Modelo estadístico para generar un scoring que permita otorgar el beneficio de tasa de interés a los asociados con vigencia crediticia de una cooperativa de ahorro y crédito.

Matilde Pulido Jaimes, Angela Katerine Rangel Leal

Trabajo de Grado para Optar el título de

ESPECIALIZACIÓN EN ESTADISTICA

Directora

Dra. Deicy Villalba Rey

Ph. D. en Estadística

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

Año actual

2020

AGRADECIMIENTOS

- *A Dios por acompañarnos, guiarnos, darnos fortaleza y salud para sobrellevar cada situación adversa presentada durante la especialización, siempre tuvimos la confianza que todo culminaría con éxito.*

- *A mi esposo, mi madre e hija por su amor, ánimo, consejos y colaboración.*

- *A mis padres por su amor incondicional, su apoyo, consejos, y los valores que me han inculcado, por su excelente ejemplo de vida a seguir.*

- *A nuestra Directora de tesis, Dra. Deicy Villalba Rey por sus enseñanzas, dedicación, paciencia, quien con su conocimiento y experiencia fueron pautas fundamentales en la consecución del proyecto.*

- *A Financiera Comultrasan, directivos y Jefes Inmediatos por contribuir en el crecimiento de nuestra formación académica.*

- *Son muchas las personas que han formado parte de esta experiencia a las que les agradecemos su amistad, consejos, apoyo, ánimo y compañía en los momentos más difíciles. Gracias por todo lo que nos han brindado y por sus bendiciones, que Dios los bendiga.*

Angela Katerine Rangel Leal

Matilde Pulido Jaimes

TABLA DE CONTENIDO

Pág.

LISTA DE GRÁFICOS

Pág.

LISTA DE TABLAS

Pág.

LISTA DE APENDICE

Pág.

Resumen

Título: Modelo Estadístico para generar un Scoring, que permita otorgar el beneficio de tasa de interés a los asociados con vigencia crediticia de una cooperativa de ahorro y crédito [*]

Autor: MATILDE PULIDO JAIMES, ANGELA KATERINE RANGEL LEAL [**]

Palabras Clave: Análisis de conglomerados, Métodos multivariados, Beneficio en la tasa de interés, Análisis Discriminante, Modelo Logit Multinomial.

DESCRPCIÓN: Con el objetivo de tener mayor conocimiento sobre el asociado que al momento del estudio del modelo presenta crédito(s) activo(s) con la Cooperativa de Ahorro y Crédito, que permita otorgarle un beneficio en la disminución de la tasa de interés, buscando fidelizarlos, así mismo le brindamos el respaldo en sus cuotas de crédito siendo más flexibles en sus proyecciones de pago. Por otro lado, se logra disminuir los prepagos o cancelaciones anticipadas porque otras entidades financieras del mercado les ofrezcan tasas más atractivas.

Este tema es más sensible a medida que pasa el tiempo, teniendo en cuenta que los asociados buscan un beneficio adicional que se les pueda brindar frente a la competencia y es de precisar que el mercado sigue evolucionando a gran velocidad con nuevos productos y con tasas de créditos atractivas acompañado de la flexibilidad en las condiciones de otorgamiento del mismo, de igual forma el mercado sigue abriendo más posibilidades a los grandes bancos del país que busca ofrecer beneficios a sus clientes y ser más competitivos Es importante el analizar muy bien a los asociados y establecer diferenciales que nos permitan anticiparnos a sus necesidades.

La conformación de los cluster se fundamenta en grupos de asociados que presenten características similares entre ellos, pero que sean diferentes entre cada cluster. Estas características permitirán a la Cooperativa establecer estrategias encaminadas al beneficio del asociado y permitir en si una mayor fidelización. Así las cosas, se corrieron las siguientes técnicas estadísticas; análisis de conglomerados a través de la Técnica no jerárquico – CLARA y un modelo Multinomial.

A través de este trabajo se busca generar un perfilamiento de los asociados de acuerdo a sus características que permita a la Cooperativa implementar estrategias de fidelización otorgándole un beneficio en la tasa de interés de crédito, mediante la construcción del modelo scoring.

---

[*] Trabajo de Grado

[**] Facultad de Ciencias. Escuela de Matemáticas. Directora: Deicy Villalba Rey. Ph D. en Estadística. Codirector: Henry Sebastián Rangel Quiñonez. Master de Ciencias de la Estadística.

Abstract

Title: Statistical Model to generate a score, which allows granting the interest rate benefit to those associated with credit validity of a credit union.[*]

Author: MATILDE PULIDO JAIMES, ANGELA KATERINE RANGEL LEAL [**]

Key Words: Cluster analysis, Multivariate methods, Interest rate benefit, Discriminant Analysis, Multinomial Logit Model.

Description: With the objective of having greater knowledge about the associate who, at the time of the model study, presents active credit(s) with the Credit Union, which allows to grant a benefit in reducing the interest rate, seeking to retain them, likewise, the support in the credit installments will be offer, being more flexible in the payment projections. On the other hand, it is possible to reduce prepayments or early cancellations because other financial entities in the market offer more attractive rates.

This subject is more sensitive as time goes by, taking into account that associates are looking for an additional benefit that can be offered to them facing the competition and it is necessary to specify that the market continues to evolve at great speed with new products and with attractive loan rates accompanied by flexibility in the conditions for granting it, in the same way the market continues to open up more possibilities for the larger banks of the country that seek to offer benefits to their clients and to be more competitive. It is important to analyze the associates very well and to establish differentials that allow us to anticipate their needs.

The formation of the clusters is based on groups of associates that present similar characteristics among themselves, but that are different between each cluster. These characteristics will allow the Cooperative to establish strategies aimed at the benefit of the member and allow greater loyalty. Thus, the following statistical techniques were run; cluster analysis through the Non-hierarchical Technique - CLARA and a Multinomial model.

Through this work, the aim is to generate a profile of the associates according to their characteristics that will allow the Cooperative to implement loyalty strategies, granting it a benefit in the loan interest rate, through the construction of the scoring model.

---

[*] Degree work

[**] Science Faculty. School of mathematics. Director. Deicy Villalba Rey Ph.D. in statistics. Cordinator: Henry Sebastián Rangel Quiñonez. Master of Science in statistics

Introducción

El sector financiero en su búsqueda por captar usuarios, encamina dicha actividad brindando diversos beneficios como la flexibilización en la compra de cartera, además de otras modalidades como lo es, ofreciendo tasas de interés atractivas y competitivas frente a las demás entidades financieras.

Ante esta situación la Cooperativa de Ahorro y Crédito sobre la cual se realizará el análisis estadístico, evalúa una a una las operaciones con el objetivo de replantear la tasa inicialmente otorgada ante las solicitudes aprobadas por los analistas de crédito, o ante los requerimientos individualizados de asociados que optan por cambiar de entidad y cancelen de manera anticipada el crédito vigente.

Por estas razones es importante conocer el perfil del asociado, cuál ha sido su comportamiento de pago, la tasa que ha manejado en otros créditos y cuál es el nivel de recursos de ahorro que mantiene en la entidad, ya que la presente investigación pretende identificar los perfiles potenciales de crédito con el objetivo de otorgarle un beneficio en la tasa de interés activa.

De esta manera se plantea construir un modelo de scoring, a través de un Logit multinomial, para analizar variables de los asociados y definir posibilidad de otorgar o no el beneficio, aplicar un beneficio básico y aplicar un beneficio Premium.

1. OBJETIVOS

1.1 Objetivo General

Aplicar un modelo de scoring que permita identificar asociados potenciales que presenten buen hábito de pago y tengan rotación de crédito en la Cooperativa, para asignarles un beneficio en la reducción de puntos porcentuales a la tasa de interés de crédito, mediante técnicas estadísticas multivariadas.

1.2 Objetivos Específicos

Realizar un análisis exploratorio y descriptivo de la información crediticia de los asociados para el periodo de estudio.

Aplicar la técnica de clúster para clasificar los asociados en tres grupos “Premium”, “Básicos” y “no aplica”, con base en las características asociadas al comportamiento crediticio del asociado.

Plantear un modelo de regresión logística multinomial con las principales variables de estudio, para determinar la predicción del beneficio de tasa de interés.

1.3 Justificación

Los mercados financieros buscan día a día estar más cerca de los usuarios y satisfacer sus necesidades financieras, ofreciendo productos atractivos y beneficios, como es la reducción de la tasa de interés de crédito que a través de la compra de cartera y otras líneas de crédito (libre consumo, libranzas, tarjetas de crédito, etc.), así mismo, unificar las obligaciones financieras, en caso de tener más de un crédito activo en el sector financiero, ampliando su plazo y mejorando el flujo de caja del deudor. Ante esta situación, la Cooperativa conformó un comité de tasas de crédito que le permita atender oportunamente las solicitudes de aquellos asociados que solicitan una

disminución en su tasa de interés del crédito vigente con la Cooperativa, o recoger cartera de otras entidades financieras, o aquellos asociados que buscan prepagar el crédito, es decir; pagar de forma anticipada la totalidad del saldo del crédito.

En estos casos se procede a evaluar la posibilidad de disminuir la tasa actual del crédito en unos puntos porcentuales, de acuerdo al perfil del asociado, comportamiento de pago, tasa de interés actual y novedades históricas, buen hábito de ahorro, esta es una forma de incentivar y retener a los asociados potenciales.

Considerando lo anterior y factores influyentes que demandan operatividad y tiempo en reevaluar las solicitudes de crédito, el presente trabajo tiene como finalidad, diseñar un modelo scoring que aportará a la cooperativa la automatización del proceso de beneficio a asignar en cada caso.

Será importante para el área de gerencia financiera para dar cumplimiento ante la superintendencia solidaria que mediante requerimiento solicita fijar los criterios técnicos y señalar los procedimientos en la asignación de la tasa de interés y la reducción de la misma; además se usará ampliamente por el área comercial, además se convierte en una herramienta útil al momento de aprobar el crédito ya que se indicarán los descuentos aplicados en la tasa de intereses. Con lo anterior, se justifica el objetivo de la entidad para fidelizar al asociado potencial a través del otorgamiento de tasa de interés preferencial previamente a la aceptación del mismo.

1.4 Antecedentes

La técnica de crédito scoring tiene una gran aplicación en el campo financiero como alternativa a las técnicas tradicionales de evaluación de crédito (poco eficientes) ante el gran volumen de solicitudes de crédito.

A continuación, se reportan algunas investigaciones relacionadas con la calificación del crédito, cobranza de crédito, simulación de nuevos productos o deserción de clientes.

Arango & Restrepo (2017) plantearon un modelo scoring idóneo en pronosticar la probabilidad de incumplimiento del cliente, basado en regresión logística con base en información de la entidad estudiada, que contiene 23 variables numéricas y categóricas de un conjunto de cliente en un mes determinado. Se realizó análisis sobre cuatro técnicas para modelos scoring como análisis discriminante, modelo probabilístico, logístico y redes neuronales artificiales, en donde el modelo seleccionado fue el de regresión logística ya que acertó en más del 99% de las veces para predecir un incumplimiento.

Delgado, Cardona y Gil (2017) realizaron el diseño de un modelo de puntuación para el manejo adecuado de la cartera en cierta entidad de cobranzas. Los investigadores aplicaron un modelo de regresión logística, utilizando 16.000 clientes morosos diversas entidades financieras y de servicios públicos. Como resultado, se derivó que, respecto a las características socioeconómicas, retraso en los pagos, los salarios como empleadores o independiente y el nivel de deudas, el 50% de la cartera comercial, de consumo y microcrédito se pueden recuperar. Los autores concluyen que este tipo de modelos pueden ser aprovechados por las entidades financieras para la ejecución de políticas de ventas en cuanto a plazos y cupos al momento del estudio de un crédito.

En la investigación realizada por Delgado (2016) se consideró una segmentación de los clientes de una cooperativa, para mejorar los procesos comerciales y de mercadeo buscando impactar los resultados de la empresa. Para ello se usó las herramientas de clústering, donde la selección de variables obedeció al conocimiento del mercado, la información que la cooperativa tiene de sus

clientes en las bases de datos y la experiencia en el manejo de las mismas. Los resultados al ejecutar el procedimiento de análisis de conglomerados en dos fases, fueron, tres clústeres "El deudor independiente", "el deudor asalariado" y "el inversionista de estrato", que se vio útil para el área comercial ya que son fuentes importantes en la redefinición de las estrategias de venta.

El trabajo de investigación realizado por Jerez (2016) explora las asociaciones existentes entre la venta de productos y las exigencias para acceder a los servicios en una entidad prestadora de ahorro y crédito. Los datos fueron tomados de una encuesta de servicio, la cual indagó con sus asociados diversos atributos de la cooperativa y se identificó las variables concluyentes en la compra del producto financiero. Y como resultado del análisis de correspondencia múltiple reflejaron la venta de productos de ahorros altamente relacionada con la venta de productos de crédito.

Tumbia, Martínez y Beltrán (2016) identificaron los segmentos más representativos de los clientes a los que se les desembolsó un crédito con destino libre inversión en una entidad financiera, mediante un análisis de correspondencia múltiple y un análisis por método jerárquico aglomerativo. El estudio concluyó que el número óptimo de clúster son tres y que los principales factores de agrupación fueron la tasa de interés, el monto desembolsado y el segmento dentro del cual el banco clasifica a sus clientes.

Martínez (2015) realizó un modelo de calificación para facilitar la normatividad de autorización de préstamos. Utilizó datos basados en requerimientos de tarjeta de crédito recepcionados por clientes extranjeros en cierta entidad bancaria, utilizando una muestra del 30 % del total de clientes a los que fue aprobada una tarjeta de crédito, 7.274 observaciones, para las variables independientes usaron el método stpewise que realiza una secuencia de estadísticos F para determinar la inclusión o no de las variables, el resultado fue enfocado en determinar los clientes

incumplidos y cumplidos, posteriormente aplicaron el cálculo de puntuaciones, a lo que Concluyeron que las variables que más aportan al score con tasa de mora, la nacionalidad y el pasivo.

Adicionalmente, en la tesis realizada por Ochoa, Galeano, Agudelo (2010), cuyo objetivo fue precisar perfiles de deudores sujetos a incumplimiento de su crédito y perfiles de deudores con buen pago. Se utilizó la técnica de análisis discriminante para el levantamiento de un modelo scoring en el otorgamiento de crédito, para cierta base de una cooperativa financiera que contiene el histórico de créditos de 24.786 personas a corte de noviembre de 2009, en el que se identificaron 30 variables relacionadas con el registro interno de cada cliente. Los autores recomiendan la actualización continua del historial crediticio de los clientes, ya que entre mayor y más actualizada sea la información con la que se ejecute, mejor será el pronóstico del modelo.

En cuanto a la aplicación de técnicas de árboles de clasificación, se destaca el trabajo de Granda y Niño (2016) que diseñaron un framework para mejorar la gestión de cobranza en el programa de microcrédito de una cooperativa financiera , la cual parte de una base de asociados con información histórica de variables sociodemográficas, financiera, otorgamiento y comportamiento de crédito, seguidamente calcularon el default o incumplimiento dado de un cliente bueno y cual es malo, luego realizaron el análisis de las variables de manera descriptiva y aplicaron los métodos estadísticos como árboles de clasificación, análisis discriminante y regresión logística. Los resultados permitieron una clasificación correcta superior al 90% en los tres modelos, además elaboraron la tabla de score distribución que permitió identificar los puntos de corte de probabilidad de default, es así como aportaron una reducción considerable en la operatividad y se aumentó la productividad en los cobros de cartera en riesgo, esta investigación ayudó en la

comprensión de la herramienta para facilitar la mejor decisión en el beneficio que se espera brindar con el trabajo.

1.5 Marco Teórico

El Banco de la República[1] siendo una autoridad monetaria, crediticia y cambiaria con base en la Ley 31 de 1992, según Rodríguez, D.A. (2017), es la que se encarga de estudiar las medidas monetarias, crediticias y bancarias para regular la circulación monetaria y el nivel de liquidez en el mercado financiero colombiano y regula el costo de los créditos. Para lograr esto, el artículo 18 de la misma ley señala que las instituciones financieras y los intermediarios en las operaciones de mercado abierto y cambiario están obligadas a suministrarle al Banco información de carácter general y particular de sus operaciones, y los datos que permitan estimar su situación financiera.

La tasa de interés de Colombia ha presentado reducciones significativas desde septiembre de 2016, del 7.75% al 4,25%. Esto significa que la economía atraviesa un ciclo recesivo, y por tanto, se requiere el ofrecimiento de créditos para incentivar la creación de empresas, lo que permitiría poner en marcha y dar dinamismo a la economía tras la crisis de petróleo de 2015 y la reforma tributaria de 2017. En la medida en que los Bancos Privados ajusten sus tasas de interés para el público y sus emprendedores, la Reserva Federal le presta dinero al Banco Central de acuerdo a lo que se denomina el "ciclo de deuda de largo plazo" dependiendo del estado de las finanzas internacionales.

[1] Banco de la República. Es un órgano de Estado de naturaleza única, con autonomía administrativa, patrimonial y técnica, que ejerce funciones del bance Central.

1.5.1 Tasa de Interés[2]. En cuanto a la política monetaria del país, es la Junta Directiva del Banco de la República quien establece la tasa de intervención que el banco cobra a los establecimientos financieros por los créditos que les realiza a través de las operaciones de mercado abierto (OMA). Esta tasa es el principal mecanismo de intervención de política monetaria que utiliza el Banco de la República para determinar la cantidad de dinero que circula en la economía.

La tasa de interés es una medida que se le aplica a los créditos, depósitos de ahorro, inversiones y colocación, entre otros, que permite generar o recibir un rendimiento. Las tasas de interés nominales de crédito o activas de los bancos sobrepasan en 10 puntos los riesgos y los plazos en relación a variadas expectativas. Los movimientos que se presentan sobre la tasa en Colombia no son tan fuertes, sin embargo, durante el tiempo ha sido importante hacerle seguimiento al promedio de la tasa de crédito sobre el mercado financiero y sus efectos, los cuales arrojan un indicador adecuado de sus movimientos y niveles.

Al haber más dinero en circulación la tasa de interés tiende a bajar y se incrementa el consumo, generando más activación en la economía colombiana, en caso contrario suben las tasas y se presenta mayor escasez, por efecto que los demandantes tienen menos apetito para consumir, solicitando menos recursos a través de préstamo a los establecimientos intermediarios financieros, mientas que los oferentes pretenden colocar más recursos en cuentas de ahorros, CDT, etc.

Existen dos tipos de tasas de interés: ***La tasa pasiva o de captación***, un costo que cancelan los intermediarios financieros a los usuarios de recursos por el dinero captado y ***la tasa activa o de***

[2] Tasa de interés: Es el precio del dinero, que representa un porcentaje del crédito o préstamo que se ha requerido y que el deudor deberá pagar a quien le presta “Es el precio por el uso del dinero”.

colocación, la que reciben los intermediarios financieros de los deudores por los préstamos otorgados. Esta última es mayor, generándose una tasa de intermediación que se utiliza para cubrir los costos de administración, y una rentabilidad sobre la operación.

1.5.2 Factores determinantes de la tasa de interés de crédito. Las tasas de interés activas las aplican las entidades financieras a los diferentes tipos de crédito, como de consumo, de vivienda, microcrédito y comerciales, así como a las tarjetas de crédito, sobregiros y créditos especiales, de acuerdo a lo comentado por Carranza, I y González, E. (2019).

Es el Banco de la República quien calcula y publica (semanal y mensualmente) las tasas de interés y los montos de las diversas modalidades de crédito de las entidades financieras de crédito vigiladas por la Superintendencia Financiera de Colombia, éstas reportan las nuevas operaciones en moneda legal realizada durante la semana, y consolidadas con periodicidad mensual.

La metodología utilizada para el cálculo se fundamenta en tasas promedio ponderadas por los desembolsos y sumatorias de los valores que permiten agregar la información por categorías de entidades, plazos, créditos, crédito y semana, la cual se representa así;

$$\frac{\sum_{i=1}^{n} t_i d_i}{\sum_{i=1}^{n} d_i}$$

Donde n = es el número de entidades financieras de crédito que reportan, t_i es la tasa promedio de la operación de crédito de la entidad i, d_i es el desarrollo del crédito a la tasa t_i de la entidad i

1.5.3 Margen de intermediación. En el mercado financiero se presenta la tasa fija y la variable, la cual se mantendrá durante la vigencia del préstamo, en el caso de la variable se modificará en cada periodo de tiempo tomando como referencia la DTF, IBR, entre otras.

1.5.4 Tipos de tasa de interés de crédito. Las tasas de interés pueden ser fijas y mantenerse estables mientras dura la inversión o se devuelve el préstamo, o variables y actualizarse cada período de tiempo con base en algún indicador de referencia, por ejemplo, la DTF.

