Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Gestión Industrial

PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA

Ing. Mynor Moshe Mendoza Chon
Asesorado por el Msc. Ing. Bryan Josué Bernal Corey

Guatemala, junio de 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

ING. MYNOR MOSHE MENDOZA CHON
ASESORADO POR EL MSC. ING. BRYAN JOSUÉ BERNAL COREY

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ARTES EN GESTION INDUSTRIAL

GUATEMALA, JUNIO DE 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

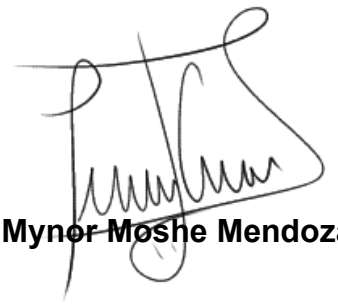
DECANA	Inga. Aurelia Anabela Cordova Estrada
DIRECTOR	Ing. Edgar Darío Álvarez Cotí
EXAMINADORA	Inga. Sindy Massiel Godinez Bautista
EXAMINADOR	Ing. Walter Darío Caal Mérida
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 17 de agosto de 2021.



Ing. Mynor Moshe Mendoza Chon

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.503.2023

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA,** presentado por: **Ing. Mynor Moshe Mendoza Chon,** que pertenece al programa de Maestría en artes en Gestión industrial después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, junio de 2023

AACE/gaoc

Guatemala, junio de 2023

LNG.EEP.OI.503.2023

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA"

presentado por **Ing. Mynor Moshe Mendoza Chon** correspondiente al programa de **Maestría en artes en Gestión industrial** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

FACULTAD DE INGENIERÍA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

Guatemala 14 de mayo 2022.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el **INFORME FINAL** titulado: "**PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA**" del estudiante **Mynor Moshé Mendoza Chon** quien se identifica con número de carné **200216473,** del programa de **Maestría en Gestión Industrial.**

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,

MA. Ing. Kenneth Lubeck Corado Esquivel
Coordinador
Maestría en Gestión Industrial
Escuela de Estudios de Postgrado

Guatemala,29 de octubre de 2020.

M.A. Ing. Edgar Darío Álvarez Cotí

Director

Escuela de Estudios de Postgrado

Presente

Estimado M.A. Ing. Álvarez Cotí

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: "**PROPUESTA DE UN PROCESO DE SEGMENTACIÓN PARA LA VENTA DE PRODUCTOS DIGITALES A CLIENTES EXISTENTES EN UNA INSTITUCIÓN BANCARIA DE LA CIUDAD DE GUATEMALA**" del estudiante **MYNOR MOSHE MENDOZA CHON** del programa de Maestría en **Gestión Industrial,** identificada con número de carné: 999003197.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Bryan Josué Bernal Corey
Ingeniero en Sistemas de Información y Ciencias de la Computación
Colegiado No. 17,323

MSc. Ing. Bryan Josué Bernal Corey

Colegiado No. 17,323

Asesor de Tesis

ACTO QUE DEDICO A:

Dios	Por permitirme realizar una más de mis metas.
Mis padres	Mynor Mendoza y Vilma Chon de Mendoza. Por traerme al mundo y guiarme a través de él.
Mis hermanos	Kemly, Javier y Heber Mendoza, por su apoyo.
Mi esposa	Estri López de Mendoza por sus sabias enseñanzas y consejos.
Mis hijos	Sharon y Hector Mendoza. Por ser la inspiración en mi vida.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser el *alma mater* que permitió nutrirme de conocimientos.
Facultad de Ingeniería	Por proporcionarme los conocimientos que me permitieron realizar este trabajo de graduación.
Familia Prado Rosal, familia y amigos en general	Por su acompañamiento y amistad incondicional

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
cm	Centímetro
kg	Kilogramo
%	Porcentaje
Q	Quetzales

GLOSARIO

Algoritmo Es una serie ordenada de instrucciones, pasos o procesos que llevan a la solución de un determinado problema.

CCV Código que trae la parte de atrás de la tarjeta. Se utiliza para validar pagos en los cuales no hay lectura de banca magnética o chop, principalmente telefónicos o por internet.

Clustering Es una tarea que consiste en agrupar un conjunto de objetos (no etiquetados), en subconjuntos de objetos llamados *clústeres* o segmentos.

Coeficiente Número o parámetro que se escribe a la izquierda de una variable o incógnita y que indica el número de veces que este debe multiplicarse.

Correlación Correspondencia o relación recíproca entre dos o más acciones o fenómenos.

Interpolación Consiste en hallar un dato dentro de un intervalo en el que se conocen los valores en los extremos.

Matriz	Arreglo bidimensional de números consistente en cantidades abstractas que pueden sumarse y multiplicarse entre sí.
NFC	Es un campo electromagnético de corto alcance 5 o 10 cm que se utiliza para comunicar principalmente a *wearables* o pagos sin contacto.
Polígono	En geometría, un polígono es una figura geométrica plana compuesta por una secuencia finita de segmentos rectos consecutivos que encierran una región en el plano. Estos segmentos son llamados lados, y los puntos en que se interceptan se llaman vértices.
Segmentación	La segmentación de mercado divide un mercado en segmentos más pequeños de compradores que tienen diferentes necesidades, características y comportamientos que requieren estrategias o mezclas de *marketing* diferenciadas.

RESUMEN

El propósito de la investigación es el de realizar y proponer una segmentación a clientes actuales de la institución bancaria en el proceso de ventas, la necesidad de ofrecer a los clientes los productos que se adecúan a cada uno de ellos es esencial principalmente porque la competencia también incursiona en la personalización de la oferta que realiza, al hacer una revisión del proceso de ofrecimiento de productos digitales a clientes existentes se encontró que no existe una segmentación efectiva y que esto ocasiona el gasto en anuncios a clientes que tienen una muy baja probabilidad de aceptar los productos objetivos.

Esta es de enfoque mixto, diseño no experimental, de alcance descriptivo porque se hizo recolección de datos para elaborar una propuesta tipo transversal con un tiempo determinado de 6 meses para la realización del estudio con toda la población que son los clientes de los cuales se cuenta con información para hacer un proceso de segmentación.

Al momento de llevar a cabo una segmentación efectiva, y de esta manera priorizar el gasto de pauta, se puede obtener hasta un 41 % más de productos vendidos, utilizando el mismo monto de pauta que se tuvo durante los primeros cinco meses, lo que significa bajar el costo de conversión de $ 0.77 a $ 0.41 y de esta manera optimizar los recursos de los que se dispone.

La principal conclusión que deja esta investigación es que, al personalizar una oferta no sólo disminuyen los costos de cada campaña dado que digitalmente se paga por cada alcance, es decir, cada anuncio que se le muestra a un cliente

que no aplica es dinero perdido, sino que se maximiza la cantidad de clientes a los cuales se puede llegar. Para esto, históricamente se han utilizado datos demográficos, sin embargo, al ir avanzando la tecnología, se han desarrollado procesos de segmentación en los cuales se pueden llegar a encontrar grupos con gustos y necesidades similares, lo que hace posible el poder predecir cuando alguien pueda ser un objetivo ideal para la campaña.

Ya no sólo depende de datos demográficos, pueden llegar a ser incluso patrones de consumo, comportamiento digital, tipos de productos que utiliza, o incluso frecuencia de uso lo que nos pueda dar una idea de lo que el cliente está buscando y pueda necesitar; la tecnología es universal, históricamente las nuevas generaciones son más propensas para utilizar tecnología sin la necesidad de una fuerte campaña sin embargo, hay segmentos en los cuales aún es necesario realizar un ofrecimiento activo por el canal que sea más adecuado, cuando se habla de canal, no necesariamente se refiere a una red social, puede ser también un mensaje de texto, una llamada telefónica, un correo electrónico, una notificación emergente en la aplicación móvil, e incluso un pop-up en la página del navegador. Esto también se puede llegar a obtener por medio de comparación de segmentos similares.

La recomendación principal es actualizar y realizar la segmentación periódicamente ya que las preferencias van cambiando muy rápidamente, y para que esta sea efectiva, no debe pasar más de tres meses sin realizar este estudio; a pesar de la cantidad de variables que se utilizan, se han demostrado las ventajas de realizar un ofrecimiento más personalizado.

PLANTEAMIENTO DEL PROBLEMA

La inexistencia de un procedimiento adecuado para el ofrecimiento de soluciones digitales a clientes acrecentando los costos de obtención de clientes.

Descripción del problema:

En la industria bancaria internacional, en la búsqueda por optimizar los recursos y generar un mayor valor agregado al servicio de los clientes, se han implementado soluciones digitales, las cuales pueden ser ofrecidas también por canales digitales, sin embargo, es un reto al día de hoy el poder dirigir eficientemente estas campañas hacia el grupo objetivo y de esta manera tener un proceso adecuado de segmentación, para poder ofrecer los productos apropiados a las personas apropiadas y de esta manera lograr un rendimiento mayor por los gastos de mercadeo y a la vez evitar el estar importunando a clientes con anuncios que no aplican a ellos.

En un banco del sistema guatemalteco se ofrecen soluciones a clientes existentes para su uso en plataformas digitales, sin embargo, no existe una metodología de ofrecimiento para saber qué clientes puedan aceptar el ofrecimiento, lo que ocasiona gastos excesivos en publicidad dirigida y poca eficiencia en el dinero invertido en publicidad, lo que no permite llegar a los crecimientos estándar del mercado.

- Pregunta central

¿Cómo un proceso de segmentación mejorará la venta de productos digitales para clientes existentes en una institución bancaria?

Preguntas de investigación

¿Cómo se hace la segmentación de ventas a los clientes existente en la institución bancaria?

¿Qué factores críticos tiene la segmentación en ventas al no contar con recursos digitales?

¿Qué ventajas tiene utilizar soluciones digitales para la modernización de atención en ventas de los clientes existentes en la institución?

- Delimitación del problema

El estudio se realizará en la gerencia de transformación digital de un banco del sistema. Durante un lapso de abril-septiembre 2021. Tomando de referencia los niveles de aceptación y alcance del año 2021.

- Viabilidad

Los recursos para la realización del proyecto serán otorgados por la empresa.

- Consecuencias de la investigación

Entre las causas que general el problema, está la falta de un procedimiento para saber qué producto ofrecer al cliente, lo que causa 2 grandes inconvenientes, que todos los clientes reciban información y esto deriva en que el cliente pierde interés en los anuncios de la empresa y segundo es que el presupuesto es limitado para los anuncios, lo que también causa que se gasten y desperdicien recursos en clientes que no aplican para el producto, pero que se gaste en ofrecérseles.

Con la aplicación de una metodología de segmentación y ofrecimiento se logrará el no ofrecer productos a sectores que no aplican, y adicional se obtendrá una mayor ratio de aceptación de los productos al ofrecerles únicamente a aquellos clientes que muestran una alta tasa de probabilidad de aceptación, lo que logrará atraer a muchos más clientes con la misma cantidad de recursos, sin invadir al cliente con publicidad.

Si no se realiza esta mejora, no solo no se tendrán resultados esperados de conversión de clientes, sino que el cliente podrá verse cansado de recibir tanta publicidad y dejará simplemente de prestar atención, lo que ocasionará que en un futuro sea aún más difícil lograr una tasa de conversión a niveles que se tienen hoy en día.

OBJETIVOS

General

Proponer un proceso de segmentación para la venta de productos digitales a clientes existentes en una institución bancaria de la ciudad de Guatemala.

Específicos

1. Identificar la forma en que la empresa hace la segmentación de clientes existentes.
2. Analizar los factores críticos para la solución de la segmentación en ventas al no contar con recursos digitales.
3. Determinar las ventajas que tiene la utilización de soluciones digitales en la segmentación de ventas en una empresa bancaria para los clientes existentes.

RESUMEN DEL MARCO METODOLÓGICO

- Enfoque

El estudio propuesto es con enfoque mixto, ya que sé que adquirirá información cuantitativa como por ejemplo la cantidad de clientes que aceptan el producto y también información cualitativa como las causas por las cuales no han aceptado la oferta ofrecida ya que sean descripciones y sucesos y, por lo tanto, estos no son cuantificables ni numéricos. El estudio es de carácter transversal al tener una fecha propuesta de inicio y fin.

- Diseño de investigación

El diseño de investigación es no experimental, porque no se intervienen con las variables, será únicamente la recolección de datos del fenómeno a estudiar.

Los datos analizar en esta investigación en particular, tanto como los datos recolectados, serán en un tiempo determinado, es decir que tiene un tiempo de inicio y un tiempo final, sin recurrir a ningún tipo de manipulación y estos datos posteriormente pasarán a evaluarse.

- Tipo

Se ha seleccionado un tipo de estudio descriptivo. Se van a recolectar datos, los cuales ayudarán a responder a las preguntas de investigación.

- Alcance

El alcance será descriptivo, ya que se refiere a una propuesta.

Se le proporcionará a la empresa todo el material para poder implementar el sistema de segmentación a clientes y que pueda ser aplicado si así lo consideran necesario

.

- Variables e indicadores

  - Variables independientes

Para el diagnóstico de los procesos que se tienen con respecto a la segmentación se evaluarán los ofrecimientos y el estatus del cliente antes y después que realizara el ofrecimiento, hasta 15 después del envío de la comunicación. Estos datos serán de tipo cuantitativos.

  - Variables dependientes

Para las ventas reconocidas como reales se revisarán los productos a los cuales fue sometido el entrenamiento del modelo. Estas variables se pueden medir a través de indicadores y poder así calcular el tiempo de respuesta para que el cliente obtenga el producto.

Tabla I. **Operativización de variables**

Objetivos	Variables	Tipo de Variable	Indicador	Instrumento
Identificar la forma en que la empresa hace la segmentación de clientes existentes	*Edad *Estado Civil *Salario *Saldo en cuentas *Categoría crédito	Cualitativa y cuantitativas	Porcentaje de ventas efectivas	Observación directa por medio de procesador de datos R Studio
Analizar los factores críticos para la solución de la segmentación de ventas de forma digital de los clientes existentes.	*Productos aceptados *Productos ofrecidos	Cualitativas nominales Y cualitativas ordinales	Cantidad de productos por cliente	Observación directa por medio de procesador de datos R Studio

Fuente: elaboración propia.

- Fases de investigación

  - Fase I. Revisión

Consiste en la revisión documental de la teoría existente para fundamentar cada uno de los planteamientos por realizar.

  - Fase II. Diagnóstico

Diagnóstico en el año 2021 para la evaluación del proceso de segmentación, cuáles son las ventajas y aprendizajes que se han tenido. Evaluar datos históricos para comparar si hay una mejora en el aprendizaje y correr el modelo para establecer un punto de partida.

- Fase III. Determinar variables de importancia

Revisión de resultados de proceso, establecer cuáles de las variables tienen correlación con la aceptación de un producto y cuáles son las mejores para establecer un modelo de segmentación que se adecue a los requerimientos y niveles de aceptación buscados.

- Fase IV. Elaborar sistema de medición

En esta fase, hay que apoyarse con los resultados obtenidos de la fase III, para establecer un sistema idóneo que permitirá tener un control efectivo en el proceso de segmentación tomando en cuenta los niveles de respuesta en cada uno de los distintos productos por ofrecer.

- Fase V. Propuesta de la metodología de segmentación

Revisar los distintos modelos creados con la información recopilada y por medio de la matriz de confusión lograr determinar cuál modelo es el idóneo, aquella que no solo tenga mayor número de aciertos, sino también menor número de aciertos en la aceptación del producto. Ya que la variable que se desea de aumentar es la de la aceptación.

- Población y muestra

La población objetivo de estudio comprende los clientes que se tienen al corte del mes anterior al inicio de nuestra investigación de la institución bancaria.

El resultado esperado es un aumento en la efectividad de ventas producto de una mejor segmentación y ofrecimiento al cliente.

- Instrumentos

La técnica por utilizar en esta investigación será la comparación sistemática en sistema del cliente, en cada una de las tablas donde se tenga almacenada la información de afiliación y aceptación de los distintos productos.

El instrumento que se utilizará será en tablas de información llamados *Datasets,* los cuales estarán almacenados en el servidor virtualizado que se contratará para este proyecto, esto debido a que cuenta con capacidad suficiente para poder guardar toda la información y también realizar la comparación del estado del cliente previo al lanzamiento de la comunicación y posterior al cliente El análisis de la información será utilizado durante el proceso para obtener el procedimiento de cambio eficiente. También se recopilará, organizará, presentará, analizará e interpretará datos. Para ello se utilizarán las siguientes herramientas de estadística descriptiva:

- Diagrama de proceso
- Tablas
- Clasificación de actividades
- Lista de control de actividades por realizar

INTRODUCCIÓN

El banco requiere de un proceso efectivo de segmentación para poder realizar una venta de productos de manera eficiente y óptima, y de esta manera mantener estándares del mercado. Es necesario implementar la segmentación para poder garantizar el funcionamiento adecuado del Departamento de Ventas.

El presente trabajo de investigación es una sistematización para el procedimiento de segmentación de clientes que presenta una propuesta que puede llegar a ser implementada en el banco para tener un proceso adecuado a los estándares globales en el ofrecimiento y venta de productos digitales a clientes existentes.

