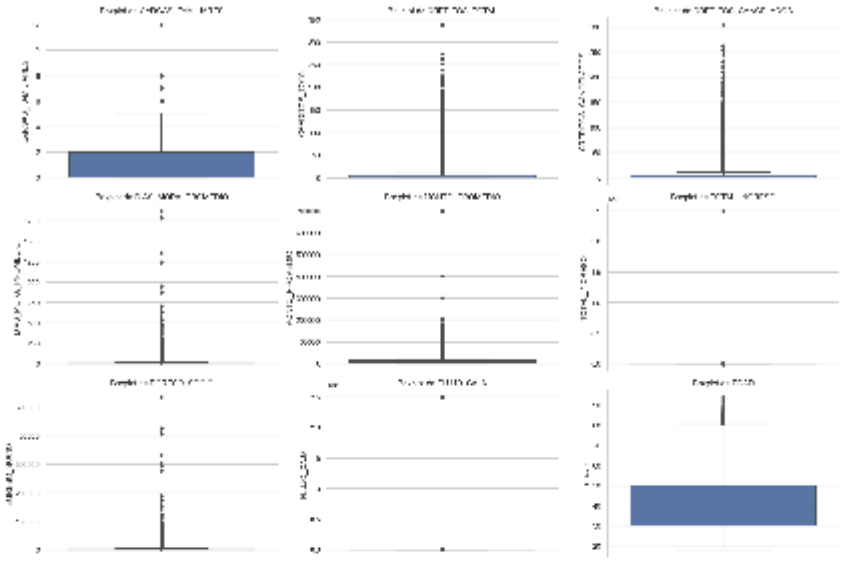
[152]:
```
print('Revision de nulos o vacios:\n',clientes_copia1.isna().sum())
```

```
Revision de nulos o vacios:
 ESTADO_CIVIL          0
CARGAS_FAMILIARES      0
CREDITOS_TOTAL         0
CREDITOS_CANCELADOS    0
DIAS_MORA_PROMEDIO     0
MONTO_PROMEDIO         0
TOTAL_INGRESO          0
EGRESO_SOCIO           0
FLUJO_CAJA             0
NIVEL_ESTUDIOS         0
SECTOR_DOMICILIO       0
EDAD                   0
TIPO_PERSONA           0
NIVEL_RIESGO_PREC      0
REGION                 0
dtype: int64
```

[153]:
```
#eliminar la columna TIPO_PERSONA YA QUE QUEDA EN DESHUSO
clientes_copia1= clientes_copia1.drop(['TIPO_PERSONA'], axis=1)
```

### 0.1.2 REVISION DE OUTLIERS EN BOXPLOT (LA PREMISA ES QUE NO NECESARIAMENTE HAY OUTLIERS YA QUE ES DATA REGISTRADA, LO QUE PUEDE EXISTIR ES ERROR OPERATIVO)

[154]:
```
# Revision de ouliers de forma gráfica
# Para cada columna numérica en los datos, genera un boxplot
#for column in clientes_copia1.select_dtypes(include=['float64', 'int64']).
 ↪columns:
#    plt.figure(figsize=(8, 4))  # Configura el tamaño de la figura
#    sns.boxplot(y=clientes_copia1[column])  # Crea un boxplot vertical
#    plt.title(f'Boxplot de {column}')  # Establece el título del boxplot
#    plt.show()  # Muestra el boxplot




# Determinar la cantidad de columnas numéricas
num_cols = clientes_copia1.select_dtypes(include=['float64', 'int64']).columns
n = len(num_cols)

# Calcular el número de filas/columnas necesarias para los subplots
n_cols = 3  # Número deseado de columnas
n_rows = n // n_cols if n % n_cols == 0 else n // n_cols + 1

# Crear una figura y un conjunto de subplots
fig, axes = plt.subplots(n_rows, n_cols, figsize=(n_cols * 6, n_rows * 4))

# Aplanar el array de ejes para una iteración más fácil
axes = axes.flatten()

# Generar un boxplot para cada columna numérica
for i, column in enumerate(num_cols):
    sns.boxplot(y=clientes_copia1[column], ax=axes[i])  # Asignar el boxplot a␣
 ↪un eje
    axes[i].set_title(f'Boxplot de {column}')  # Establecer el título para cada␣
 ↪subplot

# Esconder los ejes no utilizados (en caso de haberlos)
for j in range(i + 1, n_cols * n_rows):
    axes[j].set_visible(False)

plt.tight_layout()  # Ajustar automáticamente los parámetros de la subtrama
plt.show()  # Mostrar los boxplots
```



[155]:
```
clientes_copia1.describe()
```

[155]:

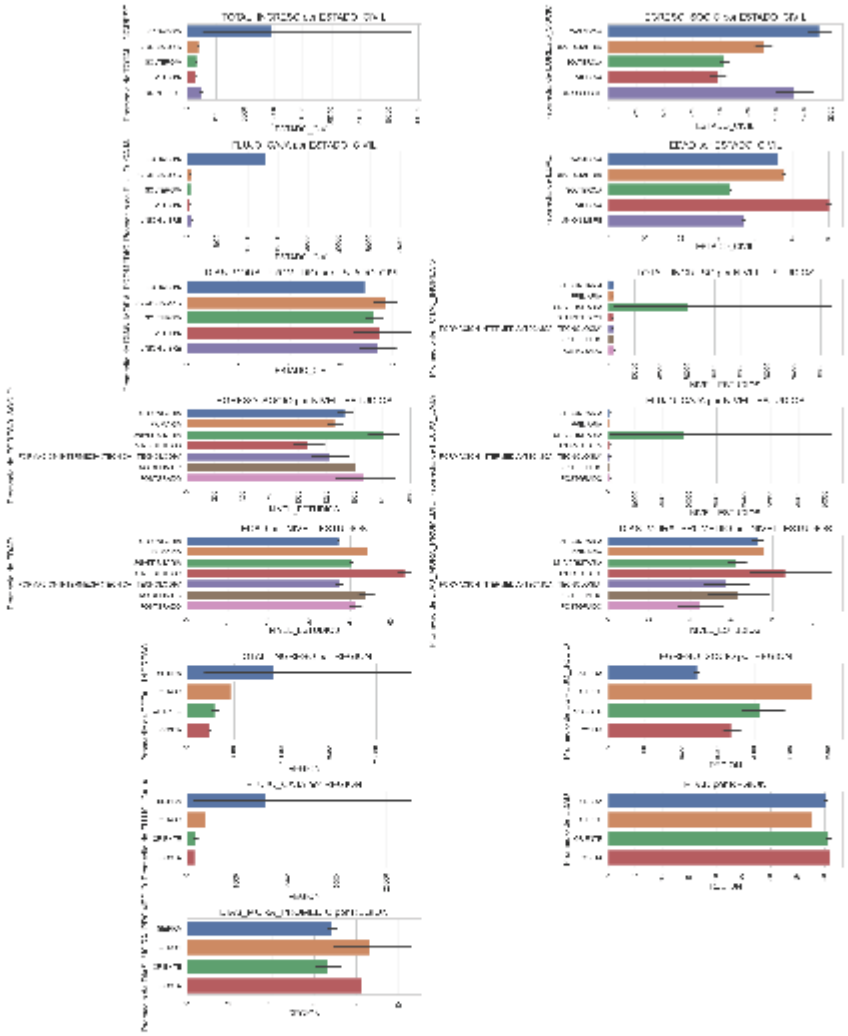
| | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | \ |
|---|---|---|---|---|
| count | 54774.000000 | 54774.000000 | 54774.000000 | |
| mean | 1.027787 | 5.115894 | 3.717695 | |
| std | 1.114284 | 10.671237 | 9.874067 | |
| min | 0.000000 | 1.000000 | 0.000000 | |
| 25% | 0.000000 | 1.000000 | 0.000000 | |
| 50% | 1.000000 | 2.000000 | 1.000000 | |
| 75% | 2.000000 | 5.000000 | 4.000000 | |
| max | 12.000000 | 338.000000 | 305.000000 | |

| | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | \ |
|---|---|---|---|---|---|
| count | 54774.000000 | 54774.000000 | 5.477400e+04 | 54774.000000 | |
| mean | 7.183226 | 8999.238238 | 6.740365e+03 | 1433.762376 | |
| std | 23.017511 | 10729.302298 | 1.066818e+06 | 5426.888185 | |
| min | 0.000000 | 18.670000 | 2.000000e+00 | 0.000000 | |
| 25% | 0.000000 | 2015.515000 | 7.300000e+02 | 390.000000 | |
| 50% | 2.000000 | 5000.000000 | 1.200000e+03 | 680.000000 | |
| 75% | 7.000000 | 12000.000000 | 2.180000e+03 | 1378.225000 | |
| max | 1871.000000 | 700000.000000 | 2.496737e+08 | 535292.290000 | |

| | FLUJO_CAJA | EDAD |
|---|---|---|

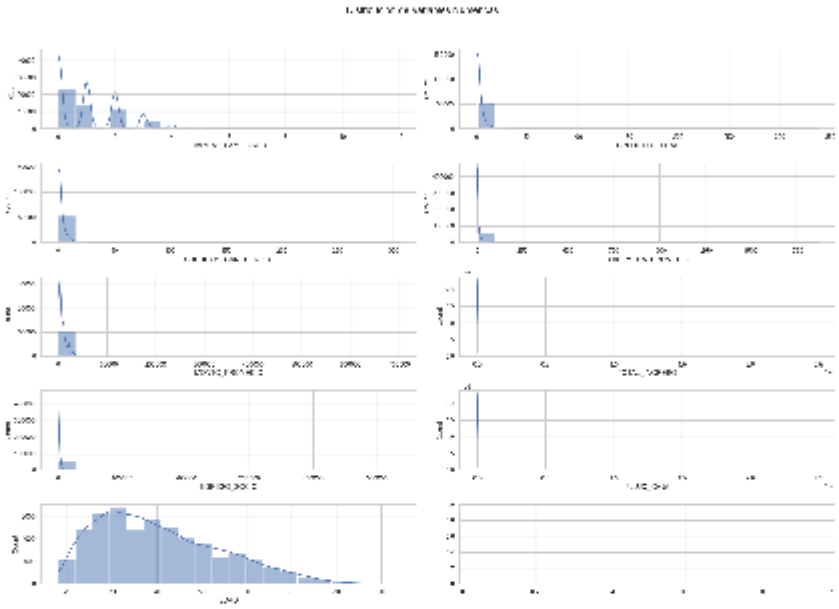
```
count  5.477400e+04  54774.000000
mean   5.306603e+03     40.491018
std    1.066664e+06     13.710070
min    0.000000e+00     18.000000
25%    2.475500e+02     30.000000
50%    4.200000e+02     38.000000
75%    7.758750e+02     50.000000
max    2.496401e+08     94.000000
```

[156]:
```
#ANALISIS DE VARIABLES CATEGORICAS

numerical_vars_extended = ['TOTAL_INGRESO', 'EGRESO_SOCIO', 'FLUJO_CAJA',␣
 ↪'EDAD', 'DIAS_MORA_PROMEDIO']
all_categorical_vars = ['ESTADO_CIVIL', 'NIVEL_ESTUDIOS', 'REGION']
# Configurando el tamaño de la figura para acomodar todos los gráficos
plt.figure(figsize=(20, 24))

# Generando gráficos para cada combinación de variable categórica y numérica␣
 ↪extendida
n_cols = 2  # Definiendo el número de columnas en la cuadrícula de gráficos
n_rows = len(all_categorical_vars) * len(numerical_vars_extended) // n_cols + 1

for i, cat_var in enumerate(all_categorical_vars):
    for j, num_var in enumerate(numerical_vars_extended):
        plt.subplot(n_rows, n_cols, i*len(numerical_vars_extended) + j + 1)
        # Para 'PROVINCIA_RESIDENCIA', limitamos la cantidad de categorías␣
 ↪mostradas para mejorar la legibilidad
        sns.barplot(x=num_var, y=cat_var, data=clientes_copia1, orient='h')
        plt.title(f"{num_var} por {cat_var}", fontsize=16)
        plt.xlabel(cat_var, fontsize=14)
        plt.ylabel(f"Promedio de {num_var}", fontsize=14)
        plt.xticks(rotation=90)

plt.tight_layout()
plt.show()
```



[157]:
```
#Visualización de distribuciones
sns.set(style="whitegrid")
fig, axes = plt.subplots(nrows=5, ncols=2, figsize=(20, 15))
fig.suptitle('Distribución de Variables Numéricas', fontsize=16)
```

```
for i, column in enumerate(clientes_copia1.select_dtypes(include=['float64',␣
 ↪'int64']).columns):
    sns.histplot(clientes_copia1[column], ax=axes[i // 2, i % 2], kde=True,␣
 ↪bins=20)

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```



### 0.1.3 SE ELIMINA EL VALOR MAXIMO DE LA COLUMNA 'TOTAL_INGRESO' YA QUE ES EL UNICO DATO QUE ESTA AFECTANDO DIRECTAMENTE EL AGRUPAMIENTO POR ESTAR DEMASIADO LEJOS DE LOS DATOS Y SE CONSIDERA COMO OUTLIER

[158]:
```
# Encontrar el valor máximo en la columna 'TOTAL_INGRESO'
max_valor_total_ingreso = clientes_copia1['TOTAL_INGRESO'].max()

# Filtrar el DataFrame para excluir el registro con el valor máximo en␣
 ↪'TOTAL_INGRESO'
clientes_copia1 = clientes_copia1[clientes_copia1['TOTAL_INGRESO'] !=␣
 ↪max_valor_total_ingreso]
clientes_copia1.describe()
```

[158]:

| | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS |
|---|---|---|---|
| count | 54773.000000 | 54773.000000 | 54773.000000 |
| mean | 1.027751 | 5.115823 | 3.717653 |
| std | 1.114262 | 10.671322 | 9.874152 |
| min | 0.000000 | 1.000000 | 0.000000 |
| 25% | 0.000000 | 1.000000 | 0.000000 |
| 50% | 1.000000 | 2.000000 | 1.000000 |
| 75% | 2.000000 | 5.000000 | 4.000000 |
| max | 12.000000 | 338.000000 | 305.000000 |

| | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO |
|---|---|---|---|---|
| count | 54773.000000 | 54773.00000 | 54773.000000 | 54773.000000 |
| mean | 7.183265 | 8999.24648 | 2182.153252 | 1433.175676 |
| std | 23.017719 | 10729.40007 | 6676.893051 | 5425.200330 |
| min | 0.000000 | 18.67000 | 2.000000 | 0.000000 |
| 25% | 0.000000 | 2015.13000 | 730.000000 | 390.000000 |
| 50% | 2.000000 | 5000.00000 | 1200.000000 | 680.000000 |
| 75% | 7.000000 | 12000.00000 | 2180.000000 | 1377.340000 |
| max | 1871.000000 | 700000.00000 | 536874.710000 | 535292.290000 |

| | FLUJO_CAJA | EDAD |
|---|---|---|
| count | 54773.000000 | 54773.000000 |
| mean | 748.977575 | 40.490862 |
| std | 2915.998323 | 13.710147 |
| min | 0.000000 | 18.000000 |
| 25% | 247.550000 | 30.000000 |
| 50% | 420.000000 | 38.000000 |
| 75% | 775.500000 | 50.000000 |
| max | 536844.710000 | 94.000000 |

## 0.2 NORMALIZACION DE DATOS

[159]:

```
#normalizacion de datos

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Variables numéricas y categóricas
columnas_numericas = ['CARGAS_FAMILIARES', 'CREDITOS_TOTAL',␣
 ↪'CREDITOS_CANCELADOS', 'DIAS_MORA_PROMEDIO', 'MONTO_PROMEDIO',␣
 ↪'TOTAL_INGRESO', 'EGRESO_SOCIO', 'FLUJO_CAJA', 'EDAD']
columnas_categoricas = ['ESTADO_CIVIL', 'NIVEL_ESTUDIOS', 'REGION',␣
 ↪'NIVEL_RIESGO_PREC']

# Pipeline de preprocesamiento
preprocesador = ColumnTransformer(
```

```
    transformers=[
        ('num', StandardScaler(), columnas_numericas),
        ('cat', OneHotEncoder(handle_unknown='ignore'), columnas_categoricas)
    ]
)
# Aplicando el preprocesamiento a los datos
datos_normalizados = preprocesador.fit_transform(clientes_copia1)
datos_normalizados.shape
```

[159]:
```
(54773, 34)
```

# 1 USANDO KMEANS

### 1.0.1 Ejecucion individual del modelo

[160]:
```
#Ejecutar previamente una reduccion de dimensionalidad con PCA
#para definir centroides de cada cluster en 2 dimensiones
# Reducción de dimensionalidad con PCA

import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
componentes_principales = pca.fit_transform(datos_normalizados)
```

[173]:
```
#Ejecucion y prueba del cluster usando K-means
from sklearn.cluster import KMeans