Cuando la tasa es variable, se le adiciona un spread o puntos adicionales a la tasa de referencia para obtener la tasa de interés, de allí que muchas veces se indique que la entidad financiera presta a la DTF + unos puntos de spread.

1.5.5 Comité de Tasas (interno)[3]. Con el objetivo de evitar que la Cooperativa otorgue créditos a asociados que presentan alto riesgo de incumplimiento, se evalúan diversas variables a través del scoring de créditos, el cual le asigna una calificación indicando si éste es aceptable o no para tomar crédito.

Estos asociados que inicialmente toman crédito, posteriormente requieren que la Cooperativa les ajuste la tasa de interés, disminuyéndola en unos puntos porcentuales, mejorar algunas de las condiciones del crédito o prepagar la obligación en su totalidad porque otra entidad financiera les presenta una mejor oferta o disponen del capital total para pagar el saldo del crédito. En estos casos; si el asociado lo requiere, el comité de tasas de crédito evalúa nuevamente el perfil del asociado, revisando su comportamiento de pago, el historial de operaciones anteriores, nivel de

[3] Comité de Tasas (Interno), Está conformado por la Gerencia Financiera, Vicepresidencia Comercial, Vicepresidencia de Riesgo, Crédito y Cartera.

ingresos, actividad económica, entre otras, para determinar si le otorga el beneficio de reducirle la tasa de interés.

1.5.6 Límites sobre las tasas de interés. Las Cooperativas de ahorro y crédito y los establecimientos bancarios, entre otros, deben aplicar en sus productos de captación y colocación, las tasas certificadas por la Superintendencia Financiera de Colombia, quien emite la tasa máxima de interés que pueden llegar a cobrar, de acuerdo a sus atribuciones legales según el Decreto 3819 de 2008

Una vez identificado que la tasa de interés cobrada este por encima de los límites establecidos, se incurre en el delito de usura, según lo señalado el artículo 305 del Código Penal, en este caso se debe presentar la denuncia ante la Fiscalía General de la Nación, quien se encargará de investigar dichas conductas, calificar los procesos y reportar a los tribunales y jueces competentes, los presuntos infractores de la ley penal. En caso de presentarse el cobro de una tasa de usura, ésta debe ajustarse inmediatamente para el periodo correspondiente, y la entidad financiera debe cubrir los dineros de dicho excedente, según lo reportado en el Boletín Jurídico (2009).

1.5.7 Clasificación de los créditos. Los créditos son clasificados de acuerdo a su destino económico.

1.5.7.1 Crédito de consumo. Las operaciones de crédito de consumo están dirigidos a personas naturales para financiar la adquisición de servicios o bienes -compras del hogar, vehículos, viajes, muebles, entre otros.

1.5.7.2 Crédito Comercial. Es una modalidad de crédito que se busca financiar en el corto plazo, a las compañías comerciales, que requieren de capital de trabajo para la adquisición de bienes o pago de servicios orientados a la operación de la misma compañía. Así mismo se puede llegar a utilizar para refinanciar pasivos con otras entidades financieras y proveedores de corto plazo.

1.5.7.3 Microcrédito. Son los créditos ofrecidos a microempresas, ya sean del sector agropecuario, industrial, comercial, manufacturero, o de servicios.

1.5.7.4 Vivienda. Son créditos dirigidos a personas naturales para la adquisición de vivienda, la cual puede ser nueva o usada, o para su construcción.

2. CAPÍTULO METODOLOGÍA

Este capítulo se concentra en la interpretación de los datos y la descripción de las variables. Además, se presenta la necesidad de evaluar la calidad de la información previamente al análisis estadístico, con la exposición de los criterios para la depuración (filtros) de la base de datos.

Igualmente se explican las técnicas estadísticas utilizadas para dar respuesta a los objetivos propuestos.

2.1 Descripción de Variables.

Basados en la información suministrada por la Cooperativa de Ahorro y Crédito, que consta de los créditos vigentes[4] (activos) a cierre de diciembre de 2019, la cual se depuró, adicionalmente se toma información socio demográfica proveniente de la base única de clientes de la cooperativa, evaluando los pesos asignados a cada una de las 19 variables a estudiar.

2.1.1 Variables Cuantitativas:

- **Edad:** Número de años de cada asociado
- **Antigüedad:** es la cantidad de años que tiene los asociados a la fecha de corte vinculado con la cooperativa
- **Activos:** corresponde a los activos líquidos que reporta el asociado en la base única de clientes.
- **Pasivos:** corresponde a los pasivos líquidos que reporta el asociado en la base única de clientes.
- **Ingresos:** es el rubro que el asociado recibe de manera mensual según su oficio
- **Saldo Depósitos:** corresponde a la suma del valor acumulado en cdat, ahorros y aportes, que el asociado presenta a la fecha de corte (diciembre 2019).
- **Apalancamiento:** Porcentaje que mide la capacidad de endeudamiento del asociado (pasivos / activos)

- **Prepagos:** Corresponde al número de créditos pagados por anticipado ante el plan de pagos establecido.
- **Créditos:** es el número de créditos que el asociado ha tenido en la cooperativa durante los últimos cinco años
- **Valor Desembolso:** es el valor del crédito desembolsado
- **Valor Promedio desembolso:** es el valor promedio de los créditos en los últimos cinco años
- **Tasa Efectiva Anual:** corresponde a la tasa aplicada al crédito desembolsado
- **Días de Mora cierre:** es el número de días en estado moroso que presenta el asociado en el crédito desembolsado.
- **Plazo:** corresponde al periodo acordado para pagar las operaciones de crédito, hace referencia al número de cuotas fijadas por el asociado para pagar la totalidad del crédito vigente.
- **Cuotas pagadas:** corresponde al número de cuotas canceladas según la frecuencia de pago pactada a la fecha de corte diciembre 2019.
- **Cuotas pendientes:** corresponde al número de cuotas por pagar, es la diferencia entre el plazo y las cuotas pagadas, a la fecha de corte diciembre 2019.

2.1.2 Variables Cualitativas

- **Género:** Característica general común que se divide en tres categorías femenino, masculino y no definido, este último hace referencia a los asociados identificados como persona jurídica.
- **Estrato:** ubicación del domicilio de los asociados dentro de la sociedad, esta variable se divide en 7 categorías: 0 no definido, 1 bajo- bajo, 2 bajo, 3 medio-bajo, 4 medio, 5 medio

-alto, 6 Alto, 7 alto personas jurídicas. Departamento Nacional de Planeación[5]. (Ley 142 de 1994)

- **Calidad de la Cartera:** corresponde al tipo de calificación de la cartera de acuerdo a los días de mora, la cual se clasifica en 4 categorías A son los créditos que presentan riesgo normal de mora hasta 30 días, B es la calificación de riesgo aceptable hasta 60 días de mora, C es la calificación de riesgo regular por mora hasta 90 días, D es la calificación del riesgo sobresaliente para mora que tiene un límite de hasta 120 días y la calificación E significa el riesgo más alto dentro de la cartera, que es incobrable para mora mayores a 120 días.

2.2 Depuración de la base de Datos

La cantidad de créditos vigentes a corte de diciembre de 2019 es de 160.760, según identificación previa se decide realizar una serie de filtros, buscando congruencia en los datos, a continuación, se describe algunos criterios que serán filtrados:

Los créditos bajo el destino económico "cupo activo", son créditos que por su plan de pagos presenta alta volatilidad, porque su saldo se activa de manera rotativa, de esta manera se eliminaron 52.620 registros.

Los créditos que al momento de establecer las condiciones de pago presentan beneficio en la misma, estos son los desembolsados bajo el código de producto "Finagro", "Vivienda empleados", "Vivienda empleados segunda vez", "Crediaportes al 80%", "Créditos al 200%", "Reciprocidad en Cdats", "Programado largo plazo" y "Crediprima" que se eliminan 9.744 registros.

[5] Dane: es la entidad responsable de la planeación, levantamiento, procesamiento, análisis y difusión de las estadísticas oficiales de Colombia..

Los créditos reestructurados es el siguiente criterio de eliminación, con 4.049, hace referencia a aquellos créditos que han sufrido cambios en las condiciones de otorgamiento o en la ampliación del plazo y monto.

No se contemplan para el estudio los créditos con saldo en cero pesos a la fecha de corte de evaluación, teniendo en cuenta que permanecen en la base y no traen información del crédito como es el saldo, la tasa de interés y el plazo, entre otros. Sobre esta información se tomaron aleatoriamente unos asociados y se pudo identificar que fueron créditos que pagaban la última cuota del crédito o prepagaron el saldo pendiente a la fecha de evaluación. Así mismo, se analizaron diferentes puntos de vista para la eliminación de los 439 registros incluyendo los prepagos, teniendo en cuenta que el asociado puede tener capacidad de pago, pero no permite cumplir el ciclo del crédito, generando una reducción en el retorno total de intereses esperados, lo ideal es ofrecer el beneficio a asociados que busquen mantener el flujo de caja de la operación en un buen comportamiento de pago.

Los créditos duplicados -8.490 registros- corresponden a créditos a nombre del mismo asociado, ya sea por el mismo destino o por otro. En este sentido permanecen los créditos que contienen la cuantía mayor y conserva sus datos iniciales, respecto a las variables de saldo depósitos, edad, género, estrato, activos, pasivos, ingresos.

Se eliminaron 1.545 registros que reportan ingresos mensuales menores a un salario mínimo legal vigente, los cuales corresponden en su mayoría a créditos Presta U, los cuales en una gran proporción son a un plazo de 6 meses y son cuantías pequeñas.

Respecto a los registros en activos y pasivos se identifican valores atípicos que no corresponde a la realidad, se asumen errores operativos en digitación por parte de los asesores que al momento

de diligenciar estos datos los realizan basados en cumplir un requisito y no se por muestra el comportamiento real de los bienes, se depuran 16.903 activos y 17.936 en pasivos.

Basados en el indicador de apalancamiento (pasivos / activos) se evidencia porcentajes atípicos, superiores al 100%, a lo cual se filtran 2.071 registros.

Se evidencia que 5.919 registros corresponden a montos desembolsados por la línea 1 (libranza), que ya presenta un beneficio en tasa, además el respaldo de la pagaduría incentiva a que estos registros se otorguen son tal beneficio.

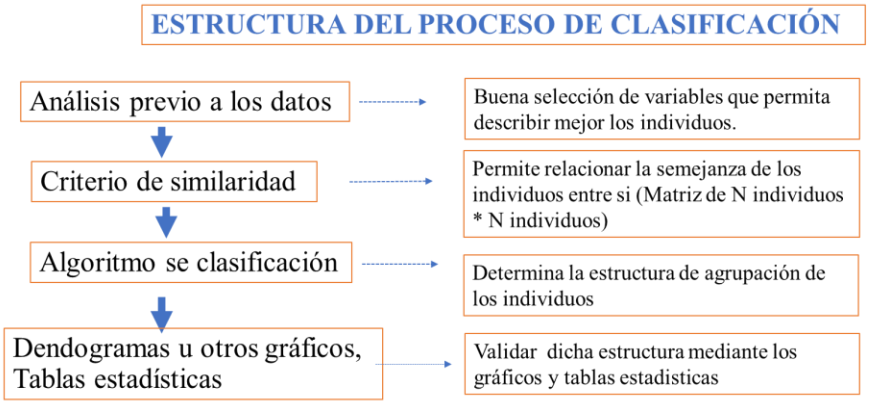
Respecto a los créditos de Presta U, se filtran 1.371 registros, toda vez que comprenden créditos con ciertas campañas en becas y auxilio económico en la apertura de aportes por parte del fondo de solidaridad de la cooperativa en estudio.

Se filtran 1.089 registros que según el plazo pactado para terminar su crédito es de seis (6) meses, ya que el retorno de los intereses es muy corto plazo. Posterior a depurar los anteriores criterios, la base de datos para realizar el estudio es de **33.428 registros**.

2.2.1 Técnicas Estadísticas Usadas para el Modelo de scoring. Desde los años 60 las entidades financieras han utilizado scoring de crédito en sus actividades como principal herramienta para evaluar el riesgo que representa un cliente cuando requiere un crédito, estas técnicas permiten analizar su historial crediticio y predecir el comportamiento del cliente en función de una característica o varias características observadas en el tiempo de vinculación con la entidad. Los modelos de scoring son claramente métodos mucho más sofisticados pues brindan mayor información acerca de las relaciones entre las variables, proporcionando la viabilidad del crédito, que monto de crédito se otorgará y bajo qué condiciones, por lo tanto, es estos modelos se cometan menos errores, al ser menos drásticos al momento de clasificar los individuos.

El presente trabajo se desarrolla con el fin de hallar un modelo que permita otorgar a los asociados preferenciales un beneficio en la tasa de interés de la cooperativa, con el objetivo de mejorar el proceso de asignación de tasa para el mejoramiento del proceso mediante tres herramientas estadísticas. A continuación, se exponen los elementos teóricos referentes a los métodos estadísticos que se usaran en el contenido del trabajo

2.2.2 Análisis de Clúster. Esta es una técnica que busca agrupar los individuos que presenten mayor homogeneidad (similar o próxima entre sí) y que mediante la medida de similitud permite ir clasificando a los individuos en unos u otros grupos, encontrando relación entre las variables.



Etapas para el análisis de clúster.

1. ***Selección de las variables***

Para la selección de variables se considera tener en cuenta, que estas sean relevantes y permitan la distribución ideal de los grupos de acuerdo al propósito del estudio. Estas variables deben ser de tipo continuo (aunque existen medidas de similaridad y disimilaridad para datos de tipo categórico, ya sean de tipo binario o de tipo multinomial y para datos mixtos).

Generalmente, al previo análisis de clúster se realiza un análisis multivariado ya sea de componentes principales (variables continuas) o de correspondencias múltiples (variables categóricas), con el objetivo de reducir dimensionalidad y aprovechar las altas correlaciones o asociaciones entre las variables. De no usarse el análisis de componentes principales, se recomienda el estandarizar las variables, para hacer comparable su métrica.

Elección de la medida de proximidad; Se debe tener claro el criterio de proximidad o de distancia, ya que es importante en la agrupación de los individuos. Una medida de distancia (o de similaridad) es una expresión en términos de distancia entre individuos. La definición de la métrica de similitud o distancia será distinta en función del tipo de dato y de la interpretación semántica que el investigador realice. Algunas de las medidas de similaridad o de distancia más comunes se explican a continuación:

Coeficiente de Correlación de Pearson; Es un índice utilizado para medir el grado de relación de dos variables cuantitativas normales. Su valor oscila entre -1 y 1. Un valor de correlación cercano a 0 es un indicador de que no hay relación lineal, un valor mayor a cero que se acerque a 1 da una mayor relación directa entre los datos, y menor a 0 (cercano a -1) da muestra de una relación inversa entre los datos.

$$r = \frac{Cov(x,y)}{\sqrt{var(x)var(y)}}$$

Coeficiente de Congruencia; Basados en el producto escalar de dos vectores, o la suma de los productos cruzados. El producto escalar puede interpretarse como el producto de la longitud del vector Xj por la longitud de la proyección de Xi sobre Xj. El coseno del ángulo es una medida de similaridad entre Xi y Xj, con valores entre -1 y 1 en virtud de la desigualdad de Schwarz.

$$C = \frac{\sum_{j=1}^{p} x_r j \, y_{s_j}}{\sqrt{\sum_{j=1}^{p} x_{rj}^2} \cdot \sqrt{\sum_{j=1}^{p} y_{s_j}^2}}$$

*Distancia Euclídea cuadrática**;*** Sabiendo que X, vector de n variables continuas medidas sobre el individuo **x** y Y vector de n variables continuas medidas sobre el individuo **y,** se define la distancia euclídea cuadrática como

$$d^2(x,y) = \sum_{i=1}^{n} (X_i - Y_i)^2$$

La raíz cuadrada se conoce como **Distancia Euclídea.**

Matriz de distancias: Se establece que todos los métodos parten de la Matriz de Distancia, la cual es de orden n*n con las distancias definidas entre cada uno de los individuos, en donde a son los asociados.

$$\Delta a = \begin{bmatrix} \delta a_{11} & \delta a_{12} \dots \delta_{a1n} \\ \delta_{a21} & \delta_{a22} \dots \delta_{a2n} \\ & \dots \\ & \dots \\ & \dots \\ \delta a_{n1} & \delta_{an2} \dots \dots \delta a_{nn} \end{bmatrix}$$

2. Elección del método de Clasificación

A partir del mismo conjunto de datos examinados, cada método puede crear diversas soluciones y emplear criterios diferentes en la agrupación de los individuos, por esto, es importante averiguar las particularidades inherentes de los diferentes métodos para seleccionar el más conveniente de acuerdo al enfoque y planteamiento del problema a resolver. Estos métodos de clasificación en: Jerárquicos, No jerárquicos y Mixtos.

En este trabajo se utilizó un Método No jerárquico con K-mediods, en particular el método CLARA (Clustering Large Aplication) implementado por (Kaufman y Rousseeuw ,1990) el cuál es una extensión de los métodos de k-medoids (PAM) para manejar datos que contienen una gran

cantidad de objetos. En tal caso, el algoritmo selecciona una muestra de tamaño r de una población de tamaño n.

i. Para *i = 1* hasta *n*, repetir los siguientes pasos:

ii. Seleccionar una muestra aleatoria de *r* individuos del conjunto total de datos, y ejecutar al algoritmo PAM para encontrar los *k medoids* en la muestra.

iii. Para cada individuo Ij del conjunto total de datos, determinar cuál de los *k medoids* es el más próximo a *Ij*.

iv. Calcular la distancia media del grupo obtenido en el paso anterior. Si este valor es menor al mínimo actual, usar este valor como el mínimo actual y retener los *k medoids* encontrados en el paso (2) como el mejor conjunto de *medoids* obtenidos.

v. Retomar al paso (1) para comenzar la próxima iteración.

Y así selecciona los individuos representativos de cada grupo (k-medoids), los individuos más cercanos a ellos y establece cada grupo o clúster.

3. Número de grupos o clúster:

La cantidad de grupos se define por el conocimiento de los individuos, que es de manera predeterminada por el investigador o por otro lado teniendo en cuenta las pautas de las técnicas de clasificación (no supervisados o supervisados). Lo anterior, en ambos casos buscando determinar y confirmar cuantos grupos se formarán al ejecutar la clasificación final.

Dentro de los criterios más relevantes, según Everit (2011), está el criterio de razón de varianzas, el cual explica la partición en un número de grupos asociada a una proporción de varianza explicada mayor en relación a otra partición con menor número de grupos

4. Validación e interpretación de los resultados:

Es importante realizar validación a los resultados de un análisis clasificatorio, teniendo en cuenta el carácter exploratorio y la diversidad de soluciones posibles que se puedan presentar. Existen varios criterios para la validación de resultados del clúster. Al respecto, Fernández (1991) destaca el criterio del coeficiente de correlación cophenético de Goodman y Kruskal, denominado como un indicador de la posición de los distintos grupos a través del cual se pueden realizar comparaciones entre diferentes grupos.

Mediante el proceso de validación se busca afirmar que la decisión sobre el agrupamiento final sea relacionado, acertado y constante. De igual forma, existen técnicas descriptivas para la interpretación de los resultados, como son las tablas estadísticas que relacionan las características derivadas con las variables que han operado de criterios en la clasificación.

2.2.3 Modelos de Regresión para variable respuesta categórica. Son modelos estadísticos que permiten identificar la relación existente entre una variable dependiente cualitativa, ya sea dicotómica (regresión logística binaria) o politómica (regresión logística multinomial o regresión logística ordinal, si se define un orden) y una serie de variables independientes, las cuales pueden ser cuantitativas o cualitativas.

Modelo de Regresión Logística Multinomial. Se utiliza en modelos con variables dependientes nominales que tengan más de dos categorías – politómica, siendo una extensión multivariante de la regresión logística binaria

Para realizar la modelación se elige una categoría referente de la variable dependiente o respuesta y se ejecutan diversas ecuaciones simultáneamente, una para cada una frente a la referencia.

En el modelo de regresión logística multinomial, la variable dependiente presenta más de dos categorías, y al considerarse una variable respuesta politómica (Y) teniendo más de dos categorías en su resultado se denota por $Y_1, Y_2, \ldots. \; Y_K$.

De cada categoría de respuesta se busca explicar la probabilidad en función de un conjunto de covariables $X = \{Y_1, Y_2, \ldots. Y_n\}$ observadas. Es decir, ajustar un modelo de la forma;

$$p_j(x) = P\left[Y = Y_J \,./X = x\right] = f_j(x) \quad \forall \quad j = 1, \ldots, k$$

Para cada vector x de valores observados de las variables explicativas X.

Cuando la variable respuesta es politómica, la distribución de Bernoulli se convierte en una distribución multinomial de parámetros las probabilidades de cada una de las categorías de respuesta. Así; (Y/ X = x) M (1; p 1(x), …, Pk (x)), siendo $\sum_{j=1}^{k} pj\ (x) = 1$.