El problema actual se debe a que, al no tener una correcta segmentación, se realizan ofrecimientos, los cuales tienen un costo monetario específico realizado por la empresa que efectúa dicho ofrecimiento, y al ofrecerse a clientes que no aplican o no tienen interés en este, se desperdician recursos, recursos valiosos que pueden ser utilizados en clientes que tengan una mayor propensión para aceptar el producto objetivo.

La importancia de la segmentación se encuentra en la capacidad de redireccionar los recursos en segmentos con un mejor perfil que permita aumentar la aceptación de los productos.

Se espera, con la investigación, crear un sistema de segmentación, el cual pueda realizarse periódicamente y que pueda mantenerla actualizada, y de esta manera, poder discriminar segmentos que no aplican o con poca probabilidad de

aceptación. Los principales aportes serán aquellos segmentos en los cuales se encuentra una gran densidad de la variable objetivo.

El esquema de solución consistió en primero tener un diagnóstico inicial sobre la forma de ofrecer el producto y cómo se realiza la segmentación, para lo cual fue necesario estudiar 5 meses atrás las bases de los clientes a los cuales se les ofreció el dinero que fue invertido y los resultados obtenidos por dichas campañas.

El trabajo fue posible dado a que se contaba con todos los recursos necesarios, recursos humanos, tecnológicos y de información, lo que permitió obtener resultados certeros y confiables. Tiene un enfoque mixto y un diseño de tipo no experimental debido a que no requirió hacer ensayos, también es un estudio descriptivo.

El primer capítulo del presente trabajo de investigación contiene el marco referencial y tiene los estudios previos de investigación, estadísticos y analíticos que ayudaron a entender el proceso de segmentación.

El segundo capítulo es el marco teórico que recopila literariamente los temas relacionados con el proceso de segmentación, correlación de variables y optimización del proceso de segmentación.

En el tercer capítulo se realiza la presentación de resultados en los cuales se observan aquellas variables que tienen una correlación y se eliminan aquellas que no la tienen, se observan los segmentos óptimos definidos mediante el cálculo de la inercia intraclases y en qué momento esta deja de disminuir considerablemente al aumentar los segmentos, así como el algoritmo óptimo a utilizar realizando una corrida y evaluación de cada algoritmo durante 25 veces.

En el cuarto capítulo si hizo la discusión de resultados conseguidos de la segmentación mediante la priorización de estos segmentos al momento de realizar un ofrecimiento por medios digitales y obtener mayores resultados comparados con los últimos cinco meses estudiados en el presente trabajo.

MARCO REFERENCIAL

En Fortune (2020) según indica que se realizó un estudio acerca de los polígonos de Thiessen, nombrados de esta manera en honor a Alfred Thiessen, quien fuera un meteorólogo estadounidense; estos polígonos se utilizan para poder construir sobre el plano euclídeo una pequeña partición a partir de una construcción geométrica. Gueorgui Voronói quién fue un matemático ruso que durante el año 1,907 hizo estudios también sobre estos objetos y es debido a estos estudios que también se le conocen con los nombres alternativos de Teselación de Voronói, o también en su versión más simple, diagramas de Voronói, aunque también está al matemático alemán Gustav Lejeune Dirichlet, quién más de 50 años antes aproximadamente en 1,850 hizo estudios sobre estos objetos y toman por nombre “Teselación de Dirichlet”.

Los diagramas de Voronói son muchas veces considerados como la forma más simple de realizar una interpolación, y se basan en la distancia euclidiana, esta toma una gran ventaja y es apropiada cuando lo que se está trabajando es información o datos cualitativos. Para poder crear los diagramas se deben unir los puntos entre sí, trazando las mediatrices de los segmentos de unión. En un espacio bidimensional, cuando se quiere determinar los polígonos, lo que se debe de utilizar son las intersecciones de estas mediatrices alrededor de un conjunto de puntos de control, con esto se está logrando que los polígonos que se generaron sean equidistantes a los puntos contiguos y así se logra designar el área de influencia que tendrán, con esta área de influencia segmentada asociada a clientes de la institución bancaria se puede generar un modelo efectivo en el cual aquellos conjuntos similares tendrán un comportamiento similar.

Estos diagramas generados podrán ser utilizados para realizar regiones de un mismo plano de clientes y de esta manera, para efectos de estudio, será más fácil analizar aquellas variables que los diferencian, también agruparlos en grupos o clasificarlos por sus preferencias.

Moreno-Seco, (2004) indica en su publicación sobre la clasificación del vecino más cercano, que es una de las técnicas que más se utilizan y esta consiste en tomar como base un conjunto de objetos que han sido etiquetados, poder asignar una etiqueta correspondiente al objeto que se desconoce la del vecino más cercano en el conjunto que se estudió, tomando en cuenta la distancia que tengan entre ellos.

En la actualidad, y gracias a las computadoras de alto rendimiento, se pueden utilizar gran variedad de algoritmos para encontrar el vecino más cercano de una manera eficiente y, tomando en cuenta que cada día se logran encontrar mejores resultados de clasificación al utilizar los k vecinos que se encuentran más cercanos cuando se toma en cuenta k mayor a la unidad, dichos algoritmos se pueden ir extendiendo cada vez más para encontrar estos k vecinos, pero habrá que tomar en cuenta que siempre que se aumente y el valor de k el tiempo que se tomará en clasificar estar aumentando de manera proporcional dependiendo de cuantas variables se evalúan y de cuantos objetos se están ingresando en el algoritmo.

Al crear los diagramas anteriormente mencionados, y también creando un algoritmo de clasificación con la intención de que el algoritmo realice un aprendizaje supervisado sobre el total de clientes existentes, se podrán predecir los nuevos clientes a manera de clasificarlos y saber cuál será el segmento por asignarles.

Tapis, (2001) en su tesis indica que la segmentación de mercado, o cómo se le conoce y qué es una segmentación de clientes o una segmentación de audiencias, habla sobre un método por el cual, tomando todos los clientes, se pueden dividir los potenciales en varios grupos, sabiendo a cuál grupo pertenece cada uno, se pueden tomar acciones específicas para tratarlos de manera especial y personalizada dependiendo del segmento en el cual sean asignados según dicha segmentación.

Las categorías principales siempre se podrán llegar a dividir en categorías mejor segmentadas y estas pueden estar definidas por cualquier tipo de variable tomando en cuenta su ubicación, su tiempo en la empresa, la edad o cualquier otra característica que se considere importante según el modelo efectuado. Cuando se logra realizar, de manera personalizada una campaña, la efectividad de esta irá aumentando, ya que se puede escoger el mensaje específico para la audiencia correcta tomando en cuenta que se irá adaptando a cada uno de los segmentos.

ANTECEDENTES

Fortune (2020) indica que se realizó un estudio acerca de los polígonos de Thiessen, los cuales toman este nombre en honor al meteorólogo estadounidense Alfred Thiessen, dicho estudio habla sobre el uso de estos polígonos para poder construir sobre el plano euclídeo una pequeña partición a partir de una construcción geométrica. Anteriormente también fueron estudiados en 1907 por el matemático ruso Gueroguio Voronío y es gracias a estos estudios que también se le conocen como Teselacion de Voronoi o Diagramas de Voronoi, aunque también se tiene al matemático alemán Gustav Lejeune Dirichlet quien, 50 años antes, aproximadamente en 1850, hizo estudios sobre estos objetos y toman por nombre Teselación de Dirichlet.

Los diagramas de Voronoi son actualmente considerados como la forma más simple de realizar una interpolación, y se basan en la distancia euclidiana, esta toma una gran ventaja y es apropiada cuando lo que se está trabajando es información o datos cualitativos. Para poder crear los diagramas se deben de unir los puntos entre sí, trazando las mediatrices de los segmentos de unión.

En un espacio bidimensional, cuando se quieren determinar los polígonos, lo que se debe utilizar son las intersecciones de estas mediatrices alrededor de un conjunto de puntos de control, con esto se está logrando que los polígonos que se generaron sean equidistantes a los puntos contiguos y así se logra designar el área de influencia que tendrán, con esta área de influencia segmentada asociada a clientes de la institución bancaria se puede generar un modelo efectivo, en el cual los conjuntos similares tendrán un comportamiento similar.

Estos diagramas generados podrán ser utilizados para realizar regiones de un mismo plano de clientes y de esta manera, para efectos de estudio, será más fácil analizar las variables que los diferencian, también agruparlos o clasificarlos por sus preferencias.

Moreno (2004), indica en su publicación sobre la clasificación del vecino más cercano, indicando que es una de las técnicas que más se utilizan, y esta consiste, tomando como base un conjunto de objetos que han sido etiquetados, en asignar una etiqueta correspondiente al objeto que se desconoce, la del vecino más cercano en el conjunto que se estudió, tomando en cuenta la distancia que tengan.

Actualmente, y gracias a las computadoras de alto rendimiento, se pueden utilizar gran variedad de algoritmos para encontrar el vecino más cercano de una manera eficiente y, tomando en cuenta que cada día se logran encontrar mejores resultados de clasificación al utilizar los k vecinos que se encuentran más cercanos cuando se toma en cuenta k mayor a la unidad, dichos algoritmos se pueden ir extendiendo cada vez más para encontrar estos k vecinos, pero habrá que tomar en cuenta que siempre que se aumente y, el valor de k, así como el tiempo que se tomará en clasificar, estará aumentando de manera proporcional, dependiendo de cuantas variables se evalúan y de cuantos objetos se están ingresando en el algoritmo.

Al crear los diagramas anteriormente mencionados y también creando un algoritmo de clasificación, con la intención de que el algoritmo realice un aprendizaje supervisado sobre el total de clientes existentes, se podrán predecir los nuevos clientes a manera de clasificarlos y saber cuál será el segmento para asignarles.

Tapis (2001) en la tesis de su autoría indica que la segmentación de mercado, o como también se le conoce como qué es una segmentación de clientes o una segmentación de audiencias, habla sobre un método por el cual, tomando todos los clientes, se pueden dividir los potenciales en varios grupos, sabiendo a qué grupo pertenece cada uno, se pueden tomar acciones específicas para tratarlos de manera especial y personalizada dependiendo del segmento en el cual sean asignados.

Según dicha segmentación las categorías principales siempre se podrán llegar a dividir en categorías mejor segmentadas y estas pueden estar definidas por cualquier tipo de variable tomando en cuenta su ubicación, su tiempo en la empresa, la edad, o cualquier otra característica que se considere importante según el modelo efectuado. Cuando se logra realizar, de manera personalizada una campaña, la efectividad de esta irá aumentando, ya que se puede escoger el mensaje específico para la audiencia correcta tomando en cuenta que se irá adaptando a cada uno de los segmentos.

1. MARCO TEÓRICO

1.1. Capítulo I: institución bancaria privada

A continuación, se presenta lo relacionado con institución bancaria privada, dando inicio con la definición.

1.1.1. Definición de institución bancaria

La definición de una institución bancaria principió por ser una empresa mercantil, la cual está constituida y bajo los estatutos de la legislación general de la República y también leyes que regulan, específicamente el sector financiero. La función de estas instituciones es:

> La intermediación financiera bancaria, consistente en la realización habitual, en forma pública o privada, de actividades que consistan en la captación de dinero, o cualquier instrumento representativo del mismo, del público, tales como la recepción de depósitos, colocación de bonos, títulos u otras obligaciones, destinándolo al financiamiento de cualquier naturaleza, sin importar la forma jurídica que adopten dichas captaciones y financiamientos. (Ley de bancos y grupos financieros, 2002, p.3)

Las actividades de una institución bancaria actualmente no se suscriben únicamente a su función fundamental, sino que también tienen actividades distintas a la intermediación financiera pueden actuar como asesores, intermediar valores, arrendamiento financiero, entre otros, lo que las convierte en grupos financieros.

1.1.2. Tipos de operaciones y servicios

La principal función de un banco es la intermediación financiera, esta consiste en actividades que se realizan normalmente en forma pública o privada tales como: colocación de bonos, títulos, recepción de depósitos o cualquier otra obligación la cual se destina al financiamiento de bienes o servicios que pudieran ser de cualquier naturaleza lícita. Las operaciones que puedan realizar variarán y pueden ser físicas como virtuales, las cuales se detallan a continuación.

1.1.2.1. Operaciones y servicios bancarios tradicionales

Tomando en cuenta las operaciones más importantes y generales que se realizan en los bancos se mencionan a continuación las siguientes:

- Operaciones activas

El propósito de estas operaciones es canalizar recursos financieros, pero también se puede otro tipo de servicios, también están las que funcionan como el derecho por parte del banco de ejercer contra terceros, sin tomar en cuenta la forma jurídica al momento de formalizar o su registro contable, un ejemplo puede ser el otorgar créditos; inversión de valores los cuales pueden variar entre el corto mediano y largo plazo; se toma cooperación activa también la constitución de depósitos en bancos extranjeros; arrendamiento factoraje, tarjeta de crédito, incluyendo la emisión, la operación y otras.

- Operaciones pasivas

Se refiere a cuyo propósito es captar recursos financieros y que a su vez el banco garantiza la seguridad y recuperación de estos, para con él cliente final

sin tomar en cuenta su forma jurídica de formalización o de su registro contable, ejemplo: depósitos monetarios, a plazo que a su vez pueden ser de corto mediano o largo plazo, emisión de deuda u obligación por medio de títulos; refinanciamiento por medio de otros bancos que puedan ser nacionales o extranjeros; y otras.

- Operaciones de confianza

Se refieren a todas las que son realizadas por las instituciones bancarias y que actúan como mandatarias, cuyo propósito es el de prestar servicios que no necesariamente implican una intermediación financiera, estás no originan derecho y tampoco crean compromiso alguno; sin embargo, esto siempre debe quedar debidamente consignado cuando se realizan los convenios o se firman contratos entre las partes contratantes. Las principales operaciones que se toman como de confianza son: cobros y pagos a terceros o por cuenta ajena; acciones; administración de fideicomisos, títulos de crédito y otras.

- Pasivos contingentes

La naturaleza de estas operaciones no crea una obligación inmediatamente, sin embargo, la falta de cumplimiento que tenga cualquiera de las partes involucradas en estas transacciones, podría derivar en problemas graves para la institución, como, por ejemplo: otorgar garantías o fianzas; emitir cartas de crédito de garantía.

Dado que el registro de estas operaciones se realiza, en cuentas llamadas de orden y son por su naturaleza transacciones expuestas a riesgos que están fuera de balances, se requiere un mayor cuidado y una constante supervisión por parte del banco ya que son administrados.

- Servicios

Los servicios son productos no tangibles que las instituciones prestan y que son complementos de sus operaciones diarias hacia los clientes. Estos servicios deben de registrarse en cuentas de orden y realizarse jurídicamente utilizando contratos donde se estipula la prestación de los servicios, que pueden ser por comisión, por mediación u otros.

Entre estos servicios se pueden mencionar: cobranza, intermediación para transferencias cablegráficas; cambio de moneda extranjera ya sea en documentos o en efectivo, arrendamiento de casillas de seguridad y otras.

1.1.2.2. Banca virtual

Esta se ha popularizado mucho principalmente a partir del año 2020, tomando en cuenta los tiempos de pandemia, puesto que permite una mayor disponibilidad para los clientes sin tomar en cuenta las restricciones de movilidad y horario, esta se ha ido entregando al cliente por 2 etapas principalmente:

La primera generación, su principal componente que eran componentes electrónicos capaces de automatizar servicios, e introduce hacia el cliente los conceptos de autoservicio, ya que la principal actividad es realizada por el cliente mismo, sus principales herramientas fueron los cajeros automáticos y la banca telefónica, atendida por respuestas programables las cuales van guiando al cliente en su interacción para con el banco y respondiendo a sus necesidades en base a las respuestas que el mismo cliente va dando.

La segunda se refiere a una digitalización completa, utilizando principalmente la computadora, en el inicio por medio de conexiones a internet y

datos, segundo el teléfono móvil con acceso a internet, lo cual terminaría de automatizar el autoservicio y en el autogestionamiento por parte del cliente minimizando en gran parte la interacción de un empleado bancario hacia el cliente, relegándolo a un papel de asesoramiento únicamente más no de servicio como tal.

La banca virtual permite minimizar los costos puesto que no es necesario espacio físico para atender al cliente y minimiza la cantidad de operarios llamados receptores o secretarias de atención al cliente, relegándolos a un papel de asesoramiento mucho más especializado y automatizando todas las gestiones o transacciones de baja complejidad masificando la capacidad de atención sin necesidad de crecer físicamente de manera proporcional, además, permite al usuario final mayores tiempos de atención que puede llegar hasta las 24 horas sin depender de un asociado o eliminando la necesidad de llegar físicamente a un lugar, exponiéndose a las dificultades que ésta plaza física pueda tener en cuestión de distancia, horario o tráfico.

