Análisis de la utilidad potencial del mercado colombiano a través de modelos de segmentación y customer life value para una empresa originadora de créditos de libranza

Estudiante: Juan José González Cano
Director: Jorge Esteban Montoya Cano
Ms.C Ingeniería
Departamento de ingeniería
Jmonto73@eafit.edu.co
Co-Director: Natalia Ochoa
Head Data Science Global – Rappi
Ms.C Analítica Avanzada
nataliaochoar@gmail.com

Maestría en ciencia de datos y analítica

Universidad EAFIT

2022

Contenido

Agradecimientos

Primero a Avista Colombia, que me permitió vivir esta fase de aprendizaje y crecimiento personal, académico y laboral a través del desarrollo de este trabajo dentro de la compañía.

Igualmente, un gran agradecimiento a mis padres por su apoyo en este camino y a mis mentores en este proyecto, Jorge Esteban Montoya y Natalia Ochoa, por su paciencia y dedicación para conseguir excelentes resultados en este proyecto.

Finalmente, gracias a la Universidad EAFIT por esta faceta de nuevos aprendizajes.

Resumen

En la actualidad las compañías definen su mercado objetivo para tener un mayor foco en ciertos individuos y grupos de la población, sin embargo, no logran entender a profundidad cual es el beneficio económico futuro que representan estos nichos de mercado, para entender si su modelo de negocio es atractivo desde el punto de vista financiero.

Este proyecto está enfocado directamente en el sector financiero Colombiano, buscando realizar una contribución directa a la forma en la cual las empresas de este sector analizan y definen el potencial económico de su mercado objetivo, mediante el uso de herramientas analíticas y financieras como modelos de segmentación y análisis del *Customer Life Value*, llegando a dar como resultado, el valor que posiblemente puede representar cada nicho en utilidad para la compañía, permitiéndole trazar una estrategia de negocio que le asegure la sostenibilidad en el tiempo y en el mercado.

Gracias a las capacidades integrales del equipo del proyecto, se usarán técnicas de segmentación que permitan soportar diferentes tipos de variables para encontrar grupos muy homogéneos en sus individuos, pero muy heterogéneos entre ellos y así llegar a saber que clústeres llevarán a la compañía a obtener un mayor beneficio.

Palabras clave: *k-means, segmentación, mercado, clv, customer life value, libranza, originación*

1. Descripción del proyecto

1.1. Planteamiento del problema

Tradicionalmente, las entidades financieras realizan análisis de su cartera mediante modelos de predicción y clasificación. Sin embargo, al momento de ganar participación en el mercado no se llevan a cabo análisis profundos que permitan a las compañías entender como enfocar sus recursos para lograr conseguir una mayor utilidad a partir de sus nichos de mercado, incluso teniendo información estadística de cómo están conformado su mercado actual alrededor del territorio nacional y cómo se comporta este según la información interna de la entidad (Marisa et al., 2019).

Es una necesidad latente para las compañías que ofrecen productos financieros el poder entender cómo se encuentra el mercado, y así poder trazar estrategias analizando la calidad de los nichos de mercado basándose en modelos que les permitan segmentar y predecir el CLV (*Customer Life Value*) para lograr maximizar su beneficio a partir de este análisis (Hassan, 2018).

1.2. Justificación

La inclusión financiera y la bancarización son temas que a través del tiempo han adquirido un alto nivel de interés, tanto para los gobiernos como para las entidades financieras, ya que una población bancarizada se traduce en un población que se interesa y se educa al respecto, y que además representa un mercado potencial que permite dar tanto a las instituciones como a las personas un crecimiento sano a través de productos financieros que se adaptan a ellos y satisfacen sus necesidades (Banco Mundial, 2018).

Aunque la bancarización es un tema muy relevante en esta era, la banca tradicional ha empezado a perder relevancia en el mercado y a dejar de ser el único canal a través del

cual las personas ingresan al mundo financiero, es por esto por lo que se ha creado un nuevo movimiento, en donde han surgido *Startups* enfocadas en productos financieros, mejor llamadas Fintech o compañías de financiación alternativa.

Estas compañías ofrecen soluciones financieras disruptivas al mercado y que permiten a la población bancarizar sin tener que acudir a las instituciones financieras tradicionales, ya que mediante el uso de la analítica para el control del riesgo pueden entender mejor al cliente y cubrir segmentos nuevos de mercado que las instituciones tradicionales han dejado desatentida como lo corroboró la gran firma Ernst & Young, que después de encuestar a 55 mil personas en 32 países, concluyeron que el 40% de los encuestados accedió al mundo financiero mediante una Fintech (Ionut, 2018).

A pesar de empezar a tener un papel relevante en la bancarización, el 70% de las Startups, categoría donde se encuentran ubicadas las *Fintech*, fracasan en los primeros 20 meses de operación (Bernal Salazar, 2019), debido a la baja capacidad para la generación de ingresos y fallas en la ejecución de su plan de negocios (Zuleta Acevedo, 2019).

Según IdeaLab, una importante aceleradora de Startups norteamericana para lograr establecerse en el mercado, es vital construir un modelo de negocio en donde el mercado objetivo represente una gran fuente de utilidad para la compañía (Schroeder, 2019), sin embargo, actualmente las Fintech únicamente usan herramientas analíticas para la gestión del riesgo y el mercadeo pero no aplican la ciencia de datos para encontrar el potencial de beneficio económico del mercado y poder llegar a maximizar este.

En este trabajo se busca llegar a entender cuál es el potencial del mercado en términos de beneficio económico para la compañía mediante el uso de herramientas analíticas, por esto la segmentación de clientes es una pieza clave y con gran relevancia ya que dentro del

mercado y la cartera de clientes existen diferentes grupos con características particulares y mediante el análisis de estos se puede llegar entender de forma mucho más asertiva el comportamiento de los clientes (Ionut, 2018).

Lograr construir un modelo de segmentación de calidad logra que una compañía financiera de cualquier naturaleza encuentre nichos dentro de sus clientes, este entendimiento permite medir el valor de cada uno de los nichos representa para la empresa en el momento actual y estimar el valor futuro de los mismos a través de modelos que permitan predecir el *Customer Life Value* (CLV) llevando a la empresa a maximizar el beneficio que obtiene de sus clientes, entendiendo cómo se comporta cada segmento a través de las variables estudiadas y aprovechando la relación con estos al máximo generando mayores utilidades a partir del uso de modelos de Segmentación por beneficios (Kahreh et al., 2014).

Finalmente, el uso de un modelo de segmentación que se alimenta de información interna y externa permite que este se esté actualizando bajo la periodicidad deseada, para que la empresa pueda visualizar cambios en el mercado y así, cambiar de estrategia buscando siempre la mayor utilidad posible y su permanencia en el tiempo.

1.3. Objetivos

1.3.1. Objetivo general

Encontrar el mercado que representa la mayor utilidad para la compañía mediante el análisis del mercado basado en segmentación de la cartera propia.

1.3.2. Objetivos específicos

- Diseñar un modelo de segmentación de clientes capaz de segmentar variables continuas y categóricas.

- Calcular el *Customer Life Value* para cada grupo de clientes encontrado segmentando adicionalmente por su zona geográfica.
- Encontrar el potencial de mercado usando información poblacional del DANE e información de la cartera de Avista encontrada bajo el modelo de segmentación.

1.4. Estado del arte y marco teórico

1.4.1. Machine Learning – Aprendizaje de Máquinas

Para entender el concepto de machine learning o aprendizaje de máquinas primero se debe ir al concepto básico de que se entiende por aprendizaje, aprendizaje es el proceso de convertir la experiencia en conocimiento por lo que el aprendizaje de máquinas el proceso en el que se logra que una máquina o computador pueda adquirir conocimiento a través del estudio de conjuntos de datos para que mediante una serie de entradas pueda generar una salida la cual por lo general es aprender a realizar una acción, sea cumpliendo una tarea, identificando un patrón, realizando una clasificación, entre otras (Janiesch et al., 2021)

Las máquinas no tienen capacidades humanas como el sentido común, por lo que su aprendizaje es diferente y siempre deben ser entrenadas aprendiendo bajo resultados obtenidos en la realidad para que así, estas pueden llegar a cumplir finalmente con la tarea que se requiere (Chang et al., 2018)

Dentro del aprendizaje de máquinas existen un sin número de técnicas con diferentes propósitos y metodologías de acuerdo con el objetivo que se tenga y el tipo de información con la cual se esté trabajando, debido a que el enfoque del trabajo es la segmentación, se hablará acerca de este campo.

Según el portal KDnuggets, las 5 técnicas más importantes al momento de realizar segmentación de datos son k-Means, mean-shift clustering, DBSCAN, Agrupación por maximización de expectativas mediante modelos de mezclas gaussianas y agrupación jerárquica aglomerativa, la definición de qué técnica usar va a depender mucho de los datos que se tengan, puesto que cada una puede tener mejores funcionamientos de acuerdo con lo mencionado anteriormente (Seif, 2018).

Reafirmando lo mencionado por KDnuggets, los autores Sehgal & Garg (Sehgal & Garg, 2014)dividen las técnicas actuales de clustering en 5 grupos principales:

- Basado en particionamiento, donde se encuentra K-Means
- Basado en jerarquía
- Basado en densidad, donde se encuentra DBSCAN
- Basados en red
- Basados en modelos

Los autores concluyen que la selección del modelo va en función del problema y los datos sobre los cuales se realiza la segmentación y entregan, por esto es importante conocer cuáles son las técnicas de segmentación para lograr construir un modelo de clustering que se ajuste de la mejor forma al problema abordado.

1.4.2.Segmentación

La segmentación es la técnica de machine learning que busca encontrar las similitudes entre observaciones al agrupar de acuerdo con una medición de distancias que permite encontrar los grupos más homogéneos en función de sus características y por el otro lado, identificar las disimilitudes – heterogeneidad- entre los grupos para hacer la distinción entre los factores. Generalmente se hace segmentación de clientes para

que las entidades puedan establecer segmentos dentro de su portafolio de clientes y pueda diseñar estrategias enfocadas en cada grupo de acuerdo con las características descubiertas (Dawood et al., 2019).

Dentro el ecosistema financiero y más aún dentro del entorno de las *Fintech*, la segmentación de clientes es una metodología que cobra vital importancia puesto que uno de los principales retos que tienen las entidades financieras es entender el comportamiento de sus clientes para maximizar su satisfacción y del mismo modo maximizar su valor (Djurisic et al., 2020).

La maximización de la utilidad de una compañía empieza a través de una buena segmentación de clientes, esto les permite a las empresas diseñar estrategias y modelos de negocio eficientes (Smeureanu et al., 2013) ya que tienen un pleno entendimiento de su cliente desde perspectivas comportamentales y sociodemográficas pudiendo comprender las necesidades de estos y atacando grupos y nichos de mercado específicos que son definidos a través de modelos de segmentación.

Tanto en la industria financiera como en el mundo académico se han realizado un sin número de estudios en cuanto a técnicas que permitan realizar segmentaciones de clientes cada vez más acertadas y eficientes que ayuden a definir grupos más homogéneos entre sus individuos, pero más heterogéneos con el resto de los grupos encontrados.

La segmentación se ha trabajado desde diferentes puntos de vista apuntando a diferentes propósitos, de acuerdo a Sharahi y Aligholi realizaron segmentación de clientes bancarios para identificar el nivel de lealtad a través de segmentación por K-

Means (Sharahi & Aligholi, 2015), de igual forma lo hicieron los autores Palaniappan et al. en 2017 realizaron modelos de segmentación a través de árboles de decisión para buscar un perfilamiento más ajustado buscando clientes que tuvieran una mayor probabilidad de tener una relación con la entidad a largo plazo (Palaniappan et al., 2017).

Las técnicas de clustering también han sido usadas para identificar riesgo crediticio buscando predecir la tasa de default en clientes de tarjeta de crédito como lo trabajaron los autores Yang y Zhang en 2018 a través de modelos de Light GBM (Yang & Zhang, 2018)y Niloy en 2018 a través de clasificadores bayesianos y árboles de decisión (Niloy, 2018).

Finalmente, para el área de interés de este trabajo se han realizado múltiples estudios sobre cómo usar la segmentación en la predicción del CLV como lo realizaron los autores Marisa et al. en 2019 a través de K-Means buscando maximizar el beneficio de clientes para pequeñas y medianas empresas (Marisa et al., 2019) o como el autor Hassan que a través de una segmentación efectiva de clientes a través de Naive Bayes definen como lograr estrategias de retención y atracción de clientes (Hassan, 2018).

1.4.3. Customer Life Value

Customer Life Value o valor de vida del cliente es el nombre que se le da al estudio de la suma de todas las utilidades generadas por una compañía proveniente de la relación con un cliente a lo largo del tiempo, normalmente se define un horizonte de tiempo para realizar este estudio y determinar cuál es el valor de CLV (Kahreh et al., 2014).

A lo largo del tiempo se han definido diferentes modelos para la estimación de este valor, como lo mencionan los autores Yoseph y AlMalaily (Yoseph & AlMalaily, 2019)

y Qismat y Fenga (Qismat & Feng, 2020), quienes en sus trabajos exponen diferentes metodologías para llegar a este resultado, sin embargo, todos estos toman entradas de datos similares para realizar el cálculo, basados en data comportamental y transaccional del cliente, y bajo los diversos modelos estima cuánto es el valor que ese individuo va a representar para la organización en un período de tiempo determinado.

El valor de vida del cliente es comúnmente usado combinándolo con modelos de segmentación como lo mencionan los autores Kahreh et al. (Kahreh et al., 2014) , pues para las entidades que poseen un alto número de clientes, estimar el valor de cada individuo se convertiría en una tarea muy tediosa. Sin embargo, cuando una compañía es capaz de segmentar a sus clientes definiendo diferentes clústeres o segmentos, puede estimar el valor futuro de los mismos bajo modelos de CLV, ya que con certeza la compañía sabe que todos los individuos ubicados en un grupo de características homogéneas (Hassan, 2018).

En ocasiones como lo hicieron los autores Kahreh et al., se realiza una segmentación por beneficios tomando como variables para la agrupación las que influyen en el cálculo del CLV de cada cliente, logrando agrupar clientes según la utilidad futura que van a proveer y no por características sociodemográficas como tradicionalmente se hace en estos modelos (Kahreh et al., 2014).

Otros autores han realizado modelos de segmentación basados en el mismo principio, Marissa et al. realizaron segmentación de clientes bajo K-Means buscando clústeres bajo el modelo de CLV LRFM (Lengh, Recency, Frequency, Monetary), usando el valor estas características en la segmentación (Marisa et al., 2019) o como Yoseph y AlMalaily que a través del uso de regresiones para predecir la tendencia de compra de

los clientes y mediante segmentación por Fuzyy C-Means hacen un análisis del comportamiento de compra de los consumidores llegando a examinar las puntuaciones mejores bajo el cálculo de CLV para conocer los clientes realmente rentables (Yoseph & AlMalaily, 2019).

1.4.4. Fintech y banca tradicional

Mundialmente ha surgido un fenómeno en donde empresas alrededor del mundo han tomado servicios provistos por compañías tradicionales y los han revolucionado usando la tecnología como su pilar fundamental, tomando las fallas en los servicios tradicionales y aprovechando estas para entrar al mercado y capturar a los clientes. Basándose en este mismo principio nacen las Fintech o empresas de financiación alternativa, estas son compañías innovadoras basadas en tecnología y en modelos de negocio que acompañan estos servicios financieros disruptivos (Mention, 2019).

En otros términos, cualquier compañía que busque formas disruptivas de mejorar, prestar y usar los servicios financieros se puede denominar una *Fintech*, puesto que, a través de estos 3 principios, estas empresas cubren y mejoran el servicio de la banca tradicional alivianando procesos y llegando a ser compañías con una cultura más juvenil y menos jerárquica que les permite tener diferentes percepciones del riesgo y cambios importantes en su cultura organizacional (Mention, 2019).

Las *Fintech* se encuentran revolucionando la industria financiera y entrando a abrir mercado en donde la banca tradicional ha dejado población sin atención o con atención muy limitada, según la encuesta del World Fintech Report de 2017 el 50,2% de los individuos encuestados aseguró haber realizado transacciones con *Fintech*, convirtiéndose en la principal amenaza para las instituciones financieras tradicionales.

Así como sucedió en el caso de la banca brasileña, donde NuBank a través de la optimización de procesos, el uso de la tecnología y la ciencia de datos logró posicionarse como un jugador poderoso en el mercado de las tarjetas de crédito logrando reducir a más de la mitad el interés que la banca tradicional cobraba a sus clientes (Startupeable, 2021).

Estas compañías se encuentran prestando servicios como pagos móviles, transferencias bancarias, préstamos entre particulares (peer-to-peer lending) y financiación colectiva (crowdfunding) a través de tecnologías como blockchain, robotización de procesos y el uso de procesos de analítica e inteligencia artificial logrando convertir el mercado en un entorno mucho más dinámico por la velocidad en la que las innovaciones y los nuevos servicios financieros disruptivos son llevados al mercado (Goldstein et al., 2019).

1.4.5. Definición de un producto financiero

El crédito es una herramienta financiera que permite a un acreedor (individuo o entidad dueña del dinero) prestarle dinero a un deudor (persona quien recibe el dinero), un monto determinado con unas condiciones pactadas entre ambas partes, entre las cuales se definen el plazo de pago, la tasa de interés y la periodicidad de pago como condiciones mínimas necesarias, sin embargo, de acuerdo al tipo de tipo de crédito, la normatividad vigente y las condiciones del mercado, el acreedor puede exigir otros cobros adicionales con miras a cubrir el riesgo de no pago por parte del deudor (Asobancaria, n.d.).

Avista al ser una compañía dedicada a la originación de créditos de libranza, posee ciertas condiciones en sus obligaciones financieras, adicionales a las anteriormente

mencionadas, con el objetivo de poseer una cobertura sobre los riesgos inherentes a sus clientes. Las condiciones con las cuales cuenta el crédito de Avista son las siguientes:

- Plazo: Período de tiempo en la que la persona debe pagar la totalidad del capital amortizado.
- Tasa de interés: Valor porcentual que se usa para calcular el monto de interés a pagar por parte del cliente en su cuota mensual.
- Seguro de vida: Cobro adicional asociado a la cobertura provista por una entidad aseguradora de riesgos, el cual, cubre la totalidad del saldo de la obligación en caso de que el titular fallezca.
- Fianza: Cobro adicional asociado a la cobertura provista por una entidad avalista. Este cobro hace las veces de codeudor para el titular de la obligación, pues pagando este cargo adicional, el cliente no requerirá ningún tipo de codeudor y Avista como acreedor, cubre el riesgo de deterioro de cartera mediante la acumulación de un saldo disponible en la entidad avalista.

1.5. Metodología

Para lograr determinar el potencial de mercado caracterizándolo según el comportamiento que ha tenido la cartera histórica de la compañía, se deben seguir los pasos expuestos a continuación para que de esta manera se llegue a la construcción de los modelos y así a la estimación del potencial de mercado para lograr la máxima utilidad posible.

Con el propósito de seguir una metodología que permita un entendimiento del problema y una solución que realmente sea satisfactoria se trabajará bajo CRISP-DM (Cross Industry

Standard Process for Data Mining), que permite asignar tareas por fases del proyecto para lograr solucionar la situación expuesta por completo.

- **Fase 1:** Definición de necesidades del cliente (comprensión del negocio)

  En la sección 1. Planteamiento del problema y sección 2. Justificación se expone la compresión del negocio y la definición de la necesidad.

- **Fase 2 y 3:** Estudio y comprensión de los datos - Análisis de los datos y selección de características

  A partir de las fuentes de datos identificadas se podrán realizar transformaciones y creación de variables cuando se considerado necesario, además se realizará un tratamiento de datos raros que garantice la integralidad en el funcionamiento de los modelos, pues es de conocimiento que la calidad de los modelos depende de la data con la que sean entrenados y validados y las transformaciones necesarias a esta para conseguir un funcionamiento óptimo. Las fuentes de datos y de información usadas son:

  - División política y administrativa (DVIPOLA) - DANE
  - Proyecciones municipales DANE 2018 – 2030
  - Datos de originación de créditos – Avista
  - Datos sociodemográficos y financieros de clientes – Avista
  - Tabla de pagos de los créditos – Avista

- **Fase 4:** Modelado

  Para la consolidación de los datos dentro del modelo se propone hacerlo mediante un modelo relacional a partir de llaves en cada conjunto de datos que permita realizar la

unión y conglomeración de la información, estas llaves están compuestas por códigos únicos y oficiales provenientes de fuentes de información como el DANE.

- Segmentación de datos

  Tomando los datos de la cartera de la compañía, se iniciará realizando modelos de segmentación usando algoritmos como PCA k-means y k-medioides debido a la presencia de variables numéricas y categóricas, esto permitirá agrupar los individuos por perfiles similares para posteriormente proceder a calcular su CLV en un escenario de tiempo definido, es por esto que la agrupación no debe ser realizada únicamente por variables sociodemográficas sino también por las características usadas en el cálculo del CLV para que luego de definir los grupos de clientes los cálculos mencionados anteriormente puedan ser realizados.

  Para el ejercicio de segmentación se crearán diferentes clústeres y cada uno debe ser perfilado por zona geográfica, por esto, en diferentes zonas geográficas pueden llegar a existir el mismo clúster. Lo anterior se debe a que, en el último paso del trabajo, al proyectar el potencial de mercado con los cálculos previamente realizados, este debe ser realizado según la zona geográfica puesto que nuestras fuentes de datos así se encuentran y de igual manera los resultados arrojados deben ser localizados por cada municipio.

- Cálculo del Customer Lifetime Value para cada clúster

  Una vez se obtengan los clústeres y según las características de cada grupo, se procederá a realizar el cálculo del CLV en un horizonte de tiempo de 5 años para cada individuo, con este cálculo podremos conocer y definir el CLV de cada clúster en cada localidad (municipio) y definir el CLV del municipio bajo un

promedio ponderado, encontrando el valor de esta municipalidad sin perder de vista los diferentes perfiles que contiene cada zona geográfica.

Definición del potencial de mercado basado en CLV

Una vez hallado el CLV ponderado por zona geográfica en el paso anterior, este será proyectado sobre el tamaño de mercado estimado bajo cifras poblacionales provistas por el DANE, encontrando así el potencial de utilidad que cada zona geográfica representa para la compañía.

- **Fase 5:** Evaluación

  Para la fase de evaluación se usarán las siguientes medidas

  - Silhouette Coefficient (separación de clústeres)
  - Elbow analysis (definición de número de clústeres)

- **Fase 6:** Despliegue

  El modelo será construido de manera local, pero su funcionamiento será en la nube a partir de las soluciones disponibles por AWS

1.6. Productos esperados

Se espera entregar un modelo de segmentación de clientes que sea capaz de realizar una agrupación usando variables continuas y categóricas para llegar a obtener segmentos de clientes, usando cualquier dato que pueda agregar asertividad en el modelo.

Este modelo será capaz de consumir fuentes de datos directamente desde las bases de datos de la compañía para que su calibración sea automática, adicionalmente un modelo que permita calcular el CLV de los diferentes clústeres encontrados a través segmentación y proyectar estos valores sobre el tamaño de mercado del período de tiempo evaluado, para que en la periodicidad establecida por los usuarios del modelo, se entregue el

potencial de mercado de cada zona geográfica y así la gerencia comercial pueda tomar decisiones en cuanto al enfoque de sus esfuerzos comerciales, buscando siempre aprovechar los mercados apuntando a obtener la mayor utilidad posible para la compañía mediante la consecución de nuevos clientes o la retención de clientes actuales.