Para un modelo lineal, se obtiene $\binom{k}{2}$ transformaciones logit que permiten comparar las variables respuestas a través de cada par de categorías, así;

$$\ln \left[\frac{\frac{p_i\ (x)}{p_i\ (x) + p_j\ (x)}}{\frac{p_j\ (x)}{p_i\ (x) + p_j\ (x)}}\right] = \ln \left[\frac{p_i\ (x)}{p_i\ (x)}\right], \quad \forall \quad i, j = 1, \ldots k\ (i \neq j)$$

Esta ecuación representa el logaritmo de respuesta Y_i en comparación a Y_j restringido las observaciones de las variables independientes, las cuales pueden caer en uno de los dos niveles.

Para la formación del modelo logit multinomial respuesta se puede tener aplicar (k-1) diversas transformaciones logit básicas, establecidas por una categoría de referencia siendo la última Yk. Así las transformaciones se especifican como;

$\text{Lj (x)} = \ln\left[\frac{p_j\ (x)}{p_k\ (x)}\right] \forall\ j = 1,\ldots.\ K - 1$, siendo Lj (x) el logaritmo de la ventaja de respuesta el logaritmo de la ventaja de respuesta Yj dado que las observaciones de las variables independientes caen en la categoría Yj o en la Yk.

Si se tiene una variable respuesta Y con k categorías y p posibles variables explicativas asociadas con las modalidades de Y, tomando de referencia a Y=k se obtendrán las siguientes k-1 ecuaciones:

$$ln\left(\frac{Y=1|x}{Y=k|x}\right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p$$

$$ln\left(\frac{Y=2|x}{Y=k|x}\right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p$$

$$\vdots$$

$$ln\left(\frac{Y=k-1|x}{Y=k|x}\right) = \beta_{(k-1)0} + \beta_{(k-1)1}x_1 + \beta_{(k-1)2}x_2 + \cdots + \beta_{(k-1)p}x_p$$

Requisitos y etapas de la regresión logística:

- Evaluar la significancia individual de cada una de las variables explicativas
- Analizar resultados de confusión e interacción del modelo explicativo.
- Evaluar la bondad de ajuste de los modelos, para lo cual se puede hacer uso, entre otro del criterio AIC, el BIC, la matriz de confusión y los seudo R cuadráticos.

- Estudiar la fuerza y significancia de los coeficientes y los exponenciales para realizar la interpretación adecuada del modelo.

Método de estimación por máxima verosimilitud

Para dar un valor aproximado de los coeficientes del modelo y errores estándar se maneja la estimación por máxima verosimilitud, que indica, que las estimaciones arrojen alta la probabilidad de conseguir los valores de la variable resultante, la cual es proporcionada por los datos de la muestra. Para la regresión logística multinomial las estimaciones no se realizan por cálculos directos, se considera aplicar los métodos iterativos como el de Newton, Raphson (1690), que aluden la iteración con un valor razonablemente cercano al cero (denominado punto de arranque o valor supuesto).

Ecuación Verosimilitud:

$$L = \prod_{i=1}^{n} \left(p_{1i}^{Y_{11}} * p_{2i}^{Y_{21}} * p_{3i}^{1-Y_{11}-Y_{2I}}\right) = \prod_{i=1}^{n} \left(\frac{P_{1i}}{P_{3i}}\right).^{Y1i} l * \left(\frac{P_{1i}}{P_{3i}}\right).^{Y2i} * P_{3i}$$

Se puede utilizar la siguiente función auxiliar:

$$\Lambda = -2 * \ln(L) = -2 * \sum_{i=1}^{n} \left(Y_{1i} * \ln\left(\frac{P_{1i}}{P_{3i}}\right) + Y_{2i} * \ln\left(\frac{P_{2i}}{P_{3i}}\right) + \ln\left(P_{3i}\right)\right) =$$

$$= 2 * \sum_{i=1}^{n} ln\left(1 + \exp\left(Z_{1i}\right) + \exp\left(\exp\left(Z_{2i}\right)\right) - Y_{1i} * Y_{1i} - Y_{2i} * Y_{2i}\right)$$

En la máxima verosimilitud al restar la función auxiliar Λ_Λ y se puede resolverse por metodologías numéricos de forma iterativa basados en la estimación inicial $\beta_{11} = \beta_{21} = \beta_{12} =$

$\beta_{22} = 0$, $\beta_{01} = \ln(n_1) - \ln(n - n_1 - n_2)$ y $\beta_{02} = \ln(n_2) - \ln(n - n_1 - n_2)$ siendo n_1 y n_2 el número de sucesos en las categorías 1 y 2 respectivamente. Estos estadísticos son obtenidos admitiendo que no se refleja influencia de las variables predictoras en el modelo planteado, por lo tanto, el valor inicial de la función auxiliar se debe minimizar es la siguiente:

$$\Lambda_0 = -2 * \left(n_1 * \ln\left(\frac{\dot{n_1}}{n}\right) + n_2 * \ln\left(\frac{n_2}{n}\right) + (n - n_1 - n_2) * \ln(\frac{n - n_1 - n_2}{n}) \right)$$

Una vez lograda la correlación del método iterativo, designaremos por Λ_0 al mínimo conseguido y por, $\hat{\beta}_{01}, \hat{\beta}_{11}$, $\hat{\beta}_{21}$, $\hat{\beta}_{02}$, $\hat{\beta}_{12}$, $\hat{\beta}_{22,\ldots}$ a todos los valores estimados para los parámetros del modelo.

- ***Interpretación del modelo***

Supóngase tres grupos o categorías de Y (a, b y c) y 2 variables explicativas X1 y X2.

Definiendo: $z_1 = \beta_{01} + \beta_{11}X_1 + \beta_{21}X_2$

$z_2 = \beta_{02} + \beta_{21}X_1 + \beta_{22}X_2$

Se tiene:

$$P_1 = P(Y = a/X) = \frac{e^{z_1}}{1 + e^{z_1} + e^{z_2}}$$

$$P_2 = P(Y = b/X) = \frac{e^{z_2}}{1 + e^{z_1} + e^{z_2}}$$

$$P_3 = P(Y = c/X) = 1 - P_1 - P_2 = \frac{1}{1 + e^{z_1} + e^{z_2}}$$

Los odds que se forman son:

$$\frac{P_1}{P_3} = e^{z_1} = e^{\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2}$$

$$\frac{P_2}{P_3} = e^{z_2} = e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2}$$

Supóngase dos individuos A y B. Supóngase que A ha seleccionado x1 y x2 y B ha seleccionado (x1+1) y x2.

¿En qué proporción incrementa B la probabilidad de seleccionar la opción b respecto a la opción c si la comparamos con el individuo A?

$$odd_A = \frac{P_2}{P_3} = e^{z_2} = e^{\beta_{02}+\beta_{12}x_1+\beta_{22}x_2}$$

$$odd_B = \frac{P_1}{P_3} = e^{z_1} = e^{\beta_{01}+\beta_{11}(x_1+1)+\beta_{21}x_2}$$

$$OR(B/A) = \frac{Odd_B}{Odd_A} = e^{\beta_{11}}$$

Es decir, que exp(B) se puede interpretar como el cociente de odds (OR) del individuo B respecto al individuo A. Si se comparan dos individuos con una diferencia de k unidades, se tiene:

$$OR(B/A) = \frac{Odd_B}{Odd_A} = e^{k\beta_{11}}$$

- ***Calidad de ajuste (o bondad) del modelo***

En cuanto a la calidad de ajuste, esta es medida por medio de los coeficientes denominados como P seudo R^2, entre los que destacan los coeficientes de Mc – Fadden, de Cox-Snell y de Nagelkerke, entre otros.

- ***Coeficiente Pseudo (R^2) de Mc-Fadden***

Al tener $\Lambda = -2\ln(V)$, identificamos por Λ_o el valor inicial de esta función, es decir el mínimo Λ bajo el modelo ajustado con todos los parámetros, obtenemos la siguiente expresión del Pseudo-R^2 es;

$$R^2_{MF} = 1 - \frac{\Lambda_f}{\Lambda_o}$$

Siendo su rango teórico de valores $0 \leq R^2_{MF} \leq 1$, pero muy raramente su valor se aproxima a 1. Suele considerarse una buena calidad del ajuste cuadrado $0{,}2 \leq R^2_{MF} \leq 0{,}4$ y es bueno para valores superiores.

- ***Coeficiente Pseudo – R^2 de Cox – Snell***

Aquí se utiliza directamente la función de verosimilitud, V, y no la función auxiliar Λ. Por lo que se denota por $V_o = \exp(-\Lambda_o/2)$ el máximo de verosimilitud bajo el modelo nulo dado sólo por un término constante y por $V_f = \exp(-\Lambda_f/2)$ el máximo de verosimilitud bajo el modelo ajustado teniendo en cuenta todos los parámetros establecidos en el coeficiente Pseudo-R^2 de Cox-Snell como;

$$R^2_{CS} = 1 - \left(\frac{V_o}{V_F}\right)^{\frac{2}{N}} = 1 - \exp\left(\frac{\Lambda_f - \Lambda_O}{N}\right)$$

El rango para el coeficiente es $0 \leq R^2_{CS} \leq 1 - V_O^{\frac{2}{N}}$, lo que hace poco interpretable al depender de V_O. Toda vez que pueden ser próximos a cero cuando hay pocos datos. En este caso para bondad de ajuste se puede utilizar el coeficiente Pseudo- R^2 de Nagelkerke.

- ***Coeficiente Pseudo- R2 de Nagelkerke.***

$$R^2_N = \frac{R^2_{CS}}{1 - V_O^{\frac{2}{N}}} = \frac{1.\exp\ \frac{\Lambda_f - \Lambda_O}{N}}{1 - \exp\left(\frac{-\Lambda_O}{N}\right)}$$

Y en este caso, su rango de valores es $0 \leq R_{CS}^2 \leq 1$, se interpreta igual que la regresión lineal clásica, muy poco los valores se acercan a 1.

3. CAPÍTULO DE RESULTADOS

A continuación, se muestran los resultados del análisis exploratorio y descriptivo de los datos. De igual forma los resultados del análisis clúster y de la regresión logística multinomial -basada en la variable respuesta arrojada en el clúster, que contiene tres categorías nombradas "Premium", "Básico" y "No aplica".

Durante el tratamiento de los datos, se utilizaron los paquetes FactoMiner, MASS, rgl, ggplot2, grid, psych, factoextra, datasets, lme4, lattice, nnet y clúster del software R versión 3.6.1

3.1. Análisis Exploratorio

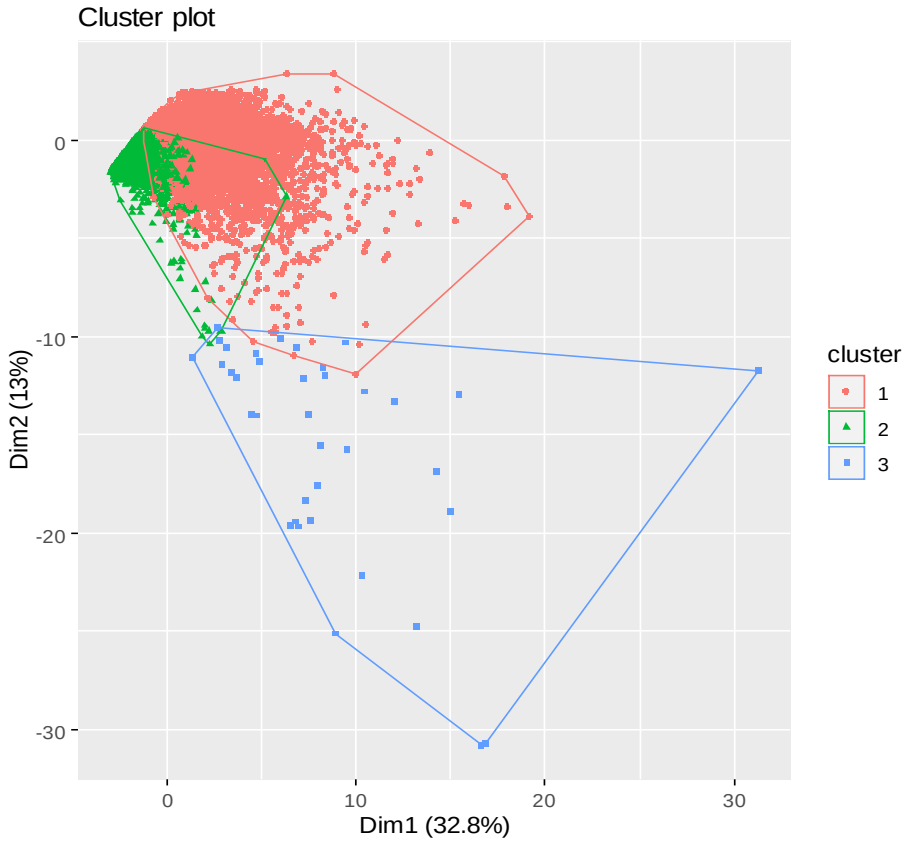
Para la exploración de los datos se usan principalmente gráficos y estadísticos que permiten examinar la distribución de los datos, identificando particularidades como los valores atípicos, concentraciones de valores y forma de distribución.

En primera medida el análisis se realizó de manera univariada, pero los resultados obtenidos no aportaban conclusiones significativas. Por lo tanto, se realiza un análisis multivariado aplicando la técnica análisis de Clúster. Cómo el objetivo principal de este trabajo es la asignación de beneficios (definidos en tres opciones: No beneficio, Beneficio Básico y Beneficio Premium) a los asociados de la cooperativa, se determinó la creación de tres grupos (o tres clúster) de asociados, los cuales según sus principales características diferenciadoras serán a posterior catalogados en cada uno de

Los grupos de Beneficio. Por ello al realizar exploración multivariada mediante la aplicación del análisis clúster (definiendo diferentes variables en su conformación) se evidenció que los asociados definidos como “personería jurídica” tenían un comportamiento muy diferente al de personas naturales.

La figura 1 muestra uno de los resultados de cluster creados con el algoritmo CLARA, por la dimensión de la base de datos-. En este clúster, por ejemplo, la totalidad de los individuos del grupo 3 (38 en total) pertenecen al género de Personería Jurídica. Al retirarlos y volver a agrupar, nuevamente apareció un reducido grupo de datos (todo género de Personería Jurídica) alejado de los demás individuos.

Figura 1.
Plot de Clúster exploratorio



Se realizó un análisis descriptivo del género Personería Jurídica y se encontró que estos asociados reflejan valores superiores en variables como Ingresos, Activos, Pasivos, Apalancamiento, promedio de créditos, entre otros, que los diferencian del resto de la población, por lo que podrían clasificarse automáticamente en el grupo óptimo para dar beneficio Premium. Por ello, se procede a excluir estos 121 datos del género de Personería Jurídica y se procede a realizar los análisis posteriores sólo con las personas naturales.

3.2 Análisis descriptivo

Con el fin de analizar los datos conceptualizados en el anterior capitulo, se realiza el análisis descriptivo para extraer conclusiones sobre el comportamiento de las variables utilizadas en el estudio.

3.2.1 Resumen de Variables estudiadas; La población en estudio presenta una edad promedio de 45,31 años (± 13.07años) -con un mínimo de 18 años y un máximo de 78 años, una antigüedad promedio en la cooperativa de 4,76 años (± 6,41años)–con mínimo de 0 y un máximo de 49 años, lo que se supone corresponden a asociados fidelizados.

Tabla 1

Resultados descriptivos de variables estudiadas

	Edad	Activos	Pasivos	Ingresos	Saldo Depósitos
Minimo	18,00	5.000.000	830.000	890.616	524
1er Quartil	35,00	35.000.000	4.374.500	1.800.000	177.566
Mediana	45,00	85.200.000	11.050.000	3.100.000	462.661
Promedio	45,31	133.533.744	25.482.204	8.127.655	1.242.828
3er Quartil	55,00	180.000.000	29.000.000	7.000.000	1.077.446
Maximo	78,00	800.000.000	612.213.658	726.000.000	370.168.144

	Apalancamiento	Créditos Prepagados	Número Créditos	Valor Desembolso	Saldo Promedio Créditos
Minimo	0.0010	0.0000	1	1.000.000	118.886
1er Quartil	0.0740	0.0000	1	5.500.000	4.451.878
Mediana	0.1700	0.0000	2	10.200.000	8.683.333
Promedio	0.2371	0.8119	2.802	16.150.846	12.700.437
3er Quartil	0.3360	1	3	20.000.000	15.346.854
Maximo	1	49	87	700.000.000	300.000.000

	Tasa_ EA	Dias de mora	Plazo	Cuotas Pagadas	Cuotas Pendientes
Minimo	0.1120	0.00	7	0.00	0
1er Quartil	0.1680	0.00	36	2	18
Mediana	0.1960	0.00	48	9	31
Promedio	0.2257	16.77	46.04	13.69	32.35
3er Quartil	0.2820	0.00	60	22	46
Maximo	0.4260	2.658	180	102	154

Años Antigüedad	
Minimo	0.000
1er Quartil	1
Mediana	3
Promedio	4.766
3er Quartil	6
Maximo	49

Calidad Cartera	Cantidad
A	31.443
B	403
C	204
D	296
E	961

Estrato	Cantidad
1	3.993
2	11.377
3	11.008
4	5.694
5	915
6	240
7	80

Genero	Cantidad
Femenimo	14.503
Masculino	18.804

Fuente: Medidas estadísticas de las variables analizadas que permiten resumir el valor de cada una.

Por otro lado, en variables que representan el comportamiento de los resultados financieros de los asociados y que al momento de una solicitud de crédito dan a conocer a la entidad financiera, los Activos que disponen los asociados tiene un valor mínimo de $5.000.000 y el máximo de $800.000.000, con un activo promedio de $133.533.744. El valor del pasivo mínimo es de $830.000 y el máximo de $612.213.658 y un pasivo promedio de $25.468.959. En cuanto a los ingresos mensuales del asociado, que son provenientes de las ventas de sus negocios, salarios, pensiones, arriendo entre otros, el promedio se evidencia $8.127.655, ocupando una fuerte concentración del 75% en montos inferiores a $7.000.000. Por último, apalancamiento es una variable calculada de los pasivos sobre los activos, es decir, las deudas que presentan frente al respaldo que tienen para cumplir con sus obligaciones, y se concentra en los índices del 7% y 34%, lo que indica que los asociados tienen un respaldo adecuado para cubrir con sus obligaciones.

Correspondiente a los recursos que tienen los asociados en la Cooperativa, a través de sus cuentas de ahorros, CDATs[6], y aportes, el saldo mínimo de ahorros son $524 y el máximo que tiene un asociado es de $370.168.144, con un promedio de $1.248.184. El 75% de los asociados presentan menos de un prepago en promedio durante los últimos 5 años, es decir; los asociados cancelan en su totalidad el saldo que presentan de sus obligaciones financieras antes de la fecha de vencimiento contractual pactada al desembolso del crédito, lo que no permite que el crédito cumpla con su flujo de caja inicialmente establecido entre la Cooperativa y el asociado. Es de precisar, que la mayoría de los asociados cumplen con el ciclo normal del flujo de caja del crédito.

[6] CDATs, Certificado de Depósito de Ahorro y Término Fijo

Durante los últimos cinco años el 75% de los asociados ha presentado 3 créditos, ya sean por diversos montos, plazos o líneas de crédito y de los asociados el máximo que ha tenido crédito durante este lapso de tiempo han sido 87 créditos.

El 75% de los asociados manejan un valor de desembolso por debajo de $20.000.000, siendo el mínimo de $1.000.000 y el máximo de $700.000.000 es cual es un dato atípico para esta variable. Respecto al promedio del valor de créditos, se evidencia que el promedio es de $12.700.437, con un mínimo de $118.886 y un máximo de $300.000.000, esto representa una gran variabilidad de los datos.

El promedio de la tasa efectiva anual es del 22.56% con una minina del 11.20% y máxima de 42.60%, las cuales son asignadas de acuerdo al destino económico, el plazo y el monto por el cual se otorgue el crédito, en cuanto a los días de mora el promedio es de 17 días , con un máximo de 2.658 días, el 75% de los asociados presentan cero días de mora, lo que indica un buen habito de pago pago dentro de los datos analizados, el plazo está representado por los meses a los que fue otorgada la operación de crédito, es decir un crédito de 1 año es equivalente a un plazo de 12 meses y se evidencia un mínimo de 7 meses y un máximo de 180 meses, en su promedio asciende a 46 meses. Respecto a las cuotas pagadas se puede explorar que la media es de 14 cuotas, y el 75% de los asociados llevan 22 cuotas pagadas. La última variable cuantitativa cuotas pendientes refleja un promedio de 32 meses y un máximo de 154 cuotas, se determina que en promedio los asociados tienen más cuotas pendientes que las cuotas pagadas, a su vez el 75% de los asociados presentan por debajo de 46 cuotas pendientes. Lo que quiere decir que estima que el recaudo de cartera cumpla con el plan de pagos acordado al inicio de cada obligación financiera.

