Acta de Correcciones al Documento de Trabajo de Grado

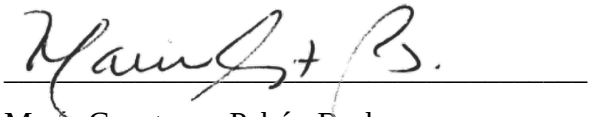
Santiago de Cali, 29 de Julio de 2024

Autor:
David Padilla López
Marisol Correa Diaz
Fabian Caballero Cortes

Título del Trabajo de Grado:
"OPTIMIZACIÓN DE LOS CANALES DIGITALES DEL BANCO COOPERATIVO BANCOOMEVA MEDIANTE MODELOS DE CLUSTERIZACIÓN"

Director: María Constanza Pabón Burbano

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



María Constanza Pabón Burbano
Director del Trabajo de Grado

Santiago de Cali, 28 de junio de 2024

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado “Optimización de los canales digitales del banco cooperativo mediante modelos de clusterización”, el cual será realizado por el (la) estudiante David Padilla López con código 8980375, Marisol Correa Diaz con código 8980182 y Fabian Caballero Cortes con código 8980342, perteneciente al énfasis en Sistemas y Computación, bajo la dirección del profesor María Constanza Pabón.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Firma
Marisol Correa Diaz
C.C. 1.070.983.511 de Facatativá

Firma
María Constanza Pabón
C.C. 34.559.22 de Popayán

Firma
David Padilla López
C.C. 1.114.060.418 de San Pedro

Firma
Fabian Caballero Cortes
C.C. 93.389.721 de Ibagué

ACUERDO DE CONFIDENCIALIDAD	
1. PARTE REVELADORA	**BANCO COOMEVA S.A.** con NIT. 900.406.150-5, representado en este acto por **Sandra Maria Orozco Agudelo**, por mayor de edad y vecino de Santiago de Cali, identificado con la cédula de ciudadanía No. 66.761.423 de Palmira (Valle), obrando en su calidad de representante legal suplente tal como consta en el certificado de existencia y representación legal expedido por la Superintendencia Financiera de Colombia
	Dirección: Calle 13 No. 57-50, teléfono: 333 0000 de Cali.
	Correo electrónico: sandram_orozco@coomeva.com.co
2. SUPERVISOR DEL CONTRATO	Claudia Johanna Polo Yepes – Jefa Analítica Bancoomeva
2. LOS RECEPTORES	**David Padilla López**, mayor de edad, identificado con cédula de ciudadanía No. 1.114.060.418 de San Pedro**, Marisol Correa Díaz**, identificada con cédula de ciudadanía No. 1.070.983.511 de Facatativá y **Fabian Rolando Caballero Cortes**, identificado con cédula de ciudadanía No. 93.389.721 de Ibagué, obrando en su propio nombre y representación, en calidad de estudiante del programa académico de Maestría En Ciencia De Datos de la Pontificia Universidad de Cali y **María Constanza Pabón** mayor de edad, identificada con cédula de ciudadanía No. 34.559.226 de Popayán obrando en su propio nombre y representación en calidad de profesor asociado del programa académico de Maestría En Ciencia De Datos de la Pontificia Universidad Javeriana Cali
	Correo electrónico: 1. daveland90@javerianacali.edu.co 2. mcorrea498@javerianacali.edu.co 3. faroca20@javerianacali.edu.co 4. mcpabon@javerianacali.edu.co
3. INFORMACIÓN A COMPARTIR	Información confidencial relacionada con la datos básicos del cliente, sociodemográfica, económica, tenencia de productos, transaccionalidad, canales de transacciones utilizados y segmentación . Toda esa información deberá estar anonimizada.

Las partes relacionadas en el encabezado anterior, en adelante conjuntamente, dependiendo de si otorgan o reciben la información confidencial se denominarán "**La Parte Reveladora**" y "**La Parte Receptora**", (en forma colectiva denominadas las "Partes"), hemos convenido celebrar el presente Acuerdo de Confidencialidad (en adelante el "Acuerdo") el cual tiene como finalidad establecer los términos y condiciones que rigen el uso y la protección de la información que sea suministrada por **La Parte Reveladora** en relación con el objeto desarrollado por **La Parte Receptora**, y con el fin de guardar la confidencialidad respecto de su contenido y sus relaciones negociales. El Acuerdo se regirá

por las estipulaciones siguientes, previas las presentes:

CONSIDERACIONES:

1. Que **BANCOOMEVA** se encuentra comprometida con la educación y la formación de sus colaboradores, quienes participan activamente en un semillero de analítica de datos, que, entre varios objetivos, busca, apoyar el desarrollo de proyectos de grado que tienen como fin generarle un beneficio al banco y culminar su proceso de grado. En este contexto, con el propósito de colaborar con la iniciativa académica, el Banco ha decidido facilitar el acceso y compartir información relevante para el desarrollo de los proyectos bajo los términos y condiciones establecidos en el presente Acuerdo de Confidencialidad.
2. Que **LOS RECEPTORES** son: tres (3) estudiantes del programa académico de posgrado en Maestría En Ciencia De Datos de la Pontificia Universidad Javeriana que como parte de las actividades académicas de la asignatura **Proyecto Aplicado**, se encuentran interesados en desarrollar un proyecto de investigación cuyos insumos de información y potencial utilidad práctica está relacionada con la actividad cotidiana y el objeto de negocio de **BANCOOMEVA,** en adelante "el Proyecto", y el Director del Proyecto de grado. El Proyecto en mención consiste en generar un modelo de aprendizaje no supervisado para la optimización de los canales digitales mediante modelos de clusterización y se desarrolla con el fin de poner en práctica los conceptos aprendidos en su proceso de formación y a su vez generar para **BANCOOMEVA** resultados y metodologías analíticas que le permitan a la entidad mejorar el desempeño del portafolio y/o segmento al cual se aplicará el desarrollo, de acuerdo con la investigación que se realice. En virtud de lo anterior, como actividades específicas, los estudiantes realizarán:

   a. Definición del problema de negocio y métricas asociadas. problema de negocio a resolver, objetivos, justificación, criterios de éxito, plan del proyecto, propuesta analítica inicial para resolver el problema.
   b. Ingeniería de datos. Generar la muestra de datos para el diseño y el test de las soluciones analíticas desarrolladas. Debe incluir como mínimo: Condiciones usadas para seleccionar la muestra de datos, fuentes usadas y su descripción general, periodos de tiempo analizados, exclusiones realizadas a la data junto con la cantidad de registros, tamaño final de la muestra generada y categorías de campos incluidas.
   c. Analítica descriptiva. comprensión a nivel exploratorio y descriptivo de los datos usados y su relación con el problema a resolver
   d. Modelado. diseño y desarrollo de los modelos analíticos seleccionados para resolver el problema. Se incluyen condiciones de la muestra de entrenamiento y testeo, validación cruzada si aplica, resultados de las pruebas estadísticas de acuerdo a la metodología implementada.
   e. Evaluación. resultados de las distintas pruebas realizadas a los modelos por fuera de la(s) muestra(s) usadas para entrenamiento y testing. Se incluyen en general los resultados del índice de estabilidad y las demás pruebas estadísticas que se hayan calculado en el entrenamiento y testing.

f. Implementación. Condiciones generales para hacer uso de las metodologías diseñadas, restricciones, alcance y principales recomendaciones.

3. Que el objetivo del Acuerdo, es que **La Parte Receptora** se comprometan a guardar la máxima reserva y secreto sobre la información clasificada como confidencial, que le sea entregada por parte de **Bancoomeva.**

4. Que es intención de **Bancoomeva**, que la información que suministre a **La Parte Receptora** sea utilizada por ellos o por su(s) director(es)/supervisor(es) de trabajo de grado, para los fines exclusivos establecidos en el Proyecto, sobre la base de que a los mismos se les extienda la obligación de confidencialidad contenida en este acuerdo y, por tanto, que dicha información sea manejada exclusivamente por ambas partes y no se divulgue a terceros, salvo bajo las condiciones de excepción incluidas en el siguiente numeral.

5. **Bancoomeva** reconoce que, como fruto del Proyecto entre las partes, existe la posibilidad de que los resultados obtenidos sean compartidos con la comunidad académica a través de ponencias y presentaciones en eventos de divulgación, o también publicados en revistas científicas especializados. Para tal fin, el acuerdo permitirá la realización de las actividades de divulgación y/o publicación científica, bajo garantía por parte los **Receptores** de mantener en reserva el nombre de **Bancoomeva** y garantizar la confidencialidad de la información recibida y transmitida en los resultados.

6. El presente Acuerdo propende establecer el manejo de la información entregada por **Bancoomeva** a **La Parte Receptora** relacionada con el generar un modelo de aprendizaje no supervisado para la optimización de los canales digitales mediante modelos de clusterización y demás información que permita diseñar la metodología.

El presente Acuerdo de Confidencialidad se regirá por las siguientes:

CLÁUSULAS:

PRIMERA. DEFINICIONES: Para efectos del presente Acuerdo, las palabras o términos que a continuación se relacionan, tendrán el significado que aquí se establece:

1.1. PARTE REVELADORA: La Parte Reveladora en este Acuerdo es Banco Coomeva S.A. quien suministrará Información Confidencial a la Parte Receptora.

1.2. PARTE RECEPTORA: Es Parte Receptora en este Acuerdo cada uno de los estudiantes que suscribe el presente Acuerdo, así como el(los) director(es)/supervisor(es) del trabajo de grado que también lo suscriben, quienes recibirán Información Confidencial de parte de Banco Coomeva S.A.

SEGUNDA. OBJETO DEL ACUERDO: El Objeto del presente Acuerdo es establecer entre las partes los términos y condiciones sobre el uso y protección de la información considerada como **CONFIDENCIAL** suministrada de parte de Banco Coomeva S.A. a **La Parte Receptora**

TERCERA. INFORMACIÓN CONFIDENCIAL: Para efectos del presente acuerdo, se considera Información Confidencial toda información verbal y/o escrita o en medios magnéticos que haya sido suministrada o que se llegue a suministrar de Banco Coomeva S.A. a **La Parte Receptora** y/o aquella a la cual tuvieron o pudieran tener acceso Las Partes. Tal información comprende:

1. Cualquier información acerca de productos, nuevas tecnologías, modelos y planes de negocios, información técnica, financiera, comercial, de mercado, estratégica, legal y cualquiera otra relacionada con las operaciones de negocios presentes y futuros de las Partes, de sus socios o accionistas, incluyendo, sus matrices, filiales, subsidiarias, afiliadas o sucursales, que haya sido entregada o comunicada a **La Parte Receptora** en virtud del Proyecto, así como documentos o archivos que se hubieran generado a partir de la información anteriormente mencionada o que la refleje.

2. Cualquier información sobre procesos internos, procedimientos, proyectos, esbozos, planes de desarrollo, mapas estratégicos, fotografías, plantas, diseños, bases de datos, directorios, tablas, conceptos de producto, especificaciones, muestras (incluso de equipos y herramientas), informativos, nombres de clientes, vendedores y/o distribuidores, información de precios, definiciones de mercado, componentes de propiedad intelectual, secretos empresariales, know-how, invenciones e ideas o, en cualquier otra forma y en cualquier otro medio, accesos remotos y/o físicos a máquinas, servicios, contraseñas, usuarios, y toda información que se suministre o divulgue en virtud del presente Acuerdo (ya sea antes o después de la fecha de suscripción del mismo) por **La Parte Reveladora** y/o cualquiera de sus respectivos directores, funcionarios, empleados o representantes a **La Parte Receptora**.

3. Todos los documentos que contengan, reflejen o reproduzcan total o parcialmente Información Confidencial (independientemente de quien los haya suministrado), así como los análisis, recopilaciones, datos, estudios, memorandos, informes y documentos, en cualquier forma y en cualquier medio que se deriven de, o se relacionen con la información que trata el párrafo anterior o que contengan o se basen en todo o en parte, en dicha información; al igual que cualquier idea, concepto, know-how, conocimiento o técnica relacionada con las actividades propias del negocio de **La Parte Reveladora** contenidas en la información que trata el mencionado párrafo, y toda la información que permanezca en la memoria de **La Parte Receptora** que han tenido acceso a la Información Confidencial bajo el presente Acuerdo

4. Cualquier información suministrada a **La Parte Receptora** por **La Parte Reveladora**, previa a la celebración del presente Acuerdo.

5. La información que se desarrolle producto de la ejecución del Proyecto.

Con ocasión del presente Acuerdo, **La Parte Reveladora** entregará y **La Parte Receptora** recibirá, la información y datos necesarios para la posible celebración y desarrollo del Proyecto, el cual, de no llegar a concretarse, no eximirá al receptor de responder por el cumplimiento del presente Acuerdo de Confidencialidad.

La Información Confidencial continuará siendo de propiedad de **La Parte Reveladora**. Las Partes acuerdan que la Información Confidencial que pueda ser revelada puede estar sujeta al derecho de Habeas Data, de propiedad y sujeta a los derechos de propiedad intelectual, incluyendo sin limitación, aquellos derechos de conformidad con las leyes aplicables a marcas, patentes, diseños, derecho de reproducción, secretos comerciales y competencia desleal. Las Partes expresamente manifiestan que el presente Acuerdo no otorga a **La Parte Receptora** ningún derecho adicional al de utilizar la Información Confidencial para efectos del Proyecto, y nada de lo expresado en este Acuerdo podrá ser interpretado como la creación tácita o expresa de una licencia para la utilización de la Información Confidencial.

CUARTA. CONFIDENCIALIDAD: Las partes que suscriben este Acuerdo, así como todas y cada una de las personas que intervengan en el objeto del mismo acuerdan que cualquier información dada a conocer entre ellas, así como la que pudiesen desarrollar de forma conjunta durante la ejecución del proyecto, será mantenida confidencial, reservada o secreta, y no podrá divulgarse o revelarse total o parcialmente a personas ajenas a las partes de este acuerdo, así como tampoco a ningún medio conocido o por conocer, según las siguientes condiciones:

1. **La Parte Receptora** de la información la mantendrá confidencial, y se abstendrá de revelarla a terceras personas, en beneficio propio o ajeno, sin el consentimiento previo y por escrito de **La Parte Reveladora**.

2. La obligación de confidencialidad se extenderá a los desarrollos que se produzcan como consecuencia del uso de la información intercambiada producto de la ejecución del Proyecto.

3. **La Parte Receptora** tendrá prohibido transferir o suministrar a terceros a cualquier título, información que obtenga de las otras partes en desarrollo del Proyecto.

4. Las partes se comprometen a adoptar todas las medidas de seguridad razonables de acuerdo con la naturaleza de la información para garantizar la confidencialidad, secreto y reserva de la misma.

5. Las partes se obligan a mantener en confidencialidad toda información dada a conocer por una parte a las demás que pueda ser objeto de protección como Propiedad Intelectual en sí misma y sin desarrollo alguno.

PARÁGRAFO PRIMERO: Si existe duda sobre la confidencialidad, reserva o secreto de alguna información suministrada por una parte a las demás, ésta será tratada como confidencial hasta que la parte que dio a conocer la información manifieste lo contrario por escrito.
PARÁGRAFO SEGUNDO: El incumplimiento a la obligación de confidencialidad que le asiste a las partes intervinientes será considera como causal de incumplimiento del acuerdo.

QUINTA. OBLIGACIONES DE LA PARTE RECEPTORA: La Parte Receptora se obliga, respecto de la Información recibida de **La Parte Reveladora**, a:

1. Guardar absoluta confidencialidad y reserva, durante la vigencia de este Acuerdo, en relación con la totalidad de la Información Confidencial, de tal forma que no sea conocida por algún tercero, y por tanto no podrá revelar o divulgar, exhibir, mostrar, comunicar, usar, utilizar en forma directa o indirecta, total o parcialmente dicha información so pena de ser responsable por la totalidad de los daños directos, perjuicios, gastos y costas que la violación de dicha confidencialidad ocasione a **La Parte Reveladora**.

2. No editar, copiar, compilar o reproducir por cualquier medio, parcial o totalmente la Información Confidencial para fines distintos a los contemplados en el presente acuerdo, salvo previa autorización expresa de **La Parte Reveladora**.

3. No utilizar en forma alguna la Información Confidencial, directamente o a través de terceros, en asuntos, negocios y/o actividades de cualquier tipo, distintos a aquellos autorizados expresamente por las partes. Esta obligación implica la de restituir a **La Parte Reveladora**, dentro de los diez (10) días hábiles siguientes a la terminación del presente Acuerdo y/o al momento en que ésta lo requiera, toda la Información Confidencial que tenga en su poder, ya sea en documentos escritos, medios magnéticos o en cualquier otro medio, e igualmente a retirarlos completamente de sus respectivos archivos físicos y electrónicos. Asimismo, deberá destruir la Información Confidencial cuando así sea requerido por **La Parte Reveladora** en el mismo término señalado para su devolución, para lo cual deberá enviar y constancia escrita de haber procedido con la destrucción.

4. No usar la Información Confidencial de manera tal que pueda ser, directa o indirectamente, perjudicial para los intereses de **La Parte Reveladora** y, en consecuencia, se compromete a adoptar medidas de seguridad adecuadas para conservarla libre de hurto o de acceso por parte de terceros no autorizados, ejerciendo sobre ésta el mismo grado de diligencia que utiliza para proteger información confidencial de su propiedad, y en todo caso, por lo menos el grado de diligencia que utilizaría un buen empresario en la protección de su información confidencial.

5. Asumir la responsabilidad por la totalidad de los daños, perjuicios, gastos y costas que genere el mal o inadecuado manejo de la Información Confidencial o la violación de las obligaciones de reserva y confidencialidad establecidas en este Acuerdo.

6. Las Partes aceptan y declaran que toda la Información Confidencial ha sido o será revelada únicamente con el propósito de permitir el cabal cumplimiento de los fines señalados en la cláusula Segunda del presente Acuerdo.

7. Dentro de la Información Confidencial se encuentra incluido el secreto empresarial, entendido de conformidad con la definición que del mismo consagra el artículo 260 de la Decisión 486 de 2000 de la Comunidad Andina o las normas que la modifiquen o sustituyan.

8. Notificar inmediatamente a **La Parte Reveladora** ante el hallazgo del uso no autorizado o revelamiento de la Información Confidencial y cooperar con **La Parte Reveladora** para mitigar cualquier consecuencia adversa.

9. Reconocer como en efecto lo hace, que la recepción de la Información Confidencial no le da, ni expresa ni implícitamente, autorización, permiso o licencia de uso de marcas comerciales, patentes, derechos de autor, o de cualquier otro derecho de propiedad industrial o intelectual de **La Parte Reveladora.**

10. No infringir ninguno de los derechos de propiedad intelectual ni otros derechos de **La Parte Reveladora**, que recaigan sobre la Información Confidencial.

11. Evitar la realización de actos, conductas, hechos constitutivos de competencia desleal, actos de descrédito o cualquier acto desleal, incluyendo actos de sustracción y explotación de secretos o información Confidencial.

12. **La Parte Receptora** reconoce que la Información Confidencial obtenida o que se obtenga con ocasión del desarrollo de las negociaciones entre las partes, tiene un valor comercial que no es del dominio público y que su divulgación podría afectar gravemente los intereses de **La Parte Reveladora**y del compromiso que se pretende suscribir formalizando el presente Acuerdo.

13. Velar por el cumplimiento del presente acuerdo de confidencialidad, además de su pronta y oportuna divulgación hacia todas las personas involucradas en el desarrollo y ejecución de las negociaciones comerciales en las que se encuentran las partes.

SEXTA. EXCEPCIONES A LA CONFIDENCIALIDAD: La confidencialidad y limitaciones establecidas en este Acuerdo no serán aplicables a la Información Confidencial, siempre:

a) Que la Información Confidencial haya sido o sea de dominio público.

b) Que antes de ser revelada, la Información estaba en posesión legítima de **La Parte Receptora**. Para estos efectos, no se considerará como posesión legítima aquella información que estaba en posesión de **La Parte Receptora** cuando tal información fue obtenida en forma de confidencialidad o sobre la base de confianza con anterioridad a la suscripción de este Acuerdo.

c) Que posteriormente a su revelación, la Información Confidencial sea legalmente recibida de un tercero que tenga derechos para distribuir la información sin notificación de ninguna restricción de su derecho a revelarla posteriormente.

d) Que sea desarrollada independientemente, o adquirida por **La Parte Receptora** a través de personas que no han tenido directa o indirectamente acceso o conocimiento de tal Información Confidencial.

e) Que deba ser revelada y/o divulgada en desarrollo o por mandato de una ley, decreto, sentencia u orden de la autoridad competente en ejercicio de sus funciones legales. En este caso, **La Parte Receptora** se obliga a avisar inmediatamente haya tenido conocimiento de esta obligación a **La Parte Reveladora**, cuando ello sea legalmente posible, para que pueda tomar las medidas necesarias para proteger su Información Confidencial y de igual manera se compromete a tomar las medidas razonables para atenuar los efectos de tal divulgación, tales como la depuración de la información suministrada mediante la eliminación de datos sensibles de la información que no interesen a la autoridad pública, cuando ello sea compatible con el requerimiento; la solicitud de reserva del expediente cuando ello proceda, o el pacto de confidencialidad en los acuerdos de colaboración e intercambio de información. El compromiso de aviso a **La Parte Reveladora** no puede constituir dilación o retardo en la atención al requerimiento de autoridad competente que debe atender la Parte que corresponda.

f) Que el período de confidencialidad haya vencido.

PARÁGRAFO: No obstante, lo previsto en esta Cláusula, **La Parte Receptora** reconoce y acepta que la Información que habitualmente se entrega, de acuerdo a los requerimientos realizados para el desarrollo del Proyecto, posee la condición de reservada, por lo cual, no podrá ser divulgada a terceros, en ningún caso, sin el consentimiento previo y escrito de **La Parte Reveladora**.

SÉPTIMA. PROPIEDAD DE LA INFORMACIÓN: Toda información intercambiada en virtud del presente Acuerdo es de propiedad exclusiva de la Parte de donde proceda. En consecuencia, ninguna de las Partes utilizará información de la otra para su propio uso, ni para fines diferentes al desarrollo del objeto del Acuerdo, ni podrá reproducir la misma sin autorización previa de su propietario, quien podrá solicitar su devolución en cualquier momento. Así mismo, dicha información deberá ser devuelta o destruida por **La Parte Receptora** a la terminación del Proyecto propuesto dentro de los diez (10) días hábiles siguientes a su terminación. La información que resulte de la ejecución del Proyecto corresponderá a quien la desarrolle, salvo que se disponga lo contrario.

OCTAVA. ENTREGA O DESTRUCCIÓN DE INFORMACIÓN CONFIDENCIAL: La información que deba considerarse confidencial, solo podrá ser entregada al personal autorizado de **La Parte Receptora**. Las Partes se reservan la facultad de auditar, en cualquier momento previa coordinación

con el líder del área a auditar, el cumplimiento de las obligaciones dispuestas en este Acuerdo. Una vez termine Proyecto, o este Acuerdo, independientemente de la causa de terminación, dentro de los diez (10) días hábiles siguientes, salvo la existencia de normas financieras sobre conservación de información a la cual esté sujeto **BANCOOMEVA**. **La parte Receptora** deberá devolver toda la información recibida de **La Parte Reveladora**, salvo que éste último solicite destrucción de la misma, caso en el cual la parte Receptora deberá acreditar su destrucción dentro de los diez (10) días hábiles siguientes a la fecha de terminación del proyecto.

NOVENA. PROPIEDAD INTELECTUAL E INDUSTRIAL: Toda información intercambiada en virtud del presente Acuerdo es de propiedad exclusiva de la Parte de donde proceda. En consecuencia, ninguna de las Partes utilizará información de la otra para su propio uso, ni para fines diferentes al desarrollo del objeto del Acuerdo, ni podrá reproducir la misma sin autorización previa de su propietario, quien podrá solicitar su devolución en cualquier momento. Así mismo, dicha información deberá ser devuelta o destruida por La Parte Receptora a la terminación del negocio jurídico propuesto dentro de los diez (10) días hábiles siguientes a su terminación. La información que se genere como resultado del proyecto de investigación propuesto, corresponderá y podrá ser utilizada de manera única y exclusiva por parte de **Bancoomeva** y no podrá ser comercializada ni utilizada por y hacia terceros, en especial a otras entidades financieras, salvo que se disponga lo contrario.

DÉCIMA. PROTECCIÓN DE DATOS PERSONALES: Cuando se comparta Información Confidencial sometida al régimen de protección de datos personales según la legislación vigente en la materia, las Partes declaran que cuentan con las respectivas autorizaciones de tratamiento por parte de los titulares de la información, así como también que conservan copias de las mismas.

La Parte Receptora se obliga a no compartir información sometida a protección de datos personales cuando el objeto del Proyecto sea distinto a las finalidades para las cuales el titular de la información dio su consentimiento.

La Parte Receptora se obliga a salir a la defensa de **La Parte Reveladora** ante cualquier reclamación por parte del titular de la información por el uso indebido que se dé a los datos personales del titular por parte de **La Parte Receptora**. Cuando el titular de los datos personales formule demanda judicial a **La Parte Receptora** por el tratamiento de sus datos, **La Parte Reveladora** se obliga a aceptar el llamamiento en garantía que le hiciere **La Parte Receptora**, o en su defecto se obliga a asumir el valor de las condenas y los gastos de su defensa, así como también las sanciones económicas que le llegaren a ser impuestas por cualquier autoridad administrativa por la misma circunstancia.

Las Partes, independientemente de su calidad de Encargado de Tratamiento o de Responsable de Tratamiento, se obligan a cumplir fielmente las finalidades para las cuales el titular del dato autorizó su tratamiento, y consecuentemente se obligan a informarse recíprocamente si la destinación de la información no contraviene las disposiciones relativas al régimen de protección de datos personales.

DÉCIMA PRIMERA. RECONOCIMIENTO DE PERJUICIOS: En caso de incumplimiento por parte de alguna de las partes de las obligaciones descritas en el presente acuerdo, responderá patrimonialmente por los daños y perjuicios generados a la otra parte.