1.1.3. Clases de bancos

La banca es un negocio universal y es por esto por lo que se puede especializar en varios segmentos, por el tipo de actividad o por el tipo de clientes, que se les está prestando un servicio, y dentro de estos segmentos se pueden encontrar:

1.1.3.1. Banca especializada

En este segmento se encuentran todas las bancas que atienden a segmentos con necesidades muy específicas y principalmente las que son excluyentes en muchos casos del mercado financiero. en estos casos se realizan

operaciones de intermediación muy específicas definidas y autorizadas por la ley financiera y se enmarcan únicamente en su giro específico.

1.1.3.2. Banca universal o multibanca

Actualmente este es el tipo de banca qué más se encuentra en el mercado actual, ya que en estos casos pueden dedicarse, no sólo a la intermediación financiera, si no que pueden englobar muchos otros servicios de valor agregado; englobando así todas las actividades financieras que tradicionalmente no pertenece a la banca, lo que permite que estas puedan ofrecer, no sólo en su oficina principal, sino en todas las agencias que puedan llegar a tener en el territorio operaciones en valores, almacenaje, seguros, fideicomiso y muchos otros servicios asociados a la intermediación financiera, pero que no tradicionalmente son productos bancarios.

1.2. Capitulo II: productos digitales

Actualmente los productos digitales han ganado, en muchos casos y principalmente en las bancas jóvenes, mayor cuota de mercado que los productos tradicionales, tomando en cuenta la penetración de teléfonos inteligentes en la población y la masificación de acceso a internet, permite que los crecimientos actuales se concentren en este segmento.

Aunado al creciente acceso a internet y teléfonos inteligentes, también se encuentra una creciente tendencia global de empresas no financieras pero que prestan servicios comúnmente llamadas Fintech, han permitido evolucionar rápidamente en productos más acercados a las tendencias de los clientes o usuarios finales, acortando las distancias físicas que antes no permitían bancarizar a dichos sectores, es por eso que muchos bancos han decidido

implementar fuertes estrategias de digitalización y volcar sus productos a una transformación digital, tomando en cuenta que estos son mucho menos costosos económicamente que sus pares tradicionales, entre estos se encuentran:

1.2.1. Pago digital – pago con el móvil

Esta es la principal atracción en términos de servicios que un banco puede llegar a prestar; puesto que no solo mejora sustancialmente la experiencia del usuario, sino que aporta altas tasas de seguridad y conveniencia al proceso de realizar un pago por medio del teléfono o cualquier dispositivo vestible, eliminando los fraudes por clonación de banda magnética o por hurto físico de tarjetas, ya que en la mayoría de bancos desarrollados, para poder realizar un pago por medio de un dispositivo móvil, tendrá que validar con la huella digital o algún otro dispositivo de seguridad que el mismo teléfono contenga, esto quiere decir que, aunque el teléfono móvil se encuentra expuesto a un hurto, este no podrá ser utilizado para realizar pagos ya que la huella digital no podrá ser duplicada.

También hay otros tipos de pagos que dependen de la tecnología NFC (*Near Field Comunication*), como lo son dispositivos vestibles que se refieren a relojes, anillos o cualquier dispositivo que se pueda adjuntar una etiqueta NFC, pero que no cuentan con la popularidad o penetración que hoy en día cuenta el teléfono móvil.

1.2.2. Tarjetas de crédito digitales

Es la que puede ser obtenida por medio 100 % digital sin necesidad de acudir a una sucursal bancaria, sino que el 100 % del proceso se realiza de manera digital, procesos que van desde identificación efectiva del cliente para

evitar suplantación de identidad, hasta la emisión de un número de tarjeta que no necesita ser troquelado en un plástico de seguridad, ni contar con elementos como chip o banda magnética puesto, que se asocia directamente a un teléfono móvil o una sucursal digital, para realizar pagos, tanto por medio de tecnología NFC o por medio de corroboración de numeración, más fecha de vencimiento más código CCV (*Card Verification Value*).

1.2.3. Pagos de servicios a terceros

Se refiere a todos los pagos que son de una empresa distinta a la de la institución financiera, pero que sin embargo se pueden realizar por medio de las sucursales físicas o digitales que el banco tiene para atender a sus propios clientes, en estos casos no hay distinción del tipo de servicio, del cual pueda originarse dicho cobro, ya que la institución únicamente presta el servicio de obtención de fondos y liquidación de pagos independientemente de cuáles hayan originado dichos cargos.

1.2.4. Transferencias internacionales

Son transferencias cablegráficas o "*wire transfers*", y son dirigidas al extranjero, sin importar el monto estas pueden ser dirigidas a todos los países que cuenten con autorización para utilizar bancos de Estados Unidos como intermediarios, ya que en casi todo el mundo se utilizan bancos intermediarios para hacer llegar a un tercer banco el dinero.

1.2.5. Doble factor de autenticación

Este es un dispositivo que permite al cliente identificarse digitalmente por medio de las plataformas del banco y es distinto al factor principal que es un

usuario y contraseña. Este puede ser físico el cual es denominado *Token* y es un dispositivo que usualmente otorga un número generado por un algoritmo, usualmente de 6 dígitos, que permite confirmar al banco que la persona que está digitando es aquella a la que se le otorgó el dispositivo, este generalmente está programado para vencer en un tiempo de 5 años, principalmente por la vida de la batería, pero también porque entre más tiempo se tiene registro de los números, más fácil sería para una persona descifrar el algoritmo que los genera, también los hay en forma digital los cuales pueden ser una aplicación instalable en un dispositivo celular, el cual no vence nunca, ya que el algoritmo puede ser cambiado remotamente y también puede ser un mensaje de texto o correo electrónico denominado OTP (*One Time Password*) por sus siglas en inglés, enviado a los registros que el cliente haya otorgado al iniciar la relación con el Banco, este número suele vencer en un tiempo de 1 a 5 minutos.

1.3. Capitulo III: segmentación

Es un método que pretende limitar los clientes objetivos posibles a los cuales dirigirse.

1.3.1. Crear modelo de aprendizaje supervisado para ofrecer productos a clientes existentes

Para crear modelos de aprendizaje supervisado se tendrá que revisar la teoría que fundamenta las bases de la presente investigación. Empezando por la normalización de la información y el trato que se debe de otorgar a las variables independientemente, cuál es su tipo para obtener patrones reales de búsqueda de los clientes, pueden ser de varios tipos.

Cualitativamente se refieren a cualidades que no se pueden tratar como números, pero que sin embargo otorga cierta diferenciación que servirá al modelo para establecer grupos específicos o segmentos, que luego serán analizados para entender sus especificaciones, cuando se tiene una variable cualitativa, ésta a su vez también puede ser nominal que se refiere a que al ser nominal pueden haber varios tipos; pero que sin embargo, ningún tipo puede llegar a tener un valor más que el otro, ya que no tienen un orden específico, por ejemplo es el estatus civil del cliente este puede ser viudo, casado, soltero, unido, divorciado.

Todos los clientes pueden ser asignados a cualquiera de las 5 categorías anteriormente mencionadas, sin embargo el hecho de que un cliente sea viudo o que otro cliente sea soltero no necesariamente significa que se puede ordenar numéricamente dando mayor peso a una o a otra, simplemente son distinciones dentro del grupo estudiado, cuando se ve este tipo de variables también se tienen las que son nominales puesto que no se muestran como un número, pero que sí pueden ser ordenadas de mayor a menor, aunque se refieran a cualidades esas cualidades pueden ser evaluadas y asignadas un valor, en estos casos.

Uno de los ejemplos más fáciles de reconocer es el grado académico que pueda llegar a tener un cliente, un grado de licenciatura será mayor que un grado de nivel medio o nivel básico, entonces, aunque se refiera a una variable cualitativa, una cualidad esta cualidad sí puede ser ordenada de mayor a menor o de menor a mayor.

Para continuar y teniendo entendimiento sobre las variables que se van a tener y que pueden llegar a ayudar a definir los segmentos, así como su representación, el siguiente paso es realizar una minería de datos, se le llama minería de datos porque ésta se refiere a una cantidad muy grande de información puesto que se está hablando de cientos de miles de datos y cada

dato a su vez puede llegar a tener una cantidad infinita de variables que se deseen estudiar.

El principal objetivo de la minería de datos es el poder descubrir patrones que puedan llegar a ser de interés y que puedan ser utilizados en el entendimiento del grupo que se va a estudiar, esos patrones en general se convertirán en reglas de asociación o clasificación.

Para esto es indispensable el poder obtener el método idóneo que únicamente puede ser realizado por medio de algoritmos utilizados por la minería de datos.

Primero se realiza un *clustering* para definir un grupo de clientes, ya sea jurídicos o personas que tengan características similares por medio de un modelo no supervisado, es decir se analizarán los registros con etiquetas de clase o categorías que puedan tener, esto para poder buscar generalidades y patrones en estos pequeños clúster de clientes para poder agruparlos en categorías más comunes y con base en cada clúster generar un modelo de recomendación, con un modelo supervisado que pueda analizar los con variables cuantitativas tanto discretas como continuas.

El objetivo de la clasificación es encontrar un modelo (una función o algoritmo) para predecir la clase a la que pertenecería cada registro, esta asignación de una clase se debe hacer con la mayor precisión posible.

En ese caso, como lo que se busca como primer paso, es lograr segmentar a todos los clientes en un clúster donde tenga características similares con otros clientes por lo que se debe de utilizar el método de las K-medias (k. *means*).

Clustering o clusterización: esta es una clasificación que se realiza sin supervisión de otras bases o también se puede definir como aprendizaje no supervisado, es bastante parecido a la discriminación o clasificación, con una pequeña variación de que estos grupos no están predefinidos. El objetivo principal es partir o segmentar los individuos o conjunto de datos en estudio y éstos a su vez pueden ser disjuntos o no. Estos conjuntos están formados en las igualdades que puedan llegar a tener en las variables estudiadas sin partir previamente de una segmentación predefinida, estos como no se tienen una segmentación al estudiarlo se deberán de dar sus principales características para entender o interpretar las razones por las que pudieron llegar a ser conformados en un mismo grupo o conjunto.

1.4. Análisis factorial

El análisis factorial es una técnica que se desprende de la estadística y que permite el poder reducir los datos estudiados y que se utilizan para poder interpretar o explicar las principales correlaciones que hay dentro de todas las variables sujetas al análisis sin necesidad de utilizar todo el conjunto de variables.

Es decir, se estarán generando nuevas variables llamadas factores que explicarán en su conjunto todas las variables sin necesidad de introducir al modelo toda la data que puede suponer cada una de las variables estudiadas; el análisis factorial tiene sus orígenes en la psicometría, y ha sido ampliamente utilizado en ciencias que estudian el comportamiento como lo es la mercadotecnia, Ciencias Sociales, demanda de productos, investigaciones operativas y otras que para su estudio requieren el análisis de una gran cantidad de datos que sin las tecnologías actuales de computación serían prácticamente imposibles de realizar.

1.4.1. Tipos de análisis factorial

Uno de los tipos de análisis factorial es el exploratorio o AFE, y éste es utilizado para entender y descubrir la estructura de una gran cantidad de variables, bajo el concepto de qué existen factores que se pueden asociar a ciertos grupos de variables. el peso de cada uno de los factores es utilizado para describir las relaciones que guardan las variables entre sí. por su facilidad este análisis factorial es el que más se utiliza.

También existe otro tipo de análisis factorial qué es el confirmatorio o AFC, y éste consiste en tratar de entender o determinar si los factores que se obtuvieron del análisis y las cargas de cada uno sobre las variables son los esperados tomando en cuenta teorías previamente definidas. Lo que busca este análisis es confirmar que tengamos factores preestablecidos que cada uno este asociado a un conjunto de variables específicos. es decir, este análisis lo que otorga es un indicador de confianza que permite aceptar o negar la hipótesis previamente establecido. en este análisis las variables son consideradas dos medidas que constantemente pueden ser cuantificados.

1.4.2. Aplicaciones

La aplicación de estos factores principalmente se centra en identificar o explicar características en conjunto que puedan llevar a entender resultados de distintas pruebas.

Un caso de uso ampliamente estudiado es aquel que define que una persona que obtiene buenas notas, en habilidad verbal obtendrá altas notas en la que las asignaturas que requieran un alto desarrollo de la habilidad verbal. esto se puede explicar con el uso de los análisis factoriales que aíslan un factor

llamado inteligencia cristalizada, que describe como una persona puede llegar a ser capaz de poder resolver un problema utilizando su habilidad verbal.

En psicología el análisis factorial se utiliza frecuentemente para investigar la inteligencia, la personalidad, las actitudes o creencias. también está asociado a la ciencia de la psicrometría, ya que está evalúa la validez que tiene un instrumento al establecer si el instrumento mide correctamente los factores que están postulados.

1.5. Correlación

La correlación es aquella relación que pueda llegar a tener una variable o numérica respecto de otra, y este puede llegar a ser cuantificado utilizando el método de Pearson para medir cuál es su coeficiente de correlación. Este proyecta un resultado que puede llegar a identificarse entre la recta empezando desde menos uno hasta más uno.

Siendo -1 una correlación inversamente proporcional y +1 una correlación proporcional a medida que una crece la otra también crece en cierta proporción, cuando se tiene una correlación de cero esta indica que ninguna de las 2 variables guarda similitud en su comportamiento linealmente hablando.

También existen relaciones no lineales y es en estos casos que ayuda bastante realizar una representación gráfica y entender visualmente si existiera algún tipo de correlación.

A continuación, se detallan algunas características del coeficiente de correlación:

- El valor que toma el coeficiente de correlación y las unidades que se utilizan en la medición de las variables son totalmente independientes.

- Al analizar el coeficiente de correlación puede llegar a ser alterado de manera importante sí hay una alta variación en la medición de las variables tal y como también ocurre cuando se analiza la desviación típica. Es altamente recomendable transformar los datos para que tengan una misma escala de variación y no se vea afectada únicamente por una variable.

- La relación de un coeficiente de correlación es únicamente medida con una línea recta. Es por esto por lo que se recomienda primero representarlas gráficamente, para entender las relaciones que pueda haber entre 2 variables y después proceder a calcular el coeficiente.

- Cuando se extrapolan los datos utilizando un coeficiente de correlación, no se deben de realizar en datos que se dan los límites estudiados entre X y Y, esto porque la relación proyectada en el coeficiente será únicamente la que hay dentro del rango llegando a variar considerablemente fuera de este.

- El hecho de que 2 variables tengan correlación no significa que una sea causalidad de la otra.

La validez del coeficiente de correlación será válida sí, las 2 muestras que se están estudiando fueron tomados aleatoriamente dentro del grupo estudiado, y que al menos 1 grupo de los que se está estudiando tenga una distribución normal.

En la tabla II se observa el peso y la talla de 20 personas de género masculino.

Para calcular la covarianza, se multiplica el peso (kg) y la talla (cm) cuya dimensión se elimina para generar un coeficiente por medio de la división de la de desviación típica de la talla y la desviación típica del peso.

El resultado es 0.885 como se vio anteriormente una relación de 1 significaba que las dos variables de estudio guardan una correlación perfecta, por lo que un valor de 0.885 indica una alta relación entre el peso de la persona y la talla de esta. Como se mencionó anteriormente una alta correlación no significa una alta causalidad.

Al elevar al cuadrado 0.885 (coeficiente de correlación) el resultado que se obtiene el coeficiente de determinación ($r^2=0.783$), el cual está indicando el peso está explicado por la talla en un 78.3 %. Es decir, si bien hay otras variables que pueden llegar a modificar o explicar la variabilidad del peso en un 22.7 % la talla es un excelente indicador para intuir el peso. Sin embargo, para entender el 22.7 % que explica el peso, se deberán introducir otras variables para explicar al 100 % el peso.

Tabla II. **Cálculo del coeficiente de correlación de Pearson entre las variables, talla y peso en 20 niños varones**

Y Peso (Kg)	X Talla (cm)	X-$\bar{X}$	y-$\bar{y}$	(X-$\bar{X}$)*(y-$\bar{y}$)
9	72	5.65	1.4	7.91
10	76	9.65	2.4	23.16
6	59	-7.35	-1.6	11.76
8	68	1.65	0.4	0.66
10	60	-6.35	2.4	-15.24
5	58	-8.35	-2.6	21.71
8	70	3.65	0.4	1.46
7	65	-1.35	-0.6	0.81
4	54	-12.35	-3.6	44.46
11	83	16.65	3.4	56.61
7	64	-2.35	-0.6	1.41
7	66	-0.35	-0.6	0.21
6	61	-5.35	-1.6	8.56
8	66	-0.35	0.4	-0.14
5	57	-9.35	-2.6	24.31
11	81	14.65	3.4	49.81
5	59	-7.35	-2.6	19.11
9	71	4.65	1.4	6.51
6	62	-4.35	-1.6	6.96
10	75	8.65	2.4	20.76

Fuente: elaboración propia.

$$\bar{x} = \frac{\sum x}{n} = 66.35$$

$$\bar{y} = \frac{\sum y}{n} = 7.6$$

$$Covarianza = \frac{\sum(\bar{x} - x) * (\bar{y} - y)}{n - 1} = \frac{290.8}{19} = 15.30$$

$$r = \frac{Covarianza}{s_x * s_y} = \frac{15.30}{8.087 * 2.137} = 0.885$$

$$s_x = Desviación\ típica\ x = 8.087$$

$$s_y = Desviación\ típica\ y = 2.137$$

1.6. Test de hipótesis de r

Luego de que se tiene la correlación de Pearson, hay que validar si este cálculo es distinto de cero estadísticamente. Se utiliza un test el cual está basado en una distribución explicada por la t de *student*.

$$Error\ estandar\ de\ r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Para afirmar que la correlación es significativa, en este caso r = 0.885 este debe de ser mayor al valor p (Error estándar de r) cuando se multiplica por la t de *student.*

Tomando el ejemplo anterior donde se evaluaron 20 personas de género masculino queda que los n – 2 es igual a 18, siendo los grados de libertad establecido y el valor de este resultado en la tabla de la t de *student* cuando se busca una seguridad mayor al 97.5 % esta es de 2.10 y si lo que se busca es una seguridad mayor al 99.5 % el valor resultante en la tabla es de 2.8784.

$$Error\ estandar\ de\ r = \sqrt{\frac{1 - 0.885^2}{20 - 2}} = 0.109$$

Al aplicar la fórmula el resultado es 0.109

Al evaluar $2.10 * 0.109 = 0.2289$ según lo que se observa anteriormente, el coeficiente de correlación puede tomarse como significativo puesto que el resultado de 0.2289 es menor al resultado obtenido al calcular la correlación (0.885).