# Ejecutando nuevamente el análisis de clustering con 3 clusters
kmeans = KMeans(n_clusters=2, random_state=42)
clusters = kmeans.fit_predict(componentes_principales)
```

[174]:
```
centroides_originales=kmeans.cluster_centers_

# Crear un DataFrame con la información de los centroides
df_centroides = pd.DataFrame(centroides_originales, columns=['CentroideX',␣
 ↪'CentroideY'])

# Añadir el ID del cluster como una columna más
df_centroides['Cluster'] = df_centroides.index

# Reordenar las columnas para que el ID del cluster aparezca primero
df_centroides = df_centroides[['Cluster', 'CentroideX', 'CentroideY']]

print(df_centroides)
```

```
   Cluster  CentroideX  CentroideY
```

```
0          0   -0.064441   -0.137700
1          1    3.359049    7.177786
```

### 1.0.2 EVALUACION DEL MODELO (SCORE DE SILUETA)

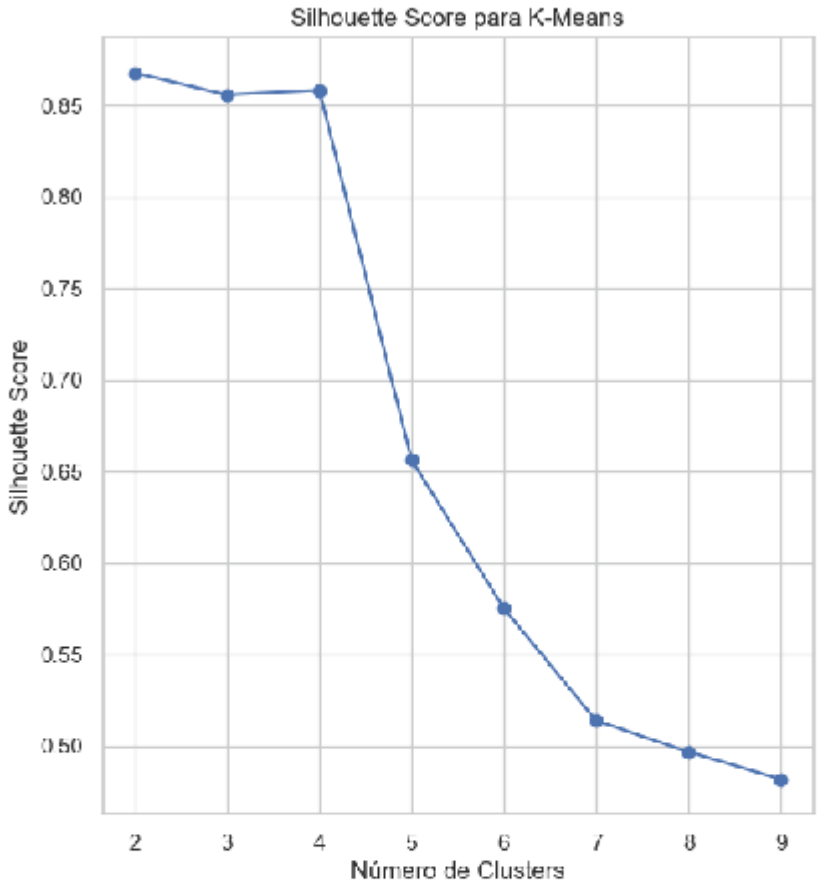
[168]:
```
#Evaluar la calidad del modelo
from sklearn.metrics import silhouette_score

# Calculando el Silhouette Score para evaluar la calidad del clustering
silhouette_avg = silhouette_score(componentes_principales, clusters)
silhouette_avg
```

[168]:
```
0.8585059026886478
```

### 1.0.3 EJECUCION ITERATIVA DEL MODELO K-MEANS DESDE 2 A 9 CLUSTERS - VISUALIZACION DE SILHOUETTE SCORE

[165]:
```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
# Rango de clusters a evaluar para K-Means
range_clusters = range(2, 10)

# Inicializando listas para guardar los resultados de Silhouette Score
silhouette_scores_kmeans = []

# Evaluación de K-Means con diferentes números de clusters
for n_clusters in range_clusters:
    kmeans_labels = KMeans(n_clusters=n_clusters, random_state=42)
    cluster_labels = kmeans_labels.fit_predict(componentes_principales)
    silhouette_avg = silhouette_score(componentes_principales, cluster_labels)
    silhouette_scores_kmeans.append(silhouette_avg)

    # Graficando los resultados de Silhouette Score para K-Means
plt.figure(figsize=(14, 7))
plt.subplot(1, 2, 1)
plt.plot(range_clusters, silhouette_scores_kmeans, marker='o')
plt.title('Silhouette Score para K-Means')
plt.xlabel('Número de Clusters')
plt.ylabel('Silhouette Score')
```

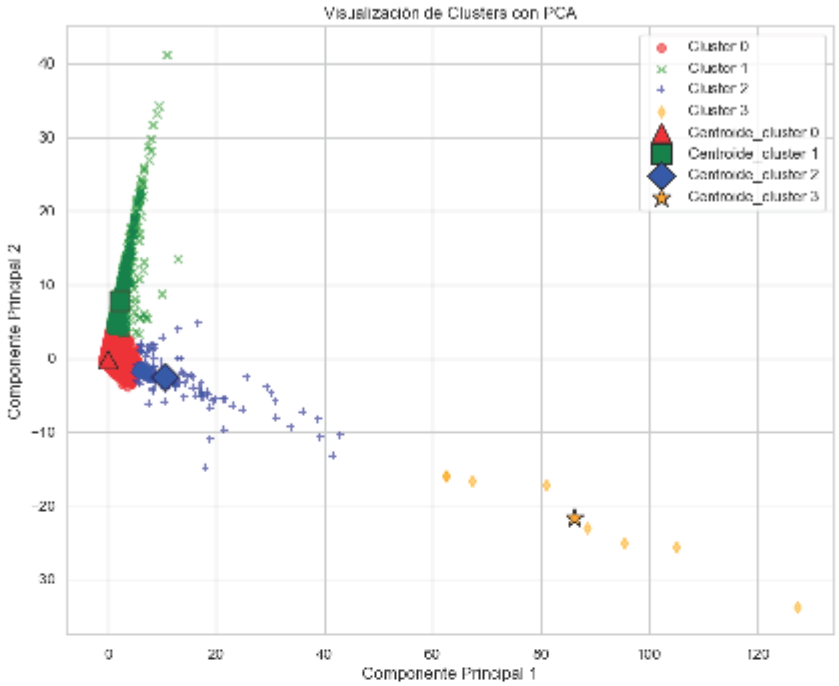
[165]:
```
Text(0, 0.5, 'Silhouette Score')
```




### 1.0.4 GRAFICO DE COMPONENTES PRINCIPALES PARA VISUALIZAR LOS AGRUPAMIENTOS; k=4

[169]:
```
import matplotlib.pyplot as plt
import pandas as pd