DÉCIMA SEGUNDA. CLAUSULA COMPROMISORIA: Las Partes acuerdan que todo litigio, discrepancia, cuestión o reclamación resultantes de la ejecución o interpretación del presente Acuerdo o relacionados con él, directa o indirectamente, se resolverán ante la justicia ordinaria cuando la cuantía fuere inferior a doscientos (200) salarios mínimos legales mensuales vigentes; sí la cuantía fuere superior se deberá someter a un Tribunal de Arbitramento que se sujetará a lo dispuesto en la Ley 1563 de 2012 y en las normas que lo reglamenten, adicionen o modifiquen, de acuerdo a las siguientes reglas: **a)** El Tribunal estará integrado por tres (3) árbitros, si la cuantía de las diferencias es superior a quinientos (500) salarios mínimos mensuales legales vigentes a ese momento. En caso contrario, el tribunal se integrará con un solo árbitro; **b)** Los árbitros o el árbitro único serán designados por Las Partes de común acuerdo. En caso de que no llegaren a un acuerdo en la elección de los árbitros, serán designados por el Centro de Conciliación, Arbitraje y Amigable Composición de la Cámara de Comercio de la ciudad donde deba funcionar el Tribunal. Los árbitros o el árbitro único deberán ser abogados con especialidad en Derecho Mercantil; **c)** La organización interna del Tribunal se sujetará a las reglas previstas para el efecto por el Centro de Conciliación, Arbitraje y Amigable Composición de la Cámara de Comercio de la ciudad donde deba funcionar el Tribunal; **d)** Los gastos, costos y honorarios del proceso arbitral se establecerán con base en las tarifas vigentes del Centro de Conciliación, Arbitraje y Amigable Composición de la Cámara de Comercio de la ciudad donde deba funcionar el Tribunal, aprobadas por el Ministerio de Justicia. **e)** El Tribunal decidirá en derecho; **f)** El Tribunal funcionará en Cali en el Centro de Conciliación, Arbitraje y Amigable Composición de la Cámara de Comercio de esta ciudad. **g)** Las costas serán de cargo de la Parte que resultare vencida en el proceso Arbitral.

DÉCIMA TERCERA. DURACIÓN: Este Acuerdo tendrá duración indefinida mientras exista relación académica entre las partes, y permanecerá vigente después de la terminación del Proyecto hasta cinco (5) años más.

DÉCIMA CUARTA. NO RENUNCIA: El incumplimiento o demora en el ejercicio del derecho de una de las Partes para solicitar el cumplimiento de cualquiera otra en cualquier momento, en ningún caso afectará el derecho de hacer cumplir dicho derecho posteriormente; la renuncia de cualquiera de las Partes al incumplimiento de cualquier término no será considerada como renuncia a cualquier incumplimiento posterior de dicho término.

DÉCIMA QUINTA. ACUERDO COMPLETO: El presente Acuerdo constituye la totalidad y el único acuerdo existente entre las Partes.

DÉCIMA SEXTA. INDIVISIBILIDAD: Si cualquier término o disposición de este Acuerdo o su aplicación a cualquier persona o circunstancia fueran invalidas, ilegales o no aplicables sea cual fuera

su alcance, el remanente de este Acuerdo permanecerá, sin embargo, válido y en plena vigencia siempre y cuando la sustancia económica o legal de dichas transacciones contempladas no afecte adversamente a cualquiera de las Partes. Si se encuentra que cualquier término o disposición es inválida, ilegal o no aplicable, las Partes negociarán de buena fe para modificar este Acuerdo para poder lograr la intención original de las Partes de la manera más exacta posible en forma aceptable con el fin de que las transacciones contempladas sean cumplidas hasta donde más sea posible.

DÉCIMA SÉPTIMA. INDEMNIDAD: La Parte Receptora se compromete con **La Parte Reveladora** a mantenerla indemne contra todo reclamo, demanda, acción legal y costo que se cause o surja de reclamarse como consecuencia de un incumplimiento por parte de **La Parte Receptora** de alguna o algunas de las obligaciones a su cargo en este Acuerdo.

DECIMA OCTAVA. CESIÓN: Este Acuerdo es personal para las partes, y no se puede ceder o transferir por ninguna de ellas sin la previa autorización por escrito de la otra, y una cesión no autorizada o transferencia será considerada nula.

DECIMA NOVENA. LEGISLACIÓN APLICABLE: Este Acuerdo se regirá e interpretará de conformidad con la legislación vigente en la República de Colombia durante su ejecución.

VIGÉSIMA. NOTIFICACIONES: Para efectos de las comunicaciones que se deban surtir entre Las Partes de este Acuerdo, referentes al cumplimiento de sus obligaciones, se acuerdan las siguientes direcciones y teléfonos:

Bancoomeva:
Nombre: Claudia Johanna Polo Yepes
Cargo: Jefa Analítica
Teléfono: 3187080789
Dirección: Cra. 61 #9-250, Piso 2
Correo: claudiaj_polo@coomeva.com.co

La parte Receptora:
Nombre: **David Padilla López**
Teléfono: 315 538 1731
Dirección: Avenida Cañasgordas # 118 - 250
Correo: daveland90@javerianacali.edu.co

Nombre: **Marisol Correa Díaz**
Teléfono: 314 464 7733
Dirección: Avenida Cañasgordas # 118 - 250
Correo: mcorrea498@javerianacali.edu.co

Nombre: **Fabian Rolando Caballero Cortes**
Teléfono: 311 221 7869
Dirección: Avenida Cañasgordas # 118 - 250
Correo: faroca20@javerianacali.edu.co

Nombre: **María Constanza Pabón**
Cargo: Director Departamento de Electrónica y Ciencias de la Computación – Pontificia Universida Javeriana Cali
Teléfono: 3218200
Dirección: Avenida Cañasgordas # 118 - 250
Correo: mcpabon@javerianacali.edu.co

VIGÉSIMA PRIMERA. FIRMA ELECTRÓNICA: Las partes acuerdan formalizar este documento mediante firma electrónica a través de Adobe Sign, medio equivalente funcionalmente y con los mismos efectos legales a la firma manuscrita en papel, cuya validez jurídica se encuentra reconocida en la Ley 527 de 1999.

El presente documento se firma electrónicamente el día y la hora que consta en el Informe de Auditoría generado por la compañía VeriSign, Inc, tercero encargado por Adobe para certificar la autenticidad de la transacción.

La Parte Reveladora,

Sandra Maria Orozco Agudelo
C.C. No. 66.761.423 de Palmira (Valle)
Representante Legal

La Parte Receptora,

DAVID PADILLA LOPEZ (6 mar.. 2024 18:56 EST)

David Padilla López
C.C. No. 1.114.060.418 de San Pedro

Fabian Rolando Caballero Cortes
C.C. No. 93.389.721 de Ibagué

Marisol Correa Díaz
C.C. No. 1.070.983.511 de Facatativá

María Constanza Pabón
C.C. No. 34.559.226 de Popayán

Elaboró: JDD- Abogado Banco
Aprobó: MCM – Directora Jurídica Banco
Aprobó: LJP – Directora Administrativa Banco

Para todos los efectos administrativos de **Bancoomeva**, se deja constancia que el presente Acuerdo fue leído y aprobado en cuanto a su objeto, duración y obligaciones de las partes, por **Claudia Johanna Polo Yepes - Jefe Analítica** quien así mismo, efectúo la correspondiente valoración sobre la idoneidad, capacidad y seriedad de los receptores y en razón de ello, es que **Bancoomeva** ha suscrito el presente Acuerdo con **La Parte Receptora.**

Claudia Johanna Polo Yepes
Supervisor del Contrato
Jefe Analítica
Banco Coomeva S.A.

Modelo Acuerdo de Confidencialidad Bancoomeva - Estudiantes PUJ VF

Informe de auditoría final 2024-03-07

Fecha de creación:	2024-03-06
Por:	Juan David Duque Quintero (juand_duqueq@coomeva.com.co)
Estado:	Firmado
ID de transacción:	CBJCHBCAABAA-9vgwDtqEmc4QaJVxOd4tLdR0aX3hiGP

Historial de “Modelo Acuerdo de Confidencialidad Bancoomeva - Estudiantes PUJ VF”

Juan David Duque Quintero (juand_duqueq@coomeva.com.co) ha creado el documento.
2024-03-06 - 19:38:58 GMT

El documento se ha enviado por correo electrónico a claudiaj_polo@coomeva.com.co para su firma.
2024-03-06 - 19:43:42 GMT

El documento se ha enviado por correo electrónico a lizbeth Jimenez Palacio (lizbeth_jimenez@coomeva.com.co) para su aprobación
2024-03-06 - 19:43:42 GMT

claudiaj_polo@coomeva.com.co ha visualizado el correo electrónico.
2024-03-06 - 20:16:33 GMT

El firmante claudiaj_polo@coomeva.com.co firmó con el nombre de Claudia Polo
2024-03-06 - 20:54:13 GMT

Claudia Polo (claudiaj_polo@coomeva.com.co) ha firmado electrónicamente el documento.
Fecha de firma: 2024-03-06 - 20:54:15 GMT. Origen de hora: servidor.

lizbeth Jimenez Palacio (lizbeth_jimenez@coomeva.com.co) ha visualizado el correo electrónico.
2024-03-06 - 22:02:19 GMT

lizbeth Jimenez Palacio (lizbeth_jimenez@coomeva.com.co) ha aprobado el documento.
Fecha de aprobación: 2024-03-06 - 22:02:33 GMT. Origen de hora: servidor.

El documento se ha enviado por correo electrónico a MARIA CLEMENCIA MARROQUIN MORENO (mariac_marroquin@coomeva.com.co) para su aprobación
2024-03-06 - 22:02:35 GMT

El documento se ha enviado por correo electrónico a Juan David Duque Quintero (juand_duqueq@coomeva.com.co) para su aprobación

2024-03-06 - 22:02:35 GMT

Juan David Duque Quintero (juand_duqueq@coomeva.com.co) ha aprobado el documento.

Fecha de aprobación: 2024-03-06 - 22:04:19 GMT. Origen de hora: servidor.

MARIA CLEMENCIA MARROQUIN MORENO (mariac_marroquin@coomeva.com.co) ha visualizado el correo electrónico.

2024-03-06 - 22:41:16 GMT

MARIA CLEMENCIA MARROQUIN MORENO (mariac_marroquin@coomeva.com.co) ha aprobado el documento.

Fecha de aprobación: 2024-03-06 - 22:42:38 GMT. Origen de hora: servidor.

El documento se ha enviado por correo electrónico a daveland90@javerianacali.edu.co para su firma.

2024-03-06 - 22:42:40 GMT

El documento se ha enviado por correo electrónico a mcorrea498@javerianacali.edu.co para su firma.

2024-03-06 - 22:42:40 GMT

El documento se ha enviado por correo electrónico a faroca20@javerianacali.edu.co para su firma.

2024-03-06 - 22:42:40 GMT

El documento se ha enviado por correo electrónico a mcpabon@javerianacali.edu.co para su firma.

2024-03-06 - 22:42:40 GMT

faroca20@javerianacali.edu.co ha visualizado el correo electrónico.

2024-03-06 - 23:29:25 GMT

daveland90@javerianacali.edu.co ha visualizado el correo electrónico.

2024-03-06 - 23:50:40 GMT

El firmante daveland90@javerianacali.edu.co firmó con el nombre de DAVID PADILLA LOPEZ

2024-03-06 - 23:56:34 GMT

DAVID PADILLA LOPEZ (daveland90@javerianacali.edu.co) ha firmado electrónicamente el documento.

Fecha de firma: 2024-03-06 - 23:56:36 GMT. Origen de hora: servidor.- Firma capturada desde el dispositivo con el número de teléfono XXXXXXXX1731

mcorrea498@javerianacali.edu.co ha visualizado el correo electrónico.

2024-03-07 - 0:08:08 GMT

El firmante mcorrea498@javerianacali.edu.co firmó con el nombre de Marisol Correa Diaz

2024-03-07 - 1:33:30 GMT

Marisol Correa Diaz (mcorrea498@javerianacali.edu.co) ha firmado electrónicamente el documento.
Fecha de firma: 2024-03-07 - 1:33:32 GMT. Origen de hora: servidor.

mcpabon@javerianacali.edu.co ha visualizado el correo electrónico.
2024-03-07 - 2:19:29 GMT

El firmante faroca20@javerianacali.edu.co firmó con el nombre de Fabian Caballero Cortes
2024-03-07 - 3:40:55 GMT

Fabian Caballero Cortes (faroca20@javerianacali.edu.co) ha firmado electrónicamente el documento.
Fecha de firma: 2024-03-07 - 3:40:57 GMT. Origen de hora: servidor.

El firmante mcpabon@javerianacali.edu.co firmó con el nombre de María Constanza Pabón Burbano
2024-03-07 - 17:51:25 GMT

María Constanza Pabón Burbano (mcpabon@javerianacali.edu.co) ha firmado electrónicamente el documento.
Fecha de firma: 2024-03-07 - 17:51:27 GMT. Origen de hora: servidor.

El documento se ha enviado por correo electrónico a Sandra Maria Orozco Agudelo (sandram_orozco@coomeva.com.co) para su firma.
2024-03-07 - 17:51:29 GMT

Sandra Maria Orozco Agudelo (sandram_orozco@coomeva.com.co) ha visualizado el correo electrónico.
2024-03-07 - 17:51:37 GMT

Sandra Maria Orozco Agudelo (sandram_orozco@coomeva.com.co) ha firmado electrónicamente el documento.
Fecha de firma: 2024-03-07 - 19:09:51 GMT. Origen de hora: servidor.

Documento completado.
2024-03-07 - 19:09:51 GMT

FICHA RESUMEN
PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

TÍTULO: OPTIMIZACIÓN DE LOS CANALES DIGITALES DEL BANCO COOPERATIVO BANCOOMEVA MEDIANTE MODELOS DE CLUSTERIZACIÓN

1. ÁREA DE TRABAJO: Sector Financiero
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicativo
3. ESTUDIANTE(S): David Padilla López, Marisol Correa Diaz, Fabian Caballero Cortes
4. CORREO ELECTRÓNICO:

a. daveland90@javerianacali.edu.co
b. faroca20@javerianacali.edu.co
c. mcorrea498@javerianacali.edu.co

5. DIRECCIÓN Y TELEFONO: Cl 18# 118-250
6. DIRECTOR: María Constanza Pabón
7. VINCULACIÓN DEL DIRECTOR:
8. CORREO ELECTRÓNICO DEL DIRECTOR: mcpabon@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): No aplica
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Banco Cooperativo
11. OTROS GRUPOS O EMPRESAS: No aplica
12. PALABRAS CLAVE (al menos 5:) Clustering, clientes, segmentación, canales digitales, entidad financiera.
13. FECHA DE INICIO: 1 de septiembre de 2023
14. FECHA DE FINALIZACIÓN: 28 de junio de 2024
15. RESUMEN: Por medio del Clustering de datos se realizó una segmentación de los clientes de una entidad financiera, generando un nuevo conocimiento del consumidor financiero para la creación de nuevas estrategias y clasificación de nuevos clientes. Este enfoque permitió identificar patrones y tendencias ocultas en grandes conjuntos de datos transaccionales y financieros, brindando a la empresa una comprensión más profunda de su base de clientes.

OPTIMIZACIÓN DE LOS CANALES DIGITALES DEL BANCO COOPERATIVO "BANCOOMEVA" MEDIANTE MODELOS DE CLUSTERIZACIÓN

Marisol Correa Diaz – Código 8980182
David Padilla López- Código 8980375
Fabian Caballero Cortes – Código 8980342

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

Director(a)
María Constanza Pabón Burbano

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 28 DE 2024

TABLA DE CONTENIDO

1 INTRODUCCIÓN 4

2 DEFINICIÓN DEL PROBLEMA 5

2.1 PLANTEAMIENTO DEL PROBLEMA 5

2.2 FORMULACIÓN DEL PROBLEMA 6

3 OBJETIVOS DEL PROYECTO 7

3.1 OBJETIVO GENERAL 7

3.2 OBJETIVOS ESPECÍFICOS 7

3.3 JUSTIFICACIÓN 7

4 MARCO TEÓRICO Y ANTECEDENTES 9

4.1 MARCO TEÓRICO 9

4.2 ANTECEDENTES 23

5 ANÁLISIS DE CARACTERÍSTICAS Y COMPORTAMIENTO DE CLIENTES 26

5.1 LIMPIEZA Y PROCESAMIENTO DE DATOS 28

5.2 ANÁLISIS EXPLORATORIO Y DESCRIPTIVO DE LOS DATOS (EDA) 36

5.3 SELECCIÓN DE VARIABLES PARA EL MODELO DE SEGMENTACIÓN 44

5.4 APLICACIÓN DE MODELOS 48

5.5 EVALUACIÓN DE LOS MODELOS POR ÍNDICE DE LA SILUETA 53

6 CATEGORIZACIÓN DE USUARIOS DIGITALES 55

6.1 CATEGORIZACIÓN 59

6.2 PERFILES 63

7 PROPUESTAS DE ESTRATEGIAS CON BASE EN LOS RESULTADOS OBTENIDOS 64

8 CONCLUSIONES Y TRABAJOS FUTUROS 66

8.1 CONCLUSIONES 66

8.2 TRABAJOS FUTUROS 67

9 REFERENCIAS BIBLIOGRÁFICAS 69

Pontificia Universidad JAVERIANA Cali

LISTA DE FIGURAS

Figura 1 Lectura Base Inicial 28
Figura 2 Descripción de Variables Iniciales 33
Figura 3.Histogramas Variables Numéricas 36
Figura 4. Boxplots Variables Numéricas 37
Figura 5. Matriz de Correlación 39
Figura 6. Eliminación Variables Correlacionadas Numéricas 39
Figura 7. Consultas Variables Categóricas 40
Figura 8. Aplicación One-Hot Encoding 41
Figura 9. Aplicación Identificación Valores Atípicos por Distancia Mahalanobis 42
Figura 10. Estandarización de Variables Numéricas 43
Figura 11. Varianza Explicada por las Componentes Principales 44
Figura 12. Componentes Principales Para el Modelado 45
Figura 13. Coeficiente de Cada Variable 46
Figura 14. Distribución de Clientes por los Primeros Dos Componentes Principales 47
Figura 15. Método del Codo 48
Figura 16. Entrenamiento 49
Figura 17. Distribución de los Segmentos 50
Figura 18. Método del Codo 51
Figura 19 Entrenamiento 52
Figura 20. Distribución de Clúster por MiniBatchKMeans 52
Figura 21. Índice de la Silueta Para K Means 53
Figura 22. Índice de la silueta Para MiniBatchKMeans 54
Figura 23. Pesos de las variables en las primeras dos componentes principales 56
Figura 24. Grafico de Segmentos por MiniBatchKMeans 57
Figura 25. Estadísticas Descriptivas por Variables Crediticias 58
Figura 26. Boxplot de Variables Numéricas por Segmentos 59
Figura 27. Frecuencia de Variables Sociodemográfica por Segmentos 62

1 INTRODUCCIÓN

La era digital ha revolucionado la forma de que los canales digitales han logrado una alta popularidad en los mercados financieros y se han posicionado como una de las vanguardias fundamentales en las instituciones bancarias, hasta el punto de que todos los bancos, sin distinción de su actividad economía o productos que ofrecen, los utilizan para mejorar y priorizar sus servicios, mejorando así una experiencia única que pueden utilizarse como Omnicanalidad, una estrategia y gestión de canales que pretende integrar y alinear todos los canales disponibles, para brindar a los clientes una experiencia homogénea a través de los mismos [1].

El fácil acceso a Internet y el pleno uso de las nuevas tecnologías han puesto en primer plano la forma en que los clientes interactúan con los servicios financieros. Por tanto, la implementación de modelos de segmentación y análisis de datos se posiciona como una herramienta esencial para comprender y satisfacer mejor las necesidades de los clientes [2].

Este proyecto se enmarca en esta tendencia, buscando ofrecer soluciones innovadoras para la mejora de los servicios digitales del banco a través del uso de técnicas avanzadas de análisis de datos. Este estudio se ha centrado en la aplicación dos algoritmos K-Means y MiniBatchKMeans para la segmentación de la base de clientes de Bancoomeva, lo que permitirá identificar patrones y características clave dentro de los diferentes segmentos de clientes, como reducir el churn, o la tasa de pérdida de clientes [3].

A lo largo del documento, se detallará el marco teórico que sustenta este estudio, la metodología empleada, los resultados obtenidos y las recomendaciones para futuras investigaciones. Se espera que los hallazgos de este proyecto contribuyan significativamente a la optimización de los canales digitales de Bancoomeva y al fortalecimiento de su relación con sus clientes, a través de una comprensión más precisa de sus necesidades y comportamientos financieros.

2 DEFINICIÓN DEL PROBLEMA

2.1 PLANTEAMIENTO DEL PROBLEMA

Actualmente, el sector financiero enfrenta desafíos significativos en el avance de la tecnología y la incorporación de esta para el uso interno y facilitar las transacciones de sus usuarios. La rápida adopción de canales digitales por parte de los clientes ha transformado la forma en que interactúan con las entidades financieras. Las aplicaciones móviles, las plataformas en línea y otras soluciones digitales se han convertido en herramientas indispensables para realizar transacciones bancarias, consultar saldos, solicitar productos y acceder a servicios financieros de manera rápida y segura.

A nivel país, las entidades financieras han aceptado e interiorizado en su foco el manejo de la información y el análisis de la data. Para el año 2021, el 84.4 % de las entidades financieras contaban con caracterización de sus consumidores [4], demostrando el interés por caracterizar, organizar y enfocar sus estrategias basándose en el conocimiento profundo de sus clientes.

En el contexto de la era digital, Bancoomeva se enfrenta a un desafío importante: el estancamiento en la transaccionalidad y el número de usuarios activos de sus canales digitales. Pese a la creciente popularidad del uso de canales no tradicionales como las App y portales electrónicos, la entidad ha observado una tendencia preocupante de falta de crecimiento en la adopción y su uso activo. Esta situación tiene un impacto directo en los ingresos generados a través de los canales digitales y limita el potencial de crecimiento y rentabilidad del banco.

Es especialmente preocupante el estancamiento del número de usuarios activos que realizan transacciones frecuentes a través de los canales digitales, ya que este indicador se ha mantenido sin cambios significativos incluso después de que se levantaron las restricciones impuestas por la pandemia de COVID-19. Este estancamiento en la adquisición y retención de usuarios activos limita la capacidad de Bancoomeva para capitalizar plenamente el potencial de los canales digitales y aprovechar las oportunidades de crecimiento en el entorno altamente competitivo de las instituciones financieras.

Con este proyecto se logró identificar y comprender las características de los clientes actuales y clasificarlos en grupos de afinidad. Esta información permitió encontrar características distintivas de los usuarios digitales y explorar estrategias efectivas para impulsar la adopción y el uso activo de los canales digitales [2].

Bancoomeva enfrenta el desafío del estancamiento en la transaccionalidad y el número

de usuarios de sus canales digitales. Este problema tiene un impacto negativo en los ingresos y la rentabilidad de la entidad, y requiere una solución basada en la ciencia de datos. La comprensión de los factores subyacentes, la identificación y categorización de los clientes actuales, así como el desarrollo de estrategias efectivas, son aspectos clave para superar este estancamiento y aprovechar plenamente el potencial de los canales digitales en el entorno financiero actual.

El enfoque principal de este proyecto es la segmentación de clientes en los canales digitales de Bancoomeva para mejorar la experiencia del usuario, aumentar la transaccionalidad y aumentar el número de usuarios activos en ellos. Para lograr esto, se aplicaron técnicas de clusterización para segmentar a los clientes en grupos homogéneos con características similares, lo que permitió proponer estrategias personalizadas la oferta de productos y servicios de acuerdo con las necesidades específicas de cada segmento. Esto permitirá a Bancoomeva tomar decisiones más informadas y desarrollar estrategias de retención y fidelización efectivas.

2.2 FORMULACIÓN DEL PROBLEMA

¿Cómo optimizar, a partir de Modelos de Clusterización, la adopción y el uso activo de los canales digitales del Bancoomeva para incrementar la capitalización y transaccionalidad, como el número de usuarios que acceden a ellos?

2.2.1 PREGUNTAS DE SISTEMATIZACIÓN

A. ¿Cuáles son las características y conductas específicas de los clientes actuales que utilizan los canales digitales?

B. ¿Cómo se pueden categorizar de manera efectiva a los usuarios digitales en grupos de afinidad para una mejor comprensión y segmentación?

C. ¿Qué estrategias basadas en ciencia de datos se pueden desarrollar para la personalización de la experiencia de nuevos usuarios de los canales digitales dentro del global de clientes de la entidad?

3 OBJETIVOS DEL PROYECTO

3.1 OBJETIVO GENERAL

Crear modelos de aprendizaje no supervisado, como punto de partida para optimizar la adopción y uso activo de los canales digitales de Bancoomeva mediante el empleo de técnicas de Clustering, con el propósito de incrementar la capitalización, transaccionalidad y aumentar el número de usuarios en estos canales.

3.2 OBJETIVOS ESPECÍFICOS

A. Analizar las características y comportamientos de los clientes actuales que utilizan los canales digitales que ofrece el Bancoomeva.

B. Categorizar de manera efectiva a los usuarios digitales en grupos de afinidad para una mejor comprensión y segmentación.

C. Desarrollar estrategias basadas en ciencia de datos para la experiencia de nuevos usuarios de los canales digitales dentro del global de clientes de la entidad.

3.3 JUSTIFICACIÓN

El sector financiero enfrenta un desafío trascendental: la adaptación y evolución a la banca digital. Bancoomeva, como actor relevante en este sector, no está exento de este desafío.

A pesar del auge de los canales digitales, la entidad se enfrenta a un retraso en la migración de transacciones y número de usuarios activos a los canales digitales. Esta situación genera un impacto negativo directo en los ingresos y la rentabilidad de Bancoomeva. La falta de crecimiento en la adopción y uso activo de los canales digitales limita el potencial de la entidad para capitalizar las oportunidades que ofrece el entorno financiero actual.

Este proyecto de grado se enfoca en comprender las causas subyacentes y proponer soluciones basadas en la ciencia de datos. De tal manera que la segmentación de clientes en los canales digitales de Bancoomeva se presenta como una herramienta

fundamental para:

- Mejorar la experiencia del usuario: Al comprender las características de cada segmento de clientes, se puede ofrecer una experiencia personalizada y satisfactoria.
- Incrementar la transaccionalidad: Dirigiendo las ofertas y promociones a los segmentos adecuados, se puede aumentar la probabilidad de que los clientes realicen transacciones a través de los canales digitales.
- Aumentar el número de usuarios activos: Implementando estrategias de marketing y comunicación dirigidas, se puede atraer a nuevos usuarios y fidelizar a los existentes.
- Generar ingresos: Con el incremento de usuarios y de transacciones, aumentaran las comisiones recibidas por cobro a operaciones especificas realizadas a través de los canales digitales

Este proyecto busca contribuir al crecimiento y la rentabilidad de Bancoomeva a través de la implementación de estrategias basadas en la segmentación de clientes en los canales digitales. La comprensión profunda de los clientes y la aplicación de técnicas de ciencia de datos permitirán a la entidad tomar decisiones informadas y aprovechar al máximo el potencial de la era digital.

4 MARCO TEÓRICO Y ANTECEDENTES

4.1 MARCO TEÓRICO

En el contexto actual de avances tecnológicos y cambios en los comportamientos de los usuarios, la adopción y el uso activo de canales digitales en el sector bancario se han convertido en aspectos cruciales para la competitividad y el crecimiento de las instituciones financieras. La transformación digital ha permitido a los bancos ofrecer servicios más eficientes, personalizados y accesibles, lo que a su vez ha llevado a una mayor satisfacción y fidelidad de los clientes [5].