En cuanto a la descripción de las variables categóricas, se involucra género conformado por 56,46% hombres, 43,54% mujeres. según el estrato socioeconómico se encuentra discriminado en:

Estrato 1 con un 11,95%, estrato 2 con 34,03%, estrato 3 con 32,93%, estrato 4 con 17,03%, estratos 5, 6 y 7 con 2,74%, 0,72% y 0,60%, respectivamente. Se observa que la mayor concentración de asociados se encuentra en los estratos 2 y 3 denominamos como Bajo y Medio-bajo. En cuanto a la calidad de la cartera, se encontró que 94,39% está clasificada en A (asociados con buenos reportes de cumplimiento de pagos), 1,21% en B, 0,61% en C, 0,89% en D y 2,91% en E.

3.2.2 Matriz de Correlaciones. La tabla 2 refleja los resultados de la matriz de correlación. Se observa que la mayoría de las variables presentan correlaciones lineales bajas. Se destacan la correlación directa entre Activos con Pasivos (0,5977), Ingresos (0,40156), valor desembolso (0,56847) y saldo promedio crédito (0,5660).

Tabla *2.*
Matriz de correlación de variables numéricas

Variables	Edad	Activos	Pasivos	Ingresos	Saldo Depositos	Apalan-camiento	Créditos Prepagos	Número Créditos
Edad	1	0.31919	0.14115	0.05108	0.09906	-0.18065	0,10657	0.13991
Activos	0.31919	1	0.59776	0.40156	0.13385	-0.20672	0,04881	0.07810
Pasivos	0.14115	0.59776	1	0.36628	0.05857	0.36778	0,04370	0.08445
Ingresos	0.05108	0.40156	0.36628	1	0.04942	-0.00804	0,02223	0.04899
Saldo Depósitos	0.09906	0.13385	0.05857	0.04942	1	-0.03465	0,05200	0.07422
Apalancamiento	-0.18065	-0.20672	0.36778	-0.00804	-0.03465	1	-0.01213	0.00288
Créditos Prepagados	0.10657	0.04881	0.04370	0.02223	0.05200	-0.01213	1	0.80937
Número Créditos	0.13991	0.07810	0.08445	0.04899	0.07422	0.00288	0.80937	1
Valor Desembolso	0.10066	0.56847	0.50686	0.34383	0.13693	0.05653	0.06920	0.12665
Saldo Promedio Crédi	0.07234	0.56608	0.48683	0.33806	0.10760	0.05808	-0.11684	-0.13413
Tasa E.A.	0.03269	-0.28659	-0.25431	-0.07942	-0.06039	-0.11394	0.00695	-0.00182
Dias de Mora	-0.02363	-0.01677	0.00087	0.00011	-0.01897	0.03110	-0.03137	-0.00502
Plazo	0.03262	0.15808	0.19658	-0.00860	0.04129	0.15491	0.01002	0.03954
Cuotas pagadas	0.13368	0.08896	0.09905	0.01245	0.01784	0.06743	-0.12832	-0.04047
Cuotas pendientes	-0.06349	0.09570	0.12736	0.01769	0.02888	0.10803	0.10277	0.06890
Años antigüedad	0.37681	0.18304	0.08415	0.05551	0.17391	-0.07620	0.18505	0.26326

Variables	Valor Desembolso	Saldo Promedio Crédito	Tasa E.A	Días de mora	plazo	uotas Pagada	Cuotas Pendientes	Años Antigüedad
Edad	0.10066	0.07234	0.03269	-0.023633	0.03262	0.13368	-0.06349	0.37681
Activos	0.56847	0.56608	-0.28659	-0.016779	0.15808	0.08896	0.09570	0.18304
Pasivos	0.50686	0.48683	-0.25431	0.000876	0.19658	0.09905	0.12736	0.08415
Ingresos	0.34383	0.33806	-0.07942	0.000111	-0.00860	0.01245	-0.01769	0.05551
Saldo Depósitos	0.13693	0.10760	-0.06039	-0.018973	0.04129	0.01784	0.02888	0.17391
Apalancamiento	0.05632	0.05808	-0.11394	-0.031102	0.15491	0.06743	0.10803	-0.07620
Créditos Prepagados	0.08692	-0.11684	0.00695	0.031972	0.01002	-0.12832	0.10277	0.18505
Número Créditos	0.12665	-0.13413	-0.00182	-0.005022	0.03954	-0.04004	0,06890	0.26326
Valor Desembolso	1	0.84827	-0.33269	0.000758	0.40692	0.14136	0.30960	0.15052
Saldo Promedio Crédi	0.84827	1	-0.34632	0.001389	0.35413	0.13651	0.25970	0.03620
Tasa E.A.	-0.33269	-0.34632	1	0.005063	-0.38090	-0.10118	-0.31228	-0.03620
Dias de Mora	-0.00074	0.00138	0.00506	1	0.07521	0.13770	-0.02331	0.01726
Plazo	0.40692	0.35413	-0.38090	0.075215	1	0.37279	0.74250	0.11778
Cuotas pagadas	0.14136	0.13651	-0.10118	0.137701	0.37279	1	-0.34474	0.19094
Cuotas pendientes	0.30960	0.25970	-0.31228	-0.023314	0.74250	-0.34474	1	-0.01868
Años antigüedad	0.15052	0.03620	-0.03357	0.017261	0.11778	0.19094	-0.01868	1

Fuente: Ejecución de la matriz de correlaciones para valorar la fuerza y orientación de la relación entre las variables numéricas.

3.3 Análisis de Cluster

Con relación al tratamiento de la información, resultado de la depuración (mencionada en el literal 2.2) y análisis exploratorio (mencionado en el literal 3.1), se contó con una base de 33.307 datos, con el propósito de originar el agrupamiento de los asociados. Se incluyeron sólo variables continuas, por lo tanto, se aplicó la métrica euclídea como medida de distancia entre los individuos.

3.3.1 Estimación del número de grupos: El número de grupos fue determinado inicialmente por las investigadoras, a su vez contemplado en uno de los objetivos específicos. Se establecieron tres grupos para diferenciar a los asociados según la similitud en el comportamiento financiero.

3.3.2 Aplicación del algoritmo de agrupación: Por la dimensión de la base de datos se usó el algoritmo CLARA (Clustering Large Aplication), el cual utiliza para la agrupación el método de k-medoids. A diferencia del algoritmo k-means, el cual utiliza como punto de agrupación un punto aleatorio del plano, el algoritmo k-medoids utiliza como punto de agrupación para cada clúster un elemento (medoides) de la base de datos. Los medoides pueden entenderse como puntos de un grupo cuya disimilaridad media a todos los demás puntos del grupo es mínima.

3.3.3 Variables Seleccionadas: Inicialmente, se pensó en realizar un análisis de componentes principales para la reducción de las variables y la inclusión de las nuevas variables latentes como información para el análisis clúster, sin embargo, no se realizó por la baja correlación observada en las variables.

Para la preparación de los datos el algoritmo CLARA estandariza las variables a usar, eliminando así la influencia en la conformación de los grupos de aquellas variables con valores muy altos. Inicialmente se incluyeron todas las variables para la conformación de los grupos, lo cual permitió la detección de los datos outliers que se retiraron previamente.

Esta primera conformación de los clústeres utilizando todas las variables continuas (edad, activos, pasivos, ingresos, saldo depósitos, créditos prepagados, número de créditos, cuotas pagadas, cuotas pendientes, antigüedad, apalancamiento, valor desembolso, saldo promedio de créditos, tasa, días de mora y plazo) proporcionó el 36.9% de explicación de los datos

Al analizar los resultados de estas agrupaciones iniciales se encontraron variables que no aportaban a la conformación de los grupos y fueron retirándose y revisando cada uno de los resultados encontrados.

Finalmente se dejaron los resultados de los clústeres formados al comparar los individuos más parecidos (y más diferentes) en términos de las variables Activos, Pasivos, Saldo de depósitos, Apalancamientos, Valor de desembolso, Tasa Efectiva, Plazo y Saldo Promedio de Créditos, que presentaron los mejores resultados frente a la variabilidad captada.

La conformación de los clústeres con estas variables (activos, pasivos, saldo depósitos, apalancamiento, valor desembolso, saldo promedio de créditos, tasa y plazo) proporcionó el 56.2% de explicación de los datos

3.3.4 Identificación del clúster asignado a cada asociado

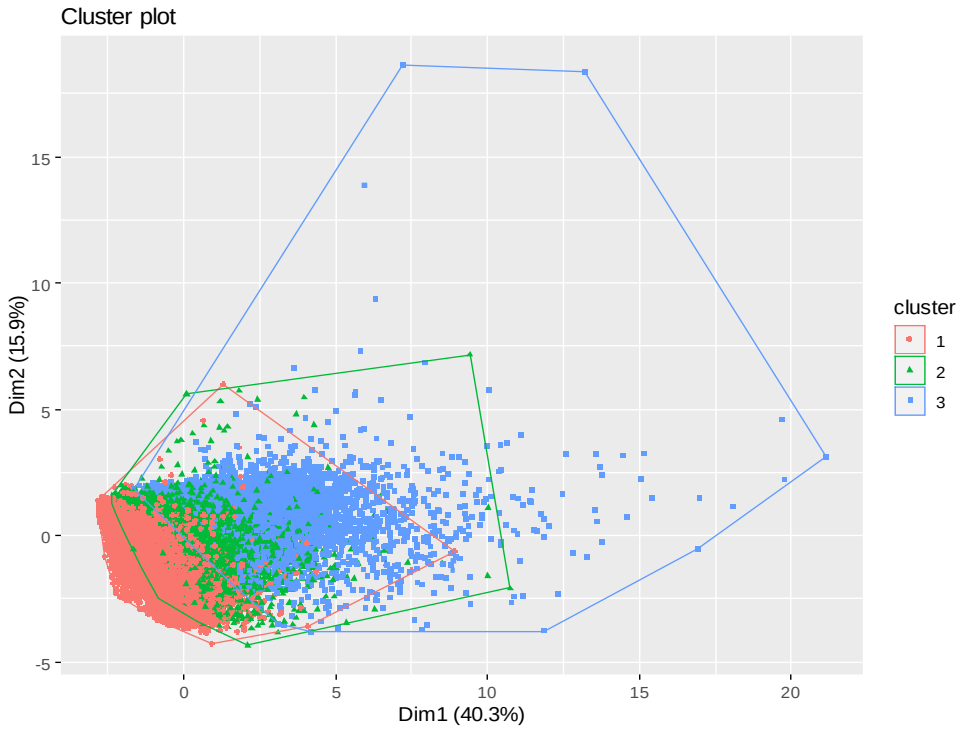
Tabla 3.
Resumen del clúster seleccionado

	Activos	Pasivos	Saldo Depósitos	Apalan-camiento	Valor Desembolso	Saldo Prom. Crédito	Tasa E.A	Plazo
15331	30.000.000	10.000.000	402.620	0,333	7.000.000	6.000.000	0,182	60
28123	142.350.000	13.692.000	1.557.416	0,096	16.500.000	11.833.333	0,319	60
20585	371.000.000	46.585.000	685.416	0,126	50.000.000	32.500.000	0,140	36

Fuente: Aplicación del análisis de clúster en el estudio de las variables seleccionadas

El grupo 1 está conformado por 18595, con un isolation de 1,545820, indicando que las distancias entre los individuos son menores frente al grupo 2, es decir se presenta una menor circunferencia.

Figura 2.
Clúster de los asociados



Fuente: Grafico donde se refleja la agrupación de los asociados en los tres grupos asignados de manera previa

El grupo 2 está conformado por 10.328 individuos con un isolation de 1,975972, indicando que las distancias entre los individuos son mayores al del grupo 1, es decir, sus individuos presentan un poco más de dispersión respecto al medio de como centro del grupo.

El grupo 3, está conformado con 4.384 asociados, los cuales presentan una mayor distancia frente a los dos grupos anteriores y se da una circunferencia más dispersa, indicando que los datos se encuentran mucho más alejados del punto central o del centroide.

A continuación, se presenta un resumen descriptivo de los clúster

- **Resumen del clúster 1**

Tabla *4.*
Resumen descriptivo del grupo o clúster 1

	Edad	Activos	Pasivos	Ingresos	Saldo Depósitos
Minimo	18,00	5.000.000	830.000	890.616	524
1er Quartil	31,00	19.200.000	2.857.500	1.500.000	121.722
Mediana	39,00	35.000.000	5.813.000	2.200.000	306.713
Promedio	40,87	38.748.004	9.558.412	3.925.512	748.451
3er Quartil	50,00	58.183.558	12.676.000	4.000.000	744.415
Máximo	78,00	87.700.000	80.000.000	260.000.000	118.517.750

	Apalan-camiento	Créditos Prepagados	Número Créditos	Valor Desembolso	Saldo Promedio Créditos
Minimo	0,0100	0,0000	1,0000	1.000.000	334.812
1er Quartil	0,1000	0,0000	1,0000	4.000.000	3.000.000
Mediana	0,2170	0,0000	2,0000	7.000.000	5.500.000
Promedio	0,2847	0,7813	2,6040	9.293.556	7.605.222
3er Quartil	0,4010	1,0000	3,0000	12.000.000	10.000.000
Máximo	1,0000	45,0000	87,0000	150.000.000	129.000.000

	Tasa_ EA	Dias de mora	Plazo	Cuotas Pagadas	Cuotas Pendientes
Minimo	0,1120	0,0000	7,00	0,00	1,00
1er Quartil	0,1820	0,0000	24,00	1,0	17,00
Mediana	0,2100	0,0000	36,00	6,0	28,00
Promedio	0,2453	0,6498	41,78	11,1	30,69
3er Quartil	0,3190	0,0000	60,00	17,0	43,00
Máximo	0,4260	15,0000	180,00	102,0	152,00

Fuente: Resumen general de las variables asociadas al cluster 1, en donde se identifican las particularidades de cada agrupamiento.

De acuerdo con las características del individuo central de este clúster (medoides) y los resúmenes descriptivos de las variables de este grupo, podemos clasificar a este grupo como el de asociados con el mayor apalancamiento, menor saldo de depósitos, más días de mora (menor calidad de cartera) y es por tanto catalogado como el grupo de asociados para los cuales **no aplica beneficio**.

- **Resumen del clúster 2**

Tabla 5.
Resumen descriptivo del grupo o clúster 2

	Edad	Activos	Pasivos	Ingresos	Saldo Depósitos
Minimo	20,0	77.332.000	830.000	890.616	524
1er Quartil	39,0	110.000.000	7.500.000	2.200.000	221.288
Mediana	49,0	150.000.000	18.000.000	3.751.000	572.052
Promedio	49,0	151.973.434	26.846.229	7.510.142	1.455.666
3er Quartil	58,0	190.000.000	37.859.750	7.500.000	1.289.785
Máximo	78,0	266.399.000	205.125.000	300.000.000	148.211.108

	Apalan-camiento	Créditos Prepagados	Número Créditos	Valor Desembolso	Saldo Promedio Créditos
Minimo	0,0030	0,0000	1,0000	1.00e+06	385.922
1er Quartil	0,0520	0,0000	1,0000	8.00e+06	5.849.909
Mediana	0,1220	0,0000	2,0000	1.30e+07	10.200.000
Promedio	0,1830	0,9024	2,9410	1.69e+07	13.193.936
3er Quartil	0,2530	1,0000	3,0000	2.10e+07	17.000.000
Máximo	1,0000	49,0000	68,0000	2.00e+08	150.000.000

	Tasa_ EA	Dias de mora	Plazo	Cuotas Pagadas	Cuotas Pendientes
Minimo	0,1120	0,0000	7,00	0,00	1,00
1er Quartil	0,1680	0,0000	36,00	3,0	21,00
Mediana	0,1960	0,0000	60,00	9,0	34,00
Promedio	0,2105	0,5981	48,80	13,9	34,87
3er Quartil	0,2390	0,0000	60,00	22,0	50,00
Máximo	0,4260	15,0000	120,00	102,0	120,00

Género	
Femenimo	4.565
Masculino	5.763

Fuente: Resumen general de las variables asociadas al cluster 2, en donde se identifican las particularidades de cada agrupamiento.

De acuerdo con las características del individuo central de este clúster (medoides) y los resúmenes descriptivos de las variables de este grupo, podemos clasificar a este grupo como el de asociados con **beneficio Básico.**

- **Resumen del clúster 3**

Se ha catalogado a éste clúster como el grupo cuyos asociados obtendrán el beneficio "**Premium"**. Este grupo se caracteriza a por asociados que presentan menos días de mora, su plazo de crédito en promedio es de 50 meses, lo que nos permite tener un retorno acorde al plazo total, son asociados representativos por el valor del desembolso de crédito con un promedio de $37.000.000, su valor de activos en promedio oscila en $411.000.000, en cuanto al saldo promedio de créditos se encuentra en $28.000.000.

Tabla 6.
Resumen descriptivo del grupo o clúster 3

	Edad	Activos	Pasivos	Ingresos	Saldo Depósitos
Minimo	22,0	238.000.000	890.000	900.000	524
1er Quartil	44,0	300.000.000	20.000.000	4.800.000	384.104
Mediana	53,0	380.000.000	49.940.500	12.000.000	945.720
Promedio	52,1	411.432.921	72.896.314	22.776.637	2.790.472
3er Quartil	60,0	481.000.000	100.000.000	27.128.842	1.918.483
Maximo	78,0	800.000.000	612.213.658	726.000.000	370.168.144

	Apalan-camiento	Créditos Prepagados	Número Créditos	Valor Desembolso	Saldo Promedio Créditos
Minimo	0,0010	0,00	1,00	1.000.000	957.200
1er Quartil	0,0530	0,00	1,00	16.000.000	12.500.000
Mediana	0,1250	0,00	2,00	30.000.000	22.175.364
Promedio	0,1766	1,05	3,41	37.129.130	28.165.586
3er Quartil	0,2490	1,00	4,00	50.000.000	37.682.609
Maximo	0,9870	42,00	68,00	400.000.000	300.000.000

	Tasa_ EA	Dias de mora	Plazo	Cuotas Pagadas	Cuotas Pendientes
Minimo	0,1120	0,0000	7,00	0,00	1,00
1er Quartil	0,1540	0,0000	36,00	4,0	20,00
Mediana	0,0176	0,0000	60,00	10,0	34,00
Promedio	0,0185	0,5370	50,11	14,5	34,59
3er Quartil	0,1960	0,0000	60,00	23,0	51,00
Maximo	0,4260	15,0000	180,00	96,0	154,00

Género	
Femenimo	1.923
Masculino	2.461

Fuente: Resumen general de las variables asociadas al cluster 3, en donde se identifican las particularidades de cada agrupamiento.