# Creando un dataframe para la visualización
df_visual = pd.DataFrame(data=componentes_principales, columns=['PC1', 'PC2'])
df_visual['Cluster'] = clusters
```

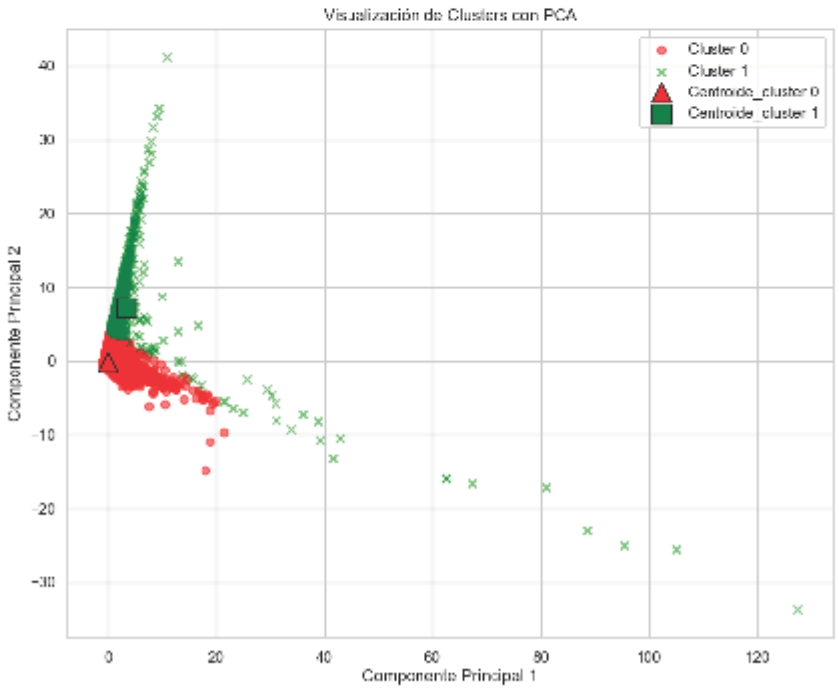
```
# Gráfico de los clusters
plt.figure(figsize=(10, 8))
#colors = ['red', 'green','blue']  # Colores para cada cluster
#markers = ['o', 'x','*']  # Marcadores para los puntos de cada cluster
#centroid_markers = ['^','s','p']  # Marcador para los centroides
colors = ['red', 'green', 'blue','orange']  # Colores para cada cluster
markers = ['o', 'x','+','d']  # Marcadores para los puntos de cada cluster
centroid_markers = ['^','s','D','*']  # Marcador para los centroides

# Dibujar los puntos de cada cluster
for i in range(4):
    plt.scatter(df_visual[df_visual['Cluster'] == i]['PC1'],␣
 ↪df_visual[df_visual['Cluster'] == i]['PC2'],
                color=colors[i], label=f'Cluster {i}', alpha=0.5,␣
 ↪marker=markers[i])

# Dibujar los centroides
for i, color in enumerate(colors):
    plt.scatter(kmeans.cluster_centers_[i, 0], kmeans.cluster_centers_[i, 1],
                color=color, edgecolor='k', marker=centroid_markers[i], s=200,␣
 ↪label=f'Centroide_cluster {i}')

plt.title('Visualización de Clusters con PCA')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend()
plt.show()
```




### 1.0.5 GRAFICO DE COMPONENTES PRINCIPALES PARA VISUALIZAR LOS AGRUPAMIENTOS; k=2

[175]:
```
import matplotlib.pyplot as plt
import pandas as pd


# Creando un dataframe para la visualización
df_visual = pd.DataFrame(data=componentes_principales, columns=['PC1', 'PC2'])
df_visual['Cluster'] = clusters

# Gráfico de los clusters
plt.figure(figsize=(10, 8))
#colors = ['red', 'green','blue']  # Colores para cada cluster
#markers = ['o', 'x','*']  # Marcadores para los puntos de cada cluster
#centroid_markers = ['^','s','p']  # Marcador para los centroides
colors = ['red', 'green']  # Colores para cada cluster
markers = ['o', 'x']  # Marcadores para los puntos de cada cluster
```

```
centroid_markers = ['^','s']  # Marcador para los centroides

# Dibujar los puntos de cada cluster
for i in range(2):
    plt.scatter(df_visual[df_visual['Cluster'] == i]['PC1'],␣
 ↪df_visual[df_visual['Cluster'] == i]['PC2'],
                color=colors[i], label=f'Cluster {i}', alpha=0.5,␣
 ↪marker=markers[i])

# Dibujar los centroides
for i, color in enumerate(colors):
    plt.scatter(kmeans.cluster_centers_[i, 0], kmeans.cluster_centers_[i, 1],
                color=color, edgecolor='k', marker=centroid_markers[i], s=200,␣
 ↪label=f'Centroide_cluster {i}')

plt.title('Visualización de Clusters con PCA')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend()
plt.show()
```




21

[177]:
```
# Añadiendo la asignación de clusters al dataframe original para análisis
clientes_copia1['Cluster'] = clusters
cluster_description_kmeans_conteo = clientes_copia1.groupby('Cluster').count()
cluster_description_kmeans_conteo
```

[177]:

| Cluster | ESTADO_CIVIL | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | FLUJO_CAJA | NIVEL_ESTUDIOS | SECTOR_DOMICILIO | EDAD | NIVEL_RIESGO_PREC | REGION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 | 53742 |
| 1 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 | 1031 |

[178]:
```
cluster_description_kmeans1_media = clientes_copia1.groupby('Cluster').mean()
cluster_description_kmeans1_media
```

[178]:

| Cluster | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | FLUJO_CAJA | EDAD |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.023203 | 3.976555 | 2.669662 | 7.258866 | 9125.568680 | 2051.084297 | 1334.566992 | 716.517305 | 40.515965 |
| 1 | 1.264791 | 64.501455 | 58.345296 | 3.242483 | 2414.563967 | 9014.265538 | 6573.260941 | 2441.004597 | 39.182347 |

[179]:
```
# Escribir el DataFrame resultante a un archivo Excel
cluster_description_kmeans1_media.to_excel('/Users/produccion-ja/Documents/
↪MAESTRIA_PUCE/TITULACION/SEGMENTACION1.xlsx', engine='openpyxl')
```

## 2 Usando variaciones de kmeans

[66]:
```
#AJUSTANDO PARAMETROS DE KMEANS
#Ejecucion y prueba del cluster usando K-means
from sklearn.cluster import KMeans

# Ejecutando nuevamente el análisis de clustering con 4 clusters
kmeans = KMeans(n_clusters=3, random_state=42, max_iter=1000,tol=1e-5)
clusters = kmeans.fit_predict(datos_normalizados)

# Añadiendo la asignación de clusters al dataframe original para análisis
clientes_copia1['Cluster'] = clusters
```

### 2.1 GENERANDO SUBSETS DE DATOS PARA EJECUTAR K-MEANS NUEVAMENTE

[180]:
```
clientes_clustser_subset0= clientes_copia1[clientes_copia1['Cluster'] == 0]
print('Total:',clientes_clustser_subset0.shape)
clientes_clustser_subset0.head()
```

```
Total: (53742, 15)
```

[180]:

| | ESTADO_CIVIL | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | FLUJO_CAJA | NIVEL_ESTUDIOS | SECTOR_DOMICILIO | EDAD | NIVEL_RIESGO_PREC | REGION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | CASADO/A | 0 | 4 | 3 | 10 | 3500.00 | 2746.55 | 2505.23 | 241.32 | SECUNDARIA | URBANO | 65 | M1 | SIERRA |
| 1 | CASADO/A | 1 | 14 | 13 | 8 | 3000.00 | 1125.00 | 708.00 | 417.00 | PRIMARIA | RURAL | 53 | B1 | SIERRA |
| 2 | DIVORCIADO/A | 1 | 4 | 3 | 0 | 10000.00 | 986.00 | 598.59 | 387.41 | UNIVERSITARIA | RURAL | 41 | M1 | SIERRA |
| 3 | CASADO/A | 0 | 17 | 14 | 60 | 472.65 | 2050.00 | 1580.00 | 470.00 | UNIVERSITARIA | URBANO | 60 | B1 | SIERRA |
| 4 | SOLTERO/A | 0 | 15 | 14 | 10 | 6000.00 | 500.00 | 207.00 | 293.00 | SECUNDARIA | URBANO | 55 | B1 | SIERRA |

```
   Cluster
0        0
1        0
2        0
3        0
4        0
```

[181]:
```
clientes_clustser_subset1= clientes_copia1[clientes_copia1['Cluster'] == 1]
print('Total:',clientes_clustser_subset1.shape)
clientes_clustser_subset1.head()
```

```
Total: (1031, 15)
```

[181]:
```
     ESTADO_CIVIL  CARGAS_FAMILIARES  CREDITOS_TOTAL  CREDITOS_CANCELADOS  \
5       SOLTERO/A                  0              71                   64
40       CASADO/A                  1             105                  103
205      CASADO/A                  2              46                   43
224  DIVORCIADO/A                  0              61                   53
232      CASADO/A                  2              67                   60

     DIAS_MORA_PROMEDIO  MONTO_PROMEDIO  TOTAL_INGRESO  EGRESO_SOCIO  \
5                     1         1279.05        1701.70       1027.00
40                    0          501.95        2085.69       1163.00
205                   1         1020.99        1350.00        400.00
224                   2         1278.07        1945.00       1447.00
232                   4          563.94        1966.62       1499.07