Bancoomeva se enfrenta al desafío de aumentar la transaccionalidad y atraer un mayor número de usuarios a sus canales digitales. La digitalización en el sector bancario no solo implica la implementación de nuevas tecnologías, sino también la adaptación de los modelos de negocio y estrategias de marketing para alinearse con las expectativas y necesidades cambiantes de los clientes [2].

La adopción de canales digitales también tiene un impacto significativo en la experiencia del cliente. Los clientes valoran la conveniencia y la accesibilidad que ofrecen estos canales, como la capacidad de realizar transacciones 24/7 desde cualquier lugar. Las aplicaciones móviles y la banca en línea son ejemplos de cómo la tecnología ha transformado la relación entre el banco y sus clientes, ofreciendo servicios más rápidos y personalizados [6].

4.1.1 Canales Digitales En El Sector Bancario

Los canales digitales en el sector bancario han transformado mucho cómo los clientes interactúan con las instituciones financieras. A continuación, se detallan los principales canales digitales y su impacto en la experiencia del cliente:

4.1.1.1 Banca En Línea

Permite a los clientes acceder a sus cuentas bancarias y realizar diversas transacciones a través de Internet. Los clientes pueden gestionar sus cuentas y realizar transacciones en cualquier momento, eliminando la dependencia de ir a las entidades bancarias y optar por los horarios de las sucursales presenciales.

Tienen grandes beneficios para el cliente en transacciones electrónicas, transferencias, pagos de facturas y consulta de saldos. Estos servicios los puede realizar desde la comodidad de su casa o cualquier lugar con acceso a Internet. Evita las largas filas en

las sucursales y permite realizar operaciones de manera rápida y eficiente.

4.1.1.2 Aplicaciones Móviles

Las aplicaciones móviles bancarias son aplicaciones específicas diseñadas para smartphones y tablets que permiten a los usuarios realizar una amplia gama de operaciones bancarias. Los clientes pueden realizar transacciones en cualquier lugar y en cualquier momento, siempre que tengan acceso a un dispositivo móvil. Las apps pueden ofrecer servicios como notificaciones de transacciones, gestión de tarjetas, localización de cajeros y sucursales, y acceso rápido a soporte [7].

Al tener acceso a estos servicios, el banco pude ofrecer recomendaciones personalizadas y promociones específicas, teniendo en cuenta su perfil, años de antigüedad y los tipos de productos que cuenta en el banco.

4.1.1.3 Cajeros Automáticos Inteligentes (Smart ATMs)

Los cajeros automáticos inteligentes ofrecen funciones avanzadas que van más allá del retiro de efectivo, como depósitos de cheques, transferencias entre cuentas y pagos de servicios. Los clientes pueden realizar una mayor variedad de transacciones sin necesidad de ingresar a una sucursal bancaria de manera presencial, están disponibles fuera del horario laboral, lo que mejora la accesibilidad de los servicios bancarios. Cuentan con una Interfaz amigable, pantallas táctiles y opciones multilingües mejoran la usabilidad y la experiencia del cliente.

4.1.1.4 Chatbots Y Asistentes Virtuales

Los chatbots y asistentes virtuales son sistemas de inteligencia artificial diseñados para interactuar con los clientes, responder preguntas y asistir en la realización de transacciones.

Los clientes pueden obtener respuestas rápidas a sus consultas sin esperar tiempos de respuesta de un humano. Cuenta con una disponibilidad 24/7, ofreciendo soporte continuo y resolviendo problemas comunes y guiar a los clientes a través de procesos complejos, mejorando la satisfacción general, como el de priorizar consultas de orden general y sin tener que disponer de otros canales con más tráfico de usuarios.

4.1.1.5 Banca Telefónica Automatizada

La banca telefónica automatizada permite a los clientes interactuar con un sistema automatizado para consultas y transacciones usando comandos de voz o teclado telefónico. Los clientes pueden obtener información sobre sus cuentas y realizar transacciones básicas sin hablar con un representante humano. Cuentan con menús de voz guiados y las opciones pregrabadas facilitan la navegación y el uso del servicio. Permite resolver consultas simples de manera rápida y eficiente, reduciendo la necesidad de visitas a sucursales o de esperar en línea para hablar con un agente.

La adopción de canales digitales en el sector bancario ha revolucionado la relación entre los clientes y los bancos. Los clientes tienen acceso a sus servicios bancarios en cualquier momento y lugar, lo que mejora significativamente la conveniencia.

La automatización y digitalización de los servicios permiten a los bancos operar de manera más eficiente y con menores costos. Los canales digitales proporcionan una experiencia más fluida, rápida y personalizada, aumentando la satisfacción y fidelidad del cliente. Los bancos que adoptan tecnologías digitales pueden ofrecer servicios innovadores, lo que les permite mantenerse competitivos en el mercado financiero [6].

4.1.2 Ciencia de Datos. Definición y Conceptos

La ciencia de datos es un campo interdisciplinario dedicado a la extracción de conocimiento y perspicacias de conjuntos de datos, tanto estructurados como no estructurados. Involucra una combinación de varias disciplinas como estadística, matemáticas, informática, ingeniería, y el conocimiento específico del dominio para analizar y comprender fenómenos complejos del mundo real. La ciencia de datos se centra en la manipulación, el análisis y la interpretación de datos para apoyar la toma de decisiones, la construcción de modelos predictivos, y la generación de estrategias basadas en datos [8].

Se enfoca principalmente en los tipos de analítica Predictiva y Prescriptiva, que utilizan datos históricos para predecir eventos futuros y recomendar acciones específicas, respectivamente. Aunque el Machine Learning no es exactamente lo mismo que la Ciencia de Datos, esta última se fundamenta en gran medida en las técnicas y modelos de Machine Learning para obtener información útil y cumplir sus objetivos. Los algoritmos de Machine Learning, esenciales para la Ciencia de Datos, se pueden clasificar en categorías principales:

Aprendizaje Supervisado: en los algoritmos de aprendizaje supervisado la tarea

principal es la de generación de un modelo, que, a partir de un conjunto de datos etiquetados, pueda ser entrenado para clasificar nuevas observaciones o predecir futuros eventos.

Aprendizaje No Supervisado: en los algoritmos de Aprendizaje No Supervisado no se cuenta con un conjunto de datos que posea un etiquetado o clasificación previa. Así que la tarea principal es buscar un modelo que permita encontrar estructuras ocultas en los datos de entrada. Por tanto, es una tarea compleja definir una respuesta correcta para los resultados obtenidos porque no se conoce de antemano una muestra contra la cual comparar.

Aprendizaje Semi-Supervisado: estos algoritmos combinan los dos enfoques anteriores. A partir de conjuntos de datos etiquetados y no etiquetados, el propósito es generar un modelo que permita realizar tareas de clasificación o predicción.

Aprendizaje por Refuerzo: en estos algoritmos el aprendizaje se logra a partir de un esquema de premios o penalidades según el nivel de error obtenido en los resultados. Un modelo generado a partir de estos algoritmos podrá interactuar con un ambiente y según la retroalimentación obtenida, reforzar su aprendizaje a partir de un proceso de ensayo y error.

4.1.3 Clustering de Datos

Esta técnica pertenece al paradigma de Aprendizaje de Máquina No Supervisado; El propósito principal es la organización de los datos no etiquetados en grupos o clusters que comparten características similares. Se orienta a identificar patrones o tendencias en los datos. En esta técnica la variable objetivo es desconocida, por tanto, no es posible ajustar el modelo buscando una salida específica o deseada. Así, se busca encontrar estructuras ocultas en los datos que no es posible mediante programación convencional [9].

4.1.4 El Análisis de Componentes Principales

- El Análisis de Componentes Principales (PCA) es un método estadístico utilizado para transformar un grupo de variables correlacionadas en un conjunto de variables no correlacionadas, denominadas componentes principales. Este proceso permite reducir la dimensionalidad del conjunto de datos, conservando la mayor parte de la variabilidad presente en los datos originales. Al hacerlo, PCA facilita el análisis y la visualización de datos complejos, mejorando la eficiencia de los modelos y reduciendo el riesgo de sobreajuste. En este proceso de transformación, se van generando iterativamente nuevas variables que son combinaciones lineales de las originales y que van recogiendo la mayor

cantidad posible de información o variabilidad de los datos. A este nuevo grupo de variables se le conoce como el conjunto de Componentes Principales [10].

- Estandarización de los datos: si las variables tienen diferentes unidades o escalas, se deben estandarizar para que tengan media cero y desviación estándar uno.

$$zij = \frac{xiy - \overline{x}j}{\sigma j}$$

donde xij es el valor de la variable j para el punto i, $\overline{x}$j es la media de la variable j, y σj es la desviación estándar de la variable j.

- Calcular la matriz de covarianza de los datos estandarizados, donde X es la matriz de datos estandarizados.

$$Cov\,(X) = \frac{1}{n-1} X^T X$$

- Ordenar los vectores propios en función de sus valores propios en orden descendente. Los vectores propios con los valores propios más grandes son los que representan las componentes principales más importantes.

$$Cov\,(X)vi = \lambda ivi$$

- Proyectar los datos originales sobre los primeros kkk vectores propios seleccionados para obtener los componentes principales.

$$Z = XW$$

donde W es la matriz de vectores propios (eigenvectors) seleccionados y Z es el conjunto de datos proyectado en el espacio de las componentes principales.

4.1.5 Algoritmo K-means

El algoritmo K-means es una técnica de Clustering que pertenece al paradigma de Aprendizaje de Máquina No Supervisado. Su objetivo principal es agrupar las observaciones o individuos del conjunto de datos basándose en una medida de similitud entre sus características. Los clusters resultantes deben ser lo más compactos posible internamente y estar bien diferenciados entre sí. Este agrupamiento se logra minimizando la suma de las distancias de los individuos al centroide del cluster más cercano. El número K de clusters se elige al inicio del algoritmo, una tarea que requiere conocimiento del área o contexto específico [16].

4.1.5.1 Pasos Generales de K-means

- Elección del número K de clusters: se determina el número K de clusters basándose en el requerimiento, la experticia y el contexto del análisis.
- Inicialización: con el número K de clusters elegido, se definen K centroides iniciales, generalmente de manera aleatoria dentro del espacio definido por los datos.
- Asignación de puntos al cluster con el centroide más cercano: cada individuo se asigna al cluster cuyo centroide esté más próximo, midiendo la distancia entre cada individuo y los centroides. La distancia más usada es la euclidiana. Para un punto de datos xi, y un centroide μj, la distancia euclidiana se calcula como:

$$d(xi, \mu j) = \sqrt{\sum_{m=1}^{M} (xim - \mu jm)^2}$$

- Actualización de centroides: se recalculan las posiciones de los centroides de cada cluster como el promedio de todos los individuos dentro del cluster.
  Para un cluster j, el nuevo centroide μj se calcula como:

$$\mu j = \frac{1}{N} \sum_{i=1}^{Nj} xi$$

  Donde Nj es el número de puntos en cluster j.
- Iteración de los pasos 3 y 4: estos pasos se repiten hasta que los centroides ya no cambian significativamente, alcanzando un mínimo en la suma de los cuadrados de las distancias de los individuos a sus respectivos centroides.

Esta iteración continua hasta que se alcanza la convergencia, es decir, cuando los centroides ya no se mueven significativamente, optimizando así la compacidad y separación de los cluster.

4.1.5.2 Distancia

La distancia entre individuos y entre clusters es fundamental en los algoritmos de clustering, ya que define el concepto de similitud y es crucial para la ejecución del algoritmo y la formación de los clusters. En un conjunto de datos con P variables y un número dado de observaciones o puntos de datos, la similitud entre dos se determina por la proximidad entre ellas, por la menor distancia que las separa. Existen diversas métricas utilizadas en clustering para medir esta distancia [11].

La selección adecuada de la métrica de distancia depende de varios factores, como la naturaleza de los datos (numéricos, categóricos, mixtos), la distribución de los datos, y los objetivos específicos del análisis de clustering. Es esencial considerar cuidadosamente estas métricas para garantizar que reflejen correctamente la similitud

entre observaciones y, por lo tanto, faciliten la formación de clusters significativos y útiles en el contexto dado [11].

4.1.5.3 Distancia Euclidiana

Es una medida fundamental en el campo de la estadística y el análisis de datos, utilizada para calcular la distancia entre dos puntos en un espacio euclidiano. Es especialmente relevante en diversos contextos, desde la geometría hasta la ciencia de datos. Esta medida se define como la longitud del segmento que une dos puntos en un espacio de múltiples dimensiones, calculada como la raíz cuadrada de la suma de los cuadrados de las diferencias entre las coordenadas de los puntos. En el análisis de clustering y otras aplicaciones de aprendizaje automático, la distancia euclidiana permite evaluar la proximidad entre observaciones, facilitando la agrupación de datos similares y la separación de datos disímiles en grupos cohesivos [12].

$$d(p,q) = \sqrt{(p1 - q1)^2 + (p2 - q2)^2 + \ldots\ldots + (pn - qn)^2}$$

Donde:

p= (p1, p2, ..., pn) es el primer punto.
p= (q1, q2, ..., qn) es el Segundo punto.
d (**p, q**) es la distancia euclidiana entre **p** y **q**

4.1.5.4 Distancia Mahalanobis

Es una medida estadística utilizada para cuantificar la distancia entre un punto y un conjunto de puntos en un espacio multidimensional, teniendo en cuenta la covarianza entre las variables. A diferencia de las distancias euclidiana y Manhattan, que consideran la relación directa entre las coordenadas de los puntos, la distancia Mahalanobis corrige la distancia euclidiana al considerar la correlación entre las variables y la varianza de cada variable por separado [13].

La medida es útil cuando las variables del conjunto de datos están correlacionadas y tienen diferentes escalas de medida. Por ejemplo, en el análisis de clustering, la distancia Mahalanobis puede identificar patrones significativos en datos multivariantes al ajustar la distancia a la estructura de covarianza del conjunto de datos. Da una medida de distancia que considera la forma y la orientación de la distribución de los datos, permitiendo una evaluación más precisa de la similitud entre observaciones en un espacio de alta

dimensionalidad [13].

$$DM(x) = \sqrt{(x - \mu)\, S^{-1}(x - \mu)}$$

Donde:
x es el vector de datos (punto) que se desea medir.
μ es el vector de medias de las variables.
S es la matriz de covarianza del conjunto de datos.
S^{-1} es la inversa de la matriz de covarianza.
$(x - \mu)^T$ es la transpuesta del vector de diferencia **x - μ**

4.1.5.5 Distancia de Correlación

La distancia de correlación es una medida utilizada en el contexto del análisis de datos para evaluar la similitud entre variables basándose en sus coeficientes de correlación. En lugar de medir la distancia espacial entre puntos individuales como lo hacen otras métricas (por ejemplo, la distancia euclidiana o la distancia Manhattan), la distancia de correlación se centra en la relación lineal entre las variables.

Para calcular la distancia de correlación entre dos variables, se utiliza el complemento de su coeficiente de correlación. Esto significa que cuanto menor sea el coeficiente de correlación entre las variables, mayor será su distancia de correlación. Esta medida es útil cuando se desea agrupar variables que tienen una relación inversa o débil entre sí en comparación con otras variables del conjunto de datos [14].

En el análisis de clustering y en otras aplicaciones de aprendizaje automático y estadística, la distancia de correlación se utiliza para explorar y comprender las relaciones entre variables, permitiendo identificar patrones y estructuras subyacentes que pueden no ser evidentes mediante otras medidas de distancia más tradicionales [14].

$$DC(x, y) = 1 - p(x, y)$$

Donde:

x y **y** son los vectores de datos
p(**x**,**y**) es la correlación de Pearson **X** y **Y**

La correlación de Pearson p(**x, y**) se calcula como:

$$p(x,\ y) = \frac{\sum_{i=1}^{n}(xi - \overline{x})(yi - \overline{y})}{\sqrt{\sum_{i=1}^{n}(xi - \overline{x})^2}\ \sqrt{\sum_{i=1}^{n}(yi - \overline{y})^2}}$$

Donde:

xi e **yi** son los elementos individuales de los vectores **X** y **Y**
$\overline{x}$ y $\overline{y}$ son las medias de los vectores **X** y **Y**

4.1.6 Algoritmo MiniBatchKMeans

El algoritmo MiniBatchKMeans es una variación del algoritmo K-means tradicional diseñado para mejorar la eficiencia y escalabilidad al trabajar con grandes conjuntos de datos. Este algoritmo utiliza pequeños lotes (minibatches) de datos en lugar de procesar el conjunto completo de datos en cada iteración, lo que reduce significativamente el tiempo de cómputo y la memoria requerida [17].

4.1.6.1 Pasos Generales de MiniBatchKMeans

- Se selecciona un número K de clusters y se inicializan K centroides, generalmente de manera aleatoria.
- En cada iteración, se selecciona aleatoriamente un minibatch de datos del conjunto completo. El tamaño del minibatch es mucho menor que el tamaño total del conjunto de datos.
- Para cada punto en el minibatch, se determina el centroide más cercano y se asigna el punto a ese cluster, usando la distancia euclidiana.

$$d(xi,\ \mu j) = \sqrt{\sum_{m=1}^{M}(xim - \mu jm)^2}$$

- Los centroides se actualizan en función de los puntos asignados en el minibatch. Esto se realiza mediante una fórmula de actualización incremental que ajusta los centroides de manera eficiente sin necesidad de recalcularlos completamente.

$$\mu j \leftarrow \mu j + \frac{1}{tj+1}(xi - \mu j)$$

Donde tj es el número de puntos asignados al cluster j hasta el momento.

- Los pasos de selección de minibatches, asignación de puntos y actualización de centroides se repiten durante un número predefinido de iteraciones o hasta que los cambios en los centroides sean menores que un umbral de convergencia.

4.1.7 Determinación Del Número Óptimo De Clusters

Como se mencionó anteriormente, la determinación del número K óptimo de clusters es

un paso fundamental en un algoritmo de clustering.

4.1.7.1 Método del Codo

El "Método del Codo" es una técnica utilizada en clustering para determinar el número óptimo de clusters en un conjunto de datos. Este método es útil cuando no se conoce de antemano el número ideal de clusters y se desea encontrar una solución basada en la estructura inherente de los datos. Proporcionar una evaluación objetiva basada en la variabilidad dentro de los clusters, este método facilita la interpretación y la validación de los resultados del clustering [15].

El WCSS mide la variabilidad interna dentro de los clusters. Se calcula como la suma de las distancias euclidianas al cuadrado entre cada punto y el centroide del cluster al que pertenece:

$$WCSS = \sum_{i=1}^{k} \sum_{x \epsilon C i} \| X - \mu_i \|^2$$

Donde:

k es el número de clusters
Ci es el conjunto de puntos en el cluster i
x es un punto dentro del cluster Ci
μi es el centroide del cluster Ci
$\|x-\mu i\|2$ es la distancia euclidiana al cuadrado entre el punto $\boldsymbol{x}$ y el centroide $\boldsymbol{\mu i}$

4.1.7.2 Método de la Silueta

El Método de la Silueta es importante porque proporciona una evaluación más cuantitativa y detallada de la calidad del clustering en comparación con otros métodos heurísticos como el Método del Codo. Permite una comprensión más profunda de cómo están distribuidos los datos y cómo se relacionan entre sí dentro de los clusters identificados, ayudando a los analistas a tomar decisiones más fundamentadas y a optimizar la configuración del clustering para obtener resultados más significativos e interpretables, contienen información sobre datos que pudo presentarse como una tabla [16].

Este método aporta un criterio numérico conocido como el Coeficiente de la Silueta que puede adquirir valores en el rango entre -1 y 1. El coeficiente indica la medida de cuán similares son los individuos de un cluster (concepto de Cohesión) y que tan diferentes son respecto a los individuos de los otros clusters (concepto de Separación) [16].

La interpretación del score o puntaje obtenido es la siguiente:

un coeficiente=1 indica que el individuo está asignado de manera óptima a un cluster y los clusters están bien separados.
Un coeficiente=0 indica traslapamiento entre los clusters.
Un coeficiente=-1 indica que el indivíduo está asignado de manera incorrecta en los clusters [16].

La puntuación de la silueta evalúa la calidad del clustering. Para cada punto i se calcula la distancia media al resto de los puntos en el mismo cluster ($a(i)a(i)$a(i)) y la distancia media al cluster más cercano ($b(i)b(i)$b(i)). La puntuación de la silueta $s(i)s(i)$s(i) se calcula utilizando estas distancias para medir la separación y cohesión del punto dentro de su cluster.

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \epsilon Ci,\, j \neq i} d(i, j)$$

$$b(i) = \min_{Ck \neq Ci} \frac{1}{|C_k|} \sum_{j \epsilon Ck} d(i, j)$$

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

4.1.8 Aplicación De Técnicas De Machine Learning Para Mejorar La Interacción Con Los Clientes En Canales Digitales Bancarios

La aplicación de técnicas avanzadas de machine learning en el sector bancario no solo permite predecir el comportamiento futuro de los clientes, sino que también facilita el análisis detallado de la retroalimentación del usuario en los canales digitales.

Estas técnicas son esenciales para mejorar la experiencia del cliente y fortalecer la relación con la entidad financiera, mediante diversas estrategias centradas en la satisfacción y fidelización del cliente.

4.1.8.1 Predicción Del Comportamiento Del Cliente

Las técnicas de predicción, tales como la regresión logística, los árboles de decisión y otros modelos de aprendizaje automático, son fundamentales para anticipar métricas clave del comportamiento del cliente:

Probabilidad de Abandono (Churn): Utilizando datos históricos de transacciones y comportamientos en los canales digitales, se identifican patrones que indican una alta probabilidad de que un cliente abandone la institución bancaria. Esto permite implementar estrategias proactivas de retención de clientes, mejorando así la lealtad y reduciendo la pérdida de clientes [18].

Bancoomeva podría identificar clientes que no han utilizado de manera frecuente sus aplicaciones en la banca móvil en los últimos meses. Utilizando un modelo predictivo de churn, el banco podría activar ofertas exclusivas o mejorar la experiencia de usuario para estos clientes, incentivándolos a continuar utilizando sus servicios digitales en la banca móvil.

Propensión a adquirir productos: Mediante modelos predictivos, se evalúan las preferencias y necesidades individuales de los clientes, permitiendo recomendar productos financieros específicos que puedan ser de su interés. Esta personalización aumenta la relevancia de las ofertas y fortalece la relación cliente-banco [19].

Bancoomeva podría predecir que ciertos clientes que utilizan frecuentemente la aplicación móvil para consultar saldos podrían estar interesados en productos de inversión a largo plazo. El banco podría lanzar campañas personalizadas o recomendaciones en la aplicación móvil para promover estos productos específicos, generando así mayor sinergia en este tipo de aplicaciones.

Satisfacción del cliente: Analizando datos de encuestas, interacciones en línea y otras formas de retroalimentación, se construyen modelos que miden la satisfacción del cliente en tiempo real. Esta información es crucial para identificar áreas de mejora y ajustar los servicios, mejorando así la experiencia del usuario en los canales digitales [20].

Al iniciar sesión de usuario en la aplicación móvil de Bancoomeva, el sistema podría utilizar un modelo de personalización basado en machine learning, para mostrar de manera destacada las funciones más utilizadas por ese cliente, un historial y trazabilidad de las transferencias rápidas o consultas de saldo, adaptando así la interfaz para una experiencia más intuitiva y satisfactoria.

4.1.8.2 Análisis De Sentimiento Y Retroalimentación Del Cliente

El análisis de sentimiento, una técnica de procesamiento del lenguaje natural (NLP), se utiliza para extraer y evaluar opiniones expresadas por los clientes en plataformas digitales. Esta técnica no solo proporciona una comprensión más profunda del comportamiento y las percepciones de los clientes, sino que también permiten a las

instituciones financieras tomar decisiones informadas y basadas en datos. Al integrar estas estrategias en la gestión diaria de los canales digitales, los bancos pueden optimizar la experiencia del cliente y fortalecer su posición en el competitivo mercado financiero actual [21].

Mediante el análisis de sentimiento de las publicaciones en redes sociales, Bancoomeva podría detectar rápidamente cualquier problema técnico o insatisfacción de los clientes con respecto a las nuevas funciones implementadas en su aplicación móvil. Con esta retroalimentación en tiempo real, el banco podría tomar medidas correctivas rápidas y mejorar continuamente la experiencia del usuario en sus canales digitales.

4.1.8.3 Importancia de Clustering para Estrategias Sector Financiero

El clustering es una técnica de aprendizaje no supervisado, utilizada en análisis de datos, para agrupar un conjunto de objetos de manera que los objetos dentro de un mismo grupo (o cluster) sean más similares entre sí, que con los objetos de otros grupos. Es muy útil para explorar patrones o segmentos en datos complejos y no etiquetados [22].

Clustering permite centrar la calidad del servicio y explorar las estrategias de marketing que pueden aportar en satisfacción de los clientes con la banca en línea, incluyendo el uso de aplicaciones móviles, banca electrónica y servicios financieros digitales. Proporcionan información valiosa sobre cómo los bancos pueden aprovechar los canales digitales para interactuar con los clientes, mejorar la experiencia del usuario y promover la adopción de servicios bancarios en línea [22].

Estos hallazgos subrayan la necesidad de que las instituciones bancarias comprendan y mejoren constantemente la experiencia del usuario en los canales digitales para satisfacer las demandas y expectativas de los clientes. Como un impacto directo en los ingresos generados a través de los canales digitales y aporta un crecimiento y rentabilidad en el sector financiero.

En el caso de los canales digitales bancarios, es crucial entender las percepciones y actitudes de los usuarios hacia estos canales, así como los factores que pueden facilitar o dificultar su uso o manejo sistemático. La relación costo beneficio se fundamenta en la mejora continua de la aplicación de estos canales digitales por parte de los clientes, ampliando sus canales y mejorando la oferta, dando como resultado una mayor cobertura en los productos y servicios que la empresa ofrece.

“Basados en técnicas de predicción se logra estimar eventos futuros o comportamientos

basados en datos históricos". Esta definición de Hastie, Tibshirani y Friedman [32] resalta el papel de la predicción en la identificación de patrones y tendencias en los datos para tomar decisiones informadas y anticiparse a las necesidades de los usuarios. “La teoría de la difusión de innovaciones propone que la adopción de una nueva tecnología está influenciada por factores sociales, psicológicos y contextuales". Esta teoría de Everett Rogers proporciona un marco para comprender cómo se adoptan y se difunden las innovaciones tecnológicas en una población. La relación entre ambos conceptos radica en la capacidad de la predicción para apoyar la difusión de innovaciones tecnológicas. Al analizar datos históricos sobre la adopción de tecnologías anteriores, es posible predecir qué segmentos de la población adoptarán una nueva tecnología y en qué momento lo harán [23].

Esto permite a las organizaciones diseñar estrategias específicas para cada grupo de adoptantes. La combinación de técnicas de predicción y la teoría de la difusión de innovaciones anticipa la adopción de nuevas tecnologías y facilita su implementación efectiva y alineada con las necesidades y comportamientos de los usuarios.