Realizando el mismo análisis, pero ahora buscando una seguridad mayor al 99.5 % en la tabla 2 se observa 2.8784 y reemplazándolo en la fórmula 2.8784 * 0.109 = 0.3137456 el resultado obtenido sigue siendo menor a la correlación calculada anteriormente por lo que se puede asegurar que el coeficiente continúa siendo significativo para una confianza mayor al 99.5 %. Este proceso puede ser aplicado indiferentemente del tamaño de la muestra al revisar la tabla 2 se tienen los valores cuando la muestra tiende a infinito para una seguridad mayor al 99.5 % el valor será de 2.576.

1.7. Intervalo de confianza del coeficiente de correlación

Para estimar correctamente los intervalos de confianza para el coeficiente de correlación de Pearson, se puede utilizar la transformación Z esto se debe a que cuando se tiene la distribución del coeficiente de correlación de Pearson esta puede no estar centrada, lo que imposibilita el poder estimar correctamente los límites de manera directa. Es por este motivo que se utiliza una transformación de Fisher del coeficiente de correlación (Z de Fisher), esta transformación se asume de distribución normal y con una desviación típica

$\frac{1}{\sqrt{n-3}}$.

La transformación es:

$$z = 1/2L_n \frac{1+r}{1-r}$$

Ln representa el logaritmo neperiano en la base e

$$El\ error\ standar\ de\ z\ es = \frac{1}{\sqrt{n-3}}$$

donde n se refiere a el tamaño muestral.

El intervalo de confianza de z se puede calcular de la siguiente forma para un 95 %:

$$z_1(limite\ inferior) = z - 1.96/\sqrt{n-3}$$

$$z_2(limite\ superior) = z + 1.96/\sqrt{n-3}$$

Luego de que se calcularon los intervalos de confianza utilizando el valor z, se debe de volver a calcular inversamente para poder calcular correctamente los intervalos de confianza del coeficiente r

$$\frac{e^{2z1}-1}{e^{2z1}+1} \quad a \quad \frac{e^{2z2}-1}{e^{2z2}+1}$$

Utilizando el ejemplo de que se está trabajando, se obtiene r = 0.885

$$z = 1/2L_n \frac{1+0.885}{1-0.885} = 1.398$$

Con un 95 % de intervalo de confianza para z

$$z_1 = 1.398 - 1.96/\sqrt{20-3} = 0.922$$

$$z_2 = 1.398 + 1.96/\sqrt{20-3} = 1.873$$

Nuevamente para calcular los intervalos de confianza en z, se realiza un cálculo inverso que darán nuevamente los intervalos del coeficiente r qué es el originalmente buscado previo a la transformación logarítmica realizada.

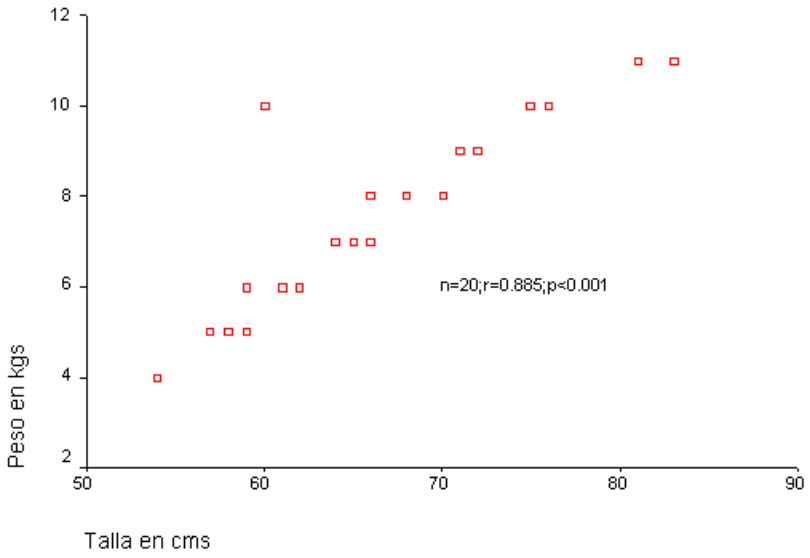
$$\frac{e^{2*0.922}-1}{e^{2*0.922}+1} \quad a \quad \frac{e^{2*1.873}-1}{e^{2*1.873}+1}$$

0.73 a 0.95 serán los intervalos de confianza a 95 % de r.

1.8. Presentación de la correlación

Tomando en cuenta lo anteriormente mencionado es recomendable en todos los casos mostrar gráficamente las 2 variables que se están estudiando y visualizar la correlación que pueda llegar a existir (figura 1). Se muestra el valor de la correlación bajo la letra r a una exactitud de 3 decimales y a su derecha el valor de la prueba de hipótesis demostrando que R es distinto de cero estadísticamente hablando. Por último, se muestran el tamaño de la muestra bajo la letra n.

Figura 1. **Correlación entre talla y peso**



Fuente: elaboración propia, realizado con Rstudio.

1.9. Interpretación de la correlación

Para interpretar correctamente el coeficiente de correlación se debe tomar en cuenta que este puede ser entre -1 y +1, sin embargo, el valor 0 es el que indica que no existe una correlación lineal entre las variables que se están estudiando.

En el caso de tener un coeficiente cercano a cero no es indicativo de la inexistencia de una correlación, puede ser que las variables tengan una relación que no necesariamente sea lineal, es decir, mientras una puede variar bajo una recta de tendencia la otra puede llegar a variar exponencialmente, cuando se tiene un caso de esta manera la correlación r subestima la asociación que las variables tienen al estar las midiendo de manera lineal, en estos casos será necesario utilizar métodos que no son paramétricos para poder entender

correctamente si las variables crecen conjuntamente o puedan estar moviéndose en direcciones distintas.

Cuando se utilizan las correlaciones es importante también estudiar la significancia de este coeficiente ya que ayudarán a evaluar sí estas son significativas o no, es decir para coeficientes de 0.7 si la muestra es realmente pequeña el intervalo de confianza demostrará que su significancia es alta tomando en cuenta que en estos casos este tiende a ser más amplio.

El coeficiente de determinación que se representa mediante r2 no es más que un porcentaje que está mostrando cuál es la variabilidad que pueden tener los datos que se están explicando por la asociación de las variables que se están evaluando.

Se debe tener en cuenta que, como se ha mencionado anteriormente, una correlación cercana a uno, aunque esta sea significativa, no necesariamente se debe a una causalidad, puede que estas 2 variables estén causadas por el mismo fenómeno, por lo cual mantengan una alta correlación, pero que no sean causantes la una del otro, un ejemplo muy antiguo es el de mencionar la correlación que hay entre las personas que tienen un caballo y su expectativa de vida, al realizar el estudio, se muestra que las personas que tienen una alta expectativa de vida, muchas comparten la misma característica que tienen un caballo, sin embargo, no es la variable de tener un caballo la de que determina su alta expectativa de vida, sino que ésta se debe al poder adquisitivo que pueda llegar a tener una persona por la cual le hace más fácil acceder a tratamientos y cuidados de salud así como a tener pasatiempos que demandan un alto gasto para obtenerlos es entonces donde se observa que tanto la variable de caballo como la variable de expectativa de vida están asociadas a una tercera variable

qué es ingresos económicos por lo cual al subir una sube la otra, pero no necesariamente una causa la otra.

Cuando se intenta medir el mismo evento mediante 2 métodos distintos, no es conveniente utilizar el coeficiente de correlación, ya que en esencia los 2 métodos están enfocados al mismo evento, por ejemplo la medición de temperaturas mediante un termómetro de mercurio versus la medición por parte de un termómetro digital, al medir la correlación que hay entre ambas medidas aunque se refieran a 2 instrumentos que utilizan tecnologías distintas, el evento evaluado es el mismo por lo que pudiera llegar a haber una correlación muy cercana a uno, sin embargo el nivel de concordancia será nulo ya que lo que mide el coeficiente de correlación es la asociación que existe entre 2 cantidades, pero no es capaz de medir el nivel de concordancia.

1.10. Coeficiente de correlación de los rangos de Spearman

Es la medición de la asociación o interdependencia lineal que es utilizada por los rangos. Para medir correctamente esta asociación o para calcular dicho coeficiente existen 2 métodos el primero estudiado por Kendal y el segundo por Spearman, cuando se habla del método de Spearman se denota bajo ρ (rho) y este es más fácil de calcular si se compara contra el método de Kendall. El ρ (rho) es idéntico al coeficiente de correlación de Pearson cuando éste se calcula sobre el rango observado. Cuando se presentan valores externos se recomienda utilizar el método ρ (rho), ya que estos valores pueden llegar afectar mucho el de Pearson, también cuando se tienen distribuciones que no son normales.

La forma de cálculo de dicho coeficiente se muestra bajo la fórmula:

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

En donde di = rxi – ryi se refiere a la diferencia de los rangos existente en X y Y.

Al colocar los valores de estos rangos deberán de llevar el orden numérico que existen entre los datos de la variable.

Los valores de estos rangos deberán de colocarse según el orden numérico de los datos de la variable.

Caso I: se busca determinar la correlación que existe entre la nicotina presente en la sangre de un grupo de estudio y la cantidad de nicotina que hay que en una conocida marca de cigarrillos.

Los rangos se representan dentro de paréntesis

Tabla III. **Nicotina existente en sangre vs. cigarrillo**

X	Y
Nicotina existente en sangre (nmol/litro)	**Nicotina existente en cigarrillo (mg)**
185.7 (2)	1.51 (8)
197.3 (5)	0.96 (3)
204.2 (8)	1.21 (6)
199.9 (7)	1.66 (10)
199.1 (6)	1.11 (4)
192.8 (6)	0.84 (2)
207.4 (9)	1.14 (5)
183.0 (1)	1.28 (7)
234.1 (10)	1.53 (9)
196.5 (4)	0.76 (1)

Fuente: elaboración propia.

En el caso de que existan valores que sean coincidentes, se debe colocar el promedio de los rangos que hubieran sido asignados, de no existir coincidencias, tómese el siguiente párrafo como ejemplo:

Tabla IV. **Edad vs. rango**

X (edad)	Rangos (Establecido)
23	1.5
23	1.5
27	3.5
27	3.5
39	5
41	6
45	7
...	...

Fuente: elaboración propia.

Escribiendo según la fórmula el cálculo estimado los valores quedarían de la siguiente forma:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad = 1 - \frac{6\left[(2-8)^2 + (5-3)^2 + (8-6)^2 + \ldots + (4-1)^2\right]}{10(10^2-1)} \quad = 1 - \frac{6(120)}{10(99)} = 0.27$$

Como se mencionó anteriormente, al utilizar la fórmula de cálculo para el coeficiente de correlación de Pearson, si obtiene el mismo resultado que con la fórmula anterior, en este caso 0.27 tomando en cuenta los rangos.

$$r_s = \frac{n\sum r_x r_y - \sum r_x \sum r_y}{\sqrt{\left[n\sum r_x^2 - \left(\sum r_x\right)^2\right]\left[n\sum r_y^2 - \left(\sum r_y\right)^2\right]}}$$

$$\sum r_x = \sum r_y = 55 \qquad \sum r_x^2 = \sum r_y^2 = 385$$

$$\sum r_x r_y = 2\,(8) + 5\,(3) + 8\,(6) + \ldots + 4\,(1) = 325$$

$$r_s = \frac{10(325) - 55(55)}{\sqrt{\left[10(385) - 55^2\right]\left[10(385) - 55^2\right]}} = 0.27$$

Al igual que se interpreta el coeficiente de correlación de Pearson el de Spearman rs también varía entre -1 y +1, exactamente igual que un Person un valor cercano a uno significa que existe una correlación fuerte y que ésta a su vez es positiva, es decir si una crece la otra crecerá de igual manera; sí el coeficiente fuera muy cercano a cero y esto nuevamente significa que la variación de una variable no está ligada o correlacionada a la otra variable mientras que al tener coeficientes cercanos a -1 esto querrá decir que las variables variarían de manera significativa una de la otra pero en sentido contrario es decir cuando crece una la otra decrece, también se puede r_s^2 estimar que el tendrá exactamente el mismo significado que con el r^2.

Por último, las distribuciones existentes entre r y rs similares por lo que, para realizar el cálculo de los intervalos de confianza en ambos casos, pueden realizarse de la misma manera que como se mostró anteriormente para el de correlación de Pearson.

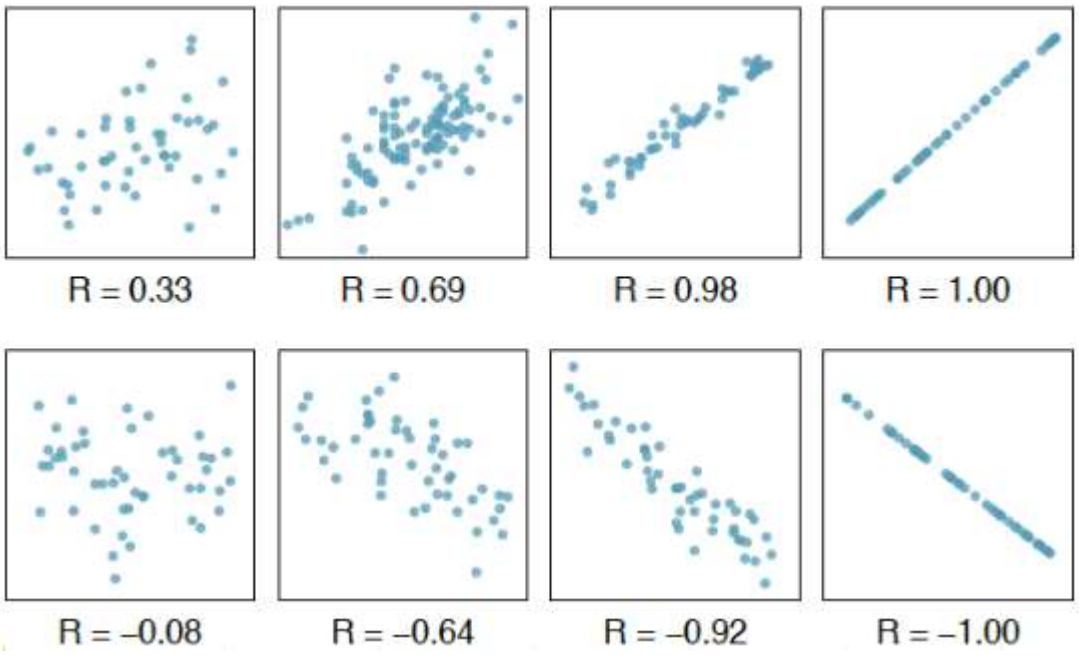
1.11. Cálculo de correlaciones

Para calcular las correlaciones, como ya se ha documentado, es el grado en el cual variarían 2 variables una de la otra, un ejemplo fácil de estudio puede ser la relación que hay entre la variable de salario comparada con la variable de experiencia laboral siendo la tendencia que al tener mayor experiencia se tendrá un mayor.

Esta se mide con un coeficiente que va de -1 a 1.

- r = 1, la relación es positiva perfecta (al crecer una la otra crece).
- 0 < r < 1 la relación es positiva (al crecer una la otra crece en menor proporción).
- r = 0 no hay relación lineal (son totalmente independientes).
- -1 < r < 0 la relación es negativa (al crecer una la otra decrece en menor proporción).
- r = -1 la relación es negativa perfecta (al crecer una la otra decrece).

Figura 2. **Representación gráfica de matriz de correlación**



Fuente: elaboración propia, realizado con Rstudio.