     FLUJO_CAJA NIVEL_ESTUDIOS SECTOR_DOMICILIO  EDAD NIVEL_RIESGO_PREC  \
5        674.70  UNIVERSITARIA           URBANO    43                B2
40       922.69  UNIVERSITARIA            RURAL    38                B2
205      950.00     SECUNDARIA           URBANO    43                M1
224      498.00  UNIVERSITARIA           URBANO    54                M2
232      467.55  UNIVERSITARIA            RURAL    41                B1

      REGION  Cluster
5      OTROS        1
40    SIERRA        1
205  ORIENTE        1
224  ORIENTE        1
232  ORIENTE        1
```

[182]:
```
#ELIMINAR LA COLUMNA DEL PRIMER AGRUPAMIENTO 'CLUSTER'

clientes_clustser_subset0_drop=clientes_clustser_subset0.
 ↪drop(['Cluster'],axis=1)
```

```
#clientes_clustser_subset2_drop=clientes_clustser_subset2.
 ↪drop(['Cluster'],axis=1)
clientes_clustser_subset0_drop.shape
```

[182]: (53742, 14)

[183]:
```
clientes_clustser_subset0_drop
```

[183]:

| | ESTADO_CIVIL | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | \ |
|---|---|---|---|---|---|
| 0 | CASADO/A | 0 | 4 | 3 | |
| 1 | CASADO/A | 1 | 14 | 13 | |
| 2 | DIVORCIADO/A | 1 | 4 | 3 | |
| 3 | CASADO/A | 0 | 17 | 14 | |
| 4 | SOLTERO/A | 0 | 15 | 14 | |
| ... | ... | ... | ... | ... | |
| 55667 | CASADO/A | 2 | 1 | 0 | |
| 55668 | CASADO/A | 1 | 1 | 0 | |
| 55669 | SOLTERO/A | 0 | 1 | 0 | |
| 55670 | SOLTERO/A | 0 | 1 | 0 | |
| 55671 | CASADO/A | 0 | 1 | 0 | |

| | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | \ |
|---|---|---|---|---|---|
| 0 | 10 | 3500.00 | 2746.55 | 2505.23 | |
| 1 | 8 | 3000.00 | 1125.00 | 708.00 | |
| 2 | 0 | 10000.00 | 986.00 | 598.59 | |
| 3 | 60 | 472.65 | 2050.00 | 1580.00 | |
| 4 | 10 | 6000.00 | 500.00 | 207.00 | |
| ... | ... | ... | ... | ... | |
| 55667 | 2 | 10000.00 | 682.98 | 390.00 | |
| 55668 | 5 | 20000.00 | 4634.60 | 2681.00 | |
| 55669 | 3 | 2500.00 | 300.00 | 259.16 | |
| 55670 | 0 | 20000.00 | 1504.79 | 626.00 | |
| 55671 | 1 | 20000.00 | 2707.66 | 900.00 | |

| | FLUJO_CAJA | NIVEL_ESTUDIOS | SECTOR_DOMICILIO | EDAD | NIVEL_RIESGO_PREC | \ |
|---|---|---|---|---|---|---|
| 0 | 241.32 | SECUNDARIA | URBANO | 65 | M1 | |
| 1 | 417.00 | PRIMARIA | RURAL | 53 | B1 | |
| 2 | 387.41 | UNIVERSITARIA | RURAL | 41 | M1 | |
| 3 | 470.00 | UNIVERSITARIA | URBANO | 60 | B1 | |
| 4 | 293.00 | SECUNDARIA | URBANO | 55 | B1 | |
| ... | ... | ... | ... | ... | ... | |
| 55667 | 292.98 | SECUNDARIA | URBANO | 40 | M1 | |
| 55668 | 1953.60 | PRIMARIA | URBANO | 46 | A2 | |
| 55669 | 40.84 | PRIMARIA | RURAL | 23 | A1 | |
| 55670 | 878.79 | SECUNDARIA | URBANO | 19 | A2 | |
| 55671 | 1807.66 | SECUNDARIA | URBANO | 63 | A2 | |

```
       REGION
0      SIERRA
1      SIERRA
2      SIERRA
3      SIERRA
4      SIERRA
...       ...
55667  SIERRA
55668  SIERRA
55669  SIERRA
55670  SIERRA
55671  SIERRA

[53742 rows x 14 columns]
```

[187]:
```
#NORMALIZACION DE DATOS

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Variables numéricas y categóricas
columnas_numericas = ['CARGAS_FAMILIARES', 'CREDITOS_TOTAL',␣
 ↪'CREDITOS_CANCELADOS', 'DIAS_MORA_PROMEDIO', 'MONTO_PROMEDIO',␣
 ↪'TOTAL_INGRESO', 'EGRESO_SOCIO', 'FLUJO_CAJA', 'EDAD']
columnas_categoricas = ['ESTADO_CIVIL', 'NIVEL_ESTUDIOS', 'REGION',␣
 ↪'NIVEL_RIESGO_PREC']

# Pipeline de preprocesamiento
preprocesador = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), columnas_numericas),
        ('cat', OneHotEncoder(handle_unknown='ignore'), columnas_categoricas)
    ]
)
# Aplicando el preprocesamiento a los datos
datos_normalizados0 = preprocesador.
 ↪fit_transform(clientes_clustser_subset0_drop)
#datos_normalizados2 = preprocesador.
 ↪fit_transform(clientes_clustser_subset2_drop)
```

[188]:
```
print('Dimensiones cluster 0:',datos_normalizados0.shape)
#print('Dimensiones cluster 2:',datos_normalizados2.shape)
```

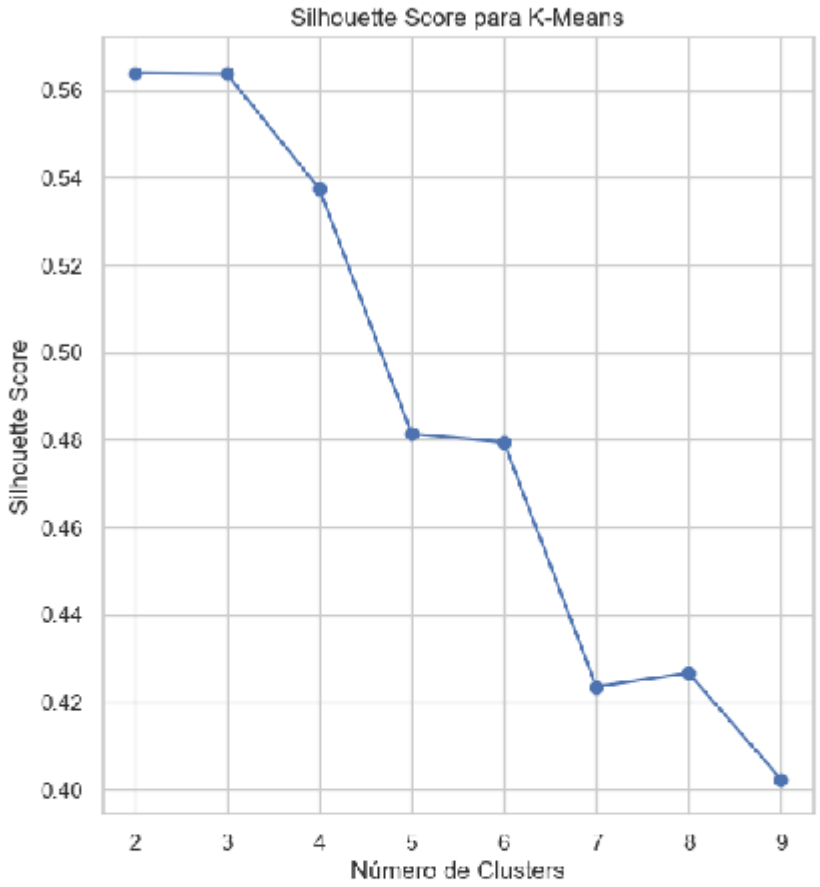
```
Dimensiones cluster 0: (53742, 34)
```

[190]:
```
# Reducción de dimensionalidad con PCA
pca_subset0 = PCA(n_components=2)
componentes_principales_subset0 = pca_subset0.fit_transform(datos_normalizados0)
```

### 2.1.1 EJECUCION ITERATIVA DEL MODELO K-MEANS DESDE 2 A 9 CLUSTERS - VISUALIZACION DE SILHOUETTE SCORE

[191]:
```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
# Rango de clusters a evaluar para K-Means
range_clusters = range(2, 10)