El modelo deberá ser corrido con cierta periodicidad que será establecida por la compañía, esto debido a que los clústeres encontrados irán presentando variaciones a través del tiempo debido a los cambios naturales por los cuales pasan los mercados y el cambio en la composición y características de la cartera de clientes de la compañía.

1.7. Plan de gestión de datos

Las fuentes de datos para elaborar el trabajo tienen orígenes públicos, como los datos poblacionales y los factores para estimación del mercado potencial, estas bases podrán ser expuestas y compartidas en el trabajo puesto que su dominio es público y se encuentran publicadas en el sitio oficial del DANE.

En cuanto a los datos referentes a información de la compañía, tanto de clientes como información transaccional, es de vital importancia que sea totalmente confidencial puesto que existe información sensible de clientes, tanto personal como financiera e información sobre la operación y recaudo de la compañía que no puede ser expuesta a terceros. Para esta data se podrá compartir las transformaciones realizadas, variables usadas y su definición, parámetros usados dentro de los modelos y resultados de estos más no el detalle de la data con la que se realice entrenamiento y validación.

1.8. Aspectos éticos

Los datos serán usados únicamente con el fin de diseñar un modelo que permita a la organización asociada al trabajo maximizar su utilidad por medio de los procedimientos

descritos en el documento. Todos los datos serán usados con el consentimiento de la compañía propietaria y los que sean de dominio público no necesitan consentimiento de uso por parte del autor.

2. Desarrollo del proyecto

2.1. Contexto e historia general de la compañía

Fundada en 2015, Avista inicio sus operaciones en el 2019. Desde entonces, la visión de la organización ha sido construir una Fintech (compañía de tecnología financiera) enfocada en desarrollar productos innovadores y simples alrededor de la pensión por medio de productos digitales.

Avista con el propósito de escalar rápidamente su operación, ha desarrollado un modelo de negocio enfocado principalmente en la originación, venta y administración de la cartera de créditos de libranza enfocados en el mercado de pensionados colombianos.

Operar en el mercado de libranza colombiano genera una gran ventaja para la compañía, puesto que este tipo de instrumento financiero, se encuentra amparado y regulado por la Ley 1527 de 2012, en la cual el acreedor tiene el derecho a matricular un descuento por el valor de la cuota del crédito, en la nómina o pensión del deudor, la cual será pagada por la entidad empleadora o fondo de pensión asociado, cada mes, lo que se traduce en que Avista está expuesto a un riesgo mínimo de recaudo de sus obligaciones.

Dentro del marco regulatorio sobre los créditos de libranza, la Superintendencia Financiera de Colombia expresa lo siguiente:

"La Superintendencia Financiera en Concepto 2008038709-002 del 7 de julio de 2008, expresa que la libranza es un "mecanismo de recaudo de cartera" mediante el cual un deudor autoriza a su empleador para que realice un descuento de su salario para atender obligaciones adquiridas con un tercero" (Villa Patiño, 2019)

Además, dentro del mismo marco regulatorio, la entidad expresa que los siguientes son los deberes de una entidad pagadora ante los créditos de esta denominación (Villa Patiño, 2019):

- Girar de manera directa los recursos a la entidad operadora de libranza a nombre del beneficiario.
- No negarse injustificadamente a la suscripción del acuerdo entre operador y beneficiario. Sin embargo, dicha suscripción podrá negarse una vez evaluada la capacidad de descuento del beneficiario.
- Efectuar las libranzas y trasladar dichas cuotas a las entidades operadoras dentro de los tres (3) días hábiles siguientes de haber efectuado el abono al asalariado, contratista, afiliado, asociado o pensionado, en el mismo orden cronológico en que haya recibido la libranza.
- Verificar que la entidad operadora o administradora se encuentra inscrita en el Registro Único Nacional de Entidades Operadores de Libranza.
- Pagar como sanción pecuniaria el doble del valor total descontado por la libranza, en caso de cobrar o descontar cuota de administración o comisión por realizar el descuento o el giro de los recursos

Tomando en cuenta lo expuesto anteriormente, el crédito de libranza, aunque es un crédito de consumo o de libre inversión, posee un funcionamiento totalmente diferente a los créditos con esta misma denominación pero con diferentes formas de recaudo, ya que aquí, el riesgo no depende de la calidad del cliente como pagador, lo que conlleva a que el análisis de crédito sea completamente diferente y existan, potencialmente, clientes que desde una perspectiva de calidad como deudor, sean malos, pero que para ser un cliente de un crédito de libranza, cuenten con las características necesarias.

2.2. Estudio y análisis de los datos

2.2.1. Fuentes de datos utilizadas

Por la naturaleza del problema abordado y para el desarrollo de la solución se usaron principalmente fuentes de datos internas de la compañía, las cuales contienen información de los clientes y las obligaciones crediticias de los mismos, y como fuentes externas se usaron cifras poblacionales obtenidas directamente de la información publicada por El Departamento Administrativo Nacional de Estadística – DANE – en su página web oficial.

Para la extracción de los datos de la compañía se realizaron una serie de consultas directas a la base de datos de Avista a través de SQL, a partir de las cuales se obtuvieron tanto tablas completas, como consultas en donde se apuntaba a resumir información específica, tanto de los clientes como de sus obligaciones, para facilitar la carga y el tratamiento mediante la reducción del volumen de información de las tablas originales.

Es importante resaltar que la base de datos con la cual opera Avista es heredada de una compañía anterior, por ende, en las tablas se encontraran un número importante de variables sin descripción y sin conocimiento de su función, ya que pertenecían a productos propios de la compañía original. Las siguientes fueron las fuentes de datos obtenidas de la base de datos corporativa:

a. Personas: Para el caso de la tabla *personas* se realizó una extracción completa de la misma, en la cual se encontraron 79.804 registros, sin realizar ningún tipo de depuración previa. En esta tabla se encuentra toda la información referente a la identidad, datos de contacto y variables sociodemográficas del cliente. El detalle de data contenida en esta tabla es la siguiente:

Tabla 1: Campos de la tabla personas

Variable	Descripción	Variable	Descripción
TIPO_DOC	documento (Identificación)	AFILIACION	Numero de afiliación del ISS
CEDULA	Cedula del cliente	CEDULA_CON	Cedula conyugue
NOMBRE1	Primer Nombre del cliente	DIRECCION2	Segunda dirección de localización
NOMBRE2	Segundo Nombre del cliente	SECCIONAL	Seccional del iss
APELLIDO1	Primer Apellido del cliente	COMISION	Código de comisión
APELLIDO2	Segundo Apellido del cliente	ACTIVIDDAD	Código de la actividad del cliente
CIUDADEXP	Ciudad de expedición del documento	CODE_BARRI	Código del barrio
FECHAEXP	Fecha de expedición	EMAILPERS	correo electrónico persona
CIUDADNAC	Código ciudad de nacimiento	CODPOSTAL	código postal
FECHANAC	fecha de nacimiento	INGRESOSCN	Sin descripción
COD_ESTCIV	código estado civil	OCUPACIOCN	código ocupación

SEXO	Genero del cliente	COD_MILITA	numero de la libreta militar
COD_NIVEDU	Código del nivel educativo	COD_BASE	Código empresa a la que pertenece
COD_PROFE	Código de la profesión del cliente	ESTRATO	código nivel estrato
PERS_CARGO	cantidad de personas a cargo	SINCRO	código de sincronización
DIRECCION	Dirección residencia del cliente	ANTIG_VIVIEN	Sin descripción
TELEFONO	Número de teléfono fijo del cliente	TIP_TELF	Sin descripción
CELULAR	Número de teléfono celular	NUMID	Sin descripción
CIUDAD	Código de la ciudad del cliente	LATITUD	Coordenada Geolocalización Latitud
COD_BARRIO	Código del barrio de residencia	LONGITUD	Coordenada Geolocalización Longitud
COD_VIVIEN	Código del tipo de vivienda	FECHA_CREA	Sin descripción
VLRPROXCA	Sin descripción	ESTATURA	Sin descripción
TIEMPO	Sin descripción	PESO	Sin descripción
ENVIOCORRE	Código de envió de correo		

b. Centrales: La tabla de la base de datos que contiene toda la información de centrales, posee uno de los volúmenes de datos más altos de toda la base de datos de la compañía, conteniendo más de 8 millones de registros. El alto volumen de registros de esta tabla se genera debido a que cada cliente es consultado en el momento de su solicitud de crédito y el resultado de esta consulta es almacenado en esta tabla línea por línea, teniendo conocimiento de que cada obligación que la persona ha adquirido durante su vida se convierte en un registro de la tabla mencionada.

Con la finalidad de reducir el volumen de información con la cual se espera trabajar, se realizó una consulta que permitiera extraer la información deseada de manera

resumida. La forma en la cual se resumió la información fue diseñada tomando en cuenta las variables de centrales de información usadas en el estudio de la solicitud de crédito y filtrando la solicitud de acuerdo con su fecha, tomando únicamente la más reciente.

La consulta diseñada para resumir la información de burós de crédito de cada cliente fue la siguiente:

```
SELECT P.NUMERO SOLICITUD, P.CLTNUMCED CEDULA,
to_char(P.FCHSOL, 'DD/MM/YYYY') FECHA_CONSULTA,
SUM(DT.SALDO_ACT) SALDO_VIGENTE, SUM(DT.SALDO_MOR)
SALDO_MORA, SUM(DT.CUOTA) CUOTAS,
                    sum(case when DT.saldo_act > 0 or
DT.estado = '01' or DT.estado = '02' then 1 else 0 end)
Obl_activas,
                    sum(case when DT.saldo_act > 0 or
DT.estado = '01' then 1 else 0 end) Obl_aldia,
                    sum(case when DT.saldo_act > 0 or
DT.estado = '02' then 1 else 0 end) Obl_mora,
                    sum(case when DT.saldo_act > 0 or
DT.estado = '08' then 1 else 0 end) Obl_casitigadas,
                    sum(case when DT.saldo_act > 0 or
DT.estado = '01' then DT.saldo_act else 0 end) Saldo_aldia,
                    sum(case when DT.saldo_act > 0 or
DT.estado = '02' then DT.saldo_act else 0 end) Saldo_mora,
                    sum(case when DT.tipo_cta = 'LBZ' then
DT.saldo_act else 0 end) Saldo_LBZ,
                    sum(case when DT.tipo_cta = 'LBZ' then
DT.saldo_mor else 0 end) SaldoMora_LBZ,
```

```
                    sum(case when DT.tipo_cta = 'CAV' or
DT.tipo_cta = 'CAU' then DT.saldo_act else 0 end) Saldo_COP,
                    sum(case when DT.tipo_cta = 'CAV' or
DT.tipo_cta = 'CAU' then DT.saldo_mor else 0 end)
SaldoMOR_COP
FROM PEDIDO P
INNER JOIN DATACAB DC ON P.CLTNUMCED = DC.CEDULA AND
P.FECHACON = DC.FECHACON
INNER JOIN DATADET DT ON DC.CEDULA = DT.CEDULA AND
DC.FECHACON = DT.FECHACON
INNER JOIN (SELECT CLTNUMCED, MAX(FCHSOL) FECHA_SOL
            FROM PEDIDO
            WHERE BASE = '77'
            GROUP BY CLTNUMCED) PM ON P.CLTNUMCED =
PM.CLTNUMCED AND P.FCHSOL = PM.FECHA_SOL
WHERE P.BASE = '77'
GROUP BY P.NUMERO, P.CLTNUMCED, P.FCHSOL;
```

Una vez ejecutada la consulta, la tabla resultante contenía 74.845 registros, con las siguientes variables:

Tabla 2: Campos de la tabla centrales

Variable	**Descripción**
SOLICITUD	Número de solicitud de crédito
CEDULA	Cédula del cliente
FECHA_CONSULTA	Fecha de consulta a buró
SALDO_VIGENTE	Saldo total de obligaciones vigentes
SALDO_MORA	Saldo de obligaciones en mora
CUOTAS	Sumatoria de las cuotas de obligaciones vigentes
OBL_ACTIVAS	Número de obligaciones vigentes
OBL_CASTIGADAS	Número de obligaciones castigadas
SALDO_LBZ	Saldo vigente de obligaciones de libranza

SALDOMORA_LBZ	Saldo en mora de obligaciones de libranza
SALDO_COP	Saldo vigente de obligaciones de cooperativas de crédito y ahorro
SALDOMOR_COP	Saldo en mora obligaciones de cooperativas de crédito y ahorro

c. **Información financiera** En esta tabla se encuentra la información financiera del cliente en cada una de sus solicitudes, dado que esta tiene una alta probabilidad de cambiar en los diferentes momentos del tiempo. Con la información consignada en esta tabla, la compañía realiza todos los cálculos correspondientes a la capacidad de endeudamiento del cliente.

La tabla fue extraída en su totalidad, obteniendo 112.388 registros sin ningún tipo de depuración y las siguientes variables:

Tabla 3: Campos de la tabla información financiera

Variable	Descripción	Variable	Descripción
NUM_SOL	numero de la solicitud	MANRECPUB	código manejo de recursos públicos
CEDULA	cedula del cliente	COD_BASE	código empresa
SAL_BASE	valor salario básico del cliente	CEDULA_CAP	Sin descripción
OTROS_INGP	valor otros ingresos del cliente	TIPO	Sin descripción
TOTALINGRE	valor total de los ingresos del cliente	CUPO	Sin descripción
TOTALEGRE	valor total de los egresos del cliente	MANRECPUB_DESC	Sin descripción
CTAS_PREPA	valor compra de cartera y prepago a realizar	GRADPODPUB	Sin descripción
MONSALDES	valor monto libre de salario a respetar	GRADPODPUB_DESC	Sin descripción

MARGENSEG	valor margen de seguridad	RECONOPUB	Sin descripción
PAZYSALVO	valor del paz y salvo	RECONOPUB_DESC	Sin descripción
CAPACIDA	valor de las capacidades del cliente	INGRESOS_DESC	Sin descripción
OPINTERNAL	código operacional internacional	INGRESOSOT_VLR	Sin descripción
OPI_VLRMES	valor operación internacional	PARENTESCO	Sin descripción
OPI_TIPO	tipo operación internacional	OTROS_BENEFICIOS	Sin descripción
OPI_DESC	descripción tipo operación internacional	CAPA_NETA	Sin descripción
MONEDAEXTR	moneda extranjera	CAPA_TABLA	Sin descripción
MEXT_PAIS	nombre país de la moneda extranjera	TOTAL_DESCUENTOS	total de los descuentos de ley
MEXT_CIUD	nombre ciudad moneda extranjera	CUOTA_NO_CR	Sin descripción
MEXT_BANCO	nombre banco de la transacción	FACTA_A	Sin descripción
MEXT_CUENT	numero de la cuenta	FACTA_B	Sin descripción
MEXT_MONED	Sin descripción	FACTA_C	Sin descripción
ACTIVOS	valor activos	FACTA_D	Sin descripción
OTR_ACTIVO	valor otros activos	FACTA_E	Sin descripción
PASIVOS	valor pasivos	INGRESOS_ADI	Sin descripción
OTR_PASIVO	valor otros pasivos		

d. Estado de créditos: Esta tabla corresponde a la clasificación de créditos saldados, dado que esta información será relevante para la caracterización sobre si un cliente es bueno o malo para la compañía en términos de utilidad. Esta información corresponde a un segmento de una consulta establecida dentro de Avista.

El segmento de la consulta usado para extraer esta información es el siguiente:

```
CASE
     WHEN C.ESTADO_CRE = '2' AND TP.PAGO NOT IN (23,50) AND
NVL(TT.CREDITO_T,'0') = '0' THEN 'Prepaid'
     WHEN C.ESTADO_CRE = '2' AND TP.PAGO = 23 THEN 'Insurance
paid'
     WHEN C.ESTADO_CRE = '3' AND TP.PAGO = 50 THEN 'Collateral
paid'
     WHEN C.ESTADO_CRE = '0' AND NVL(MC.DIASMORA,0) = 0 THEN
'Current'
     WHEN C.ESTADO_CRE = '0' AND NVL(MC.DIASMORA,0) > 0 THEN
'Delinquent'
     WHEN C.ESTADO_CRE = '2' AND NVL(TT.CREDITO_T,'0') > '0'
THEN 'Refinanced'
     WHEN C.ESTADO_CRE = '2' THEN 'Paid Off' ELSE 'Revisar' END
LOAN_STATUS
```

El resultado de la consulta arrojó una tabla con 30.556 registros, únicamente con 3 variables como se expone a continuación:

Tabla 4: Campos de la tabla estado creditos

Variable	Definición
CREDITO	Número de crédito
FECHA_AP	Fecha de apertura del crédito
ESTADO	Estado

Los posibles estados y sus definiciones son las siguientes:

- Current: Créditos vigentes y al día

- Prepaid: Créditos pagados directamente por el cliente antes del plazo original de la obligación sin el objetivo de ser refinanciado con otro crédito de Avista.
- Insurance paid: Créditos pagados por la aseguradora de riesgos por el fallecimiento del titular de la obligación.
- Collateral paid: Créditos pagados por la entidad proveedora de la fianza de los créditos por el deterioro y no pago de la obligación por parte del titular.
- Refinanced: Créditos cerrados con motivo de retanqueo (Apertura de una nueva obligación de mayor valor que consolida el saldo la obligación cerrada más un monto adicional otorgado al cliente)

e. Solicitudes: En esta tabla se encuentra información asociada a la solicitud de crédito en sí. Fue extraída completa de la base de datos corporativa sin ningún tipo de tratamiento, por ello se obtuvieron 282.055 registros con las siguientes variables:

Tabla 5: Campos de la tabla solicitudes

Variable	Definición	Variable	Definición
NUMERO	Numero de Pedido	FECHAEXP	Sin descripción
FCHSOL	Fecha solicitud	ANTIG_VIVIEN	Sin descripción
BASE	cartera	ANTIG_LABOR	Sin descripción
CODOFIC	Código de oficina	TIPO_LIQ	Sin descripción
CLTNUMCED	numero de cedula titular	PORLINEA	Sin descripción
CLTCIUCED	ciudad de expedición de cedula titular	FONDOG	Sin descripción
CLTAPELL1	primer apellido titular	VLR_FONDOG	Sin descripción
CLTAPELL2	segundo apellido titular	COD_NIVRIE	Sin descripción

CLTNOMBR1	primer nombre titular	SCOREP	Sin descripción
CLTNOMBR2	segundo nombre titular	ASES_EXT	Sin descripción
CLTDIRRES	dirección de residencia titular	IMAGEN_PRO	Sin descripción
CLTCIURES	código ciudad de residencia titular	INGRESOS	Sin descripción
CLTBARRES	código barrio de residencia titular	TIPSOLICITA	Tipo de solicitante (Dependiente/Independientes)
CLTTELEFO	número telefónico fijo titular	COD_UNIVERS	Código de Universidad
CLTCELULA	numero celular titular	COD_UNIVCIU	Ciudad Universidad
ESTADCIVI	código estado civil titular	COD_UNIVPRG	Código programa Universitario
CLTEMAIL	correo electrónico titular	UNIVJORNADA	Jornada educativa (Diurna/Nocturna)
CNYNUMCED	numero de cedula conyugue	UNIV_ANOPER	Año y Periodo educativo a cursar
CNYCIUCED	ciudad de expedición de cedula conyugue	PARENTESCO	Parentesco del Estudiante con el Deudor Solidario
CNYAPELL1	primer apellido conyugue	PUESTOICFES	Puesto ICFES
CNYAPELL2	segundo apellido conyugue	COD_ALUMNO	Código estudiante Universidad
CNYNOMBR1	primer nombre conyugue	PROM_NOTAS	Promedio notas estudiantes antiguo
CNYNOMBR2	segundo nombre conyugue	TIPINSTBAS	Tipo Institución Básica Secundaria
CNYDIRRES	dirección de residencia conyugue	IMAGEN_PRO2	Sin descripción
CNYCIURES	código ciudad de residencia conyugue	IMAGEN_PRO3	Sin descripción

CNYBARRES	código barrio de residencia conyugue	EGRESOS	Sin descripción
CNYTELEFO	número telefónico fijo conyugue	CEDULA_ES	Sin descripción
CNYCELULA	numero celular conyugue	CEDULA_DS	Sin descripción
CNYEMAIL	correo electrónico conyugue	TIELOCAL	Sin descripción
NEGOC_NIT	nit del negocio	VLR_VEHICULO	Sin descripción
NEGOC_NOM	nombre del negocio	NUEVO	Sin descripción
NEGOC_TEL	teléfono del negocio	FECHA_CIMG	Sin descripción
NEGOC_ANT	antigüedad del negocio (años)	ANT_LAB_ANT	Sin descripción
NEGOC_DIR	dirección del negocio	TIPO_USO	Sin descripción
NEGOC_CIU	código de la cuidad del negocio	PERFIL_PROF	Sin descripción
NEGOC_BAR	código del barrio del negocio	GASTOS_FAM	Sin descripción
NEGOC_ACT	código de la actividad del negocio	VLR_ARRIENDO	Sin descripción
NEGOC_TIP	código del tipo de negocio	COD_MATRICULA	Sin descripción
NEGCMONTO	valor solicitado	GASTOS_NEG	Sin descripción
NEGC_VENT	ventas solicitadas	VL_EFECTIVO	Sin descripción
NEGC_UBIC	descripción de la ubicación del negocio	VL_ADICIONAL1	Sin descripción
COMOSUPO	código como supo de nosotros	TIPO_LUGAR	Sin descripción
FORMULA	código de donde lleno el formulario	CANTIDAD	Sin descripción
APROBO	Código de aprobación	COD_MATRICULADS	Sin descripción
NEGADO	Código de negación	COD_CONTRA	Sin descripción

RECHAZO	Código Cauda de rechazo	NROTAXIS	Sin descripción
CODE_DEPA	Código del departamento	VLR_OBSEQ	Sin descripción
DESTINO	Código de destino del crédito	OTROS_INGP	Sin descripción
VLR_CREDIT	valor de crédito	OTROS_BENEFICIOS	Sin descripción
PORCEMORA	porcentaje de mora	FECHACON	Sin descripción
FECHA	fecha de referencia interna	CIUDAD_OPER	Sin descripción
ESTADO	estado del crédito	VLR_MAXAPAGAR	Valor de cuota máximo que puede pagar el cliente
ASESOR	Código del asesor que monta la solicitud	USUARIO_DIG	Sin descripción
USUARIO	Código del usuario que monta la solicitud	SEDE_DIG	Sin descripción
ESTRACTO	Código del estrato	IMAGEN_PRO4	Imagen 4 de la garantía utilizada inicialmente para el SOAT
COD_NIVEDU	código del nivel educativo	NUM_EMPLE	Sin descripción
SEXO	código del sexo del cliente	FEC_SCORELOG	Sin descripción
COD_VIVIEN	código del tipo de vivienda	FORMAPAGO	Sin descripción
FECHANAC	fecha de nacimiento del cliente	REFERENCIA_SP	Sin descripción
PERS_CARGO	código del número de personas a cargo	COD_MARCA	Sin descripción
NUMSOLASE	Código de solicitud sincronizada de un asesor (Plantilla)	COMPRAS_CC	Sin descripción