En cuanto a los algoritmos que se pueden identificar para lograr mejorar la utilización de servicios digitales de los clientes en el sector bancario se encuentran: el Clustering k-means, MiniBatchKMeans o el Clustering jerárquico, los cuales permiten agrupar a los clientes bancarios en segmentos homogéneos en función de sus características, comportamientos y necesidades financieras. Esto permite una personalización más efectiva de los canales digitales, ya que se pueden adaptar los mensajes, las ofertas y los servicios según las características de cada segmento de clientes.

Por la naturaleza del objetivo, este trabajo se enmarca en los algoritmos de Aprendizaje No Supervisado, ya que no se conocen los patrones o estructuras subyacentes en las variables relacionadas con los clientes que utilizan los canales digitales de Bancoomeva. En este contexto, se describe a continuación de manera general la técnica de Clustering, la cual se ajusta y contribuye al diseño e implementación del modelo de analítica propuesto.

Estas técnicas permiten agrupar a los clientes en segmentos homogéneos basados en sus características y comportamientos. El Clustering puede identificar patrones en grandes conjuntos de datos, revelando insights que no son evidentes a simple vista. Esto es útil para personalizar las estrategias de adopción y promoción de los canales digitales, ya que permite adaptar las campañas de marketing y las ofertas de servicios a las necesidades de cada segmento de clientes [17].

Además, el Clustering facilita la identificación de grupos de clientes con similares

características y con bajo uso de canales digitales, permitiendo al banco desarrollar programas específicos para aumentar su participación, como el de capitalizar oportunidades de negocio, en el sentido de precisar nuevos productos y servicios ajustados a sus necesidades y en sintonía a su perfil financiero. Esto no solo incrementa la transaccionalidad, sino que también mejora la eficiencia operativa al reducir la dependencia de los canales tradicionales y costosos, como las sucursales físicas [24].

4.2 ANTECEDENTES

Se destaca el informe de la importancia de la calidad del servicio, en el que se presenta el concepto de calidad de servicio electrónico y se propone un modelo conceptual con dimensiones clave como facilidad de uso, eficiencia, capacidad de respuesta y seguridad. Sosteniendo que la calidad del servicio electrónico influye en la satisfacción y la lealtad del cliente, lo que a su vez afecta la adopción y el uso activo de los canales digitales en el sector bancario [25].

Además, el concepto de calidad del servicio tecnológico (TTQ), reconoce la importancia de la calidad tecnológica en la experiencia del usuario. Los clientes evalúan la calidad de las características tecnológicas y las interfaces de usuario de los canales digitales, y esta percepción influye en su satisfacción y en su intención de uso. Así mismo, los estudios, sobre la adopción de tecnologías financieras (FinTech) [26] destacan la relevancia de la utilidad financiera, la facilidad de uso, la compatibilidad y el riesgo percibido en la adopción de servicios financieros digitales.

La propuesta de investigación cuenta con artículos de referencia donde destacan la importancia de los canales digitales en el sector bancario. Se enfocan en que los avances tecnológicos han transformado la forma en que los clientes interactúan con los bancos y realizan transacciones financieras. Los canales digitales ofrecen comodidad, accesibilidad las 24/7 y una amplia gama de servicios y opciones personalizadas para los clientes bancarios. El modelo SERVQUAL destaca la importancia de la calidad del servicio percibida por los clientes. Las dimensiones del modelo, como confiabilidad, capacidad de respuesta, seguridad, empatía y tangibilidad, son relevantes para los canales digitales. Los usuarios esperan que los canales digitales bancarios sean confiables, rápidos, seguros, fáciles de usar y brinden una sensación tangible de calidad en términos de interfaces y características tecnológicas [27].

Estos factores también son aplicables a los canales digitales bancarios y contribuyen a la experiencia del usuario, proporciona una visión general concisa de FinTech y se define el término como "la innovación tecnológica en la industria financiera que conduce a nuevos modelos de negocio, aplicaciones, procesos o productos con un impacto material

o disruptivo en la provisión de servicios financieros" [26]. Se cuentan con estudios los cuales dan cuenta en la calidad percibida del servicio en los portales web, que también se aplica a los canales digitales bancarios. La calidad percibida del servicio tiene una influencia significativa en la satisfacción del usuario y en su intención de uso. Respaldando la importancia de la experiencia del usuario en la adopción y el uso activo de los canales digitales en la banca [25].

Si bien el manejo y la fidelización de los clientes es importante para la aplicabilidad de estos modelos de predicción, ejerce una doble vía o beneficio el hecho de integrarlos para la segmentación de clientes y la predicción de la rentabilidad en el sector bancario, utilizando técnicas de Clustering para segmentar a los clientes en grupos homogéneos, en función de su comportamiento financiero, y luego aplican algoritmos de predicción para estimar la rentabilidad futura de cada segmento. Estudios muestran cómo esta metodología puede ayudar a los bancos a tomar decisiones estratégicas y dirigir sus recursos de manera más efectiva [19].

Se han evidenciado estudios en los cuales se revisan la aplicación de técnicas de Clustering y clasificación en el sector bancario, se indaga como estas técnicas pueden utilizarse para segmentar clientes, detectar fraudes, predecir el riesgo crediticio y mejorar la toma de decisiones en el sector bancario. Proporcionan ejemplos de estudios empíricos que demuestran la eficacia de estas técnicas en la industria bancaria. Como el aporte de algoritmos de aprendizaje automático, para predecir la rotación de clientes en el sector bancario, utilizando técnicas de clasificación y predicción para identificar patrones de comportamiento que puedan indicar la probabilidad de que un cliente abandone el banco [24].

La principal técnica de ciencia de datos que va a ser usada durante el proyecto es el Clustering de datos, este método ha sido implementado en otras ocasiones con fines similares como el caso de Roncallo en su proyecto “Segmentación de clientes de una empresa utilizando recencia, frecuencia, monto (rfm) y métodos de Clustering” [28] donde su objetivo principal es la organización de datos para obtener un mayor provecho, al igual que la presente propuesta, Rocallo utilizó sobre transacciones, en este caso, de ventas. Antes de la implantación del clutering, Rocallo se efectuó el método RFM (Recency, Frequency, Monetary Value) con el fin de discretizar las variables para facilitar el Clustering. Como resultado en este estudio, se obtuvo una segmentación de clientes basada en variables globales que permitió mejorar la relación entre cliente y empresa [28].

De esta misma manera Becerra en su trabajo “La innovación tecnológica en el contexto de los clusters regionales” expone el uso del Clustering para la comprensión de relaciones entre entidades. Por medio de este artículo se evidencia la utilización de modelos gráficos

construidos para el análisis de los clusters, como se pueden identificar tipologías de innovación como las innovaciones de tipo socioeconómico e innovaciones de tipo empresarial [29]. Estudios de "Fraud detection in online credit card transactions using big data analytics" empleó MiniBatchKMeans para agrupar transacciones de tarjetas de crédito y detectar patrones de comportamiento anómalos, permitiendo identificar posibles fraudes de manera eficiente [30].

Otro estudio realizado por Behavior Analysis in Mobile Banking Apps utilizó MiniBatchKMeans para segmentar a los usuarios de una aplicación de banca móvil en función de su comportamiento, como la frecuencia de inicio de sesión, tipos de transacciones y tiempo de uso. Los resultados ayudaron a mejorar la interfaz de usuario y las funcionalidades de la aplicación [31].

Estos antecedentes demuestran que la aplicación de técnicas de Clustering en el sector bancario no solo optimiza la segmentación de clientes y la personalización de servicios, sino que también mejora la eficiencia operativa y la experiencia del cliente. Al identificar y analizar patrones en los datos de clientes, los bancos pueden anticipar las necesidades y comportamientos de los usuarios, permitiendo una mejor adaptación a las expectativas cambiantes y una mayor competitividad en el mercado.

5 ANÁLISIS DE CARACTERÍSTICAS Y COMPORTAMIENTO DE CLIENTES

Para lograr este primer objetivo se llevaron a cabo los siguientes pasos, que se presentarán con información detallada en las siguientes secciones de este capítulo:

- **PROCESO RECOPILACIÓN DE DATOS**

Los datos empleados en el proyecto fueron recopilados por Bancoomeva, debido a esto, se llevó a cabo un proceso formal de solicitud a la empresa, incluyendo el acuerdo de confidencialidad junto con el área jurídica de la Universidad.

- **COMPRENSIÓN Y LECTURA DE BASES DE DATOS MEDIANTE PYTHON**

Para comprender las bases de datos, fue fundamental solicitar y estudiar los diccionarios de datos que describen cada variable contenida en ellas. Utilizando Python, se lleva a cabo la lectura de la base de datos y se procede a vincular cada variable correspondiente.

- **LIMPIEZA Y PROCESAMIENTO DE DATOS**

Utilizando el entorno de Google Colab, se importaron las librerías necesarias como pandas, numpy y matplotlib. Se identificaron las listas y los tipos de datos que contenían cada una de las variables, inicializándolas para identificar su nombre, tipo de dato, categorías y valores nulos. Luego se depuraron las variables a trabajar según el criterio del conocimiento de mercado.

- **GESTIÓN DE VALORES “NO DEFINIDOS”**

Se identificaron varias variables que contenían el valor "No Definido". Se procedió a reemplazar estos valores con pd.NA (representación de NaN en Pandas) para mejorar la calidad de los datos y facilitar su manipulación en análisis posteriores. Se manejaron los datos faltantes, que representaban un porcentaje mínimo del total, mediante imputación. Para los datos categóricos se utilizó la moda y para los datos numéricos la mediana.

- **ANÁLISIS EXPLORATORIO Y DESCRIPTIVO DE LOS DATOS (EDA)**

- Análisis de correlación entre características numéricas y categóricas, Inicialmente, se realizó un análisis exhaustivo de correlación entre las características numéricas y categóricas del conjunto de datos para identificar y eliminar colinealidades. Esto aseguró que las características utilizadas en el proceso de Clustering fueran independientes y contribuyeran significativamente a la segmentación de los datos.

- Detección de outliers con distancia mahalanobis: Se aplicó la distancia Mahalanobis para detectar outliers en las variables numéricas. Este método estadístico multivariante permitió identificar puntos de datos atípicos que podrían afectar negativamente el desempeño de los algoritmos de clustering.

- Estandarización de variables numéricas y transformación de variables categóricas por OneHotEncoder: Se llevó a cabo la estandarización de las variables numéricas, ajustándolas a una escala comparable con media cero y desviación estándar de uno.

- Además, se transformaron las variables categóricas por OneHotEncoder, es una herramienta específica proporcionada por bibliotecas como scikit-learn en Python para realizar la codificación de variables categóricas en variables dummy de manera más automatizada y eficiente.

- Reducción de dimensionalidad con análisis de componentes principales (PCA): Se utilizó el análisis de componentes principales (PCA) para reducir la dimensionalidad del conjunto de datos sin perder información significativa. Este proceso transformó las características originales en un conjunto de componentes principales que capturaron la mayor variabilidad en los datos, facilitando una representación más eficiente y efectiva para los algoritmos de clustering.

- **APLICACIÓN DE ALGORITMOS DE CLUSTERING Y EVALUACIÓN**

Algoritmos de clustering, como K-Means y MiniBatchKMeans, fueron implementados en el conjunto de datos preprocesado. Estos algoritmos fueron evaluados utilizando métricas como el índice de silueta para determinar la calidad de los clusters generados y seleccionar el número óptimo de clusters mediante métodos como el "codo" en la curva de variación explicada.

- **PERFILES DE CLIENTES**

Basado en la evaluación comparativa de los algoritmos de clustering y la interpretación

de los clusters resultantes, se seleccionó el modelo que mejor se ajustó a los objetivos del estudio. Se generaron perfiles de clientes para cada clúster y se analizaron las características demográficas y transaccionales asociadas a cada segmento. Estos perfiles proporcionan una comprensión profunda de los diferentes segmentos de clientes, permitiendo desarrollar estrategias personalizadas de atención y fidelización.

5.1 LIMPIEZA Y PROCESAMIENTO DE DATOS

La base de datos contiene 57 variables y 496,590 registros de clientes. Se realiza lectura del archivo “DatosModelo.CSV” y se muestran parte de las 5 primeras filas del Dataframe, el cual está representado en la ilustración 1.

1

	FechaFinMes	TotalCreditosActivos	TotalCreditosHistoria	Mora_Maxima_Historica	FechaPrimerCredito	EdadPrimerCredito	SaldoDeudaTotal
0	2024-02-29	1.0	14.0	398.0	2006-07-19	17.0	991342.0
1	2024-02-29	0.0	1.0	52.0	2015-10-14	8.0	0.0
2	2024-02-29	3.0	11.0	37.0	2017-09-04	6.0	1662587.0
3	2024-02-29	1.0	1.0	0.0	2018-10-06	5.0	8100.0
4	2024-02-29	1.0	1.0	12.0	2021-03-11	2.0	2297435.0

Figura 1 Lectura Base Inicial

A continuación, se detallan las 57 variables que serán analizadas para el modelo, con su respectiva descripción:

Nombre Variable	Descripción
TotalCreditosActivos	Número total de créditos activos que posee el cliente en Bancoomeva.
TotalCreditosHistoria	Números de créditos que el cliente ha tenido en su historia crediticia en Bancoomeva.
Mora_Maxima_Historica	Es el mayor período de atraso en días que un cliente ha tenido en el pago de sus créditos a lo largo de su historial crediticio con Bancoomeva.
EdadPrimerCredito	Identifica la antigüedad del cliente con respecto a su historial crediticio en Bancoomeva.
SaldoDeudaTotal	Saldo total de la deuda del cliente. Esta deuda puede estar relacionada con créditos, préstamos u otras obligaciones financieras que el cliente tiene con la entidad bancaria.
ValorCuota	Valor de la cuota de crédito del cliente. Si el

	cliente tiene más de un crédito, el valor de "ValorCuota" sería la suma de todas las cuotas mensuales de esos créditos.
MoraActual	Mora actual del cliente en sus créditos.
STRTIPDOC	Tipo de documento de identificación del cliente.
SEXO	Género del cliente.
EDAD	Edad del Cliente
STRREGION	Región geográfica de residencia del cliente.
STREST_FIN	Estado financiero del cliente (Activo, Inactivo).
PERMEUA	Indicador de si el cliente tiene residencia o permiso de permanencia en Estados Unidos.
ACTEUA	Indicador de si el cliente tiene activos en Estados Unidos.
SALDO	Saldo de la cuenta del cliente. Se refiere al dinero disponible en la cuenta del cliente, que puede usarse para hacer transacciones.
CanalTransaccional	Canal transaccional utilizado por el cliente. Si realiza transacciones físicamente en sucursales ("Presencial") o a través de medios electrónicos y digitales ("Virtual"). Hibrido en caso de utilizar canales virtuales o presenciales.
MontoPromedioTransacciones	Monto promedio de las transacciones del cliente. Se toma en cuenta los últimos doce meses.
NumTransaccionesMensuales	El número de transacciones mensuales del cliente se calcula promediando las realizadas en doce meses.
SaldoPromedioCuenta	Saldo promedio de la cuenta del cliente al final de cada día. Este indicador se obtiene sumando el saldo al final de cada día durante el mes y dividiéndolo por el número de días del mes; el valor resultante se tiene en cuenta para ser promediados en los últimos 12 meses.
IngresosMensuales	Ingresos mensuales o salarios devengados percibidos por el cliente al momento de apertura un producto de ahorro, inversión o consumo.
GastosMensuales	Gastos mensuales del cliente al momento apertura un producto de ahorro, inversión o consumo.
SaldoTarjetaCredito	Saldo de la tarjeta de crédito del cliente.
NumPrestamosActivosDescripción	Número de préstamos activos del cliente.

NumInversiones	Número de inversiones del cliente en los últimos doce meses.
NumProductosFinancieros	Número de productos financieros activos del cliente se refiere a la cantidad total de productos financieros (como cuentas bancarias, tarjetas de crédito, préstamos, inversiones.) que el cliente ha mantenido abiertos y en uso en los últimos doce meses.
Antigüedad	Antigüedad del cliente en la entidad financiera.
BancaMovil	Uso de la banca móvil por parte del cliente NAN: Sin acceso a los servicios digitales o preferencia por otros canales. Durmientes: Clientes que han descargado la aplicación de banca móvil, pero nunca han activado la aplicación para realizar alguna transacción. Inactivos: Clientes que han utilizado la aplicación de banca móvil en el pasado, pero no la utilizan de manera regular ni realizan transacciones recientes. Activos: Clientes que utilizan la aplicación de banca móvil de manera regular para realizar transacciones básicas, como consultas de saldo, transferencias y pagos.
CanalesPrincipales	Los principales canales por parte del cliente, se clasifican en: No tienen: Identifica a clientes que aún no han adquirido o activado servicios bancarios digitales. Inactivos: Identifica a los clientes que han dejado de utilizar los servicios del banco. Durmientes: Identifica a los clientes que han descargado aplicaciones o se han registrado en servicios digitales, pero nunca los han utilizado activamente. Activos Bajos: clientes que utilizan los servicios de manera ocasional o con frecuencia limitada. Activos: clientes que utilizan activamente los servicios del banco, tanto virtuales como presenciales, para una amplia gama de transacciones financieras.
Estrato	Estrato socioeconómico del cliente (01-02-03-04-

	05-06).
EstadoCivil	Estado civil del cliente (Soltero, Casado, Separado, unión Libre).
NivelAcademico	Nivel académico del cliente (Profesional, Técnico, Tecnólogo, Bachiller, Primaria).
Ocupación	Ocupación del cliente (Empleado, Independiente, Pensionado, Informal).
FechaFinMes	Representa el último día de un mes calendario.
Target	Perfil detallado de los clientes ideales a los que se desea atraer, basado en características como ingresos, edad, ubicación geográfica.
Activo	Cliente que mantiene una relación activa y continua con el banco o inactivo cuando deja de realizar transacciones o actividades bancarias durante 12 meses.
Score	Es una evaluación numérica que se asigna a un consumidor o entidad basada en su historial de crédito.
Decil	Forma de dividir una muestra de datos en diez partes iguales.
Grupo de Riesgo	Grupo de riesgo, resultado de modelo interno.
STRTIPCLI	División de los clientes para determinar si es persona natural o persona jurídica.
STROFI_VIN	Oficina donde se vinculó el cliente.
NOMBRE	Nombre del cliente.
APELLIDO1	Apellido del cliente.
SECECO	Código del sector económico al cual pertenece el cliente según su vinculación inicial.
STRSECECO	Nombre del sector económico al cual pertenece el cliente.
STRNATJUR	Indicador de cliente con naturaleza jurídica.
ORIFON	Código del origen de los fondos del cliente.
STRORIFON	Descripción del origen de los fondos del cliente.
NUMSOC	De ser cliente con naturaleza jurídica, indicador del número de socios del cliente.
STRTIPEMP	Definición del tipo de empresa
NCIONA1	Nacionalidad del cliente
COD_BANCA	Código de la banca Personas
DESC_BANCA	Descripción de la banca a la cual pertenece el

	cliente: Personal o Empresarial.
ID CLIENTE	Código único de cliente
FechaPrimerCredito	Fecha en que registra el primer producto del activo tipo crédito.
FECPROC	Ultima fecha de proceso de la información.
EdadCuent	Tiempo desde la creación de su cuenta de ahorros o corriente.
ClienteCredito	Indicador de si el cliente tiene crédito

Tabla 1 Variables Dataset

A través del conocimiento del negocio, no se tuvieron en cuenta 23 variables, las cuales no aportaron para la construcción del modelo, por las siguientes razones:

Nombre Variable	Justificación
FechaFinMes	Fecha de corte igual información 29-02-2024
Target	Variable no representativa
Activo	Se cuenta solo con clientes activos
Score	Variable no representativa
Decil	Variable no representativa
Grupo de Riesgo	Variable no representativa
STRTIPCLI	Personas naturales
STROFI_VIN	Personas naturales
NOMBRE	Datos personales no inciden en el comportamiento del cliente
APELLIDO1	Datos personales no inciden en el comportamiento del cliente
SECECO	No elegimos el Código del sector económico, seleccionamos la variable STRSECECO nombre del sector económico
STRNATJUR	Descripción naturaleza jurídica, todos los clientes son naturales y no empresas.
ORIFON	No elegimos Código de origen de los Fondos del cliente, seleccionamos la variable STRORIFON descripción de origen de fondos
NUMSOC	“Número de Socios”. El cliente es persona natural y no empresas
STRTIPEMP	“Tipo de Empresa”. Personas naturales
NCIONA1	Nacionales Colombianos
COD_BANCA	Se analizará solo la banca personas
DESC_BANCA	“Descripción clasificación banca", todos son Banca Personas
ID CLIENTE	Al eliminar variables como el ID del cliente, se puede enfocar la atención en las variables predictoras que son realmente relevantes para el problema en cuestión.
FECPROC	“Ultima fecha de proceso de la información”. Fecha 29-02-2024

FechaPrimerCredito	Seleccionamos para el modelo la variable EdadPrimerCredito
EdadCuent	Seleccionamos para el modelo la variable EdadPrimerCredito
ClienteCredito	Indicador si el cliente tiene crédito, no se toma en cuenta ya que se identifica que todos los clientes cuentan con al menos un crédito en el Banco.

Tabla 2 Justificación Variables Eliminadas

Se seleccionaron 34 variables consideradas relevantes para abordar el problema específico. A continuación, se detallan las variables seleccionadas:

	Variable	TipoDato	CatUnicos	VectorUnicos	Nulos	PorcentajeNulos
0	TotalCreditosActivos	float64	15	[3.0, 0.0, 1.0, 6.0, 2.0, 4.0, 5.0, 12.0, 16.0...	153580	30.93
1	TotalCreditosHistoria	float64	206	[12.0, 57.0, 48.0, 24.0, 21.0, 31.0, 27.0, 8.0...	153580	30.93
2	Mora_Maxima_Historica	float64	2571	[18.0, 4.0, 30.0, 41.0, 31.0, 34.0, 22.0, 37.0...	153580	30.93
3	EdadPrimerCredito	float64	15	[7.0, nan, 12.0, 4.0, 3.0, 5.0, 14.0, 10.0, 13...	216658	43.63
4	SaldoDeudaTotal	float64	127722	[2372220.0, 0.0, 184079.0, 1179207.0, 309258.0...	153580	30.93
5	ValorCuota	float64	108887	[384477.0, 0.0, 28585.0, 110141.0, 108923.0, 1...	153580	30.93
6	MoraActual	float64	362	[0.0, 15.0, 29.0, 14.0, 1.0, 18.0, 3.0, 25.0, ...	153580	30.93
7	EDAD	float64	496590	[35.649029860181294, 35.09034514924469, 38.558...	0	0.00
8	SEXO	object	2	[MASCULINO, FEMENINO, nan]	153632	30.94
9	STRREGION	object	6	[Palmira, Cali, Caribe, Bogota, Medellin, Eje ...	0	0.00
10	STREST_FIN	object	2	[Activo, Inactivo]	0	0.00
11	PERMEUA	object	2	[No, Si]	0	0.00
12	ACTEUA	object	2	[No, Si]	0	0.00
13	SALDO	float64	496590	[1168864.540365517, 957299.6380046464, 1838292...	0	0.00
14	CanalTransaccional	object	2	[Virtual, Presencial, nan]	49895	10.05
15	MontoPromedioTransacciones	float64	253492	[328051.8207333333, 204171.00503, 106201.26048...	153580	30.93
16	NumTransaccionesMensuales	float64	2519	[12.0, 33.0, 29.0, 32.0, 30.0, 6.0, 52.0, 9.0,...	153580	30.93
17	SaldoPromedioCuenta	float64	246838	[2952466.3866, 2041710.0503, 637207.5629, 1073...	153580	30.93
18	IngresosMensuales	float64	496590	[1722894.7684796853, 2926093.8722217316, 15618...	0	0.00
19	GastosMensuales	float64	496590	[2394027.2240131972, 1955559.1001787712, 17067...	0	0.00
20	SaldoTarjetaCredito	float64	496590	[638857.0751101093, 1383430.7513987957, 185634...	0	0.00
21	NumPrestamosActivos	int64	4	[1, 3, 4, 2]	0	0.00
22	NumInversiones	int64	2	[1, 0]	0	0.00
23	EdadCuenta	float64	496590	[2.0566891016031925, 2.277240352888064, 1.3910...	0	0.00
24	NumProductosFinancieros	int64	4	[2, 3, 4, 1]	0	0.00
25	Antiguedad	object	8	[10-14, 0-4, 5-9, 20-24, 15-19, 25-29, 35-39, ...	0	0.00
26	BancaMovil	object	3	[Activos, Inactivos, nan, Durmientes]	220283	44.36
27	CanalesPrincipales	object	5	[Activos, No Tiene, Inactivos, Durmientes, Act...	0	0.00
28	Estrato	float64	6	[3.0, 2.0, 4.0, 5.0, 1.0, 6.0, nan]	153910	30.99
29	EstadoCivil	object	6	[Soltero, Union Libre, <NA>, Casado, Separado,...	49511	9.97
30	NivelAcademico	object	8	[Profesional, Tecnico, <NA>, Bachiller, Tecnol...	78162	15.74
31	Ocupacion	object	7	[PENSIONADO, INDEPENDIENTE INFORMAL, EMPLEADO,...	153725	30.96
32	STRSECECO	object	5	[Sector Primario, Sector Secundario, Sector Te...	0	0.00
33	STRORIFON	object	7	[Salarios, Bonificaciones, Pensiones, Honorari...	0	0.00

Figura 2 Descripción de Variables Iniciales

5.1.1 Proceso De Imputación Variables Categóricas:

Para las variables categóricas con valores nulos son se optó por utilizar la moda como método de imputación de datos faltantes en estas variables. La moda representa el valor más frecuente en un conjunto de datos, lo que la convierte en una elección adecuada para variables categóricas.

Variables	Nulos iniciales	Nulos finales
SEXO	153632	0.00
STRREGION	0	0.00
STREST_FIN	0	0.00
PERMEUA	0	0.00
ACTEUA	0	0.00
CanalTransaccional	49895	0.00
Antiguedad	0	0.00
BancaMovil	220283	0.00
CanalesPrincipales	0	0.00
Estrato	153910	0.00
EstadoCivil	49511	0.00
NivelAcademico	78162	0.00
Ocupacion	153725	0.00
STRSECECO	0	0.00
STRORIFON	0	0.00

Tabla 3. Imputación de variables categóricas

5.1.2 Proceso De Imputación Variables Numéricas:

Las variables numéricas, también conocidas como variables cuantitativas, son tipos de datos que representan cantidades numéricas y permiten realizar operaciones matemáticas sobre ellas. Estas variables se refieren a medidas o valores numéricos que se pueden medir, contar o calcular.

Se realizó un proceso de imputación utilizando la mediana. Se usa la mediana debido a que no se ve afectada por valores atípicos, lo que la hace más confiable para representar los valores centrales en distribuciones sesgadas.