# Inicializando listas para guardar los resultados de Silhouette Score
silhouette_scores_kmeans_subset0 = []

# Evaluación de K-Means con diferentes números de clusters
for n_clusters in range_clusters:
    kmeans_labels = KMeans(n_clusters=n_clusters, random_state=42)
    cluster_labels = kmeans_labels.fit_predict(componentes_principales_subset0)
    silhouette_avg = silhouette_score(componentes_principales_subset0,␣
 ↪cluster_labels)
    silhouette_scores_kmeans_subset0.append(silhouette_avg)

    # Graficando los resultados de Silhouette Score para K-Means
plt.figure(figsize=(14, 7))
plt.subplot(1, 2, 1)
plt.plot(range_clusters, silhouette_scores_kmeans_subset0, marker='o')
plt.title('Silhouette Score para K-Means')
plt.xlabel('Número de Clusters')
plt.ylabel('Silhouette Score')
```

[191]:
```
Text(0, 0.5, 'Silhouette Score')
```




[192]:
```
#Ejecucion y prueba del cluster usando K-means para subset cluster 0
from sklearn.cluster import KMeans

# Ejecutando nuevamente el análisis de clustering con 3 clusters
kmeansSubset0 = KMeans(n_clusters=2, random_state=42)
clustersSubset0 = kmeansSubset0.fit_predict(componentes_principales_subset0)
```

[193]:
```
#Evaluar la calidad del modelo
from sklearn.metrica import silhouette_score

# Calculando el Silhouette Score para evaluar la calidad del clustering
```

```
silhouette_avgSubset0 = silhouette_score(componentes_principales_subset0,␣
 ↪clustersSubset0)
silhouette_avgSubset0
```

[193]: 0.5639394404452174

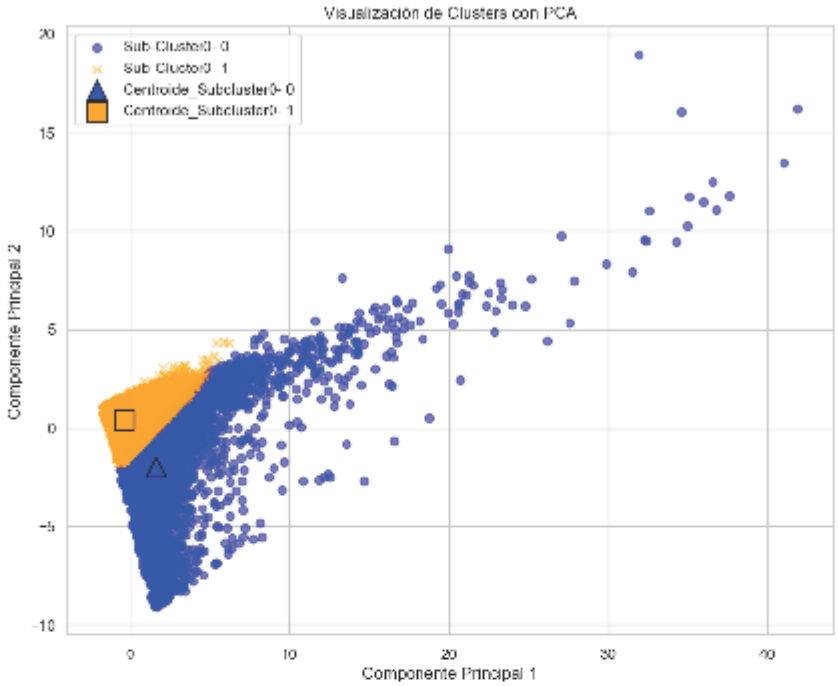
[194]:
```
centroides_originales_subset0=kmeansSubset0.cluster_centers_

# Crear un DataFrame con la información de los centroides
df_centroides_subset0 = pd.DataFrame(centroides_originales_subset0,␣
 ↪columns=['CentroideX', 'CentroideY'])

# Añadir el ID del cluster como una columna más
df_centroides_subset0['Cluster'] = df_centroides_subset0.index

# Reordenar las columnas para que el ID del cluster aparezca primero
df_centroides_subset0 = df_centroides_subset0[['Cluster', 'CentroideX',␣
 ↪'CentroideY']]

print(df_centroides_subset0)
```

```
   Cluster  CentroideX  CentroideY
0        0    1.649442   -1.983598
1        1   -0.349257    0.420012
```

### 2.1.2 GRAFICAR EL SUBSET OBTENIDO CON K-MEANzS (SUBSET CLUSTER INICIAL 0)

[195]:
```
#GRAFICAR RESULTADO CLUSTER USANDO ANALISIS DE COMPONENTES PRINCIPALES
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# Convirtiendo la matriz dispersa en una matriz densa
#datos_normalizados_densos = datos_normalizados.toarray()

# Creando un dataframe para la visualización
df_visual_subset0 = pd.DataFrame(data=componentes_principales_subset0,␣
 ↪columns=['PC1', 'PC2'])
df_visual_subset0['Cluster'] = clustersSubset0

# Gráfico de los clusters
plt.figure(figsize=(10, 8))
#colors = ['red', 'green', 'blue', 'purple', 'orange']
#colors = ['red', 'green', 'blue']
colors = ['blue', 'orange']
markers = ['o', 'x']  # Marcadores para los puntos de cada cluster
centroid_markers = ['^','s']  # Marcador para los centroides
```

```
#colors = ['blue', 'orange','purple']  # Colores para cada cluster
#markers = ['o', 'x','*']  # Marcadores para los puntos de cada cluster
#centroid_markers = ['^','s','p']  # Marcador para los centroides


# Dibujar los puntos de cada cluster
for i in range(2):
    plt.scatter(df_visual_subset0[df_visual_subset0['Cluster'] == i]['PC1'],␣
 ↪df_visual_subset0[df_visual_subset0['Cluster'] == i]['PC2'],
                color=colors[i], label=f'Sub-Cluster0- {i}', alpha=0.5,␣
 ↪marker=markers[i])

# Dibujar los centroides
for i, color in enumerate(colors):
    plt.scatter(kmeansSubset0.cluster_centers_[i, 0], kmeansSubset0.
 ↪cluster_centers_[i, 1],
                color=color, edgecolor='k', marker=centroid_markers[i], s=200,␣
 ↪label=f'Centroide_Subcluster0- {i}')


plt.title('Visualización de Clusters con PCA')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend()
plt.show()
```




[196]:
```
# Añadiendo la asignación de clusters al dataframe original para análisis
clientes_clustser_subset0_drop['Cluster'] = clustersSubset0
cluster_description_kmeans_subset0 = clientes_clustser_subset0_drop.
 ↪groupby('Cluster').count()
#print('Descripción del cluster1: ',cluster_description_dbscan)
cluster_description_kmeans_subset0
```

[196]:

| Cluster | ESTADO_CIVIL | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | \ |
|---|---|---|---|---|---|
| 0 | 9354 | 9354 | 9354 | 9354 | |
| 1 | 44388 | 44388 | 44388 | 44388 | |

| Cluster | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | \ |
|---|---|---|---|---|---|
| 0 | 9354 | 9354 | 9354 | 9354 | |
| 1 | 44388 | 44388 | 44388 | 44388 | |

FLUJO_CAJA NIVEL_ESTUDIOS SECTOR_DOMICILIO EDAD \

```
Cluster
0              9354            9354             9354   9354
1             44388           44388            44388  44388

         NIVEL_RIESGO_PREC  REGION
Cluster
0                     9354    9354
1                    44388   44388
```

[197]:
```
cluster_description_kneans_subset0_mean = clientes_clustser_subset0_drop.
 ↪groupby('Cluster').mean()
cluster_description_kneans_subset0_mean
```

[197]:
```
         CARGAS_FAMILIARES  CREDITOS_TOTAL  CREDITOS_CANCELADOS  \
Cluster
0                 1.212209       11.389031             9.208360
1                 0.983374        2.414504             1.291746

         DIAS_MORA_PROMEDIO  MONTO_PROMEDIO  TOTAL_INGRESO  EGRESO_SOCIO  \
Cluster
0                  8.116528     9178.010408    4186.175164   2893.075174
1                  7.078129     9114.517497    1601.150982   1006.138463