TPCREDGARA	Código del tipo de crédito	VALOR_CC	Sin descripción
CREDPREP	Número del crédito a prepagar	TURNO	Sin descripción
VLRPREPA	valor del prepago	TURNO_PROP	Sin descripción
NUMFIC	numero	COD_COLOCA	Sin descripción
SCORENEW	Valor del nuevo Scoring	ENFERMEDAD	Sin descripción
CEDULA_CO	numero de cedula del codeudor	IMAGEN_SEG	Sin descripción
SINCRO	Código de estado de sincronización	OBSERVACION1	Sin descripción
VL_PRODUCTO	Sin descripción	OBSERVACION2	Sin descripción
MODELO_VEH	Sin descripción	VALORCUO_CC	Sin descripción
TIPOCRE	Sin descripción	DIRECCION	Sin descripción
VL_CUOTA	Sin descripción	TELEFONO	Sin descripción
COD_EMPRE	Sin descripción	CELULAR	Sin descripción
NUM_CUO	Sin descripción	TIPO_PENSION	Sin descripción
COD_FC	Sin descripción	CLAVE_PENSION	Sin descripción
TIPO_DOC	Sin descripción	PORC_INVALIDEZ	Sin descripción
COD_PROFE	Sin descripción	Sin descripción	Sin descripción

f. Créditos: En esta tabla se encuentra toda la información referente a los créditos desembolsados. Para la extracción de esta data, no se realizó ningún tipo de limpieza o depuración, obteniendo como resultado 29.815 registros y las siguientes variables:

Tabla 6: Campos de la tabla créditos

Variable	Definición	Variable	Definición
ANU_CONCE	Concepto de anulación	FEC_VTAJUR	fecha venta jurídico
ANU_COTCRE	Estado de la anulación	CAP_VTAJUR	valor capital venta jurídica

CEDULA_CLI	Cedula del cliente	EMP_COMPRA	Código empresa de compra
CEDULA_CO1	Cedula del 1 codeudor	CAP_COMPRA	valor capital de compra
CEDULA_CO2	Cedula del 2 codeudor	FEC_COMPRA	fecha de compra
CHASIS	Número del chasis	FEC_DATCRE	fecha de consulta data crédito
CIUDAD	Código de la ciudad	TIPO_VTA	código tipo venta
COD_VEHICU	Código del vehículo	TIP_VTAJUR	código tipo venta jurídico
COLOR	Color del vehículo	CAP_COMPRAORI	valor de capital de compra original
CORREO	Correo	TIP_COMPRA	código tipo de compra
CREDITO	número del crédito	DESCUOTA_C	numero de cuota a partir de que se vende
CUOTA_INIC	Valor cuota inicial	COD_BASE	código de la empresa
DIAS_PAGO	Numero días de pago	VLR_FONDOG	valor del fondo garantías
EMPRESA	Código empresa	CONAPORTE_V	Estado de venta con aporte
ENTIDAD	Código entidad	VALOR_CUO_V	Valor cuota venta
ESTADO_CR	Código estado de crédito cartas	NUMCREDCOMPRA	Numero de crédito que se compra
FCH_EST_CRE	Fecha estado crédito	DESFACE_V	valor del desfase venta
ESTADO_CRE	código estado del crédito	CONDESFACE_V	código desfase venta
IMP_DOC_JUR	código importación documentación jurídico	FECH_ENVFONDOG	fecha envió fondo garantías
TRAMIDEMAN	Sin descripción	ARCH_ENVFONDOG	nombre archivo envió fondo garantías

ESTADO_JUR	Código estado jurídico	CAP_VTAJURCON	valor venta capital jurídico contable
FCHARCHIVO	fecha generación de archivo plano	TIPO_SCORE	código tipo score
USUARCHIVO	código usuario que genero el archivo	SCORING	valor del scoring
FACTURA	numero de factura	SCOREP	Sin descripción
FECHA_ANU	fecha de anulación	NUMERO	numero de la solicitud
FECHA_AP	fecha de aprobación	FEC_ENVFG_PREPA	fecha de envió prepago
FECHA_VENT	fecha de venta	ARC_ENVFG_PREPA	nombre archivo prepago
FORMA_P	Código forma de pago	COD_NIVRIE	código nivel de riesgo
F_PAGADO	fecha de pago	VALOR_CASC	valor del castigo
F_RESTRUC	fecha de restructuración	VLRMARREC	valor margen recompra
INT_NORMAL	interés normal	SCOREM	calificación score mas
LINEA	Código de línea	CANT_VTA	numero de cuotas
MARCA	Nombre marca	COD_EMPRE	Código del distribuidor
MODELO	Nombre modelo	FIRMARUEGO	estado de autorización de firma
MOTOR	Nombre motor	FEC_COSTOADI	fecha aplicación costo adicional
NOTA	nota	FECHA_DESC	Sin descripción
NO_PRENDA	numero de prenda	VERIF_IMG	Nombre de la imagen
NUMERO_CUO	numero de cuotas	BLOQUEOVTA	estado boqueo venta
OFICINA	Código oficina	EMP_NOV	Sin descripción
PLACA	numero de placa	COD_BASE_ORI	Sin descripción
SALDO_FINA	Valor saldo final	VLR_MICROSEGURO	Sin descripción
SEGURO	valor seguro	COD_CAMPA	Sin descripción

SERVICIO	Código tipo de servicio	MARGENDIF	Sin descripción
SUPLAN	Sin descripción	PORCEDIF	Sin descripción
TIPO	Código tipo de crédito	CEDULA_ASES	Sin descripción
USUARIO	usuario que ingreso el registro	COD_MATRICULA	Sin descripción
VALORCIA	valor mercancía	COD_MATRICULADS	Sin descripción
VALOR_CUO	valor cuota	NOVEDAD	Sin descripción
VALOR_SOL	valor solicitado	FEC_REALREC	Sin descripción
VALOR_TOTA	valor total	USUA_DCAS	Sin descripción
VENDEDOR	código vendedor	COD_CANAL	Código del canal al cual pertenece el asesor
VENTANA	Sin descripción	USUARIO_DIG	Sin descripción
VLR_PREPA	valor prepago	SEDE_DIG	Sin descripción
VLR_REST	valor reestructurado	CEDULA_DIG	Sin descripción
ZONA	código de zona	CEDULA_LEG	cedula del usuario que legaliza
DPTO	código departamento	FECHA_LEG	fecha de legalización
SEDE	código sede	SERFAC_V	Numero de documento en la facturación en venta de cartera créditos
ASES	código asesor	NUMERO_CUOCTE	Numero de cuota máxima para definir como valor corriente
VLR_ADMIN	valor administración	VLRCAPRECCTE	valor corriente del capital recomprado
PROVEEDOR	Número de identificación proveedor	CAP_VTAJURCTE	Valor para el capital jurídico corriente
ESTADO_NUE	Sin descripción	CAP_VTAJURCONCTE	Valor para el capital jurídico contable corriente

VLR_SEGURO	valor seguro	NUMERO_CUOCTE_V	Numero de cuota máxima para definir como valor
FUERZA	Código fuerzas militares	VLRMARRECCTE	Sin descripción
VLR_ESTUDI	valor estudio	COSTOTRAN	Sin descripción
MODELO_LIQ	código modelo liquido	CENTROCB	código del centro de costo o beneficio SAP del gestor
VLR_COMI	valor comisión	NUM_FOPEP	Numero asignado por fopep a crédito de positiva
NETO	valor neto	PERIODO_PAGO	Sin descripción
FECHA_ESC	fecha del escenario	EMP_COBRANZA	Código identifica Casa de Cobranzas
ESCENARIO	Código del escenario	FEC_COBRANZA	Fecha vinculación Casa de Cobranzas
ASESOLD	Código asesor anterior	TIPO_JUDICIAL	Identificación registro tipo judicial (Datos 262)
FECHA_APO	Fecha de aprobación	FEC_REFINA	Sin descripción
EMPRESA_V	código empresa venta	INT_NORMALNEW	Sin descripción
NUMFAC_V	factura venta	COD_COLOCA	Sin descripción
FECHA_V	fecha venta	COMISION_COL	Sin descripción
CAPITAL_V	valor capital venta	CREDITO_ORI	Crédito original de la venta
CUOTA_V	valor cuota venta	INT_VTAJUR	Sin descripción
DESCUOTA_V	Valor descuento cuota	FECHA_SYS	Fecha del sistema
INTERES_V	valor interés venta	ENVIADO	Archivo enviado a pichincha
FECHA_CAS	fecha castigo	NUM_OPERACION	Numero de operación pichincha
VALOR_CAS	valor castigo	VLR_COMISIONEMP	Valor cobrado por comisión

VLR_RECOMPRA	valor recompra	PORC_EMPCOLOCA	Porcentaje empresa colocadora
FECHA_REC	Fecha restructuración	COMISION_EMPINI	Valor de Comisión inicial del correspondiente 30%
CONSECUTIVO_V	consecutiva venta	PROCESO1_ESP	Valor por proceso1
CEDULA_ALT	cedula alterna	OPERA	Indica si el crédito opera o no
FECHA_DCAS	Fecha levantamiento castigo	SEDE_EXT	Sin descripción
VLR_RESTRU	valor reestructurado	COD_VENTA	Sin descripción
VLR_DESFA	valor del desfase	TIPO_CAS	Sin descripción
TIPDESEM	código tipo de desembolso	COB_CARTERA	Sin descripción
FEC_RECINI	Fecha recompra inicial	DIR_CARTERA	Sin descripción
TIPO_REC	Código tipo de recompra	ESTADO_CAR	Sin descripción
PORLINEA	código porcentaje de línea	PORC_SEGURO	Sin descripción
ESTADOCUST	código estado custodio	PORC_SEGURONEW	Sin descripción
VLRCAPREC	Valor de capital recomprado	PORC_FONDOG	Sin descripción
FEC_PAGOV	fecha de pago venta	VLR_SEGINI	Sin descripción
FEC_REALV	fecha real venta	PORC_SEGINI	Sin descripción
TIPOCRE	código tipo de crédito	PROC_DESFA	Marcador de proceso diario devolución intereses
PYME	valor pyme	PORCE_CORRETA	Porcentaje equivalente al corretaje
IVA	valor IVA	VLR_CORRETA	Valor equivalente al corretaje

g. Reestructurados: Esta es una tabla provista por el área de cobranzas de la compañía, en la cual se detallan los créditos que han sido reestructurados (Cambio del plazo o la tasa original mediante un acuerdo con el cliente para el saneamiento de la obligación). Esta tabla contiene únicamente 229 registros y las siguientes variables:

Tabla 7: Campos de la tabla reestructurados

Variable	**Definición**
CEDULA	Cédula del cliente
CREDITO	Número de crédito
MES APLICACIÓN	Mes en el cual se aplicó la reestructuración
POSICION	Entidad dueña de los derechos económicos de la obligación
SALDO	Saldo de capital al momento de la reestructuración
CUOTA ANTERIOR	Valor original de la cuota del crédito
CUOTA NUEVA	Valor de la cuota del crédito posterior a la reestructuración
% Operaba	Porcentaje de la cuota que se encontraba operando a través del descuento de nómina

h. Ventas Fondeador A[1]: Esta es una consulta extraída de la base de datos corporativa en la cual se detallan los créditos vendidos al Fondeador A, con el propósito de realizar la caracterización de clientes buenos en términos de utilidad para la compañía, la única variable contenida en esta tabla es el número de crédito.

2.2.2. Análisis de los datos

Como se menciona y evidencia en las fuentes de datos, al ser una base de datos heredada de otra compañía, existen un gran número de variables que no poseen

[1] Por motivos de confidencialidad, ninguna entidad con la cual Avista posee una relación comercial o contractual será llamada por su nombre propio.

relevancia para el negocio actual, no se tiene conocimiento de la información que existe en ellas o se encuentran sin datos.

Por lo anterior, lo primero que se realizó, fue la limpieza de los datos, fuente a fuente de la siguiente manera:

- Tabla ***Personas***
    - En primer lugar, se filtraron los registros para obtener únicamente los registros con tipo de documento igual a 1, correspondiente a cédula de ciudadanía, para eliminar registros con otro tipo de documento diferente al mencionado.
    - Se calculó la longitud en número de caracteres de la cédula de cada registro y se eliminaron todos los registros con longitud menor a 6, puesto que, al analizarlos, correspondían a errores dentro de la tabla.
    - Se realizó un análisis de duplicados, en donde se encontró que existían 17 registros duplicados de una base de 79.804 y que la eliminación de estos no suponía una pérdida de información.
    - Por último, se eliminaron los registros que se encontraban con la fecha de nacimiento vacía en la tabla, puesto que la edad pasará a ser una variable relevante en cálculos posteriores.

Luego de aplicar estos filtros obtuvimos una base de 68.583 a partir de una base original de 79.804 registros. Es importante resaltar que en esta base existe un número importante de individuos que fueron solicitantes de crédito, pero no lograron obtener una obligación con Avista.

- Tabla ***Centrales***

  En primer lugar, se realizó un análisis descriptivo para conocer la estructura de los datos. A partir de este, se logró identificar que existían datos raros y muy extremos dentro de la información, como se evidencia en la Tabla ,7 y los histogramas de cada una de las variables (Gráfico 1)

  Tabla 8: estadística descriptiva de los campos de la tabla centrales

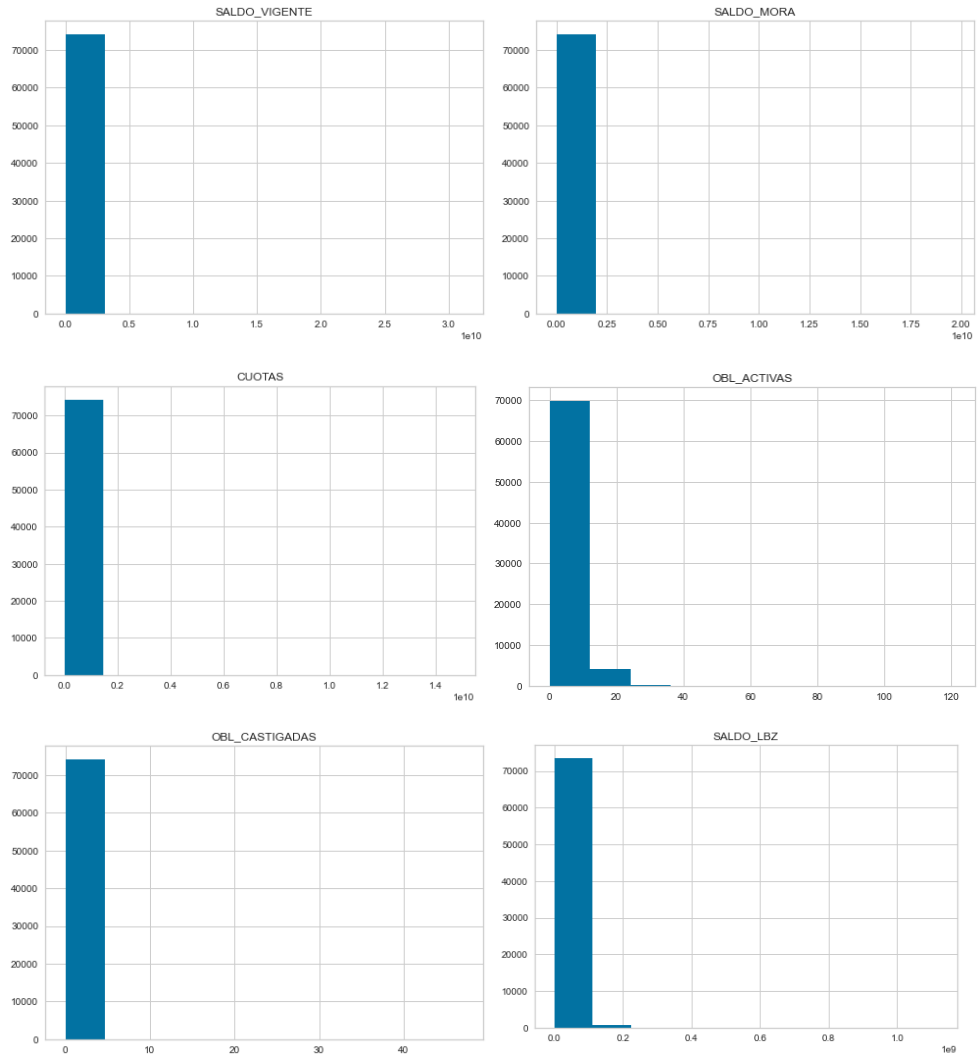
	SOLICITUD	CEDULA	SALDO_VIGENTE	SALDO_MORA
count	74.191	74.190	74.191	74.191
mean	796.866	54.657.639	52.987.505	11.554.962
std	35.504	166.805.607	181.954.909	101.089.930
min	717.158	85.008	-2	0
25%	769.393	10.158.729	15.431.000	0
50%	802.109	22.326.485	30.096.000	515.000
75%	827.541	39.561.946	63.931.500	5.788.000
max	848.755	1.234.990.069	31.115.456.000	19.633.870.000

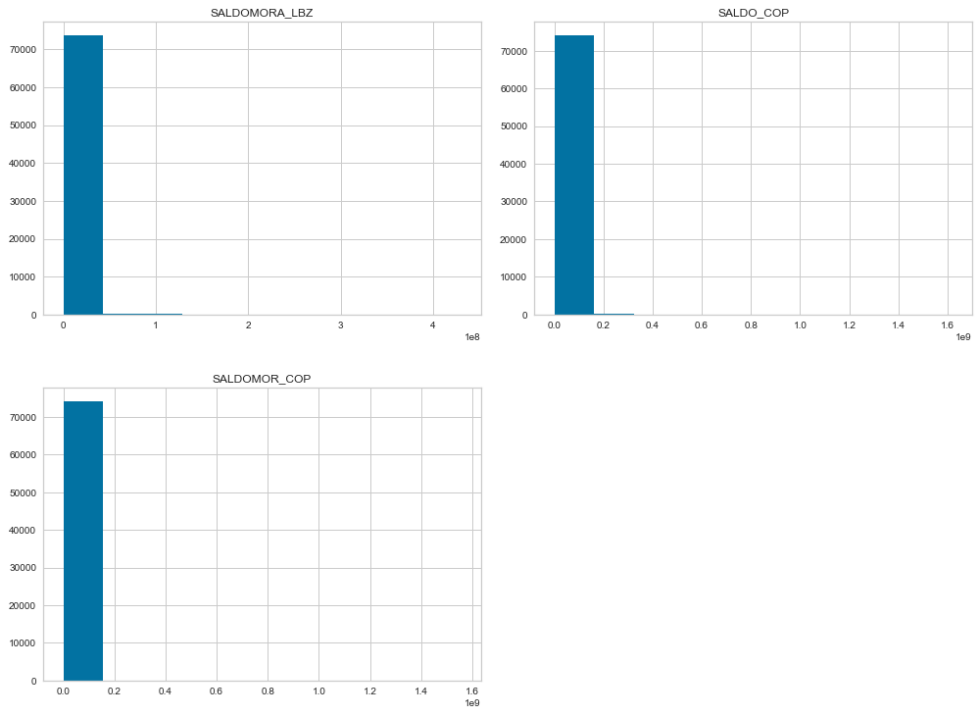
	CUOTAS	OBL_ACTIVAS	OBL_CASTIGADAS	SALDO_LBZ
count	74.191	74.191	74.191	74.191
mean	7.358.030	6	0	10.259.182
std	67.962.245	4	0	24.288.395
min	0	0	0	0
25%	560.000	3	0	0
50%	1.295.000	5	0	0
75%	3.612.500	7	0	9.613.000
max	14.752.394.000	121	47	1.118.415.000

	SALDOMORA_LBZ	SALDO_COP	SALDOMOR_COP
count	74.191	74.191	74.191
mean	933.722	2.211.438	399.764

std	7.916.596	17.620.233	9.174.068
min	-3	0	0
25%	0	0	0
50%	0	0	0
75%	0	0	0
max	430.174.000	1.618.347.000	1.556.158.000

Ilustración 1: Histogramas de las variables de la tabla centrales



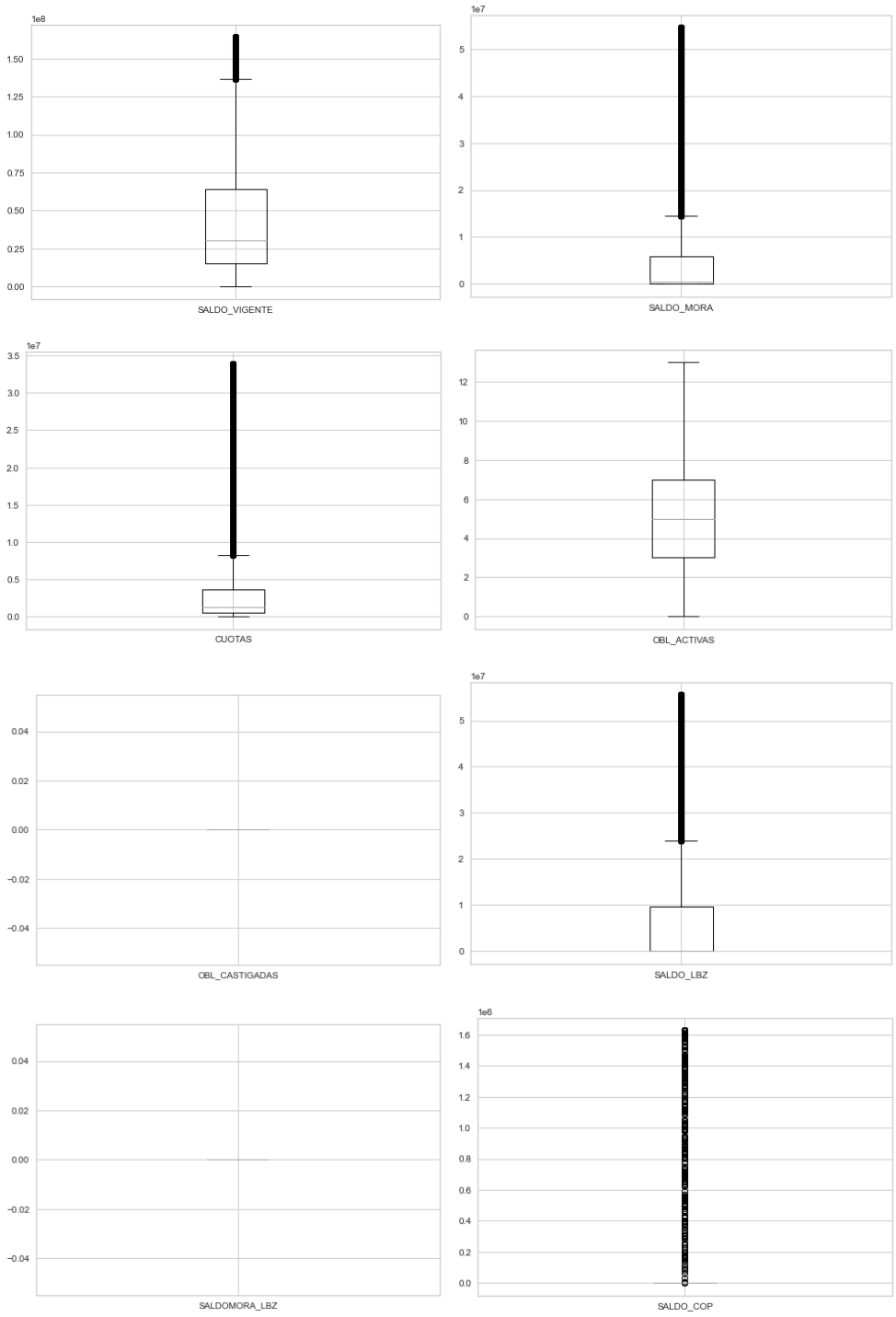


Al visualizar la información de la tabla de manera gráfica, mediante histogramas, se pudo evidenciar la necesidad de una limpieza en los datos, ya que, en todas las variables de la tabla, se encontró una concentración muy fuerte de data en el lado izquierdo del histograma. Lo anterior hace referencia a la presencia de datos atípicos y extremos, que no permiten visualizar la distribución real de cada variable en su histograma, ya que los pocos datos atípicos de la muestra no permiten una visualización de la distribución en la cual se encuentra concentrada la información.

Debido a la detección de datos atípicos mediante los histogramas de las variables, se llevó a cabo una limpieza del set de datos, usando el percentil 95, reemplazando los datos outliers que superan al 95% de la muestra, con el valor de este percentil.