3.3.5 Comportamiento de las variables según el clúster.

3.3.5.1 Valor desembolsado; La figura 3 muestra el valor de desembolso. Donde se aprecia que los tres cluster presentan una distribución de las observaciones con asimetría a la derecha, con datos más dispersos en el cluster 3, y este presenta mayor valor de desembolso. Y el cluster 1, releja una observación atípica

Figura 3

Diagrama de caja para valor de desembolso por clúster

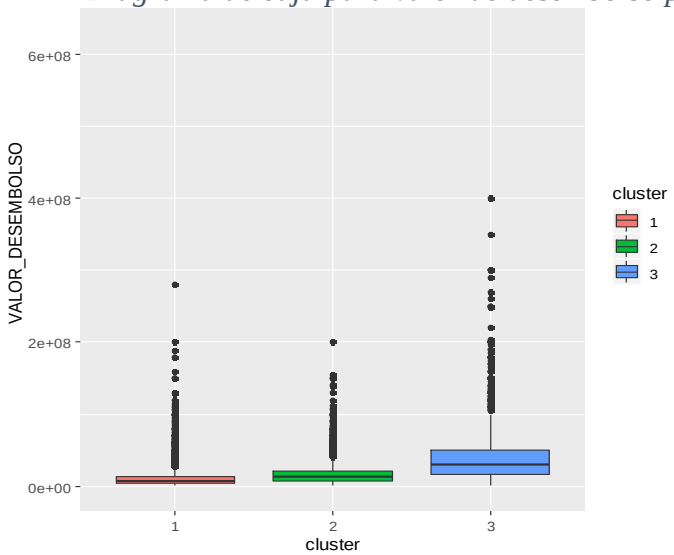


Figura 4. Diagrama de caja para la variable valor de desembolso, que obedece al dinero que la Cooperativa entregó al asociados por el créditos aprobado corte de Diciembre de 2019, y que se descrita para cada uno de los clúster conformados

3.3.5.2 Apalancamiento

Figura 5.
Diagrama de caja para Apalancamiento por clúster

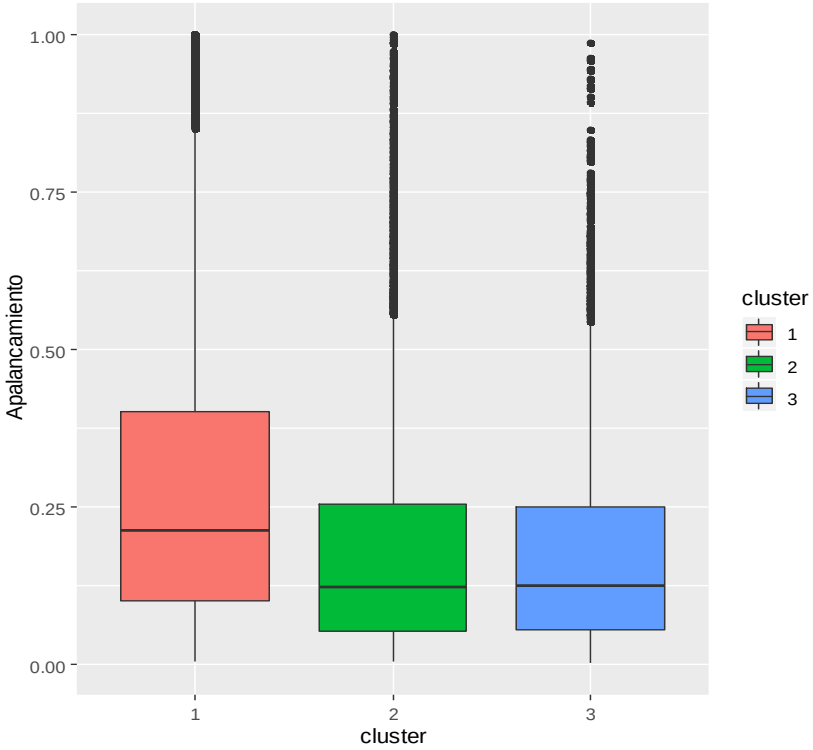


Figura 6. Diagrama de caja del indicador Apalancamiento, aplicado a los datos de pasivos sobre activos de cada asociado incluido en el estudio, a su vez, refleja la descripción del indicador por cada uno de los clúster conformados.

La figura 4 muestra el comportamiento de la variable apalancamiento. Se observa que los tres clústeres presentan una distribución de las observaciones con asimetría a la derecha. El clúster 1 es el que presenta el mayor apalancamiento (elemento negativo para la confianza del cliente)

3.3.5.3. ***Plazo: La figura 5,*** se observa el comportamiento de plazo de crédito en los tres cluster, los cuales comparten la misma mediana del tercer cuartil, y se da mayor dispersión en el cluster 1, este junto con el cluster 3, presentan datos atípicos.

Figura 7.
Diagrama de caja para plazo por clúster

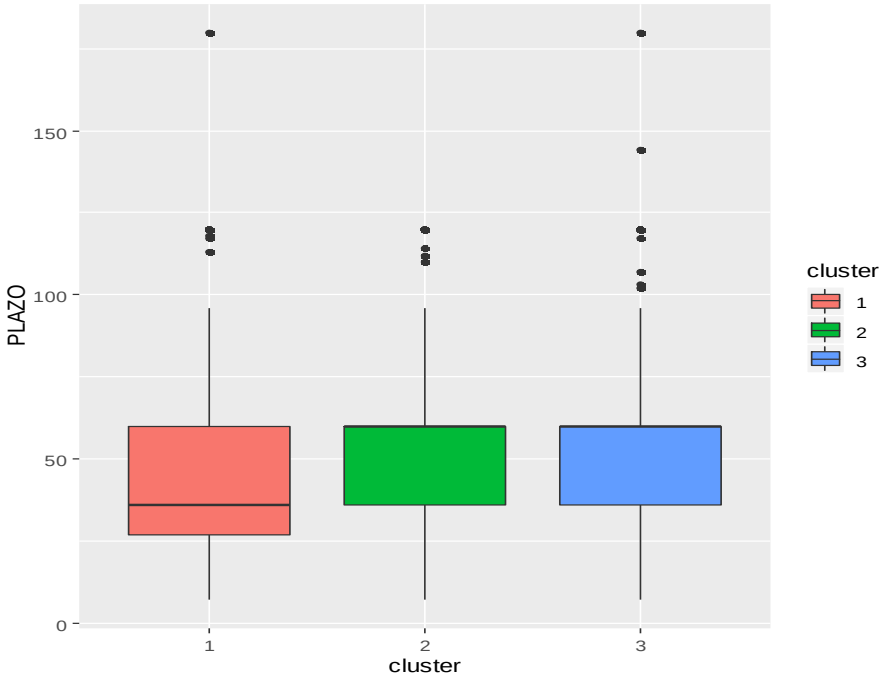


Figura 8. Diagrama de caja para la variable plazo, referente al tiempo señalado para cancelar de manera normal la obligación financiera, en donde se describe para cada clúster conformado

3.3.5.4 Tasa de Interés: La figura 6, se observa el comportamiento de la tasa de interés, donde todos los cluster presentan un comportamiento asimétrico a la derecha, y el cluster 1 es el que presenta la mayor distribución en la observación de los datos, el cluster 2 y 3 refleja datos más extremos de su media, en comparación con el cluster

.

Figura 9.
Diagrama de caja para tasa de interés por clúster

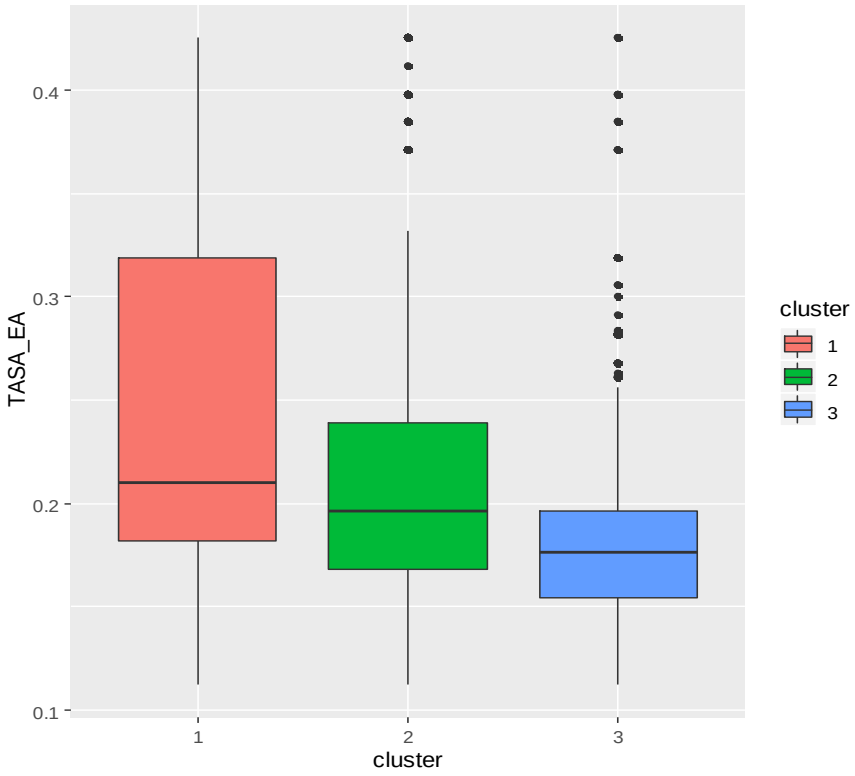


Figura 10. *Diagrama de caja que representa la variable tasa de interés, el cual es el porcentaje que el asociado asume pagar como contraprestación por utilizar los recursos de la cooperativa, por cada clúster definido.*

3.3.5.5 Ingresos. Mediante la figura 7, se aprecia que el comportamiento de los ingresos de los asociados en los tres cluster presenta un comportamiento asimétrico a la derecha. En el cluster 3 se refleja mayor ingreso de los asociados. Así mismo, mediante este cluster se da una mayor dispersión de los datos de ingresos.

Figura 11.
Diagrama de caja para Ingresos por clúster

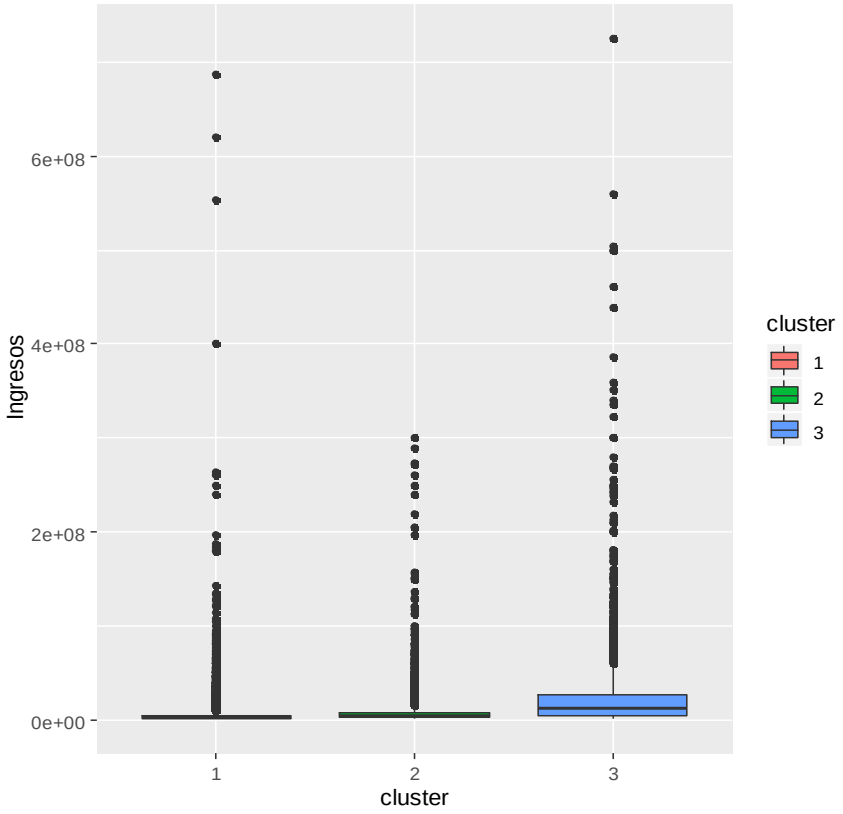


Figura 12. Diagrama de caja para la variable Ingresos, atribuidos a la información mensual de los asociados del estudio y que se describe en el agrupamiento de los tres cluster.

3.3.5.6 Activos: En la figura 8. Se observa que en los tres cluster los activos de los asociados presentan una distribución asimétrica a la derecha, siendo más pronunciada en el cluster 1 y 3. Este último tiene mayores activos frente a los otros dos cluster, lo que beneficia a los asociados, el cluster 2 presenta mayores activos.

Figura 13.
Diagrama de caja para Activos por clúster

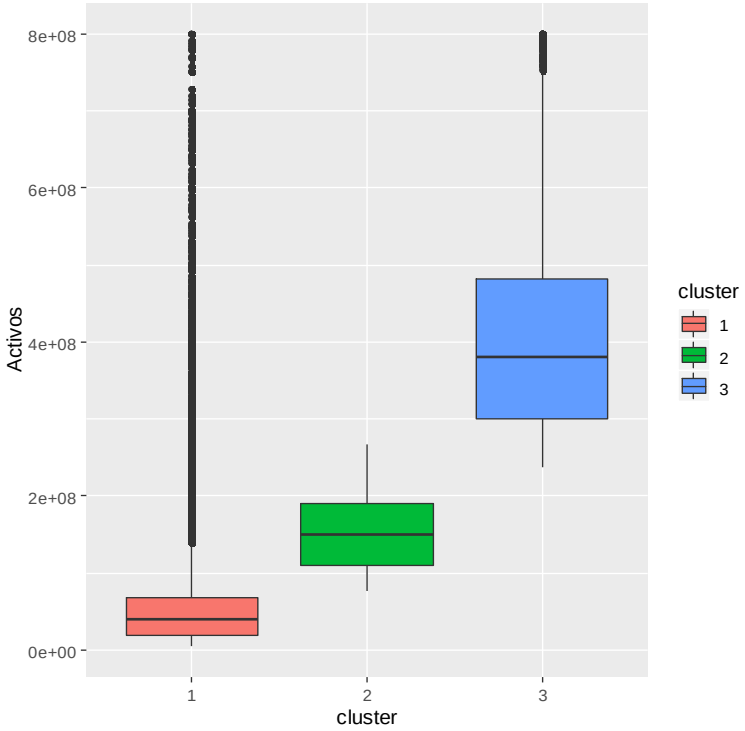
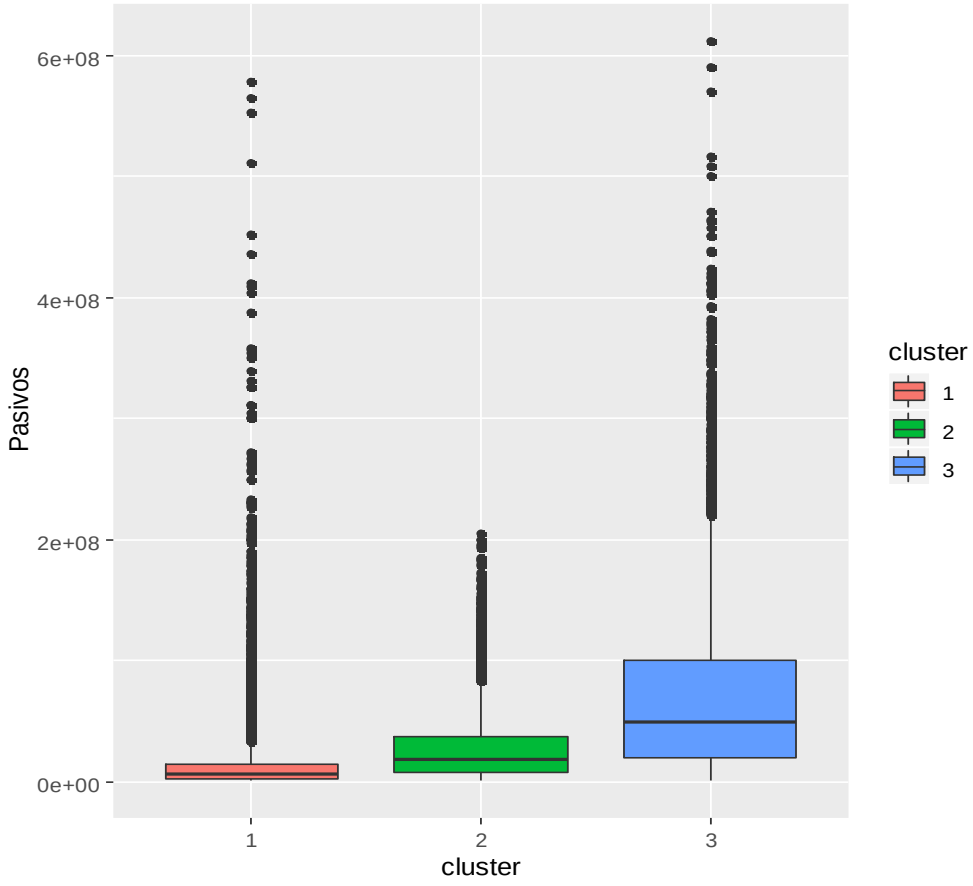


Figura 14. Diagrama de caja para Activos que hacen referencia por clúster

3.3.5.7 ***Pasivos:*** En la figura 9 muestra el comportamiento de la variable pasivos, se observa que los tres cluster presentan una distribución de las observaciones con asimetría a la derecha, dándose mayor dispersión en el cluster 3 y el cluster 1. El cluster 3 es el que presenta mayores pasivos.

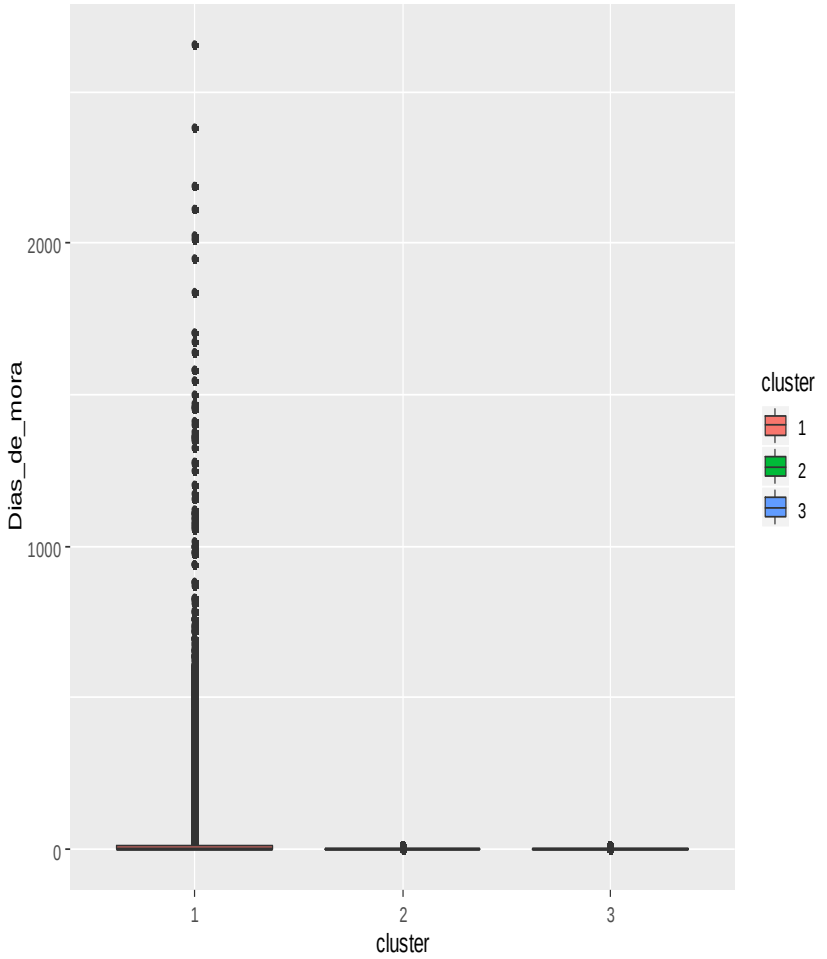
Figura 15.
Diagrama de caja para Pasivos por clúster



3.3.5.8 Días de Mora El cluster 1, se observa que entre el Q1 y Q3, hay una alta concentración de los días de mora. Sin embargo, el 25% de los datos tienen una alta dispersión entre los días de mora, y tienen datos atípicos, hay asociados que presentan días de mora alta, los cuales son créditos que tienen alta morosidad y se encuentran en estados pre jurídico o jurídico.

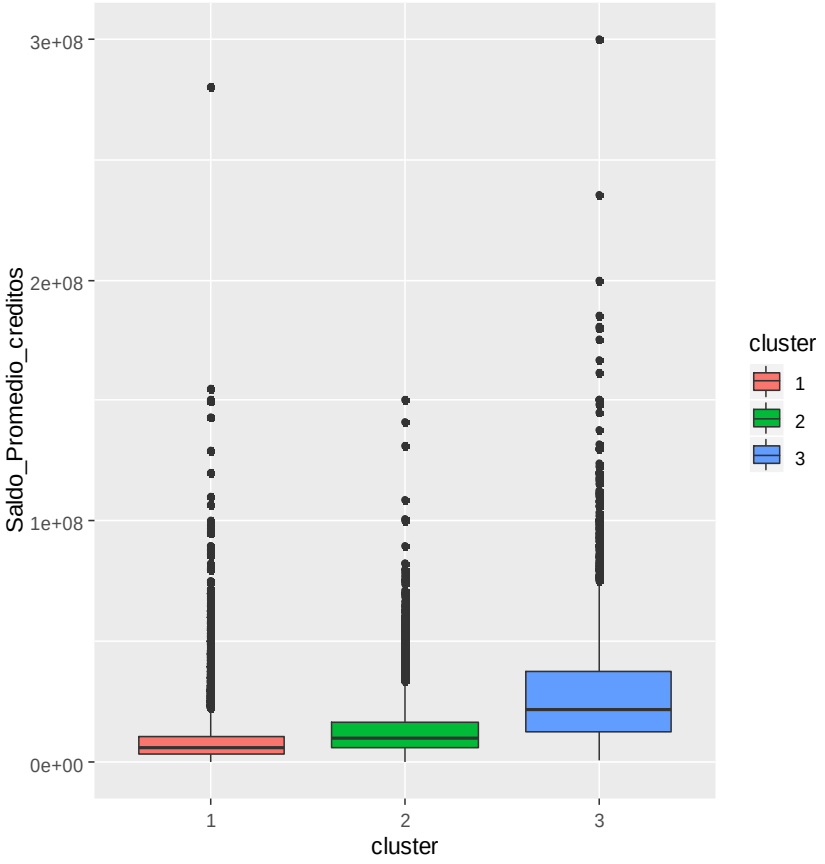
El cluster 2 y 3, son asociados que tienen una alta concentración de asociados que no presentan mora, así mismo, el día de mora máximo son 15 días. Siendo estos asociados que mantienen una cartera saludable.