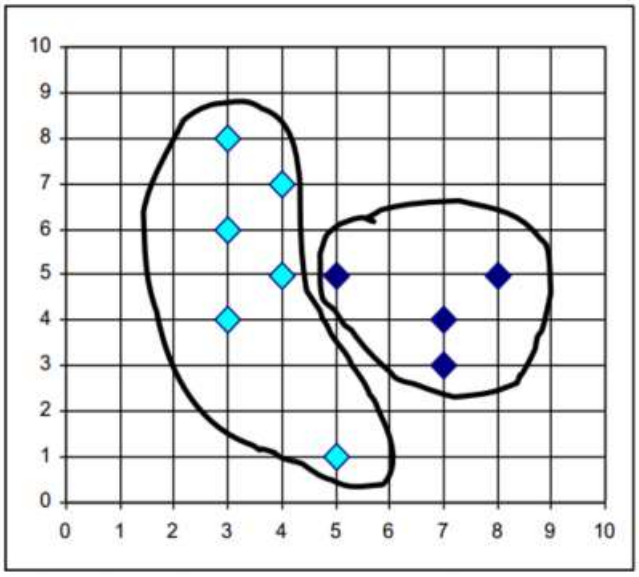
1.12. K-means

Cuando se refiere a este algoritmo es del tipo de clasificación no supervisado o clusterización donde se van agrupando los objetos en una k cantidad de grupos tomando como base las características de las variables estudiadas. Dicha clasificación se logra mediante el proceso de minimizar la suma de las diferencias que existen entre cada objeto y el centroide propuesto, en general se pueden utilizar varias distancias desde la euclidiana hasta la.

Dicho algoritmo se puede realizar en los siguientes pasos los cuales son 3:

Inicio: cuando ya se tiene la cantidad de grupos que se desea, se buscan los centroides esto en un principio pueden ser escogidos aleatoriamente.

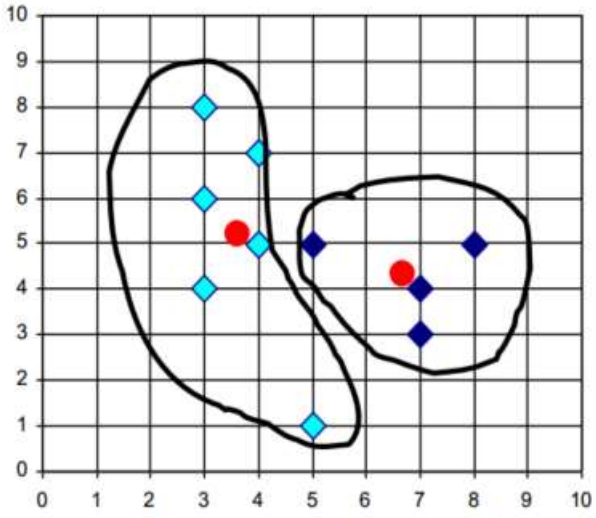
Figura 3. **Segmentos con centroides escogidos aleatoriamente**



Fuente: elaboración propia, realizado con Rstudio.

Cuando ya se tiene el centroide establecido tomando en cuenta la cantidad de segmentos que se desean realizar, el siguiente paso es evaluar todos los puntos y calcular las distancias entre el punto y los centroides y asignarlo al centroide que quede más cercano mediante el cálculo de la distancia establecido para el ejercicio.

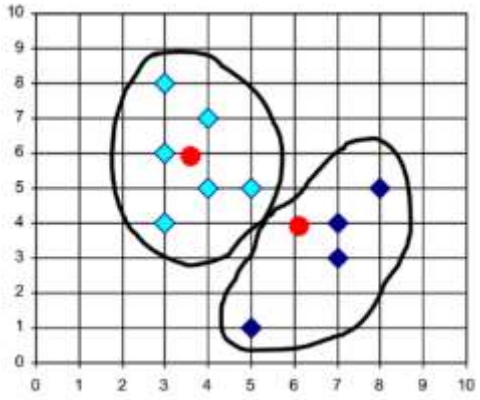
Figura 4. **Cálculo de centroide**



Fuente: elaboración propia, realizado con Rstudio.

Actualización de los centroides: luego de que ya se tienen los 2 segmentos en el caso de este ejemplo, se procederá a calcular el centroide sin tomar en cuenta el centroide que se estableció aleatoriamente al principio del ejercicio y esto nos dará los nuevos centroides.

Figura 5. **Nuevos centroides calculados**



Fuente: elaboración propia, realizado con Rstudio.

Luego de que ya se tienen los nuevos grupos de centroide se procederán a realizar los pasos 2 qué es el cálculo y asignación de cada valor a un segmento específico y como paso 3 calcular nuevamente el centroide de estos grupos, al continuar iterando se observará que por cada iteración el centroide empezará a disminuir su variación respecto de la iteración anterior por lo que se podrá detener cuando se logre una variación mínima establecida para cada paso.

Este algoritmo lo que busca es resolver un problema de optimización, que es el total de las distancias sumadas del grupo de objetos en un clúster y el centroide de este.

Definiciones utilizadas en el método de k-means

Inercia total: es la suma de la inercia de todos los puntos evaluados

$$\text{I} = \frac{\text{M}}{\text{I}} \sum_{\text{M}}^{\text{i=I}} \|\mathbf{x}^{\text{i}} - \mathfrak{a}\|_{3}$$

Inercia inter-clases: se utiliza para calcular la inercia que existe entre los centros de gravedad (en este caso estudiado 2 centros) y el centro de gravedad que genera el 100 % de los puntos:

$$B(P) = \sum_{k=1}^{K} \frac{|C_k|}{n} ||\mathbf{g}_k - \mathbf{g}||^2$$

Inercia intra-clases: esto se refiere a la inercia existente dentro de cada uno de los segmentos establecidos

$$W(P) = \sum_{k=1}^{K} I(C_k) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$$

Una de las principales ventajas que se tienen con este método es la sencillez que representa y dependiendo de la cantidad de datos este puede ser bastante rápido tomando en cuenta las capacidades de computación existentes en la actualidad, sin embargo una de las desventajas que existen es que se debe de escoger primeramente la cantidad de segmentos que se desean buscar y el resultado final pueda llegar a variar significativamente dependiendo en qué lugar del plano se colocan los primeros centroides ya que como se recordara se tomó de manera aleatoria, es decir, el resultado final variará siempre y en algunos casos significativamente dependiendo donde se coloquen los centrales.

1.13. Teorema igualdad de Fisher

I. total = i. inter-clases + i. intra-clases

I = B(P) + W(P)

Objetivo: el objetivo principal es que la inercia que existe entre cada una de las clases sea mínima, mientras que la inercia que existen entre las clases evaluadas sea máxima, es decir matemáticamente que los clustering sean lo más separado posible uno del otro y que cada clúster esté lo más concentrado posible dentro del clúster mismo.

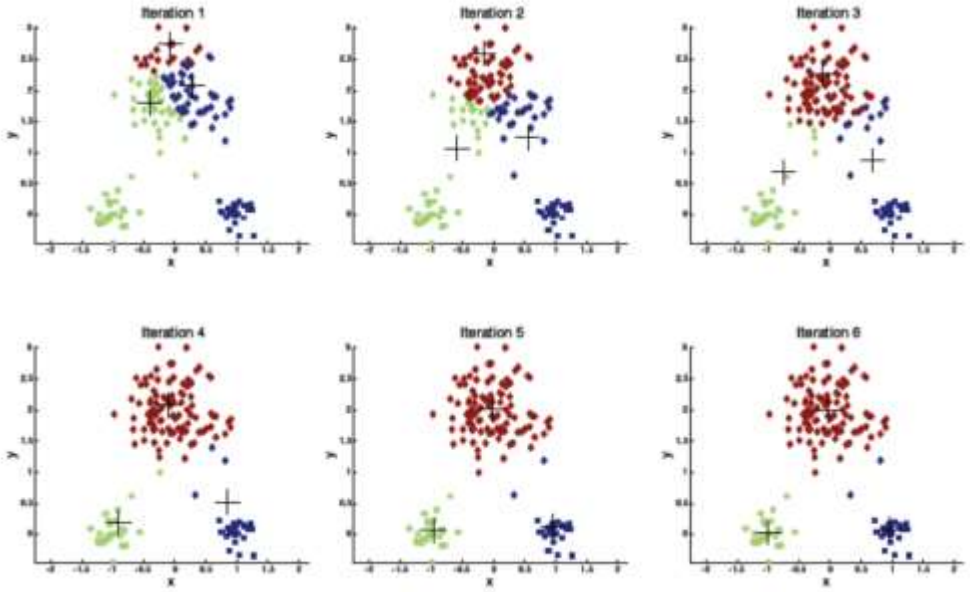
La inercia total no cambia dado que los puntos siempre serán los mismos, por lo que a lograr maximizar la inercia interclases automáticamente se logrará minimizar la inercia intraclase por lo que con enfocarse en minimizar la inercia intraclase se logrará automáticamente maximizar la inercia interclases logrando un grupo más homogéneo dentro de cada clúster.

Se deben tomar en cuenta que si lo que se busca es una segmentación de k número de clases en una nube de n individuos, por definición no se obtendrá ningún beneficio el tener una partición k clases = n individuos.

1.14. Objetivo del método K-means

Como se mencionó el principal objetivo de este método es encontrar una partición que logre que W(P) sea lo más mínima posible

Figura 6. **Evolución en la segmentación en cada iteración**



Fuente: elaboración propia, realizado con Rstudio.

Tomando en cuenta la figura anterior la principal decisión a tomar es cuántas segmentaciones o clases se realizarán, ya que de esto parte la homogeneidad que se pueda lograr en los grupos que se quieren establecer, en este caso se utilizarán la técnica gráfica a la cual se le conoce como “codo de Jambú”

Observando el método se puede notar que todo parte de cuantos clústeres debemos de generar para lograr una cantidad óptima, para esto se puede utilizar una técnica gráfica que se le conoce comúnmente como “codo de Jambú”

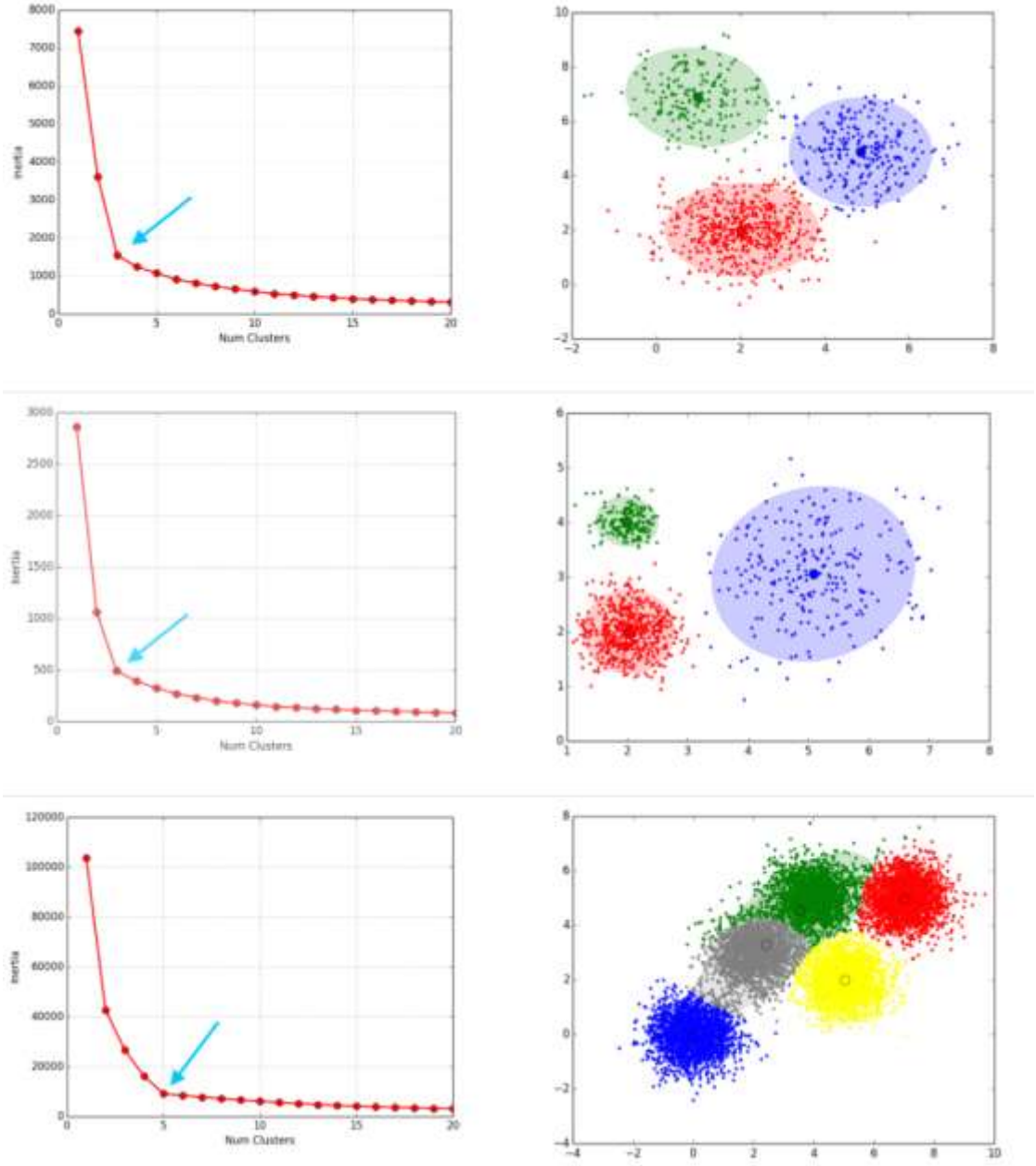
1.15. Método del codo de Jambú

Para utilizar el método lo primero que se realiza es el cálculo de la inercia sí k fuera igual a uno, luego él se calcula sí k fuera igual a 2 y se hace crecer k hasta un número n de clúster, eso dará un valor de inercia dependiendo de la cantidad de clúster que se vayan generando al graficar esta inercia en un plano cartesiano siendo x la cantidad de clúster y llegue la inercia total generada se podrá ir observando una gráfica que va disminuyendo exponencialmente a medida que la cantidad de clúster va creciendo.

Siendo en los últimos valores o cuando la cantidad de clúster tiende al infinito la variación en la inercia será muy bajo por lo que la recomendación es buscar el último valor que hace de crecer considerablemente la inercia y tomar es como la cantidad de clúster que se van a trabajar puesto que es lo que se entiende es que si a partir de esos *clustering* se sigue creciendo se aumentará considerablemente la necesidad de cálculos siento el beneficio que se obtiene de esa clusterización despreciable.

En la siguiente figura se puede observar visualmente a lo que se refiere este método, del lado izquierdo se tiene el codo mientras del lado derecho se muestra ya la clusterización segmentada por colores lo que permite validar de forma gráfica que dicha cantidad de segmentos tiene coherencia con la inercia total generada por estos.

Figura 7. **Muestra de segmentos vs. codo de Jambú**



Fuente: elaboración propia, realizado con Rstudio.

2. PRESENTACIÓN DE RESULTADOS

Se validaron los últimos 5 meses de las campañas digitales que se han realizado respecto a un solo producto, ya que al definir la segmentación para un producto luego se puede replicar para cualquier otro producto es por eso por lo que se enfocó en uno específicamente, de los cuales se obtuvieron los siguientes resultados.

2.1. Diagnóstico situacional

Objetivo 1. Identificar la forma en que la empresa hace la segmentación de clientes existentes.

En esta primera fase se realizó el análisis de la segmentación anterior y los datos históricos de cómo se realizaba un ofrecimiento a cada cliente, de acuerdo con el histórico se presentan los siguientes resultados.

En correspondencia al primero objetivo propuesto se presentan los siguientes resultados.

2.2. Procedimiento

Se realizó la revisión del proceso actual y como es que la empresa segmenta el ofrecimiento de productos digitales a sus clientes.

Se determinó que el ofrecimiento se segmenta únicamente en base a el estatus del cliente referente a las 2 variables digitales de producto, estas son si

el cliente cuenta con usuario digital y segundo si el cliente cuenta con un doble factor de autenticación.

Si el cliente cuenta con estos dos productos, los cuales son los mínimos necesarios para poder obtener más productos de manera digital. Se realiza un ofrecimiento limitado únicamente por el presupuesto asignado a cada campaña digital.

2.3. Campaña digital

Una campaña digital es aquella que se envía 100 % por medios digitales y utiliza herramientas principalmente de redes sociales, en el caso estudiado y al cual se busca una propuesta se refiera a las campañas que se realizaron en Facebook y Google, y el presupuesto se limitó por campaña no por cuenta, es decir, al producto A se le asignó un monto específico independientemente de cuanto se tuvieron en total la suma de todos los productos. Se definió el pago del anuncio por conversión no por impresión y no se utilizó segmentación de la plataforma, sino que se cargó una base de clientes previamente definidos según el procedimiento anteriormente descrito, se definió un límite de $1,000 gastados durante el tiempo de la campaña el cual fue de 3 semanas y al momento de terminarse el presupuesto se concluyó con el ofrecimiento por parte de la red social.

2.4. Medición de resultados

La medición de resultados se realizó midiendo los siguientes pasos:

- Cantidad de clientes que se cargaron a la campaña
- Alcance de clientes que lograron ver el anuncio
- Conversión o clientes que hicieron clic en el anunció que se presentó

- Ventas, clientes que terminaron el proceso de compra durante el mismo mes que se les mostró el anuncio.