Con el fin de conocer la distribución de los datos y poder ejecutar una revisión para saber si las variables que aún permanecían con ruido, se realizó un segundo análisis, bajo el uso de *boxplots,* obteniendo los siguientes resultados:

Ilustración 2: Boxplots de las variables de la tabla centrales



Los *boxplots* previamente expuestos, permiten observar que gran parte de estas variables carecen de información relevante para modelar, ya que por la estructura de sus datos, únicamente generarían ruido para el modelo y no aportarían información. Las únicas variables relevantes a partir de este análisis serían SALDO_VIEGENTE, SALDO_MORA, OBL_ACTIVAS y SALDO_LBZ.

Adicionalmente desde la visión del negocio, las variables previamente seleccionadas, son las usadas por los analistas de crédito para realizar la asignación del perfil de riesgo y de esta manera, definir valores máximos a términos de montos, plazos y tasas.

Luego de la limpieza a los datos, se llevó a cabo la creación de nuevas variables, mediante el cálculo de la media y la mediana de toda la muestra para cada una de las variables, para posteriormente, realizar la resta del valor que poseía cada uno de los registros en su respectiva variable con las medidas estadísticas mencionadas anteriormente. Lo anterior se realizó buscando posibles variables que puedan llegar a ser relevantes al momento de realizar el modelo segmentación de datos.

Luego de realizar limpieza y tratamiento de datos sobre esta tabla, se obtuvo una tabla con 74.191 registros y 18 variables en total, partiendo de una tabla que originalmente poseía 74.845 registros y 12 variables.

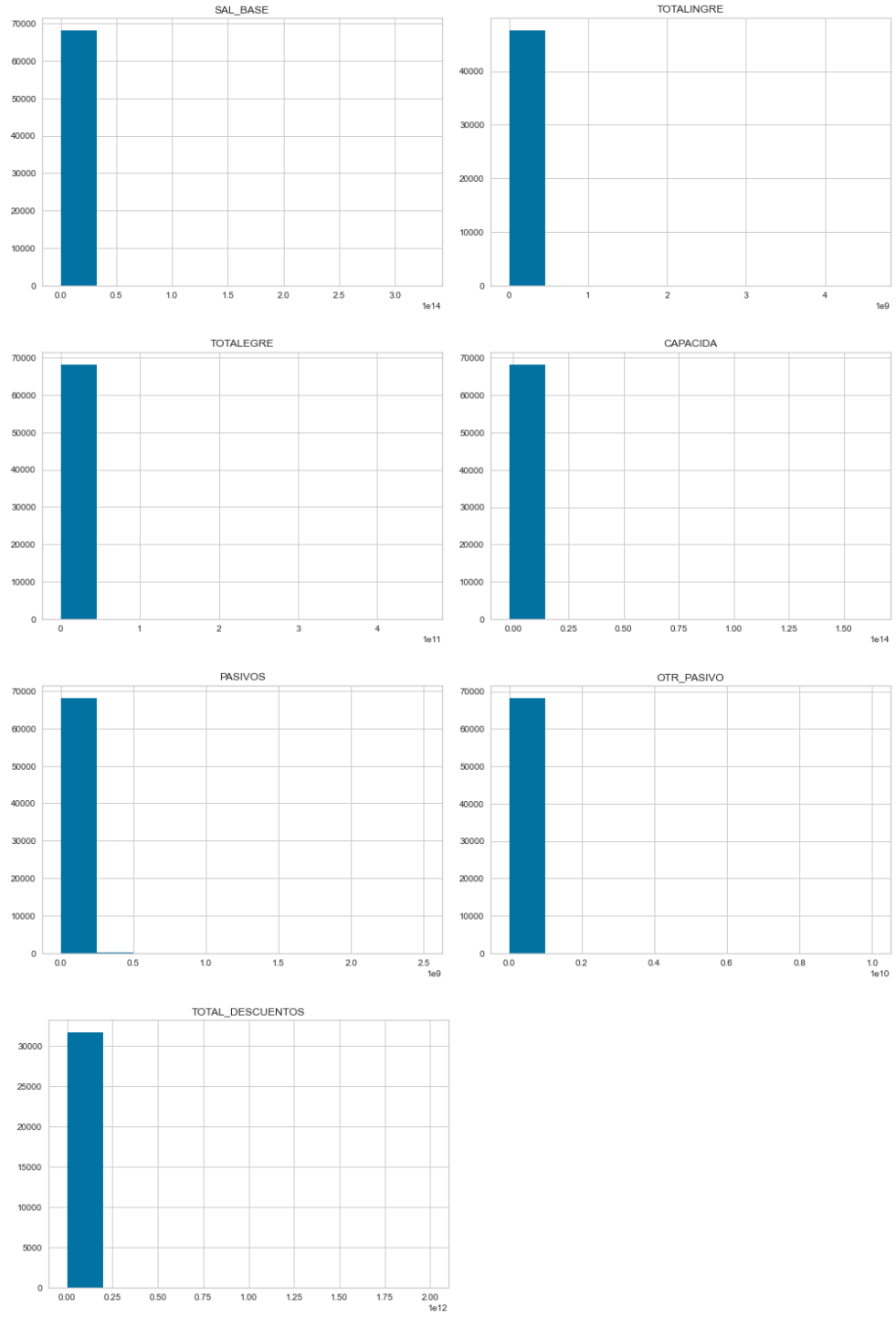
- Tabla ***Información Financiera*** y ***Solicitudes***

  La información de esta tabla permite tener un reflejo de la situación de ingresos y egresos de la persona al momento de la solicitud, por lo que un mismo cliente puede llegar a tener tantos registros como solicitudes de crédito haya realizado. Es por esto, que como criterio de tratamiento de datos y con el objetivo de tener la información más reciente posible para cada cliente, se seleccionó la solicitud con la fecha más reciente usando la tabla *solicitudes* como fuente de esta fecha.

  La tabla *solicitudes* es una tabla que contiene 202 variables, pero para el caso de estudio, la única variable de utilidad será la fecha de la solicitud, ya que de las demás variables hacen referencia a datos de identificación del cliente en diferentes aspectos tales como nombres, apellidos, cédula, número de afiliación, dirección, entre otros; estos datos están almacenados de igual forma en otras tablas, por lo que, de ésta, únicamente se usará el número de la solicitud y la fecha de la misma.

  Por lo anterior solo se realizó el análisis de datos extraños para la tabla financiera, encontrando los siguientes histogramas para las variables sin ningún tipo de limpieza:

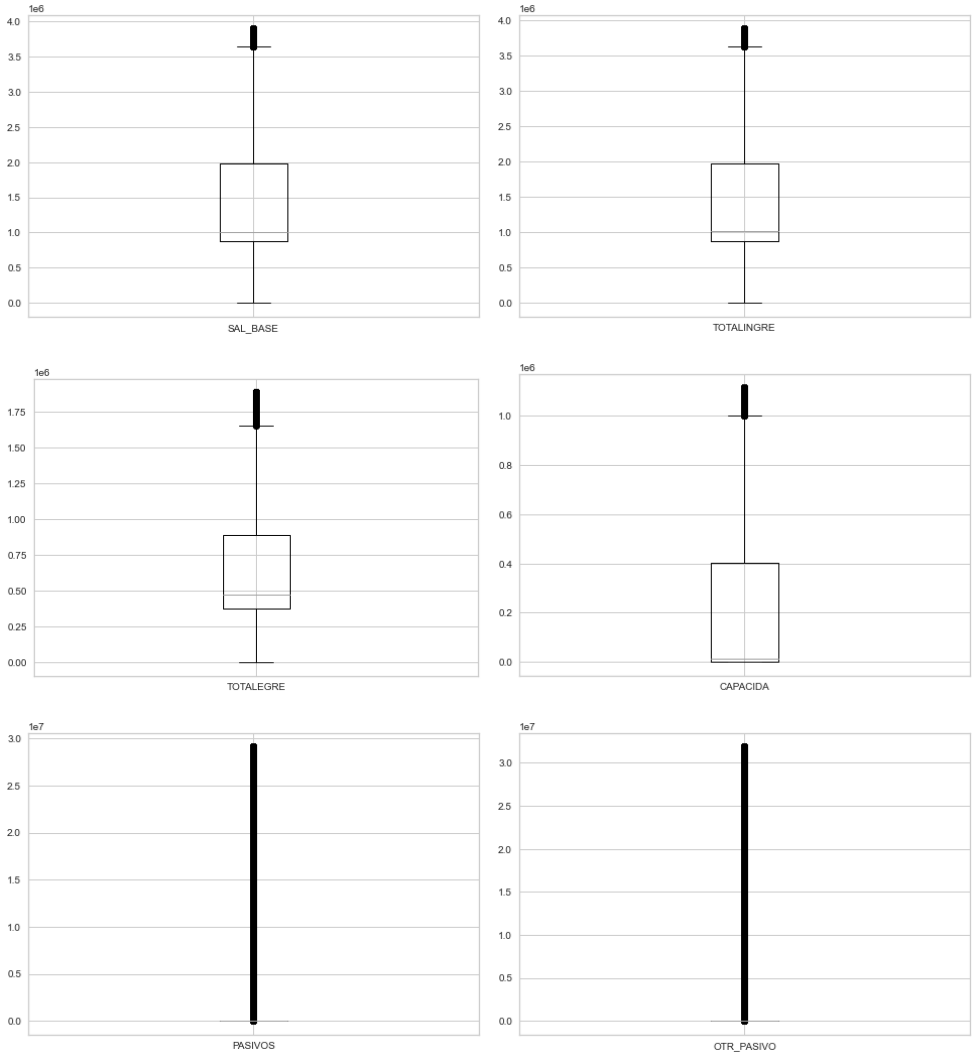
Ilustración 3: Histogramas de la tabla Información Financiera y Solicitudes

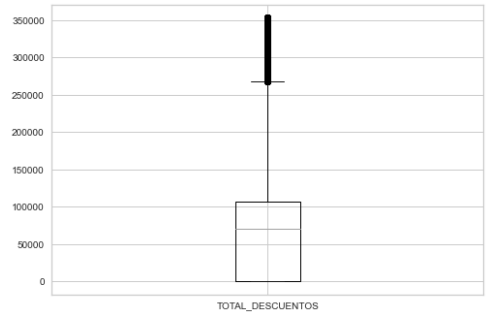


Como se evidencia en los histogramas presentados anteriormente, los datos requerían de una limpieza para aportar información al modelo, es por esto, que

como que, en el caso anterior, se aplicó una limpieza de datos outliers mediante el reemplazo de valores extremos con el valor del percentil 95 de cada una de las variables y el percentil 5 únicamente para la variable capacidad la cual tenía valores menores a 0, obteniendo las siguientes distribuciones:

Ilustración 4: Boxplots de la tabla Información Financiera y Solicitudes





Según los gráficos expuestos arriba se puede determinar que hay variables que carecen de información real o datos que puedan ser convertidos en información por el modelo, por esto las variables a usar a partir de ese análisis son SAL_BASE, TOTALINGRE, TOTALEGRE, TOTAL_DESCUENTOS y CAPACIDA.

Adicionalmente y por su distribución, se puede evidenciar tanto desde los datos, como desde la relevancia para el negocio, que variables como PASIVOS y OTR_PASIVOS, son datos que no representan la realidad del negocio y no agregan información desde ninguna perspectiva, ya que si así fuera, presentarían un comportamiento en el cual, se podría extraer información, pero al evidenciar una distribución poblada de datos outliers, podemos concluir que estas variables pueden ser eliminadas.

- Tabla ***Créditos***

  La tabla mencionada posee una estructura en donde se detalla cada uno de los créditos con sus características, por esto un cliente puede tener múltiples registros en esta tabla. Debido a que la base para la modelación tiene una estructura en donde cada cliente posee un único registro, se usó la información de esta tabla para obtener estadísticas por cliente de acuerdo con las obligaciones históricas que ha

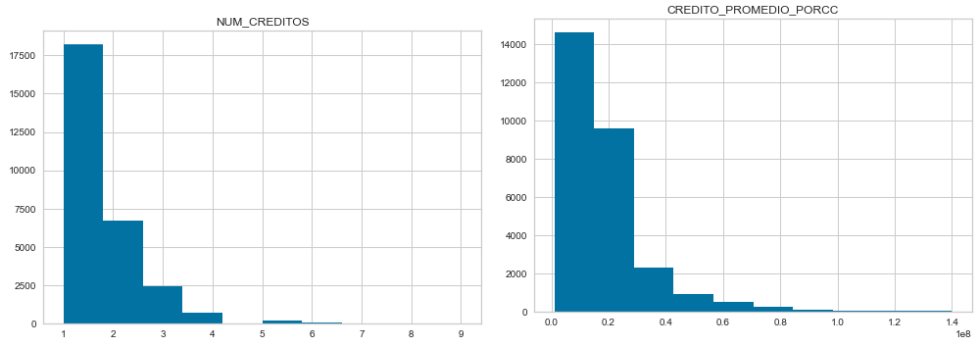
adquirido con la compañía, logrando construir un perfil del individuo en términos de su histórico.

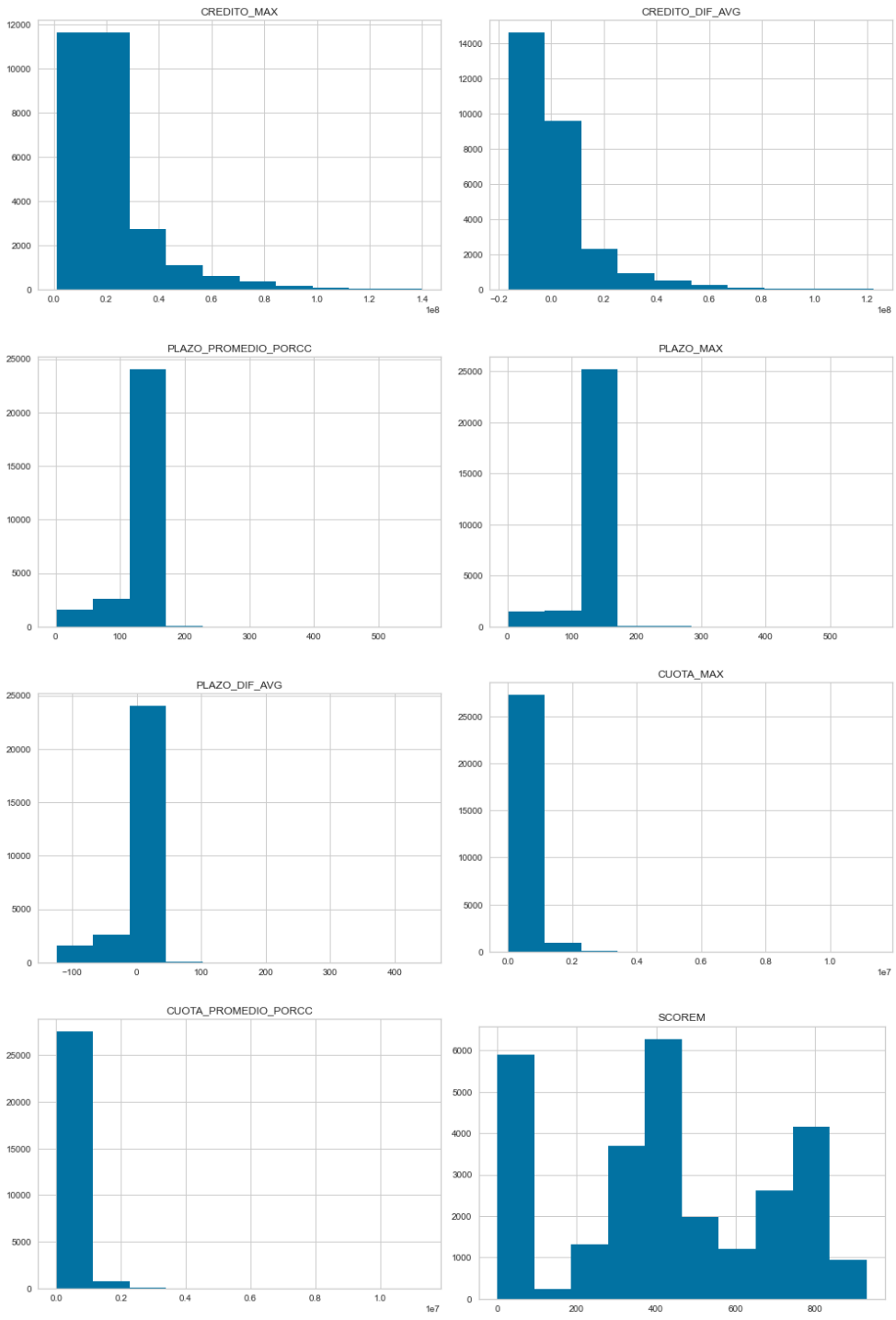
Las medidas estadísticas usadas para la construcción de este fueron máximo, promedio, mediana, desviación estándar, y diferencias de los individuos contra las dos últimas mencionadas para generar una información con variabilidad referente a los individuos de la muestra.

Adicionalmente se convirtió la diferencia entre la fecha de originación y la fecha de pago de los créditos saldados (plazo de pago), en número de meses, para aplicar las medidas estadísticas mencionadas anteriormente, esta variable de plazo de pago se usó también para construir una variable categórica en donde se marcaban los clientes que poseen créditos saldados y los que aún no han terminado de pagar sus obligaciones con la entidad.

Los histogramas obtenidos a partir del set de datos construido fueron los siguientes:

Ilustración 5: Histogramas de la tabla créditos



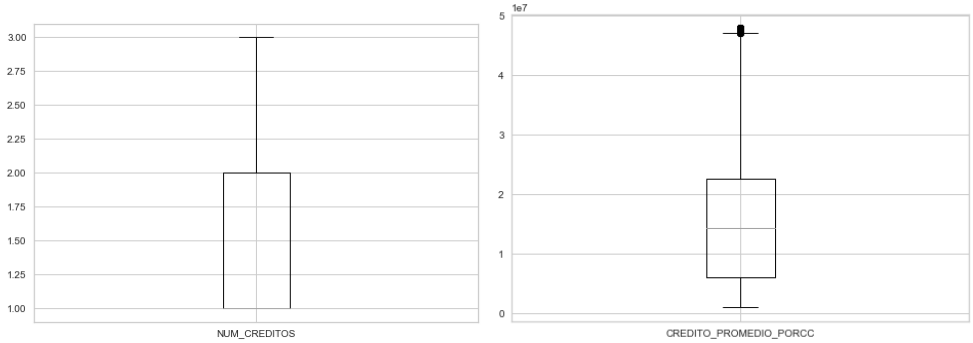


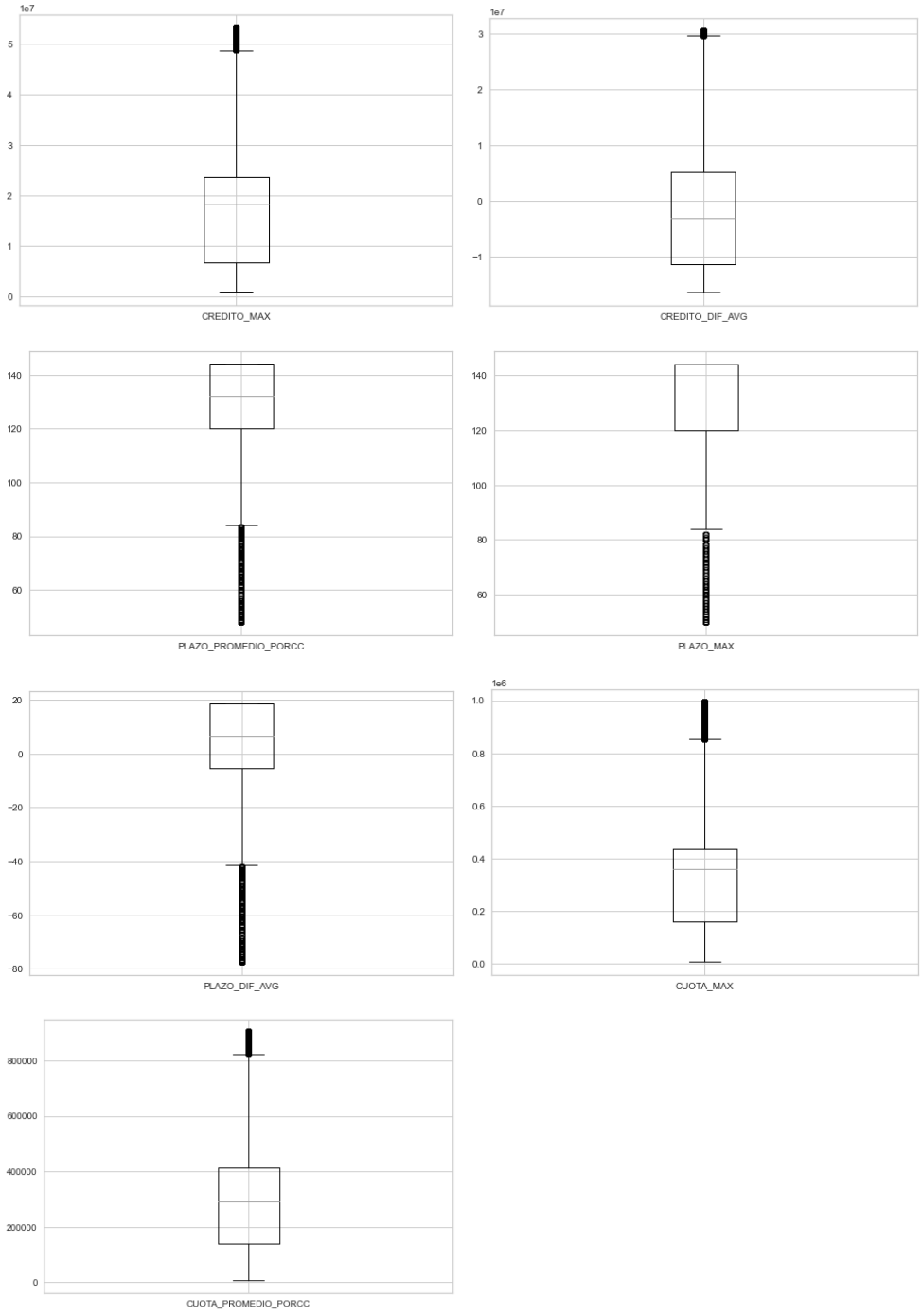
Al observar la representación gráfica de cada una de las variables mediante histogramas, se encontraron datos con distribuciones sesgadas a la izquierda, lo que sugiere presencia de datos atípicos y extremos en la muestra.

Un histograma en el cual se puede observar una distribución coherente con el negocio es en el caso de SCOREM, en donde claramente, se identifica mediante el gráfico, la presencia de 3 poblaciones de clientes, clientes sin experiencia crediticias (score de 0 puntos), clientes con riesgo medio (score alrededor de 400 puntos) y clientes con riesgo bajo (clientes alrededor de 800 puntos). Tomando en cuenta este ejemplo, hay fuerte evidencia que sugiere realizar una limpieza de datos para las demás variables, pero teniendo en cuenta que la concentración en ciertos valores para el caso de los plazos es conocida desde el lado del negocio.