Figura 16.
Diagrama de caja para Días de mora por clúster



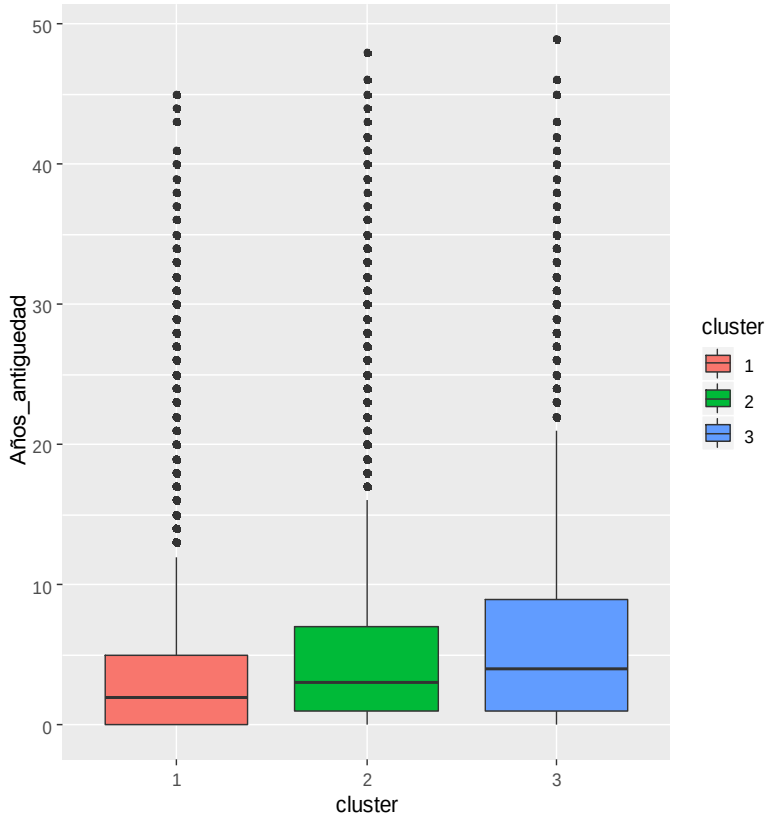
3.3.5.9 Saldo Promedio de Créditos. En la figura 11 muestra el comportamiento de la variable saldo promedio de crédito, se observa que los tres cluster presentan una distribución de las observaciones con asimetría a la derecha. El cluster 3 es el que presenta mayor saldo promedio de crédito, lo que indica que son los asociados que más tienen obligación financiera con la Cooperativa.

Figura 17.
Diagrama de caja para saldo promedio de créditos por clúster



3.3.5.10 Años de Antigüedad: En la figura 12 muestra el comportamiento de la variable años de antigüedad, se observa que los tres cluster presentan una distribución de las observaciones con asimetría a la derecha. El cluster 3 es el que presenta mayor año de antigüedad de los asociados en la Cooperativa, lo que beneficia su relación con la Cooperativa.

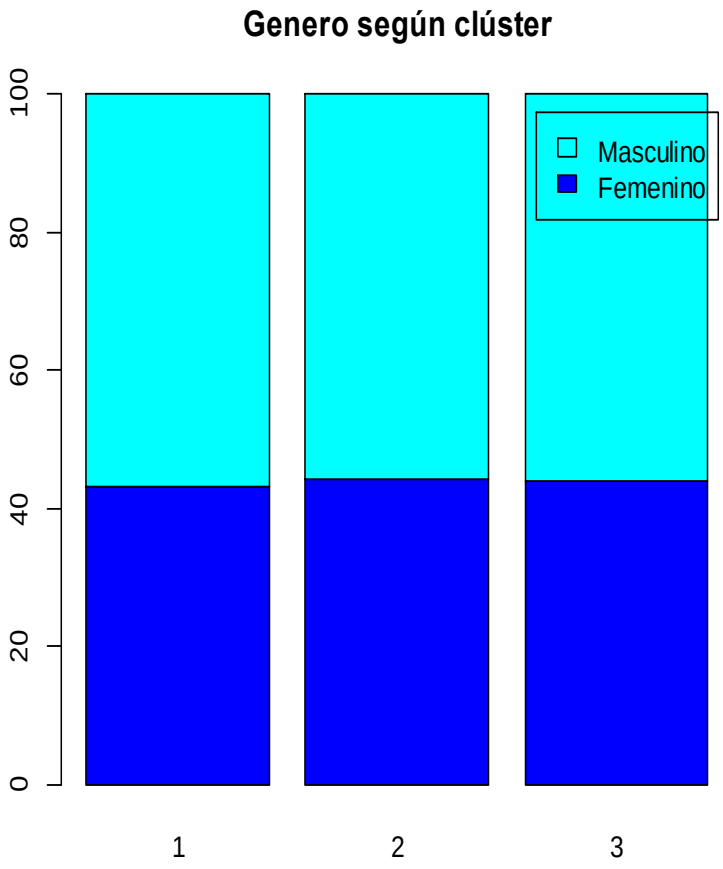
Figura 18.
Diagrama de caja para años de antigüedad por clúster



3.3.6. Análisis para perfilamiento de Asociados

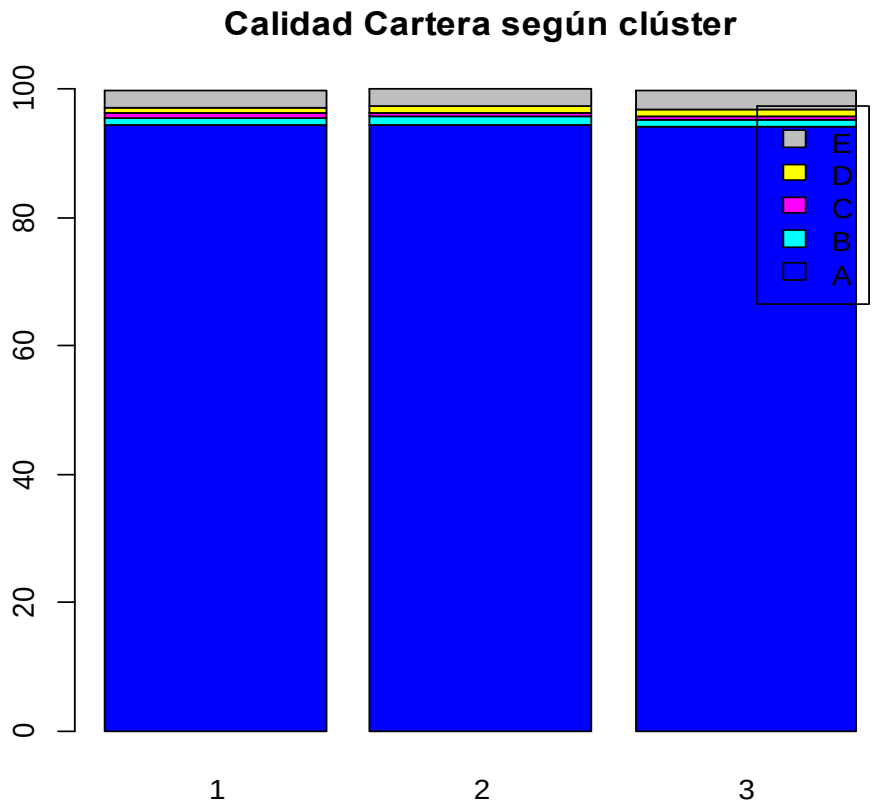
3.3.6.1. Género: La distribución de los asociados según el género la mayor proporción corresponde al personal masculino en los tres grupos, para el grupo 1 con el 56.89%, el grupo 2 con el 55.79% y el 3 con el 56.13% y las mujeres representan una proporción similar, en el grupo 1 con el 43.10%, el grupo 2 con el 44.20% y el grupo 3 con el 43.86%.

Figura 19.
Gráfico de barras para género por clúster



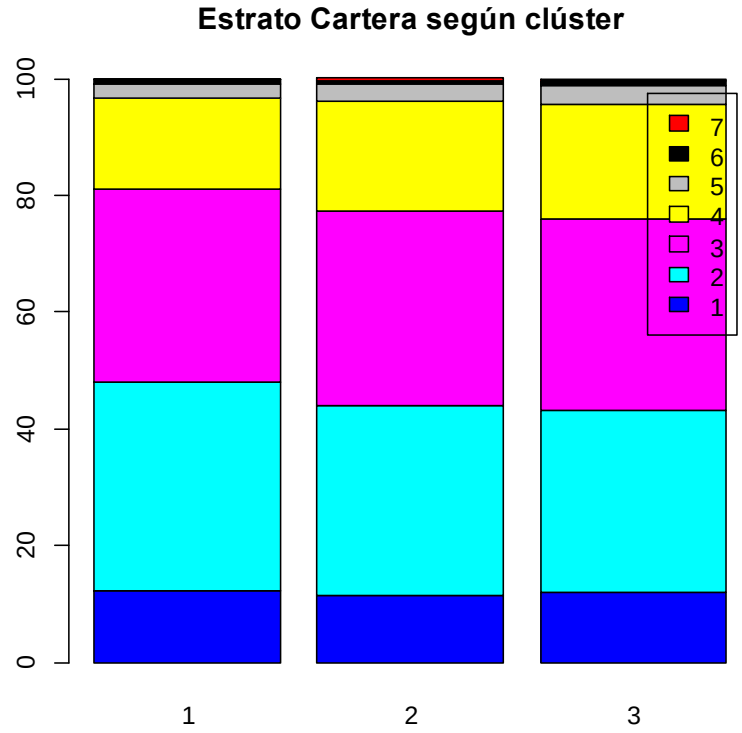
3.3.6.2 Calidad de la Cartera: Respecto a la calidad de la cartera los grupos están concentrados en los asociados con un buen habito de pago con un 94.40% para el grupo 1, 94.51% para el grupo 2 y 94.13% para el grupo 3. Este comportamiento es catalogado como cartera saludable, basados en las políticas internas de la cooperativa en estudio. No se observan diferencias significativas en esta variable.

Figura 20.
Gráfico de barras para calidad de la cartera por clúster



3.3.6.3 Estrato Socioeconómico: En cuanto a la distribución del estrato socioeconómico, los grupos se distribuyen también en similares proporciones, concentrando la mayor población en dos niveles "Bajo" con el 35.77% para el primer grupo, 32.45% para el segundo grupo y 31.29% para el tercer grupo y en el "Medio Bajo" con el 33.01% para el primer grupo, 33.20% y 32.82% respectivamente.

Figura 21.
Gráfico de barras para estrato por clúster



3.4 Análisis Modelo Logit Multinomial

Conformado el clúster y decidido el tipo de beneficio a otorgar según las características de los asociados, se procedió al planteamiento del modelo scoring mediante un modelo logit multinomial.

Sin embargo, como el grupo 3 tiene un número de individuos bastante menor que los otros dos grupos , para evitar el sesgo que este desbalance pueda causar en el modelamiento, se realizó una división en la base de datos (esto mismo se hizo para el ordinal y no fue solución) de la siguiente manera: Se tomó una base de entrenamiento (con la cual se formularon los diferentes modelos) de 12.286 individuos, conformados por el 80% del clúster 3 (3507 individuos), 40% del cluster 2 (4131 individuos) y el 25% del cluster 1 (4648 individuos) y una base de datos para la evaluación

del modelo de 21.021 datos. Estos 21021 asociados fueron usados para la evaluación de la bondad de ajuste de cada uno de los modelos creados.

En primer lugar, se procedió a realizar la modelación de la variable tipo de beneficio (o clúster en nuestro caso) mediante regresión logistica ordinal, la cual se emplea cuando la variable respuesta tiene un orden y se desea detectar la forma en que la respuesta está relacionada con las variables independientes. El modelo más común a usar es el de Odds proporcionales, el cual se basa en el supuesto de que el efecto de las covariables $X1,\ldots,Xp$ es igual para todas las categorías en la escala logarítmica (supuesto de rectas paralelas)

A continuación se muestra uno de los modelos de regresión ordinal cuyas variables tuvieron significancia estadística:

Modelo: cluster ~ Apalancamiento + Años_antiguedad, data = entrenamiento

Tabla 7.
Significancia y coeficientes de un modelo ordinal

Variable	Coeficiente	Error estándar	Valor t
Apalancamiento	-204.777	0.089634	-22.85
Años antigüedad	0.04552	0.002496	18.24

Interceptos	Coeficiente	Error estándar	Valor t
1|2	-0.7234	0.0303	-239.000
2|3	0.7808	0.0305	255.728

Sin embargo, al validar el modelo mediante el supuesto de rectas paralelas (cumplimiento de la proporcionalidad de odds) éste no se cumplió (ver tabla 8). Ante el no cumplimiento del supuesto de rectas paralelas el modelo no es válido.

H0: Se cumple el supuesto de rectas paralelas

Tabla 8.
Evaluación test de rectas paralelas (odds proporcionales)

Test Para	X2	df	probability
Omnibus	9.640.809	2	1,16E - 15
Apalancamiento	7.546.041	1	3,73E - 12
Años_antiguedad	1.893.033	1	1,36E + 01

Esta situación fue común en cada uno de los modelos de regresión ordinal en los que se encontraron variables significativas (a pesar de haber eliminado el sesgo del tamaño de los grupos o categorías). Por ello se tomó la decisión de realizar la modelación mediante logit multinomial.

A continuación, se presentan los coeficientes del modelo final con un AIC de 21.738,67 (el menor de todos los modelos trabajados) y cuyas variables con significancia estadística del 5% son Ingresos, Días de mora, Años de antigüedad y Plazo. En los anexos se muestran algunos de los resultados de diversos modelos ajustados.

Tabla 9.
Significancia y coeficientes del modelo final

multinom (formula = Cluster ~ Ingresos + Días de mora + Años Antigüedad + Plazo + data = Entrenamiento)

Coeficientes:	**(Intercepto)**	**Ingresos**	**Días de mora**	**Años de Antigüedad**	**Plazo**
2	- 1,4642	8,325E-08	- 0,13801	0,044489	0,022075
3	- 2,7927	1,347E-07	- 0,16944	0,065186	0,031074

Std. Errores:	**(Intercepto)**	**Ingresos**	**Días de mora**	**Años de Antigüedad**	**Plazo**
2	3,183149e-16	3,064078e-09	5,563029e-16	1,3262409e-15	1,380193e-14
3	2,281209e-16	3,059268e-09	4,470999e-16	1,142167e-15	1,025956e-14

Residual Deviance: 21718,67
AIC: 21738,67

Analysis of Deviance Table (Type II tests)

Response: Cluster

	LR Chisq	Df	Pr (> Chisq)	
Ingresos	2434,2	2	< 2,2e -16	***
Días de Mora	2400,15	2	< 2,2e -16	***
Años Antigüedad	260,57	2	< 2,2e -16	***
Plazo	469,17	2	< 2,2e -16	***

Signif. Codes: 0 "*" 0,001 " * * " 0,01 "*" 0,1**

3.4.1 Interpretación de los coeficientes: El signo positivo de las variables Ingresos, Años de antigüedad y Plazo muestran una relación directa entre el aumento de estas variables y la probabilidad de ser asignados a los clúster 2 o 3 (los clúster que obtendrán beneficio). Es decir que, al aumentar los ingresos aumenta la probabilidad de estar en la categoría dos o tres respecto a la categoría uno (no aplica), de igual forma, dicha probabilidad aumenta el aumentar los años de antigüedad del asociado o mayor tiempo de plazo de pago del crédito. En el caso de los días de mora, la relación es inversa (signo negativo de los coeficientes), por lo tanto, al disminuir los días de mora, aumenta la probabilidad de estar en la categoría dos o tres.

Este modelo tiene en cuenta la fidelización del cliente (años de antigüedad) y el cumplimento en sus obligaciones (días de mora). De igual forma, altos ingresos dan confianza a la entidad financiera y un mayor plazo en el pago del crédito aumenta la rentabilidad.

3.4.2 Interpretación de los exponenciales de los coeficientes: Al aumentar en un peso los Ingresos de los asociados, aumenta en 0,00008% (exp (0,0000000832497) = 1.00000008) la probabilidad de obtener el beneficio básico (categoría 2) en comparación con no aplicársele beneficio. Para mejorar la interpretación, se asumen un valor de aumento en un millón de pesos,

en tal caso tenemos, dos asociados A y B, tales que el asociado B tiene un millón de pesos más en sus Ingresos que el asociado A (y todas las demás variables iguales):

$$OR(B/A) = \frac{Odd_B}{Odd_A} = e^k = e^{1000000*0,0000000832497} = e^{0.0832497} = 1.086813$$

Es decir, que al aumentar los ingreso en un millón de pesos, aumenta aproximadamente un 8,68% la probabilidad de obtener el beneficio Básico (categoría 2) en comparación con no aplicársele beneficio.

De igual forma, al analizar la probabilidad de obtener el beneficio Premium, en comparación con no obtener beneficio, se asume un valor de aumento en un millón de pesos en los ingresos, es decir, dos asociados A y B, tales que el asociado B tiene un millón de pesos más en sus ingresos que el asociado A (y todas las demás variables iguales):

$$OR(B/A) = \frac{Odd_B}{Odd_A} = e^k = e^{1000000*0,0000001347279} = e^{0.1347279} = 1.144225$$

Es decir, que por cada aumento en un millón de pesos en los ingresos, aumenta aproximadamente un 14,42% la probabilidad de obtener el beneficio Premium, en comparación con no tener Beneficio, suponiendo las demás variables constantes.

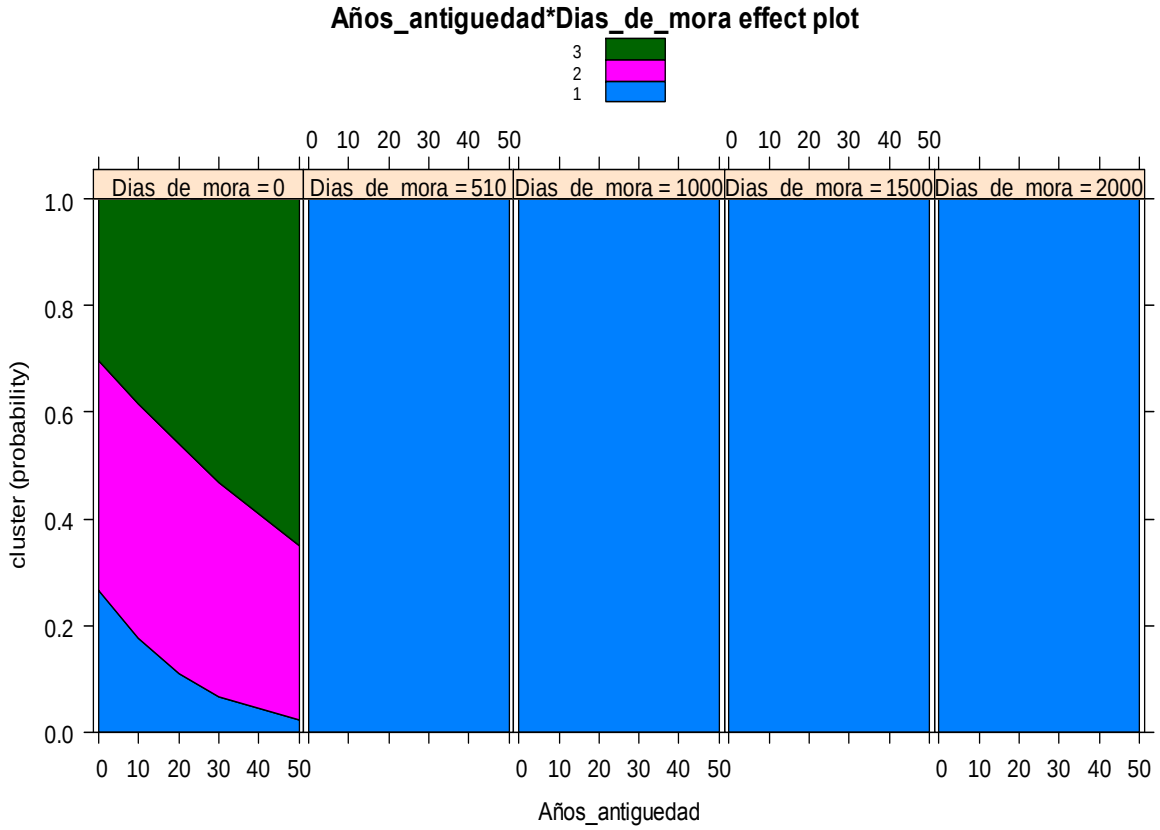
También se observa que, suponiendo las demás variables constantes, por cada año de más en la antigüedad en la cooperativa, es 4,55% más probable de obtener un beneficio Básico y un 6,73% más probable que se obtenga el Beneficio Premium.