          FLUJO_CAJA       EDAD
Cluster
0        1293.099989  48.057409
1         595.012519  38.926737
```

[199]:
```
# Escribir el DataFrame resultante a un archivo Excel
cluster_description_kneans_subset0_mean.to_excel('/Users/produccion-ja/
 ↪Documents/MAESTRIA_PUCE/TITULACION/SEGMENTACION2.xlsx', engine='openpyxl')
```

[200]:
```
clientes_clustser_subset_s1=␣
 ↪clientes_clustser_subset0_drop[clientes_clustser_subset0_drop['Cluster'] ==1]
print('Total:',clientes_clustser_subset_s1.shape)
clientes_clustser_subset_s1.head()
```

```
Total: (44388, 15)
```

[200]:
```
    ESTADO_CIVIL  CARGAS_FAMILIARES  CREDITOS_TOTAL  CREDITOS_CANCELADOS  \
0       CASADO/A                  0               4                    3
2   DIVORCIADO/A                  1               4                    3
6       CASADO/A                  1               1                    0
9       CASADO/A                  0               3                    3
10      CASADO/A                  3               5                    4

    DIAS_MORA_PROMEDIO  MONTO_PROMEDIO  TOTAL_INGRESO  EGRESO_SOCIO  \
```

```
0                  10          3500.0        2746.55       2505.23
2                   0         10000.0         986.00        598.59
6                   0         10000.0        1798.83       1003.15
9                   5          1000.0         800.00        540.00
10                  2         25000.0        1828.00       1267.23

    FLUJO_CAJA NIVEL_ESTUDIOS SECTOR_DOMICILIO  EDAD NIVEL_RIESGO_PREC  \
0       241.32     SECUNDARIA           URBANO    65                M1
2       387.41  UNIVERSITARIA            RURAL    41                M1
6       795.68     SECUNDARIA           URBANO    43                A1
9       260.00       PRIMARIA           URBANO    58                M1
10      560.77       PRIMARIA            RURAL    44                B2

    REGION  Cluster
0   SIERRA        1
2   SIERRA        1
6   SIERRA        1
9   SIERRA        1
10  SIERRA        1
```

[201]:
```
clientes_clustser_subset_s1_drop=clientes_clustser_subset_s1.
  ↪drop(['Cluster'],axis=1)
clientes_clustser_subset_s1_drop
```

[201]:
```
       ESTADO_CIVIL  CARGAS_FAMILIARES  CREDITOS_TOTAL  CREDITOS_CANCELADOS  \
0          CASADO/A                  0               4                    3
2       DIVORCIADO/A                 1               4                    3
6          CASADO/A                  1               1                    0
9          CASADO/A                  0               3                    3
10         CASADO/A                  3               5                    4
...             ...                ...             ...                  ...
55667      CASADO/A                  2               1                    0
55668      CASADO/A                  1               1                    0
55669     SOLTERO/A                  0               1                    0
55670     SOLTERO/A                  0               1                    0
55671      CASADO/A                  0               1                    0

       DIAS_MORA_PROMEDIO  MONTO_PROMEDIO  TOTAL_INGRESO  EGRESO_SOCIO  \
0                      10         3500.0        2746.55       2505.23
2                       0        10000.0         986.00        598.59
6                       0        10000.0        1798.83       1003.15
9                       5         1000.0         800.00        540.00
10                      2        25000.0        1828.00       1267.23
...                   ...            ...            ...           ...
55667                   2        10000.0         682.98        390.00
55668                   5        20000.0        4634.60       2681.00
55669                   3         2500.0         300.00        259.16
```

```
55670                 0         20000.0        1504.79        626.00
55671                 1         20000.0        2707.66        900.00

       FLUJO_CAJA NIVEL_ESTUDIOS SECTOR_DOMICILIO  EDAD NIVEL_RIESGO_PREC  \
0          241.32     SECUNDARIA           URBANO    65               M1
2          387.41  UNIVERSITARIA            RURAL    41               M1
6          795.68     SECUNDARIA           URBANO    43               A1
9          260.00       PRIMARIA           URBANO    58               M1
10         560.77       PRIMARIA            RURAL    44               B2
...           ...            ...              ...   ...              ...
55667      292.98     SECUNDARIA           URBANO    40               M1
55668     1953.60       PRIMARIA           URBANO    46               A2
55669       40.84       PRIMARIA            RURAL    23               A1
55670      878.79     SECUNDARIA           URBANO    19               A2
55671     1807.66     SECUNDARIA           URBANO    63               A2

       REGION
0      SIERRA
2      SIERRA
6      SIERRA
9      SIERRA
10     SIERRA
...       ...
55667  SIERRA
55668  SIERRA
55669  SIERRA
55670  SIERRA
55671  SIERRA

[44388 rows x 14 columns]
```

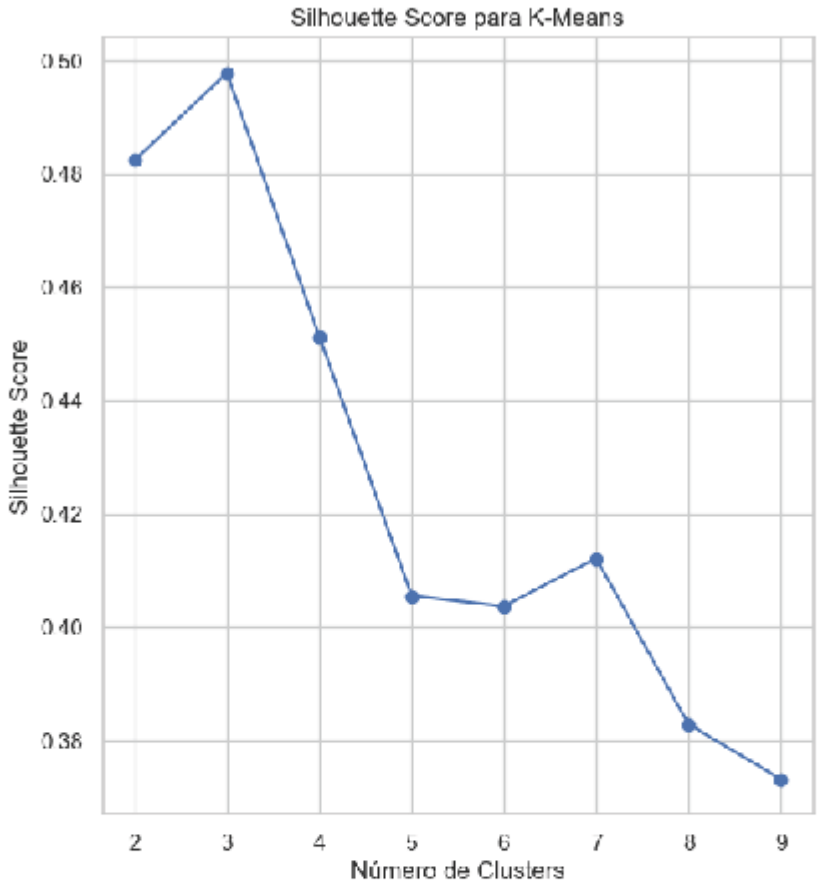
[209]:
```
#NORMALIZACION DE DATOS

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Variables numéricas y categóricas
columnas_numericas = ['CARGAS_FAMILIARES', 'CREDITOS_TOTAL',␣
 ↪'CREDITOS_CANCELADOS', 'DIAS_MORA_PROMEDIO', 'MONTO_PROMEDIO',␣
 ↪'TOTAL_INGRESO', 'EGRESO_SOCIO', 'FLUJO_CAJA', 'EDAD']
columnas_categoricas = ['ESTADO_CIVIL', 'NIVEL_ESTUDIOS', 'REGION',␣
 ↪'NIVEL_RIESGO_PREC']

# Pipeline de preprocesamiento
preprocesador = ColumnTransformer(
    transformers=[
```

```
        ('num', StandardScaler(), columnas_numericas),
        ('cat', OneHotEncoder(handle_unknown='ignore'), columnas_categoricas)
    ]
)
# Aplicando el preprocesamiento a los datos
datos_normalizados_s1 = preprocesador.
 ↪fit_transform(clientes_clustser_subset_s1_drop)
```

[210]:
```
# Reducción de dimensionalidad con PCA
pca_subset0_s1 = PCA(n_components=2)
componentes_principales_subset0_s1 = pca_subset0_s1.
 ↪fit_transform(datos_normalizados_s1)
```

[204]:
```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
# Rango de clusters a evaluar para K-Means
range_clusters = range(2, 10)