Finalmente, con el propósito de eliminar sesgos y aportar variables con mejor calidad de datos al modelo, se realizó una limpieza de *outliers* mediante el uso del percentil 95 para todas las variables excepto para SCORE_M, la cual mostró una distribución coherente con los niveles de riesgo de los clientes desembolsados por la compañía, y para las variables referentes al plazo del crédito, se hizo una limpieza inferior usando el percentil 5, ya que existían valores muy bajos que no reflejan la realidad de los plazos otorgados. L os *boxplots* resultantes posteriores a la limpieza de los datos fueron los siguientes:

Ilustración 6: Boxplots de la tabla créditos





Este set de datos fue construido a partir del resumen de la historia crediticia de cada cliente con la entidad, mediante el uso de diferentes medidas estadísticas, como se mencionó anteriormente. Cada una de las variables analizadas, muestra como la compañía realiza su colocación concentrando las condiciones de sus obligaciones

alrededor de ciertas cifras, como en el caso de los plazos, donde el 75% la muestra se concentra en plazos superiores a 120 meses, o en el caso de la cuota, donde el 75% de la población se ubica en cuotas inferiores a COP$400.000. Es por esto, por lo que estos gráficos permiten conocer la realidad tanto del negocio como del mercado.

De este set de datos se usarán todas las variables analizadas previamente, pues muestran un comportamiento coherente en términos de lo que representa cada una y podrían aportar información valiosa al modelo, adicional a las variables continuas de la muestra se usaran las variables categóricas creadas, las cuales se crearon a partir del uso de medianas y nos indican en qué casos la observación la supera o no, buscando de igual forma, aportar más información para el momento de la modelación.

2.2.3.Construcción de la marca de clientes malos para la compañía

El objetivo del trabajo consiste en poder encontrar en donde se concentra la utilidad de la compañía a través de la identificación de segmentos de clientes en el territorio colombiano. Por lo anterior, se llevó a cabo la construcción de una marca de clientes malos, a partir de criterios definidos con expertos del negocio, ya que realizar un cálculo de la rentabilidad crédito a crédito no fue posible debido a un gran número de cambios que ha experimentado Avista en términos de estructura de costos y cobros a sus clientes.

Los criterios definidos para saber si un cliente representa una buena fuente de utilidad para la compañía son los siguientes:

- **Clientes con una edad inferior a 75 años**, puesto que no representan un mayor costo para Avista en términos de seguro de vida. Esto se da debido a que Avista dentro de su propuesta de valor, cubre durante toda la vida del crédito, el cobro que realiza la entidad aseguradora por concepto de la póliza contra fallecimiento, la cual tiene un mayor valor para los clientes que superan los 75 años.
- **Clientes que no hayan fallecido** y por ende sus obligaciones no hayan sido pagadas por la aseguradora por fallecimiento del titular de la obligación. Lo anterior es óptimo para la compañía ya los fallecimientos representan un riesgo de encarecimiento o incluso perdida de la póliza asignada a los créditos desembolsados.
- **Clientes que no hayan tenido un nivel de deterioro** alto el cual haya implicado realizar una reclamación de saldo total a la entidad proveedora de la fianza, esto debido a que los clientes reclamados representan una reducción de saldo disponible en la entidad avalista para cubrir el deterioro de cartera, además de que no representan ningún ingreso real para Avista en los diferentes cargos del crédito ya que estos realizaron pocos o ningún pago.
- **Clientes que no hayan tenido reestructuración en las condiciones originales de su crédito**. Una reestructuración representa un cliente incapaz de atender su obligación, por lo que acarrea costos adicionales en su proceso de cobranza, además de una modificación en las condiciones de su

crédito, generalmente en la tasa, representando así un menor ingreso para la compañía. El proceso de reestructuración puede provenir tanto de una solicitud del cliente para normalizar su obligación o como resultado del proceso interno de cobranza, el cual busca normalizar los clientes cuyas obligaciones se encuentra deterioradas.

- **Clientes con una tasa superior al 1.6% mensual**. Los créditos con tasas de interés inferiores la mencionadas son desembolsados en campañas comerciales específicas en donde se pretende capturar mercado, sin embargo, no representan un ingreso para Avista.
- Clientes que cumplan dos condiciones al mismo tiempo:
    - Fecha de desembolso posterior a diciembre de 2019
    - Entidad compradora de la obligación diferente a Fondeador A[2]

  Esto debido a que en ese momento existía un acuerdo con esta entidad que aseguraba una fuente de liquidez, pero se traducía en un costo alto para Avista, lo que llevaba a que estos créditos no fueran una fuente de utilidad. Este acuerdo fue renegociado y por ende créditos vendidos a esta misma entidad a partir de enero de 2020 si representan obligaciones rentables para la compañía.

Con la definición de estos criterios se procedió a construir una base en la cual se obtuvieron las cédulas de los clientes con su marca de malos, representando un **32,96%** de la muestra total de clientes.

[2] Por motivos de confidencialidad, ninguna entidad con la cual Avista posee una relación comercial o contractual será llamada por su nombre propio.

2.2.4.Consolidación de los datos para modelación

Para la fase de modelación, se decidió usar toda la información consolidada en la tabla ***Personas***, a excepción de las variables que a través de las diferentes uniones entre tablas quedaron incluidas, pero no poseían ningún tipo de dato conocido o simplemente no poseían información.

Dentro de la tabla previamente mencionada, las variables que si contenían información de cualquier tipo fueron las siguientes:

Tabla 9: Variables consolidada para modelación

Nombre	Tipo	Definición
CIUDADEXP	Categórica	Ciudad de expedición del documento del titular
CIUDADNAC	Categórica	Ciudad de nacimiento del documento del titular
COD_ESTCIV	Categórica	Código que representa el estado civil del titular
COD_NIVEDU	Categórica	Código que representa el nivel de educación del titular
COD_PROFE	Categórica	Código que representa la profesión del titular
PERS_CARGO	Continua	Número de personas que dependen económicamente del titular
CIUDAD	Categórica	Ciudad donde fue originada la obligación
ESTRATO	Categórica	Estrato socioeconómico del titular
CODOFIC	Categórica	Código de la oficina donde fue tramitada la solicitud
ESTADCIVI	Categórica	Estado civil del titular
ASESOR	Categórica	Código del asesor que tramitó la solicitud
SALDO_VIEGENTE	Continua	Saldo total de las obligaciones del titular en centrales de riesgo
SALDO_MORA	Continua	Saldo en mora de las obligaciones del titular en centrales de riesgo
OBL_ACTIVAS	Continua	Número de obligaciones activas en centrales de riesgo
SALDO_LBZ	Continua	Saldo de las obligaciones clasificadas como libranza del titular en centrales de riesgo
SAL_BASE	Continua	Salario base del titular
TOTALINGRE	Continua	Total de ingresos del titular
TOTALEGRE	Continua	Total de egresos del titular
TOTAL_DESCUENTOS	Continua	Total de descuentos en el desprendible de nómina/pensión del titular
NUM_CREDITOS	Continua	Número de créditos que el titular ha tenido con Avista

CREDITO_PROMEDIO_PORCC	Continua	Monto promedio de crédito que el titular ha tenido con Avista
CREDITO_MAX	Continua	Monto máximo de crédito que el titular ha tenido con Avista
CREDITO_DIF_AVG	Continua	Diferencia entre el crédito promedio del titular con el crédito promedio de todos los clientes de Avista
PLAZO_PROMEDIO_PORCC	Continua	Plazo promedio de crédito que el titular ha tenido con Avista
PLAZO_MAX	Continua	Plazo máximo de crédito que el titular ha tenido con Avista
PLAZO_DIF_AVG	Continua	Diferencia entre el plazo promedio del titular con el plazo promedio de todos los clientes de Avista
CUOTA_MAX	Continua	Cuota máxima que el titular ha tenido con Avista
CUOTA_PROMEDIO_PORCC	Continua	Cuota promedio que el titular ha tenido con Avista
PLAZO_PAGO2	Categórica	Marca binaria para conocer si el titular ya pagó o no sus obligaciones
CREDITO_MARCAMEDIANA	Categórica	Marca binaria para conocer si el crédito promedio del titular se encuentra por encima de la mediana de créditos de todos los clientes de Avista
PLAZO_MARCAMEDIANA	Categórica	Marca binaria para conocer si el plazo promedio del titular se encuentra por encima de la mediana de créditos de todos los clientes de Avista
CUOTA_MARCAMEDIANA	Categórica	Marca binaria para conocer si la cuota promedio del titular se encuentra por encima de la mediana de créditos de todos los clientes de Avista
CAPACIDA	Continua	Capacidad de descuento del titular en su nómina/pensión
SCOREM	Continua	Score de buró del titular
EDAD	Continua	Edad del titular
ANOS_EXPEDICION	Continua	Número de años desde la expedición del documento del titular
SEXO	Categórica	Genero del titular
MARCA_MALO	Categórica	Marca binaria para conocer si el titular es un cliente rentable o no para Avista

Como aclaración las variables referentes al sexo y la marca de malos no se usarán dentro del modelo para evitar sesgos en los clústeres.

Una vez seleccionadas las variables (Tabla 8), se procede a realizar una clasificación de estas en diferentes sets de datos, dividiendo estas entre continuas y categóricas, para proceder a realizar una limpieza final de datos atípicos que pudieran afectar la precisión del modelo posteriormente.

La clasificación de las variables se dio de la siguiente forma:

- Variables continuas

Tabla 10: Variables continuas

PERS_CARGO	CREDITO_MAX
SALDO_VIEGENTE	CREDITO_DIF_AVG
SALDO_MORA	PLAZO_PROMEDIO_PORCC
OBL_ACTIVAS	PLAZO_MAX
SALDO_LBZ	PLAZO_DIF_AVG
SAL_BASE	CUOTA_MAX
TOTALINGRE	CUOTA_PROMEDIO_PORCC
TOTALEGRE	CAPACIDA
TOTAL_DESCUENTOS	SCOREM
NUM_CREDITOS	EDAD
CREDITO_PROMEDIO_PORCC	ANOS_EXPEDICION

- Variables discretas

Tabla 11: Variables discretas

CIUDADEXP	CODOFIC
CIUDADNAC	ESTADCIVI
COD_ESTCIV	ASESOR
COD_NIVEDU	PLAZO_PAGO2
COD_PROFE	CREDITO_MARCAMEDIANA
CIUDAD	PLAZO_MARCAMEDIANA
ESTRATO	CUOTA_MARCAMEDIANA

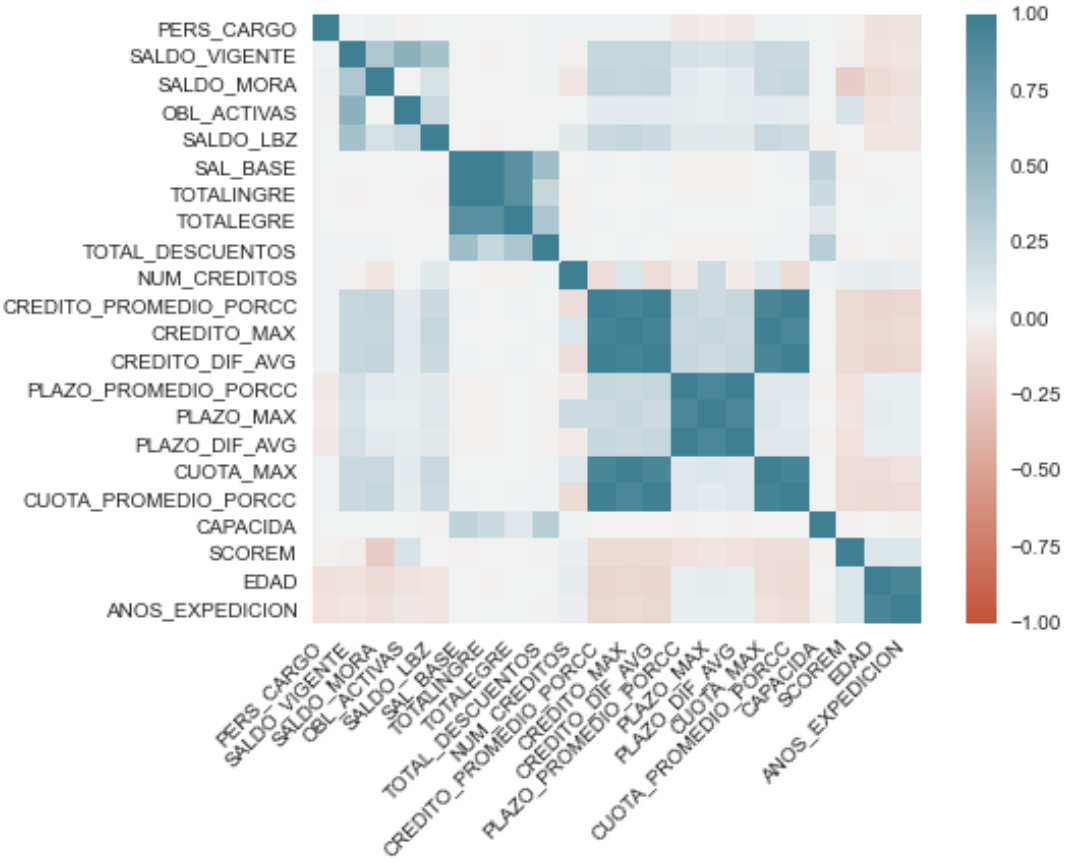
Es importante resaltar que la marca de clientes buenos no está incluida dentro de las variables para modelación ya que estar será usada posteriormente para la caracterización de los clústeres.

Una vez constituidos los sets de datos, se realizó la última fase de limpieza de estos, específicamente en las variables categóricas, donde se evidenciaron valores que no correspondían a una categoría valida dentro de la variable, específicamente en los siguientes casos:

- Estrato: Registros con valor diferente a 1, 2, 3, 4, 5 o 6.
- Ciudad de nacimiento, ciudad de expedición del documento o ciudad de desembolso con longitudes mayores a las estipuladas por tabla DIVIPOLA[3] expedida por el DANE.

Para el caso de las variables continuas, se calculó el coeficiente de correlación de Pearson para conocer la relación entre las variables seleccionadas obteniendo el siguiente mapa de calor:

Ilustración 7: Mapa de calor de correlaciones



[3] DIVIPOLA: División política y administrativa provista por el Departamento Administrativo Nacional de Estadística (DANE), en la cual se muestran todos los departamentos y municipios del territorio nacional.

A partir de este se pudo evidenciar una fuerte correlación entre las variables relacionadas con las estadísticas de créditos de cada cliente además de la información referente al resumen realizado de burós de crédito.

Finalmente, tomando toda la matriz de datos para modelación, se realizó un recuento de valores nulos por variable, para eliminar aquellas que donde los datos faltantes representaran una cantidad considerablemente alta, buscando que estas variables no reduzcan la muestra en el momento de modelar. Los resultados del número de valores nulos por columna fueron los siguientes (las variables con cero valores nulos no se muestran en la lista):

Tabla 12: Recuento de valores nulos en las variables

Variable	No. Nulos	Variable	No. Nulos
TOTAL_DESCUENTOS	12202	COD_ESTCIV	176
TOTALINGRE	6925	COD_NIVEDU	175
ANOS_EXPEDICION	2945	TOTALEGRE	34
CIUDADEXP	439	SAL_BASE	32
CIUDADNAC	425	CODOFIC	26
SALDO_VIGENTE	260	ESTADCIVI	24
SALDO_MORA	260	CAPACIDA	5
OBL_ACTIVAS	260	ESTRATO	3
SALDO_LBZ	260	ASESOR	2
COD_PROFE	215	SEXO	1

Tomando en cuenta el alto número de valores nulos que existen en cada variable, se tomaron las siguientes decisiones de acuerdo con la naturaleza de cada una:

- TOTAL_DESCUENTOS: Los valores nulos serán reemplazados por 0.
- TOTALINGRE: Esta variable será eliminada, puesto que desde el análisis descriptivo se encontró un alto grado de similitud con la variable SAL_BASE.

- ANOS_EXPEDICION: Esta variable será eliminada, dado que posee una fuerte correlación con la variable EDAD.

Respecto a las demás variables, no se ejecutará ninguna acción ya que no representan una gran pérdida de registros respecto a la base.

2.2.5.Transformación de los datos para modelación

Para el uso de las variables categóricas dentro de la fase de modelación se realizaron dos transformaciones al set de datos:

- Distancia de Gower

  En la modelación de escenarios de la vida real, la existencia set de datos con variables de diferente tipología es un hecho, es por estos que se deben implementar métodos como el propuesto por Gower (Gower, 1971) en donde a través de la similitud entre las observaciones se asigna una distancia a partir de la cual el modelo posteriormente podrá interpretar esta como una variable continua.

  La medida de similitud para un par de objetos de datos mixtos d-dimensionales x_i y x_j se define como:

  Ecuación 1: Medida de similitud de Gower

$$S\left(x_{il}(t), x_{jl}(t)\right) = \frac{\sum \delta_{ijl}(t) S_{ijl}(t)}{\sum \delta_{ijl}(t)}$$

  Donde $S_{ijl}(t)$ indica la similitud para la característica *l* entre dos observaciones, y $\delta_{ijl}(t)$ es un coeficiente 0-1 basado en si la medida de los

dos objetos falta en el momento t. Para variables categóricas el componente de similitud es obtenido de la siguiente forma:

Ecuación 2: Componente de similitud

$$S_{ijl}(t) = \begin{cases} 1 \; si \; x_{il}(t) = \; x_{jl}(t) \\ 0 \; si \; x_{il}(t) \neq \; x_{jl}(t) \end{cases}$$

(Akay & Yüksel, 2018)

- Variables Dummy

  "Una variable dummy es un variable numérica que representa un hecho cualitativo o una proposición lógica proposición… Además de los beneficios directos para el análisis estadístico, la representación de la información en forma de variables dummy facilita la conversión del modelo en una herramienta de decisión." (Sharma et al., n.d.)

  Con esta metodología cada posible valor contenido en las variables categóricas toma un valor binario y se convierte en una variable para la interpretación del modelo, por ejemplo, si una de las variables del set de datos tiene 3 posibles valores, esta metodología la transformara en 3 diferentes columnas en donde cada registro obtendrá un 1 o 0 de acuerdo con el valor que posea en esta característica.

Una vez se obtenidas las transformaciones de las variables categóricas, se concatenaron los sets de datos para consolidar una sola base de modelación. El conjunto de datos obtenido a partir de las distancias de Gower y las variables

continuas fue normalizado para asegurar una mejor compresión de los datos en el momento del modelado y unicidad en las escalas de los valores.

2.3. Modelación

En la etapa de modelación la metodología a usar será la aplicación de un modelo de K-Means, una vez se realicé el modelado de los datos, se evaluarán los resultados mediante un *Silhouette Score* el cual permitirá identificar la separación de los clústeres y conocer si el modelo es el indicado para la muestra de datos trabajada.

El *silhouette score* es una métrica que ayuda a entender la separación de los datos dentro de un modelo de segmentación, ya que lo que se espera de los *clústeres* encontrados es que presenten heterogeneidad entre ellos para que así los individuos de cada clúster tengan características que similares entre ellos, pero diferentes con los de otros segmentos (Shahapure & Nicholas, 2020).

Para el cálculo del score mencionado se usó la función de la librería Scikit-learn, la cual da las siguientes indicaciones para la interpretación de este coeficiente:

> *"El mejor valor es 1 y el peor es -1. Los valores cercanos a 0 indican que los clústeres se solapan. Los valores negativos suelen indicar que una muestra ha sido asignada al clúster equivocado, ya que otro clúster es más similar."* (Scikit-Learn, n.d.)

Como paso inicial se realizó una fase iterativa del modelo (K-means), en la cual se ejecutó con un rango de clústeres entre 1 y 6, tomando tanto el set de datos normalizado como sin normalizar, para posteriormente analizar sus resultados. Al finalizar la iteración y a través

de un *Elbow Analysis* se llevó a cabo el análisis del número óptimo de clústeres para el modelo.

> *"La ilustración del Elbow Analysis en los algoritmos de K-Means muestra de forma gráfica la relación de clúster con el error, a medida que va aumentando el valor de K el gráfico disminuirá lentamente hasta que el resultado del valor de K sea estable y allí se podrá encontrar un número óptimo de clústeres para usar en el modelo."* (Syakur et al., 2018)

Luego de ejecutar de forma iterativa el K-Means, se obtuvieron las siguientes representaciones gráficas para el *Elbow Analysis*:

Ilustración 8: Elbow Analysis 1 K-means

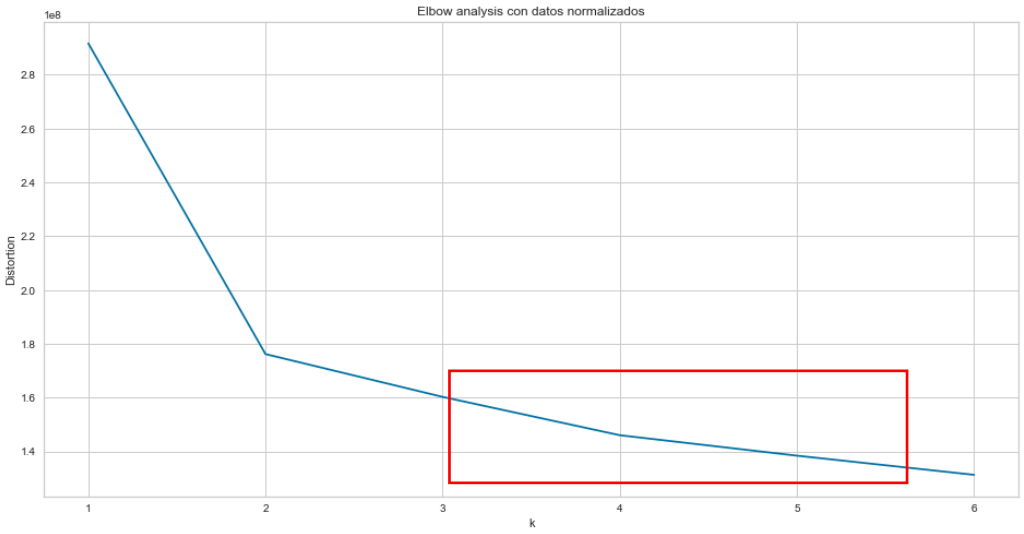
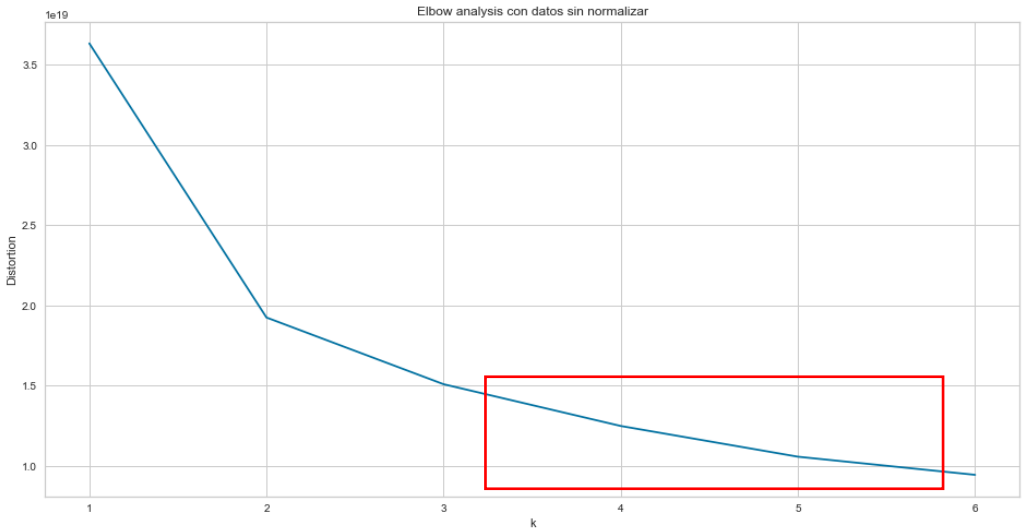


Ilustración 9: Elbow Analysis 2 K-means



A partir de la representación gráfica, se pudo identificar claramente que, para el conjunto de datos normalizados, el punto a partir del cual la línea empieza a tener un decrecimiento lineal se encuentra entre 3 y 5 clústeres, al igual que para el caso de los datos sin normalizar.