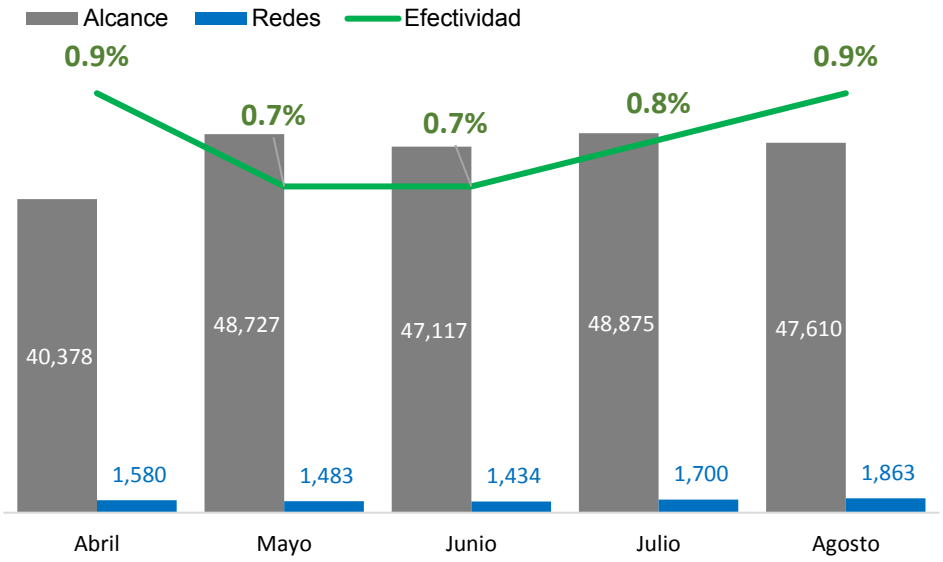
Tabla V. **Resultados por segmentación anterior**

Mes	Abril	Mayo	Junio	Julio	Agosto
Base	175,556	211,857	204,857	212,500	207,000
Alcance	40,378	48,727	47,117	48,875	47,610
Presupuesto	$ 1,250.00	$ 1,250.00	$ 1,250.00	$ 1,250.00	$ 1,250.00
Redes	1,580	1,483	1,434	1,700	1,863
Efectividad	0.9 %	0.7 %	0.7 %	0.8 %	0.9 %
Costo por Venta	$ 0.71	$ 0.84	$ 0.87	$ 0.74	$ 0.67

Fuente: elaboración propia.

A continuación, se muestran los resultados de la campaña en función al Alcance, o cantidad de clientes que vieron el anuncio e hicieron clic y se elimina el factor del tamaño de la base, ya que es la pauta o presupuesto el que limita la campaña y no la segmentación.

Figura 8. **Resultados campañas en redes sociales durante el tiempo estudiado**



Fuente: elaboración propia.

En general al revisar los 6 meses de estudio el costo por venta promedio es de $0.78, el cual proviene de sumar todo el gasto que se tuvo durante las campañas el cual asciende a $6,250.00 y dividirlo por todas las ventas realizadas durante el mismo período 8,060.

Objetivo 2: analizar los factores críticos para la solución de la segmentación en ventas al no contar con recursos digitales.

De acuerdo con el segundo objetivo planteado en esta fase se muestra el estudio realizado para determinar cuáles son los factores críticos que permitirán realizar una correcta segmentación en ventas.

2.5. Definición de metodología para determinar factores críticos

Se realizó un barrido de todo el universo por utilizar, el cual consiste en todos los clientes de los cuales se puede extraer registro de los productos con los que cuenta en el banco, así como del uso de plataformas digitales. Se identificaron 38 variables distintas separadas por las características de origen de la siguiente manera:

Tabla VI. **Cantidad de variables por evaluar**

	Variables	
Origen	Producto	Sub-Producto
Banco	9	3
Tarjeta de crédito	3	4
Aplica a Ambos		5
Comportamiento Digital	2	11
Comportamiento Tradicional	2	
Total de variables		39

Fuente: elaboración propia.

2.6. Variables banco

Son todas aquellas variables que están asociadas a un producto emitido por el Banco, estas pueden ser cuentas bancarias separadas por sus características, préstamos, depósitos a plazo.

2.7. Variables tarjeta de crédito

Son aquellas que están asociadas a tarjeta de crédito; específicamente se refieren a tipos de tarjeta separadas por el color o categoría a la que pertenecen.

2.8. Variables subproducto

Estos son productos secundarios que no están asociados directamente, sino que dependen de un producto principal este puede ser un producto bancario o un producto de tarjeta de crédito, aunque también están las variables que pueden estar asociadas a cualquiera de los 2 productos

2.9. Variables comportamiento digital

De estas variables se encontraron 2 de producto que se refiere principalmente al producto de banca electrónica y el producto de doble factor de autentificación. También se encontraron 11 de subproducto las cuales pueden estar asociadas a uno o los 2 productos principales del comportamiento digital. Las variables de comportamiento digital se refieren a acciones que el cliente ha realizado dentro de las páginas y aplicaciones del mismo banco estas pueden llegar a ser consulta sus saldos por internet, realiza compras dentro de la página de cupones y descuentos, realiza transferencias por medio de internet y cualquier otro comportamiento que el cliente realice en medios exclusivamente digitales.

2.10. Variables comportamiento tradicional

Estas variables estudiaron el comportamiento del cliente en los canales denominados tradicionales, estos se refieren a visita a agencias bancarias, usos del cliente en cajeros automáticos y uso de corresponsal no bancario, de esta se utilizaron las 3 variables.

2.11. Correlación

Del barrido realizado se identificaron 549,660 registros que contenían datos en cuanto a productos de banco o productos de tarjeta. Y se realizó el análisis de correlación respecto de la variable objetivo. Esto con el fin de identificar las variables que tienen muy poca relación y eliminarlas del análisis, con la intención de mejorar los tiempos de análisis tomando en cuenta que hay 549,660 registros.

A continuación, se muestran las correlaciones entre todas las variables y la variable objetivo.

Tabla VII. **Correlación de la variable objetivo con respecto al resto de variables**

VARIABLE	Correlación	VARIABLE	Correlación	VARIABLE	Correlación
COMP_DIG_1	0.12759481	PROD_A_2	0.01554395	PRODUCTO_OBJETIVO	1
COMP_DIG_10	0.26725479	PROD_A_3	0.00637063	SUB_PROD_1	0.19965417
COMP_DIG_2	0.22631826	PROD_A_4	0.0230071	SUB_PROD_2	0.00680111
COMP_DIG_3	0.22631826	PROD_A_5	0.00637063	SUB_PROD_3	-0.00016912
COMP_DIG_4	-0.00171671	PROD_A_6	-0.00858507	SUB_PROD_4	-0.01089212
COMP_DIG_5	0.00099715	PROD_A_7	0.07385837	SUB_PROD_5	-0.00085735
COMP_DIG_6	0.05706453	PROD_A_8	0.00140438	SUB_PROD_A_1	0.07012822
COMP_DIG_7	0.14510799	PROD_A_9	-0.01570135	SUB_PROD_A_2	0.01243225
COMP_DIG_8	0.12759481	PROD_B_1	-0.01582429	SUB_PROD_A_3	0.00373927
COMP_DIG_9	0.26725479	PROD_B_2	0.00392325	SUB_PROD_B_2	-0.00300838
COMP_TRAD_1	0.20305214	PROD_B_3	-0.01712622	SUB_PROD_B_3	-0.00276308
COMP_TRAD_2	0.02957281	PROD_DIG_1	0.16659614	SUB_PROD_B_4	0.00963808
PROD_A_1	-0.0036832	PROD_DIG_2	0.26832139	SUB_PROD_B_5	0.00023231

Fuente: elaboración propia.

Tomando el valor absoluto de las correlaciones identificadas se eligen aquellas que son menores a 0.0001 ya que no llevan correlación con la variable objetivo.

Tabla VIII. **Variables por eliminar del análisis por no tener una correlación significativa**

VARIABLES ELIMINADAS	Correlación
SUB_PROD_3	-0.00016912
SUB_PROD_B_5	0.00023231
SUB_PROD_5	-0.00085735
COMP_DIG_5	0.00099715

Fuente: elaboración propia.

2.12. Análisis descriptivo de las variables

Se generaron los estadísticos descriptivos para las 35 variables restantes, en la tabla se muestra a manera de ejemplo los datos calculados para cada una de las variables.

Tabla IX. **Análisis descriptivo de todas las variables en estudio**

PROD_A_1	PROD_A_2	PROD_A_3	PROD_A_4	PROD_A_5
Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.0000
Median :0.00000	Median :0.0000	Median :0.0000	Median :0.00000	Median :0.0000
Mean :0.01731	Mean :0.1415	Mean :0.1196	Mean :0.03457	Mean :0.1196
3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:0.0000
Max. :1.00000	Max. :7.0000	Max. :8.0000	Max. :6.00000	Max. :8.0000

SUB_PROD_2	SUB_PROD_B_2	SUB_PROD_A_2	SUB_PROD_4	SUB_PROD_B_3
Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.000000
1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.000000
Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.000000
Mean :0.01826	Mean :0.08433	Mean :0.01022	Mean :0.01006	Mean :0.002678
3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.000000
Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.000000

PROD_A_6	PROD_A_7	PROD_A_8	PROD_A_9	COMP_DIG_1
Min. :0.00000	Min. :0.00000	Min. :0.00000	00 Min. :0.00000	Min. :0.0000

Continuación tabla IX.

1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	00 1st Qu.:0.00000	1st Qu.:0.0000
Median :0.00000	Median :0.00000	Median :0.00000	00 Median :0.00000	Median :0.0000
Mean :0.01653	Mean :0.02861	Mean :0.00079	32 Mean :0.01777	Mean :0.1063
3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	00 3rd Qu.:0.00000	3rd Qu.:0.0000
Max. :3.00000	Max. :2.00000	Max. :2.00000	00 Max. :4.00000	Max. :1.0000

COMP_DIG_6	COMP_DIG_7	COMP_DIG_8	COMP_DIG_9	COMP_DIG_10
Min. :0.00000	Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.00000	Median :0.00000	Median :0.0000	Median :0.0000	Median :0.0000
Mean :0.00302	Mean :0.03179	Mean :0.1063	Mean :0.2389	Mean :0.2389
3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.0000
Max. :1.00000	Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :1.0000

SUB_PROD_1	PROD_DIG_1	SUB_PROD_A_3	PROD_DIG_2	COMP_DIG_4
Min. :0.0000	Min. :0.0000	Min. :0.000000	Min. :0.0000	Min. :0.000000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.000000	1st Qu.:0.0000	1st Qu.:0.000000
Median :0.0000	Median :1.0000	Median :0.000000	Median :0.0000	Median :0.000000
Mean :0.1111	Mean :0.6193	Mean :0.003901	Mean :0.3771	Mean :0.001521
3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:0.000000	3rd Qu.:1.0000	3rd Qu.:0.000000
Max. :1.0000	Max. :1.0000	Max. :4.000000	Max. :1.0000	Max. :1.000000

Fuente: elaboración propia.

También se evaluaron la varianza de todas las variables presentes en la base quedando de la siguiente manera:

Tabla X. **Varianza de variables por estudiar**

VARIABLE	Varianza
PROD_A_1	0.017
PROD_A_2	0.140
PROD_A_3	0.137
PROD_A_4	0.042
PROD_A_5	0.137
SUB_PROD_A_1	0.355
PROD_B_1	1.526
SUB_PROD_1	0.099
SUB_PROD_2	0.018
SUB_PROD_B_2	0.077
SUB_PROD_A_2	0.010
SUB_PROD_4	0.010
COMP_DIG_6	0.003
COMP_DIG_7	0.031
COMP_DIG_8	0.095
COMP_DIG_4	0.002
COMP_DIG_10	0.182
PROD_B_3	0.140

VARIABLE	Varianza
COMP_DIG_9	0.182
PRODUCTO_OBJETIVO	0.041
usuario	0.236
activo	0.235
SUB_PROD_B_3	0.003
SUB_PROD_B_4	0.133
PROD_B_2	0.332
SUB_PROD_A_3	0.005
PROD_A_6	0.018
PROD_A_7	0.028
PROD_A_8	0.001
PROD_A_9	0.020
COMP_DIG_1	0.095
COMP_DIG_2	0.102
COMP_DIG_3	0.102
COMP_TRAD_1	0.217
COMP_TRAD_3	0.020

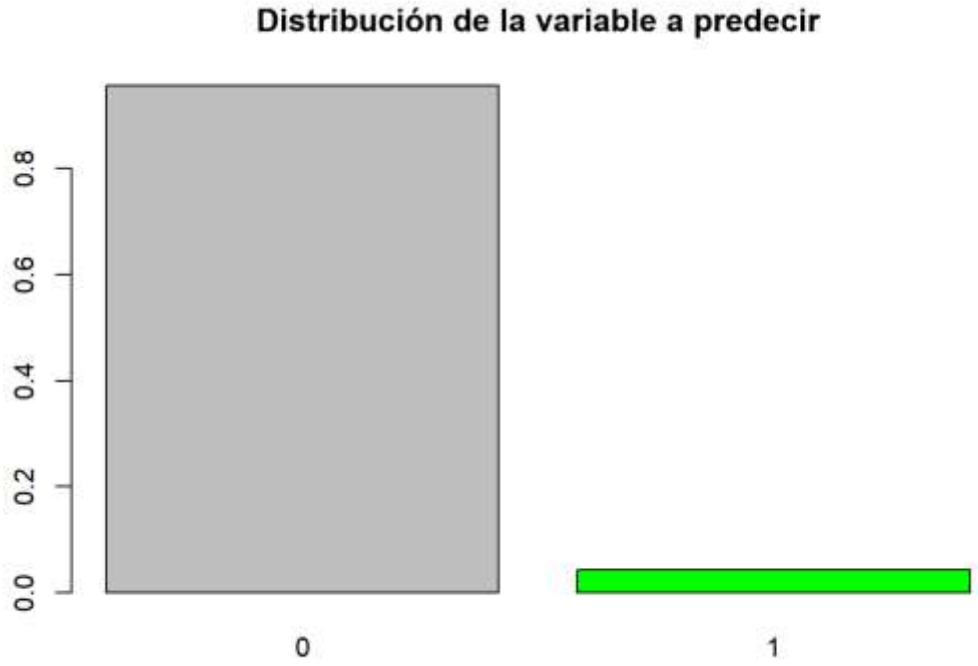
Fuente: elaboración propia.

Al realizar la interpretación de resultados tomando en cuenta la varianza se observó que casi todas las variables se encuentran entre 1 y 0, a diferencia de la variable PROD_B_1 que tiene un máximo de 44, sin embargo, al revisar la mediana 0 y la media 0.8249 se consideró que no es necesario escalarlas y centrarlas, adicional todas las variables son numéricas por lo que no habrá necesidad de transformarlas.

Para el análisis de la variable objetivo se evaluó la siguiente gráfica de frecuencia, y se pudo observar que no es un modelo balanceado ya que únicamente el 4.29 % de los clientes tienen el producto objetivo, por lo que se

decidió por un modelo de aprendizaje no supervisado, basado en un modelo de densidad de la variable objetivo dentro del segmento creado.

Figura 9. **Gráfica de frecuencia de la variable objetivo**

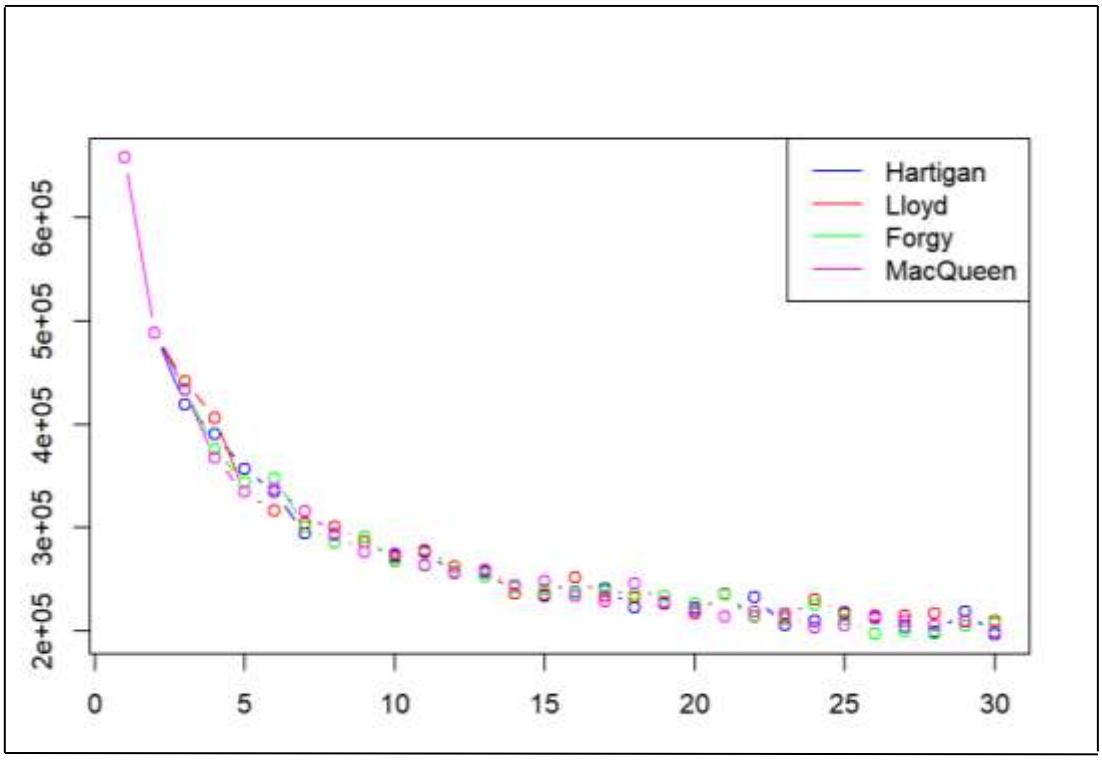


Fuente: elaboración propia.