Variables	Nulos iniciales	Nulos finales
TotalCreditosActivos	153580	0.00
TotalCreditosHistoria	153580	0.00

Mora_Maxima_Historica	153580	0.00
EdadPrimerCredito	216658	0.00
SaldoDeudaTotal	153580	0.00
ValorCuota	153580	0.00
MoraActual	153580	0.00
EDAD	0	0.00
SALDO	0	0.00
MontoPromedioTransacciones	153580	0.00
NumTransaccionesMensuales	153580	0.00
SaldoPromedioCuenta	153580	0.00
IngresosMensuales	0	0.00
GastosMensuales	0	0.00
SaldoTarjetaCredito	0	0.00
NumPrestamosActivos	0	0.00
NumInversiones	0	0.00
EdadCuenta	0	0.00
NumProductosFinancieros	0	0.00

Tabla 4. Imputación de variables numérica

5.2 ANÁLISIS EXPLORATORIO Y DESCRIPTIVO DE LOS DATOS (EDA)

Es un enfoque inicial y esencial en el análisis de datos que permite comprender la estructura, las relaciones y las principales características de los datos antes de proceder a técnicas de modelado o inferencia estadística más complejas.

El EDA es un paso fundamental en cualquier proyecto de análisis de datos o ciencia de datos. Permite a los analistas familiarizarse con los datos, comprender sus peculiaridades y, a menudo, guiar el camino para los análisis subsiguientes.

5.2.1 Análisis Univariado de Variables Numéricas

Se llevó a cabo un análisis detallado de la variable numéricas

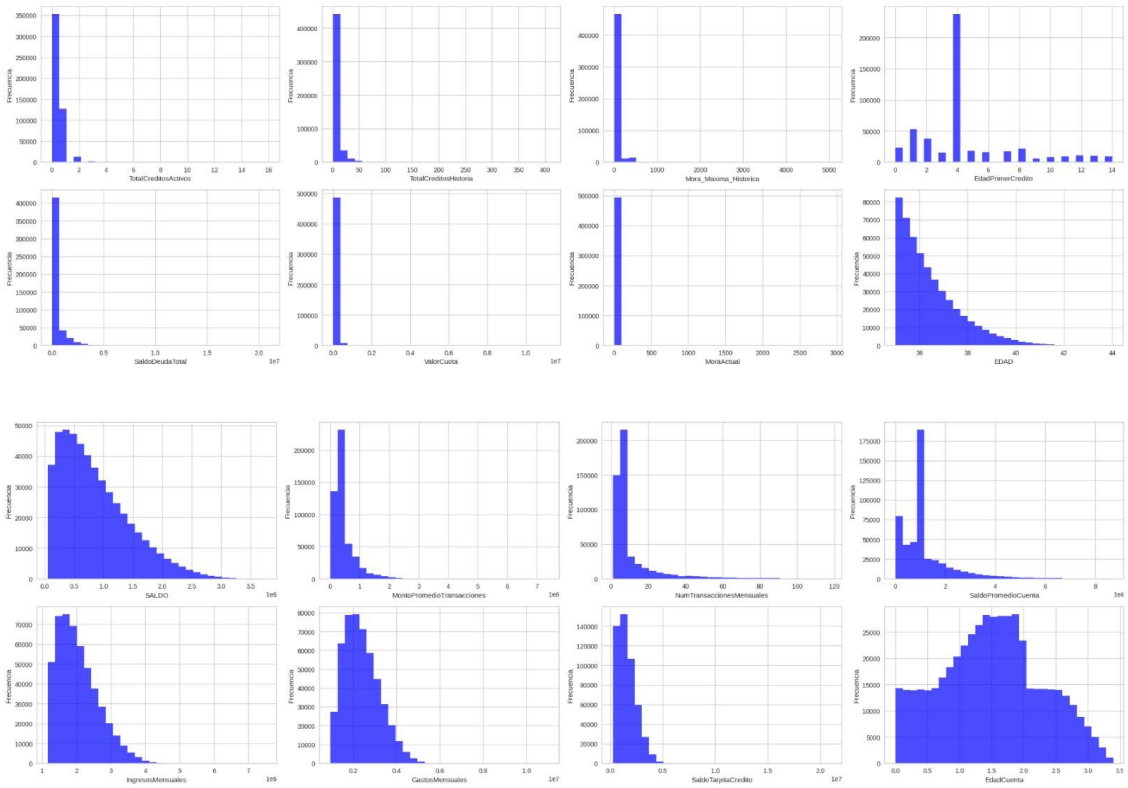


Figura 3.Histogramas Variables Numéricas

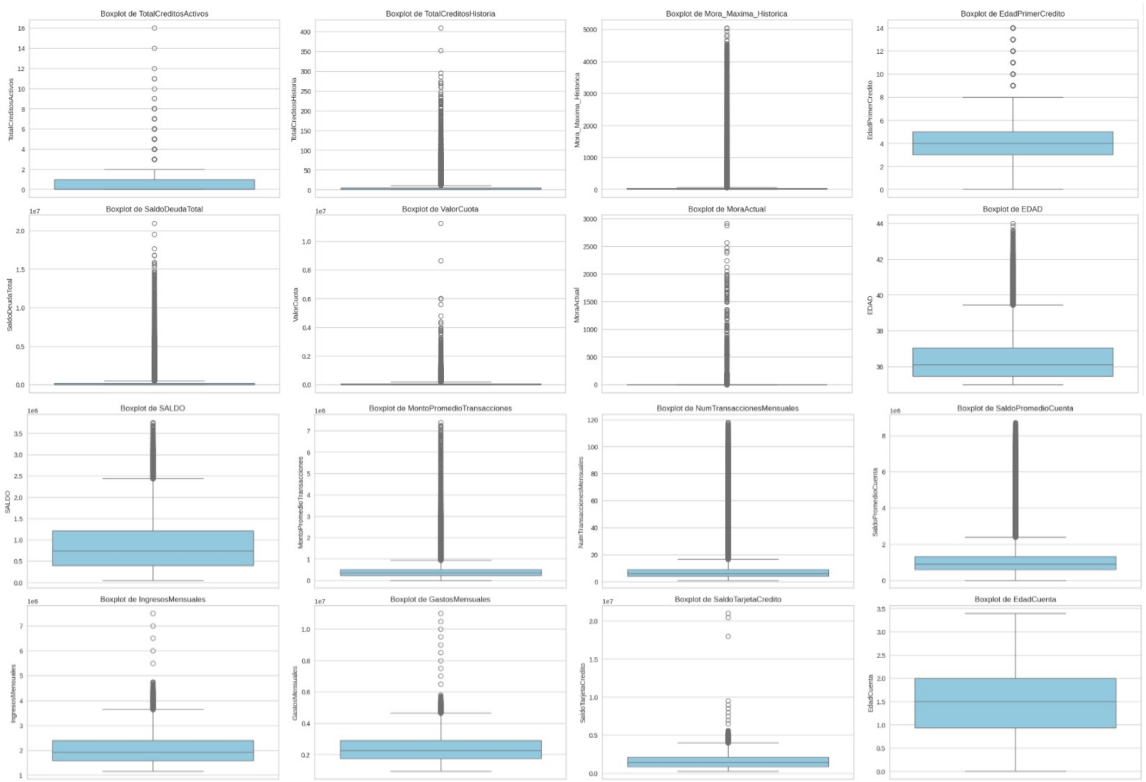


Figura 4. Boxplots Variables Numéricas

El análisis descriptivo de las variables revela aspectos clave sobre el perfil financiero y demográfico de los clientes. La ValorCuota muestra una media de 51,772.04, indicando variabilidad en los pagos mensuales de los clientes, con una desviación estándar de 116,183.01. La MoraActual, con una media de 2.85 y una desviación estándar significativa de 34.74, sugiere que algunos clientes enfrentan niveles elevados de morosidad, llegando hasta 2,920 en el caso máximo.

En cuanto a la edad, la variable EDAD tiene una media de 34.00 años y una desviación estándar de 1.48 años, reflejando una dispersión moderada en las edades de los clientes, con un rango intercuartílico (IQR) de 3 a 5 años. Por otro lado, el SALDO muestra una media de 840,669.38 y una desviación estándar de 541,733.46, indicando amplias variaciones en los saldos de cuenta entre los clientes, desde mínimos de 50,069.57 hasta un máximo de 3,694,534.01.

El MontoPromedioTransacciones tiene una media de 438,488.81 y una dispersión notable con una desviación estándar de 406,349.34, lo que subraya diferencias significativas en los patrones de transacciones financieras de los clientes. La NumTransaccionesMensuales, con un promedio de 10.43 transacciones mensuales, también muestra variabilidad (desviación estándar de 14.42), indicando diferentes niveles

de actividad financiera.

Los ingresos mensuales (IngresosMensuales) muestran una media de 1,152,283.23 y gastos mensuales (GastosMensuales) de 2,258,127.59, reflejando un patrón común de ingresos y gastos entre los clientes. El SaldoTarjetaCredito, con una media de 1,773,172.20 y una desviación estándar de 715,442.34, sugiere variabilidad en los saldos de tarjeta de crédito mantenidos por los clientes. Finalmente, la EdadCuenta tiene una media de 1.50 años, indicando que en promedio, las cuentas tienen relativamente poca antigüedad.

5.2.2 Análisis de Correlación entre Variables Numéricas

Para explorar la relación entre las variables numéricas, primero se realizó un análisis de correlación utilizando el coeficiente de correlación de Pearson. Esta medida proporciona una indicación de la fuerza y la dirección de la relación lineal entre dos variables.

La correlación de Pearson mide la dependencia lineal entre dos variables X e Y. El coeficiente resultante es un valor entre -1 y 1 inclusive, donde:

- 1: correlación lineal positiva total.
- 0: Sin correlación lineal, las dos variables probablemente no se afectan entre sí.
- -1: correlación lineal negativa total

En la matriz de correlación resultante se identificaron relaciones significativas entre varias variables del conjunto de datos. En particular, se observa una alta correlación entre Ingresos Mensuales y Saldo Promedio en Cuenta (0.76), así como entre Gastos Mensuales y Saldo Promedio en Cuenta (0.72). Además, Gastos Mensuales muestra una correlación muy fuerte con Ingresos Mensuales (0.95)

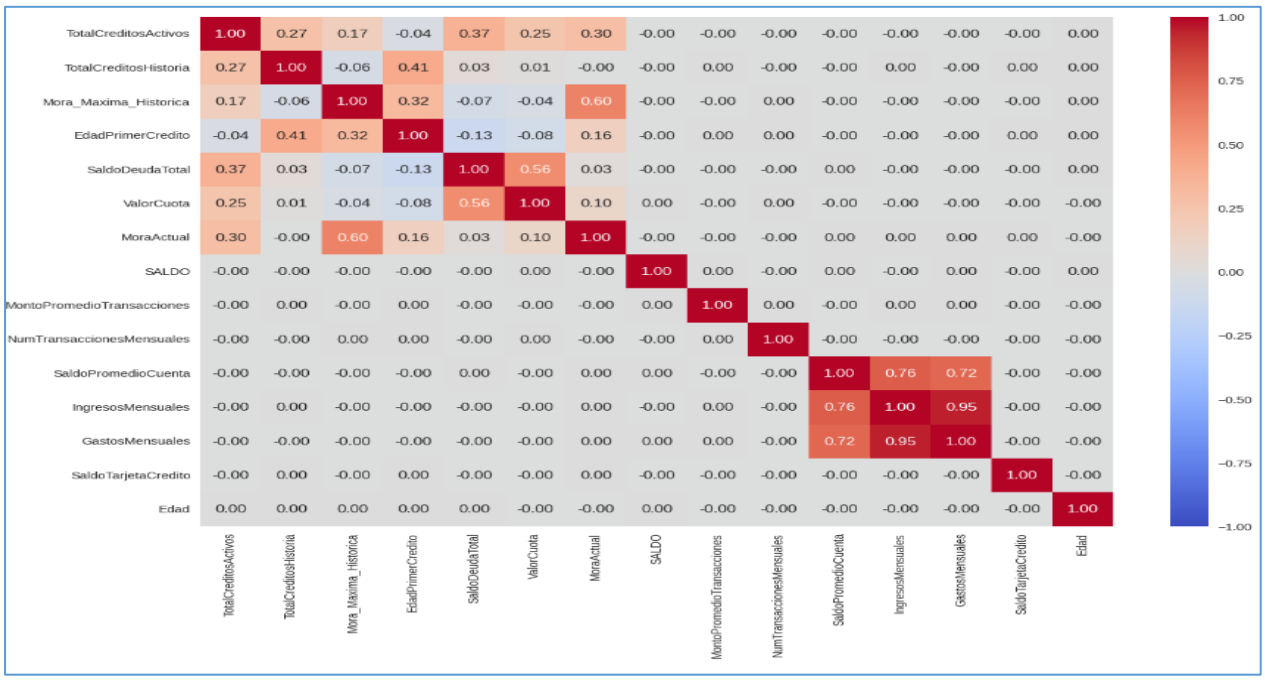


Figura 5. Matriz de Correlación

Estas correlaciones indican la existencia de algunas asociaciones lineales fuertes entre las variables correspondientes. Sin embargo, para evitar la multicolinealidad y garantizar la robustez del análisis, se tomó la decisión de eliminar una de las variables altamente correlacionadas en cada par. Para mitigar este efecto, se seleccionaron las variables 'SaldoPromedioCuenta' y 'GastosMensuales' debido a su coeficiente de correlación.

```
[ ] VariablesNumericas,Eliminadas=utils.remove_collinear_features(VariablesNumericas,0.65)

    IngresosMensuales | SaldoPromedioCuenta | 0.76
    GastosMensuales | SaldoPromedioCuenta | 0.72
    GastosMensuales | IngresosMensuales | 0.95
```

Figura 6. Eliminación Variables Correlacionadas Numéricas

5.2.3 Análisis de independencia entre Variables Categóricas

Para medir la independencia entre variables categóricas se usa el estadístico de Chi-cuadrado. Codificación de variables categóricas: Usa One-Hot Encoding para convertir las variables categóricas en variables dummy.

```
cat_columns
```

```
['ClienteCredito',
 'STRTIPDOC',
 'SEXO',
 'STRREGION',
 'STREST_FIN',
 'PERMEUA',
 'ACTEUA',
 'CanalTransaccional',
 'Antiguedad',
 'BancaMovil',
 'CanalesPrincipales',
 'Estrato',
 'EstadoCivil',
 'NivelAcademico',
 'Ocupacion']
```

Figura 7. Consultas Variables Categóricas

Evaluar la relación entre variables categorías:

handle_unknown='ignore' en OneHotEncoder: Esta opción maneja categorías desconocidas en los datos de prueba, asignándolas a una columna especial en lugar de lanzar un error.

Manejo de la matriz superior de asociación: Usamos np.triu con astype(bool) para asegurarnos de que obtenemos una matriz booleana adecuada.

La transformación One-Hot Encoding de variables categóricas sirve para manejar de manera completa y efectiva las categorías o niveles de una variable, convirtiéndolas en vectores binarios. Cada categoría se representa como una columna distinta, donde el valor binario indica la presencia o ausencia de esa categoría para cada observación. Esto es útil en modelos de machine learning porque permite que el algoritmo interprete correctamente las variables categóricas como características numéricas.

```
[281]  1 # Aplica One-Hot Encoding
       2 from sklearn.preprocessing import OneHotEncoder
       3 encoder = OneHotEncoder(sparse=False, drop='first', handle_unknown='ignore')
       4 encoded_array = encoder.fit_transform(df[categorical_cols])
       5 encoded_df = pd.DataFrame(encoded_array, columns=encoder.get_feature_names_out(categorical_cols))
       6 print(f'Tamaño de la base original: {len(encoded_df.columns)}')
       7 # Añade las columnas codificadas al DataFrame original (opcional)
       8 df_encoded = encoded_df#pd.concat([df, encoded_df], axis=1)
       9
      10 # Calcular la matriz de correlación
      11 corr_matrix = encoded_df.corr().abs()
      12
      13 # Umbral para determinar alta correlación
      14 threshold = 0.8
      15
      16 # Encuentra las columnas que están correlacionadas por encima del umbral
      17 upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool))
      18
      19 # Selecciona las columnas que tienen correlación mayor que el umbral
      20 to_drop = [column for column in upper.columns if any(upper[column] > threshold)]
      21
      22 # Elimina las columnas correlacionadas del DataFrame original
      23 df_reduced = df_encoded.drop(columns=to_drop)
      24
      25 # Opcional: eliminar también las columnas originales categóricas si no son necesarias
      26 #df_reduced = df_reduced.drop(columns=categorical_cols)
      27 print(f'Tamaño de la base reducinada: {len(df_reduced.columns)}')

→ Tamaño de la base original: 53
  Tamaño de la base reducinada: 53
```

Figura 8. Aplicación One-Hot Encoding

Después de aplicar el proceso de eliminación de variables categóricas con fuerte asociación, se evidenció que no existían asociaciones significativas entre las variables categóricas. Por lo tanto, no se eliminó ninguna variable categórica.

5.2.4 Detección de Outliers con Distancia Mahalanobis para las variables numéricas.

Para asegurar la calidad de los datos y mejorar la precisión del análisis de clustering, se llevó a cabo la detección de valores atípicos utilizando la distancia Mahalanobis. Este método es ideal para identificar outliers multidimensionales debido a su capacidad de considerar la correlación entre variables.

Se incluyeron todas las variables numéricas de la base de datos en este análisis: 'TotalCreditosActivos', 'TotalCreditosHistoria', 'Mora_Maxima_Historica', 'EdadPrimerCredito', 'SaldoDeudaTotal', 'ValorCuota', 'MoraActual', 'EDAD', 'SALDO', 'MontoPromedioTransacciones', 'NumTransaccionesMensuales', 'SaldoPromedioCuenta', 'IngresosMensuales' y 'SaldoTarjetaCredito' y se estableció un nivel de significancia de 0.10 y se calcularon los grados de libertad, que corresponden al número de variables.

El umbral para identificar los outliers se determinó usando la distribución chi-cuadrado. Los puntos de datos cuyas distancias Mahalanobis excedían este umbral fueron

identificados como outliers. Finalmente, se generaron esta pareja de gráficos como ejemplo de dispersión antes y después de la detección de outliers. En el primer gráfico, todos los datos se representaron en azul, mientras que, en el segundo gráfico, los outliers identificados se resaltaron en rojo, permitiendo una visualización clara de los valores atípicos detectados.

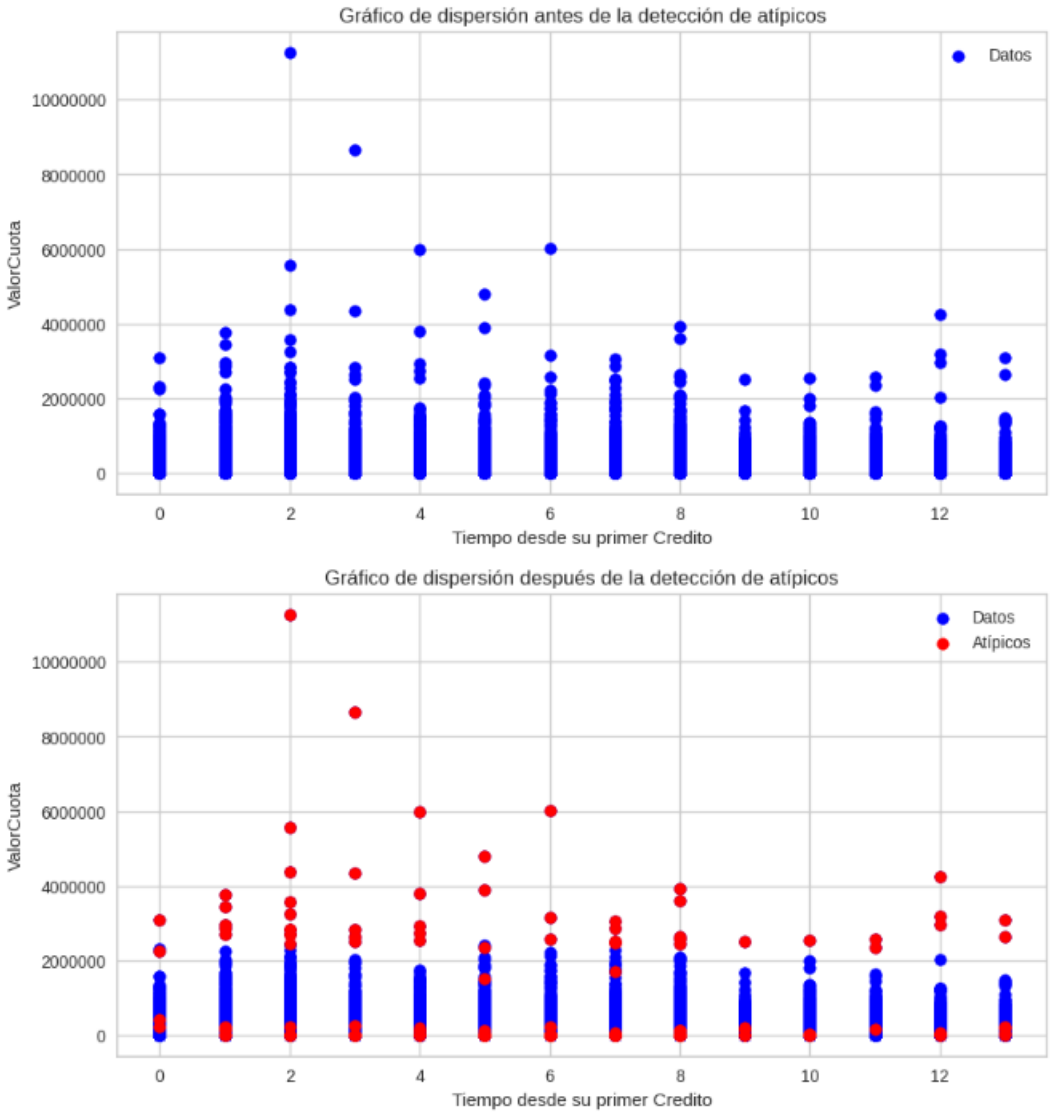


Figura 9. Aplicación Identificación Valores Atípicos por Distancia Mahalanobis

En nuestro análisis, identificamos que aproximadamente 1,046 clientes presentaban características atípicas, representando menos del 1% del total de la muestra. Decidimos excluir estos datos debido a su baja representación en el conjunto de datos, lo que minimiza su impacto en la generalización de nuestros resultados.

Además, nuestra decisión de excluir los outliers está justificada por el uso de algoritmos como K-means para la segmentación de clientes. Los algoritmos de clustering, como K-means, son especialmente sensibles a los valores atípicos, ya que pueden distorsionar la formación de los clusters y afectar la precisión del modelo. Los outliers pueden provocar que los centroides se desplacen, lo que resulta en una segmentación menos precisa y representativa del comportamiento general de los clientes.

Por estas razones, consideramos que la exclusión de los clientes atípicos es una medida necesaria para mejorar la fiabilidad y precisión de nuestros análisis y modelos.

5.2.5 Estandarización de Variables Numéricas

La estandarización de datos desempeña un papel crucial en el análisis de datos y el machine learning al asegurar que todas las características tengan una escala comparable. Este proceso es fundamental porque muchos algoritmos de machine learning, como PCA (Análisis de Componentes Principales) y K-means, son sensibles a la escala y la varianza de las características. La elección entre StandardScaler y otros métodos de escalado, como MinMaxScaler, depende de las necesidades específicas del problema y las características de los datos.

StandardScaler es preferido en este caso, ya que se adapta bien a algoritmos que requieren que todas las características estén en la misma escala. Esto es particularmente útil en algoritmos como PCA, donde la magnitud de las características puede afectar significativamente la variabilidad capturada por los componentes principales.

```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
# Ajustar y transformar las variables numéricas
VariablesNumericas_escaladas = scaler.fit_transform(df_Original[num_columns])
```

```
# Convertir las variables numéricas escaladas a DataFrame con nombres de columnas
VariablesNumericas_escaladas_df = pd.DataFrame(VariablesNumericas_escaladas, columns=num_columns)
VariablesNumericas_escaladas_df
```

	TotalCreditosActivos	TotalCreditosHistoria	Mora_Maxima_Historica	EdadPrimerCredito	SaldoDeudaTotal	ValorCuota	MoraActual	EDAD	SALDO	MontoPromedioTransacciones	NumTransaccionesMensuales	SaldoPromedioCuenta	IngresosMensuales	SaldoTarjetaCredito
0	4.62	0.59	-0.20	0.79	2.73	3.13	-0.10	-0.83	-1.19	-0.27	0.11	1.66	-0.43	-1.19
1	-0.57	4.88	-0.29	-0.17	-0.42	-0.48	-0.10	1.04	0.07	-0.58	1.57	0.82	-1.61	-0.65
2	1.16	4.03	-0.12	-0.17	-0.18	-0.21	-0.10	1.70	-0.10	-0.82	1.29	-0.48	-0.65	1.34
3	9.80	1.74	-0.05	2.39	1.15	0.56	-0.10	0.27	0.04	-0.92	1.50	-0.07	-0.94	-1.71
4	2.89	1.45	-0.12	-0.17	-0.01	0.54	0.72	0.16	0.62	-0.65	1.36	0.57	1.16	0.31
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495539	-0.57	-0.36	-0.21	-0.17	-0.42	-0.48	-0.10	0.49	-1.37	-0.23	-0.31	-0.25	-0.49	-0.31
495540	-0.57	-0.36	-0.21	-0.17	-0.42	-0.48	-0.10	0.05	0.39	-0.23	-0.31	-0.25	-0.35	-0.44
495541	-0.57	-0.36	-0.21	-0.17	-0.42	-0.48	-0.10	1.65	-0.90	-0.23	-0.31	-0.25	-0.71	1.02
495542	-0.57	-0.36	-0.21	-0.17	-0.42	-0.48	-0.10	1.54	-1.47	-0.23	-0.31	-0.25	1.61	-0.40
495543	-0.57	-0.36	-0.21	-0.17	-0.42	-0.48	-0.10	0.05	0.28	-0.23	-0.31	-0.25	1.69	0.48

495544 rows × 14 columns

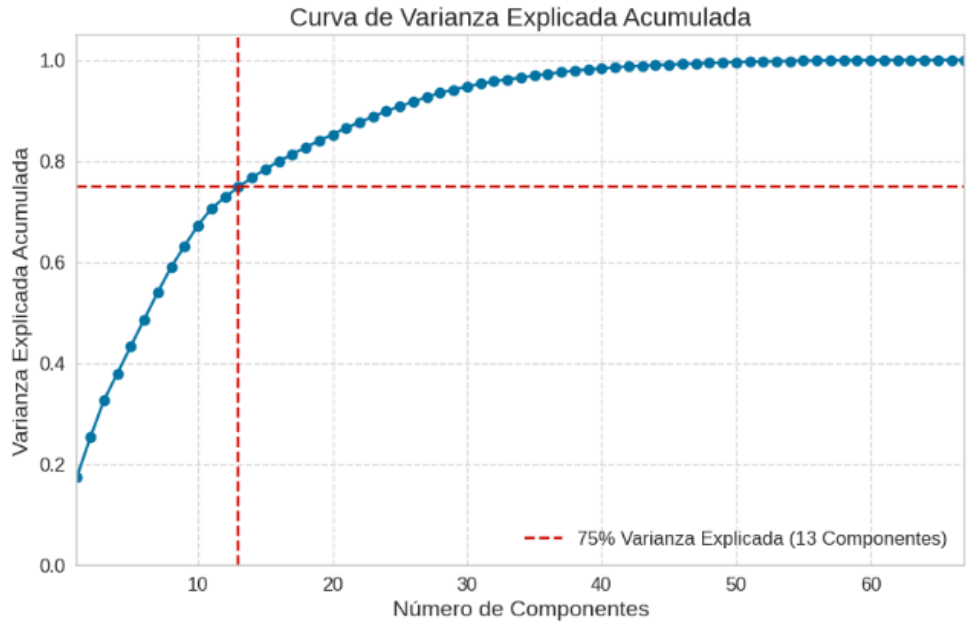
Figura 10. Estandarización de Variables Numéricas

5.3 SELECCIÓN DE VARIABLES PARA EL MODELO DE SEGMENTACIÓN

Después de la estandarización de las variables numéricas y la codificación de las variables categóricas, nuestro conjunto de datos resultante contiene un total de 67 variables. Estas variables representan tanto aspectos cuantitativos como cualitativos de nuestros datos, preparadas de manera óptima para el siguiente paso en nuestro análisis, que es la selección de características.