Una vez obtenidos el rango de clústeres para cada uno de los sets de datos, se procedió a calcular el silhouette score para cada una de las muestras de datos con cada uno de los valores del rango

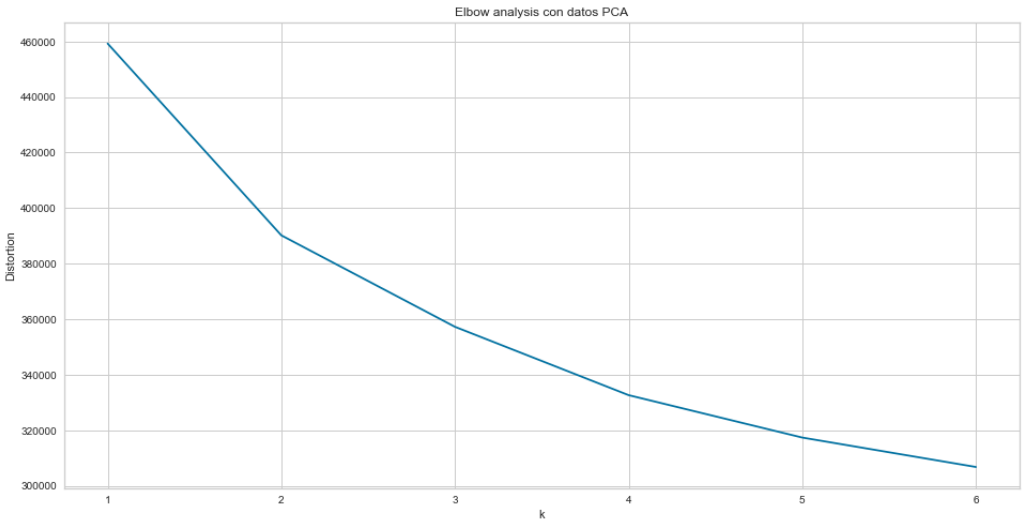
- Datos normalizados con 3 *clústeres*: 0,2586
- Datos normalizados con 4 *clústeres*: 0,1872
- Datos normalizados con 5 *clústeres*: 0,1661
- Datos sin normalizar con 3 *clústeres*: 0,4080
- Datos sin normalizar con 4 *clústeres*: 0,3897
- Datos sin normalizar con 5 *clústeres*: 0,3024

Luego de la modelación a partir de los datos sin transformaciones y también con normalización de los mismos y sus respectivos análisis de desempeño, se eligió como segunda opción para modelar, la aplicación del método de variables dummys, para el uso de las variables categóricas dentro del modelo (Blaufuks, 2021), y análisis de componentes principales (PCA), para la optimización del set de datos completo para finalmente realizar la modelación con este, puesto que la reducción de dimensionalidad que nos otorga esta metodología, puede llegar a que K-Means tenga un mejor ajuste.

Mediante PCA se encontró un nuevo set de dimensiones de acuerdo con la variación de los datos dentro del mismo, logrando reducir la dimensionalidad, puesto que el número de dimensiones será menor al número total de variables de la data original (Jamal et al., 2018). Al aplicar la metodología previamente mencionada, la dimensionalidad del set de datos se reduce de 3.754 variables a 272 dimensiones.

Al aplicar K-Means se obtuvieron los siguientes resultados:

Ilustración 10: Elbow Analysis 3

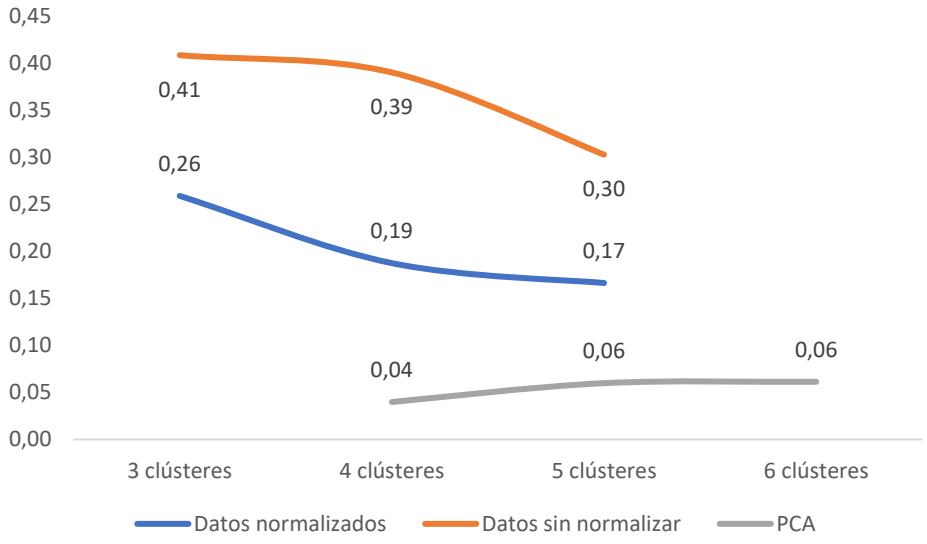


A partir del Elbow Analysis se pudo determinar que el error se estabiliza entre 4 y 6 clústeres y en línea con el procedimiento realizado anteriormente en las otras muestras de datos, se procedió a modelar y calcular el silhouette score para estos modelos con 4, 5 y 6 clústeres, obteniendo los siguientes resultados:

- Datos PCA 3 clústeres: 0,0397
- Datos PCA 4 clústeres: 0,0598
- Datos PCA con 5 clústeres: 0,0612

Una vez evaluadas las 3 opciones de transformaciones para modelación, estimando el número de clústeres a través del elbow analysis podemos ver el comportamiento del score de siluetas de cada uno en la siguiente gráfica:

Ilustración 11: Comparación de elbow analysis



Según los resultados obtenidos, se pudo concluir que el modelo obtiene el mejor score al ser ajustado con 3 clústeres como parámetro y usando los datos sin normalización,

aplicando como única transformación la Distancia de Gower para las variables categóricas. La distribución de individuos a lo largo de los clústeres fue la siguiente:

Tabla : Distribución por clusters

Clúster	Observaciones	%
0	3899	22,89%
1	1479	8,68%
2	11652	68,42%

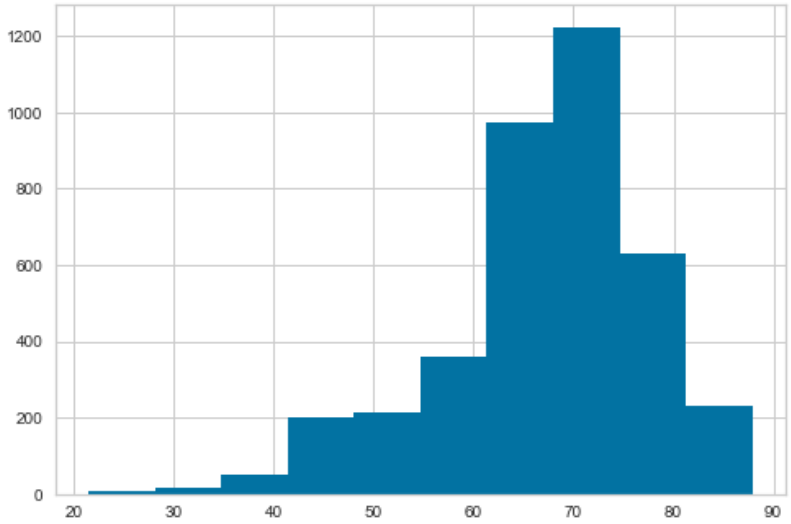
2.4. Caracterización de los clústeres

- Middle age – Low penetration

  Este primer clúster está conformado en su mayoría por hombres ya que el 41,14 % de las observaciones corresponden al género femenino y el 58,86% restante al masculino.

  En términos de edad, las personas de este grupo poseen una edad promedio de 67,24 años y el 50% de esta población supera los 68,63 años, lo que nos permite observar una distribución un poco sesgada a la derecha haciendo referencia a mayor concentración de individuos en edades superiores, adicionalmente estas personas tienen una desviación estándar de 10,28 años, lo que nos permite ver la siguiente distribución:

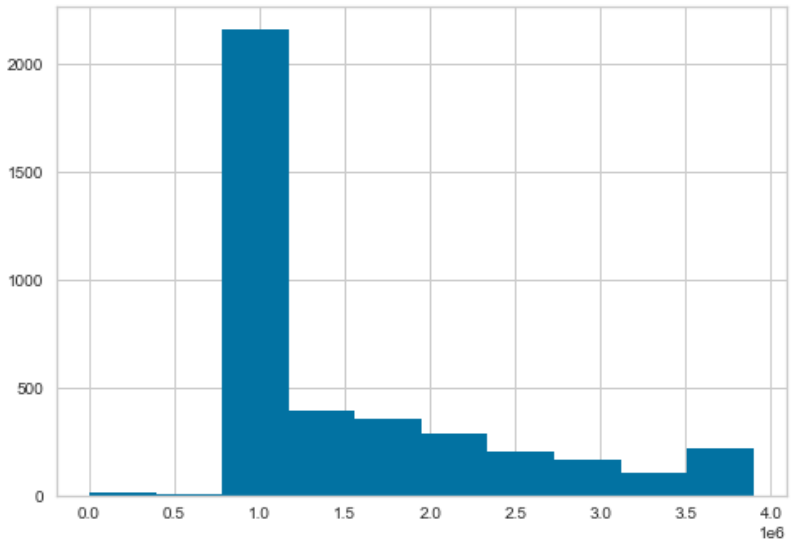
Ilustración 12: Distribución por edad Middle age – Low penetration



En este grupo de individuos, los hombres y las mujeres, en promedio, tiene edades muy similares, ubicadas en 67,18 años y 67,34 años respectivamente.

El 50% de estos individuos tiene un ingreso mensual superior a COP$980.982, lo que se traduce en que la mitad del grupo posee ingresos cercanos al mínimo salario vigente colombiano, sin embargo su salario mensual promedio es de COP$1.511.542, pero existiendo una diferencia entre géneros, donde las mujeres tienen un ingreso medio inferior ubicado en COP$1.504.784 y los hombres en COP$1.516.265, lo que permite concluir que existe una gran concentración de individuos en un nivel de ingresos inferior tomando en cuenta también que su desviación estándar es de COP$887.472, equivalente casi a un salario mínimo colombiano. Respecto a esta variable obtenemos la siguiente distribución:

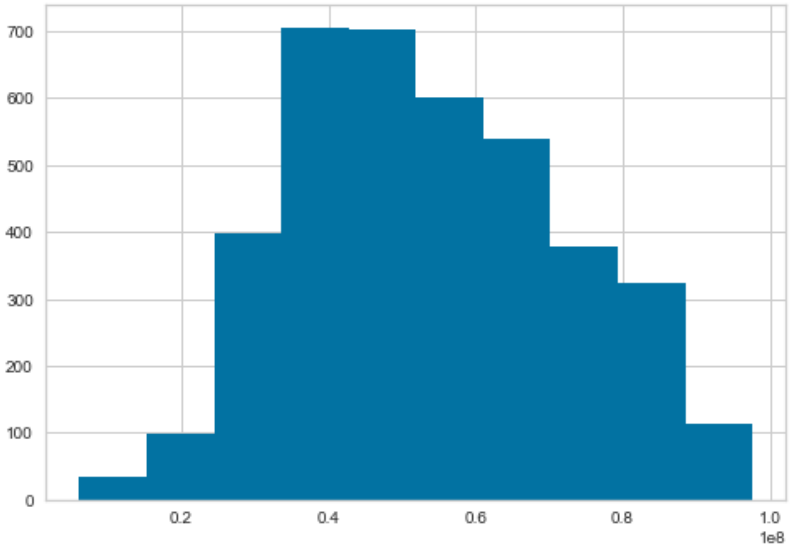
Ilustración 13: Distribución por ingresos Middle age – Low penetration



El histograma que muestra gráficamente la distribución de los ingresos muestra claramente como Avista se centra en atender clientes con un nivel socioeconómico medio bajo y por ende sus ingresos están concentrados alrededor del salario mínimo.

Estos clientes cuentan con un saldo adeudado promedio de COP$53.755.569, presentando un mayor nivel de endeudamiento medio en el género femenino puesto que los hombres tienen un saldo activo promedio de COP53.399.835 y las mujeres de COP$54.264.554. Este grupo presenta gran uniformidad en su distribución respecto a su nivel de endeudamiento, ya que su mediana es de COP$51.999.000, por lo que la mayor parte de los individuos del grupo cuentan con un saldo más alto en cuanto a obligaciones se refiere.

Ilustración 14: Distribución por saldo adeudado – Low penetration



Adicional a lo anterior vemos personas con poca tendencia al deterioro ya que el 50% de las personas poseen un saldo en mora en burós de crédito inferior a COP$1.866.000, que corresponde al 3,59% de su saldo mediano. Además, podemos ver que es un grupo con poca penetración en cuanto a productos de libranza, ya que solo el 30% de todas estas personas tiene un saldo de libranza superior a COP$20.651.000.

El 28,1% de estas personas, representan clientes que no son rentables para la compañía por diferentes causales, en donde los hombres ocupan el primer lugar con una tasa de malos del 28,6% y las mujeres con un 27,3%. Los clientes no rentables se dividen de la siguiente forma bajo las causales definidas:

Tabla 13: Distribución marcas de malo Middle age – Low penetration

Causal	No. Clientes	%
Edad	535	48,90%
Tasa	351	32,08%
Fondeador A	87	7,95%
Seguro	69	6,31%

Reestructurados	37	3,38%
Fianza	15	1,37%
Total general	**1.094**	**100,00%**

Los clientes categorizados como *Middle age – Low penetration* se encuentran distribuidos en 318 municipios colombianos, pero el 80% de ellos están concentrados en solo 32 y 19 departamentos de la siguiente forma:

Tabla 14: Distribución por goegráfica Middle age – Low penetration

Departamento	**No. Clientes**	**%**
Bogotá, D.C.	686	17,59%
Antioquia	367	9,41%
Valle Del Cauca	333	8,54%
Cundinamarca	289	7,41%
Atlántico	219	5,62%
Tolima	182	4,67%
Bolívar	181	4,64%
Santander	178	4,57%
Magdalena	133	3,41%
Cesar	115	2,95%
Córdoba	97	2,49%
Huila	96	2,46%
Quindío	77	1,97%
Risaralda	48	1,23%
Caldas	28	0,72%
Meta	27	0,69%
Norte De Santander	26	0,67%
La Guajira	20	0,51%
Boyacá	18	0,46%

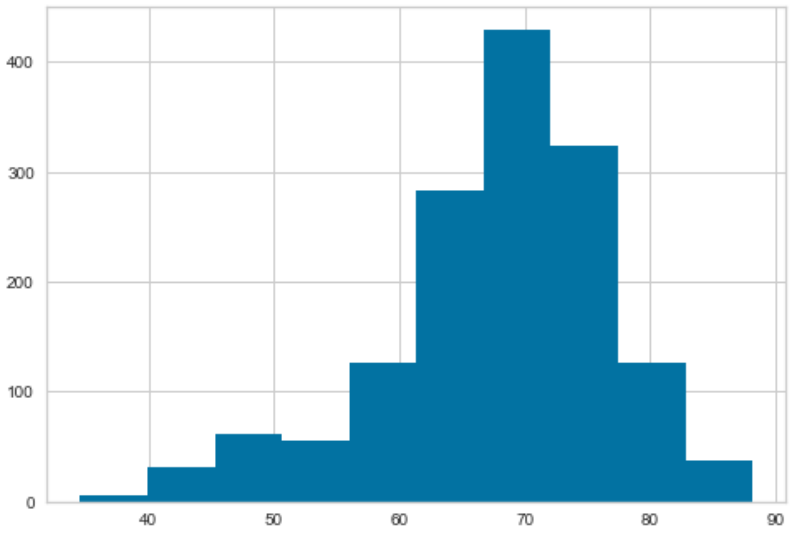
- Lower age – High endebtness

  Para este segundo grupo de personas, observamos una distribución de géneros similar a la del anterior, perteneciendo el 42,12% al género femenino el 57,88% al masculino.

  Son personas más jóvenes con una edad media muy similar de 67,81 años, en donde el género femenino promedio 67,67 años y el masculino 67,91 años. Tomando en cuenta

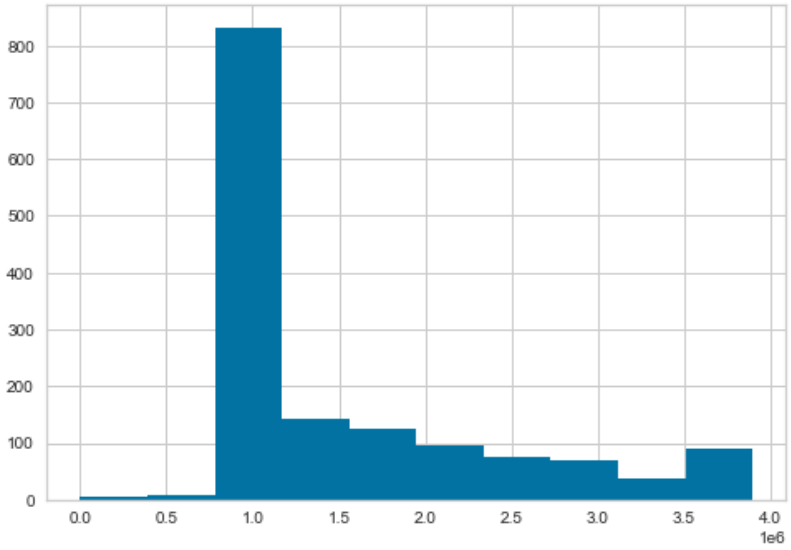
que la mitad de las personas tienen una edad superior a 68,90 años, vemos de la misma forma, que las observaciones se concentran en edades superiores y muy centradas alrededor de los 70 años, alejándose en promedio 9 años de la media de todo el grupo.

Ilustración 15: Distribución por edad Lower age – High endebtness



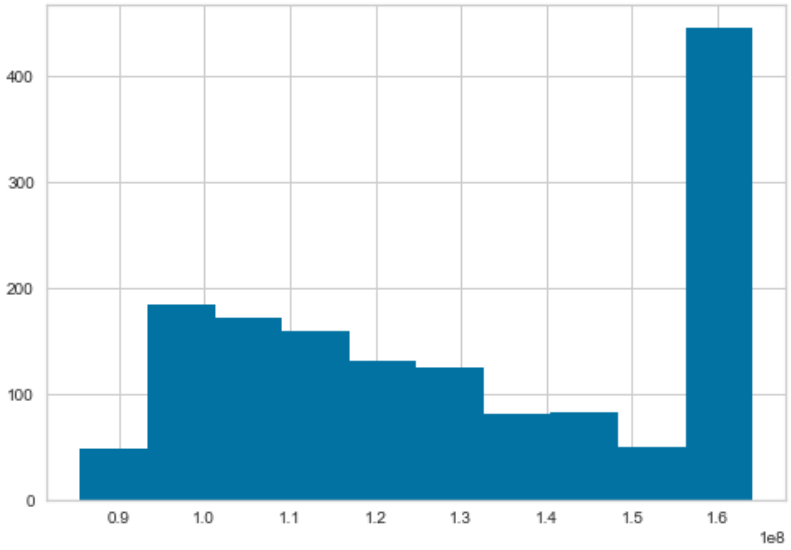
Estas personas al igual que la edad media poseen un ingreso promedio similar ubicado en COP$1.512.162, además, la mitad de estos individuos ganan una menor cantidad de dinero mensual que los del clúster *Middle age – Low penetration*, ya que su mediana de ingreso es de COP$951.762, concentrándose de igual forma en ingresos inferiores, además de tener una mayor desviación estándar ubicada en COP$907.762, lo que se traduce en mayor diferencia en esta variable para los individuos de este grupo.

Ilustración 16: Distribución por ingresos Lower age – High endebtness



Estas personas siendo más jóvenes, cuentan con un mayor saldo promedio adeudado, ubicado en COP131.097.388, en donde los hombres son superados por las mujeres con un saldo medio de COP$130.083.578 y COP$132.490.361 respectivamente, aunque conservando un comportamiento similar al grupo anterior ya que el 50% de las personas ubicadas en el este clúster supera COP$127.474.000 de saldo activo en burós de crédito, obteniendo la siguiente distribución respecto a esta variable:

Ilustración 17: Distribución por saldo adeudado Lower age – High endebtness



Estas personas al mostrar un nivel superior de endeudamiento presentan también menor cuantía en cuanto al deterioro de sus obligaciones, ya que el 50% de estas personas posee un saldo en mora mayor a $1.465.000 en centrales de riesgo, siendo clientes menos morosos que los del clúster anterior. El 50% de las personas poseen un saldo inferior a COP$5.294.000 en obligaciones de libranza, esto deja concluir que poseen poca penetración en cuanto a este tipo de obligaciones en su portafolio financiero.