2.13. Utilización codo de Jambú

Para establecer la cantidad de segmentos óptimos a utilizar en la segmentación se utilizó el programa R bajo el entorno RStudio, y con el comando *kmeans* se calculó la inercia intraclase variando la cantidad de segmentos de 1 en 1 hasta llegar a 30 segmentos, esto para los cuatro algoritmos, Hartigan-Wond, Lloyd, Forgy y MacQueen, los cuales se pueden observar en la siguiente tabla.

Figura 10. **Representación gráfica de la inercia intraclase cuando se varían los segmentos de 1 a 30**



Fuente: elaboración propia.

Al realizar una observación se establece que a partir de la realización de 6 segmentos la inercia comienza deja de disminuir considerablemente por lo que se decidió realizar 6 segmentos, bajo los preceptos que al seguir creciendo en segmentos se aumenta exponencialmente la necesidad de recursos de computación así como el tiempo que tarda en realizarse la segmentación y no es redituable, ya que la inercia dentro del segmento no disminuye considerablemente, es decir, los segmentos no serán tan homogéneos.

2.14. Determinación de algoritmo óptimo

Tomando en cuenta que ya se tiene la cantidad de segmentos óptimos por realizar se calcularon nuevamente los 4 algoritmos 25 veces cada uno y se calculó el promedio de cada algoritmo.

Tabla XI. **Promedio de los 25 cálculos por cada algoritmo para determinar el algoritmo óptimo**

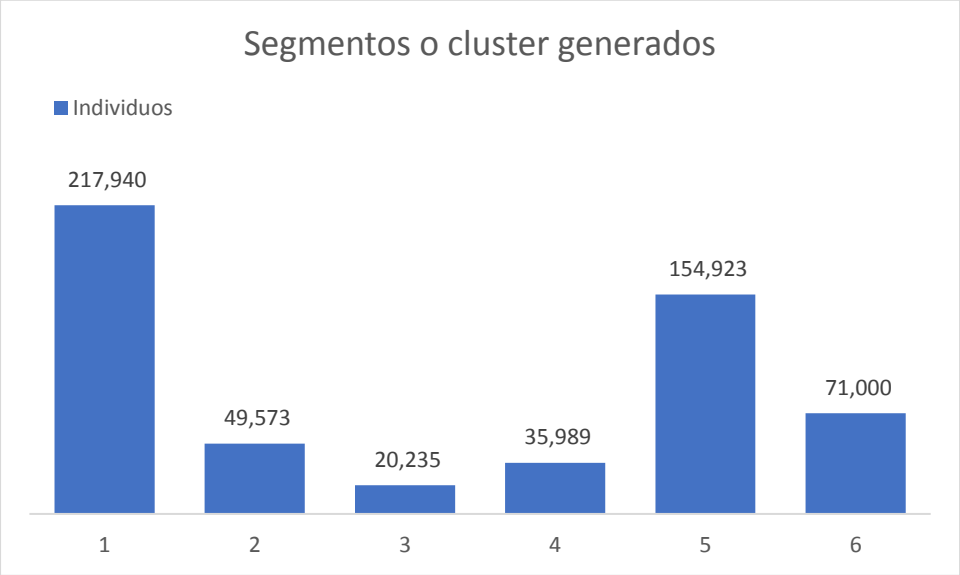
Algoritmo	Promedio de los 25 cálculos
HARTIGAN	275,853.40
LLOYD	271,575.40
FORGY	273,672.40
MacQueen	280,006.20

Fuente: elaboración propia.

2.15. Segmentación de base

Se procedió a realizar la segmentación tomando como base la cantidad de 6 segmentos óptimos y el algoritmo MacQueen. Estos se muestran en la figura 11.

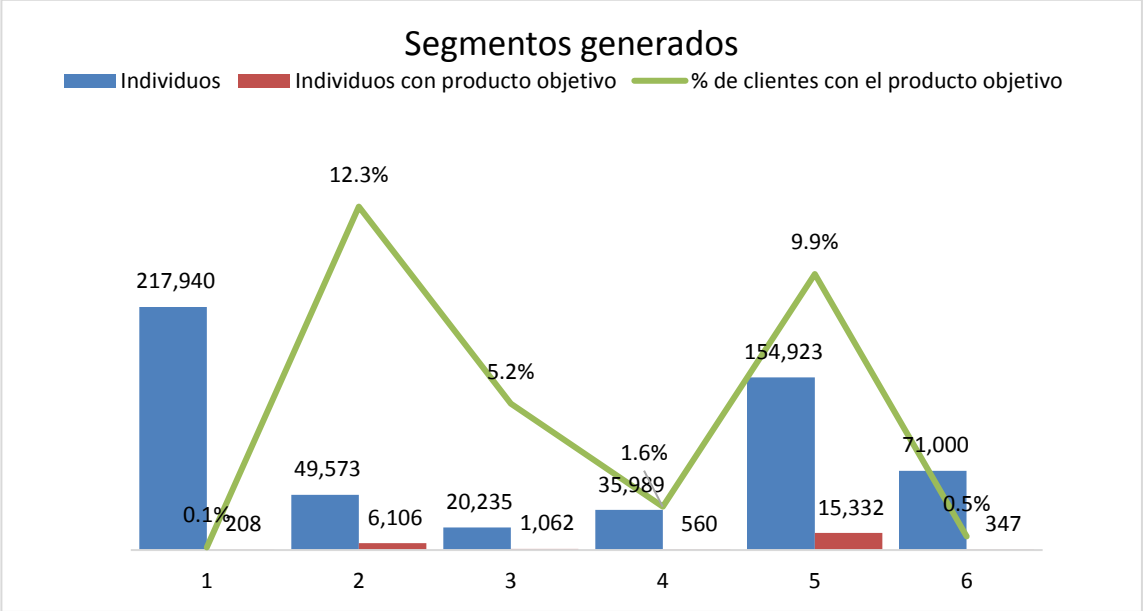
Figura 11. **Segmentos generados bajo el algoritmo de MacQueen**



Fuente: elaboración propia.

En general los segmentos quedaron disparejos unos de otros por lo que se procedió a revisar la densidad de la variable objetivo, es decir cuántos de los individuos en cada segmento tienen el producto objetivo.

Figura 12. **Densidad de la variable objetivo en los segmentos generados**



Fuente: elaboración propia.

Al evaluar la densidad de la variable objetivo se logró identificar que el segmento 2 el segmento 5 y el segmento 3 son aquellos segmentos a los cuales se tiene que priorizar un ofrecimiento ya que es donde más clientes tienen este producto, de la misma manera se puede descartar fácilmente el primer segmento sabes que para ofrecer el producto ya que únicamente el 0.1 % de los individuos en ese segmento tienen el producto, también se recomendó eliminar de la base de ofrecimiento el segmento cuatro y el segmento 6.

Objetivo 3: Determinar las ventajas que tiene la utilización de soluciones digitales en la segmentación de ventas en una empresa bancaria para los clientes existentes.

De acuerdo con el tercer objetivo planteado se presentan los resultados obtenidos.

2.16. Resultado

Se muestra la presentación de resultados de las ventajas que tiene la utilización de soluciones digitales en la segmentación de ventas

2.17. Ventas efectivas bajo segmentación propuesta

Ventas de productos nuevos por segmento creado bajo propuesta de segmentación.

Tabla XII. **Ventas de producto objetivo por segmentos**

Segmento	1	2	3	4	5	6
Individuos	217,940	49,573	20,235	35,989	154,923	71,000
Ventas de producto objetivo	73	535	109	72	1,443	45
% de ventas	0.03 %	1.08 %	0.54 %	0.20 %	0.93 %	0.06 %

Fuente: elaboración propia.

Los resultados de ventas por segmento muestran que aquellas que tuvieron un mejor desempeño porcentualmente es el segmento 2 el segmento 3 y el segmento 5 el resto de los segmentos tuvieron ventas, sin embargo, porcentualmente el desempeño es mucho más bajo el desempeño que más se acerca es el segmento cuatro siendo menos de la mitad de alcance que el segmento más bajo recomendado que en este caso es el segmento 3.

Objetivo general: proponer un proceso de segmentación para la venta de productos digitales a clientes existentes en una institución bancaria de la ciudad de Guatemala.

Luego de haber realizado la segmentación propuesta y habiendo priorizado las bases para mostrar anuncios aquellos clientes que se encuentran en segmentos que tienen una mayor concentración de producto se llegó a la colocación de 2,277 productos nuevos con la misma inversión en pauta publicitaria de $ 1,250.00 esto representa un aumento en la colocación de ventas del 41 %, el costo promedio por cada venta bajo el método anterior el cual no era una segmentación sino simplemente se dejaba de mostrar el anuncio al acabarse el crédito establecido es de $0.71 mientras que bajo la segmentación estudiada el nuevo costo es de $0.41.

3. DISCUSIÓN DE RESULTADOS

Con el uso de la técnica de segmentación para obtener grupos de clientes con una mayor densidad o penetración de la variable objetivo se logró mejorar los índices de alcance en ventas por medio de medios digitales, los cuales se muestran a continuación.

3.1. Análisis interno

Al revisar los resultados obtenidos se comprueba la efectividad de la implementación de la segmentación propuesta para el ofrecimiento de productos digitales a clientes de una entidad bancaria. tomando en cuenta que se obtuvo una mejora del 41 % en las ventas totales que se tuvieron, cuando se evalúa la efectividad bajo el proceso anterior el mes que mejores resultados se tuvieron porcentualmente fue abril y agosto llegando casi al 0.90 % sin embargo revisando la efectividad del segmento 2 para el mes de septiembre se obtiene una efectividad del 1.08% así como del segmento 5 el cual se tiene una efectividad del 0.93 %. esta efectividad es mayor que es la mejor efectividad obtenida durante los últimos 5 meses en los cuales se realizó el ofrecimiento sin la segmentación propuesta.

Para la solución de la investigación se cumplió con las cuatro fases establecidas, la primera fase se refiere a la revisión documental donde se revisó toda la documentación relacionada al tema de investigación en fuentes primarias secundarias y terciarias, llegando a estudiar hasta 5 meses atrás de resultados.

En la segunda fase, se logró responder el primer objetivo de la investigación, que era investigar e identificar la forma en que la empresa hace la segmentación de clientes existentes analizando datos como el gasto en publicidad, el alcance que se tuvo al público objetivo y las ventas totales para lo cual se revisaron las bases de los últimos 5 meses del año y se evaluaron los resultados obtenidos durante cada mes.

En la tercera fase, se logró responder el segundo objetivo de la investigación, analizar los factores críticos para la solución de la segmentación en ventas al no contar con recursos digitales, se estudiaron 39 variables de las cuales se lograron eliminar cuatro por haberse demostrado que no aportaban al modelo de segmentación al no contar con una correlación específica con la variable estudiada o variable objetivo, así como la cantidad de segmentos óptimos a realizar y el algoritmo óptimo que representa una mayor homogeneidad en los segmentos realizados, esto luego de haberse demostrado que el método a realizar es un aprendizaje no supervisado dado que la variable objetivo no era una variable balanceada en el universo tomando en cuenta que el 4 % sí tenía el producto y el 96% no.

La cuarta fase, respondió al tercer objetivo. para ello se realizó un ofrecimiento del producto priorizando aquellos segmentos que se establecieron tenían una mayor densidad de la variable o producto objetivo, esperando mejores resultados que sin la segmentación demostrada comparando el segmento 1 con un alcance de tan solo 0.03% versus el alcance en el segmento 2 Qué fue de 1.03 % considerablemente mayor.

En cuanto al procedimiento de segmentación se logró, no sólo mejorar los resultados, sino que se cuidó la optimización de los recursos utilizando la menor cantidad de segmentos necesarios, así como de las variables a utilizar evitando

con esto derrochar o sobrecargar el recurso computacional disponible, lo cual favorece el proceso no sólo del equipo de ventas cine sino del equipo que generará las bases.

3.2. Análisis externo de la investigación

En cuanto a el mayor porcentaje de alcance que se logró mediante segmentar las bases se logró aumentar en un 41 % la cantidad total de ventas que se realizaron por medio de canales digitales. Se logró una disminución importante de $ 0.23 por cada venta realizada lo cual permite tener un mayor alcance y mejorar la rentabilidad del producto a largo plazo.

Durante el desarrollo de la investigación se tuvieron varios aspectos positivos entre ellos el apoyo y respaldo de la empresa para la realización del trabajo investigativo, el compromiso del equipo involucrado para aportar mejoras en cuanto al proceso de segmentación i el conocimiento adquirido por medio de la experiencia que tenía en cada uno de los integrantes del equipo.

Entre los aspectos negativos, se tuvo la dificultad de obtener más registros allá de los 5 meses que se entregan en este documento ya que habría aportado mucho más el tener mayor tiempo de registros, también las dificultades para escalar los servidores y de esta manera poder ampliarlo las iteraciones realizadas y tener mejores segmentaciones en un tiempo más adecuado.

Según lo estudiado por Fortune (2020) acerca de los polígonos de Thiessen, que eran utilizados para poder construir sobre el plano euclídeo una pequeña partición a partir de una construcción geométrica de la misma manera se realizó la construcción de los segmentos en los clientes, lo que permitió determinar aquellos segmentos óptimos para realizar ofrecimientos dado que

tendrían una mayor aceptación esto asumido porque la variable objetivo se encontraba en mayor densidad en estos segmentos mientras que se logró eliminar o bajar de prioridad en el ofrecimiento aquellos segmentos que mostraban una baja densidad de la variable objetivo.

Al igual que en el estudio de Moreno-Seco, (2004) en su publicación sobre la clasificación del vecino más cercano se toma la distancia de sus vecinos en el conjunto estudiado y como se mencionó en dicho estudio, gracias a las computadoras de alto rendimiento se puede procesar una mayor cantidad de datos no sólo de variables sino también objetos, en el presente estudio se evaluaron cuatro algoritmos distintos siendo el algoritmo MacQueen el escogido esto luego de haber evaluado una segmentación variable desde uno a 30 en cada uno de los cuatro algoritmo para escoger la cantidad de segmentos óptimos y haber evaluado cada uno de los cuatro algoritmos 25 veces para escoger el que mejor resultado daba, evaluando siempre el universo completo obtenido siento este de 549,660 individuos y 35 variables distintas por cada uno de estos individuos, lo que demuestra el avance de la tecnología ya que hace un par de años esto hubiera sido prácticamente imposible con las capacidades de computación disponibles en su tiempo o hubiera tomado demasiado tiempo tanto para hacer práctica la realización de la segmentación.

Realizando una segmentación de clientes o audiencias como lo menciona Tapis, (2001), se dividieron grupos potenciales de individuos e identificando los a cuál grupo pertenece sí fueron segmentando categorías para crear una campaña personalizada y esto permitió pasar de una efectividad de 0.78 % en promedio durante los primeros 5 meses estudiados a una efectividad de hasta el 1.08 % vista en el segmento 2 lo que representa 30 puntos base más que con el método anterior, aumentando en un 41 % los resultados de la campaña de venta

u ofrecimiento utilizando exactamente los mismos recursos que se utilizaron los 5 meses anteriores.

CONCLUSIONES

1. Se implementó proceso de segmentación mediante el análisis de los comportamientos del cliente, utilizando técnicas de clusterización que identifican segmentos apropiados para el ofrecimiento del producto objetivo a clientes dentro de una institución bancaria de la ciudad de Guatemala.

2. Se estableció que la forma idónea para segmentar adecuadamente a los clientes existentes es utilizar 35 variables de comportamiento dentro de la institución bancaria mediante la utilización de técnicas estadísticas que comprueban correlación directa con el producto objetivo a ofrecer.

3. Se comprobó que realizando el análisis mensual o trimestral permite tener un desempeño superior y un alcance del 100 % de la base segmentada mediante la identificación de la densidad del producto en los clústeres existentes, optimizando el presupuesto asignado a cada campaña específica que resuelve el problema de ofrecer productos a clientes que no son parte del público objetivo.