En el caso del plazo, por cada mes de más, aumenta en 2,23% probabilidad de obtener el beneficio Básico y en un 3,15% la probabilidad de obtener Beneficio Premium, en comparación con no tener Beneficio.

En el caso de los días en mora, al ser el coeficiente negativo, el OR es menor que 1, ya que exp(-0.1380112) =0.8710889. Por lo tanto, por cada día de mora, disminuye en un 12,9% (1-0.8710889) la probabilidad de obtener Beneficio Básico y en un 15,59% la probabilidad de tener Beneficio Premium.

3.4. 3 Visualización gráfica del fenómeno. En la figura 16 se observa que en clientes no morosos o con pocos días de mora, al aumentar los años de permanencia en la cooperativa la probabilidad de obtener el beneficio Premium aumenta considerablemente, disminuyendo de igual forma la probabilidad de no obtener beneficio. Los clientes con alta morosidad no se ven beneficiados a pesar de la antigüedad en la cooperativa.

Figura 22.
*Probabilidad de asignación a un clúster según años de antigüedad * días de mora*



3.4.4 Planteamiento de las ecuaciones para cada clúster. Con el modelo logit multinomial, se brinda una nueva alternativa que nos permita identificar a los asociados por su comportamiento y fidelización que se les puede otorgar el beneficio de tasa de interés.

Para analizar las elecciones de un asociado con buen habito de pago, considerando las variables del modelo final, tendremos la probabilidad del beneficio del asociado que se analice, el cual lo ubicará al grupo 2 o al grupo 3. A través de este puntaje se busca identificar el perfil del asociado y minimizar la operatividad que se puede estar presentando al interior de la Cooperativa, frente a este proceso.

Así mismo, el modelo proyecta un coeficiente para las variables independientes "Ingreso", "Días de mora", "Años de antigüedad" y "plazo", y que resultaron ser significativas para la aplicación.

3.4.4.1 Clúster 2

$$Ln\left[\frac{p\,(x)}{1-p\,(x)}\right] = -1{,}47 + 8{,}33\,X\,Ingreso - 0{,}14\,X\,Dias\,de\,mora + 0{,}05\,X\,Años\,de\,Antiguedad + 0{,}02\,X\,Plazo$$

$$p(x) = \frac{1}{1 + e - (-1{,}47 + 8{,}33\,X\,Ingreso - 0{,}14\,X\,Dias\,de\,mora + 0{,}05\,X\,Años\,de\,Antiguedad + 0{,}02\,X\,Plazo)}$$

3.4.4.2 Cluster 3

$$Ln\left[\frac{p\,(x)}{1-p\,(x)}\right] = -2{,}80 + 1{,}35\,X\,Ingreso - 0{,}17\,X\,Dias\,de\,mora + 0{,}07\,X\,Años\,de\,Antiguedd + 0{,}03\,X\,Plazo$$

$$p(x) = \frac{1}{1 + e - (-2{,}80 + 1{,}35\,X\,Ingreso - 0{,}17\,X\,Dias\,de\,mora + 0{,}07\,X\,Años\,de\,Antiguedd + 0{,}03\,X\,Plazo)}$$

La probabilidad de pertenecer al clúster 1 será: $p(i \in clúster\ 1) = 1 - [p(i \in clúster\ 2) + p(i \in clúster\ 3)]$. El modelo le permitirá a la cooperativa determinar si un asociado puede o no obtener beneficio adicional en su tasa de interés, y si es así, de cuánto (Básico o Premium) a partir del perfil del asociado, es decir, conocidos los años de antigüedad, sus ingresos y los días de mora y plazo de su crédito, determinar la probabilidad de pertenencia a un clúster y clasificarlo en aquella cuya probabilidad sea mayor.

3.4.5 Bondad de Ajuste del modelo final. La bondad de ajuste del modelo se puede evaluar mediante la tabla de clasificación (que tan bien logra discriminar el modelo un grupo de otro) y mediante los seudo R cuadráticos.

Tabla 10. *Tabla de buena clasificación*

Clúster Predicho	No.	Cluster Observado		
		1	2	3
	1	9.466	1989	124
	2	4.024	3351	347
	3	457	857	406

En cuanto a la tabla de clasificación (tabla 10) tenemos un porcentaje global de buena clasificación del 62,9%. Al revisar la clasificación correcta por grupos tenemos que el modelo clasifica correctamente en un 67.87% (9466/13947) a la población de No recibir beneficio (clúster1), clasifica correctamente en un 54,07% (3351/6197) a la población clasificada en el Beneficio Básico (clúster2) y en un 46,29% (406/877) a quienes están clasificados en Premium.

Los valores de Seudo R cuadrados del modelo encontrados son: el de McFadden 19,06%, el de Máxima Verosimilitud (ML) de 34,06% y el de Cragg y Uhler's (CU) de 38,37% , que son valores típicos en una modelación multinomial múltiple.

CONCLUSIONES Y RECOMENDACIONES

Al aplicar los modelos estadísticos planteados (Análisis univariado, Análisis de conglomerados – CLARA para la construcción de los Cluster, Análisis de regresión Logística Multinomial), permitió encontrar un modelo de Scoring para la aplicación de un beneficio en la tasa de interés.

La conformación inicial de conglomerados como método de exploración multivariante permitió la identificación de los asociados con Personería Jurídica como datos atípicos multivariantes (121 en total) y por lo tanto se hizo el análisis con asociados identificados como personas naturales.

El análisis clúster se realizó con el método CLARA (Clustering Large Aplications) por la alta dimensión de la base de datos, que trabaja con variables continuas o mixtas. Los conglomerados finales recogen el 56,2% de la variabilidad observada entre los individuos.

El primer clúster conformado por 18595 asociados presenta el menor número de activos, pasivos, valor desembolsado, saldo promedio de depósitos, menor tasa EA y el mayor apalancamiento. De igual forma es el grupo con el mayor promedio de días de mora (cerca de la cuarta parte del grupo presenta mora de más de 15 días). Este clúster fue catalogado como el grupo al cual no aplica beneficio en la tasa de interés. El segundo clúster conformado por 10328 asociados fue catalogado como el grupo con el beneficio Básico. El tercer clúster conformado por 4384 asociados presenta el mayor número de activos, pasivos, menor tasa EA, mayor desembolso.

De igual forma es el grupo con el menor promedio de días de mora. Este clúster fue catalogado como el grupo al cual se aplica beneficio Premium en la tasa de interés.

Para la asignación de los puntos de reducción en la tasa de interés, se efectuarán basados al estudio en el comité de tasas de la entidad.

Con esta información previa se realizó una modelación logit multinomial con variables respuesta el grupo o clúster y con variables explicativas las características de los asociados. Para evitar el sesgo que pueda generar el desbalance en el tamaño de los grupos (tercer grupo representa solo el 13,16% del total) y por lo tanto de la variable respuesta, se realizó un muestreo aleatorio del 80% del grupo tres (3507 individuos seleccionados), 40% del grupo dos (4131 individuos seleccionados) y 25% del grupo uno (4648 individuos seleccionados) conformando una base de datos de entrenamiento de 12286 asociados y dejando como base de prueba del modelo a un total de 21021 asociados, estos últimos no entraron en el muestreo toda vez que son utilizados para analizar la bondad de ajuste del modelo.

El modelo multinomial seleccionado fue el que arrojó el menor criterio AIC (21738,86) de todos los modelos corridos, teniendo en cuenta siempre el principio de parsimonia, con pocas variables que logren captar la variabilidad observada. Este modelo tuvo un Pseudo R^2 entre 19,06% y 38,37% y una correcta clasificación global del 66,35%. Las variables significativas en el modelo premian a los clientes más fieles (Años de antigüedad) y con mejor historial de pago (menor número de días de mora), así como garantiza a la cooperativa la retribución del crédito a partir de un mayor plazo en el pago del mismo y relacionado con el nivel de ingresos del mismo.

Finalmente, se recomienda a la cooperativa realizar jornadas de actualización de información (formulario único de clientes) para que los datos que fueron necesariamente depurados por las

inconsistencias encontradas se minimicen. De esta manera se realizaría un seguimiento continuo y con la calibración del modelo, esta proporción o cierta proporción de asociados (no incluidos en el análisis) posiblemente lograrían obtener algún beneficio en la tasa de intereses.

Las características observadas en los asociados pueden ayudar a la cooperativa a determinar el perfilamiento de los asociados; de este modo el área comercial implementaría campañas definidas para ciertos nichos de mercado.

REFERENCIAS BIBLIOGRÁFICAS

Arango, L. y Restrepo, D. (2017). *Diseño de un modelo de scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento Colombiana. Escuela de Economía y Finanzas*. Medellín: Universidad Eafit, Recuperado de https://repository.eafit.edu.co/bitstream/handle/10784/12434/Laura_ArangoDuque_Daniel_RestrepoBaena_2017.pdf?sequence=2.

Banco de la República. (2019). *Informe de Tasas de interés de política monetaria*. Recuperado de https://www.banrep.gov.co/es/estadisticas/tasas-interes-politica-monetaria

Barreiro, J., Ruzo, E., y Lozada, F. (2014). *"Modelo logit multinomial y regresion con variables ficticias: una aplicacion regional al sector lacteo,"* Economic Development 81, University of Santiago de Compostela. Santiago de Chile, Chile.

Boletín Jurídico. (2009) *Superintendencia Financiera de Colombia. Relación de Conceptos*. Recuperado de https://www.superfinanciera.gov.co/jsp/Publicaciones/publicaciones/loadContenidoPublicacion/id/16026/dPrint/1/c/00

Carranza, I. J & González, E. (2019). *Informe semanal de tasas de interés activas o de colocación*: *Manual de usuario de tasas de colocación del Banco de la Republica.* Recuperado de https://www.banrep.gov.co/sites/default/files/manual-usuario-tasas-de-colocacion.pdf

Dabós, M. (2015). *Credit Scoring.* Belgrano. Recuperado de ttps://mba.americaeconomia.com/sites/mba.americaeconomia.com/files/credit_scoring.pdf

Delgado, D. (2015). *Segmentación de clientes mediante análisis Conglomerados* (tesis de postgrado). Universidad Industrial de Santander, Bucaramanga, Colombia.

Dueñas, R. (2008). *Introducción al sistema financiero y Bancario*. Bogotá. Recuperado de https://crear.poligran.edu.co/publ/00008/SFB.pdf

Fernández, S. (2011). *Análisis conglomerados*. Madrid. Recuperado de 2020 http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/CONGLOMERADOS/conglomerados.pdf

Pando, V., y Fernández, V. (2004). *Regresión Logística Multinomial*. Madrid. Recuperado de https://www.academia.edu/18911198/Regresi%C3%B3n_log%C3%ADstica_multinomial

Hand, D.J. and Henley, W.E. (1997) *Statistical Classification Methods in Consumer. Credit Scoring:* A Review. Journal of Royal Statistical Society, 160, 523-541.

Hernández, N., Jaimes, G. y Mosquera, J. (2015), *Caracterización sectorial y laboral de las cooperativas en el municipio de Pamplona*. Pamplona. Recuperado de http://revistas.unipamplona.edu.co/ojs_viceinves/index.php/FACE/article/view/2650/0

Kaufman, L. y Rousseeuw, P. (2005). *Finding Groups in Data an introduction to Clúster Ananlysis*. Belgium. Recuperado de https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

López, P., y Fachelli, S. (2015). *Análisis de regresión logística. Metodología de la Investigación Social Cuantitativa.* Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. Capítulo III.10. 1ª edición. Recuperado de https://www.fundacionmapfre.org/documentacion/publico/i18n/catalogo_imagenes/imagen_id.cmd?idImagen=1102733

Ochoa, C. (2015). *Muestreo probabilístico, muestreo por conglomerados*. Chile. Recuperado https://www.netquest.com/blog/es/blog/es/muestreo-probabilistico-muestreo-conglomerados

Peña, D. (2002). *Análisis de Datos Multivariantes.* Cuadras. Recuperado de https://www.academia.edu/6134000/An%C3%A1lisis_de_Datos_Multivariantes_-_Daniel_Pe%C3%B1a

Posada, C & Misas, M. *La tasa de interés en Colombia.* Banco de la Republica. Recuperado de https://www.banrep.gov.co/es/tasa-interes-colombia

Rodríguez, D.A (2017). *Cómo funcionan las tasas de interés del Banco Central. Economía tu dinero*. Recuperado de https://latinamericanpost.com/es/17340-como-funcionan-las-tasas-de-interes-del-banco-central

Tukey, J. (1977). *Exploratory Data Analysis.* New York. Recuperado de libro Exploratory Data Analysis Capitulo 4

Vásconez, G. (2012). *Scoring, una herramienta para la evaluación de crédito*. Alemania. Recuperado https://es.slideshare.net/gustavovasconez/scoring-una-herramienta-para-evaluacin-de-microcrdito?next_slideshow=2

APENDICES

APENDICE A

ALGUNOS MODELOS MULTINOMIAL AJUSTADOS

Inicialmente se presentan algunos modelos univariados, ya que, por principio de parsimonia, entre menos variables se requieran para una modelación con buen ajuste, mejor. Se revisó entonces que tan bien el modelo discrimina a los asociados según el clúster al que pertenece de acuerdo a alguna de sus características.

MODELO: cluster ~ Apalancamiento. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Apalancamiento, data = entrenamiento)

Coefficients:
  (Intercept) Apalancamiento
2   0.4559315      -2.555423
3   0.2998645      -2.598572

Std. Errors:
  (Intercept) Apalancamiento
2  0.03304041      0.1161172
3  0.03441616      0.1231626

Residual Deviance: 26079.2
AIC: 26087.2
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	7955	2367	342
2	5992	3830	535
3	0	0	0

Clasificación correcta global= 43.94%

El modelo no es capaz de discriminar a los asociados del grupo 3. Ninguno tendría beneficio Premium.

MODELO: cluster ~ Valor de desembolso. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ VALOR_DESEMBOLSO, data = entrenamiento)

Coefficients:
  (Intercept) VALOR_DESEMBOLSO
2  -0.8042279     5.215444e-08
3  -2.2462374     1.055948e-07

Std. Errors:
   (Intercept) VALOR_DESEMBOLSO
2 7.752966e-17     1.357485e-09
3 5.516350e-17     1.362022e-09

Residual Deviance: 22960.51
AIC: 22968.51
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	11451	3765	206
2	1641	1340	177
3	855	1092	494

Clasificación correcta global= 36.8%

El modelo mejora la capacidad de discriminación, en particular en el grupo 3. La mayoría de los asociados del grupo 2 son catalogados erróneamente en el grupo 1.

MODELO: cluster ~ Ingresos. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Ingresos, data = entrenamiento)

Coefficients:
  (Intercept)     Ingresos
2   -0.426710 5.127117e-08
3   -1.156144 9.896939e-08

Std. Errors:
   (Intercept)     Ingresos
2 2.062252e-16 2.441002e-09
3 1.549613e-16 2.369674e-09

Residual Deviance: 24692.92
AIC: 24700.92
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	12179	4802	363
2	915	701	147
3	853	694	367

Clasificación correcta global= 36.98%

El modelo mejora la capacidad de discriminación, en particular en el grupo 1. La mayoría de los asociados del grupo 2 son catalogados erróneamente en el grupo 1.

MODELO: cluster ~ Días de mora. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Dias_de_mora, data = entrenamiento)

Coefficients:
  (Intercept) Dias_de_mora
2  0.17616883   -0.1133052
3  0.01279151   -0.1142928

Std. Errors:
  (Intercept) Dias_de_mora
2   0.0233515  0.005883100
3   0.0242859  0.006364402

Residual Deviance: 24962.02
AIC: 24970.02
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	3682	373	43
2	10265	5824	834
3	0	0	0

Clasificación correcta global= 45.22%

El modelo mejora la capacidad de discriminación del grupo 2. Pero discrimina muy mal al grupo 1 (clasificándolos erróneamente en el grupo 2) y no es capaz de discriminar a los asociados del grupo 3 (clasificados en su mayoría en el grupo 2).

MODELO: cluster ~ Años de antigüedad. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Años_antiguedad, data = entrenamiento)

Coefficients:
  (Intercept) Años_antiguedad
2  -0.3234999      0.04470436
3  -0.6184036      0.06624928

Std. Errors:
  (Intercept) Años_antiguedad
2  0.02722987     0.003711450
3  0.02897180     0.003671786

Residual Deviance: 26459.45
AIC: 26467.45
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	11819	4663	597
2	1540	892	143
3	588	642	137

Clasificación correcta global= 38.88%

El modelo mejora la capacidad de discriminación del grupo 1. Pero discrimina muy mal al grupo 2 y al grupo 3 (clasificándolos erróneamente en el grupo 1)

Estos modelos univariados se realizaron con todas y cada una de las variables explicativas consideradas en este trabajo. A continuación se presentan algunos modelos de los modelos

multivariados planteados y analizados. Se presentan sólo aquellos en los que todas las variables explicativas fueron significativas.

MODELO: cluster ~ Apalancamiento + Años de antigüedad. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Apalancamiento + Años_antiguedad,
    data = entrenamiento)

Coefficients:
  (Intercept) Apalancamiento Años_antiguedad
2  0.25455798      -2.513255       0.0418596
3 -0.04112619      -2.507057       0.0633510

Std. Errors:
  (Intercept) Apalancamiento Años_antiguedad
2  0.03753502      0.1165448     0.003795065
3  0.03953580      0.1242094     0.003754628

Residual Deviance: 25756.83
AIC: 25768.83
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	8527	2491	340
2	4915	3099	407
3	505	607	130

Clasificación correcta global= 44.07%

El modelo mejora la capacidad de discriminación del grupo 2. Pero discrimina mal al grupo 1 y al grupo 3. En el caso del grupo 3 sólo el 14,82% de los asociados clasificados en el grupo 3 son correctamente clasificados

MODELO: cluster ~Ingresos + Años de antigüedad. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Ingresos + Años_antiguedad, data = entrenamiento)

Coefficients:
  (Intercept)     Ingresos Años_antiguedad
2  -0.5874791 4.787161e-08      0.03929712
3  -1.4440942 9.545141e-08      0.06004359

Std. Errors:
   (Intercept)     Ingresos Años_antiguedad
2 2.005889e-16 2.423462e-09    9.390559e-16
3 1.479571e-16 2.349870e-09    7.604871e-16

Residual Deviance: 24415.7
AIC: 24427.7
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	11562	4195	303
2	1500	1213	185
3	885	789	389

Clasificación correcta global= 37.77%

El modelo mejora la capacidad de discriminación del grupo 1. Pero discrimina mal al grupo 2. En el caso del grupo 2 sólo el 19.58% de los asociados clasificados en el grupo 2 son correctamente clasificados.

MODELO: cluster ~ Apalancamiento + Años de antigüedad + Días de mora. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Apalancamiento + Años_antiguedad +
    Dias_de_mora, data = entrenamiento)

Coefficients:
  (Intercept) Apalancamiento Años_antiguedad Dias_de_mora
2   0.5071727      -2.531104       0.05565098   -0.1184777
3   0.2153221      -2.526411       0.07685553   -0.1219144

Std. Errors:
  (Intercept) Apalancamiento Años_antiguedad Dias_de_mora
2  0.04009276      0.1212834     0.004275656  0.006030067
3  0.04200935      0.1288673     0.004259667  0.006593861

Residual Deviance: 23893.96
AIC: 23909.96
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	7964	1665	215
2	5608	3917	529
3	375	615	133

Clasificación correcta global= 42.84%

El modelo mejora la capacidad de discriminación del grupo 2. Pero discrimina mal al grupo 1 y al grupo 3.

MODELO: cluster ~ Apalancamiento + Años de antigüedad + Plazo.