Dentro de las personas del clúster *Lower age – High endebtness* se presenta una tasa de clientes no rentables más baja que la del grupo anterior, ubicada en 26,2% en donde hombres y mujeres posee tasas similares de 26,6% y 25,5% respectivamente. La distribución de clientes malos entre las causales definidas es la siguiente:

Tabla 15: Distribución por marca de malos Lower age – High endebtness

Causal	No. Clientes	%
Edad	190	49,10%

Tasa	123	31,78%
Fondeador A	35	9,04%
Seguro	27	6,98%
Reestructurados	6	1,55%
Fianza	6	1,55%
Total general	**387**	**100,00%**

Finalmente, este grupo se encuentra distribuido a lo largo de 212 municipios en el territorio nacional, concentrándose el 80% de los individuos en 33 de ellos y 18 departamentos de la siguiente forma:

Tabla 16: Distribución geográfica Lower age – High endebtness

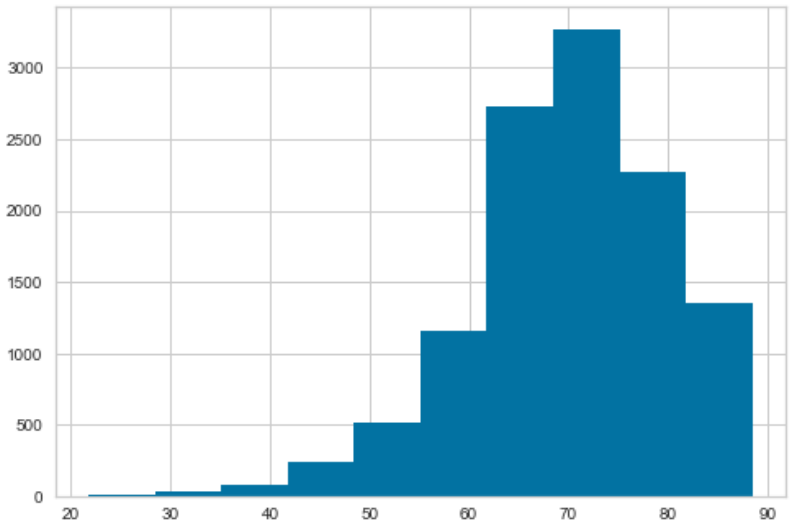
Departamento	**No. Clientes**	**%**
Bogotá, D.C.	226	15,28%
Valle Del Cauca	118	7,98%
Bolívar	110	7,44%
Magdalena	96	6,49%
Antioquia	88	5,95%
Cesar	76	5,14%
Tolima	75	5,07%
Atlántico	68	4,60%
Santander	67	4,53%
Cundinamarca	53	3,58%
Córdoba	48	3,25%
Quindío	46	3,11%
Huila	40	2,70%
La Guajira	19	1,28%
Risaralda	18	1,22%
Meta	13	0,88%
Norte De Santander	13	0,88%
Caldas	12	0,81%

- High Income – Low indebtness

  En este clúster es donde encontramos mayor presencia femenina, conformado en un 48,82% por mujeres y un 51,18% por hombres, pero edades medias superiores que los grupos anteriores, ya que le edad media femenina se ubica en 68,63 años y la masculina

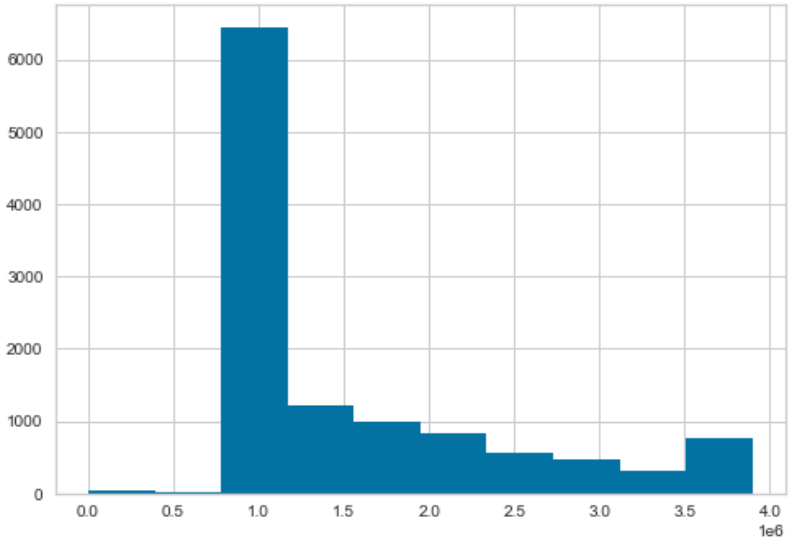
en 71,27 años, además el 50% de toda esta muestra se ubica por encima de los 70,56 años, convirtiéndose en el grupo con edades mayores y conservando la distribución asimétrica a la derecha de la siguiente forma:

Ilustración 18: Distribución por edad High Income – Low indebtness



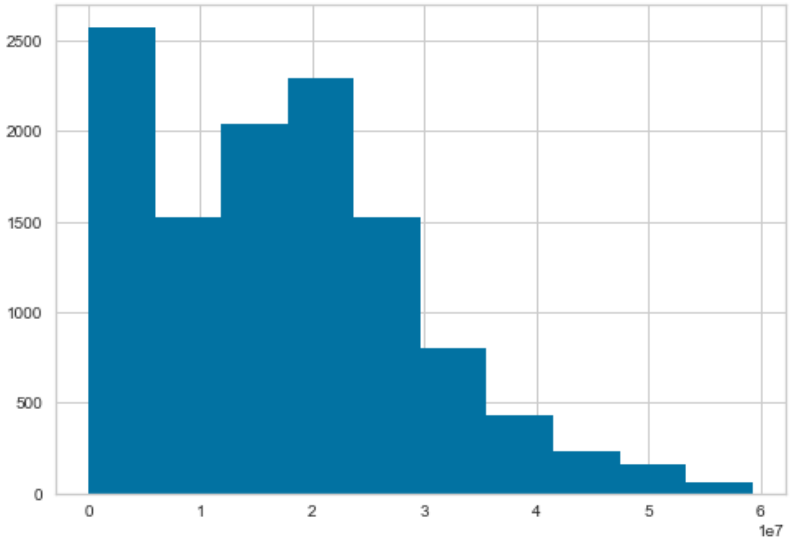
Este grupo es el que posee los ingresos medios más altos de los 3 ubicados en COP$1.519.397, en donde las mujeres están ubicadas en COP$1.519.477 y los hombres en COP$1.519.320, siendo el grupo en donde los géneros tienen el ingreso medio más similar, además el 50% del grupo posee ingresos superiores a COP$975.555, esto da como conclusión que, en términos de ingresos, este grupo conserva la distribución de los dos anteriores, posicionándose como el grupo con mejor ingreso medio y mediano La distribución del ingreso para estos individuos es la siguiente:

Ilustración 19: Distribución por ingresos High Income – Low indebtness



A diferencia de los otros individuos, estos, posee el nivel de endeudamiento más bajo de los 3 grupos, ya que su media es de COP$17.291.661 y su percentil 50 se encuentra ubicado en COP$16.995.500, rompiendo la tendencia en cuanto a géneros, donde en los otros clústeres las mujeres presentan mayor nivel de endeudamiento que los hombres, aquí que sus promedios son COP$17.046.460 y COP$17.525.595 respectivamente. Su saldo vigente en cuanto a obligaciones se encuentra fuertemente sesgado a la izquierda, lo que quiere decir que estas personas poseen saldos bajos en cuanto a deudas, como se puede observar en la siguiente gráfica:

Ilustración 20: Distribución por saldo adeudado High Income – Low indebtness



Además de tener el endeudamiento más bajo, son los clientes con menor tendencia al deterioro de obligaciones ya que el 50% de ellos posee saldos en mora inferiores a COP$178.500 y en promedio toda la muestra tiene el menor valor en mora en sus obligaciones ubicado en COP$2.073.875, así mismo existe baja penetración de productos de libranza en este clúster, ya que la mitad de los individuos no tiene ningún tipo de saldo activo referente a este tipo de obligaciones y solo el 20% supera los COP$4.596.400.

Inverso a su tendencia al deterioro, el clúster *High Income – Low indebtness*, presenta la mayor tasa de clientes no rentables ubicada en 32,6% y en línea con los otros grupos en donde las mujeres tienen una menor tasa de malos que los hombres, alcanzando 31,0% y 34,2% respectivamente. Los clientes con la marca mencionada previamente se distribuyen de la siguiente forma:

Tabla 17: Distribución por marca de malos High Income – Low indebtness

Causal	No. Clientes	%
Tasa	2.420	63,63%
Fondeador A	763	20,06%
Seguro	330	8,68%
Fianza	195	5,13%
Reestructurados	50	1,31%
Edad	45	1,18%
Total general	**3.803**	**100,00%**

Por último, estas personas se encuentran ubicadas en 421 municipios, de los cuales 33 contienen el 80% de ellas, ubicándose así en 16 departamentos como lo ilustra la siguiente tabla:

Tabla 18: Distribución geográfica High Income – Low indebtness

Departamento	No. Clientes	%
Antioquia	1515	13,00%
Bogotá, D.C.	1315	11,29%
Cundinamarca	1110	9,53%
Valle Del Cauca	1094	9,39%
Tolima	649	5,57%
Santander	619	5,31%
Atlántico	617	5,30%
Bolívar	443	3,80%
Magdalena	441	3,78%
Risaralda	314	2,69%
Córdoba	302	2,59%
Cesar	260	2,23%
Quindío	222	1,91%
Huila	221	1,90%
Caldas	161	1,38%
Meta	54	0,46%

2.4.1.Selección de variables del modelo

Una ejecutado el modelo y caracterizados los individuos mediante la clasificación realizada a partir del mismo, se continuó con la ejecución de un modelo que

permitiera conocer la relevancia de cada una de las variables incluidas el K-means previamente ejecutado y que permitiera realizar una selección de estas para optimizar el gasto computacional y en futuras corridas, omitir las variables que no resulten relevantes para este proceso.

Para el fin mencionado anteriormente se eligió un modelo XGBoost, este tipo de modelo fue elegido gracias a que presenta un excelente desempeño en términos de rapidez, eficiencia y escalabilidad para la selección de variables usadas la clasificación de un resultado, adicionalmente este mostrará cuales son las variables necesarias para llegar a la respuesta, que, en este caso, sería el clúster en el cual la observación quedó catalogada (Wang & Ni, n.d.).

Previo a la modelación con K-means, se realizaron una serie de iteraciones, buscando el mejor resultado a través del *Silhouette Score*, entre las cuales se ejecutaron las siguientes:

- Transformación de las variables categóricas con la distancia de *Gower*, para posteriormente unir la matriz resultante con las variables continuas y a través de un proceso de normalización obtener una matriz para ejecutar la clasificación de través de K-means.
- Transformación de las variables categóricas con la distancia de *Gower*, para posteriormente unir la matriz resultante con las variables continuas y ejecutar la clasificación de través de K-means sin ninguna transformación adicional.
- Transformación de las variables categóricas a través de un proceso de conversión binaria (Variables *Dummys*), para posteriormente unir la matriz

resultante con las variables continuas y ejecutar un proceso de reducción de dimensionalidad a través de *análisis de componentes principales (PCA)* para finalmente realizar la clasificación a través de K-means.

De las 3 iteraciones, con la que mejor desempeño tuvo K-means fue con la transformación mediante distancia de *Gower* sin normalización posterior de datos. Tomando en cuenta lo anterior y debido a que el modelo XGBoost requiere de valores numéricos para su ejecución, se ejecutó un *Label Encoding*, procedimiento que permitió asignar categorías numéricas a cada uno de los valores contenidos en las variables categóricas, para que así, el XGBoost, pudiera definir qué características llevan a una observación a ser clasificada bajo cada uno de los clústeres definidos.

Una vez ejecutado el XGBoost, los siguientes fueron los resultados de las variables en términos de importancia para la clasificación de las observaciones:

Tabla 19: Importancia de las variables con XGBoost

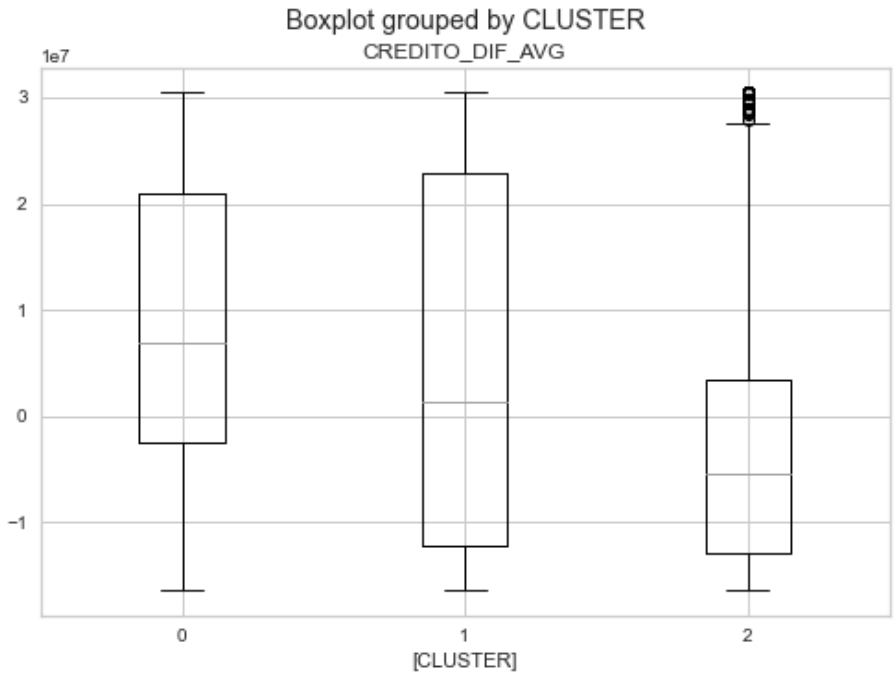
Variable	Importancia	Variable	Importancia
OBL_ACTIVAS	51,18%	ESTRATO	0,29%
CREDITO_DIF_AVG	17,40%	CODOFIC	0,28%
PLAZO_PROMEDIO_PORCC	13,57%	COD_PROFE	0,26%
TOTALEGRE	4,27%	COD_NIVEDU	0,26%
SALDO_LBZ	2,43%	CIUDAD	0,26%
SCOREM	2,08%	COD_ESTCIV	0,24%
CAPACIDA	1,86%	TOTAL_DESCUENTOS	0,21%
CREDITO_MAX	1,41%	PLAZO_DIF_AVG	0,20%
SAL_BASE	0,53%	CUOTA_MAX	0,19%
SALDO_MORA	0,48%	ESTADCIVI	0,04%
CIUDADEXP	0,44%	SALDO_VIGENTE	0,00%
EDAD	0,40%	CUOTA_PROMEDIO_PORCC	0,00%
PLAZO_MAX	0,38%	ASESOR	0,00%
CIUDADNAC	0,38%	PLAZO_PAGO2	0,00%
CREDITO_PROMEDIO_PORCC	0,32%	CREDITO_MARCAMEDIANA	0,00%

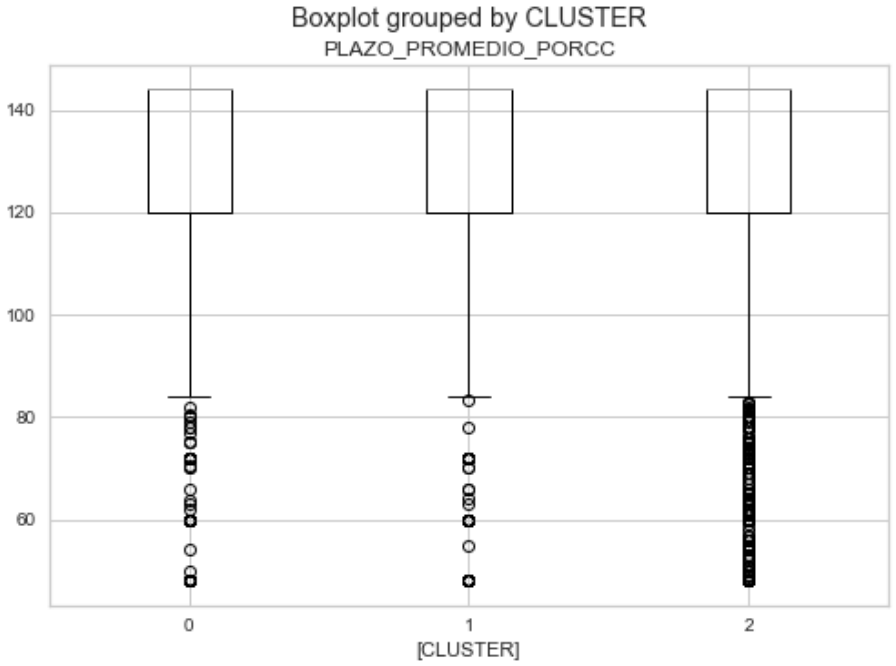
NUM_CREDITOS	0,32%	PLAZO_MARCAMEDIANA	0,00%
PERS_CARGO	0,32%	CUOTA_MARCAMEDIANA	0,00%

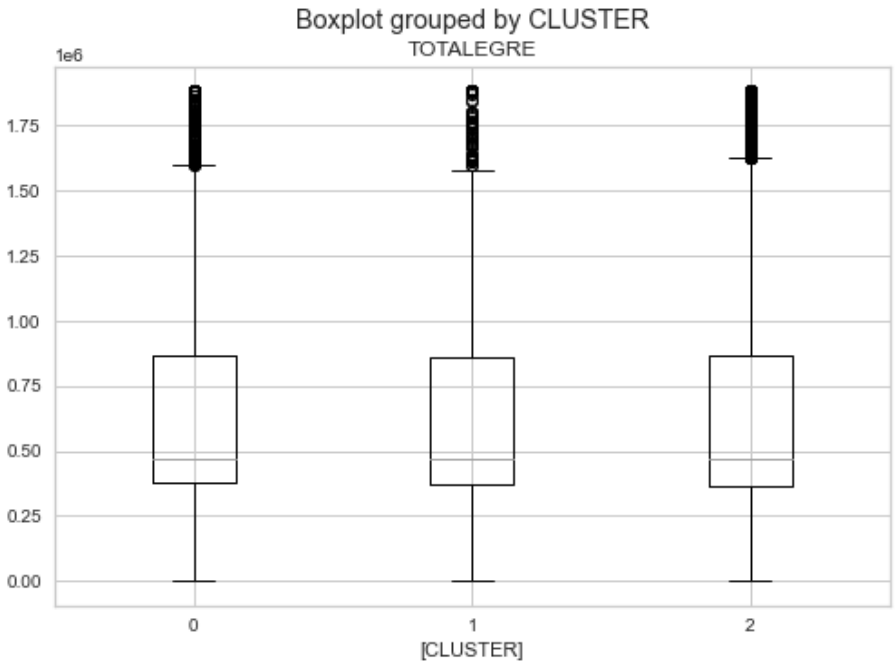
De lo anterior se puede concluir de las 34 variables usadas para la modelación con K-means, 8 de ellas aportan información al modelo en términos de clasificación, por lo tanto 26 de ellas pueden ser omitidas al momento de la modelación, sin embargo, todas aportan información en términos del perfilamiento de las personas incluidas en cada uno de los clústers.

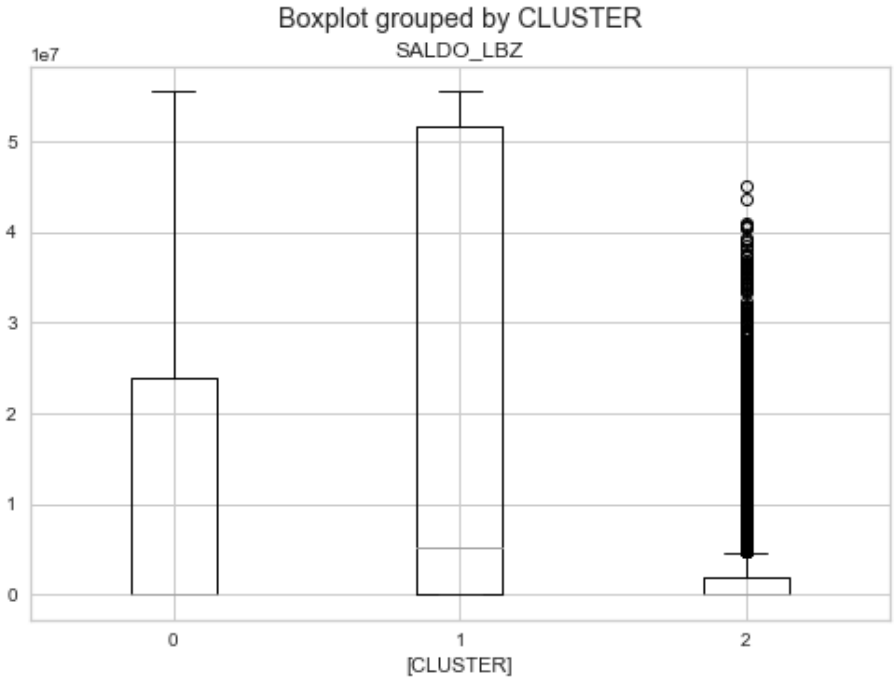
Dado lo anterior, podemos observar el siguiente comportamiento para cada una de las variables continuas del modelo:

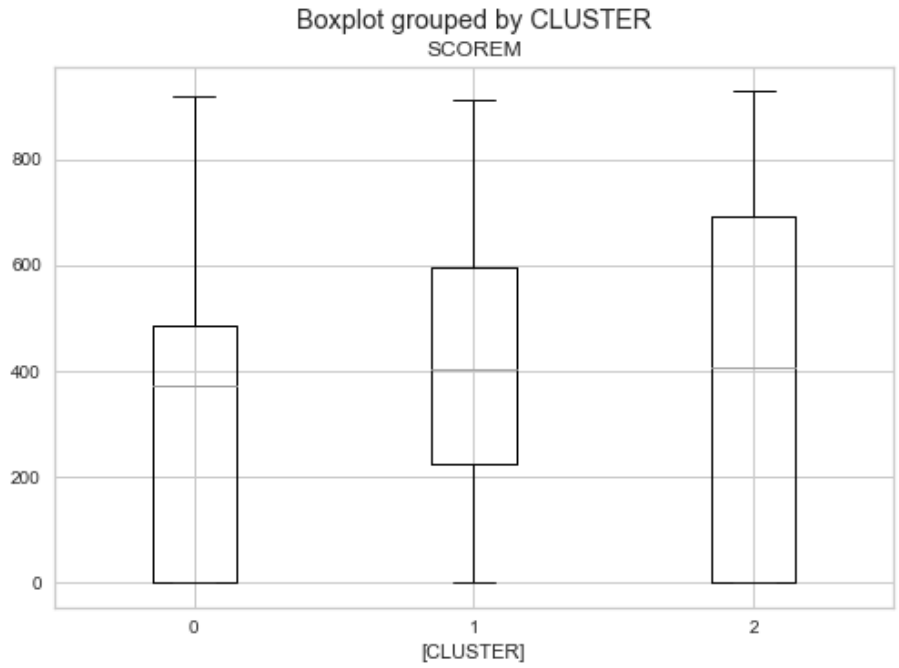
Ilustración 21: Comparativo de clústeres por sus variables

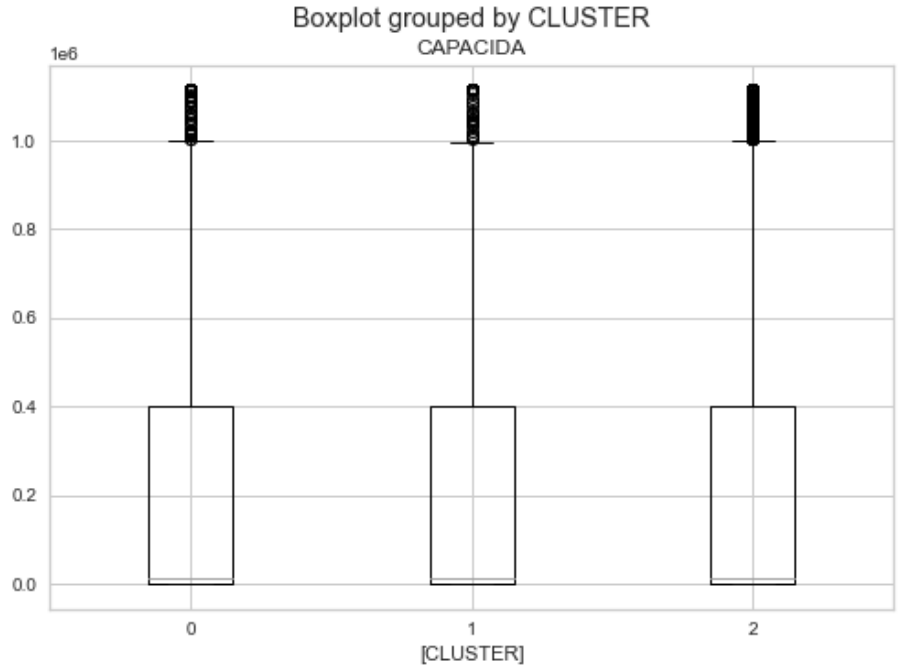

Boxplot grouped by CLUSTER
CREDITO_DIF_AVG
1e7
3
2
1
0
−1
0
1
2
[CLUSTER]

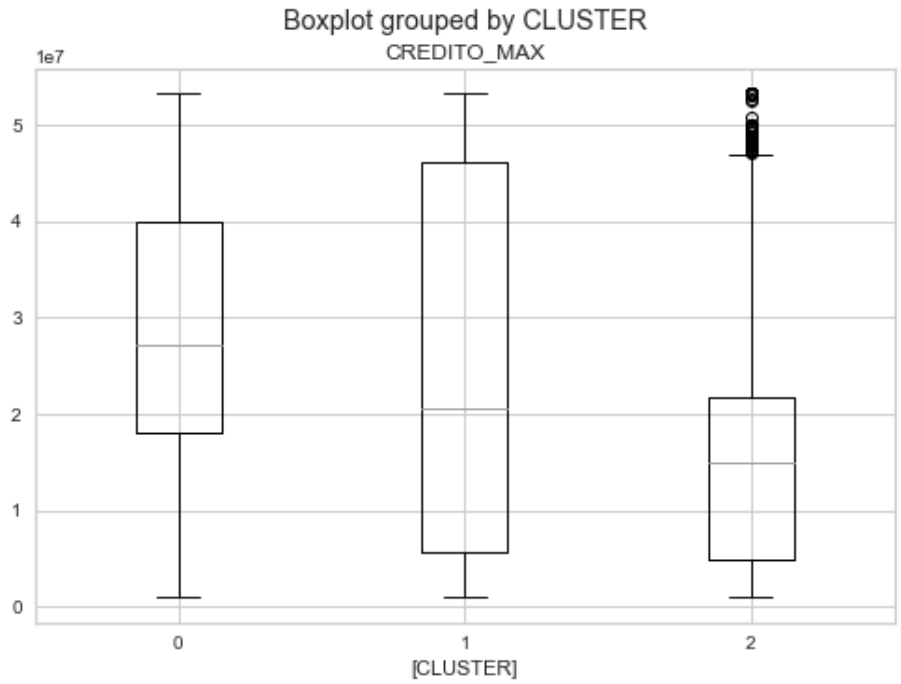

Boxplot grouped by CLUSTER
PLAZO_PROMEDIO_PORCC
140
120
100
80
60
0
1
2
[CLUSTER]


Boxplot grouped by CLUSTER
TOTALEGRE
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
1
2
[CLUSTER]


Boxplot grouped by CLUSTER
SALDO_LBZ
1e7
5
4
3
2
1
0
0
1
2
[CLUSTER]


Boxplot grouped by CLUSTER
SCOREM
800
600
400
200
0
0
1
2
[CLUSTER]


Boxplot grouped by CLUSTER
CAPACIDA
1e6
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
[CLUSTER]



En la mayoría de las variables continuas relevantes dentro del modelo, podemos encontrar separación entre los individuos de los 3 clústeres, lo que permite concluir que existen diferencias relevantes entre los mismos, lo que permite que el modelo realice una agrupación apropiada de las observaciones.