5.3.1 Reducción De Dimensionalidad Mediante El Método De Análisis De Componentes Principales (PCA)

La reducción de dimensionalidad mediante el método de Análisis de Componentes Principales (PCA) es una técnica fundamental en el análisis de datos y el aprendizaje automático. PCA busca transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas llamadas componentes principales, que capturan la máxima varianza en los datos. Al seleccionar un número menor de componentes principales que retengan la mayor cantidad de varianza explicada, podemos reducir la complejidad del conjunto de datos original sin perder demasiada información importante. Esta técnica es particularmente útil en conjuntos de datos con muchas variables, mejorando la eficiencia computacional y facilitando la visualización y la interpretación de los datos transformados en aplicaciones posteriores de modelado predictivo o clustering.



Número de componentes para explicar al menos el 75.0% de la varianza: 13

Figura 11. Varianza Explicada por las Componentes Principales

La elección de utilizar 13 componentes principales para explicar al menos el 75.0% de la varianza se fundamenta en el análisis de la curva de varianza explicada acumulada obtenida mediante PCA.

En el gráfico mostrado, se puede observar claramente cómo la varianza explicada acumulada aumenta con el número de componentes. Se estableció un umbral del 75.0% de varianza explicada para determinar cuántos componentes eran necesarios.

A partir del análisis gráfico, se identifica que con 13 componentes principales se logra alcanzar o superar este umbral, lo cual permite conservar una cantidad sustancial de la información original del conjunto de datos mientras se reduce significativamente la dimensionalidad. Esta decisión se basa en encontrar un equilibrio entre la reducción de la complejidad del modelo y la retención de la información relevante para aplicaciones de modelado por clustering.

5.3.2 Análisis de las variables para el modelo

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
0	5.63	0.57	-0.63	-1.14	-0.24	-1.53	0.02	0.32	0.64	-2.78	-1.01	-0.29	0.22
1	0.76	2.50	-2.78	1.46	-0.16	-1.40	0.19	0.19	2.68	2.38	1.02	-0.00	0.30
2	0.96	2.62	-2.21	1.68	1.26	0.30	0.82	-0.21	2.83	0.75	0.18	-0.48	-0.42
3	5.64	3.35	-1.80	0.40	-1.16	-1.53	0.06	0.02	1.40	-3.51	-1.65	-1.30	0.79
4	2.43	1.66	-0.81	-0.07	-0.37	1.26	0.21	-0.98	1.20	-0.38	0.34	-1.21	-0.46
...	...	...	...	...	...	...	...	...	...	...	...	...	...
495539	-1.05	-0.30	0.01	-0.17	0.56	-1.12	0.95	-0.16	-0.27	0.03	0.19	0.52	0.92
495540	-1.05	-0.31	0.02	0.34	-0.45	-0.31	-0.23	-0.22	-0.27	0.05	0.18	0.48	0.20
495541	-1.05	-0.30	0.01	1.19	1.44	-0.27	1.25	-0.12	-0.28	0.04	0.19	0.46	-0.47
495542	-1.05	-0.30	0.00	-0.59	-0.09	0.35	2.61	-0.21	-0.27	0.03	0.19	0.50	0.22
495543	-1.05	-0.32	0.02	-0.59	-0.22	1.58	0.48	-0.28	-0.25	0.04	0.19	0.48	-0.54

495544 rows × 13 columns

Figura 12. Componentes Principales Para el Modelado

Los datos mostrados en la tabla anterior representan la proyección de las observaciones originales en el espacio reducido de componentes principales (PC1 a PC13) después de aplicar PCA. Cada fila corresponde a una observación transformada, donde cada valor en las columnas PC1 a PC13 representa la coordenada de esa observación a lo largo de cada componente principal. Estas coordenadas son los nuevos atributos que capturan la mayor parte de la variabilidad en los datos originales, reduciendo así la dimensionalidad mientras se conserva la información más relevante para el análisis posterior o la modelización.

Estos componentes principales son combinaciones lineales de las variables originales y se ordenan según la cantidad de varianza explicada por cada uno. Por ejemplo, PC1 captura la mayor parte de la varianza, seguido por PC2 y así sucesivamente. Visualizar estas coordenadas puede ayudar a entender cómo se distribuyen las observaciones en el espacio transformado y cómo se relacionan entre sí bajo las nuevas dimensiones derivadas de PCA.

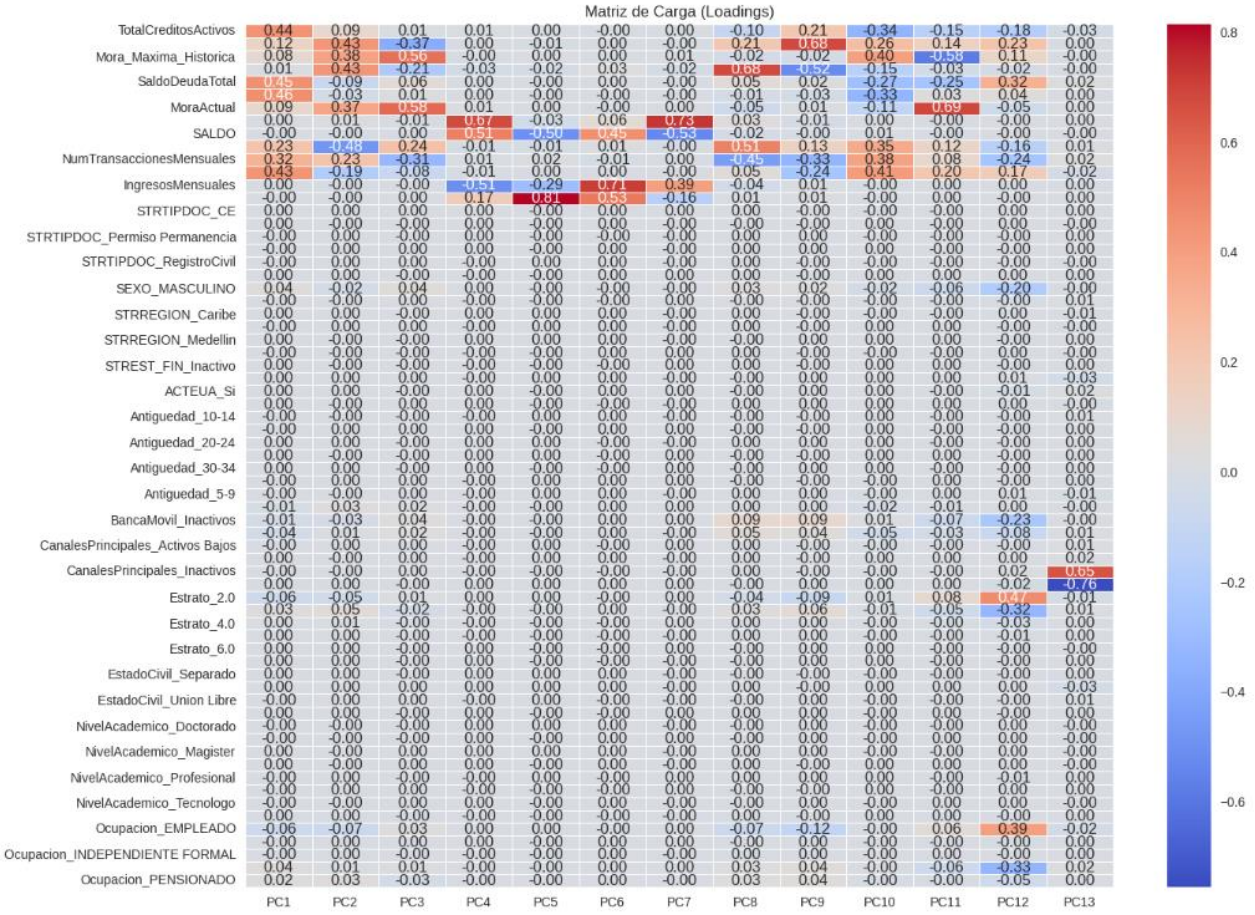


Figura 13. Coeficiente de Cada Variable

La ilustración 15 representan las cargas o coeficientes de cada variable original en cada componente principal (PC) obtenido después de aplicar PCA. Las cargas indican la contribución relativa de cada variable original a la definición de cada componente principal. Aquí algunas observaciones sobre los resultados:

- Interpretación de las cargas: Las cargas más altas en magnitud (ya sea positivas o negativas) indican que esas variables tienen una mayor influencia en la definición de ese componente principal. Por ejemplo, en PC1, las variables con las mayores cargas positivas son "TotalCreditosActivos" y "SaldoDeudaTotal", lo que sugiere

que PC1 está influenciado significativamente por la cantidad total de créditos activos y la deuda total.

- Relaciones entre Variables y Componentes: Al observar las cargas, podemos inferir cómo las variables están relacionadas entre sí bajo la estructura de los componentes principales. Por ejemplo, en PC2, "TotalCreditosHistoria" tiene la carga más alta, seguida por "EdadPrimerCredito". Esto podría indicar una relación donde la cantidad total de créditos históricos y el tiempo transcurrido desde su primer crédito están relacionadas de alguna manera que se captura en PC2.

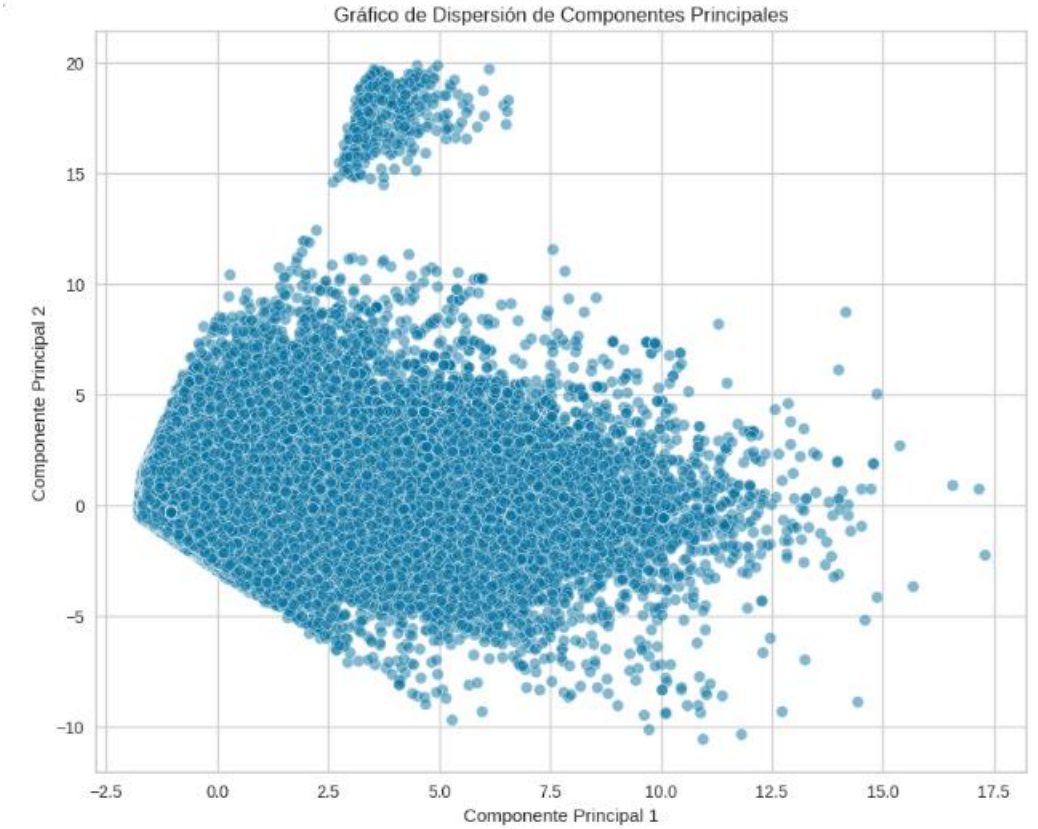


Figura 14. Distribución de Clientes por los Primeros Dos Componentes Principales

Al visualizar las coordenadas de las observaciones en el espacio reducido de los componentes principales (como PC1 y PC2), podemos observar cómo las observaciones se agrupan o separan claramente en conjuntos distintos y bien definidos.

Esta separación puede ser muy beneficiosa para tareas como la segmentación de clientes o la detección de patrones, ya que permite identificar grupos homogéneos de clientes o patrones específicos en los datos. La clara diferenciación entre los grupos en el espacio de los componentes principales sugiere que estos componentes capturan características subyacentes significativas de los datos originales, facilitando así la

interpretación y la toma de decisiones basadas en estos análisis dimensionales.

5.4 APLICACIÓN DE MODELOS

Para la aplicación de los algoritmos K-means y MiniBatchKMeans, se procede con el objetivo de segmentar los datos transformados por PCA en grupos o clusters significativos. Inicialmente, se utiliza el método del codo para determinar el número óptimo de clusters en ambos algoritmos. Este método evalúa la inercia o la suma de las distancias cuadradas dentro de cada clúster para diferentes valores de k, buscando el punto donde la inercia comienza a estabilizarse. Una vez determinado el número de clusters, se procede a evaluar la calidad de la agrupación utilizando el índice de silueta. Este índice mide cuán similar es un objeto a su propio clúster en comparación con otros clusters, proporcionando una medida de cuán bien separados están los clusters.

5.4.1 Aplicación del K-Means

El siguiente paso en nuestro análisis implica la aplicación del modelo K-Means utilizando las 13 componentes principales obtenidas mediante PCA

Selección de cantidad de grupos para la segmentación: Método del codo

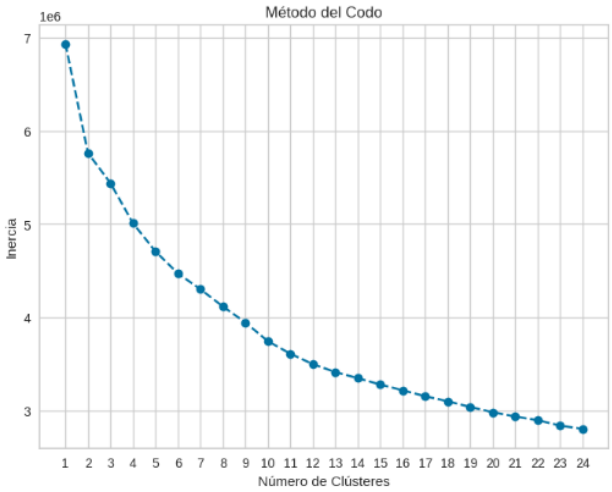


Figura 15. Método del Codo

Para determinar el número óptimo de grupos para la segmentación de clientes, utilizamos el método del codo con el algoritmo K-Means. Este método implica ajustar K-Means a nuestros datos para diferentes valores de k y calcular la suma de las distancias al cuadrado de cada punto al centroide más cercano. Graficamos estas sumas en función de k y observamos el punto donde la reducción en la suma de distancias se vuelve más

gradual, formando un "codo" en el gráfico. Este punto nos indica el número óptimo de grupos, ya que añadir más no mejoraría significativamente la segmentación.

Según el método del codo, el número óptimo de grupos para nuestra segmentación de clientes es 4. Este resultado se determinó al identificar el punto en el gráfico donde la reducción en la suma de distancias al cuadrado se hizo más lenta, indicando que k=4 proporciona una buena división de los datos sin necesidad de más grupos. Por lo tanto, procedemos con la creación de 4 grupos para nuestro modelo de segmentación de clientes.

5.4.1.1 Entrenamiento del Modelo

```
# Entrenar el modelo de K-Means con 4 clusters
kmeans = KMeans(n_clusters=N_culster, random_state=42)
clusters = kmeans.fit_predict(data_scaled)

# Agregar columna de clusters al DataFrame original
data['Cluster'] = clusters

# Mostrar el conteo de observaciones por cluster
print(data['Cluster'].value_counts())
```

Figura 16. Entrenamiento

Entrenamos el modelo K-Means utilizando 4 clusters con la implementación de scikit-learn, utilizando una semilla aleatoria para garantizar la reproducibilidad de los resultados. Luego, asignamos a cada cliente su cluster correspondiente y añadimos esta información como una nueva columna al DataFrame original. Al revisar la distribución de observaciones por segmento, obtuvimos una visión inicial de cómo se agrupan los clientes según sus características.

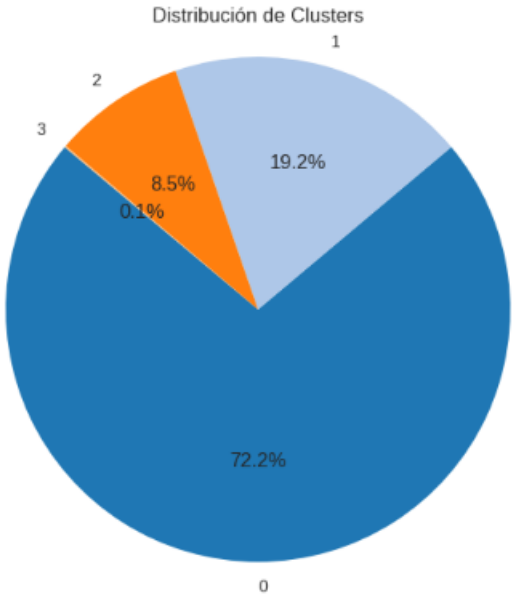


Figura 17. Distribución de los Segmentos

Después de aplicar el modelo K-Means con 4 clusters a nuestros datos, observamos la siguiente distribución de clientes en cada segmento: el Cluster 0 contiene 357,918 clientes, el Cluster 1 tiene 95,112 clientes, el Cluster 2 cuenta con 42,036 clientes, y finalmente, el Cluster 3 está compuesto por 478 clientes. Esto representa aproximadamente el 72% de los clientes en el Cluster 0, el 19% en el Cluster 1, el 8% en el Cluster 2, y el 0% en el Cluster 3.

5.4.2 Aplicación del MiniBatchKMeans

El método MiniBatchKMeans es una variante eficiente del algoritmo de agrupamiento K-Means tradicional, diseñado para manejar grandes conjuntos de datos de manera más rápida sin sacrificar demasiada precisión en la segmentación. A diferencia de K-Means estándar, que utiliza el conjunto de datos completo en cada iteración para actualizar los centroides de los clusters, MiniBatchKMeans utiliza pequeños lotes de muestras de datos seleccionadas al azar (de ahí su nombre "MiniBatch") para realizar actualizaciones más frecuentes de los centroides. Esto permite una convergencia más rápida del algoritmo, lo cual es especialmente beneficioso en conjuntos de datos grandes o en aplicaciones donde la velocidad de procesamiento es crucial.

Selección de cantidad de grupos para la segmentación: Método del codo

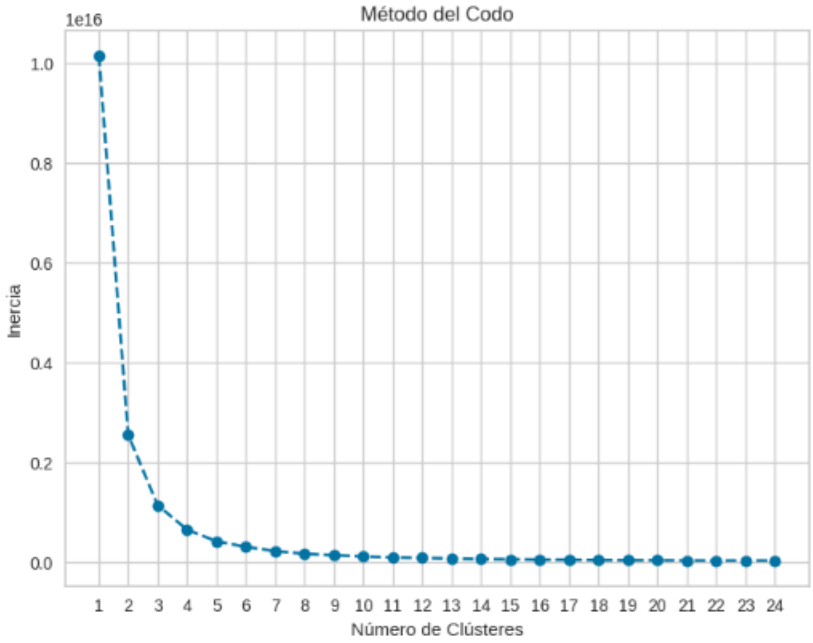


Figura 18. Método del Codo

Para el algoritmo MiniBatchKMeans, identificamos que el número óptimo de clusters utilizando el método del codo fue 3. Este método implica ajustar el algoritmo MiniBatchKMeans a nuestro conjunto de datos para diferentes valores de k (número de grupos) y luego calcular la suma de las distancias al cuadrado de cada punto al centroide más cercano. Al graficar estos valores en función de k, observamos un punto donde la disminución de la distancia al cuadrado se hace significativamente más lenta, lo cual indica que añadir más grupos no proporcionaría una mejora sustancial en la segmentación. Por lo tanto, decidimos proceder con la creación de 3 grupos para nuestro modelo de segmentación de clientes utilizando MiniBatchKMeans.

5.4.2.1 Entrenamiento del Modelo

```
num_clusters=3
```

```
mbkmeans = MiniBatchKMeans(n_clusters=num_clusters, batch_size=100, random_state=42)
mbkmeans.fit(data_final_pca_df)
labels = mbkmeans.labels_
```

```
# Agregar columna de clusters al DataFrame original
df_Original['Cluster']=labels
```

Figura 19 Entrenamiento

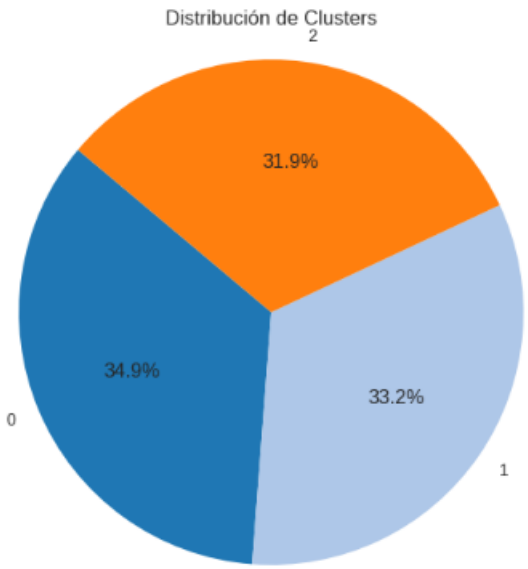


Figura 20. Distribución de Clúster por MiniBatchKMeans

Al ejecutar el algoritmo MiniBatchKMeans tenemos que: el Cluster 0 tiene 172,942 clientes, el Cluster 1 tiene 164,431 clientes y el Cluster 2 tiene 158,171 clientes. En términos de porcentaje, esto se traduce en que el Cluster 0 abarca el 35% de los clientes, el Cluster 1 el 33%, y el Cluster 2 el 32%. Esto revela una distribución bastante equilibrada entre los grupos identificados por el algoritmo MiniBatchKMeans.

5.5 EVALUACIÓN DE LOS MODELOS POR ÍNDICE DE LA SILUETA

En la evaluación del modelo, utilizaremos el índice de silueta para medir la cohesión y separación de los segmentos. El índice de silueta proporciona una medida de cuán similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación). Este valor oscila entre -1 y 1, donde un valor cercano a 1 indica que el objeto está bien clasificado en su propio grupo, mientras que un valor cercano a -1 indica que podría estar mejor ubicado en un grupo diferente. Una puntuación global alta del índice de silueta sugiere una buena calidad de agrupamiento.