4. Se determinó que utilizando la segmentación aplicada incrementó la productividad de la campaña al priorizar el ofrecimiento en el grupo objetivo que tenía una mayor probabilidad, mejorando de esta manera las ventas en un 41 % y disminuyendo el costo por venta de $ 0.77 a $0.41. La información para los resultados se obtuvo dividiendo la cantidad del presupuesto asignado a la campaña publicitaria sobre la cantidad de ventas efectivas realizadas al finalizar dicha campaña.

RECOMENDACIONES

1. Evaluar periódicamente el uso de nuevas variables y realizar el proceso de correlacionarlas para determinar si es conveniente el ir aumentando variables que aporten en el proceso de segmentación como una búsqueda de ir mejorando constantemente el proceso establecido.

2. Realizar el análisis del algoritmo cuando se cambie la variable objetivo, para garantizar el óptimo resultado. esto luego de haber incluido o eliminado las variables que aporten al modelo de segmentación.

3. Evaluar los resultados obtenidos en ventas en función del dinero asignado a cada campaña para establecer un costo por venta y determinar correctamente cuál es la mejora económica de utilizar la segmentación propuesta i todas aquellas mejoras que se vayan realizando al modelo para garantizar que se continúa mejorando en el tiempo.

4. Realizar una segmentación en un período de uno a tres meses máximo para mantener el modelo en óptimas condiciones, aunque aún no se hayan definido nuevas variables o un cambio de algoritmo la actualización de las variables existentes determinará siempre los segmentos a los cuales un individuo pueda pertenecer esto garantizará una adecuada asignación de recursos a las campañas por cada uno de los objetivos establecidos.

REFERENCIAS

1. Andrieu, C., de Freitas, N., Doucet, A., y Jordan, M. (10 de septiembre de 2001). *An Introduction to MCMC for Machine Learning. Machine Learning.* [Mensaje en un blog]. Recuperado de e http://www.cs.ubc.ca/~nando/papers/mlintro.pdf.

2. Asensi, E., Boronat, L., Tomás, D. y Vicedo, J. (2006*). Desarrollo de un corpus de entrenamiento para sistemas de búsqueda de respuestas basados en aprendizaje automático.* España: Sociedad Española para el Procesamiento del Lenguaje Natural. Recuperado de http://sepln.org/revistasepln/revista/37/08.pdf.

3. Bernal, D. (4 de julio de 2017). *Ajuste de particiones planas mediante diagramas de Voronoi discretos*. [Mensaje en un blog]. Recuperado de http://oa.upm.es/47124.

4. Bringas, P., Lopez, I. y Grueiro, I. (marzo 2014). Visión artificial basada en aprendizaje automático para la categorización de defectos superficiales en fundición. *Dyna,* 89(3), 325-332. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=4677519.

5. Díaz, C., Morales, E. y Pérez, C. (septiembre 2009). Support vector machine model for regression applied to the estimation of the creep ruptura stress in ferritic steels. *Revista Facultad de Ingenieria-Universidad de Antioquia.* (47), 53-58. Recuperado de

http://scielo.org.co/ scielo.php?script=sci_arttext&pid=s0120-62302009000100005.

6. Elvira, J., Cívico, F., Cabrera, R., Osuna, M., Cabrera, J. y Olivares, R. (agosto 2017). Actividades extraescolares y rendimiento académico en alumnos de Educación Secundaria. *Electronic Journal of Research in Educational Psychology,* 4(8), 35-46. Recuperado de http://investigacion-psicopedagogica.org/revista/articulos/8/espannol/art_8_82.pdf.

7. Figuera, D. (12 de enero de 2004). *Métodos cuantitativos para la toma de decisiones*. [Mensaje en un blog]. Recuperado de http://exa.unne.edu.ar/ informatica/evalua_ant/metodos cuatitativo.pdf.

8. Gonzalez, D. (20 de mayo de 2011). *Algoritmos de clustering paralelos en sistemas de recuperación de información distribuidos*. [Mensaje en un blog]. Recuperado de https://riunet.upv.es/handle/10251/11234.

9. Hernández, A. (12 de septiembre de 2019). *Online Clustering con STREAMING K-MEANS usando SPARK Streaming*. [Mensaje en un blog]. Recuperado de http://oa.upm.es/56397.

10. Illescas, P., Rizo, D., Iñesta, J. y Ramirez, R. (2011). *Learning melodic analysis rules*. España: Universitat Pompeu Fabra. Recuperado de http:// grfia.dlsi.ua.es/repositori/grfia/pubs/276/mml2011-melan-final.pdf.

11. Mecánico, I. y Diaz, A. (2006). *Selección de un servomotor y transmisión por el método de las potencias transitorias*. España: Universidad de La Rioja. Recuperado de https://dialnet.unirioja.es/descarga/articulo/ 4832259.pdf.

12. Moreno, F. (2004). *Clasificadores eficases basados en algoritmos rápidos de busqueda del vecino más cercano*. (Tesis de doctorado). Universidad de Alicante, España. Recuperado de https://rua.ua.es/dspace/bitstream/10045/11790/1/Moreno-Seco-Francisco.pdf

13. Núñez, E., Steyerberg, E. y Núñez, J. (junio 2011). Estrategias para la elaboración de modelos estadísticos de regresión. *Revista Espanola de Cardiologia,* 64(6), 501-507. Recuperado de https://revespcardiol.org/es-estrategias-elaboracion-modelos-estadisticos-regresion-articulo-s0300893211003502.

14. Ortega, M. (2008). *Kermelized graph matching and clustering*. España: Universidad de La Rioja. Recuperado de https://dialnet.unirioja.es/servlet/tesis?codigo=21378.

15. Quevedo, F. (marzo 2011). Medidas de tendencia central y dispersión. *Medwave, 11*(03). Recuperado de https://medwave.cl/link.cgi/medwave/ series/mbe04/4934.

16. Roca, J., Bueno, A. y Sancho, J. (mayo 2013). Exploiting Diversity of Neural Network Ensembles based on Extreme Learning Machine. *Neural Network World, 23*(5), 395-409. Recuperado de http://nnw.cz/obsahy13.html.

17. Rubio, A., Fernández, V., Unanue, R., y Herranz, S. (marzo 2005). Evaluación del clustering de páginas web mediante funciones de peso y combinación heurística de criterios. *Procesamiento Del Lenguaje Natural,* 35(35), 417-424. Recuerado de http://sepln.org/revistasepln/revista/35/51.pdf.

18. Sánchez, A., Fernández, F., Valero, C., Muñoz, M., Rodríguez, A. López, y Espejo, I. (14 de enero de 2009). *Estadística Descriptiva y Probabilidad: (Teoría y problemas)*. [Mensaje en un blog]. Recuperado de https://libros.metabiblioteca.org/handle/001/140.

19. Steve Fortune (junio de 1986). Numerical stability of algorithms for line arrangements*. Actas del séptimo simposio anual sobre geometría computacional,* 1(1), 334-341. Recuperado de https://dl.acm.org/doi/10.1145/109648.109685

20. Tapis, X. (agosto 2001). El marketing aplicado a las ONGD: coherencias e incoherencias en relación con la educación para el desarrollo. *Comunicar,* 8(16), 103-114. Recuperado de https://dialnet.unirioja.es/descarga/articulo/185291.pdf.

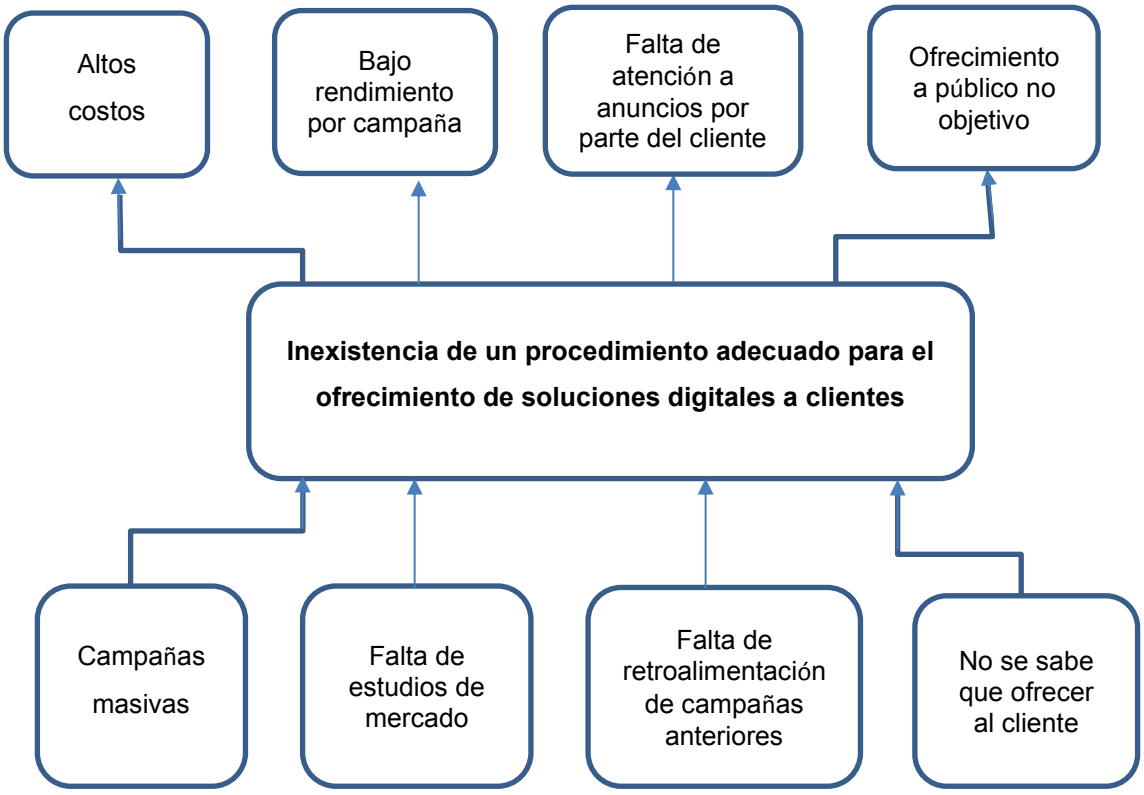
APÉNDICES

Apéndice 1. **Matriz de coherencia**

Formulacion del problema	Objetivos	Hipotesis	Variables	Indicadores	Tecnicas e intrumentos	Metodologia
La inexistencia de un procedimiento adecuado para el ofrecimiento de soluciones digitales a clientes acrecentando los costos de obtención de clientes	General: Proponer un proceso de segmentación para la venta de productos digitales a clientes existentes en una institución bancaria Identificar la forma en que la empresa hace la segmentación de clientes existentes Determinar las ventajas que tiene la utilización de soluciones digitales en la segmentación de ventas en una empr3esa bancaria para los clientes existentes	La propuesta de un nuevo proceso de segmentación para la venta de productos digitales a clientes existentes aumentara la mejora de los resultados y la rentabilidad del área	Cuales son las mejoras caracterísiticas a tomar en cuenta para evaluar una correcta segmentación Tipos de productos a tomar en cuenta en la evaluacion para ofrecimiento	Cantidad de produtos aceptados versus cantidad de productos ofrecidos Cantidad de productos vendidos por tipo vrs total de productos vendidos.	Los polígonos de Thiessen Diagramas de Voronoi	**Metodo de investigación** El metodo de investigación es descriptivo **Diseño de investigacion** El diseño correlacional

Fuente: elaboración propia.

Apéndice 2. **Árbol de problemas**



Fuente: elaboración propia.

Apéndice 3. **Código R utilizado para generar segmentación**

```
#Librerias a utilizar

library(readr)
library(tidyverse)
library("FactoMineR")
library("factoextra")
library("corrplot")
library(kknn)
library("dplyr")
#Carga de base a trabajar en formato CSV
baseinicial <- read_csv("C:/universo.csv")
universo<-baseinicial

#Identificar tamaño de la base

tam<-dim(baseinicial)
n<-tam[1]
n
#Se buscan correlaciones

correl<-baseinicial
#REEMPLAZA LOS 'NA' DEL PRODUCTO OBJETIVO POR 0
correl <- mutate_all(correl, ~replace(., is.na(.), 0))
# Se escala y centra la data
correl <- as.data.frame(scale(correl[,1:39]))
correlaciones <- as.data.frame(cor(correl))
#revisando correlación se eliminan aquellas que tengan menos de 0.0009 de
correlación
universo$SUB_PROD_3 <- NULL
universo$SUB_PROD_B_5 <- NULL
universo$SUB_PROD_5 <- NULL
universo$COMP_DIG_5 <- NULL
#REEMPLAZA LOS 'NA' DEL PRODUCTO OBJETIVO POR 0
universo <- mutate_all(universo, ~replace(., is.na(.), 0))
#REVISION LECTURA CORRECTA DE BASE
summary(universo)
str(universo)
#REVISIÓN DE LA VARIANZA
apply(universo,2,var)
# DISTRIBUCION DE LA VARIABLE A PREDECIR
barplot(prop.table(table(universo$PRODUCTO_OBJETIVO)),col=c("gray","gree
n"),main="Distribución de la variable a predecir")
# Fabricar codo de Jambu para evaluar K óptimo
```

Continuación apéndice 3.

```
InerciaIC.Hartigan = rep(0, 30)
InerciaIC.Lloyd = rep(0, 30)
InerciaIC.Forgy = rep(0, 30)
InerciaIC.MacQueen = rep(0, 30)
for (k in 1:30) {
  grupos = kmeans(universo, k, iter.max = 100, algorithm = "Hartigan-Wong")
  InerciaIC.Hartigan[k] = grupos$tot.withinss
  grupos = kmeans(universo, k, iter.max = 100, algorithm = "Lloyd")
  InerciaIC.Lloyd[k] = grupos$tot.withinss
  grupos = kmeans(universo, k, iter.max = 100, algorithm = "Forgy")
  InerciaIC.Forgy[k] = grupos$tot.withinss
  grupos = kmeans(universo, k, iter.max = 100, algorithm = "MacQueen")
  InerciaIC.MacQueen[k] = grupos$tot.withinss
}
plot(InerciaIC.Hartigan, col = "blue", type = "b")
points(InerciaIC.Lloyd, col = "red", type = "b")
points(InerciaIC.Forgy, col = "green", type = "b")
points(InerciaIC.MacQueen, col = "magenta", type = "b")
legend("topright", legend = c("Hartigan", "Lloyd", "Forgy", "MacQueen"), col =
c("blue",                                                    "red", "green",
"magenta"), lty = 1, lwd = 1)
#Se utilizará K=6 ya que es donde se estabiliza el codo de Jambu
# Se evalúa cual será el algoritmo que mejor se desempeñe
# Dado que cambia se correo 25 veces y se promedia para estar seguros de
cual es el óptimo
Hartigan <- 0
Lloyd <- 0
Forgy <- 0
MacQueen <- 0
for (i in 1:25) {
  grupos <- kmeans(universo, 4, iter.max = 100, algorithm = "Hartigan-Wong")
  Hartigan <- Hartigan + grupos$betweenss
  grupos <- kmeans(universo, 4, iter.max = 100, algorithm = "Lloyd")
  Lloyd <- Lloyd + grupos$betweenss
  grupos <- kmeans(universo, 4, iter.max = 100, algorithm = "Forgy")
  Forgy <- Forgy + grupos$betweenss
  grupos <- kmeans(universo, 4, iter.max = 100, algorithm = "MacQueen")
  MacQueen <- MacQueen + grupos$betweenss
}
Hartigan/25
Lloyd/25
Forgy/25
```

Continuación apéndice 3.

```
MacQueen/25
#Se utiliza el algoritmo "Lloyd" es el que mejores resultados muestra

#Se prepara toda la data
universof <- baseinicial
universof$No <- NULL
universof$PRODUCTO_OBJETIVO_CANTIDAD <- NULL
universof$ID_A <- NULL
universof$ID_B <- NULL
universof$us_a <- NULL
universof$us_b <- NULL
universof$act_a <- NULL
universof$COMP_TRAD_2 <- NULL
universof$act_b <- NULL
universof$COMPRA_A <- NULL
universof$COMPRA_B <- NULL
universof$compra_actual_a <- NULL
universof$compra_actual_b <- NULL
universof$compra_actual_a <- NULL
universof$Producto_final <- NULL
universof$SUB_PROD_3 <- NULL
universof$SUB_PROD_B_5 <- NULL
universof$SUB_PROD_5 <- NULL
universof$COMP_DIG_5 <- NULL
#REEMPLAZA LOS 'NA' DEL PRODUCTO OBJETIVO POR 0
universof <- mutate_all(universof, ~replace(., is.na(.), 0))
clustering_universo <- kmeans(universof,centers = 6,iter.max = 100,algorithm =
"Lloyd")
names(clustering_universo)
#validación visual de congruencia
clustering_universo$cluster
#Revisión de resultados
#plot(clustering_universo$activo,clustering_universo$comp_dig_10, col =
clustering_universo$cluster,
 #    xlab = "Digital", ylab = "Comportamiento digital" )
final<-cbind(baseinicial,clustering_universo$cluster)
#write.csv(final,file="clusterizacion.csv")
```

Fuente: elaboración propia, empleando Rstudio.