# Inicializando listas para guardar los resultados de Silhouette Score
silhouette_scores_kmeans_subset0_s1 = []

# Evaluación de K-Means con diferentes números de clusters
for n_clusters in range_clusters:
    kmeans_labels = KMeans(n_clusters=n_clusters, random_state=42)
    cluster_labels = kmeans_labels.
 ↪fit_predict(componentes_principales_subset0_s1)
    silhouette_avg = silhouette_score(componentes_principales_subset0_s1,␣
 ↪cluster_labels)
    silhouette_scores_kmeans_subset0_s1.append(silhouette_avg)

    # Graficando los resultados de Silhouette Score para K-Means
plt.figure(figsize=(14, 7))
plt.subplot(1, 2, 1)
plt.plot(range_clusters, silhouette_scores_kmeans_subset0_s1, marker='o')
plt.title('Silhouette Score para K-Means')
plt.xlabel('Número de Clusters')
plt.ylabel('Silhouette Score')
```

[204]:
```
Text(0, 0.5, 'Silhouette Score')
```



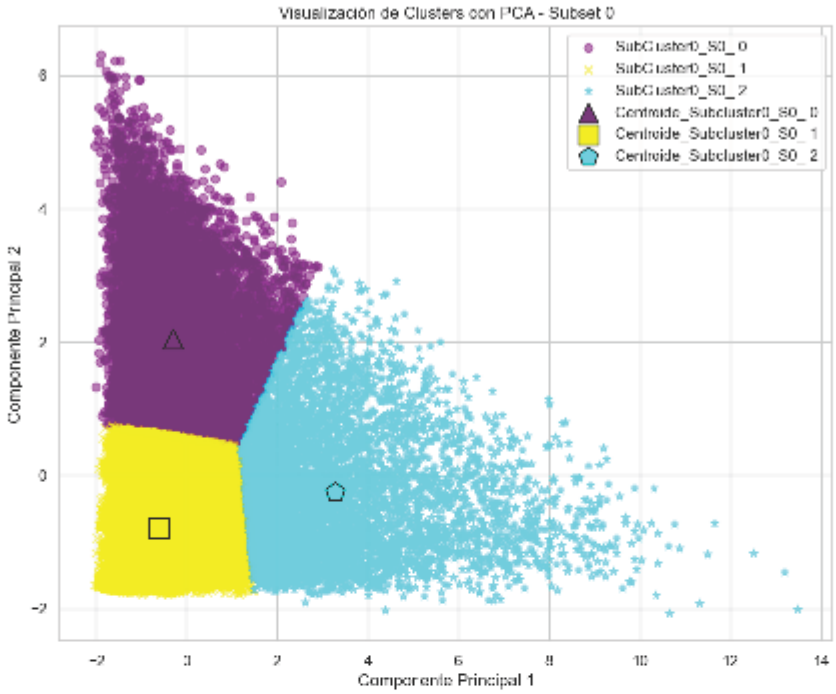

[211]:
```
#Ejecucion y prueba del cluster usando K-means para subset cluster 0
from sklearn.cluster import KMeans

# Ejecutando nuevamente el análisis de clustering con 3 clusters
kmeansSubset0_s1 = KMeans(n_clusters=3, random_state=42)
clustersSubset0_s1 = kmeansSubset0_s1.
 ↪fit_predict(componentes_principales_subset0_s1)
```

[212]:
```
#Evaluar la calidad del modelo
from sklearn.metrica import silhouette_score
```

```
# Calculando el Silhouette Score para evaluar la calidad del clustering
silhouette_avgSubset0_s1 = silhouette_score(componentes_principales_subset0_s1,␣
 ↪clustersSubset0_s1)
silhouette_avgSubset0_s1
```

[212]:
```
0.49777739679902294
```

[213]:
```
#GRAFICAR RESULTADO CLUSTER USANDO ANALISIS DE COMPONENTES PRINCIPALES
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# Convirtiendo la matriz dispersa en una matriz densa
#datos_normalizados_densos = datos_normalizados.toarray()

# Creando un dataframe para la visualización
df_visual_subset0_s1 = pd.DataFrame(data=componentes_principales_subset0_s1,␣
 ↪columns=['PC1', 'PC2'])
df_visual_subset0_s1['Cluster'] = clustersSubset0_s1
# Gráfico de los clusters
plt.figure(figsize=(10, 8))
#colors = ['red', 'green', 'blue', 'purple', 'orange']
colors = ['purple', 'yellow', 'cyan']
markers = ['o', 'x','*']  # Marcadores para los puntos de cada cluster
centroid_markers = ['^','s','p']  # Marcador para los centroides
#for i in range(5):
for i in range(3):
    plt.scatter(df_visual_subset0_s1[df_visual_subset0_s1['Cluster'] ==␣
 ↪i]['PC1'], df_visual_subset0_s1[df_visual_subset0_s1['Cluster'] == i]['PC2'],
                color=colors[i], label=f'SubCluster0_S0_ {i}', alpha=0.5,␣
 ↪marker=markers[i])


# Dibujar los centroides
for i, color in enumerate(colors):
    plt.scatter(kmeansSubset0_s1.cluster_centers_[i, 0], kmeansSubset0_s1.
 ↪cluster_centers_[i, 1],
                color=color, edgecolor='k', marker=centroid_markers[i], s=200,␣
 ↪label=f'Centroide_Subcluster0_S0_ {i}')


plt.title('Visualización de Clusters con PCA - Subset 0')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend()
plt.show()
```




[214]:
```
# Añadiendo la asignación de clusters al dataframe original para análisis
clientes_clustser_subset_s1_drop['Cluster'] = clustersSubset0_s1
cluster_description_kmeans_subset0_s1 = clientes_clustser_subset_s1_drop.
  ↪groupby('Cluster').count()
cluster_description_kmeans_subset0_s1
```

[214]:

| Cluster | ESTADO_CIVIL | CARGAS_FAMILIARES | CREDITOS_TOTAL | CREDITOS_CANCELADOS | \ |
|---|---|---|---|---|---|
| 0 | 11272 | 11272 | 11272 | 11272 | |
| 1 | 26927 | 26927 | 26927 | 26927 | |
| 2 | 6189 | 6189 | 6189 | 6189 | |

| Cluster | DIAS_MORA_PROMEDIO | MONTO_PROMEDIO | TOTAL_INGRESO | EGRESO_SOCIO | \ |
|---|---|---|---|---|---|
| 0 | 11272 | 11272 | 11272 | 11272 | |
| 1 | 26927 | 26927 | 26927 | 26927 | |
| 2 | 6189 | 6189 | 6189 | 6189 | |

```
         FLUJO_CAJA  NIVEL_ESTUDIOS  SECTOR_DOMICILIO   EDAD  \
Cluster
0             11272           11272             11272  11272
1             26927           26927             26927  26927
2              6189            6189              6189   6189

         NIVEL_RIESGO_PREC  REGION
Cluster
0                    11272   11272
1                    26927   26927
2                     6189    6189
```

[215]:
```
cluster_description_kneans_subset0_s1_mean = clientes_clustser_subset_s1_drop.
  ↪groupby('Cluster').mean()
cluster_description_kneans_subset0_s1_mean
```

[215]:
```
         CARGAS_FAMILIARES  CREDITOS_TOTAL  CREDITOS_CANCELADOS  \
Cluster
0                 1.211231        4.686657             3.352555
1                 0.858172        1.539830             0.507483
2                 1.113104        2.081758             0.950557

         DIAS_MORA_PROMEDIO  MONTO_PROMEDIO  TOTAL_INGRESO  EGRESO_SOCIO  \
Cluster
0                  9.339957     7172.021415    1320.352415    857.352258
1                  6.059234     7400.490979    1036.686798    623.023328
2                  7.391663    20109.736088    4568.429792   2943.973225

          FLUJO_CAJA       EDAD
Cluster
0         463.000157  43.303229
1         413.663469  36.662086
2        1624.456566  40.808854
```

[216]:
```
# Escribir el DataFrame resultante a un archivo Excel
cluster_description_kneans_subset0_s1_mean.to_excel('/Users/produccion-ja/
  ↪Documents/MAESTRIA_PUCE/TITULACION/SEGMENTACION2.xlsx', engine='openpyxl')
```