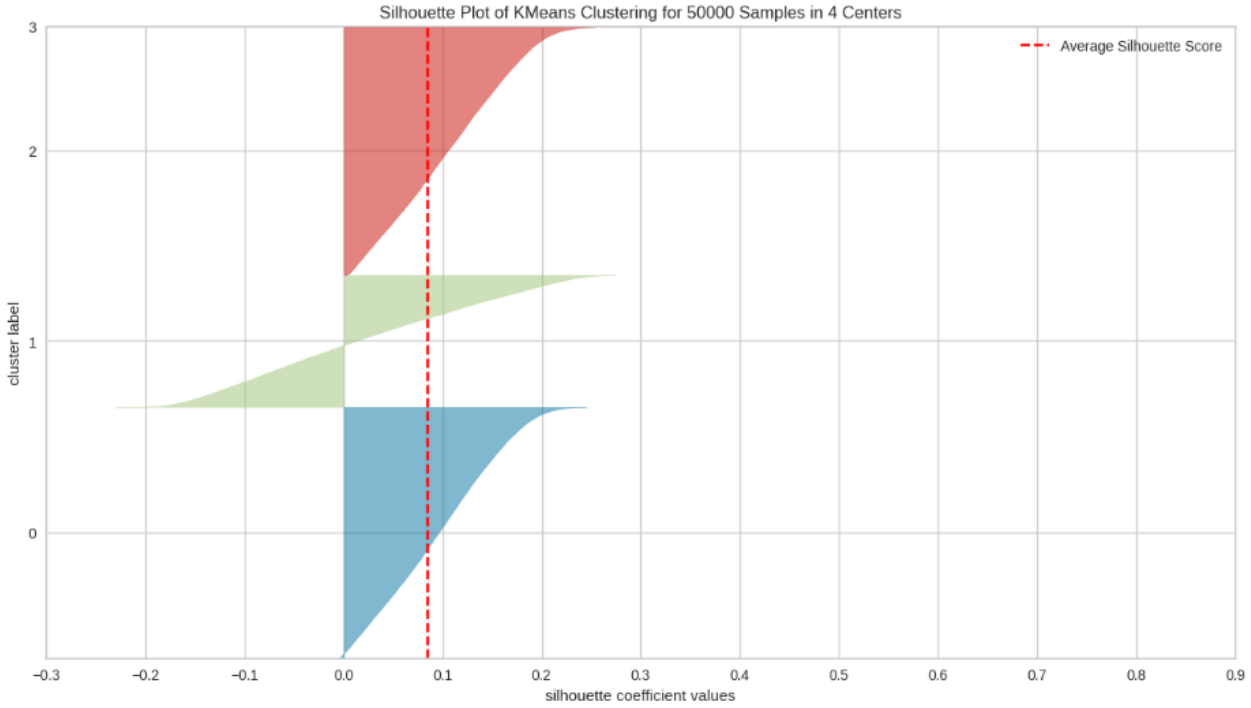


Figura 21. Índice de la Silueta Para K Means

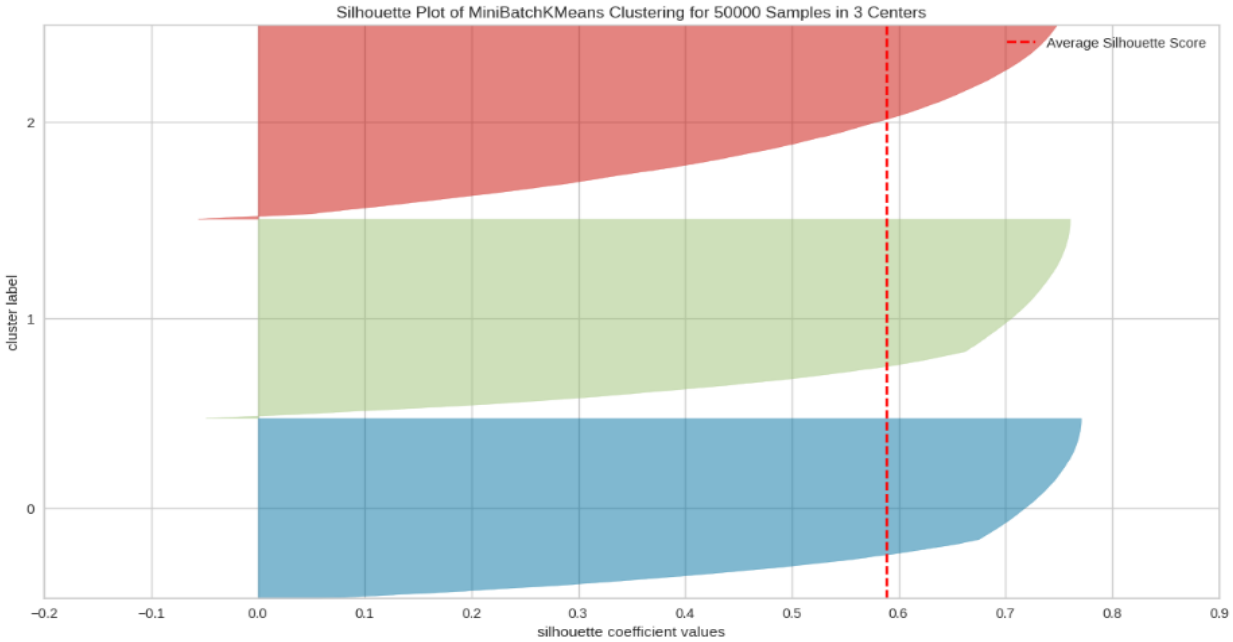


Figura 22. Índice de la silueta Para MiniBatchKMeans

Basado en los resultados obtenidos, podemos observar que el algoritmo MiniBatchKMeans muestra un índice de silueta significativamente más alto (0.589) en comparación con K-Means (0.085). Este resultado indica que MiniBatchKMeans logra una mejor cohesión intra-cluster y una separación inter-cluster más clara en nuestros datos. Además, el hecho de que MiniBatchKMeans haya identificado tres clusters frente a los cuatro de K-Means puede sugerir una estructura de datos más clara y menos dispersa, lo cual es favorable para tareas de segmentación como la nuestra. Por lo tanto, considerando el índice de la silueta como métrica de evaluación, podemos concluir que MiniBatchKMeans es el modelo preferido para nuestra aplicación de segmentación de clientes en este caso específico.

6 CATEGORIZACIÓN DE USUARIOS DIGITALES

En el contexto del análisis de perfiles de clientes, la aplicación de técnicas de clustering como MiniBatchKMeans ha demostrado ser fundamental para agrupar clientes en segmentos homogéneos basados en sus características financieras y de comportamiento. Este modelo ha sido seleccionado debido a su capacidad para identificar patrones claros y distintivos entre grupos de clientes, facilitando la personalización de estrategias de marketing y servicios. En este análisis, exploraremos cómo MiniBatchKMeans nos permite definir perfiles de clientes, mejorando nuestra capacidad para ofrecer experiencias personalizadas que maximicen la satisfacción del cliente y optimicen el rendimiento del negocio bancario.

Vamos a analizar la distribución de los cluster por las dos componentes principales, las dos principales componentes obtenidas del análisis PCA proporcionan una visión detallada de cómo diferentes variables contribuyen a la variabilidad de los datos. La primera componente principal (Componente 0) muestra una fuerte asociación positiva con variables como ValorCuota, TotalCreditosActivos, SaldoDeudaTotal, y NumTransaccionesMensuales, indicando una actividad financiera activa con créditos, deudas y transacciones frecuentes. Por otro lado, esta componente tiene una asociación negativa con variables como Sexo femenino y Empleados, sugiriendo posibles diferencias en términos de género y estatus laboral en el comportamiento financiero observado.

En contraste, la segunda componente principal (Componente 1) exhibe influencias positivas significativas de variables como Durmientes e Inactivos en banca móvil, Ocupación informal y Pensionados. Al mismo tiempo, muestra una influencia negativa de variables como Sexo femenino, Activos en banca móvil, NumTransaccionesMensuales y SaldoPromedioCuenta. Esto sugiere que esta componente captura diferencias demográficas y patrones de uso bancario, destacando roles ocupacionales específicos y preferencias en el uso de servicios financieros móviles.

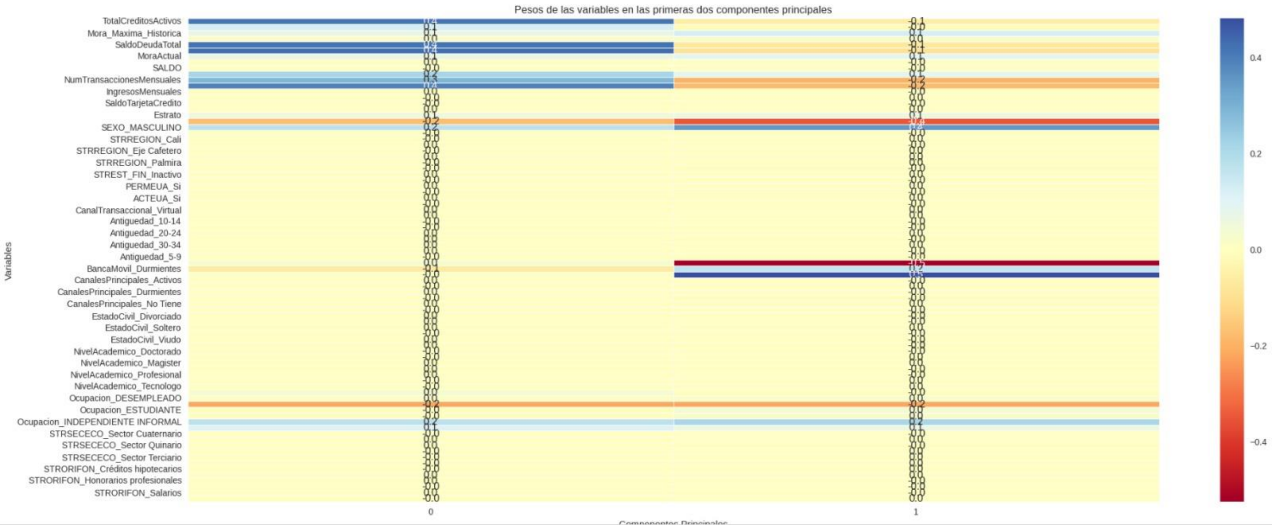


Figura 23. Pesos de las variables en las primeras dos componentes principales

En el gráfico de dispersión de los datos proyectados en las dos primeras componentes principales, se identifican tres centroides distintos que representan los clusters obtenidos mediante el algoritmo MiniBatchKMeans. Cada centroide indica el centro de un grupo de clientes que comparten características similares basadas en las variables originales transformadas por PCA.

Cluster 0: El primer centroide se encuentra en la dirección positiva de la primera componente principal y negativa en la segunda. Este cluster se caracteriza por valores elevados en variables como TotalCreditosActivos, SaldoDeudaTotal, ValorCuota, y MontoPromedioTransacciones. Además, muestra una menor influencia de variables como SEXO_FEMENINO y BancaMovil_Activos.

Cluster 1: El segundo centroide está en la dirección negativa de la primera componente principal y positiva en la segunda. Este grupo exhibe altos valores en NumTransaccionesMensuales, SaldoPromedioCuenta, SEXO_MASCULINO, y Ocupacion_INDEPENDIENTE INFORMAL, lo cual sugiere perfiles demográficos y comportamientos financieros diferentes en comparación con el Cluster 0.

Cluster 2: El tercer centroide se sitúa en la dirección negativa de ambas componentes principales. Este cluster podría estar asociado con valores más bajos en variables como Mora_Maxima_Historica, MoraActual, y BancaMovil_Inactivos, indicando un comportamiento financiero potencialmente más conservador o estable.

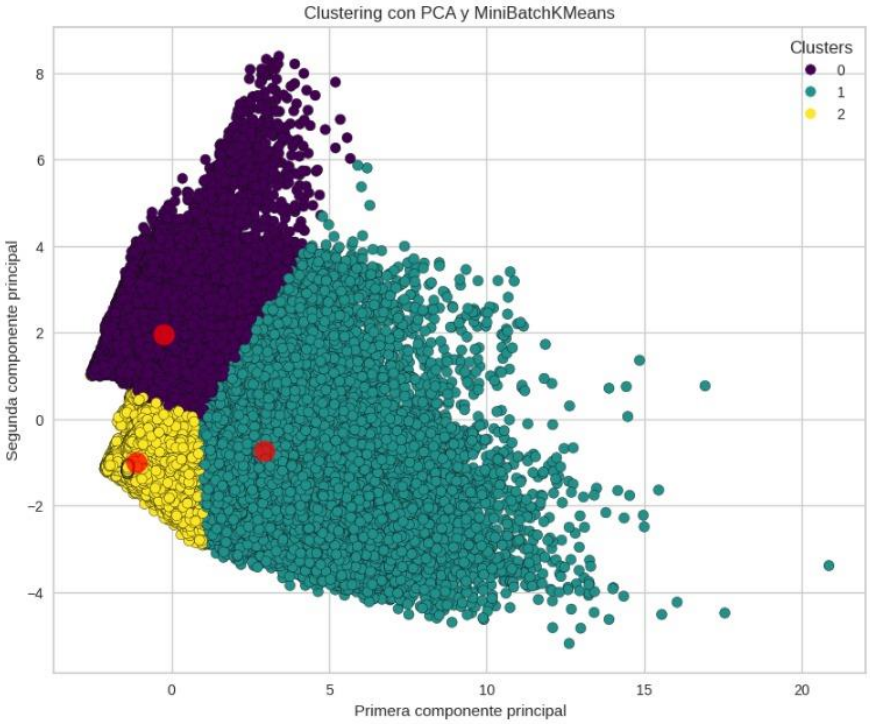


Figura 24. Grafico de Segmentos por MiniBatchKMeans

El análisis de las variables a través de gráficos de Boxplot, agrupadas por clúster, revela varias tendencias importantes. En cuanto a TotalCreditosActivos, observamos que el clúster 0 presenta una mediana baja y una dispersión mínima, lo que indica que la mayoría de los individuos en este grupo tienen pocos créditos activos. Similarmente, los clústeres 1 y 2 también muestran una baja mediana y dispersión, reflejando comportamientos consistentes en esta métrica.

Para SaldoDeudaTotal, los Boxplots indican que el clúster 0 tiene una mediana significativamente alta y una amplia dispersión, con numerosos valores atípicos, sugiriendo una variabilidad considerable en el endeudamiento de este grupo. El clúster 1 muestra una mediana mucho más baja y menor dispersión, mientras que el clúster 2 tiene una mediana intermedia y una variabilidad notablemente alta, aunque no tanto como el clúster 0.

En ValorCuota, los Boxplots revelan que el clúster 0 tiene la mayor mediana y una gran dispersión, nuevamente con muchos valores atípicos, lo que indica altos pagos de cuotas con variabilidad significativa. Los clústeres 1 y 2 tienen medianas y dispersiones más bajas, indicando pagos de cuota más pequeños y menos variables.

Los Boxplots de Mora_Maxima_Historica muestran que el clúster 2 tiene la mediana más

alta y la mayor dispersión, sugiriendo que este grupo experimenta una morosidad histórica máxima considerablemente más alta y variable en comparación con los otros clústeres. El clúster 0 también presenta una mediana alta, aunque con menor dispersión que el clúster 2.

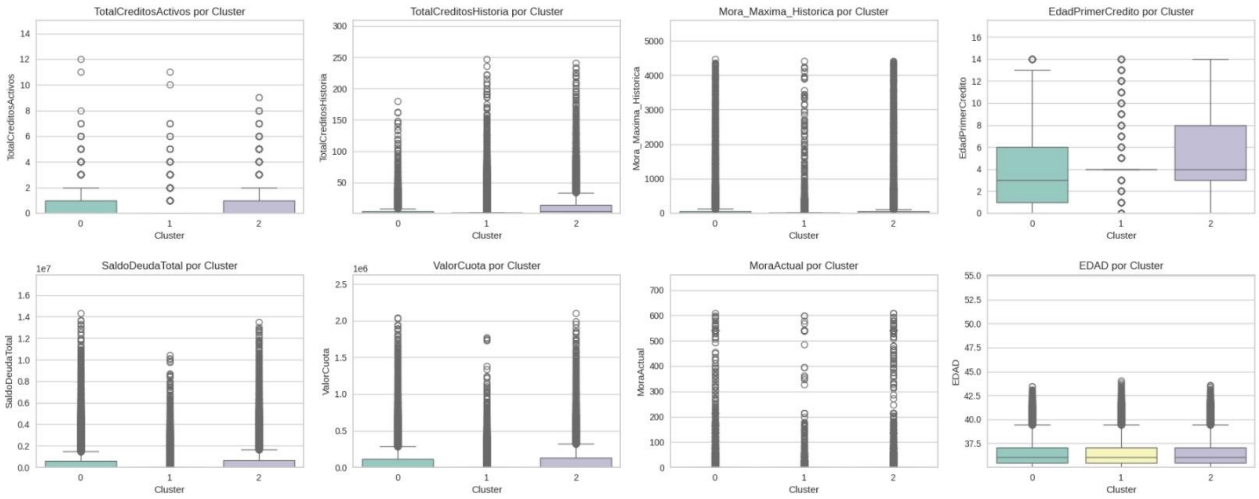
La variable EDAD muestra una mediana consistente alrededor de 36 años en todos los clústeres, con baja dispersión, indicando que la antigüedad de las cuentas es homogénea entre los grupos.

Los ingresos y gastos mensuales (IngresosMensuales y GastosMensuales) presentan Boxplots que indican que el clúster 0 tiene las medianas más altas y la mayor dispersión, sugiriendo capacidades económicas superiores, pero con gran variabilidad. Los clústeres 1 y 2 muestran medianas significativamente más bajas y menor dispersión, indicando comportamientos financieros más consistentes.

Finalmente, los Boxplots de SaldoTarjetaCredito revelan que el clúster 0 tiene la mediana más alta y una dispersión considerable, reflejando saldos altos en tarjetas de crédito con variabilidad significativa. Los clústeres 1 y 2 tienen medianas y dispersiones más bajas, sugiriendo saldos de tarjeta de crédito más pequeños y menos variables.

	SaldoDeudaTotal				ValorCuota				MontoPromedioTransacciones				Mora_Maxima_Historica			
	min	max	mean	median	min	max	mean	median	min	max	mean	median	min	max	mean	median
Cluster																
0	0.00	13688977.00	440980.38	0.00	0.00	1772140.00	69901.04	0.00	250.00	7159600.84	517389.60	401278.68	0.00	3482.00	72.70	16.00
1	0.00	13463364.00	476889.24	0.00	0.00	1987424.00	78399.60	0.00	216.09	7391805.04	444151.26	294034.61	0.00	3492.00	52.91	20.00
2	0.00	8531372.00	12249.85	0.00	0.00	1754760.00	2006.95	0.00	893.36	6827764.71	347800.72	345630.56	0.00	3326.00	17.86	17.00

Figura 25. Estadísticas Descriptivas por Variables Crediticias



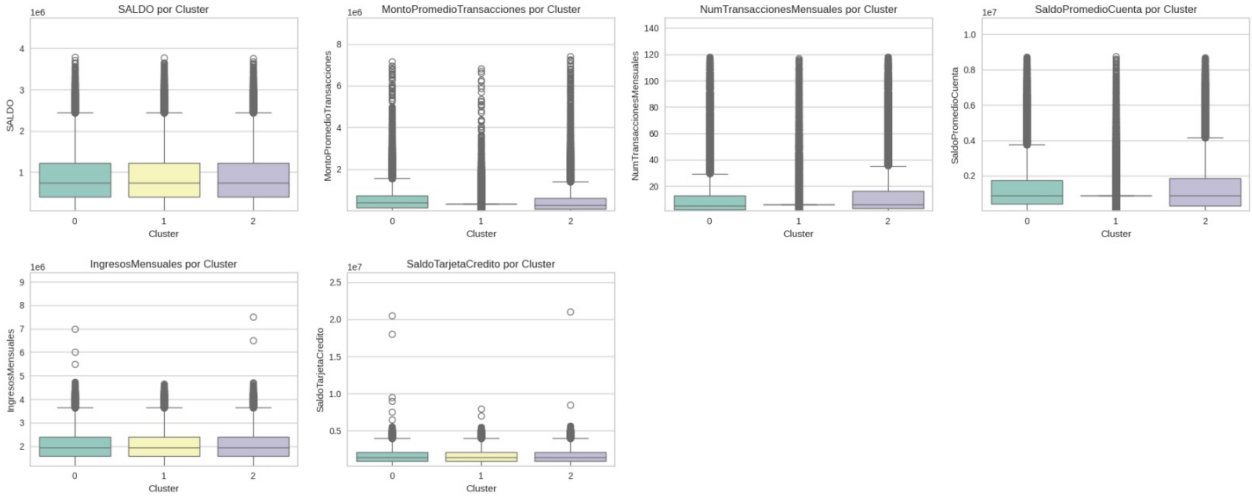
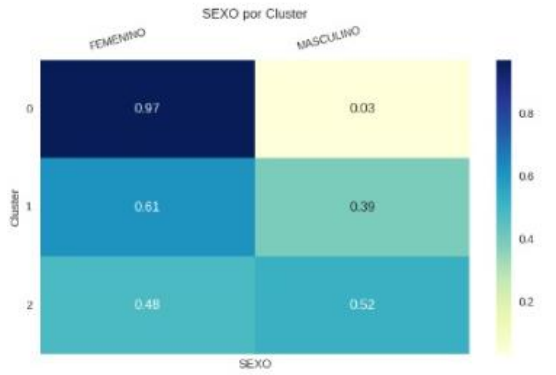
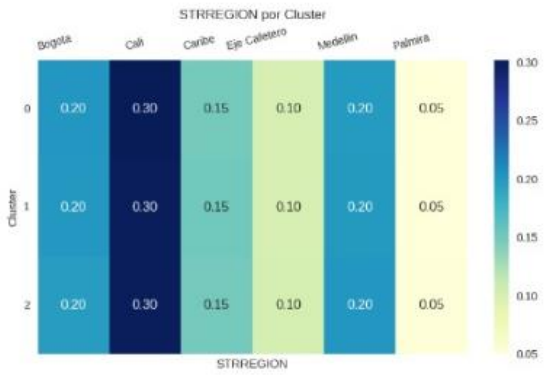


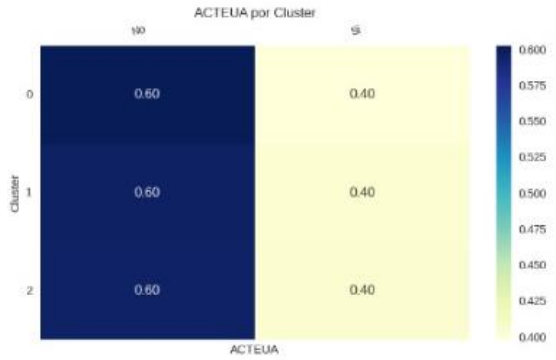
Figura 26. Boxplot de Variables Numéricas por Segmentos

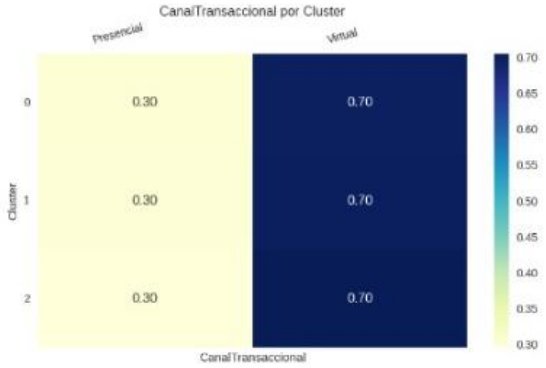
6.1 CATEGORIZACIÓN

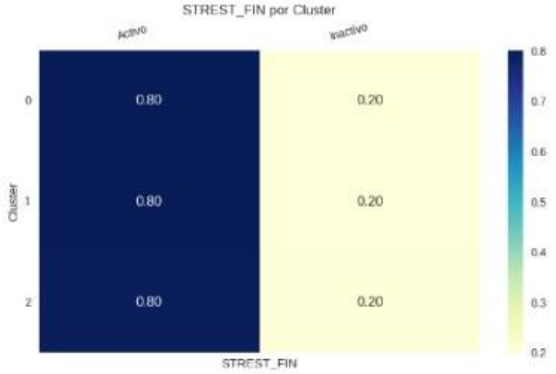
A continuación, se describen brevemente los perfiles de cada segmento según análisis de variables Sociodemográficas para Bancomeva.

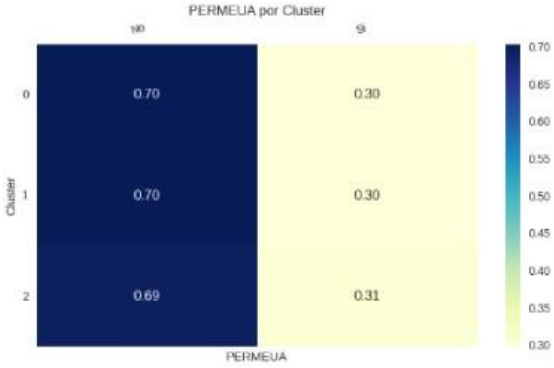

SEXO por Cluster
FEMENINO
MASCULINO
0.97
0.03
0.61
0.39
0.48
0.52
Cluster
SEXO

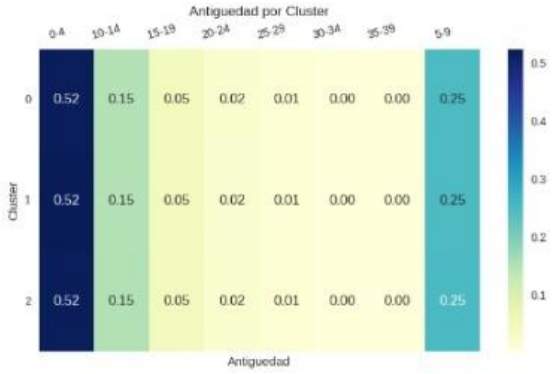

STRREGION por Cluster
Bogotá
Cali
Caribe
Eje Cafetero
Medellín
Palmira
0.20 0.30 0.15 0.10 0.20 0.05
0.20 0.30 0.15 0.10 0.20 0.05
0.20 0.30 0.15 0.10 0.20 0.05
Cluster
STRREGION

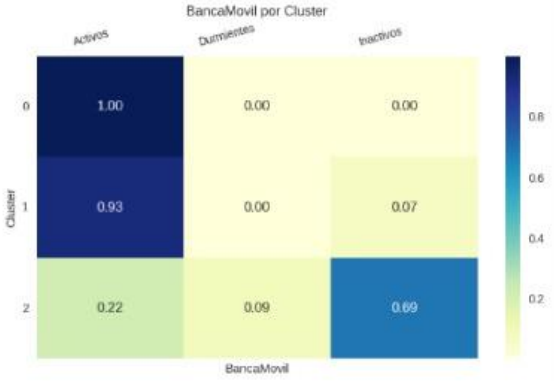

ACTEUA por Cluster
No
Si
0.60 0.40
0.60 0.40
0.60 0.40
Cluster
ACTEUA

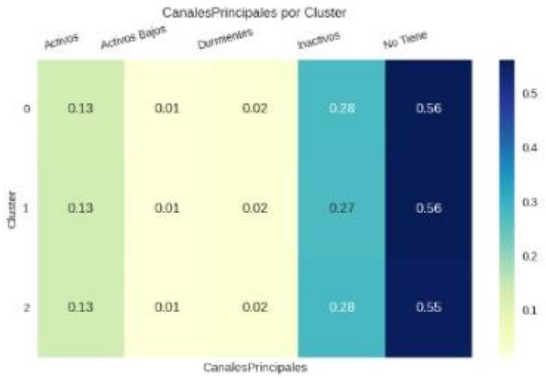

CanalTransaccional por Cluster
Presencial
Virtual
0.30 0.70
0.30 0.70
0.30 0.70
Cluster
CanalTransaccional

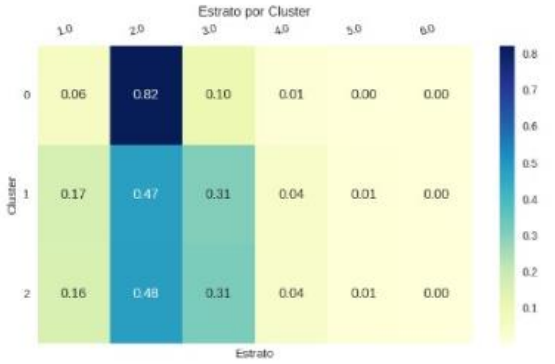

STREST_FIN por Cluster
Activo
Inactivo
0.80 0.20
0.80 0.20
0.80 0.20
Cluster
STREST_FIN