Datos: entrenamiento

```
multinom(formula = cluster ~ Apalancamiento + Años_antiguedad +
    PLAZO, data = entrenamiento)

Coefficients:
  (Intercept) Apalancamiento Años_antiguedad      PLAZO
2  -0.5533442      -2.882964      0.03485476 0.02010231
3  -1.0556502      -2.973130      0.05569661 0.02472109

Std. Errors:
  (Intercept) Apalancamiento Años_antiguedad       PLAZO
2  0.06337277      0.1213174     0.003777727 0.001276179
3  0.06815119      0.1301776     0.003742052 0.001340151

Residual Deviance: 25344.02
AIC: 25360.02
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	9436	2493	337
2	3853	2968	383
3	658	736	157

Clasificación correcta global= 40.24%

El modelo mejora la capacidad de discriminación de los grupos 1 y 2. Pero discrimina mal al grupo 3 (sólo el 17,9% de los asociados del grupo 3 están correctamente clasificados)

MODELO: cluster ~ Ingresos + Años de antigüedad + Días de mora. Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Ingresos + Dias_de_mora + Años_antiguedad,
    data = entrenamiento)

Coefficients:
  (Intercept)     Ingresos Dias_de_mora Años_antiguedad
2  -0.4725355 7.891090e-08   -0.1273457      0.05037367
3  -1.3514665 1.291832e-07   -0.1514254      0.07163839

Std. Errors:
   (Intercept)     Ingresos Dias_de_mora Años_antiguedad
2 3.129910e-16 3.024486e-09 5.472158e-16    1.346781e-15
3 2.278797e-16 3.006493e-09 4.745230e-16    1.125333e-15

Residual Deviance: 22187.84
AIC: 22203.84
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	10457	2800	169
2	3012	2601	319
3	478	796	389

Clasificación correcta global= 61.96%

Este modelo discrimina muy bien al grupo 1, pero no tanto al grupo 2.

MODELO: cluster ~ Apalancamiento + Años de antigüedad + Valor de desembolso

Datos: entrenamiento

```
Call:
multinom(formula = cluster ~ Apalancamiento + Años_antiguedad +
    VALOR_DESEMBOLSO, data = entrenamiento)

Coefficients:
  (Intercept) Apalancamiento Años_antiguedad VALOR_DESEMBOLSO
2  -0.4070748      -3.022437      0.02965805     6.243888e-08
3  -1.7073433      -4.505807      0.04169964     1.180884e-07

Std. Errors:
   (Intercept) Apalancamiento Años_antiguedad VALOR_DESEMBOLSO
2 8.255506e-17   2.106713e-17    3.743481e-16     1.422861e-09
3 5.822555e-17   1.335166e-17    3.029469e-16     1.449813e-09

Residual Deviance: 21673.97
AIC: 21689.97
```

Bondad de ajuste del modelo ¿qué tan bueno es el modelo discriminando entre los tres grupos? Datos: test

predicted_class	1	2	3
1	8266	2909	432
2	5239	2988	396
3	442	300	49

Clasificación correcta global= 53.77%

El modelo discrimina mal al grupo 3

APENDICE 2

Código en R

```
library(ade4)
library(MASS)
library (psych)
library(rgl)
library (ggplot2)
library("FactoMineR")
library("factoextra")
library (datasets)
library (lme4)
library (lattice)
library(cluster)
library(ordinal)


############################
LECTURA DE LOS DATOS
############################

datos=read.csv("C:/Users/Casa/Desktop/base 15 marzo wjo.csv", sep=";",header=TRUE, dec = ",")
names(datos)
dim(datos)

attach(datos)
datos$Años_antiguedad=datos$A.os_Antig.edad
datos$Estrato=factor(datos$Estrato)

datos2=datos[,-c(1,3) ]

str(datos2)
summary(datos2)

continuas=datos2[, c(1:15,19)]
cor(continuas)

#############################################
#EXPLORACIÓN CON ANÁLISIS  CLÚSTER
#############################################

#análisis con todas las variables continuas
```

```
clara.res2=clara(continuas, 3, metric = "euclidean", stand = FALSE,    samples = 50, pamLike = TRUE)

print(clara.res2)

dd2 = cbind(datos2, cluster = clara.res2$cluster)

write.table(dd2, file="resultados cluster con todas las variables.txt")

head(dd2, n = 4)

clara.res$medoids

head(clara.res2$clustering, 10)

fviz_cluster(clara.res2, stand = T, geom = "point",   pointsize = 1)


clara.res2$clusinfo

fviz_cluster(object = clara.res2, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +
  theme_bw() +
  labs(title = "Resultados clustering CLARA") +
  theme(legend.position = "none")


#con otras variables sin escalar

continuas2=datos2[, c(2:6, 9:13)]

clara.res3=clara(continuas2, 3, metric = "euclidean", stand = FALSE,       samples = 50, pamLike = TRUE)
print(clara.res3)

dd3 = cbind(datos2, cluster = clara.res3$cluster)

write.table(dd3, file="resultados cluster sin y menos variables.txt")

head(dd2, n = 4)

clara.res3$medoids
clara.res3$clusinfo

fviz_cluster(clara.res3, stand = T, geom = "point",    pointsize = 1)
```

```
fviz_cluster(object = clara.res3, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +
  theme_bw() +
  labs(title = "Resultados clustering CLARA") +
  theme(legend.position = "none")

#retirando los datos de personería jurídica

datos3=datos2[ which(Genero=='Femenino'),]
datos4=datos2[ which(Genero=='Masculino'),]
datos5=rbind(datos3,datos4)

summary(datos5)

attach(datos5)

continuas=datos2[, c(1:15,19)]


#####################################################
#ANÁLISIS  CLÚSTER SOLO CON PERSONAS NATURALES
#####################################################

#análisis con todas las variables continuas

clara.res2=clara(continuas, 3, metric = "euclidean", stand = FALSE,
     samples = 50, pamLike = TRUE)
print(clara.res2)

dd2 = cbind(datos5, cluster = clara.res2$cluster)

write.table(dd2, file="resultados cluster con todas las variables.txt")

head(dd2, n = 4)

clara.res$medoids

head(clara.res2$clustering, 10)

fviz_cluster(clara.res2, stand = T, geom = "point",       pointsize = 1)

clara.res2$clusinfo

fviz_cluster(object = clara.res2, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +
  theme_bw() +
```

```
  labs(title = "Resultados clustering CLARA") +
  theme(legend.position = "none")
#con otras variables sin escalar
continuas2=datos5[, c(2:6, 9:13)]

clara.res3=clara(continuas2, 3, metric = "euclidean", stand = FALSE,
     samples = 50, pamLike = TRUE)
print(clara.res3)

dd3 = cbind(datos5, cluster = clara.res3$cluster)

write.table(dd3, file="resultados cluster sin y menos variables.txt")

head(dd2, n = 4)

clara.res3$medoids
clara.res3$clusinfo

fviz_cluster(clara.res3, stand = T, geom = "point",   pointsize = 1)


fviz_cluster(object = clara.res3, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +
  theme_bw() +
  labs(title = "Resultados clustering CLARA") +
  theme(legend.position = "none")

#con nuevas variables sin escalar
continuas5=datos5[, c(2,3,5,6, 9:11,13)]

clara.res5=clara(continuas5, 3, metric = "euclidean", stand = FALSE,
     samples = 50, pamLike = TRUE)
print(clara.res5)

dd5 = cbind(datos5, cluster = clara.res5$cluster)

write.table(dd5, file="resultados cluster sin y mucho menos variables.txt")

clara.res5$medoids
clara.res5$clusinfo

fviz_cluster(clara.res5, stand = T, geom = "point",   pointsize = 1)


fviz_cluster(object = clara.res3, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +
```

```
  theme_bw() +
  labs(title = "Resultados clustering CLARA") +
  theme(legend.position = "none")
##############################################
#ANÁLISIS DE LOS CLÚSTER
##############################################
datosfinal2=read.table("C:/Users/Casa/Documents/resultados cluster sin y mucho menos variables.txt",header=TRUE)
names(datosfinal2)
dim(datosfinal2)
str(datosfinal2)


datosfinal2$cluster=factor(datosfinal2$cluster)
datosfinal2$Estrato=factor(datosfinal2$Estrato)

attach(datosfinal2)

summary(datosfinal2)

cluster1=datosfinal2[ which(cluster=='1'),]
dim(cluster1)

cluster2=datosfinal2[ which(cluster=='2'),]
dim(cluster2)

cluster3=datosfinal2[ which(cluster=='3'),]
dim(cluster3)

summary(cluster1)
summary(cluster2)
summary(cluster3)

write.table(datosfinal2, file="resultados cluster abril 25.txt")

datosfinal3=read.table("C:/Users/Casa/Documents/resultados cluster abril 25.txt",header=TRUE)
names(datosfinal3)
dim(datosfinal3)
attach(datosfinal3)
summary(datosfinal3)

str(datosfinal3)
datosfinal3$Estrato=factor(datosfinal3$Estrato)
datosfinal3$cluster=as.ordered(datosfinal3$cluster)

str(datosfinal3)
```

```
#gráficas exploratorias y descriptivas de cada clúster
library(ggplot2)
boxplot(Apalancamiento~cluster)
library(ggplot2)
g1 = ggplot(datosfinal3, aes(cluster, Apalancamiento))
g1 + geom_jitter(width = 0.05) +
    stat_summary(fun.y = "mean", geom = "point", col = "red", shape = 15, size = 3)


g2=ggplot(datosfinal3, aes(x=cluster, y=Pasivos, fill=cluster))
 g2+ geom_boxplot()

g2=ggplot(datosfinal3, aes(x=cluster, y=Apalancamiento, fill=cluster))
 g2+ geom_boxplot()

g2=ggplot(datosfinal3, aes(x=cluster, y=Años_antiguedad, fill=cluster))
 g2+ geom_boxplot()

boxplot(Saldo_Promedio_cr.ditos~cluster)
g1 = ggplot(datosfinal3, aes(cluster, Saldo_Promedio_cr.ditos))
g1 + geom_jitter(width = 0.05) +
    stat_summary(fun.y = "mean", geom = "point", col = "red", shape = 15, size = 3)

boxplot(PLAZO~cluster)
g1 = ggplot(datosfinal3, aes(cluster,PLAZO))
g1 + geom_jitter(width = 0.05) +
    stat_summary(fun.y = "mean", geom = "point", col = "red", shape = 15, size = 3)

boxplot(Dias_de_mora~cluster)
g1 = ggplot(datosfinal3, aes(cluster, Dias_de_mora))
g1 + geom_jitter(width = 0.05) +
    stat_summary(fun.y = "mean", geom = "point", col = "red", shape = 15, size = 3)


p1 <- ggplot(datosfinal3, aes(x=cluster, y=Activos, fill=Genero)) +
   geom_boxplot() +     facet_wrap(~Genero)
p1

tabla2=table(Estrato,cluster)
prop.table(tabla2,2) # proporción columna

# Gráfico de barras
barplot(round(prop.table(tabla2,2)*100,1),legend=TRUE,  col=c(4:10),  main="Estrato  Cartera
según clúster")
```

```
ggplot(data = datosfinal3, aes(x = Años_antiguedad, fill=cluster)) +
geom_histogram(position = "identity")

ggplot(data = datosfinal3, aes(Apalancamiento, fill = cluster)) + geom_histogram(binwidth = 0.5)
+
  facet_wrap(~cluster, ncol = 1)

options(scipen=999)
levels(cluster) <- c("No_aplica", "Básico", "Plus")


##############################################
#MODELOS LOGIT ORDINAL
##############################################

library(MASS)
library(foreign)
library(effects)
library(ordinal)
library(pscl)
library(brant)

cluster1 = datosfinal3[ which(datosfinal3$cluster=='1'), ]
dim(cluster1)

cluster2 = datosfinal3 [ which(datosfinal3$cluster=='2'), ]
dim(cluster2)

cluster3 = datosfinal3 [ which(datosfinal3$cluster=='3'), ]
dim(cluster3)


trainingRows1 = sample(1:nrow(cluster1), 0.25*nrow(cluster1))
training1 = cluster1[trainingRows1, ]
test1 = cluster1[-trainingRows1, ]
dim(training1)
dim(test1)

trainingRows2 = sample(1:nrow(cluster2), 0.4*nrow(cluster2))
training2 = cluster2[trainingRows2, ]
test2 = cluster2[-trainingRows2, ]
dim(test2)
dim(training2)

trainingRows3 = sample(1:nrow(cluster3), 0.8*nrow(cluster3))
training3 = cluster3[trainingRows3, ]
```

```
test3 = cluster3[-trainingRows3, ]

dim(training3)

#base de entrenamiento
entrenamiento=rbind(training1,training2,training3)
dim(entrenamiento)

#base para validación del modelo
test=rbind(test1,test2,test3)
dim(test)

#modelos univariados
modelo1 = polr(cluster ~ Ingresos,entrenamiento)
summary(modelo1)

plot(Effect(focal.predictors = c("Ingresos"), modelo1), rug = FALSE,
    style="stacked")

library (VGAM)
modelo1 = vglm(cluster ~  Apalancamiento,
family = cumulative(link = "logit",  parallel = TRUE, reverse = TRUE))
summary (modelo1)


modelo2 = polr(cluster ~  Edad)
summary(modelo2)
plot(Effect(focal.predictors = c("Edad"), modelo2), rug = FALSE,
    style="stacked")

modelo2 = vglm(cluster ~ VALOR_DESEMBOLSO,
family = cumulative(link = "logit",  parallel = TRUE, reverse = TRUE))
summary (modelo2)

#otra opción
library(rms)
modelo13= lrm(cluster ~  Años_antiguedad,entrenamiento)
modelo13

#modelos multivariantes varios

modelo12 = vglm(cluster ~ Activos+Pasivos+Ingresos+Saldo_Depositos+
Apalancamiento+creditos_prepagados+ numero_creditos+VALOR_DESEMBOLSO+
Saldo_Promedio_creditos+ TASA_EA+Dias_de_mora+PLAZO+Cuotas_pagadas+
Cuotas_pendientes+Calidad_cartera+ Años_antiguedad,entrenamiento,
family = cumulative(parallel=FALSE~Dias_de_mora, reverse = TRUE))
```

```
summary (modelo12)

#test de rectas paralelas
brant(modelo12)


modelo13 = vglm(cluster ~ Ingresos+Saldo_Depositos+ Apalancamiento+ Años_antiguedad,entrenamiento,
family = cumulative(link = "logit",  parallel = TRUE, reverse = TRUE))

summary (modelo13)
#test de rectas paralelas
brant(modelo13)


modelo14= lrm(cluster ~ Ingresos+ Apalancamiento+ Años_antiguedad,entrenamiento)
modelo14

brant(modelo14)

#paso a paso: función step

fit0 = polr(cluster ~ 1 , data=datosfinal3, method="logistic")
fit.sup = polr(cluster ~ Ingresos+ Apalancamiento+ Años_antiguedad+TASA_EA+Dias_de_mora+PLAZO+VALOR_DESEMBOLSO+Saldo_Depositos,
data=entrenamiento, method="logistic")

fit.sup2 = polr(cluster ~ Ingresos+ Apalancamiento+ Años_antiguedad+Dias_de_mora+PLAZO+VALOR_DESEMBOLSO,
data=entrenamiento, method="logistic")


fit.1 <- stepAIC(fit0,scope = list(upper = fit.sup, lower = fit0))
fit.final <- stepAIC(fit0, scope = list(upper = fit.sup2, lower = fit0))

fit.final2=polr(cluster ~Ingresos, data=entrenamiento)
summary(fit.final2)

brant(fit.final2)


library(VGAM)

# Modelo con ventajas proporcionales
```

```
modelo.vglm <- vglm(cluster ~Apalancamiento, cumulative(parallel = TRUE), data=entrenamiento)
# Comprobamos si se cumple la asunción de ventajas proporcionales
#en general
modelo.vglm2 <- vglm(cluster ~Apalancamiento, cumulative(parallel = FALSE), data=datosfinal3)
pchisq(deviance(modelo.vglm) - deviance(modelo.vglm2), df = df.residual(modelo.vglm) - df.residual(modelo.vglm2), lower.tail = FALSE)

(vglmFit = vglm(cluster ~ Ingresos+ Apalancamiento+ Años_antiguedad+Dias_de_mora, family=propodds, data=entrenamiento))

(lrmFit <- lrm(cluster ~ Apalancamiento+ Años_antiguedad,  data=entrenamiento))

fit.final=polr(cluster ~ Apalancamiento+ Años_antiguedad, data=entrenamiento)
summary(fit.final)

brant(fit.final)

plot(Effect(focal.predictors = c("Apalancamiento","Años_antiguedad"), fit.final), rug = FALSE,
    style="stacked")

##############################################
#MODELOS LOGIT MULTINOMIAL
##############################################

library(nnet)
library(car)

#modelos univariados

multi1=multinom(cluster~Apalancamiento, data=entrenamiento)
summary(multi1)

multi2=multinom(cluster~Ingresos, data=entrenamiento)
summary(multi2)

multi3=multinom(cluster~Dias_de_mora, data=entrenamiento)
summary(multi3)

multi4=multinom(cluster~Años_antiguedad, data=entrenamiento)
summary(multi4)

multi5=multinom(cluster~Edad, data=entrenamiento)
summary(multi5)
```

```
multi6=multinom(cluster~Activos, data=entrenamiento)
summary(multi6)

multi7=multinom(cluster~Pasivos, data=entrenamiento)
summary(multi7)

multi8=multinom(cluster~Saldo_Depositos, data=entrenamiento)
summary(multi8)

multi9=multinom(cluster~creditos_prepagados, data=entrenamiento)
summary(multi9)

multi10=multinom(cluster~numero_creditos, data=entrenamiento)
summary(multi10)

multi11=multinom(cluster~VALOR_DESEMBOLSO, data=entrenamiento)
summary(multi11)

multi12=multinom(cluster~Saldo_Promedio_creditos, data=entrenamiento)
summary(multi12)

multi13=multinom(cluster~TASA_EA , data=entrenamiento)
summary(multi13)

multi14=multinom(cluster~PLAZO, data=entrenamiento)
summary(multi14)

multi15=multinom(cluster~Cuotas_pagadas, data=entrenamiento)
summary(multi15)

multi16=multinom(cluster~Cuotas_pendientes, data=entrenamiento)
summary(multi15)

#modelos multivariables

multinomModel1 <- multinom(cluster ~Ingresos+Dias_de_mora, data=entrenamiento) #
summary (multinomModel1) # model summary

#predicción

predicted_scores = predict (multinomModel1, test, "probs")

predicted_class = predict (multinomModel1, test)

table(predicted_class, test$cluster)
```

```
mean(as.character(predicted_class) != as.character(test$cluster))


multinomModel2 <- multinom(cluster ~Ingresos+Dias_de_mora + TASA_EA + PLAZO, data=entrenamiento) # multinom Model
summary (multinomModel2) # model summary

#predicción

predicted_scores = predict (multinomModel2, test, "probs")

predicted_class = predict (multinomModel2, test)

table(predicted_class, test$cluster)

mean(as.character(predicted_class) != as.character(test$cluster))


multinomModel3 <- multinom(cluster ~ Ingresos+ Apalancamiento+ Años_antiguedad+Dias_de_mora+PLAZO+VALOR_DESEMBOLSO, data=entrenamiento)
summary (multinomModel3) # model summary

#predicción

predicted_scores = predict (multinomModel3, test, "probs")

predicted_class = predict (multinomModel3, test)

table(predicted_class, test$cluster)

mean(as.character(predicted_class) != as.character(test$cluster))


multinomModel <- multinom(cluster ~Ingresos+Dias_de_mora + Años_antiguedad+ PLAZO, data=entrenamiento) # multinom Model
summary (multinomModel) # model summary

#predicción

predicted_scores = predict (multinomModel, test, "probs")

predicted_class = predict (multinomModel, test)

table(predicted_class, test$cluster)

mean(as.character(predicted_class) != as.character(test$cluster))
```

```
coef(multinomModel)

exp(coef(multinomModel ))

#seudo R2

library(pscl)
pR2(multinomModel)

Anova(multinomModel)
Anova Table (Type II tests)

#presencia de datos atípicos
influenceIndexPlot(multinomModel)

influencePlot(model)
summary(influence.measures(model =model))

plot(effect("Ingresos",multinomModel))
plot(effect("Dias_de_mora","Ingresos","Años_antiguedad",multinomModel))
plot(effect("bse",multinomModel))
plot(effect("Años_antiguedad",multinomModel),style="stacked")

plot(Effect(focal.predictors = c("Años_antiguedad","Dias_de_mora","PLAZO"),
multinomModel), rug = FALSE)

plot(Effect(focal.predictors = c("Ingresos","Dias_de_mora"), multinomModel), rug = FALSE,
    style="stacked")

plot(Effect(focal.predictors = c("Años_antiguedad","Ingresos"), multinomModel), rug = FALSE,
    style="stacked")

plot(Effect(focal.predictors = c("Años_antiguedad","Dias_de_mora"), multinomModel), rug =
FALSE,     style="stacked")

plot(Effect(focal.predictors = c("Dias_de_mora","PLAZO"), multinomModel), rug = FALSE,
    style="stacked")

plot(Effect(focal.predictors = c("Dias_de_mora","Ingresos"), multinomModel),multiline=T, rug =
FALSE)


plot(Effect(focal.predictors = c("Apalancamiento","Años_antiguedad"), multinomModel), rug =
FALSE,     style="stacked")
```

```
plot(Effect(focal.predictors = c("Apalancamiento","PLAZO"), multinomModel),multiline=T, rug
= FALSE,      style="stacked")
```