2.5. Cálculo del Customer Life Value por cluster

Como se mencionó en la metodología, el cálculo del *Customer Life Value* será realizado para cada uno de los clústeres, sin embargo, se usará una metodología diferente a la usada tradicionalmente por las entidades financieras, ya que debido al gran número de cambios que ha tenido la compañía en su estructura de costos y cobros a clientes no es posible determinar crédito a crédito la utilidad de cada obligación.

Para realizar este cálculo se tomarán 4 parámetros, estos serán:

- Cuota de la obligación (CO): Este será determinado a partir del valor en el cual se ubique el percentil 50 del clúster.

- Edad del crédito (EC): Este será determinado a partir del percentil 50 de cada uno de los clústeres
- Tasa de default (TD): Esta tasa de default, será la tasa de clientes malos encontrada para cada uno de los clústeres.
- Número de clientes (NC): Número de clientes que se encuentran en el clúster

Una vez encontrados estos 3 valores, se aplicará la siguiente fórmula:

Ecuación 3: Cálculo del CLV

$$CLV\ Cluster_i = CO * (1 - TD) * EC * NC$$

El cálculo fue ejecutado para cada uno de los departamentos del territorio nacional diferenciado por cada uno de los clústeres, de igual forma, los parámetros fueron calculados a nivel de departamento y clúster para no generalizar los posibles comportamientos, obteniendo los siguientes resultados:

Tabla 20: Cálculo del CLV por cluster

	Cluster			
Departamento	**0**	**1**	**2**	**Total general**
Amazonas	130.199.616		46.554.480	176.754.096
Antioquia	19.336.310.400	4.960.218.816	38.864.833.128	63.161.362.344
Arauca	226.987.920	157.954.896	133.758.720	518.701.536
Archipiélago De San Andrés, Providencia Y Santa Catalina	-	111.806.208		111.806.208
Atlántico	11.065.489.344	3.616.604.640	21.091.466.448	35.773.560.432
Bogotá, D.C.	39.803.083.128	10.428.861.024	47.824.794.000	98.056.738.152
Bolívar	9.830.950.668	4.170.842.280	13.143.720.960	27.145.513.908
Boyacá	4.368.735.360	402.915.168	1.751.894.400	6.523.544.928
Caldas	2.210.608.512	634.730.976	3.838.750.380	6.684.089.868
Caquetá	386.148.816	29.643.984	-	415.792.800
Casanare	749.988.360	38.735.880	346.216.896	1.134.941.136
Cauca	1.723.407.840	384.725.376	1.964.075.616	4.072.208.832

Cesar	6.293.318.400	2.309.857.200	6.758.412.672	15.361.588.272
Chocó	423.013.185	130.199.616	74.147.766	627.360.567
Córdoba	5.856.303.548	865.678.320	8.312.085.000	15.034.066.868
Cundinamarca	16.131.467.520	4.157.393.508	33.721.904.502	54.010.765.530
Huila	7.891.485.696	3.376.409.880	10.690.008.912	21.957.904.488
La Guajira	3.270.994.704	2.469.189.096	1.662.209.208	7.402.393.008
Magdalena	7.966.516.896	3.770.476.920	13.247.782.080	24.984.775.896
Meta	2.257.637.400	711.544.392	2.445.398.208	5.414.580.000
Nariño	1.451.293.200	580.625.856	1.786.035.456	3.817.954.512
Norte De Santander	1.851.136.848	795.587.040	1.276.496.136	3.923.220.024
Putumayo	104.354.784	119.349.648	47.965.680	271.670.112
Quindío	3.763.279.080	1.437.016.581	5.253.930.480	10.454.226.141
Risaralda	2.798.595.072	415.611.000	6.869.274.624	10.083.480.696
Santander	7.462.413.420	2.086.456.086	15.060.651.606	24.609.521.112
Sucre	1.562.908.608	520.798.464	1.176.971.904	3.260.678.976
Tolima	9.085.847.628	2.328.650.928	16.029.990.900	27.444.489.456
Valle Del Cauca	20.403.675.084	3.814.133.400	31.326.463.584	55.544.272.068
Vichada	316.606.752	69.090.840	181.597.680	567.295.272

2.5.1. Proyección sobre el crecimiento del mercado nacional de pensionados

En Colombia no hay una fuente que provea información sobre el número exacto de pensionados que existe en el país, es por esto por lo que, para la estimación de esta cifra, se realizaron cálculos de forma particular tomando fuentes verificadas sobre información poblacional y datos sobre el mercado de personas pensionadas en Colombia provistas por el DANE y Colpensiones respectivamente.

En primer lugar, se calculó el número de personas pensionadas en el país a cierre del año 2020, para esto se usó la información provista por Colpensiones en su Informe de gestión 2020, en donde la entidad expone que posee una tercera parte de los afiliados al sistema de pensiones colombiano, alcanzando 6.811.214 pensionados afiliados a su institución (Colpensiones, 2020), lo que permite conocer que la cifra total de personas afiliadas al sistema colombiano de pensiones para 2020 fue de 20.640.042.

El Departamento Administrativo Nacional de Estadística (DANE), estimó que la población total colombiana para 2020 era de 50.372.424 personas y el número correspondiente a personas mayores de 18 años, edad en la cual la persona puede acceder a un empleo, era de 36.078.248 (Departamento Administrativo Nacional de Estadística (DANE), 2018). Contrastando estas cifras con las obtenidas a partir del Informe de Gestión 2020 de Colpensiones, se puede estimar que el 57% de los habitantes colombianos con edad suficiente para trabajar, aportan al sistema pensional colombiano y por ende tienen o tendrán el derecho de acceder a una mesada pensional.

Bajo los parámetros previamente expuestos se realizó el siguiente cálculo para estimar el crecimiento del segmento de personas pensionadas en el país.

Tabla 21: Proyección del mercado de pensionados colombiano

Año	Total población	Población 18+	Hombres 62+	Mujeres 57+	Total objetivo	Pensionados	Tasa de crecimiento
2020	50.372.424	36.078.248	2.643.697	4.558.853	7.202.550	4.120.514	-
2021	51.049.498	36.759.634	2.752.567	4.756.891	7.509.458	4.296.094	4,26%
2022	51.609.474	37.351.267	2.864.169	4.954.727	7.818.896	4.473.120	4,12%
2023	52.156.254	37.927.237	2.979.733	5.152.688	8.132.421	4.652.485	4,01%
2024	52.691.440	38.516.633	3.066.832	5.305.101	8.371.933	4.789.508	2,95%
2025	53.216.592	39.085.429	3.182.466	5.493.326	8.675.792	4.963.343	3,63%
2026	53.732.415	39.648.373	3.299.725	5.678.455	8.978.180	5.136.336	3,49%

Según lo anterior podemos encontrar que el segmento de personas pensionadas en Colombia tendrá un crecimiento compuesto entre 2020 y 2026 del 3,2% para el total del país, pero debido a la necesidad de conocer el crecimiento compuesto para cada uno de los departamentos del territorio nacional, se desagregó la información poblacional de forma departamental, tomando la población mayor a 18 años entre

2020 y 2026 y multiplicándola por el factor de pensionados previamente encontrado (51%) para así obtener el crecimiento compuesto en cada zona geográfica, obteniendo los siguientes resultados:

Tabla 22: Crecimiento departamental compuesto

Departamento	**Crecimiento compuesto**
Amazonas	2,81%
Antioquia	1,71%
Arauca	1,53%
Archipiélago de San Andrés	0,87%
Atlántico	1,52%
Bogotá, D.C.	1,16%
Bolívar	1,38%
Boyacá	0,88%
Caldas	1,04%
Caquetá	1,61%
Casanare	1,03%
Cauca	1,09%
Cesar	1,81%
Chocó	1,50%
Córdoba	0,98%
Cundinamarca	2,78%
Guainía	2,50%
Guaviare	2,39%
Huila	1,09%
La Guajira	1,92%
Magdalena	1,46%
Meta	1,12%
Nariño	0,45%
Norte de Santander	0,94%
Putumayo	1,76%
Quindío	1,29%
Risaralda	1,13%
Santander	1,04%
Sucre	1,35%
Tolima	0,66%
Valle del Cauca	1,05%
Vaupés	5,59%
Vichada	2,01%

Por último, para conocer el potencial de cada departamento, caracterizado por la cartera actual de Avista, se multiplicó el Customer Life Value de cada uno de los clústeres por (1 + Crecimiento Compuesto), para así, finalmente encontrar el potencial de mercado en un horizonte de 5 años respecto al 2021. Los resultados obtenidos en orden descendente fueron los siguientes:

Tabla 23: Proyección departamental del CLV

Departamento	CLV	Crecimiento compuesto	Proyección
Bogotá, D.C.	98.056.738.152	1,16%	99.195.638.066
Antioquia	63.161.362.344	1,71%	64.239.228.015
Valle Del Cauca	55.544.272.068	1,05%	56.125.042.622
Cundinamarca	54.010.765.530	2,78%	55.513.893.731
Atlántico	35.773.560.432	1,52%	36.315.815.104
Tolima	27.444.489.456	0,66%	27.624.611.862
Bolívar	27.145.513.908	1,38%	27.519.800.113
Magdalena	24.984.775.896	1,46%	25.350.096.778
Santander	24.609.521.112	1,04%	24.865.560.158
Huila	21.957.904.488	1,09%	22.196.186.558
Cesar	15.361.588.272	1,81%	15.639.193.728
Córdoba	15.034.066.868	0,98%	15.181.247.130
Quindío	10.454.226.141	1,29%	10.589.517.245
Risaralda	10.083.480.696	1,13%	10.197.556.573
La Guajira	7.402.393.008	1,92%	7.544.773.068
Caldas	6.684.089.868	1,04%	6.753.561.857
Boyacá	6.523.544.928	0,88%	6.580.929.085
Meta	5.414.580.000	1,12%	5.475.322.380
Cauca	4.072.208.832	1,09%	4.116.653.515
Norte De Santander	3.923.220.024	0,94%	3.960.176.426
Nariño	3.817.954.512	0,45%	3.835.295.859
Sucre	3.260.678.976	1,35%	3.304.811.725
Casanare	1.134.941.136	1,03%	1.146.673.192
Chocó	627.360.567	1,50%	636.764.996
Vichada	567.295.272	2,01%	578.680.807
Arauca	518.701.536	1,53%	526.636.025
Caquetá	415.792.800	1,61%	422.479.076
Putumayo	271.670.112	1,76%	276.442.043

Amazonas	176.754.096	2,81%	181.721.760
Archipiélago De San Andrés, Providencia Y Santa Catalina	111.806.208	0,87%	112.776.231

3. Conclusiones

A partir del desarrollo del trabajo para la encontrar las zonas geográficas que representen un mejor mercado en términos de utilidad para la compañía podemos concluir lo siguiente:

- Los clientes de la compañía pueden ser caracterizados con pocas variables, sin embargo, al originar un único producto, la segmentación de estos puede verse un poco sesgada, ya que, en las características sociodemográficas de la muestra analizada, no se puede observar alta variabilidad. Lo anterior se traduce en grupos altamente uniformes que llegan a ser segmentados más por su perfil financiero que por su perfil sociodemográfico.
- La compañía debe realizar esfuerzos en la estandarización de su data para lograr construir capacidades analíticas robustas, ya que se encontraron diversos problemas con diferentes datos y variables. Dentro de la base de datos se encontraron muchas tablas con variables en donde la información no correspondía a la variable analizada o simplemente se encontraba vacía sin posibilidad de hacer una inferencia para encontrar su valor.
- La compañía posee grandes volúmenes de datos, sin embargo, muchos de ellos no poseen utilidad en términos de análisis, por ende y en línea con la conclusión anterior, se recomienda hacer un análisis de la data para definir desde la perspectiva del negocio, la utilidad que puede tener y así, realizar una depuración de los datos que estén generando ruido y no tengan ningún papel relevante dentro de los procesos internos.

4. Discusión

En primer lugar, Avista cuenta con una base de datos donde existen múltiples tablas con problemas de diseño, ya que albergar demasiadas variables, de las cuales muchas se encuentran totalmente en desuso o se tiene pleno desconocimiento desde el negocio de la utilidad de las mismas debido a que toda infraestructura fue heredada de una compañía con un modelo de operación diferente y por ende, aunque se adapta y funciona para la operación actual, existen muchos datos que a primera vista no son de utilidad para extraer información.

En línea con lo anterior, desde una mirada técnica, la compañía debe trabajar en analizar, limpiar y extraer los datos valiosos, realizar una reclasificación de los campos y en conjunto con expertos en data, pensar en migrar las tablas y datos existentes a nuevas estructuras que le permitan a Avista aprovechar estos y avanzar hacía una infraestructura analítica potenciando su negocio a partir del uso de la información obtenida producto de la operación.

Además de desarrollar una infraestructura analítica corporativa, Avista podría empezar a pensar en la creación de modelos referentes al sector de libranza, que históricamente no ha sido abarcado desde una perspectiva analítica. Los modelos basados en inteligencia artificial y machine learning, le podrían dar a la empresa una ventaja frente a sus competidores, logrando entender de una mejor forma las estacionalidades asociadas al sector y al tipo de clientes, además de poder aprender cada vez más de sus clientes y lograr orientar sus estrategias de mercadeo de una forma más eficiente y por último en una etapa más avanzada, lograr una comercialización de su conocimiento del sector a través de diferentes modelos analíticos.

Adicionalmente los resultados obtenidos está totalmente alineados con la realidad del negocio, debido a que la concentración de los posibles y actuales clientes está relacionada directamente

con los tamaños poblacionales de cada territorio y las oportunidades más grandes de la compañía se dan en los mercados más poblados, algo que se encuentra sucediendo actualmente en la compañía y que confirma que las estimaciones realizadas son válidas desde una perspectiva de generación de utilidades presentes y futuras.

Desde el foco comercial y como se puede evidenciar en los resultados finales, la compañía debe enfocar su estrategia en los mercados localizados en los departamentos que contienen las principales ciudades del país, ya que enforcarse en otros mercados podría representar esfuerzos comerciales innecesarios ya que su valor futuro no llevaría a la empresa a encontrar un mejor nivel de utilidad, puesto que el 50% de esta se concentra en tan solo 3 departamentos.

Sería ideal para la compañía invertir estos esfuerzos en continuar desarrollando esos mercados, ya que, al tener presencia en los mismos, podría simplemente aumentar su capacidad comercial, logrando alcanzar cada vez más clientes y generar más utilidad a partir de estas zonas.

De igual forma se recomienda invertir en esfuerzos comerciales en departamentos como Atlántico, Tolima, Bolívar, Magdalena, Santander y Huila, que representan el 30% de su valor futuro, ya que, ante una coyuntura en los 3 departamentos previamente mencionados, el negocio de la compañía podría verse seriamente comprometido y con problemas para asegurar su continuidad en el tiempo.

Finalmente, según los clústeres encontrados, Avista podría diseñar esfuerzos comerciales segmentados, ya que dentro de los resultados del trabajo (Anexo 1) se puede conocer la distribución de cada clúster a nivel departamental, permitiéndole a los equipos comerciales, optimizar la forma en la que abordan sus clientes, asegurando un mayor número de desembolsos mensuales buscando impactar personas que aún no han sido incluidas en el mundo de los créditos de libranza.

5. Anexos

- GitHub para el código realizado para el desarrollo del modelo: https://raw.githubusercontent.com/juanjosegonzalez94/trabajodegrado/main/modelo

6. Referencias

Akay, Ö., & Yüksel, G. (2018). Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms. *Communications in Statistics: Simulation and Computation*, *47*(10), 3031–3041. https://doi.org/10.1080/03610918.2017.1367806

Asobancaria. (n.d.). *Componentes del crédito*.

Banco Mundial. (2018). *Inclusión financiera*. https://www.bancomundial.org/es/topic/financialinclusion/overview

Bernal Salazar, A. A. (2019). *El precariado en el desarrollo de emprendimientos digitales y el fenómeno de las startups.* http://expeditiorepositorio.utadeo.edu.co/handle/20.500.12010/7894

Blaufuks, W. (2021, May 25). *FAMD: How to generalize PCA to categorical and numerical data*. Towards Data Science.

Chang, S., Cohen, T., & Ostdiek, B. (2018). What is the machine learning? *Physical Review D*, *97*(5). https://doi.org/10.1103/PhysRevD.97.056009

Colpensiones. (2020). *Informe de gestión 2020*.

Dawood, E. A. E., Elfakhrany, E., & Maghraby, F. A. (2019). Improve Profiling Bank Customer's Behavior Using Machine Learning. *IEEE Access*, *7*, 109320–109327. https://doi.org/10.1109/access.2019.2934644

Departamento Administrativo Nacional de Estadística (DANE). (2018). Proyecciones de población por genero y edad 2018 - 2070. In *DANE*. DANE.

Djurisic, V., Kascelan, L., Rogic, S., & Melovic, B. (2020). Bank CRM Optimization Using Predictive Classification Based on the Support Vector Machine Method. *Applied Artificial Intelligence*, *34*(12), 1–15. https://doi.org/10.1080/08839514.2020.1790248

Goldstein, I., Jiang, W., & Karolyi, G. A. (2019). To FinTech and beyond. *Review of Financial Studies*, *32*(5), 1647–1661. https://doi.org/10.1093/rfs/hhz025

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4). https://doi.org/10.2307/2528823

Hassan, M. M. (2018). Customer Profiling and Segmentation in Retail Banks Using Data Mining Techniques. *International Journal of Advanced Research in Computer Science*, *9*(4), 24–29. https://doi.org/10.26483/ijarcs.v9i4.6172

Ionut, P. A. (2018). *EVOLUTION OF CUSTOMERS' SEGMENTATION TECHNIQUES IN RETAIL BANKING*. *I*, 194–199.

Jamal, A., Handayani, A., Septiandri, A. A., Ripmiatin, E., & Effendi, Y. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 192. https://doi.org/10.24843/lkjiti.2018.v09.i03.p08

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Springer Link*. https://doi.org/10.1007/s12525-021-00475-2/Published

Kahreh, M. S., Tive, M., Babania, A., & Hesan, M. (2014). Analyzing the Applications of Customer Lifetime Value (CLV) based on Benefit Segmentation for the Banking Sector. *Procedia - Social and Behavioral Sciences*, *109*(October 2015), 590–594. https://doi.org/10.1016/j.sbspro.2013.12.511

Marisa, F., Ahmad, S. S. S., Yusof, Z. I. M., Fachrudin, & Aziz, T. M. A. (2019). Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model. *International Journal of Integrated Engineering*, *11*(3), 169–180. https://doi.org/10.30880/ijie.2019.11.03.018

Mention, A. L. (2019). The Future of Fintech. *Research Technology Management*, *62*(4), 59–63. https://doi.org/10.1080/08956308.2019.1613123

Niloy, N. (2018). Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients. *American Journal of Data Mining and Knowledge Discovery*, *3*. https://doi.org/10.11648/j.ajdmkd.20180301.11

Palaniappan, S., Mustapha, A., Foozy, C. F. M., & Atan, R. (2017). Customer profiling using classification approach for bank telemarketing. *International Journal on Informatics Visualization*, *1*(4–2), 214–217. https://doi.org/10.30630/joiv.1.4-2.68

Qismat, T., & Feng, Y. (2020). *Comparison of classical RFM models and Machine learning models in CLV prediction*. 0–52.

Schroeder, B. (2019). What Is The Most Important Element Of A Successful Startup? Hint, It's Not The Idea, Team, Business Model Or Funding Dollars. *Forbes*. https://www.forbes.com/sites/bernhardschroeder/2019/09/23/what-is-the-most-important-element-of-a-successful-startup-hint-its-not-the-idea-team-business-model-or-funding-dollars/?sh=761c6d8c727c

Scikit-Learn. (n.d.). *sklearn.metrics.silhouette_score*. Https://Scikit-Learn.Org/Stable/Modules/Generated/Sklearn.Metrics.Silhouette_score.Html?Highlight=silhouette#sklearn.Metrics.Silhouette_score.

Sehgal, G., & Garg, K. (2014). *Comparison of Various Clustering Algorithms*. www.cs.waikato.ac.

Seif, G. (2018). *The 5 Clustering Algorithms Data Scientists Need to Know*. KDnuggets.

Shahapure, K. R., & Nicholas, C. (2020). *Cluster Quality Analysis Using Silhouette Score*. https://www.

Sharahi, M., & Aligholi, M. (2015). *Classify the Data of Bank Customers Using Data Mining and.pdf*. *5*(5), 458–464.

Sharma, A., Trần, A. A., & Garavaglia, S. (n.d.). *A SMART GUIDE TO DUMMY VARIABLES: FOUR APPLICATIONS AND A MACRO*.

Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*, *14*(5), 923–939. https://doi.org/10.3846/16111699.2012.749807

Startupeable. (2021). *Nubank: Historia y Futuro del Banco Digital Más Grande del Mundo*. Startupeable.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Villa Patiño, E. A. (2019). La libranza: una mirada actual, más allá de la ley 1527 del 2012. *Ces Derecho*, *10*(2), 535–565. https://doi.org/10.21615/cesder.10.2.1

Wang, Y., & Ni, X. S. (n.d.). *A XGBOOST RISK MODEL VIA FEATURE SELECTION BAYESIAN HYPER-PARAMETER OPTIMIZATION*.

Yang, S., & Zhang, H. (2018). Comparison of Several Data Mining Methods in Credit Card Default Prediction. *Intelligent Information Management*, *10*(05), 115–122. https://doi.org/10.4236/iim.2018.105010

Yoseph, F., & AlMalaily, M. (2019). New Market Segmentation Methods Using Enhanced (Rfm), Clv, Modified Regression and Clustering Methods. *International Journal of Computer Science and Information Technology*, *11*(01), 43–60. https://doi.org/10.5121/ijcsit.2019.11104

Zuleta Acevedo, J. C. (2019). El Índice Global de Fracaso. *La República*. https://www.larepublica.co/analisis/juan-carlos-zuleta-acevedo-532896/el-indice-global-de-fracaso-2783653