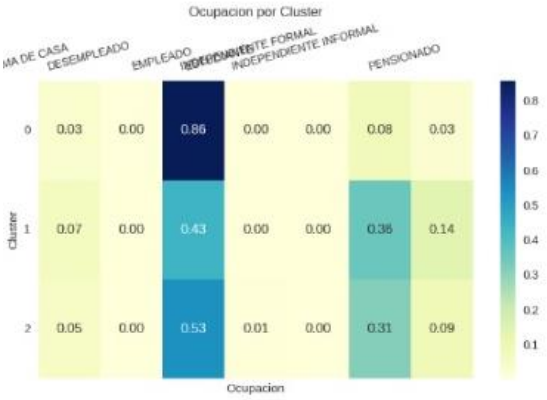

PERMEUA por Cluster
No
Si
0.70 0.30
0.70 0.30
0.69 0.31
Cluster
PERMEUA

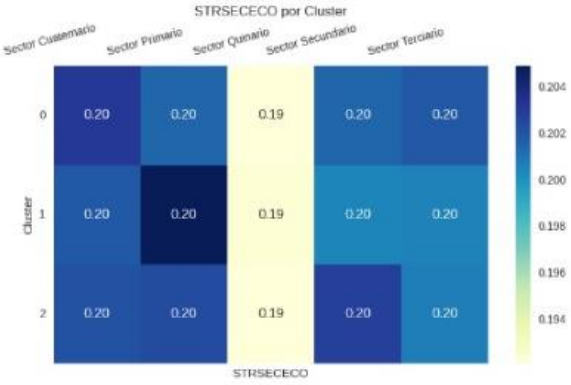

Antiguedad por Cluster
0-4
10-14
15-19
20-24
25-29
30-34
35-39
5-9
0.52 0.15 0.05 0.02 0.01 0.00 0.00 0.25
0.52 0.15 0.05 0.02 0.01 0.00 0.00 0.25
0.52 0.15 0.05 0.02 0.01 0.00 0.00 0.25
Cluster
Antiguedad


BancaMovil por Cluster
Activos
Dormientes
Inactivos
1.00 0.00 0.00
0.93 0.00 0.07
0.22 0.09 0.69
Cluster
BancaMovil


CanalesPrincipales por Cluster
Activos
Activos Bajos
Dormientes
Inactivos
No Tiene
Cluster
0
1
2
0.13 0.01 0.02 0.28 0.56
0.13 0.01 0.02 0.27 0.56
0.13 0.01 0.02 0.28 0.55
CanalesPrincipales
0.5
0.4
0.3
0.2
0.1


Estrato por Cluster
1.0
2.0
3.0
4.0
5.0
6.0
Cluster
0
1
2
0.06 0.82 0.10 0.01 0.00 0.00
0.17 0.47 0.31 0.04 0.01 0.00
0.16 0.48 0.31 0.04 0.01 0.00
Estrato
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Ocupacion por Cluster
DESEMPLEADO
EMPLEADO
INDEPENDIENTE FORMAL
INDEPENDIENTE INFORMAL
PENSIONADO
Cluster
0
1
2
0.03 0.00 0.86 0.00 0.00 0.08 0.03
0.07 0.00 0.43 0.00 0.00 0.38 0.14
0.05 0.00 0.53 0.01 0.00 0.31 0.09
Ocupacion
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


STRSECECO por Cluster
Sector Cuaternario
Sector Primario
Sector Quinario
Sector Secundario
Sector Terciario
Cluster
0
1
2
0.20 0.20 0.19 0.20 0.20
0.20 0.20 0.19 0.20 0.20
0.20 0.20 0.19 0.20 0.20
STRSECECO
0.204
0.202
0.200
0.198
0.196
0.194

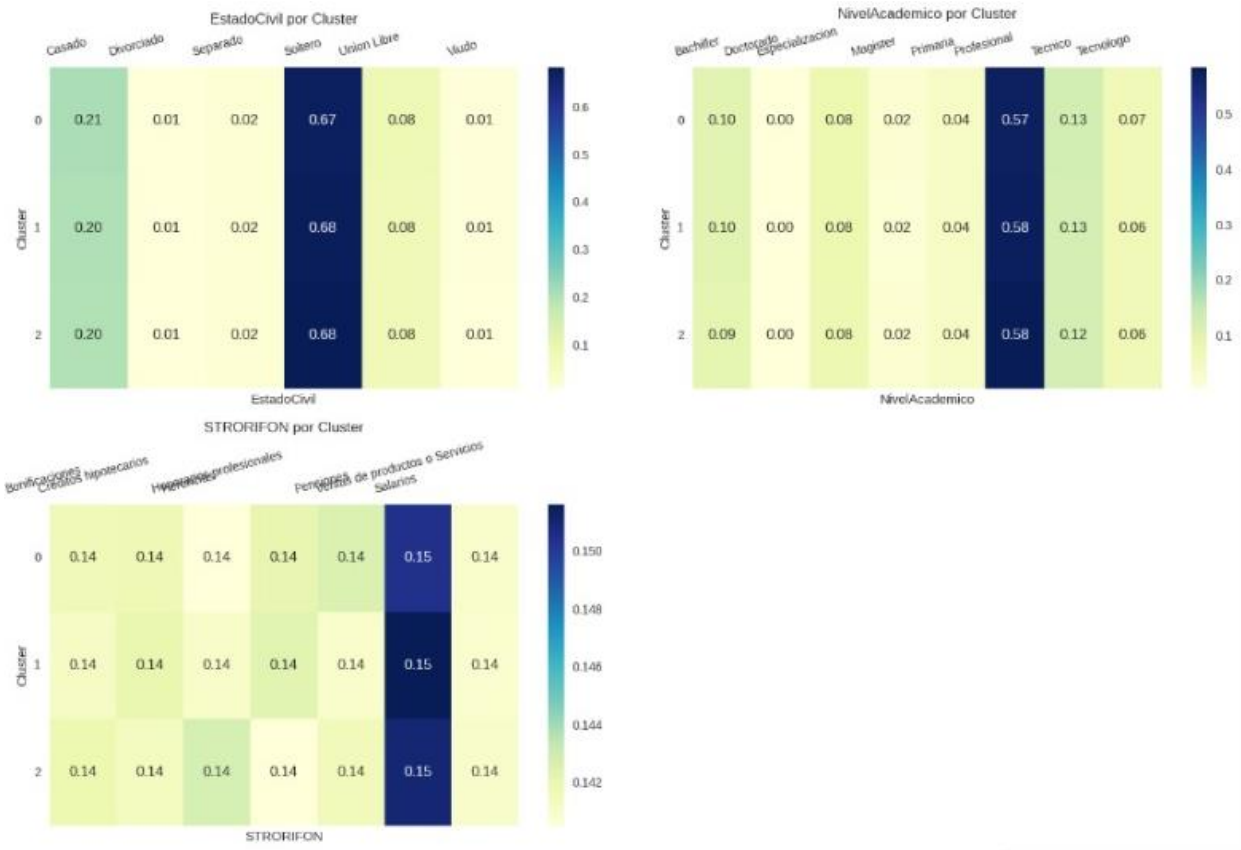


Figura 27. Frecuencia de Variables Sociodemográfica por Segmentos

El análisis revela segmentos distintos de clientes a través de los clusters. Tanto el Cluster 0 como el Cluster 1 muestran tasas moderadas a altas de adopción de banca móvil entre clientes activos, con un 40% y un 46%, respectivamente. Estos grupos también tienen una proporción notable de clientes inactivos (35% y 31%, respectivamente), lo que sugiere oportunidades para estrategias de reactivación. El Cluster 1 se destaca por tener un porcentaje más alto de clientes que no tienen acceso a banca móvil (18%), lo que señala posibles barreras para la participación digital.

En contraste, el Cluster 2 muestra una adopción excepcional de banca móvil, con casi todos los clientes activos (98%) utilizando estos servicios. Además, este cluster tiene una cantidad mínima de clientes inactivos (1%) y una proporción insignificante de clientes sin acceso a banca móvil (1%), lo que indica un fuerte compromiso digital y una audiencia receptiva para ofertas avanzadas de banca digital. Estos insights pueden guiar iniciativas de marketing dirigidas y mejoras en servicios para satisfacer las necesidades únicas y comportamientos de cada cluster, optimizando así la participación y satisfacción de los clientes en los servicios de banca móvil.

En cuanto al estrato socioeconómico, el Cluster 0 presenta una distribución donde predominan los clientes de estratos 2 y 3, con un 49% y 28% respectivamente. Los estratos más altos (4, 5 y 6) tienen una representación mucho menor en este grupo. Por otro lado, el Cluster 1 muestra una distribución similar, aunque con una ligera diferencia en la proporción de estratos más bajos (1 y 2) en comparación con el Cluster 0. En el Cluster 2, casi todos los clientes pertenecen al estrato 2, con una mínima presencia en otros estratos.

Con relación a la ocupación de los clientes en los diferentes clusters, podemos observar que el Cluster 0 está compuesto principalmente por empleados (67%) y también incluye una proporción considerable de personas en ocupaciones informales (29%). En contraste, el Cluster 1 muestra una distribución más diversa, con empleados representando el 36%, seguidos por trabajadores informales (34%) y amas de casa (11%). Por último, el Cluster 2 está dominado por empleados (97%), con una pequeña presencia de pensionados (2%) y una mínima proporción de amas de casa (1%).

Los datos muestran una clara variación en la distribución por género en tres clusters distintos. En el primer cluster, aproximadamente el 57% son mujeres y el 43% son hombres. El segundo cluster presenta una mayoría más marcada de mujeres, con un 68%, mientras que los hombres constituyen el 32%. Por otro lado, el tercer cluster exhibe una discrepancia significativa, donde el 99% de los individuos son mujeres y solo el 1% son hombres. Estas cifras revelan patrones distintivos en la composición por género dentro de cada cluster, evidenciando diferentes perfiles demográficos o comportamentales entre grupos identificados en el análisis.

6.2 PERFILES

Cluster 0: Clientes Oro

Características Principales: Los clientes en el Cluster 0 presentan un saldo de deuda total alto, con una media de 440,980, y realizan transacciones elevadas, con una media de monto promedio de transacciones de 517,389. La mora máxima histórica es baja, con una media de 72 días. Predominan los empleados, con un total del 67% de los clientes en esta categoría. Además, hay una notable presencia de independientes informales (29%) y una pequeña proporción de amas de casa (2%), estudiantes (1%) y pensionados (1%). En términos de género, el 57% son mujeres y el 43% son hombres. Respecto a la Banca Móvil, el 40% de los clientes son activos, el 4% son durmientes, el 35% son inactivos y el 21% no tienen Banca Móvil.

Cluster 1: Clientes Dinámicos

Características Principales: Los clientes en el Cluster 1 también tienen un saldo de deuda total alto, con una media de 476,889.24, y realizan transacciones moderadas, con una media de 444,151.26. La mora máxima histórica es moderada, con una media de 52.91 días. Este clúster muestra una distribución más diversa en términos de ocupaciones, con un 36% de empleados y un 34% de independientes formales. Hay una presencia considerable de amas de casa (11%) y pensionados (18%). Los desempleados, estudiantes e independientes informales tienen una presencia menor o casi nula en este grupo. En términos de género, el 68% son mujeres y el 32% son hombres. Respecto a la Banca Móvil, el 46% de los clientes son activos, el 4% son durmientes, el 31% son inactivos y el 18% no tienen Banca Móvil.

Cluster 2: Clientes Vital

Características Principales: Los clientes en el Cluster 2 tienen un saldo de deuda total bajo, con una media de 12,249.85, y realizan transacciones moderadas, con una media de 347,800.72. La mora máxima histórica es muy baja, con una media de 17.86 días. Este cluster está dominado por empleados, representando el 97% de los clientes. Hay una pequeña proporción de pensionados (2%) y una presencia mínima de amas de casa (1%). No hay desempleados, estudiantes ni independientes informales en este grupo. En términos de género, el 99% son mujeres y el 1% son hombres. Respecto a la Banca Móvil, el 98% de los clientes son activos, el 0% son durmientes, el 1% son inactivos y el 1% no tienen Banca Móvil.

7 PROPUESTAS DE ESTRATEGIAS CON BASE EN LOS RESULTADOS OBTENIDOS

Dada la intención de este trabajo, donde se pretende optimizar los canales digitales a través de incentivar el uso de estos, las estrategias van dirigidas en mayor medida a aquellos clientes que de acuerdo con el resultado de la categorización, son más atractivos en términos financieros y de mercado, pero en contraste son usuarios poco o nada frecuentes de los canales digitales.

Estrategias Recomendadas: **Clúster Clientes Oro**

Dado el perfil de alto saldo de deuda y transacciones financieras significativas de los Clientes Oro del Clúster 0, proponemos desarrollar productos de crédito con tasas preferenciales y límites amplios que puedan ser activados únicamente a través de canales digitales del banco. Estos productos deben estar diseñados específicamente para empleados con alto saldo de deuda, ofreciendo condiciones competitivas que

fomenten la consolidación de deudas y la expansión de líneas de crédito según las necesidades financieras individuales.

Implementar estrategias de fidelización centradas en la personalización y la excelencia en el servicio a través de los canales digitales. Esto puede incluir presentar un look and feel diferente como parte de la experiencia oro para los clientes de este segmento, esta personalización puede estar volcada a presentar una interfaz personalizada con productos de interés para estos clientes como sus productos de inversión cupos o créditos.

Integrar alertas personalizadas en los canales digitales para notificar a los clientes sobre oportunidades de inversión, vencimientos de créditos y ofertas especiales que se alineen con sus necesidades y perfiles financieros, así como ofrecer herramientas para establecer y seguir metas financieras, para ayudar a los clientes a alcanzarlas.

Presentar oportunidades exclusivas a los clientes de este segmento. Para los clientes oro se pueden ofrecer tasas preferenciales para los productos rotativos solo usándolos a través de los canales digitales; podrían ser tasas diferenciales al avanzar con tarjeta de crédito o usar un sobregiro o cupos rotativos.

Utilización de programas de fidelización; para estos clientes que tienen elevadas transacciones promedio se podrían utilizar recompensas diferenciales a través de programas como “Puntos”, “Millas”, “Pinos”, etc., siempre y cuando estos “Puntos” sean generados o alcanzados por realizar transacciones a través de los canales digitales.

Estrategias Recomendadas: **Cluster Clientes Dinámicos**

Estos clientes caracterizados por tener un saldo de deuda alto, transacciones moderadas, una presencia mayoritaria de clientes mujeres, con una ocupación diversa donde predominan empleados y trabajadores formales y una adopción de la banca móvil interesante; requieren estrategias que no los desestimen como clientes importantes, pero sin llegar a tener los mismos beneficios de los clientes “oro”.

Para estos clientes sugerimos estrategias basadas en productos del activo de tasa media, a corto plazo y de montos bajo que puedan ser activados únicamente por el canal digital.

La experiencia dentro de la aplicación debe ser de navegación intuitiva y menos sugestiva que para los clientes del perfil anterior; para estos clientes no es necesario presentar ofertas de productos de inversión o créditos de alto monto, por el contrario, sería pertinente enseñarle una interfaz visual con sus obligaciones: detalles de saldo y fechas de vencimiento.

De igual manera ya que más del 50% de usuarios son empleados o pensionados y se conocen los hábitos transaccionales de estos clientes, se propone tener en los canales virtuales campañas u ofertas especiales preaprobadas para ser usadas a través del canal, de cupos rotativos de bajo valor para los días previos a nominas o fechas identificadas como salariales.

Estrategias Recomendadas: **Cluster Clientes Vital**

Las principales características de estos clientes encierran la predominancia de población femenina, un alto indicador de población empleada y una considerable adopción de los canales digitales hacia el uso de los canales digitales, son muestra de la aceptación por los canales digitales del banco. Las estrategias para este cluster deben de ir volcadas más en renovar la experiencia sin sacrificar la usabilidad a la cual están acostumbrados los usuarios.

Un cambio en el look and feel para los canales digitales de estos usuarios significaría que Bancoomeva presta atención a los detalles y se renueva sin sacrificar la usabilidad. Modificar la apariencia de los canales sin afectar los flujos de procesos impactaría positivamente a los clientes de este segmento quienes percibirían cambios también para ellos y no se sentirían excluidos de los demás segmentos.

Dado que son clientes en su mayoría empleados, con alto uso de los canales, baja mora histórica y con una deuda histórica poco significativa, se proponen créditos de nómina de acuerdo con el apetito de riesgo de la organización, activados directamente por los canales digitales.

8 CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

En este proyecto, se implementó el algoritmo MiniBatchKMeans para la segmentación de clientes basada en datos financieros y sociodemográficos, obteniendo un índice de silueta de 0.58. Esta métrica es considerablemente superior al índice obtenido con el algoritmo K-Means estándar, que fue de 0.08, indicando que MiniBatchKMeans proporcionó agrupamientos más cohesivos y distintivos.

8.1.1 Aspectos Clave del Proyecto

Selección del Modelo: La elección de MiniBatchKMeans se basó en su capacidad para

manejar grandes volúmenes de datos de manera eficiente, ideal para conjuntos de datos extensos como el utilizado en este estudio. Esta elección fue respaldada por su buen rendimiento y la capacidad de generar resultados robustos en términos de segmentación de clientes.

Número de Clústeres: Se determinó que el número óptimo de clústeres para este conjunto de datos fue 3, lo cual permitió una segmentación efectiva de los clientes. Esta decisión se basó en criterios como la cohesión intra-cluster y la separación inter-cluster, optimizando la interpretación de los resultados.

Análisis de Componentes Principales (PCA): Previo a la aplicación de MiniBatchKMeans, se realizó una reducción de dimensionalidad mediante PCA, seleccionando 13 componentes principales que capturaron la variabilidad de las variables originales en 75%. Esta técnica facilitó la eficiencia computacional del modelo.

8.1.2 Ajuste a Perfiles de Clientes:

Los resultados del clustering fueron coherentes con los perfiles socio-demográficos y financieros identificados previamente. Cada clúster representó grupos distintivos de clientes con características similares en términos de ocupación, género, niveles de deuda y comportamiento financiero, permitiendo así una personalización de las estrategias de gestión de cliente.

8.2 TRABAJOS FUTUROS

En este proyecto, se exploró la aplicación de los algoritmos K-Means y MiniBatchKMeans para la segmentación de clientes basada en características financieras. Para futuros trabajos, se propone la evaluación de otros algoritmos de agrupamiento que podrían ofrecer resultados alternativos y complementarios.

8.2.1 Exploración de Otros Algoritmos de Agrupamiento

K-Medoids: El algoritmo K-Medoids es una variante de K-Means que utiliza puntos de datos reales como medoides en lugar de centroides. Esta técnica podría ser beneficiosa para conjuntos de datos con presencia de datos atípicos, ya que los medoides son más robustos ante valores extremos. Sin embargo, es importante considerar que K-Medoids puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes debido a su necesidad de calcular todas las distancias entre pares de puntos.

Agrupamiento Jerárquico: Los métodos jerárquicos agrupan los datos en una

estructura de árbol jerárquico, lo cual permite explorar agrupaciones a diferentes niveles de granularidad. Esto proporciona una visión más detallada de las relaciones entre los grupos. Sin embargo, el agrupamiento jerárquico puede ser más lento y requerir más recursos computacionales, especialmente cuando se aplica a grandes volúmenes de datos debido a su naturaleza exhaustiva.

DBSCAN (Agrupamiento Basado en Densidad): DBSCAN es un algoritmo de agrupamiento basado en densidad que puede identificar agrupaciones de formas arbitrarias y detectar puntos atípicos como ruido. Esto lo hace adecuado para conjuntos de datos donde los grupos tienen formas y tamaños irregulares. Aunque DBSCAN es eficiente para conjuntos de datos grandes, puede requerir ajustes cuidadosos de parámetros y una comprensión profunda del dominio para obtener resultados óptimos.

8.2.2 Consideraciones sobre el Costo Computacional

Es crucial destacar que la implementación de estos algoritmos alternativos conlleva desafíos adicionales, especialmente en términos de costo computacional. Cada uno de estos métodos tiene requisitos diferentes en cuanto a memoria y tiempo de ejecución, lo cual puede impactar significativamente en la capacidad de escalar el análisis a grandes conjuntos de datos. Por lo tanto, antes de adoptar estos algoritmos, será necesario realizar evaluaciones exhaustivas de rendimiento y optimización para asegurar que sean viables desde el punto de vista computacional.

9 REFERENCIAS BIBLIOGRÁFICAS

[1] G. C. Miranda, «LA EVOLUCIÓN Y DIGITALIZACIÓN DEL SECTOR BANCARIO,» 04 2020. [En línea]. Available: https://repositorio.comillas.edu/rest/bitstreams/411219/retrieve.

[2] L. Rodriguez Calderon, "Aceptación tecnológica; análisis del comportamiento del consumidor financiero hacia la transformación digital bancaria," Universidad Nacional de Colombia, 2023.

[3] Gamco, «Gamco,» [En línea]. Available: https://gamco.es/reduciendo-churn-y-mejora-la-satisfaccion-de-los-clientes/. [Último acceso: 21 06 2024].

[4] Asobancaria, "Informe de Sostenibilidad 2021," [online]. Disponible en: https://www.asobancaria.com/wp-content/uploads/2022/08/Asobancaria_Informe_Sostenibilidad_2021.pdf.

[5] S. C. y. J. Barberis, «Fintech: Guía para inversores, emprendedores y visionarios para entender la nueva revolución tecnológica,» 2016. [En línea]. Available: https://www.studocu.com/co/document/institucion-universitaria-antonio-jose-camacho/administracion/el-futuro-es-fintech-una-guia-para-inversores-emprendedores-y-visionarios-para-entender-la-nueva-revolucion-tecnologica/45837900. [Último acceso: 20 06 2024].

[6] A. Arduino, «Comportamiento de los clientes bancarios frente a los servicios electrónicos,» *Actualidad Contable FACES,* vol. 21, nº 37, 2018.

[7] Pixelpro, «Pixelpro,» [En línea]. Available: https://pixelpro.com.co/el-futuro-del-marketing-movil-conectando-con-los-consumidores-en-la-era-digital/. [Último acceso: 02 06 2024].

[8] María Teresa Ballestar de las Heras, «Qué es la ciencia de datos: la alquimia de la era de la inteligencia artificial,» 2023. [En línea]. Available: https://www.eleconomista.com.mx/tecnologia/Que-es-la-ciencia-de-datos-la-alquimia-de-la-era-de-la-inteligencia-artificial-20231202-0028.html. [Último acceso: 14 06 2024].

[9] Lerena, O. “Métodos y aplicaciones de la ciencia de datos para las políticas de CTI: redes sociales, minería de textos y clustering”. Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación. 2019.

[10] Almenara J, González JL, García C, Peña P. “¿Qué es el Análisis de Componentes ¿Principales?” 1998; 1268:58-60.

[11] I. Cassol, «icassol.github.i,» [En línea]. Available: https://icassol.github.io/CursoR/Mod8.html. [Último acceso: 24 06 2024].

[12] Marin, J. “Analisis de Componentes Principales”. Universidad Carlos III de Madrid.

2006. Recurso disponible en: https://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema3am.pdf. Consultado en abril de 2024.

[13] caminosaleatorios, «caminosaleatorios,» [En línea]. Available: https://caminosaleatorios.wordpress.com/2020/04/16/tan-cerca-y-tan-lejos-la-distancia-de-mahalanobis/ . [Último acceso: 20 06 2024].

[14] Libertex. [En línea]. Available: https://libertex.org/es/blog/coeficiente-de-correlacion. [Último acceso: 24 06 2024].

[15] Amat Rodrigo, Joaquín. "Clustering y heatmaps: Aprendizaje no supervisado". cienciadedatos.net. 2017. Recurso disponible en: https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps#Medidas_de_distancia. Accedido en mayo de 2023.

[16] F. Sanz, «The Machine Learners,» [En línea]. Available: https://www.themachinelearners.com/k-means/. [Último acceso: 24 06 2024].

[17] https://scikit-learn.org/, «https://scikit-learn.org/,» [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html.[Último acceso: 25 06 2024].

[18] Nunez, L. (19 de 06 de 2024). Obtenido de https://emma.io/blog/que-es-el-app-churn-rate-tasa-abandono/

[19] FasterCapital. (21 de 06 de 2024). *https://fastercapital.com/.* Obtenido de https://fastercapital.com/es/tema/utilizaci%C3%B3n-del-an%C3%A1lisis-predictivo-en-el-modelado-de-clientes-potenciales.html

[20] Hammond, M. (21 de 06 de 2024). *Hubspot.* Obtenido de blog.hubspot.e: https://blog.hubspot.es/service/satisfaccion-del-cliente

[21] branch. (21 de 06 de 2024). *branch.com.co.* Obtenido de https://branch.com.co/marketing-digital/analisis-de-sentimiento-como-escuchar-a-tus-clientes/#:~:text=El%20an%C3%A1lisis%20de%20sentimiento%20del,y%20a%20responder%20de%20forma%20eficaz.

[22] wabe. (21 de 06 de 2024). *https://wabe.es/.* Obtenido de https://wabe.es/marketing-digital/clustering-descubre-como-segmentar-tu-audiencia-en-marketing-digital/

[23] FasterCapital. (21 de 06 de 2024). *https://fastercapital.com/.* Obtenido de https://fastercapital.com/es/tema/la-teor%C3%ADa-de-la-difusi%C3%B3n-de-las-innovaciones.html

[24] Iuvity. (22 de 06 de 2024). *Iuvity.* Obtenido de

https://www.iuvity.com/es/blog/clustering-el-camino-hacia-un-sector-financiero-eficaz-y-seguro#:~:text=El%20Clustering%20permite%20la%20creaci%C3%B3n,para%20la%20prevenci%C3%B3n%20de%20fraudes.

[25] Oliv, E. J. (2005). Revisión del concepto de calidad del servicio y sus modelos de medición. *Innovar*.

[26] Peña, J. W., & Baños, M. A. (2022). Factores asociados a la adopción de la banca electrónica en México. *Revista mexicana de economía y finanzas*.

[27] Nishizawa, R. M. (2014). Desarrollo del Modelo Servqual para la medición de la calidad del servicio en la empresa de publicidad Ayuda Experto. *Revista Perspectivas*.

[28] Roncallo Mulford, D. E., Acevedo Callejas, A., & Arango Mendoza, A. (20 de 06 de 2024). *https://manglar.uninorte.edu.co/.* Obtenido de https://manglar.uninorte.edu.co/handle/10584/10535

[29] Rodríguez, F. B. (2008). La innovación tecnológica en el contexto de los clusters regionales. *Cuadernos de Administración*.

[30] Alshammari, A., Alshammari, R., Altalak, M., & Alshammari, K. (20 de 06 de 2024). *https://www.researchgate.net/.* Obtenido de https://www.researchgate.net/publication/366703547_Credit-card_Fraud_Detection_System_using_Big_Data_Analytics

[31] Huseynli, N., Kandemir, G., & Huseynli, B. (20 de 06 de 2024). *https://www.researchgate.ne*. Obtenido de https://www.researchgate.net/publication/371721250_Analysis_of_Consumer_Behavior_Variables_Influencing_the_Adoption_of_Mobile_Banking

[32] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Expplainable", 2019. https://christophm.github.io/interpretable-ml-book/.